UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Irony and Sarcasm Detection in Twitter: The Role of Affective Content

PhD Candidate

Delia Irazú Hernández Farías

Thesis Advisors

Paolo Rosso
Universitat Politècnica de València, Spain

Viviana Patti
Università degli Studi di Torino, Italy

Valencia, September 2017

UNIVERSITÀ DEGLI STUDI DI TORINO

Dipartimento di Informatica

Dottorato di ricerca in Informatica
Ciclo XXX

Irony and Sarcasm Detection in Twitter: The Role of Affective Content

Tesi presentata da:
Delia Irazú Hernández Farías

Tutors:
Paolo Rosso
Universitat Politècnica de València, Spain
Viviana Patti
Università degli Studi di Torino, Italy

Coordinatore del dottorato:
Marco Grangetto

Settembre 2017

Settore scientifico-disciplinare di afferenza: INF/01

Acknowledgments

I would like to like to express my most sincere gratitude to all of those who have made this work possible.

Firstly, to my advisors: Paolo Rosso and Viviana Patti, without their help, it would not have been possible to conclude this thesis. Thanks a lot for all the time dedicated to our interesting and fascinating research topic #withoutsarcasm :D.

Paolo, thank you for all the opportunities you have given me since more than five years ago. Thanks a lot for encouraging me to be a better PhD student and also for all your advice and patience. I am very thankful for your help and support during these years. I just want to say this in all the languages I speak: thank you! grazie! gracias!

Viviana, thank you so much for all the support and help that you've given me. I really appreciate that you have made me collaborate in different projects. Thank you for inviting me to spend part of my PhD in a beautiful city such as Torino (giving to me the opportunity to learn a new language: Italian). Sinceramente, Grazie mille!

I'm really thankful to the reviewers of this thesis: Rachel Giora, Horacio Saggion, and Pavel Braslavski; thanks for your valuable comments about my thesis. Thank you very much to the members of the evaluation tribunal of this thesis: Horacio Saggion, Elisabetta Fersini, and Roberto Basili.

Thank you so much to Universitat Politècnica de València (UPV) and Università degli Studi di Torino (UniTo) for all the facilities and support provided to me. And also to the people in the Pattern Recognition and Human Language Technology (PRHLT) research center.

Thanks to all the people from different countries and cultures that shared some time in the laboratory at UPV with me. A special mention is for Maite: thank you so much for the time and experiences we share during this period: moltes gràcies!

I also want to say GRAZIE to the people at UniTo, especially to Emilio Sulis, Cristina Bosco, and Mirko Lai (who learned to speak his own version of Spanish with me).

Thank you to all the people who have shared not only good (also bad and stressful) moments but also their lives with me in Valencia and Torino.

Thanks to my grandfather, aunts, cousins, and friends in Mexico for always having words of encouragement for me.

Last but not least, I would say thank you to the most important people in my life: my mom and my brother. Thank you for being always there supporting, helping, and encouraging me no matter the distance. Mami: Thank you so much for taking care of us and also for always having a smile even in rough times.

Delia Irazú Hernández Farías
València, July 2017.

Funding

This work has been funded by the National Council for Science and Technology (CONACyT - Mexico) with the Grant No. 218109/313683. Part of the research was carried out in the framework of the SomEMBED TIN2015-71147-C2-1-P MINECO project.

Abstract

Investigating how people express themselves in social media has attracted the attention of several disciplines due to the great potential for research that it represents. Social media platforms, like Twitter, offer a face-saving ability that allows users to express themselves employing figurative language devices such as irony to achieve different communication purposes. Ironic utterances in such platforms are generated by users that most of the time have only an intuitive definition of what irony is. Dealing with such kind of content represents a big challenge for computational linguistics. Irony is closely associated with the indirect expression of feelings, emotions and evaluations, intended as the writer's attitude or stance towards a particular target entity involved in the ironic utterance. Thus, interest in detecting the presence of irony in social media texts has grown significantly in the recent years, also for the impact on natural language processing (NLP) areas related to sentiment analysis, where irony detection is important to avoid misinterpreting ironic statements as literal.

In this thesis, we introduce the problem of detecting irony in social media under a computational linguistics perspective. We propose to address this task by focusing, in particular, on the role of affective information for detecting the presence of such figurative language device. Attempting to take advantage of the subjective intrinsic value enclosed in ironic expressions, we present a novel model, called emotIDM, for detecting irony relying on a wide range of affective features. For characterising an ironic utterance, we used an extensive set of resources covering different facets of affect from sentiment to finer-grained emotions. We address irony detection by casting it as a binary classification problem. To evaluate our model, we collected a set of Twitter corpora used by scholars in previous research, to be used as benchmarks with a two-fold purpose: to compare the performance of our model against other approaches in the state of the art, and to evaluate its robustness across several different aspects related to the characteristics of the corpora, such as collection mode, size and imbalance degree. Results show that emotIDM has a competitive performance across the experiments carried out, validating the effectiveness of the proposed approach. In most cases, our outcomes outperform those from the related work confirming that affective information helps in distinguishing between ironic and non-ironic tweets. Another objective of the thesis is to investigate the differences among tweets labeled with #irony and #sarcasm. Our aim is to contribute to the less investigated topic in computational linguistics on the separation between irony and sarcasm in social media, again, with a special focus on affective features. We also studied a less explored hashtag that has been used by scholars for collecting samples of sarcastic intention: #not. We find data-driven arguments on the differences among tweets containing

these hashtags, suggesting that the above mentioned hashtags are used to refer different figurative language devices. We identify promising features based on affect-related phenomena for discriminating among different kinds of figurative language devices and our classification results outperform the state of the art. We also analyse the role of polarity reversal in tweets containing ironic hashtags, observing that the impact of such phenomenon varies. In the case of tweets labeled with #sarcasm often there is a full reversal (varying from a polarity to its opposite, almost always from positive to negative polarity), whereas in the case of those tagged with #irony there is an attenuation of the polarity (mostly from negative to neutral).

Detecting irony in user-generated content could have a broad range of applications. Undoubtedly, one of the areas that can benefit most from irony detection is sentiment analysis. We analyse the impact of irony and sarcasm on sentiment analysis, observing a drop in the performance of NLP systems developed for this task when irony is present. Therefore, we explored the possible use of our findings in irony detection for the development of an irony-aware sentiment analysis system, assuming that the identification of ironic content could help to improve the correct identification of sentiment polarity. To this aim, we incorporated emotIDM into a pipeline for determining the polarity of a given Twitter message. We compared our results with the state of the art determined by the 'Semeval-2015 Task 11: Sentiment Analysis of Figurative Language in Twitter' shared task, demonstrating the relevance of considering affective information together with features alerting on the presence of irony for performing sentiment analysis of figurative language for this kind of social media texts. To summarize, we demonstrated the usefulness of exploiting different facets of affective information for dealing with the presence of irony in Twitter.

Resumen

La manera en la cual se expresan los usuarios en redes sociales ha atraído la atención de varias disciplinas debido al gran potencial de investigación que esta representa. Las plataformas de redes sociales, como Twitter, ofrecen a los usuarios la posibilidad de expresarse de forma libre y espontanea haciendo uso de diferentes recursos lingüísticos como la ironía para lograr diferentes propósitos de comunicación. Las expresiones irónicas en dichas plataformas son generadas por usuarios que, la mayoría de las veces, tienen solamente una definición intuitiva de lo que es la ironía. Manejar ese tipo de contenido representa un gran reto para la lingüística computacional. La ironía está estrechamente vinculada con la expresión indirecta de sentimientos, emociones y evaluaciones (entendidas como la actitud o postura del autor hacia la entidad específica involucrada en la expresión irónica). Por consiguiente, el interés en detectar la presencia de ironía en textos de redes sociales ha aumentado significativamente en los últimos años, también por el impacto en áreas del procesamiento de lenguaje natural (PLN) relacionadas con el análisis de sentimientos, donde detectar la presencia de la ironía es importante para evitar malinterpretar expresiones irónicas como literales.

En esta tesis, introducimos el problema de detección de ironía en redes sociales desde una perspectiva de la lingüística computacional. Proponemos abordar dicha tarea enfocándonos, particularmente, en el rol de información relativa al afecto y las emociones para detectar la presencia de dicho recurso lingüístico. Con la intención de aprovechar el valor intrínseco de subjetividad contenido en las expresiones irónicas, presentamos un modelo para detectar la presencia de ironía denominado emotIDM, el cual está basado en una amplia variedad de rasgos afectivos. Para caracterizar instancias irónicas, utilizamos un amplio conjunto de recursos que cubren diferentes ámbitos afectivos: desde sentimientos (positivos o negativos) hasta emociones específicas definidas con una granularidad fina. Abordamos la detección de ironía interpretándola como un problema de clasificación binaria. Para evaluar nuestro modelo, recopilamos un conjunto de datos de Twitter previamente utilizados en trabajos relacionados a la detección de ironía teniendo en cuenta dos objetivos: por un lado, comparar el desempeño de nuestro modelo con otros enfoques en el estado del arte, y por el otro, evaluar su robustez considerando varios aspectos diferentes, los cuales están relacionados con las características de los conjuntos de datos como modalidad de obtención, tamaño y grado de desbalance entre clases. Los resultados obtenidos muestran que emotIDM tiene un desempeño competitivo en los experimentos realizados, validando la efectividad del enfoque propuesto. En la mayoría de los casos, nuestros resultados superaron el estado del arte confirmando que la información relativa al afecto y las emociones ayuda para distinguir entre tweets irónicos y no irónicos.

Otro objetivo de la tesis es investigar las diferencias entre tweets etiquetados con #irony y #sarcasm. Nuestra finalidad es contribuir a un tema menos investigado en lingüística computacional: la separación entre el uso de ironía y sarcasmo en redes sociales, con especial énfasis en rasgos afectivos. Además, estudiamos un hashtag que ha sido menos analizado: #not, el cual ha sido utilizado en investigaciones relacionadas a la detección de contenidos irónicos para recopilar datos que expresan una intención sarcástica. Nuestros resultados parecen evidenciar que existen diferencias entre los tweets que contienen dichos hashtags, sugiriendo que son utilizados para hacer referencia de diferentes recursos lingüísticos. Identificamos un conjunto de características basadas en diferentes fenómenos afectivos que parecen ser útiles para discriminar entre diferentes tipos de recursos lingüísticos, además nuestros resultados superaron aquellos en el estado del arte. Adicionalmente analizamos la reversión de polaridad en tweets que contienen hashtags irónicos, observamos que el impacto de dicho fenómeno es diferente en cada uno de ellos. En el caso de los tweets que están etiquetados con el hashtag #sarcasm, a menudo hay una reversión total (que va desde una polaridad a su opuesta, casi siempre de un sentimiento positivo a uno negativo), mientras que en el caso de los tweets etiquetados con el hashtag #irony se produce una atenuación de la polaridad (principalmente de negativo a neutral).

Detectar ironía en contenido generado por usuarios podría tener un amplio rango de aplicaciones. Sin duda alguna, una de las áreas que puede resultar más beneficiada por la detección de ironía es el análisis de sentimientos. Llevamos a cabo un estudio del impacto de la ironía y el sarcasmo en el análisis de sentimientos, observamos una disminución en el rendimiento de los sistemas de PLN desarrollados para dicha tarea cuando la ironía está presente. Por consiguiente, exploramos la posibilidad de utilizar nuestros resultados en detección de ironía para el desarrollo de un sistema de análisis de sentimientos que considere de la presencia de ironía, suponiendo que la detección de contenido irónico podría ayudar a mejorar la correcta identificación del sentimiento expresado en un texto dado. Con este objetivo, incorporamos emotIDM como la primera fase en un sistema de análisis de sentimientos para determinar la polaridad de mensajes en Twitter. Comparamos nuestros resultados con el estado del arte establecido en la tarea de evaluación 'Semeval-2015 Task 11: Sentiment Analysis of Figurative Language in Twitter', demostrando la importancia de utilizar información afectiva en conjunto con características que alertan de la presencia de la ironía para desempeñar análisis de sentimientos en textos con lenguaje figurado que provienen de redes sociales. En resumen, demostramos la utilidad de aprovechar diferentes aspectos de información relativa al afecto y las emociones para tratar cuestiones relativas a la presencia de la ironía en Twitter.

Resum

La forma en la qual s'expressen els usuaris en les xarxes socials ha atret l'atenció de diverses disciplines a causa del gran potencial d'investigació que aquesta representa. Les plataformes de xarxes socials, com Twitter, oferixen als usuaris la possibilitat d'expressar-se de forma lliure i espontània fent ús de diferents recursos lingüístics com la ironia per aconseguir diferents propòsits de comunicació. Les expressions iròniques en les dites plataformes són generades per usuaris que, la majoria de les voltes, tenen solament una definició intuïtiva del que és la ironia. Manejar aquest tipus de contingut representa un gran repte per a la lingüística computacional. La ironia està estretament vinculada amb l'expressió indirecta de sentiments, emocions i avaluacions (enteses com l'actitud o postura de l'autor cap a l'entitat específica involucrada en l'expressió irònica). Per consegüent, l'interés a detectar la presència d'ironia en textos de xarxes socials ha augmentat significativament en els últims anys, també per l'impacte en àrees del processament del llenguatge natural (PLN) relacionades amb l'anàlisi de sentiments, on detectar la presència d'ironia és important per evitar malinterpretar expressions iròniques com literals.

En aquesta tesi, introduïm el problema de detecció d'ironia en xarxes socials des de la perspectiva de la lingüística computacional. Proposem abordar aquesta tasca enfocant-nos, particularment, en el rol d'informació relativa a l'afecte i les emocions per detectar la presència d'aquest recurs lingüístic. Amb la intenció d'aprofitar el valor intrínsec de subjectivitat contingut en les expressions iròniques, presentem un model per a detectar la presència d'ironia denominat emotIDM, el qual està basat en una àmplia varietat de trets afectius. Per caracteritzar instàncies iròniques, utilitzàrem un ampli conjunt de recursos que cobrixen diferents àmbits afectius: des de sentiments (positius o negatius) fins emocions específiques definides de forma molt detallada. Abordem la detecció d'ironia interpretant-la com un problema de classificació binaria. Per avaluar el nostre model, recopilàrem un conjunt de dades de Twitter que prèviament s'havia utilitzat en treballs relacionats amb la detecció d'ironia tenint en compte dos objectius: per un costat, comparar l'acompliment del nostre model amb altres enfocaments de l'estat de l'art, i per l'altre, avaluar la seua robustesa considerant diversos aspectes diferents, els quals estan relacionats amb les característiques dels conjunts de dades com la modalitat d'obtenció, tamany i grau de desequilibri entre classes. Els resultats obtinguts mostres que emotIDM té un rendiment competitiu en els experiments realitzats, validant l'efectivitat de l'enfocament proposat. En la majoria dels casos, els nostres resultats han superat l'estat de l'art confirmant doncs que la informació relativa a l'afecte i les emocions és d'ajuda per a distingir entre tweets irònics i no irònics.

Un altre objectiu de la tesi és investigar les diferències entre tweets etiquetats com a #irony i #sarcasm. La nostra finalitat és contribuir a un tema menys investigat en lingüística computacional: la separació entre l'ús d'ironia i sarcasme en xarxes socials, amb especial èmfasi amb els trets afectius. A més, estudiem un hashtag que ha sigut menys estudiat: #not, el qual s'ha utilitzat en investigacions relacionades amb la detecció de continguts irònics per recopilar dades que expressen una intenció sarcàstica. Els nostres resultats pareixen evidenciar que existixen diferències entre els tweets que contenen els hashtags esmentats, cosa que suggerix que s'utilitzen per fer referència de diferents recursos lingüístics. Identifiquem un conjunt de característiques basades en diferents fenòmens afectius que pareixen ser útils per a discriminar entre diferents tipus de recursos lingüístics, a més que els nostres resultats superaren aquells en l'estat de l'art. Addicionalment analitzem la reversió de polaritat en tweets que continguen hashtags irònics, observant que l'impacte del fenomen esmentat és diferent per a cadascun d'ells. En el cas dels tweet que estan etiquetats amb el hashtag #sarcasm, a sovint hi ha una reversió total (que va des d'una polaritat a l'oposada, quasi sempre des d'un sentiment positiu a un negatiu), mentre que en el cas dels tweets etiquetats amb el hashtag #irony es produïx una atenuació de polaritat (principalment de negatiu a neutral).

Detectar ironia en contingut generat per usuaris podria tindre un ampli rang d'aplicacions. Sens dubte, una de les àrees que poden resultar més beneficiades per la detecció d'ironia és l'anàlisi de sentiments. Duem a terme un estudi de l'impacte de la ironia i el sarcasme en l'anàlisi de sentiments, on observem una disminució en el rendiment dels sistemes de PLN desenvolupats per a aquestes tasques quan la ironia està present. Per consegüent, vam explorar la possibilitat d'utilitzar els nostres resultats en detecció d'ironia per a desenvolupar un sistema d'anàlisi de sentiments que considere la presència d'ironia, suposant que la detecció de contingut irònic podria ajudar a millorar la correcta identificació del sentiment expressat en un text donat. Amb aquest objectiu, incorporem emotIDM com la primera fase en un sistema d'anàlisi de sentiments per determinar la polaritat de missatges en Twitter. Hem comparat els nostres resultats amb l'estat de l'art establert en la tasca d'avaluació 'Semeval-2015 Task 11: Sentiment Analysis of Figurative Language in Twitter', demostrant la importància d'utilitzar informació afectiva en conjunt amb característiques que alerten de la presència de la ironia per exercir anàlisi de sentiments en textos amb llenguatge figurat que provenen de xarxes socials. En resum, hem demostrat la utilitat d'aprofitar diferents aspectes d'informació relativa a l'afecte i les emocions per tractar qüestions relatives a la presència d'ironia en Twitter.

Abstract

Lo studio delle modalità di espressione nei social media è un tema che ha suscitato molto interesse in diverse discipline per via del grande potenziale che queste scritture rappresentano per la ricerca in vari ambiti. Le piattaforme di social media, come Twitter, offrono agli utenti la possibilità di esprimersi in modo spontaneo, anche utilizzando parlar figurato e in particolare figure retoriche come l'ironia per perseguire scopi comunicativi diversi. Gli enunciati ironici in queste piattaforme sono espressione di utenti che per lo più fanno riferimento a una nozione intuitiva di ciò che l'ironia è. Quello dell'ironia rimane un fenomeno interessante da esplorare sotto diverse prospettive, che pone in particolare una sfida difficile per la linguistica computazionale. L'ironia è spesso in stretta relazione con l'espressione indiretta di sentimenti, emozioni e valutazioni, intese come atteggiamenti valutativi o prese di posizione di chi scrive verso una particolare entità target coinvolta nell'enunciato ironico. Di conseguenza l'interesse per la possibilità di identificare la presenza di ironia nei testi di social media è cresciuta in modo significativo negli ultimi anni, anche per l'impatto su aree di ricerca nell'ambito dell'elaborazione del linguaggio naturale (NLP) legate all'analisi del sentiment e delle emozioni, dove identificare l'ironia è importante per evitare di interpretare erroneamente in senso letterale enunciati ironici.

In questa tesi, il problema dell'identificazione dell'ironia nei social media viene affrontato in una prospettiva di Linguistica Computazionale. Nell'affrontare questo problema, ci focalizzeremo in particolare sul ruolo dell'informazione legata alle emozioni e all'*affect*, investigando il ruolo di questa informazione nell'identificazione di questo particolare tipo di linguaggio figurato. Cercando di sfruttare l'intrinseca carica soggettiva delle espressioni ironiche, verrà presentato un nuovo modello computazionale chiamato EmotIDM, sviluppato per identificare l'ironia sulla base di una vasta gamma di caratteristiche legate alla sfera dell'affect. Per caratterizzare un'espressione ironica viene usato un insieme consistente e vario di risorse che coprono diversi aspetti della sfera dell'*affect*, dal sentimento positivo o negativo alle emozioni specifiche di granularità più fine. Il problema del riconoscimento dell'ironia viene interpretato come un problema di classificazione binaria. Per valutare il modello proposto è stato raccolto un insieme di corpora Twitter, già utilizzati dagli studiosi del settore in ricerche precedenti, che sono stati utilizzati come dataset di riferimento con un duplice scopo: confrontare le performance del modello qui proposto con quelle di altri approcci in letteratura, e valutare la sua robustezza secondo diverse dimensioni legate alle caratteristiche dei corpora, come la modalità di raccolta dei dati, le dimensioni e il grado di 'imbalance' fra le classi. I risultati e gli esperimenti effettuati mostrano che emotIDM ha prestazioni competitive, a conferma dell'efficacia dell'approccio proposto. Nella maggior parte dei casi, i risultati sono migliori di quelli ottenuti, come documentato in letteratura, da altri sistemi di identificazione dell'ironia, il che conferma che le informazioni legate alla sfera dell'*affect* aiutano a distinguere tra tweet ironici e non ironici.

Un altro obiettivo della tesi è quello di studiare le differenze tra i tweet etichettati con #irony e #sarcasm. Lo scopo è contribuire a fare luce su un argomento poco studiato nell'ambito della Linguistica Computazionale, ovvero la separazione tra ironia e sarcasmo nei social media, ancora una volta, ponendo una particolare attenzione al ruolo dell'informazione legata all'*affect*. Lo studio ha coinvolto anche un hashtag meno esplorato, che è stato utilizzato in letteratura per raccogliere esempi di espressioni sarcastiche: #not. Lo studio ha portato alla luce argomenti data-driven a favore dell'esistenza di differenze tra i tweet che contengono questi hashtags, suggerendo che i diversi hashtags sono utilizzati per fare riferimento a differenti dispositivi linguistici del parlar figurato relativo all'ironia. In particolare sono state identificate features promettenti basate su fenomeni legati alla sfera dell'affect per discriminare tra i diversi tipi di dispositivi linguistici figurativi, e i risultati di classificazione automatica del modello proposto migliorano lo stato dell'arte. E' stato analizzato anche il ruolo dell'inversione di polarità nei tweet contenenti diversi hashtag ironici, e si osserva che ci sono variazioni interessanti. Nel caso di tweet etichettati con #sarcasm, spesso l'ironia determina un'inversione completa (con variazione da una polarità al suo opposto, in particolare quasi sempre da polarità positiva letterale a negativa intesa), mentre nel caso di messaggi etichettati con #irony si osserva un'attenuazione della polarità (per lo più da negativa letterale a neutra intesa).

La rilevazione dell'ironia nei contenuti generati dagli utenti di social media ha una vasta gamma di possibili applicazioni. Senza dubbio, una delle aree che possono trarre un notevole beneficio dal riconoscimento automatico dell'ironia è l'analisi dei sentimenti. E' stato analizzato l'impatto della presenza di ironia e sarcasmo sull'analisi del sentimento, osservando una calo nelle prestazioni dei sistemi NLP sviluppati per questo compito quando l'ironia è presente. Pertanto, è stato investigata la possibilità di usare i nostri risultati sul riconoscimento dell'ironia per lo sviluppo di un sistema di analisi del sentimento che tenga conto di questo fenomeno, ipotizzando che l'identificazione di contenuti ironici possa avere un impatto positivo e migliorare la corretta identificazione della polarità del sentimento. A questo scopo, il modello emotIDM è stato inserito all'interno di una pipeline per lo sviluppo di un sistema automatico per determinare la polarità di messaggi Twitter. I risultati sono stati confrontati con quelli dello stato dell'arte, determinato dallo shared task "Semeval-2015 Task 11: Sentiment Analysis of Figurative Language in Twitter", e dimostrano che considerare le informazioni affettive insieme a features relative alla presenza di ironia è importante e utile per l'analisi del sentimento in presenza di linguaggio figurato, nei testi di social media considerati. In sintesi, è stata dimostrata l'utilità di sfruttare diverse dimensioni dell'informazione legata alla sfera dell'*affect* per riconoscere la presenza di ironia in Twitter.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Analyzing written language is an interesting topic that has been studied by several disciplines. Recently, due to the explosive growth of Internet, social media has become an attractive source of information for research purposes on written communication. Social media allow people to express themselves in various and varied ways. These platforms promote the use of figurative language devices such as irony and sarcasm for communicative purposes. Dealing with irony and sarcasm in social media represents a big challenge for computational linguistics.

In this thesis, we introduce the problem of detecting irony and sarcasm in social media under a computational linguistics perspective. We propose to address this task by considering one of the most important aspects of such attractive figurative language device: the affective component of irony. To do so, we take advantage of a wide range of lexical resources covering several facets of affect ranging from sentiment to fine-grained emotions.

This chapter introduces first theoretical concepts about irony and sarcasm. We describe the use of these figurative language devices in social media and some of the areas that could benefit from the results of being able to recognize them. Next, we briefly introduce current approaches in literature to address this task. Finally, we present the research questions, objectives, contributions and structure of this document.

1.1 Irony and Sarcasm

Studying how the language is used for communication purposes has been a concerning topic for many years. As human beings, we have the ability to express ourselves in complex ways. We are able to interpret when words in a sentence have a meaning beyond to the literal one, i.e., when language is used in a figurative sense. Figurative language can be described as the use of words to say something more than its literal meaning in a creative way. There are many figurative language devices. Among them, there is one that is widely used to achieve different communication purposes: *irony*.

Irony is a concept difficult to define in formal terms. It has been widely investigated by many research areas ranging from linguistics, philosophy, psychology, cognitive science, to computational linguistics. Each one has tried to define what irony is. However, there is a lack of agreement between them. Broadly speaking, irony is a term covering two concepts: *situational irony* and *verbal irony*.

Situational irony is stated to be a condition of events opposite to what was, or might be naturally expected, or a contradictory outcome of events [96]. Since we are interested in studying the use of irony in social media texts, we focus on the different perspectives for describing verbal irony. Several theories have been proposed attempting to define verbal irony. According to Grice [66] this concept is a trope where the speaker intends to communicate the opposite meaning of what is literally said. Verbal irony has been defined by Wilson and Sperber [151] as an "echoic mention" that refers to some proposition to demonstrate its absurdity. The difference between the context surrounding an utterance and its literal meaning was considered by Attardo [10] to define what irony is. In addition to this, Utsumi [141], Kummun-Nakamura and Glucksberg [88] have taken into account the speaker's position (approval or disapproval) on the result of something; they consider that an ironic utterance is triggered due to a "failed expectation". Furthermore, Giora et al. [63] have considered the role of negation to achieve a non-literal (i.e., ironical) intention. However, none of the above theories provide an exact definition of what verbal irony is.

Verbal irony is also related to another interesting concept: *sarcasm*. Irony is often deemed as a synonym of sarcasm due to the subtle distinction between both terms [59, 62, 63]. Even though some authors consider irony as an umbrella term that also covers sarcasm [28, 56, 87], others propose arguments to separate both concepts. Many authors have identified that when irony involves stressed negative evaluation towards a particular target with the intention to offend, it can be understood as sarcasm [10, 26, 38, 44, 100]. Therefore, sarcasm could be perceived as more aggressive than irony; the former is used to express biting criticism of a particular target.

The fuzzy separation between these concepts could lead to confusion. Therefore, when we are speaking about their everyday use, the differences between irony and sarcasm depend on how they are used rather than on theoretical definitions. Commonly, the term "irony" is used as an umbrella term also covering sarcasm. In what follows we shall use "irony" in the same perspective.

Irony is commonly used to express an evaluative judgment or attitude towards some particular target such as an utterance, an event or a situation by using the language in a creative and non-literal sense [5, 44, 66, 151]. In some way, irony allows us to convey very subjective ideas and opinions in an indirect way, going beyond the literal meaning of the words. Although there is no agreement on a single definition of verbal irony, most theorists have claimed towards the critical role of affective aspects when this figurative language device is used. Irony is closely associated with the expression of feelings, emotions, attitudes, and evaluations [5, 66, 151] toward a particular target.

Irony represents an important challenge not only from a linguistic point of view but also from a cognitive one. Since early childhood, most people can recognize and produce ironic utterances even without a strict definition of what is or may be considered as an irony [68]. These competencies are often related to the ability to correctly inferring and interpreting others' communicative intentions as well as their attitude toward a given situation. Investigating how this mode of communication is used can help to understand better the cognitive and linguistic processes involved in such interesting language device. Irony comprehension implies dealing with both literal and figurative sense of an utterance. Thus, understanding irony requires a more complex set of cognitive abilities than literal meaning [40].

The following section introduces the use of irony and sarcasm in social media as well as some of the potential applications that could benefit from irony detection.

1.2 Irony and Sarcasm in Social Media

Internet has influenced almost every aspect of our daily life during the recent years. It has become an essential part of almost everything we do. We use Internet for searching information, buying products, watching movies, and so on. One of the main changes involves the communication mode used by people around the world. Internet provides many options for interacting with others; probably the most popular one is the use of social media platforms. These media have become a powerful communication channel. People use them not only to keep in contact with their relatives and friends but also to share their opinions and ideas.

Investigating how people express themselves in social media has attracted the attention of several disciplines due to the great potential for research that it represents. User-generated contents in social media offer a gold mine of useful data for investigating users by studying their online activities, postings, attitudes and behavior [1]. Social media provide diverse kinds of content ranging from video, audio, images, and texts, the latter being one of the most investigated. Texts produced in such media are an invaluable source of information that can be exploited by several areas. To cite only one example, by taking advantage of social media written content it is possible to identify profiling aspects such as gender, age, or personality of a given author [122].

People tend to use irony in social media to achieve different communication purposes. Ironic utterances in such platforms are generated by users that most of the time have only an intuitive definition of what irony is. Therefore, ironic content in social media reflects what people considers such figurative language device. Let us introduce some posts from social media where users employ irony[1]:

[1]Examples (i) and (ii) were extracted from a Twitter dataset collected by Ghosh et al. [51]. Example (iii) was retrieved from the Reddit dataset described in [146]. Example (iv) belongs to the Amazon Reviews corpus collected by Filatova [48].

(i) Dad is giving me pointers on how to grill when he burns almost everything he grills #irony

(ii) It's totally fine to blare music in the house when I have school in the morning ... #sarcasm

(iii) I honestly do not know if this article is for real... no one is that stupid... right??

(iv) Customer review on the product "Baby Call Nursery"
I was chuckling to myself, reading all the great reviews of this product, thinking, "do I have the same monitor? Am I doing something wrong?". Then I got to the reviews at the bottom of the page in which other customers were having the same beeping problem as me. Thank goodness! We have the baby in a room on the same floor as us – the monitor is maybe 15 feet away...and it beeps randomly all the time. God forbid I move it maybe 20 feet away from the base - forget it! I am out of range (and the monitor doesn't let me forget it - which is good - but I don't feel like I should be out of range at this distance). Unfortunately, I cannot recommend this monitor to anyone, unless you enjoy not sleeping - not due to a baby crying, but rather your monitor beeping!

Very often, everybody has a predefined idea, expectation or bias for a certain situation. As illustrated by the examples above, when evaluating the outcome of a situation, people can demonstrate their attitude by an ironic utterance. Irony can be used to express an evaluation indirectly (examples (ii) and (iv)); to reveal a failed expectation (example (i)); to underline a context incongruity (example (iii)), and so on. Unlike investigating irony in other areas, when dealing with social media texts, we have a real depiction of what people consider irony instead of designed material. Those instances represent a big challenge since they seldom assume a template containing all (or at least the majority of) the linguistic properties suggested by the experts.

Irony detection is a fascinating and challenging task. Correctly identifying real intention behind social media posts could have a broad range of applications. An environment for communication such as social media provides an ideal scenario of several security issues such as threatening messages. When a suspicious message is detected, it is needed to verify whether it is genuine or not. A way to communicate a malicious message could be hiding it behind a figurative language device. Therefore, a system should be able to identify when the meaning of the words used in a message is not the literal one[2]. Furthermore, recognizing irony could be useful for many other purposes not necessarily related to social media texts. For example

[2]A misinterpreted tweet provoked the detection of two Irish at an airport in Los Angeles. The user wrote: "Free this week, for quick gossip/prep before I go and destroy America". The message was considered as a threat against the U.S. because of the use of figurative language ("destroy" was to meant "party"). `http://abcnews.go.com/Blotter/pair-held-twitter-homeland-threat-mix-reports/story?id=15472918`

in "human-computer interaction systems", it is needed to properly understand the real intention of a given sentence. Therefore, identifying figurative language devices could help to avoid misunderstandings and to improve the user experience with such systems.

Undoubtedly, one of the areas that may result in more beneficiated by irony detection is sentiment analysis. Generally speaking, sentiment analysis is the task of automatically determining the polarity of a piece of text (whether it is positive, negative or neutral) [102]. In other words, sentiment analysis helps us to determine the evaluative judgment expressed towards a particular target. Dealing with subjective contents often rich in irony especially in some domains [48, 143], systems analyzing sentiment should avoid misinterpreting the sentiment polarity expressed when ironic statements are used. In other words, they need to go further the literal meaning of the text in hand. Indeed, often irony provokes an interesting effect: polarity reversal; i.e., leading the overall polarity valence from positive to negative, or vice-versa. The presence of irony and sarcasm could indeed undermine the accuracy of sentiment analysis systems [25]. It has been recognized that identifying the presence of irony improves the performance of sentiment analysis systems [99]. The monitoring and control of irony and sarcasm in sentiment analysis systems is an ongoing task. We are just at the beginning of developing systems able to correctly handle the presence of these figurative language devices.

Just to mention two of the sentiment analysis related applications that could benefit from irony detection, let us introduce the following ones:

- *Social media monitoring services*
  These kinds of services may take advantage of irony detection by using it to filter and classify the content of interest. Then, when information about a given target is found, those comments having a meaning different from the literal one could be correctly identified. For marketing and political purposes, misunderstanding those messages where people subtly communicate dissatisfaction could impact on the perception of the success or progress of a given campaign.

- *Predicting stock market movements*
  The sentiment expressed in social media texts has been proven to have some influence on stock market prediction [24]. Therefore, to correctly study how people's opinion influences the market movements, it is needed to have reliable indicators of social media content's sentiments. The presence of irony could undermine these results.

Several approaches to deal with such a difficult objective have been proposed. The next section presents an overview of the current methods for detecting irony in social media texts. A most complete review of the state of the art in irony and sarcasm detection is described at Chapter 2.

1.3 Computational Linguistic Approaches

Interest in detecting the presence of irony in social media texts has grown significantly in the recent years. Irony detection is not a trivial task due to the compound nature of this kind of figurative language device. Being a tricky mode of communication, irony covers every aspect of language from lexical words to conceptualization. This means that when irony detection is addressed as a natural language processing task, it is needed to analyze not only the surface structure of a text but also the semantic and pragmatic meaning of the used words. Dealing with this kind of content requires an in-depth understanding of how the language is used beyond its literal meaning. It implies the need to acquire and process knowledge about the real world, a non-trivial task for computational linguistics. Therefore, detecting irony involves the most challenging areas of natural language processing.

Research in user-generated content is a very challenging task. Despite the social media texts convey an invaluable source of information, they are difficult to process because they are noisy, informal, with little context, and plenty of grammatical mistakes. Unlike in face-to-face interaction, when computer-mediated communication is used, we do not have the extra information of facial expressions, gestures, postures, and voice tone (denoted as paralinguistic cues); these clues have been recognized as important when people use irony [11]. Notwithstanding the lack of such cues, social media offer a face-saving way that promotes the use of ironic language. When using social media, the users may take advantage of different resources provided by these platforms such as emoticons, images, and alike, to cover the necessity mentioned above.

Several approaches have been proposed to deal with irony detection. Most of the current methods conduct a procedure involving first the usage of linguistic patterns used to communicate ironic messages in social media, i.e., a set of features attempting to capture the real meaning of an utterance. As in the majority of the natural language processing tasks, developing corpora for experimental purposes is an issue. Therefore, the current trend is to generate an in-house dataset to assess the performance of a given model. There are two main methods to create ironic corpora: (i) gathering user-generated data, and then annotating the presence of irony, and (ii) taking advantage of specific labels used by social media users to point out an ironic intention such as Twitter hashtags. The variety of corpora used for irony detection includes Amazon reviews[3] [48], comments from debate sites such as 4forums.com[4] [2, 97], entries submitted to Reddit[5] [146], dialogues from TV series [79], quotes on GoodReads[6] [80], and mostly data retrived from Twitter[7] [13, 41, 81, 104, 120, 121, 126, 127]. Being considered a classification task, irony detection has been addressed by taking advantage of a broad range of machine learning classifiers already used in such kinds of tasks, and the most frequent are: decision tree

[3] http://www.amazon.com
[4] http://www.4forums.com
[5] http://www.reddit.com
[6] http://www.goodreads.com
[7] http://www.twitter.com

and support vector machine [18, 126]. Recently, deep learning algorithms have also been used for detecting irony in social media [80, 115]. Nevertheless, there is no particular evaluation framework for assessing the performance of irony detection models.

The use of irony in social media has been investigated mainly in Twitter. Twitter represents a powerful real-time communication tool that is extremely popular. It provides an interesting source of information regarding how people perceive events, products, and so on; besides, it enables an increasing accessibility to large amounts of data. Twitter texts, named tweets, are short: a maximum of 140 characters is allowed. Therefore, users write their messages in a concise manner, and if needed they can also exploit the markers provided by Twitter (such as hashtags, emoticons, images, and URL) to accomplish their communicative intent. For irony detection, perhaps the most important mark of Twitter texts is the hashtag. It allows users to point out their ironical intention clearly. Having ironic comments annotated by their authors enable us to capture the real use of irony on Twitter. Indeed, the readability of using hashtags as golden labels has been experimentally confirmed [70, 89].

Current irony detection research relies on information such as the text itself, social media metadata, and some knowledge-based resources. Irony detection has been addressed as a classification problem where the main aim is to distinguishing between ironic and non-ironic texts. For such kind of task, taking advantage of those elements that are more salient to be exploited automatically become a crucial part. An important issue concerns the lack of agreement on the differences between irony and sarcasm. Under a computational linguistics perspective, both concepts are often considered either as synonyms or being irony an umbrella term covering sarcasm. Thus, the current models that have been proposed, address irony as a broad term. Indeed, in the related works on irony detection the term used by the authors to refer to this figurative language device is: irony or sarcasm. Below, we briefly introduce some of the proposed approaches for detecting ironic content in social media without distinguishing the term used to name verbal irony.

Among the different proposals for detecting irony, the most applied one is related to exploit features coming mainly from the text in hand. Kreuz and Caucci [86] pointed out the fundamental role that lexical factors play for both communicating and recognizing ironic intent. Taking advantage of this, several authors [29, 35, 41, 47, 65, 89, 120, 139] have attempted to identify relevant features by exploting the text itself. This kind of approach relies mainly on punctuation marks, uppercase characters, emoticons, and alike, together with n-grams and textual markers such as interjections, synonyms, adjectives, verbs, adverbs, part-of-speech labels, among others to capture the ironic intention in social media texts. Some of these features allow the user to cover the lack of paralinguistic cues that serve to alert about a possible use of irony.

Being a very subjective language device, irony is rich in sentiment content. This feature has also been exploited to address irony detection. There are some approaches where some features coming from sentiment lexica have been included as potential clues to characterize irony in texts. Studies such as those conducted

by Reyes et al. [126] and Barbieri et al. [18] have sought to take advantage of some sentiment features to perform irony detection. One less explored strategy is to identify utterances matching a particular pattern. In Riloff et al. [127], the authors attempted to recognize instances where a positive sentiment contrasts with a negative situation.

There are some cases where a stand-alone text does not provide enough information to identify the presence of irony. There exist some approaches aiming to go beyond the textual content in a social media message for recognizing irony. The idea here is to capture further information about the user and her environment, i.e., the context surrounding an ironic utterance. Rajadesingan et al. [121], Khattri et al. [83] and Bamman et al. [13] took advantage of historical Twitter posts of a user and the relationship with her audience. Wallace et al. [145] exploited the structure of Reddit to identify irony in conversational threads. An approach to capture pragmatic context in French tweets was proposed by Karoui et al. [81].

As mentioned before, one of the areas that could benefit from irony detection is sentiment analysis. Recently, some evaluation campaigns focused in sentiment analysis have considered the presence of ironic content to measure the performance of the participating systems. Since 2014, the task on sentiment analysis in the framework of SemEval has included a set of sarcastic tweets as part of the testset in the competition [105, 128, 129]. In 2015, for the first time, a task dedicated to sentiment analysis with a particular focus on figurative language devices in Twitter was also organized in the framework of SemEval [51]. Investigating the impact of irony in sentiment analysis tasks has also been addressed in Italian. The last two editions of SENTIPOLC (SENTIment POLarity Classification) in the framework of *EVALITA* also included as part of the evaluation the presence of irony in Italian tweets [14, 20]. Overall, the performance of the participant systems drops when figurative language is involved. It serves to confirm that detecting irony is, in fact, a critical part to avoid misleading polarity in social media messages. Despite the importance of detecting irony before carrying out sentiment analysis, it is needed to explore this issue further.

Research on irony detection has been focused mainly on English in a similar fashion than in other natural language processing tasks, although recently there is a growing interest in investigating irony detection and its impact on sentiment analysis in other languages such as: Chinese [139], Czech [120], Dutch [89], French[8] [81], Italian [25], and Portuguese [35].

In this section, we briefly describe some of the main approaches that have been proposed to deal with irony detection. More complete surveys can be found in the current literature [72, 78, 144].

The research conducted so far has been focused on several aspects of irony in social media. However, there is a particular aspect of this kind of figurative language

[8] In 2017 for the first time a shared task on sentiment analysis on the presence of figurative language devices written in French will be organized: `https://deft.limsi.fr/2017/indexEng.html`

device representing an attractive starting point for researching: the affective content. In the next section we will address the importance of taking into account affective information in irony detection.

1.4 Affective Information in Irony

As already discussed (Section 1.1), a major characteristic of irony is its emotionally-laden content. Ironical expressions serve not only as a way to convey an opposite of the literal meaning but to show certain attitudes [69]. Leggitt and Gibbs [91] and Shamay-Tsoory et al. [133] have recognized that emotions and affective states are crucial for both communicating and comprehending irony. Furthermore, Alba-Juez and Attardo [5] consider that a key for identifying among various types of verbal irony lies precisely in its evaluative component. Irony may mitigate an evaluation, softening from a negative observation to a kindest one, and, least common from a compliment to a fuzzy judgment. While on the other hand, sarcasm is intimately associated with particular negative affect states [100]. Other authors have also shared this perspective: sarcasm is offensive [90] and it may enhance negative emotions such as anger, irritation and disgust [91].

We believe the affective information in such kind of figurative language devices plays a key role. Filik et al. [49] suggest that phrasing a message ironically may influence the emotional response that is elicited. In order to investigate more the value of affect in language, and concretely certain forms of ironic language, especially in social media such as Twitter, we aim to analyze emotions in ironic texts. It has been recognized that textual communication expresses affective states [30]. Research has been carried out attempting to investigate how people express emotions through text and also how text can elicit different emotions. When attempting to identify the emotional charge in texts, it is important to cover as much affect-related information as possible.

Since irony is considered as an affective manner of communication, taking advantage of affect-related information may help to identify ironic content in social media. It is important to highlight that our aim is not to identify the emotions in texts but to exploit affective content for detecting irony. We proposed a model that relies mainly on affect-based information to characterize ironic content in Twitter.

A key factor when dealing with affect in ironic tweets is how to represent it. In the literature, various theories attempt to describe affect. We decide to make use of a wide range of English lexical resources covering several facets of affect. Probably the simplest way to depict affective states is through two broad aspects: positive and negative sentiment. We select a set of resources assigning an overall sentiment to words.

However, a complex and multifaceted phenomenon such irony merits to be addressed more in-depth. Therefore, we decided to consider also the presence of affect in a finer-grained perspective. Each language has a set of words used to express and

describe emotions. Taking as starting point the different models of emotions, some researchers have tried to investigate which words and how they are related to a given emotion. This has led to the development of different resources containing words related to different kinds of affective content. These resources have been widely used in psychological, linguistic, and psycholinguistic studies. Moreover, due to the proliferation of data generated in social media, several methods have been proposed to collect resources reflecting how people express affect in such media.

The study of emotions is one of the most challenging (and still open) chapters in the history of psychology. Emotions are elicited by a cognitive evaluation of antecedent situations and events [114]. Two main approaches are proposing a framework to describe emotions: the *Categorical model of emotions* and the *Dimensional model of emotions*. In the first place, the *Categorical model of emotions* involves the use of labels to identify affective states. Under this perspective, emotions are conceived as discrete entities having a particular tag such as boredom, frustration, anger, etc. The categorical approach argues the existence of a small number of basic emotions, having its own physiological, expressive, and behavioral reaction patterns as well as specific conditions to be triggered. There are various approaches defining categorical models of emotions such as those of Arnold [7], Ekman [45], Parrot [111], and Plutchik [114]. All these models include different basic emotions such as anger, aversion, courage, dejection, desire, despair, fear, hate, hope, love, sadness, disgust, joy, surprise, love, trust, and anticipation. Alternatively, the *Dimensional model of emotions* propose the idea that an emotional state is determined by its position in a space of independent dimensions such as valence, activation, power, and alike. Each theory in this approach proposes a set of different dimensions that serve to define a particular emotional state.

We used various lexical resources which refer to each of the approaches describing emotions. All these affective resources together allow us to attempt to catch the underlying emotionally-laden content in ironic texts from social media. We evaluated our model for detecting irony by casting it as a classification task. We experimented with a set of Twitter corpora in the state of the art composed by five dataset collected by Reyes et al. [126], Barbieri et al. [18], Riloff et al. [127], Mohammad et al. [104], and Ptáček et al. [120]. Our results outperform those from the literature.

Furthermore, using the same set of features, we analyze a set of tweets rich in figurative language content labeled with different hashtags #irony, #sarcasm, and #not. We are interested in determining whether or not these hashtags are used to accomplish different communication purposes by Twitter's users. Besides, bearing in mind the importance of irony detection for sentiment analysis tasks, we investigated also the differences in polarity reversal terms of such tweets. Finally, the results of our analysis allow us to probe distinctions and similarities between tweets labeled as #irony and #sarcasm. Our results allow us to contribute to the assumption that there is a separation between these figurative language devices.

1.5 Research Questions and Objectives

Concretely, the main purpose of this thesis is to validate the importance of affect-related information for irony detection. Following this objective, we carried out a research aimed to address the following main research questions which have been also entailed in derived questions:

I. Could information about different facets of affect be useful for detecting irony in social media?

   - How to build an irony detection model for social media texts taking advantage of affective information?

II. Are the #irony and #sarcasm hashtags used to label different ironic intentions?

   - Have different tweets labeled with #irony and #sarcasm the same degree of polarity reversal?
   - Is the #not hashtag used to label some sort of irony or sarcasm in Twitter?

III. Is it possible to improve the performance of sentiment analysis systems being aware of the presence of ironic content?

Aiming to answer the research questions previously presented, the following objectives have been defined:

- The main aim of this research is to propose a model based on affective information for detecting irony in Twitter. Several tasks must be performed to complete this objective:
  - Analyzing the presence of different aspects of affect in ironic utterances.
  - Identifying a set of features capturing affective characteristics such as sentiment and finer-grained emotions from texts.
  - Evaluating the performance of the model on different datasets. Taking advantage of several corpora having different properties allows us to validate the robustness of the proposed model.
- In the literature, some hashtags have been considered as reliable marks of ironic intention. We are also interested in to study the differences in affective-related terms among tweets containing with these labels. The tasks related to this objective are:
  - Determining which kinds of affect-related features serve to characterize the use of irony and sarcasm on Twitter.
  - Analyzing the polarity reversal effect on different kinds of ironic tweets.
  - Investigating the role of the hashtag #not as an ironic hashtag.

- Irony is a phenomenon having an impact on the results of sentiment analysis. We are interested in study such impact under two perspectives: first when a sentiment analysis system is not aware of irony, and second how its performance is considering such information.
  - Investigating the performance of sentiment analysis systems when they deal with the presence of irony and sarcasm.
  - Developing a sentiment analysis system that considers the presence of irony as a previous step before assigning a polarity degree.

1.6 Contributions

Irony has been recognized as a complex mode of communication closely related to the expression of feelings. Thus, considering the key part of sentiment and emotions in this kind of language device represents an interesting research opportunity. Hence, this thesis aimed at investigating the role of affect-related information in irony detection. We decided to take advantage of several features related to affective information to characterize irony in social media. In this line, the following contributions were made within the development of the present research:

- We presented a brief description of the proposed approaches in the literature together with an analysis of shared tasks regarding sentiment analysis where the participating systems were evaluated on the presence of figurative language devices. Overall, there is a drop in performance of the systems. It allows validating the assumption concerning to the importance of irony detection for determining the sentiment in a text.

- A novel model for identifying irony in social media, called emotIDM, was proposed. emotIDM takes advantage of different facets of affective information from overall sentiment to fine-grained aspects of emotions. emotIDM is the first irony detection model that is largely based on a broad range of affective information to capture ironic intention in Twitter. We experimented with a set of Twitter corpora in the state of the art. The obtained results overall outperform those from the literature.

- We investigated the use of the #irony and #sarcasm hashtags that have been recognized as a trustworthy label of figurative intent in Twitter. Besides, we also included a set of tweets labeled with #not that has been used to collect sarcastic instances. We performed a distribution and correlation analysis over a corpus rich in figurative messages considering a broad variety of psycholinguistic and affective features concerning emotional information. We conclude that these hashtags are indeed used to label different phenomena. We explored the controversial subject to separate irony from sarcasm outperforming the state of the art. We investigated the behaviour of tweets labeled with such hashtags in terms of polarity reversal.

- We incorporated emotIDM in a pipeline of sentiment analysis that relies mainly on sentiment and emotional-related resources. For the sake of comparison, we evaluated the proposed approach on the dataset developed for the SemEval-2015 Task 11. We compared our results against those of the shared task. Our proposal shows comparative results validating the relevance of exploiting affective related features as well as the presence of irony for determining the sentiment in a given tweet.

1.7 Structure of the Thesis

This document comprises a compendium of research articles published during the researching period of this PhD. Two international journal papers, an international conference paper, and a chapter in a book, as well as not published content constitute this thesis.

A brief overview of the sections included in this document is introduced below.

Chapter 2. Irony, Sarcasm, and Sentiment Analysis

The content of this chapter is part of the book titled "*Sentiment Analysis in Social Networks*". In this section, an overview of automatic irony detection is given, together with an introduction to some sentiment analysis shared task where the presence of ironic utterances was considered. This chapter introduces some state-of-the-art approaches that have been proposed to deal with irony and sarcasm detection in social media. Furthermore, we analyzed different shared tasks dedicated to sentiment analysis where the impact on the performance of irony was evaluated. Besides, we describe the first shared task on sentiment analysis fully dedicated to figurative language devices.

Chapter 3. Applying Basic Features from Sentiment Analysis for Automatic Irony Detection

This chapter presents the research work published in the "*Proceedings of the Pattern Recognition and Image Analysis: 7th Iberian Conference, IbPRIA 2015*". In this paper, we proposed an irony detection model focused on Twitter. Our model exploits two groups of features: surface patterns (such as the frequency of textual markers) and lexical-based features (exploiting external resources). We take advantage of two sentiment analysis related features: "sentiment score" and "polarity value" to characterize each tweet. We experimented with a set of classifiers outperforming the state-of-the-art results. The sentiment analysis features in our model emerged among the most relevant ones, thereby giving insights into the importance of such information for detecting irony in Twitter.

Chapter 4. Irony Detection in Twitter: The Role of Affective Content

This chapter introduces the research work published in the "*ACM Transactions on Internet Technology (TOIT)*" journal. An irony detection model for Twitter, called emotIDM, is described in this paper. emotIDM extends the model described in Chapter 3 with new features coming from a wide range of lexical resources reflecting different facets of affect. It is the first irony detection model that considers such kind of information covering aspects of affect ranging from sentiment to fine-grained models of emotions. emotIDM was evaluated over a set of Twitter corpora previously proposed in the literature. Overall, the obtained results outperform the performance of the state-of-the-art approaches. We proved the importance of affect-related information for irony detection in Twitter.

Chapter 5. Figurative Messages and Affect in Twitter: Differences between #irony, #sarcasm and #not

This chapter presents the research work published in the "*Knowledge-Based Systems*" journal. An analysis of different figurative language devices in Twitter is described. With the aim to distinguish between tweets labeled with #irony, #sarcasm and #not, we analyzed the role of features related to different facets of affective information in such tweets. We found data-driven arguments for the separation between texts labeled with these particular hashtags. Furthermore, we carried out classification experiments between tweets labeled with these hashtags. Our results for what concerns to classify between #irony and #sarcasm outperform those of the state of the art. With respect to #not, it seems that it is used to represent a different figurative language device, although closer to #sarcasm than #irony.

Chapter 6. Sentiment Polarity Classification of Figurative Language: Exploring the Role of Irony-Aware and Multifaceted Affect Features

This chapter presents the research work published in the *Proceedings of the 18th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2017)*. In this paper, we described an approach for performing sentiment analysis in tweets with figurative language. We proposed a pipeline that comprises two phases: first, we exploited emotIDM for identifying irony; then, by taking advantage of several affective resources we determine a polarity degree also considering the presence of ironic content. The obtained results are competitive with the state of the art.

Chapter 7. Discussion of the Results

This chapter summarizes the obtained results. With the aim to enhance this research work, we extended the experiments carried out. We analyze the impact of different lexical resources for classifying irony. We evaluated our irony detection model in Twitter corpora considering various degrees of imbalance distribution between classes. Finally, we carried out a qualitative analysis of the affective related resources contained in our irony detection model in order to calculate the overlap between them and have an idea of the grade of redundancy of the affective features of emotIDM.

Chapter 8. Conclusions and Future Work

This chapter outlines the conclusions of this thesis. Besides, it includes a summary of the contributions of this research work, as well as some possible directions for future work.

Chapter 2

Irony, Sarcasm, and Sentiment Analysis

In this chapter, we describe a literature review on irony and sarcasm detection. We introduce several state-of-the-art approaches for dealing with figurative language devices in social media. We describe the contributions and results of each method. Some shared tasks on sentiment analysis have started to take into account the impact of irony for evaluation purposes. We analyze the performance of the participating systems that is different when they deal with ironic content.

Published in:

Hernańdez Farías, D.I. and Rosso, P. (2017). Irony, Sarcasm and Sentiment Analysis. Chapter 7 in Pozzi, F.A., Fersini, E., Messina, E., and Liu, B., eds.: Sentiment Analysis in Social Networks. Morgan Kaufmann, pages 113-128.
ISBN: 978-0-12-804412-4.
DOI: `http://dx.doi.org/10.1016/B978-0-12-804412-4.00007-3`

Abstract

Irony and sarcasm are sophisticated forms of speech act in which the authors write the opposite of what they mean. They have been studied in linguistics, psychology and cognitive science. While irony is often used to emphasize occurrences that deviate from the expected, sarcasm is commonly used to convey implicit criticism. However, the detection of irony and sarcasm is a complex task, even for humans. The difficulty in recognizing irony and sarcasm causes misunderstanding in everyday communication and poses problems to many natural language processing task such as sentiment analysis. This is particularly challenging when dealing with social media messages, where the language is concise, informal and ill-formed. The detection of irony and sarcasm is a complex task, even for humans.

Keywords: Irony Detection, Sarcasm Detection, Figurative Language Processing

2.1 Introduction

Everyday people make judgments about their environment. This is an inherent behavior as human beings. There are different ways to express our opinions, one of the most interesting is by figurative language devices such as irony and sarcasm. This allows us to express ourselves in a particular way using the words not only in its most salient meaning but also in a creative and funny sense. The use of words or expressions with a meaning that is different from the literal interpretation, is known as figurative language.

Irony and sarcasm are two interesting and strongly related concepts. Usually people do not have a clear idea of what they are. However since early childhood we begin to use them in our daily life. They have been a topic studied by different disciplines such as: linguistics, philosophy, psychology, psycholinguistics, cognitive science, and recently computational linguistics. Each discipline has tried to define what they are, how they are produced, why they are used. These figurative devices give us the opportunity to explore the interaction between cognition and language.

Broadly speaking, irony and sarcasm refer to figurative language devices that serve to achieve different communication purposes. The most common definition of irony refers to an utterance by which the speaker expresses the opposite meaning for what is literally said. There are different theories which attempt to explain

what irony is. Grice's theory [66] points out that the speaker intentionally violates the "Maxim of Quality" (the speaker does not say what she believes to be false) when expresses an ironic utterance. Some theories like the one described in [151] propose to define it beyond the literal sense of the words: for Wilson & Sperber an ironic utterance is an "echoic mention" that alludes to some real or hypothetical proposition to demonstrate its absurdity. Attardo [10] considers an ironic utterance as a form of "relevant inappropriateness" in which the speaker relies on the ability to the listener to reject the literal meaning based on the disparity between what is literally said and the context where it is said. On the other hand, the "failed expectation" intention (i.e., the speaker's approval or disapproval of the entity or situation in hand) behind an ironic expression has been studied by Utsumi [141] and Kumon-Nakamura and Glucksberg [88].

Usually, irony is considered as a broader term that covers also sarcasm [54, 149]. Irony may be positive (i.e., non-critical) while sarcasm usually is not [5, 58]. Sarcasm is commonly more aggressive and offensive than irony. In this work, irony and sarcasm are treated as two different concepts.

Social media offer a face-saving way to expression for people, who sometimes choose to use ironic or sarcastic utterances to communicate their attitude or evaluative judgment towards a particular target (e.g. a public person, a product, a movie, an event, etc.). The presence of ironic or sarcastic content in human communication may cause misunderstandings. Even for humans to identify this intention is not a trivial task: different cognitive processes are involved as well as an environment knowledge is needed. For Natural Language Processing (NLP) tasks such Sentiment Analysis (SA), this kind of subjective user-generated content represents a big challenge. In some cases, the presence of ironic content plays a particular role: "polarity reversal". This means, for instance, that an utterance seems to be positive, but its real intention is negative (or vice versa).

Let us to introduce the following example, extracted from an ironic set of Amazon Reviews collected by Filatova [48]:

"i would recomend this book to friends who have insomnia or those who i absolutely despise"[1]

For a SA system that exploits the basic approach of considering the frequency of positive and negative terms to assign a polarity, this sentence could be considered as positive. The words "*recomend (recommend)*, *book* and *friends*" are positive terms, while "*insomnia* and *despise*" denote a negative sense. Therefore, in this sentence there are three positive and two negative terms, and it could be identified as positive. However, this review conveys a meaning far from positive. The author expresses a negative judgment against the book in an imaginative way. On one hand, the author speaks about to "recommend" the book, that can be considered as a positive aspect

[1] In this ironic utterance it can be noted an example of misspelling writing in social media texts. The author uses "recomend" instead of "recommend".

about the target (the book), but at the same time there is a point about "friends who have insomnia" or "those who i absolutely despise". Thus, the author's hidden intention could be to state that the "book" is enough boring to provoke sleep (even to those who have insomnia).

Research in irony could not only improve the performance of sentiment analysis systems, but it also can help us to understand the cognitive process involved and how the human beings process and produce this kind of utterances. After introducing the state-of-the-art in irony and sarcasm detection, the aim of this chapter is to investigate the impact that the use of these figurative language devices may have on sentiment analysis.

This chapter is organized as follows. In Section 2.2 we describe the state-of-the-art in irony and sarcasm detection. In Section 2.3 we address the impact that figurative language has on sentiment analysis. We analyze three shared tasks that recently have been organized. Section 2.4 discusses future trends and directions. Finally, in Section 2.5 we draw some conclusions.

2.2 Irony and Sarcasm Detection

Irony and sarcasm detection are considered as special cases of text classification, where the main goal is to distinguish between ironic (or sarcastic) texts from non-ironic (or non-sarcastic) ones. In order to analyze this kind of figurative devices, it is needed not only to consider the syntactic and lexical textual level (to extract salient features such as words position, punctuation marks, etc.), but also semantics (literal vs. non-literal meaning of the words), pragmatics (words matching with the appropriate context) and discourse analysis (relation between the utterance in hand with the discourse in which it is expressed). However, the progress so far achieved has been a result of exploiting mainly syntactic, lexical and shallow semantic.

Dealing with social media text is a challenging task. It has specific characteristics: it is informal and it uses ill-formed language. People express themselves in a face-saving way by unstructured content. Usually, social media texts contain spelling mistakes, abbreviations and slang. In Twitter, the text should be written in a maximum of 140 characters, therefore, figurative language is expressed in a very concise manner, which causes an additional issue. When people express their opinions by ironic or sarcastic utterances, they can choose how to use the language to achieve their communicative goals. There is no particular structure in order to construct ironic or sarcastic utterances. In a such way, the main objective of irony and sarcasm detection task is to discover features that allow us to discriminate ironic (or sarcastic) from non-ironic (or non-sarcastic) texts.

The interest in irony and sarcasm detection in social media provokes the necessity to have user-generated data that allow us to capture the real use of this kind of figurative language devices. As in the majority of NLP tasks, the lack of corpora is

an issue. There are two main approaches for ironic/sarcastic corpus construction: "self-tagging" and "crowd-sourcing". The first one considers as positive instances those texts in which the author points out her intention using an explicit label (e.g. the hashtags #irony or #sarcasm). Therefore, in this case we rely on the author's definition about what irony or sarcasm are. The "crowd-sourcing" involves human interaction by labeling the content as ironic (or sarcastic). Mainly, the labeling process is carried out without any strict definition or guideline. Therefore, it represents a subjective task, where the agreement between annotators is often very low. By this way, it is possible to obtain potential ironic and sarcastic texts produced by people in social media.

For computational linguistic purposes, irony and sarcasm are often considered as synonyms. The following sub-sections describe some proposed approaches to address irony and sarcasm detection. The first one is focused on works where the ironic intention was considered as an overall term, while the second one is focused on the research were sarcasm was considered as a different concept.

2.2.1 Irony Detection

One of the first works in irony detection was carried out by Carvalho et al. [35]. They worked on the identification of a set of surface patterns to identify ironic sentences in a Portuguese on-line newspaper. The most relevant features were the use of punctuation marks and emoticons. Veale and Hao [143] carried out an experiment by harvesting the web looking for a commonly used framing device for linguistic irony: "the simile" (two queries "as * as * " and "about as * as * " were used to retrieve snippets from the web). They analyzed a very large corpus to identify characteristics of ironic comparisons, and presented a set of rules in order to classify a simile as ironic or non-ironic.

Reyes et al. [125] analyzed tweets tagged with the hashtag #irony and #humor in order to identify textual features for distinguishing them. They proposed a model that includes structural, morphosyntactic, semantic and psychological features. Additionally, their considered the polarity expressed in a tweet using the Macquarie Semantic Orientation Lexicon (MSOL)[2]. The authors experimented with different feature sets and a decision tree classifier, obtaining encouraging results (F-measure 0.80 approximately).

Afterwards, Reyes et al. [126] collected a corpus composed by 40,000 tweets, relying on the "self-tagged" approach. Four different hashtags were selected: #irony, #education, #politics and #humor. Their model is organized accordingly to four types of conceptual features: signatures (such as punctuation marks, emoticons and discursive terms), unexpectedness (opposition, incongruency and inconsistency in a text), style (recurring sequences of textual elements) and emotional scenarios (elements which symbolize sentiment, attitude, feeling and mood) by exploiting

[2] `http://www.saifmohammad.com/Release/MSOL-June15-09.txt`

the Dictionary of Affect in Language (DAL)[3]. They addressed the problem as a binary classification task, distinguishing ironic from non-ironic tweets by using naïve bayes and decision tree as classifiers. The authors achieved an average of 0.70 in F-measure terms. Barbieri and Saggion [17] proposed a model to detect irony using lexical features, such as: frequency of rare and common terms, punctuation marks, emoticons, synonyms, adjectives, positive and negative terms. They compared their approach with the one of Reyes et al. [126] on the same corpus using a decision tree, slightly outperforming the previously obtained results. The authors concluded that rare words, synonyms and punctuation marks seem to be the most discriminating features. Hernández et al. [74] described an approach for irony detection that uses a set of surface text properties enriched with SA features. The authors exploited two widely applied SA lexicons: Hu&Liu[4] and AFINN[5]. They experimented with the same dataset used in [126, 17]. Their proposal was evaluated using a set of classifiers composed by naïve bayes, decision tree, support vector machine, multilayer perceptron and logistic regression. The proposed model improves the previous results (F-measure 0.79 approximately). The features related to SA were the most relevant.

Buschmeier et al. [29] presented a classification approach using the Amazon review corpus collected by Filatova [48], which contains both ironic and non-ironic reviews annotated by Mechanical Turks crowdsourcing. The authors proposed a model that takes into account features such as: n-grams, punctuation marks, interjections, emoticons and the start-rating of each review (a particular feature from Amazon reviews, that according to the authors, seems to help obtaining a good performance in the task. They experimented with a set of classifiers (composed by naïve bayes, logistic regression, decision tree, random forest and support vector machine) achieving a F-measure rate of 0.74.

Wallace et al. [145] attempted to undertake the study of irony detection using contextual features, specifically by combining noun phrases and sentiment extracted from comments. They propose exploiting information regarding to the conversational threads to which comments belong. Their approach capitalizes the intuition that members of different user communities are likely to be sarcastic about different things. A dataset of comments posted to Reddit[6] was used[7].

Karoui et al. [81] have recently presented an approach to separate ironic from non-ironic tweets written in French. They proposed a two-stage model. In the first part they addressed the irony detection as a binary classification problem. Then, the misclassified instances are processed by an algorithm that tries to correct them by querying Google to check the veracity of tweets with negation. They represented

[3] `http://www.cs.columbia.edu/~julia/papers/dict_of_affect/`

[4] The resource is freely available: `http://www.cs.uic.edu/~liub/FBS`

[5] The resource is freely available: `http://github.com/abromberg/sentiment_analysis/blob/master/AFINN/AFINN-111.txt`

[6] `http://www.reddit.com`

[7] Particularly comments posted to two pairs of polarized user communities (or subreddits) were selected: progressive and conservative subreddits (related to US political spectrum respectively) and atheism and christianity subreddits.

each tweet with a vector composed of six groups of features: surface (such as punctuation marks, emoticons and capital letters), sentiment (positive and negative words), sentiment shifter (positive and negative words in the scope of an intensifier), shifter (presence of an intensifier, a negation word or reporting speech verbs), opposition (sentiment opposition or contrast between a subjective and an objective proposition) and internal contextual (the presence/absence of personal pronouns, topic keywords and named entities). The authors experimented with a support vector machine as classifier achieving an F-measure of 0.87.

To sum up, several are the approaches that have been proposed to detect irony as a classification task. Many of the employed features have been already used in various tasks related to sentiment analysis such as polarity classification. The ironic intention is captured by exploiting mainly surface features such as punctuation marks and emoticons. These kinds of lexical cues have been shown to be useful to distinguish ironic content, especially in tweets. It may confirm in some way the necessity of the users to add textual markers to deal with the absence of paralinguistic cues. Besides, many authors point out the importance to capture the inherent incongruity in ironic utterances. To achieve this goal the presence of opposite polarities (positive and negative words) and the use of semantically unrelated terms (synonyms and antonyms) have been considered in many approaches. Both kinds of features seem to be relevant for characterize ironic from non-ironic utterances. Decision trees have been among the classifiers that obtained the best results.

2.2.2 Sarcasm Detection

In order to determine whether specific lexical factors (e.g. the use of some part of speech or punctuation marks) play a role in the sarcasm detection, Kreuz and Caucci [86] asked to some college students to read excerpts from paragraphs that originally contains the "said sarcastically" sentence (removed before this task). The participants were able to distinguish sarcastic from non-sarcastic utterances. This work represents a key in order to consider the influence that lexical factors can have to analyse social media content.

One of the first approaches which considers the #sarcasm hashtag as indicator of sarcastic content, was developed by Davidov et al. [41]. The authors introduced a semi-supervised algorithm for sarcasm detection that considers as features frequent words, punctuation marks and syntactic patterns in order to identify sarcastic utterances. They collected a dataset from both Amazon and Twitter; their results seem to be promising, with F-measures close to 0.80.

Gonzalez et al. [65] carried out an experiment over two datasets: a set of self-tagged tweets and a manually annotated one. They considered as sarcastic instances a set of self-tagged tweets containing #sarcasm or #sarcastic hashtags, and as non-sarcastic instances some positive and negative tweets (retrieved using different hashtags such as: #happy, #joy, #lucky and #sadness, #angry, #frustrated, respectively). As features they considered interjections and emoticons as

well as some resources like LIWC[8] and WordNetAffect[9]. The authors attempted to distinguish between sarcastic, positive and negative tweets. They applied support vector machine and logistic regression as classifiers. Their reported results are related to both datasets; the overall accuracy rate was around 0.57. The authors suggested that their results demonstrate the difficulty of sarcasm detection for both humans and machine learning methods.

According to Riloff et al. [127], a common form of sarcasm in Twitter consists of a positive sentiment contrasting with a negative situation (e.g. *absolutely adore it when my bus is late #sarcasm*). The goal of the authors' research was to recognize sarcasm instances containing this pattern[10]. The authors presented a bootstrapping algorithm that automatically learns phrases corresponding to negative situations. As sarcastic instances for the learning process, tweets that contained a sarcasm hashtag were retrieved. From the bootstrapping process authors collected some positive sentiment verb phrases, predicative expressions and negative situation phrases. They also carried out some binary classification experiments using a support vector machine (SVM) classifier. They used a set of features that contain not only their list of phrases but also n-grams and three sentiment and subjectivity lexicons (Hu&Liu, AFINN and MPQA[11]). The best result (F-measure 0.51) was achieved by an hybrid approach where a tweet is considered as sarcastic if either it contains a contrast (according to their list of phrases) or it is identified as such by the SVM (with unigram and bigram features).

Wang [147] presented a study to identify similarities and distinctions between irony and sarcasm. The study was addressed by a quantitative sentiment analysis and a qualitative content analysis. A set of sarcastic and ironic tweets collected by "self-tagging" approach was used. She found that sarcastic tweets were more positive that ironic ones.

Barbieri et al. in [18] attempted to study the differences between ironic and sarcastic tweets. They addressed the problem as a binary classification task between tweets tagged with #irony and #sarcasm hashtags. Their system is similar to the one presented in [17] for irony detection; they included two new features in their model: if a tweet contains an URL and named entities. The model was evaluated using a decision tree as a classifier. The authors obtained a F-measure of 0.62; this result emphasizes the difficulty to distinguish between irony and sarcasm. The authors mention the two most relevant features to distinguish between ironic and sarcastic tweets: the use of adverbs (more intense one in sarcastic samples) and the sentiment value (sarcastic tweets are denoted by more positive words than ironic ones).

Fersini et al. [47] addressed the sarcasm detection by introducing an ensemble approach (the Bayesian Model Average (BMA)). As features the authors used emoti-

[8] `http://www.liwc.net`

[9] `http://wndomains.fbk.eu/wnaffect.html`

[10] To identify "stereotipically" perceived negative situations represents per se a big challenge.

[11] `http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/`

cons, punctuation marks, onomatopoeic expressions, part-of-speech labels, as well as bag-of-words. The authors collected a set of tweets using the #sarcasm and #sarcastic hashtags, then three annotators were asked to determine the presence of sarcastic content in tweets. Besides, the authors evaluated the ensemble method over the corpus presented in [126]. Their results, around to 0.75 in both corpora, seem to indicate that the use of this strategy outperforms those obtained by traditional classifiers.

The approach in [121], a framework for sarcasm detection that uses a behavioral modelling approach, was developed by Rajadesingan et al. It defines some criteria in order to determine whether a tweet is sarcastic or not, by leveraging behavioral traits (using some of the user's past tweets) and textual-content features (such as punctuation marks, uppercase words, part-of-speech, among others). The authors collected tweets that contain the #sarcasm and #not hashtags as sarcastic instances; as negative instances the last 80 tweets from each sarcastic sample's author were retrieved. A binary classification task was performed between the sarcastic and non-sarcastic instances using decision tree, logistic regression and support vector machine as classifiers. Their results seem to be good, reaching rates above 0.70 in accuracy terms.

A similar approach is the one of Bamman and Smith in [13], who stated that modelling the relationship between a sarcastic tweet and the author's past tweets can improve accuracy. They presented some experiments to discern the effect of sarcasm by using features derived not only from the local context of the message itself (words in the tweet, part-of-speech, among others). They used also information about the author, the relationship with her audience and the immediate communicative context they both share (such as salient historical terms and topics and profile information). For evaluation purposes, all tweets with #sarcasm or #sarcastic in the GardenHose sample of tweets in the August 2013 - July 2014 period were used as sarcastic instances, while for the non-sarcastic ones the 3,200 most recent tweets from each "sarcastic author" (i.e. the user who posted a tweet labeled with #sarcasm or #sarcastic in the subset) were retrieved. As classifier a binary logistic regression was employed, achieving an accuracy of 0.851.

To sum up, there is a consistent body of work focused on sarcasm detection. It is a controversial issue whether irony and sarcasm are considered or not similar linguistic phenomena. Almost the same features used for irony detection have been employed also for sarcasm detection. Among the most widely applied features we can mention punctuation marks and part-of-speech labels. As classifiers, logistic regression and support vector machine have been the most used ones for sarcasm detection. Recent approaches on sarcasm detection consider information beyond the text itself exploiting contextual information and information about the user.

2.3 Figurative Language and Sentiment Analysis

In the recent years the interest in understanding the role of irony and sarcasm in SA has derived in different evaluation campaigns. Their main objective is not to identify ironic or sarcastic content, but to develop systems that will be able to correctly classify the polarity of figurative language social media texts. In fact the presence of figurative language devices such irony and sarcasm usually causes a polarity reversal. Irony and sarcasm detection become a necessary and important part for a sentiment analysis system because the performance of the latter is affected by the former one. In [99] the authors carried out an experiment to measure the effect of sarcasm on the polarity of tweets. They proposed a set of rules to improve the accuracy of sentiment analysis when sarcasm is present.

Following, three different evaluation campaigns are introduced. In Section 2.3.1 we describe a pilot subtask to identify ironic content. A sentiment classification task in Twitter over both sarcastic and non-sarcastic social media text is presented in Section 2.3.2. Finally, the recent sentiment analysis task wholly dedicated to figurative language in Twitter is described in Section 2.3.3.

2.3.1 Sentiment Polarity Classification at Evalita 2014

In the context of Evalita[12] 2014, the SENTIment POLatity Classification task (SENTIPOLC) [20] was organized. Its main focus was the sentiment classification at the message level of Italian tweets. The task was divided in three independent subtasks: (i) subjectivity classification; (ii) polarity classification and (iii) irony detection. Participants were provided with a dataset composed by a collection of 6,448 tweets in Italian (70% for training and 30% for test) derived from two existing corpora: SENTI-TUT [25] and TWITA [21]. Each tweet in the dataset was labeled according to subjectivity (subjective or objective), polarity (positive, negative, neutral or mixed) and to the presence of ironic content. The systems were evaluated using the F-score measure for each subtask. A total of eleven teams participated in the SENTIPOLC task (further information about each system can be found in [22]). Table 2.1 summarizes the results obtained[13] by the systems that participated in the irony detection task.

All the participants outperform the established baseline. The performance rates in F-measure for both subjectivity and polarity classification, achieved values near to 0.70 while on subtask 3 the higher values were below 0.60. This confirms the difficulty of the ironic-content related subtask. The best ranked team for the first two subtasks (UNIBA2930 [19]) did not participate in the irony detection task (see Table

[12]Evalita is an initiative devoted to the evaluation of Natural Language Processing and Speech tools for Italian: http://www.evalita.it/

[13]For each task, two runs could be submitted: Constrained (using the provided training data only) and unconstrained (using additional data for training). Table 2.1 presents results for the Constrained run; only three teams (UNIBA2930, UNITOR and IRADABE) participated with both a constrained and unconstrained run on the three subtasks.

2.1). No system was developed for addresing particularly the irony detection subtask.

Table 2.1. SENTIPOLC task results in F-measure terms

Team	Task 1	Task 2	Task 3
UNIBA2930	**0.71**	**0.67**	–
UNITOR	0.68	0.62	**0.57**
IRADABE	0.67	0.63	0.54
SVMSLU	0.58	0.60	0.53
itagetaruns	0.52	0.51	0.49
mind	0.59	0.53	0.47
fbkshelldkm	0.55	0.56	0.47
UPFtaln	0.64	0.60	0.46
baseline	0.40	0.37	0.44

Most systems used supervised learning and the support vector machine algorithm was the most popular. One further challenge for this task was the lack of Italian resources as well as NLP tools (such as tokenizers, part-of-speech (PoS) taggers, etc.); however, some systems (for instance UNIBA2930 and IRADABE) translated some of the resources available in English into Italian. For classification purposes a variety of features were used like bag-of-words, punctuation marks, emoticons and Twitter language markers (such as hashtags and mentions). UNITOR [36], the best ranked system in irony detection, proposed an "ironic vector" that captures the presence of some features such as punctuation marks, emoticons, bag-of-words and Sentix[14], to train a SVM classifier. IRADABE [75] exploited two different set of features: textual (n-grams, emoticons, PoS, uppercase words, among others), and information extracted from the in house Italian version of English resources such as: AFINN, SentiWordNet (SWN)[15], Hu&Liu, DAL, and Temporal Compression and Counterfactuality terms[16] together with a SVM classifier. The SVMSLU [6] system addressed the problem using a SVM for classifying binary vectors of tokens together with punctuation marks, hashtags and re-tweet marks. In itagetaruns [42] a set of linguistic rules was defined to classify the tweets; the author considered some markers such as: intensifiers and diminishers and modal verbs. The Mind system [46] is based on a multi-layer Bayesian ensemble learning; the authors addressed the task under a hierarchical framework. If a given sentence is detected as ironic, then its positive or negative polarity is reversed. On the other side, if the sentence is ironic, but its polarity has been classified as mixed, then it is switched to negative. The system only takes into account a vector composed by terms for which a boolean weight was computed, no additional information was added. Finally, the UPFtaln [15] system addressed the task by a decision tree classifier. Their approach is similar to the one presented in [17] for irony detection. The main difference is the use of Italian resources: the Italian WordNet 1.6 [17], Sentix and the CoLFIS corpus[18].

[14] http://wikis.fu-berlin.de/pages/viewpage.action?pageId=671548598

[15] http://sentiwordnet.isti.cnr.it/

[16] The last three resources have been previously used for irony detection in English by Reyes et al. [126]

[17] http://multiwordnet.fbk.eu/english/home.php

[18] http://linguistica.sns.it/CoLFIS/Homeeng.htm

2.3.2 Sentiment Analysis in Twitter at SemEval 2014 and 2015

During the last years, as part of SemEval[19], a task on sentiment analysis in Twitter has been organized [129, 128, 106]. The participating systems were required to assign one of the following labels: positive, negative or objective (neutral). The organizers provided two datasets[20] for training and test, composed by social media texts, mainly from Twitter.

In both 2014 and 2015 editions participating systems have been evaluated also on a subset of sarcastic tweets. In 2014 a small set of tweets that contained the #sarcasm was added to the test set, whereas in 2015 a set of tweets were manually labeled as "sarcastic" by human annotators. In Table 2.2 among the 44 participating systems we show the seven best performing ones.

TABLE 2.2. Sentiment analysis task in F-measure terms for both regular and sarcastic tweets in 2014 edition.

System	Twitter2014	Sarcasm2014
TeamX	**70.96**	56.50
coooolll	70.14	46.66
RTRGO	69.95	47.09
NRC-Canada	69.85	**58.16**
TUGAS	69.00	52.87
CISUC_KIS	67.95	55.49
SAIL	67.77	57.26
...	...	...

The results obtained for the best ranked teams in the 2015 edition are shown in Table 2.2. The overall drop in F-measure between regular and sarcastic tweets is slightly minor than in 2014. From the tables it can be appreciated an important drop in the performance when the systems were evaluated on the sarcastic tweets.

Generally sentiment analysis systems obtain good results for regular content, but when the same systems is evaluated with sarcastic content its overall performance is affected. It has to be said that none of the proposed approaches directly tried to capture the sarcastic intention. All systems addressed the task as a supervised approach, taking into account features widely applied in SA tasks such as: bag-of-words, part-of-speech tags and punctuation marks. Some of the systems used well known resources such as: AFINN, Hu&Liu and SWN. A more detailed description of the shared task and the participating systems can be found in [129, 128].

[19]SemEval (Semantic Evaluation) is an ongoing series of evaluations of computational semantic analysis systems.

[20]More details about it can be found in [129, 128]

Table 2.3. Best results in sentiment analysis task results in F-measure terms for both regular and sarcastic tweets in 2015 edition

System	Twitter2015	Sarcasm2015
Webis	**64.84**	53.59
unitn	64.59	55.01
lsislif	64.27	46.00
INESC-ID	64.17	64.91
Splusplus	63.73	60.99
wxiaoac	0.63	52.22
IOA	62.62	**65.77**
...	...	...

2.3.3 Sentiment Analysis of Figurative Language in Twitter at SemEval 2015

Task 11 at SemEval-2015[21] was the first sentiment analysis task addressing figurative language devices such as: irony, sarcasm and metaphor. The goal of the task was not to directly detect any of the previously mentioned devices, but to perform sentiment analysis in a fine-grained scale ranging from -5 (very negative) to +5 (very positive). In fact, since irony and sarcasm are typically used to criticize or to mock, and thus skew the perception of sentiment toward the negative, it is not enough for a system to simply determine whether the sentiment of a given tweet is positive or negative [51]. The participants were asked to determine the degree in which a sentiment was communicated, rather than to assign a more general score (such as in the previously described tasks).

A corpus composed by three subsets of tweets was supplied to the participants: trial (1,025), training (8,000) and test (4,000). The corpus construction involved crowdsourcing and some tweets explicitly tagged with hasghtags as: #irony, #sarcasm, #not, #yeahright; or that contained words commonly associated with the use of metaphor (e.g. "literally" and "virtually"). Further information can be found in [51]).

Fifteen teams participated in the task on sentiment analysis of figurative language [22]. Table 2.4 shows the results of the seven best ranked systems according to the overall cosine similarity measure.

The best ranked system, called ClaC [109], showed robustness across different sentiment analysis related tasks [110][23]. ClaC is based on a pipeline framework that groups different phases, from pre-processing to polarity induction. It exploits

[21] `http://alt.qcri.org/semeval2015/`

[22] Some systems such as Clac, UPF-taln and EliRF participated also in the related task on sentiment analysis in Twitter in Semeval 2015.

[23] ClaC obtained the 9^{th} best performance in both regular and sarcastic tweets in the task on sentiment analysis in Twitter [128]

TABLE 2.4. Best results in the task on sentiment analysis of figurative language in Twitter (cosine similarity measure)

Team	All	Irony	Sarcasm
ClaC	**0.758**	0.904	0.892
UPF	0.711	0.873	0.903
LLT_PolyU	0.687	**0.918**	0.896
EliRF	0.658	0.905	**0.904**
LT3	0.658	0.897	0.891
ValenTo	0.634	0.901	0.895
HLT	0.630	0.907	0.887
...	...	...	...

some resources such as NRC-lexicon[24], Hu&Liu and MPQA. In addition the authors developed a new resource called Gezi (for more details see [109, 110]). The main difference between their proposal for both tasks was the machine learning algorithm used for polarity assignment, a SVM for the regular one and M5P (a decision tree regressor) for figurative language tweets. Nevertheless it did not achieve the best performance neither for ironic nor sarcastic tweets in the figurative language task. The UPF-taln [16] system presented an extended approach that considered frequent, rare, positive and negative words and also exploited bag-of-words as features. To assign the polarity degree, the authors used a regression algorithm (Random-Sub-Space with M5P). Their system achieved the second place in the overall ranking.

Two similar and efficient approaches were the ones proposed by LLT_PolyU [155] and EliRF [57] both scored the best results in irony and sarcasm detection, respectively. LLT_PolyU and EliRF considered as features n-grams, negation scope windows and sentiment resources (LLT_PolyU exploited Hu&Liu, MPQA, AFINN and SWN; while EliRF used Pattern[25], AFINN, Hu&Liu, NRC-lexicon and SWN). In both systems, regression models (RepTree in LLT_PolyU and regression SVM in EliRF) were used to calculate the polarity value.

LT3 [142] and ValenTo [73] systems included in their set of features the presence of punctuation marks, emoticons and hasthags. To capture potential clues of figurative content in tweets LT3 took advantage of features to detect changes in the narrative as well as contrasting, contradictory and polysemic words. In the LT3 system a SVM classifier was used to determine the polarity value of tweets. Furthermore, ValenTo system exploits sentiment analysis resources (such as AFINN, Hu&Liu, General Inquirer[26] and SWN) as well as some containing emotional and psycholinguistic information (ANEW[27], DAL, SenticNet[28] and LIWC[29]. Besides,

[24] http://www.saifmohammad.com/WebPages/ResearchInterests.html
[25] http://www.clips.ua.ac.be/pattern
[26] http://www.wjh.harvard.edu/~inquirer/
[27] http://csea.phhp.ufl.edu/media/anewmessage.html
[28] http://sentic.net/
[29] http://liwc.wpengine.com/

a feature to reverse the polarity valence of a tweet when it contains a sarcastic intention was considered. In ValenTo a linear regression model was used to assign the polarity value. Finally, the HLT system [51] used a SVM approach together with lexical features such as negation and intensifier and some markers of amusement and irony.

2.4 Future Trends and Directions

Irony and sarcasm detection have been addressed as a text classification task. Salient features such as lexical marks are mainly used to characterize ironic and sarcastic utterances. As figurative language devices, irony and sarcasm need to be studied beyond the scope of the textual content of the utterance. In this regard both the context in which utterances are expressed and common knowledge should be considered to identify the real intention behind an ironic or sarcastic expression. There are some attempts to take advantage of this kind of information. Wallace et al. [145] exploited contextual information of the forum where a comment was posted. Information about users who wrote sarcastic tweets (such as their past tweets) has been considered by Rajadesingan et al. [121] and Bamman and Smith [13] in order to distinguish between sarcastic and non-sarcastic tweets.

Besides, it is necessary to consider how affective and emotional content is implicitly embedded in irony and sarcasm. Some works in literature started already exploiting affective information by using sentiment and affective lexica such as the Dictionary of Affect in Language [126, 74], Afinn and Hu&Liu [74], and SentiWordNet [17, 18].

With regards to the impact on irony and sarcasm detection on sentiment analysis, before to determine the polarity of an utterance it would be helpful to identify if the utterance expresses either ironic or sarcastic intention. Further investigation is needed in order to develop approaches that could efficiently identify ironic and sarcastic content to avoid misclassifying the polarity score of a subjective text.

2.5 Conclusions

People communicate their ideas in complex ways. Figurative language devices such as irony and sarcasm are often used in order to express evaluative judgments in an unconventional way. Irony and sarcasm are concepts difficult to define, however they are often used in social media. In this sense user-generated content represents a big challenge. The progress so far achieved in irony and sarcasm detection has been a result of exploiting mainly the syntactic, lexical and semantic levels of natural language processing. Similar approaches have been proposed for addressing the task as a binary classification. At this moment, the biggest effort concerns to identify the most salient features that allow to determine when the intended content of an utterance is ironic or sarcastic.

From the sentiment analysis perspective, the presence of irony and sarcasm affects the performance of the task. As we pointed out, state-of-the-art systems generally have good results when dealing with regular content, but when they are evaluated with ironic or sarcastic content its overall performance is affected. Therefore, robust sentiment analysis systems will need to understand when human communications in social media make use of figurative language devices such as irony and sarcasm.

Acknowledgments

The National Council for Science and Technology (CONACyT Mexico) has funded the research work of Delia Irazú Hernández Farias (Grant No. 218109/313683 CVU-369616).

Chapter 3

Applying Basic Features from Sentiment Analysis for Automatic Irony Detection

In this chapter we introduce an irony detection model that considers a set of lexical markers together with two sentiment-related resources. We include as features for characterizing irony in Twitter two values related to the overall sentiment expressed in a given tweet: sentiment score and polarity value. We experiment with different classifiers. The obtained results outperforms the state of the art showing the importance of using sentiment-related features.

Published in:

Hernández Farías, D.I., Benedí, J.M., and Rosso, P. (2015). Applying Basic Features from Sentiment Analysis for Automatic Irony Detection. In: *Proceedings of the Pattern Recognition and Image Analysis: 7th Iberian Conference, IbPRIA 2015, Santiago de Compostela, Spain, June 17-19, 2015*. Springer International Publishing, pages 337-344.
DOI: `http://dx.doi.org/10.1007/978-3-319-19390-8_38`

Abstract

People use social media to express their opinions. Often linguistic devices such as irony are used. From the sentiment analysis perspective such utterances represent a challenge being a polarity reversor (usually from positive to negative). This paper presents an approach to address irony detection from a machine learning perspective. Our model considers structural features as well as, for the first time, sentiment analysis features such as the overall sentiment of a tweet and a score of its polarity. The approach has been evaluated over a set classifiers such as: Naïve Bayes, Decision Tree, Maximum Entropy, Support Vector Machine, and for the first time in irony detection task: Multilayer Perceptron. The results obtained showed the ability of our model to distinguish between potentially ironic and non-ironic sentences.

Keywords: automatic irony detection, figurative language proceesing, sentiment analysis.

3.1 Introduction

The ability to recognize ironic intent in utterances is performed by humans in a relatively easy way although not always. We develop this ability since childhood and, over years with social interaction we increase it. In many cases we are able both to understand and to produce such utterances without a strict definition of what is or may be considered an ironic expression. Irony is a sophisticated, complex and prized mode of communication; it is intemately connected with the expression of feelings, attitudes or evaluations [10]. Moreover, irony can be considered as a strategy, which is intented to criticise or to praise. Sometimes but not always, it means the opposite of the literal meanings; generally irony shows or express some kind of contradiction [4].

Recently interest for discover information in social media has been growing. Twitter, offers a face-saving ability that allows users to express themselves employing linguistic devices such as irony. User-generated content is difficult to analyse: Internet language is hard to analyze due to the lack of paralinguistic cues; in addition one needs to have a good understanding of the context of the situation, the culture in question, and the people involved [99]. For research areas such sentiment analysis (SA), irony detection is important to avoid misinterpreting ironic statement as literal [144].

For computational linguistic purposes, most of the time irony and sarcasm are often viewed as the same figurative language device. Irony is often considered as an umbrella term that covers also sarcasm [147]. Previous works are mainly based on the classification of tweets as ironic or sarcastic and rely solely on text analysis.

This paper presents an approach for irony detection using a set of features that combine both surface text properties and information exploited from sentiment analysis lexicons. The main contribution of this paper is to take advantage of the classification of utterances according to their polarity. We consider in order to detect irony it is important to take into account the sentiment expressed in a tweet. Our model improves state-of-the-art results. The rest of this article is organized as follows: previous works on automatic irony detection are introduced in Section 3.2. In Section 3.3 we describe the set of features used. In Section 3.4, dataset, classifiers, experimental setting and evaluation of our approach are presented. Finally, in Section 3.5 we draw some conclusions and discuss future work.

3.2 Related Work

Recently automatic irony detection has attracted the attention of researchers from both machine learning and natural language processing [144]. A shared task on figurative language processing has been organized at SemEval 2015[51][1].

A survey that includes both philosophical and literary works investigating ironic communication and some computational efforts to operationalize irony detection is presented by Wallace in [144]. Reyes et al. [126] address the problem of irony detection as a classification task; the authors proposed a model employing to four types of conceptual features: signatures, unexpectedness, style and emotional scenarios. Bosco et. al. in [25] present a study that investigates sentiment and irony in online political discussion social media in Italian. Buschmeier et al. [29] present an analysis of 29 features (such as punctuation marks, emoticons, interjections and bag-of-words); the authors' main goal is to investigate the impact of features removal on the performance of their approach. Barbieri and Saggion [17] used six groups of lexical features (frequency, written-spoken, intensity, structure, sentiments, synonyms, ambiguity), in order to classify ironic tweets (the same dataset of [126] was used).

3.3 Proposed Features

We address irony detection as a classification problem, considering different types of features. In our model, we consider some features previously applied in irony detection. Moreover, we propose two sentiment analisys features (*Sentiment Score*

[1]Given a set of tweets the task consist in determining whether the user has expressed a positive, negative or neutral sentiment; more information is available at: `http://alt.qcri.org/semeval2015/task11/`

and *Polarity Value*) in order to take advantage of resources that allow to measure the overall sentiment expressed in each tweet. We can distinguish the set of features into *Statistical-based* and *Lexical-based*. *Statistical-based* are surface patterns that can be obtained taking into account the frequency of some words or characters in the tweet. *Lexical-based* are obtained by using information beyond the textual content of the tweet, i.e. applying external resources.

The first set, ***Statistical-based*** features is composed of four dimensions: **a)** *Textual Markers (TM)*, features widely used in this task, which include frequency of visual cues as: lenght of tweet, capitalization, punctuation marks, and emoticons[2]; **b)** *Counter-Factuality (CF)*[3], the frequency of discursive terms that hint at opposition or contradiction in a text such as "nevertheless"[4]; **c)** *Temporal Compression (TC)*[3], the frequency of terms that identify elements related to opposition in time, i.e. terms that indicate an abrupt change in a narrative; and **d)** *POS-based features (POS)*, where each tweet has been processed using a POS-tagger developed for this kind of texts called ARK[5]; we take into account frequency of verbs, nouns, adjectives and adverbs.

Our second set of features, ***Lexicon-based***, exploits different knowledge bases to represent each tweet: **a)** *Semantic Similarity (SIM)*[3], consists in obtaining the degree of inconsistency measuring the relationship between the concepts contained in each tweet using the WordNet::Similarity[6] module; **b)** *Emotional Value (EV)*[3], where the emotional value is calculated taking into account the categories described by Whissel [150], in her Dictionary of Affect in Language (DAL)[7]. **c)** *Sentiment Score(SS)*, in order to catch the overall sentiment (positive, negative or neutral) expressed in a tweet. We applied a lexicon developed by Hu-Liu in [76] [8]; and **d)** *Polarity Value(PV)*, this feature allows to identify the rate of evaluation, either to criticize (negative) or to praise (positive). We use AFINN[9] lexicon, which contains a list of words labelled with a polarity valence value between minus five (negative) and plus five (positive) for each word.

The last two features in this set (*Sentiment Score(SS)* and *Polarity Value(PV)*) have not been previously used in irony detection. Our main motivation to use sentiment analysis features is that an ironic utterance is subjective, hence contains a

[2]Using emoticons, with few characters is possible to display one's true feeling; sometimes they are virtually required under certain circumstances in text-based communication, where the absence of some kind of cues can hide what was originally intended to be humorous, sarcastic, ironic, and often negative [153].

[3]Feature previously applied by Reyes et al. [126]

[4]The complete list of words can be downloaded from `http://users.dsic.upv.es/grupos/nle`

[5]`http://www.ark.cs.cmu.edu/TweetNLP/`

[6]`https://codegoogle.com/p/ws4j/`. This module allows to calculate a set of seven different similarity measures.

[7]DAL is composed by 8,000 English words, distributed in three categories: *Activation*, refers to the degree of response, either passive or active, that humans exhibit in an emotional state; *Imagery*, quantifies how easy or difficult is to form a mental picture for a given word; and *Pleasantness*, quantifies the degree of pleasure suggested by a word.

[8]`http://ww.cs.uic.edu/~liub/FBS/`

[9]`http://github.com/abromberg/sentiment_analysis/blob/master/AFINN/AFINN-111.txt`

positive or negative opinion. On the other hand, we taking into account a feature that allows us obtaining a polarity value from each tweet, so we have both the "overall" sentiment and a score of the polarity. In sentiment analysis, there are several resources that could help to improve the detection of ironic tweets.

3.4 Experiments and Results

The dataset used in this work was compiled by Reyes et al. [126] and consists of a total of 40,000 tweets written in English, distributed in four different classes: Irony, Education, Humor and Politics. The corpus was built retrieving 10,000 tweets that contain one of the following hashtags: #irony, #education, #humor and #politics. These hashtags allow to have tweets in which users explicitly declare their ironic attempt, and a large sample of non-ironic tweets. In order to perform classification process, we apply a set of classifiers widely used in text classification tasks. Some of them has been used in irony identification. The set of classifiers[10] is composed by: Decision Tree (*DT*), Maximum Entropy (*ME*), Naïve Bayes (*NB*), Random Forest (*RF*) and Support Vector Machine (SVM, with a RBF kernel)[11] and Multilayer Perceptron (*MLP*, we used a backpropagation based multilayer perceptron, with sigmoid functions, a learning rate of 0.3 and 500 epochs in each run; we did not perform any parameter tuning.). In this paper we propose to apply *MLP*, that has never been used for irony detection.

As in [17] and [126], we perform a set of binary classifications between Irony and Education/Humor/Politics. Each experiment has been performed in a 10-fold-cross-validation setting. We run experiments for one baseline: *Bag Of Words (BOW)*. We exploit only most frequent unigrams per class (1,000) in order to represent each tweet. This baseline relies on standard text classification features. According to [144], words counts alone offer an insufficient representation for verbal irony detection.

We apply two different vector representation approaches for experimental purposes. Each tweet was converted to a vector composed by 16 features. No feature selection technique was performed. In the first approach the features belonging to *Statistical-based* were taking into account the frequency of each one; while *Lexicon-based* are represented in different ways: the semantic similarity is the value obtained using the above-mentioned module; emotional value is calculated taking into account values in DAL over words that compose each tweet; the sentiment score can be *positive* (more positive than negative terms), *negative* (more negative than positive terms) or neutral (same amount of positive and negative terms); finally, the polarity value is assigned by calculating the difference between the positive and the negative polarity of each tweet according to AFINN lexicon.

[10] We used Weka toolkit's version of each classifier available at `http://www.cs.waikato.ac.nz/ml/weka/downloading.html`

[11] Default parameters for each algorithm were used

In the second approach we applied the representativeness criterion presented by Reyes et al [126] in order to assign a value for *Statistical-based* features; the representativeness of a given document d_k (e.g. a tweet) is computed according to:

$$\delta_{i,j}(d_k) = \frac{f_{ij}}{|d_k|} \tag{3.1}$$

where i is the i-th feature; j is the j-th dimension; f is the feature dimension frequency; and $|d_k|$ is the lenght of the k-th document d_k. If $\delta_{i,j}(d_k)$ is $\geq$0.5, a value of 1 is assigned; otherwise, a representativeness value of 0 (not representative at all) is assigned; and the *Lexicon-based* features were represented as the same way above described for the first approach.

Three experiments were carried out using the classification algorithms mentioned above. Each experiment are constructed under different criteria. Two of them (***Lesk*** and ***Wu-Palmer***) are based in the first representation approach while the third (***Rep, Representativeness***) takes into account the second approach. The difference between *Lesk* and *Wu-Palmer* is the semantic similarity[12], that take into account, using Lesk and Wu-Palmer measures respectively.

In Table 3.1, we report F-measure results of our classification experiments. It can be observed that all results overcome the baseline. The bold values are used to highlight those F-measures greater than state-of-the-art (See Table 3.3). The best result is achieved by *SVM* in the three sub-tasks (binary classification Irony vs. Education, Irony vs. Humor and Irony vs. Politics). As reported by [17] and [126], higher results in F-measure are achieved by *ironic-vs-politics* classification, while lower F-measure lie in *ironic-vs-humor*. We carried out the t-test (with a 95% confidence level) in order to see if the best results are statistically significant.

Moreover, we calculated the *Classification Error Rate (CER)*. In Table 3.2 CER values for each binary classification (*Iro-Edu*, *Iro-Hum* and *Iro-Pol*) are presented. As can be seen, our model obtains satisfactory CER rates. The best results (bold values in Table 3.2) are obtained by: SVM, MLP and RF.

[12]We performed experiments using each similarity measure of the WordNet::Similarity module. Due to lack of space, we report only the results with highest classification rates. The similarity measures are described in detail in [112].

TABLE 3.1. Results in F-measure for the baseline and each representation approach corresponding to binary classification. The underlined values are statistically significant.

	BOW	Lesk	Wu-Palmer	Rep
DT	0.34	**0.78**	**0.78**	0.68
ME	0.37	**0.75**	**0.75**	0.66
MLP	0.50	**0.78**	**0.78**	0.67
NB	0.44	0.70	0.70	0.66
RF	0.16	**0.79**	**0.79**	0.68
SVM	0.63	**0.80**	**0.80**	0.68

(A) **Irony-Education**

	BOW	Lesk	Wu-Palmer	Rep
DT	0.34	0.75	0.74	0.70
ME	0.37	0.74	0.74	0.69
MLP	0.50	0.75	0.76	0.70
NB	0.46	0.69	0.70	0.65
RF	0.16	0.76	0.76	0.70
SVM	0.59	**0.77**	**0.78**	0.69

(B) **Irony-Humor**

	BOW	Lesk	Wu-Palmer	Rep
DT	0.34	**0.79**	**0.79**	0.63
ME	0.36	**0.76**	**0.76**	0.95
MLP	0.50	**0.79**	**0.79**	0.61
NB	0.45	0.70	0.71	0.57
RF	0.16	**0.81**	**0.81**	0.63
SVM	0.64	**0.81**	**0.80**	0.63

(C) **Irony-Politics**

As mentioned above, the dataset has been used before ([17] and [126]). The results reported by their authors are shown in Table 3.3. In both works a Decision Tree classifier was used. The last two rows in the table correspond to our results using the Decision Tree classifier.

As Table 3.3 shows, our approach improves the F-measure obtained previously by state-of-the-art approches. In order to determine which features are more relevant in our model, Information Gain[11] was calculated. There are some features that seem to contribute more than others in our model to discriminate between classes (see Figure 3.1). As can be seen, the textual markers (TM) features are a good indicator of this kind of utterances. Moreover, also the sentiment analysis features (SS and PV) showed to have an important impact on irony detection. This strenght the idea that irony detection is strongly related to sentiment analysis. According to Figure 3.1, features related to SA seem to be quite important to identify ironic from non-ironic tweets. From this we may say that using features and resources for SA could improve performance of models for irony detection.

TABLE 3.2. Results in terms of CER.

	BOW	Lesk	Wu-Palmer	Rep
DT	66.01	43.65	43.75	63.12
ME	62.58	49.88	49.93	67.46
MLP	50	43.8	42.87	64.76
NB	55.18	59.62	59.31	66.31
RF	84.11	40.5	40.71	63.22
SVM	55.18	**40.1**	**40.15**	63.65

(A) **Irony-Education**

	BOW	Lesk	Wu-Palmer	Rep
DT	65.1	49.67	51.72	59.58
ME	62.48	50.51	50.64	60.17
MLP	50	48.25	**42.87**	60.07
NB	53.68	60.43	59.27	68.82
RF	84.31	46.19	45.97	49.71
SVM	55.08	**44.17**	44.07	60.7

(B) **Irony-Humor**

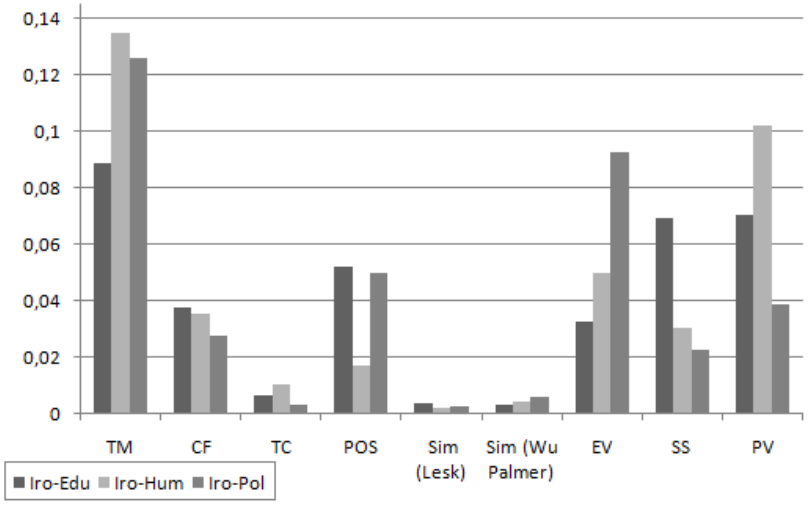
	BOW	Lesk	Wu-Palmer	Rep
DT	65.41	41.05	41.36	72.22
ME	63.13	46.82	46.59	79.24
MLP	50	40.82	40.96	76.53
NB	54.91	53.36	57.46	77.55
RF	84.2	**37.09**	**36.88**	72.59
SVM	54.46	37.93	37.88	73.82

(C) **Irony-Politics**

TABLE 3.3. Results in F-measure of our model against state-of-the-art

	Irony vs.		
	Education	***Humor***	***Politics***
Reyes et al.	0.70	0.76	0.73
Barbieri and Saggion	0.73	0.75	0.75
Our approach *Lesk*	**0.78**	0.75	**0.79**
Our approach *Wu-Palmer*	**0.78**	**0.79**	**0.79**

FIGURE 3.1. Information Gain for our set of features



3.5 Conclusions

Given the growing interest in exploiting knowledge generated in social media, irony detection has attracted the attention of different research areas. Different approaches have been proposed to tackle this task. In this paper we proposed a model for ironic tweets classification, taking advantage for the first time of sentiment analysis features. One of the best results was obtained by *MLP*, a method has not been previously used for irony detection. Also in terms of CER, our model showed good performance in classification rates of ironic tweets in the experiments we carried out. As future work an in-depth analysis of the impact of the proposed features is needed. We plan to exploit further features and resources from sentiment analysis.

Acknowledgments

The National Council for Science and Technology (CONACyT Mexico) has funded the research work of the first author (Grant No. 218109/313683, CVU-369616). The research work of third author was carried out in the framework of WIQ-EI IRSES (Grant No. 269180) within the FP 7 Marie Curie, DIANA-APPLICATIONS (TIN2012-38603-C02-01) projects and the VLC/CAMPUS Microcluster on Multimodal Interaction in Intelligent Systems.

Chapter 4

Irony Detection in Twitter: The Role of Affective Content

In this chapter we introduce emotIDM, a novel irony detection model that exploits affective information for characterizing ironic utterances in Twitter. emotIDM takes advantage of a wide range of resources covering several aspects of affective phenomena: from sentiment to finer-grained emotions. emotIDM performs well under different conditions related to corpora. The obtained results validate the importance of affect-related features for detecting irony in social media.

Published in:

Hernández Farías, D.I., Patti, V., and Rosso, P. (2016). Irony Detection in Twitter: The Role of Affective Content. *ACM Transactions on Internet Technology* 16(3), pp. 1-24.
DOI: `http://dx.doi.org/10.1145/2930663`

Abstract

Irony has been proven to be pervasive in social media, posing a challenge to sentiment analysis systems. It is a creative linguistic phenomenon where affect-related aspects play a key role. In this work, we address the problem of detecting irony in tweets, casting it as a classification problem. We propose a novel model which explores the use of affective features based on a wide range of lexical resources available for English, reflecting different facets of affect. Classification experiments over different corpora show that affective information helps in distinguishing among ironic and non-ironic tweets. Our model outperforms the state-of-the-art in almost all cases.

Keywords: Irony Detection, Figurative Language Processing, Affective Resources

4.1 Introduction

The huge amount of information streaming from online social networking and micro-blogging platforms such as Twitter, is increasingly attracting the attention of researchers in the area of sentiment analysis. Twitter communications include a high percentage of ironic devices [41, 143, 65, 126, 124], and platforms monitoring the sentiment in Twitter messages experienced the phenomenon of wrong polarity classification of ironic messages [25, 51]. Indeed, the presence of ironic devices in a text can flip the polarity of an opinion expressed with positive words to the intended negative meaning (one says something "good" to mean something "bad") – or vice versa – working as an unexpected polarity reverser. This can undermine systems' accuracy. The automatic detection of irony is, therefore, crucial for the development of irony-aware sentiment analysis systems, but at the same time it is also an interesting conceptual challenge from a cognitive point of view and can help to shed some light on how human beings use irony as a communicative tool.

Irony has been a topic studied by various disciplines, such as linguistics, philosophy, and psychology, but it is difficult to define it in formal terms. There is no consensus on a single definition and different accounts shed light on relevant aspects of a creative and complex linguistic phenomenon. However, most theorists would agree that emotions play a role in the use of irony in different respects, and the important role of affective information for irony communication-comprehension is also emphasized by recent psychological findings [91, 133].

Linguistic devices such as irony and sarcasm allow users to express themselves by using words in a creative and non-literal sense. They are intimately connected with the expression of affective contents such as feelings, emotions, attitudes or evaluations [66, 151, 5] towards a particular target (e.g. a person, an event, but also a product or a movie when we consider social media texts). In irony people express affective contents in an indirect way, since the critical or praising attitudinal load they communicate is on top of what they explicitly say. According to the Gricean tradition [66], the function of irony is to effectively communicate the opposite of the literal interpretation of the utterance. Furthermore, an ironic statement can elicit affective reactions. For instance, ironic criticism (or sarcasm) has been recognized in [26] with a specific target to attack, offensive [90], and "intimately associated with particular negative affective states" [100]. It may enhance the negative emotions felt by the recipient, such as anger, irritation, disgust [91], and it can be hypothesized that the use of such figurative device also conveys information on the speaker's attitude towards the target. On the other hand, there are cases where irony may reduce the strength of a statement, that is, criticism becomes gentler or less negative, and praise less positive or more ambivalent, if phrased ironically [43]. Overall, the affective information involved in ironic communications is multi-faced, involving aspects related to the emotional state of the ironist and of the recipient, and issues related to the evaluative meaning of the ironic utterance, i.e. to the expression of a positive or negative opinion towards a target.

There is now a consistent body of work on computational models for irony and sarcasm detection in social media [65, 126, 147, 127, 18, 120, 74], and in particular in Twitter, which can be considered the most widely used source of information to experiment with irony detection. In this article we also address the task of detecting irony in tweets, by identifying a set of discriminative features to automatically differentiate an ironic text from a non-ironic one. In line with most of the current approaches and with some theoretical accounts [54, 149], irony is here considered an umbrella term that covers also sarcasm, being the issue of discriminating between the two devices a further challenge for figurative language processing. Our irony detection model, called emotIDM, extends the model proposed in [74] with new features, in particular experimenting the use of a wide range of psycholinguistic and affective features concerning affective information, with the main aim to answer to our main research questions: (1) Does information about different facets of affect help in distinguishing among ironic and non-ironic tweets? (2) Which facets of affect seem to be more important in order to address our classification task? Affective information expressed in our texts is multi-faceted. Both sentiment and emotion lexicons, and psycholinguistic resources available for English, refer to various affective models. In our view, all such resources represent a rich and varied lexical knowledge about affect, under different perspectives. Therefore, we propose here a comprehensive study of their use in the context of our analysis, in order to test if they convey relevant knowledge to distinguish between ironic and non-ironic messages. To our knowledge, this is the first work that addresses the issue by considering different facets of the affective content, taking advantage of the wide availability of lexical resources for English covering the various perspectives. Such facets include sentiment polarity aspects related to the polarity of words, but also finer-grained ones, related

to the writer's emotional state or to emotions evoked in the reader, which can be captured according to different categorical or dimensional models of emotions.

Another novelty of our proposal is that we evaluated our model over six different Twitter corpora developed in previous work on irony and sarcasm detection, without creating our own dataset. This is important not only in order to carry out a fair evaluation of our model against the state-of-the-art approaches, but also to test the robustness under different datasets, were samples of ironic utterances were collected by using different criteria (i.e. different hashtags).

The evaluation of our model for irony detection over a set of Twitter corpora already used in the same task confirms the significance of affective features for irony detection. Experimental results show that emotIDM outperforms the irony detection models presented in [127, 126, 18, 74] over the same datasets.

Contributions. Summarizing, the main contributions of this paper are the following: a) We propose a new approach to irony detection emotIDM based on [74] that exploits affective information as features to represent ironic tweets; b) We evaluate emotIDM carrying out a battery of binary classification experiments over a set of Twitter corpora, developed in different way both for what concerns the selection criteria for samples of irony/sarcasm and the annotation methodology. This is important in order to validate the robustness of the model and to better compare results with state-of-the-art; c) We demonstrate that affective information helps in distinguishing among ironic and non-ironic tweets, presenting a comparative evaluation of the performances over the various corpora, and a feature analysis in order to identify the most useful features in emotIDM.

Organization. The paper is structured as follows. Section 4.2 describes related work in irony detection. Section 4.3 presents a set of Twitter corpora developed in literature for evaluating previously proposed models in irony detection. Section 4.4 introduces our starting point, the IDM model in [74], and the new proposal, emotIDM, which enriches IDM with affective features. In Section 4.5, we describe a set of experiments carried out over the set of corpora by using both models for irony detection, as well as an information gain analysis to identify the most relevant features in emotIDM. Finally, in Section 4.6 we conclude with final remarks and future work.

4.2 Related Work

Different approaches to the task of recognizing verbal irony in texts have been developed. The majority of them take advantage only of the textual content itself, since in textual messages other paralinguistic cues, like for instance the tone or corporal movements, are not available. Twitter is the most widely used source of information to experiment with irony detection. This is mainly due to availability of a large set of samples of ironic texts, which are easy to be collected relying on the behavior of

Twitter users, who often explicitly mark their ironic messages by using hashtags such as '#irony' or '#sarcasm'. The pretty good reliability of the user-generated hashtags as golden labels for irony has been experimentally confirmed by [89]. Moreover, it seems that, due to the interaction model underlying the micro-blogging platform, irony expressed here could be somehow easier to analyze. Indeed, Twitter users have to be sharp and short, having only 140 characters for expressing their comments, and most of the times the ironic posts do not require knowledge about the conversational context to be understood. Several works have been carried out using tweets for experimental purposes [41, 65, 126, 147, 127, 18, 120, 74, 121, 13, 79, 81]. Furthermore, there are some efforts in other social media such as customer reviews from Amazon[1] [48, 29]; comments from the online debate sites such as 4forums.com[2] [2, 97] and, recently, Reddit[3] [145].

The majority of the research in irony detection has been addressed in English, although there is some research in other languages, such as: Dutch [89], Italian [25], Czech [120], French [81], Portuguese [35] and Chinese [139]. A shared task for English on sentiment analysis of figurative language in Twitter has been organized at SemEval-2015 for the first time [51], and a pilot shared task for Italian on irony detection has been proposed in Sentipolc-2014 within the periodic evaluation campaign EVALITA [20, 8]. This confirms the growing interest for this task in the research community, especially for understanding the impact of the ironic devices on sentiment analysis.

Irony detection has been modeled as a binary classification problem, where mostly tweets labeled with certain hashtags (i.e. #irony, #sarcasm, #sarcastic, #not) have been considered as ironic utterances. Following this framework, different approaches have been proposed [41, 65, 126, 127, 18, 120, 74, 47]. The authors proposed models that exploit mainly textual-content such as: punctuation marks, emoticons, part-of-speech labels, discursive terms, specific patterns (e.g., according to [127], a common form of sarcasm in Twitter consists of a positive sentiment contrasting with a negative situation), among others.

Another key characteristic for irony is *unexpectedness* [10]. According to many theoretical accounts people infer irony when they recognize an incongruity between an utterance and what is known (or expected) about the speaker and/or the environment. This is something that can be referred to as the pragmatic context. Recent approaches started to address such issue, taking into account information about context [121, 13, 145].

For what concerns the affective information, some approaches already used in their models some kind of sentiment and emotional information. Reyes et al. [126] included in their model some features to characterize irony in terms of elements related to sentiments, attitudes, feelings and moods exploiting the Dictionary of

[1] `http://www.amazon.com/`
[2] `http://www.4forums.com/political/`
[3] `http://www.reddit.com`

Affect in Language proposed by [150]. Barbieri et al. [18] considered the amount of positive and negative words by using SentiWordNet [12]. Hernández et al. [74] exploited two widely applied sentiment lexicons: Hu&Liu and AFINN[4] as features in their model. However, no previous work focused specifically on studying the role of affective information in a comprehensive manner, by exploring the use of a wide range of lexical resources available for English, reflecting different aspects of a multi-faceted phenomenon.

4.3 Evaluation Datasets

Annotated data are a crucial source of information to capture the real use of irony in social media. Large corpora providing annotations marking whether an expression is ironic or not are scarce [29, 139]. Therefore, in general, the authors have built their own corpora for evaluating the proposed models. This constitutes a problem for establishing a baseline, and this is the reason we decided to follow here a different approach, by evaluating our model against a set of already available Twitter corpora that have been developed in related work on irony detection. We observed that there are two main approaches which have been used for creating corpora for irony detection: self-tagging and crowd-sourcing.

Self-tagging. Twitter allows users to communicate ideas in short messages and to assign labels (i.e. hashtags) to their own messages. The "Self-tagging" approach considers as positive instances those tweets in which the author points out her intention using an explicit label. For instance the hashtags '#irony' and '#sarcasm' can be considered as markers of irony, which rely on the author's definition about what irony is. The underlying assumption is that the best judge of whether a tweet is intended to be ironic is the author of the tweet [65, 126]. Furthermore, some experiment shows that self-labeled tweets allow to produce good quality gold standards [89]. However, it is worth to be noticed that not in all languages Twitter users are used to mark explicitly by hashtags the intention to be ironic. For instance, both Czech and Italian users generally do not use the sarcasm (i.e. '#sarkasmus', in Czech; '#sarcasmo' in Italian) or irony ('#ironie' in Czech or '#ironia') hashtag variants, thus in such cases relying on simple self-tagging is not an option [120, 25].

Crowd-sourcing. The "Crowd-sourcing" approach involves human interaction by labeling the content as ironic or non-ironic. Mainly, the labeling process is carried out without any strict definition or guideline. Therefore, it represents a subjective task, where the agreement between annotators is often very low.

Below, we describe six corpora which have been created by using the methodologies depicted above. In [126, 18, 120] authors took advantage of the presence of hashtags to create the corpus and evaluate their models. Likewise, in [104] data were manually annotated by using crowdsourcing with information related to irony, and annotators were asked to decide whether a tweet is ironic or not, whereas in [127, 104] a mixed approach has been taken.

[4]Hu&Liu: `http://www.cs.uic.edu/~liub/FBS`; AFINN: `http://github.com/abromberg/sentiment_analysis/blob/master/AFINN/AFINN-111.txt`

TwReyes2013. In [126] the authors retrieved a set of 40,000 tweets by using the "Self-tagging" criterion. They selected four hashtags: #irony to get ironic instances (or at least tweets wrote by Twitter users with an intuitive definition of what an irony is) and #education, #humor and #politics to retrieve a large sample of non-ironic tweets on different topics. This corpus (henceforth TwReyes2013) contains 10,000 ironic tweets and 30,000 non-ironic tweets.

TwBarbieri2014. Barbieri et al. [18] introduced a Twitter dataset constructed following a methodology similar to [126]. Overall, it includes 60,000 tweets equally divided into six different classes: education, humor, politics, newspaper, irony and sarcasm. For what concerns the first three categories (education, humor and politics), authors reused samples from the TwReyes2013. The irony and sarcasm tweets were collected by using the #irony and #sarcasm hashtags, respectively. In the following, we will use TwIronyBarbieri2014 to refer to a corpus where irony-laden tweets are sampled by the irony class of TwBarbieri2014, whereas we will use TwSarcasmBarbieri2014 to denote a different corpus where they are sampled by the sarcasm class. In both corpora the non-ironic samples are tweets from the education, humor, politics, and newspaper classes.

TwRiloff2013. Riloff et al. [127] created a manually annotated corpus from Twitter including 3,200 tweets (henceforth TwRiloff2013). They followed a mixed approach for developing a corpus of samples including ironic and non-ironic tweets. First, a set of tweets tagged with the #sarcasm and #sarcastic hashtags as well as tweets without these hashtags were retrieved (self-tagging methodology). Then, three annotators were asked to manually annotate the collected tweets by omitting the hashtags. Annotation guidelines asked users to label a tweet as sarcastic, if it contains comments judged to be sarcastic based solely on the content of that tweet.

TwPtáček2014. In the work by Ptáček et al. [120] two datasets were collected: in Czech and English. The first one involved manual annotation of tweets[5]. Instead, for the English dataset the hashtag #sarcasm was used as an indicator of sarcastic tweets (henceforth TwPtáček2014); for the non-sarcastic samples the authors collected tweets from the general Twitter stream using as parameter only the language (English). Two different distribution scenarios were created for the English dataset: balanced (composed by 50,000 sarcastic and 50,000 non-sarcastic tweets) and imbalanced (composed by 25,000 sarcastic and 75,000 non-sarcastic tweets).

TwMohammad2015. The TwMohammad2015 corpus [104] contains a set of tweets with a multi-layer annotation concerning different aspects: sentiment (positive or negative), emotions (anger, anticipation, disgust, fear, joy, sadness, surprise and trust), purpose (to point out a mistake, to support, to ridicule, etc.) and style (simple statement, sarcasm, hyperbole, understatement). Noticed that, only the 23.01% of the tweets were labeled with a style tag pertinent to the expression of irony, whereas most of them were annotated with the label *simple statement*, which can be interpreted as a tag for marking non-ironic expressions. The authors collected tweets labeled with a set of hashtags pertaining to the 2012 US presidential elections[6]. The tweets were annotated by relying on crowdsourcing platforms.

[5]For more details about Czech dataset see [120].

[6]Some of the hashtags used are: #election2012, #election, #campaing2012, #president2012.

The next sections describe the experimental setting and results obtained over these corpora. A summary of their features in reported in Table 4.1[7]. Most of the above described corpora were created for evaluating irony detection models presented in related work. TwMohammad2015 is the only one designed for purposes which go beyond to irony detection, i.e. for predicting emotion and purpose labels in tweets. Most of the corpora rely on self-annotation of tweets, but we have also samples of corpora manually annotated by using crowdsourcing platforms. The datasets were developed based on criteria, which are different for what concerns the choice of the hashtags or the guidelines for manual annotation. Such variety of aspects makes particularly interesting to use all the datasets in order to evaluate our proposal, which is described in the next section. Indeed, our model will be evaluated by using tweets coming from different scenarios (for instance tweets in TwMohammad2015 pertain the political domain), collected with different methodologies. This allows us to test the robustness of the approach across a wide set of irony samples, which represent a rich variety of use of ironic devices.

Table 4.1. Evaluation datasets

Ironic	**Non-ironic**	**Labeling criterion**	**Hashtag**
TwReyes2013			
10,000	10,000 (#education) 10,000 (#humor) 10,000 (#politics)	Self-tagging	#irony
TwIronyBarbieri2014			
10,000	10,000 (#education) 10,000 (#humor) 10,000 (#politics) 10,000 (#newspaper)	Self-tagging	#irony
TwSarcasmBarbieri2014			
10,000	10,000 (#education) 10,000 (#humor) 10,000 (#politics) 10,000 (#newspaper)	Self-tagging	#sarcasm
TwPtáček2014			
19,026	51,860	Self-tagging	#sarcasm
TwMohammad2015			
532	1,397	Crowd-sourcing	-
TwRiloff2013			
474	1,689	Self-tagging Crowd-sourcing	#sarcasm #sarcastic

[7]Note that for some corpora only the IDs of the tweet coupled with the annotation were available. Thus, we had to retrieve again the text of the tweet by Twitter API at experiment time, but some data were not available anymore (deleted tweets or canceled accounts).

4.4 Our Approach: The emotIDM model

We addressed the task of irony detection as a classification problem, applying supervised machine learning to the set of corpora described in the previous section. To represent each tweet we use different group of features: some of them (structural features, henceforth) are designed to detect common patterns in the structure of the ironic tweets, e.g. type of punctuation, length, emoticons; others are designed to detect affective information (affective features, henceforth).

In this section we will recall the main characteristics of the irony detection model to identify ironic tweets [74], which is our starting point (IDM henceforth). Then, we will present emotIDM, which enriches IDM with additional features, with a special focus on features which exploits information about affect.

It is important to highlight that in this work irony and sarcasm are considered as synonyms, a common assumption in computational linguistic approaches to irony detection [41, 48, 126, 99, 120]. Moreover, the approach proposed here does not rely on bag-of-words (BOW). We consider that irony detection should be addressed by models based mainly on features that allow to capture irony disregarding domain or topic, since our aim is to develop a model able to identify irony in social media texts capturing inherent characteristics of this kind of device. Some authors share a similar perspective on this issue [18, 29, 144].

4.4.1 Irony Detection Model (IDM)

Let us describe the set of features used in IDM [74].

Structural Features

Structural features are the following:

Punctuation Marks. Punctuation marks have been widely applied in irony detection [35, 41, 126]. Some lexical marks help the writer to point out the sense and meaning in a text. According to [86] the use of some textual factors (e.g. punctuation marks) may provide reliable clues for identifying ironic intent in social media content. In short texts like tweets this kind of visual cues can help to achieve the real intention behind the literal content in the utterance. In IDM, the punctuation marks and uppercase words are considered as lexical markers to distinguish ironic from non-ironic utterances.

Length of Words. Twitter users must communicate their messages in 140 characters and express their ideas in a concise and direct manner. We consider a feature to catch the length in words (lengthWords) of each tweet, under the assumption that, thanks to a creative use of language, ironic tweets may achieve a communicative goal probably with less words than non-ironic tweets.

Emoticons. In social media, emoticons ("emotional icon") are used to display a feeling in as few characters as possible. It can be used as visual cues to show the real intention of the speaker in order to achieve a particular effect: humor, sadness, despair, confusion, to apologize, or to express solidarity/support. Sometimes the emoticons are required under certain circumstances in text-based communication, where the absence of some sort of cues can hide what was originally intended (to be humorous, sarcastic, ironic, and often negative) [153]. In IDM the frequency of emoticons is considered as a feature.

Discourse Markers. People use different discourse markers for writing. They have certain functions and help to express ideas. In IDM, there are two different kinds of discursive terms[8]: Counter-Factuality and Temporal Compression. A list of terms that hint an opposition or contradiction in a text (such as 'nevertheless') was considered to calculate a Counter-Factuality score. Furthermore, the frequency of terms that identify elements related to opposition in time (i.e. terms that indicate an abrupt change in a narrative, like 'suddenly') refers to the Temporal Compression score.

Part-of-speech. To capture the structure used in a tweet, we consider the frequency of different part-of-speech (POS) labels. According to [86] adjectives and adverbs can be also considered as lexical markers in ironic expressions. In IDM four POS tags were taken into account: verbs, nouns, adjectives and adverbs. These sets of labels allow us to identify the presence of certain kinds of words in ironic utterances.

Semantic Similarity. Ironic texts are often expressed by using words with a different meaning. According to [60] at the initial stage irony comprehension involves to get the literal sense of the words and then involves incompatible meanings. In order to obtain the degree of inconsistency in a tweet the [154] semantic similarity measure was calculated using the WordNet::Similarity module. [9]

Affective Features

The use of some few features related to affect was already investigated by [74] in IDM:

Dictionary of Affect in Language (DAL). Such resource (see Table 4.6) was exploited in a first attempt to capture some kind of affective information related to a tweet. Three different values were calculated: Activation (degree of response that humans have under an emotional state); Imagery (how difficult is to form a mental picture of a given word); Pleasantness (degree of pleasure produced by words).

Sentiment Lexicons: Hu&Liu and *AFINN.* Giving negative (or positive) evaluations towards some targets is inherent to ironic utterances [5]. In this sense the

[8]These discursive terms have been used previously by Reyes et al. [126]. Both lists are available at: `http://users.dsic.upv.es/grupos/nle`

[9]This module allows to calculate a set of seven different similarity measures. According to the experiments carried out in [74] this semantic similarity performed better than the others.

sentiment score of a tweet may help to distinguish between different types of tweets [147], i.e. ironic and non-ironic. In order to catch the writer's attitude two features were considered: (i) the score, which refers to the overall sentiment (positive, negative or neutral) expressed in a tweet, taking into account a well known sentiment analysis resource developed by Hu&Liu; and (ii) the valence, which is used to compute the rate of evaluation expressed, i.e., a criticism (negative) or a praise (positive), by using the AFINN lexicon.[10] Both features related to the sentiment score and to the polarity value were strongly relevant to irony classification, according to an information gain analysis reported in [74]. This encouraged us to better investigate the use of features related to affect.

In [74] some experiments were carried out with the corpus developed by [126], obtaining encouraging results. As experimental setting five different classifiers were applied (Naïve Bayes, Decision Tree, Support Vector Machine, Multilayer Perceptron and Maximum Entropy) under a ten fold-cross validation. The results outperformed those from [126]. In Section 4.5, we will extend the evaluation for this model, by presenting the results obtained applying the IDM model over all the other corpora mentioned in Section 4.3, for comparison purposes with the results obtained by using the extended model emotIDM.

4.4.2 emotIDM: Irony Detection Model + Emotional Information

In this section we introduce emotIDM, which extends IDM considering a much wider set of features exploiting information related to emotions for irony detection. In particular, as a novelty with respect to other approaches, we sought that could be useful to incorporate in emotIDM information about the psychological and emotional content of tweets by means of (i) a variety of sentiment and emotion lexicons that can offer information about sentiment and emotions expressed in text according to different levels of granularity (e.g. referring simply to positive or negative sentiment, or to emotional categories such as joy, sadness, fear, and so on); (ii) a variety of psycholinguistic resources that could give some additional measure about the emotional disclosure in our sample, according to different theoretical perspectives on emotions. We organize the description of affect-related features to catch such different aspects in three groups: the first group is related to information about *sentiment polarity*, the second group is related to information about *emotions* by referring to a finer-grained categorization model (beyond the polarity valence), and the third one to different perspectives related to emotions according to dimensional approaches to emotion modeling. Affect related features rely on the use of various lexical resources. This is needed with the purpose to increase the coverage of different affective aspects in textual content. Moreover, also new structural features were considered. Next we describe in detail each group of features as well as the resources involved.

TABLE 4.2. Structural features in emotIDM

Features	Description
colon exclamation question PM	The frequency of each punctuation mark in a tweet.
lengthWords lengthChars	The total amount of words and characters in a tweet, respectively.
verbs nouns adjectives adverbs	The frequency of each pos-tag in the tweet.
upperCaseChars	The total amount of uppercase characters in a tweet.
totalEmoticons	The total number of emoticons in a tweet.
val_counter val_temporal	Frequency of Counter-factuality and Temporal compression terms defined in Section 4.4.1.
semantic_similarity	The degree of inconsistency in a tweet (Wu&Palmer semantic similarity measure).
hashtagsFreq mentionsFreq rt	The frequency of each specific Twitter marker in a tweet.

Structural Features

This group includes structural features in the IDM model and, in addition, 8 new features: the length in characters (lengthChars), colon, exclamation, question, the amount of uppercase characters (upperCaseChars); and a set of specific markers of Twitter content: hashtagsFreq, mentionsFreq and rt (retweets). The complete group of features is described and summarized in Table 4.2[11]. As we are proposing a model specifically for Twitter, we consider that in ironic tweets these markers could provide important clues.

Affective Features

Sentiment-related Features

As we already mentioned, irony can be used to express an evaluative judgment and sentiment resources can be useful in order to capture the positive or negative polarity of words in a sentence. Three different scores were used to catch the sentiment expressed in tweets: positive, negative and a total value (that considers both positive and negative values). The sentiment resources we exploited can be split in two categories: those composed by simple lists of positive and negative words, and those where each word is labelled with a sentiment strength in a range of polarity values (from positive to negative). In the first case, in order to obtain the positive and negative score for each tweet we sum the number of words belonging to each category (positive or negative expressions). For resources assigning a numerical score varying in a range of intensity for the polarity valence, the positive/negative score is the sum of all the positive/negative values in a tweet. In both cases, the total value is defined as the difference between the positive and negative score. In total 24 sentiment features were obtained from nine different resources. Table 4.3 summarizes the features and the resources exploited to calculate their values[12].

Table 4.3. Sentiment features in emotIDM

Features	Description
AFINN_total AFINN_pos AFINN_neg	AFINN[13] is a resource collected by Finn Arup Nielsen [107]. The most recent available version of the dictionary contains 2,477 English words. Each one has been manually labelled with a sentiment strength in a range of polarity from -5 up to $+5$. The list includes a number of words frequently used on the Internet, like obscene words and Internet slang acronyms such as LOL (laughing out loud).

[10] See Table 4.3 for a description of the sentiment lexicons mentioned.

[11] PM is defined as the sum of colon, exclamation and question.

[12] Normalization was carried out in order to adjust the values of all resources in a range between 0 and 1.

[13] https://github.com/abromberg/sentiment_analysis/blob/master/AFINN/AFINN-111.txt

HL_total HL_pos HL_neg	The Hu&Liu's lexicon (HL) is a well-known resource originally developed for opinion mining [76]. The final version of the dictionary includes 6,789 words divided in two groups: 4,783 negative (HL_neg) and 2,006 positive (HL_pos)[14].
GI_total GI_pos GI_neg	The Harvard General Inquirer (GI)[15] developed by [136] is a resource for content analysis that attaches syntactic, semantic and pragmatic information to 11,788 part-of-speech tagged words. A total of 182 categories are included in the GI. Two of them, i.e., positive words (1,915) and negative words (2,291), are exploited in our model (GI_pos and GI_neg, respectively).
SWN_total SWN_pos SWN_neg	SentiWordNet[16] (SWN) is a lexical resource based on WordNet developed by [12]. It assigns to each of the about 117,000 synsets of WordNet three sentiment numerical scores (in a range between 0 up to 1): positivity, negativity, and objectivity.
EWN_total EWN_pos EWN_neg	EffectWordNet[17], developed by [37], is a lexicon created on the basis of WordNet. The main idea is that the expressions of sentiment are often related to states and events which have positive or negative (or null) effects on entities. It contains more than 11,000 events distributed in three groups: positive (3,288), negative (2,427) and null (5,296).
SO	Taboada and Grieve [138][18] annotated a list of adjectives with Semantic Orientation (SO) values. The resource is made of 1,720 adjectives and their "near bad" and "near good" values according to the Pointwise Mutual Information - Information Retrieval measure (PMI-IR).
SUBJ_str_pos SUBJ_weak_pos SUBJ_str_neg SUBJ_weak_neg	The Subjectivity lexicon (SUBJ) includes 8,222 terms (labeled as subjective expressions) collected by [152]. It contains a list of words, along with their POS-tagging, labeled with polarity (positive, negative, neutral) and intensity (strongly or weakly subjective). This resource is part of the Multi-Perspective Question-Answering (MPQA) lexicon[19].

[14] http://www.cs.uic.edu/~liub/FBS

[15] http://www.wjh.harvard.edu/~inquirer/homecat.htm

[16] http://sentiwordnet.isti.cnr.it/download.php

[17] http://mpqa.cs.pitt.edu/

[18] We considered the "near good" as positive and "near bad" as negative to calculate the SO value. http://www.sfu.ca/~mtaboada/research/nserc-project.html

[19] http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/

EmoLex_positive EmoLex_negative	EmoLex[20] is a word-emotion association lexicon developed by Mohammad and Turney [103], which include also manual annotations about the polarity value of words, negative or positive. The dictionary contains 14,182 words.
SN_Pol SN_Formula	SenticNet[21] (SN) SenticNet is a recent semantic resource for concept-level sentiment analysis [33]. The current version (SenticNet 3) contains 30,000 words. A value of polarity is provided directly by the resource for each word (SN_Pol). Each concept is associated with the four dimensions of the Cambria's hourglass of emotions model [32], and a polarity measure can be defined in terms of the four affective dimensions, according to the formula in [32]. We will also consider such measure in our study (SN_Formula).

Emotional Categories

Theories in the nature of emotion suggested the existence of basic or fundamental emotions such as anger, fear, joy, sadness and disgust. Different approaches propose different sets of basic or fundamental emotions, each having its own specific eliciting conditions and its own specific physiological, expressive, and behavioral reaction patterns. The emotional categories included in emotIDM are based on 4 resources: EmoLex, EmoSenticNet, SentiSense and LIWC (see Table 4.4). Different resources related to various theories were considered with the purpose to increase the coverage of emotions in textual content. Indeed, the resources we used refer to different emotion models well-grounded in psychology, such as the ones proposed by Robert Plutchik [114], Paul Ekman [45], Magda Arnold [7] and Gerrod Parrot [111]. In particular, emotional labels of EmoLex refer to the eight basic emotions of the Plutchik circumplex model, the ones of EmoSenticNet to the six emotions from the Ekman's model, whereas SentiSense proposes a wider set of emotional labels inspired by different models, including Arnold and Parrot. We compute the frequency of words in a tweet belonging to an emotional category according to information encoded in the various resources. In total 10 different emotions were considered as features (see Table 4.5). Moreover, we also consider in this group of features the coarser-grained classification of emotional words w.r.t. positive and negative emotions provided by LIWC. Table 4.4 summarizes the resources included in emotIDM.

Dimensional Models of Emotions

There are some theories proposing that the nature of an emotional state is determined by its position in a space of independent dimensions. According to a dimensional approach, emotions can be defined as a coincidence of values on a number of different

[20] http://www.saifmohammad.com/WebPages/lexicons.html
[21] http://sentic.net/

Table 4.4. Emotional categories features in emotIDM

Features	Description
EMOLEX_emotion[a]	EmoLex[b] is a word-emotion association lexicon [103] containing 14,182 words labelled according to the eight Plutchik's primary emotions [114]: joy, sadness, anger, fear, trust, surprise, disgust and anticipation.
EmoSN_emotion[a]	EmoSenticNet[c] (EmoSN) is a lexical resource [118] that assigns WordNet-Affect emotion labels related to the Six Ekman's basic emotions to SenticNet concepts. The whole list includes 13,189 entries annotated with the six Ekman's emotions: disgust, sadness, anger, joy, fear and surprise.
SentiSense_emotion[a]	SentiSense[d], developed by [34] attaches emotional meanings to concepts from the WordNet lexical database. It is composed by a list of 5,496 words words tagged with emotional labels from a set of 14 emotional categories, which refer to a merge of models by Arnold, Plutchik and Parrot. In emotIDM we considered a subset[e] composed by: joy, fear, surprise, anger, disgust, love, anticipation, sadness and like.
LIWC_total LIWC_pos LIWC_neg	Linguistic Inquiry and Word Counts[f] dictionary [113] (LIWC) contains 4,500 words distributed in categories for analyze psycho-linguistic features in texts. One of the categories is related to positive and negative emotions.

[a] "Emotion" is parametric to the various instances of emotion, i.e., anger, joy, ... etc.

[b] `http://www.saifmohammad.com/WebPages/lexicons.html`

[c] `http://www.gelbukh.com/emosenticnet/`

[d] `http://nlp.uned.es/~jcalbornoz/SentiSense.html`

[e] Due to the very limited size of word lists related to emotions, some of them were removed from SentiSense features.

[f] `http://www.liwc.net`

Table 4.5. Emotions in emotIDM

Emotion	Resource
Anger	EmoLex, EmoSenticNet and SentiSense
Anticipation	EmoLex and SentiSense
Disgust	EmoLex, EmoSenticNet and SentiSense
Fear	EmoLex, EmoSenticNet and SentiSense
Joy	EmoLex, EmoSenticNet and SentiSense
Sadness	EmoLex, EmoSenticNet and SentiSense
Suprise	EmoLex, EmoSenticNet and SentiSense
Trust	EmoLex
Like	SentiSense
Love	SentiSense

strategic dimensions [27]. Dimensional views of emotions have been advocated by a large number of theorists. emotIDM considers the Pleasantness-Activation-Imagery dimensions of the Dictionary of Affect in Language (DAL), already exploited in IDM. Moreover, it considers dimensions from the ANEW resource, which refers to the the VAD model (Valence-Arousal-Dominance), and from SenticNet, which relies on the Hourglass of Emotions model [32] and reinterprets the Plutchik's model by organizing primary emotions around four independent but concomitant dimensions (Pleasantness-Attention-Sensitivity-Aptitude). In Table 4.6 the resources related to dimensional models used in emotIDM are summarized.

Table 4.6. Emotional dimensions features in emotIDM

Features	Description
ANEW_val ANEW_aro ANEW_dom	Affective Norms for English Words[a] (ANEW) is a set of words associated with emotional ratings [27]. In ANEW each concept in the dictionary is rated in terms of the Valence-Arousal-Dominance (VAD) model.
DAL_ple DAL_act DAL_ima	The Dictionary of Affect in Language[a] (DAL) developed by Whissell [150] contains 8,742 English words rated in a three-point scale along three dimensions: Pleasantness, Activation and Imagery.
SN_Pleas SN_Atten SN_Sensit SN_Apti	SenticNet[a,b] (SN) is a semantic resource where each concept is associated with the four dimensions of the Cambria's hourglass of emotions model [32]: Pleasantness, Attention, Sensitivity and Aptitude.

[a]Normalization was carried out in order to adjust the values of all resources in a range between 0 and 1.

[b]`http://sentic.net/`

In emotIDM ten features related to dimensional models of emotions were considered. It is important to mention that ANEW and DAL were constructed by human-manual rating of words while SenticNet by an automatic process that merges different resources. To calculate the degree of each dimension, the sum of the values for each word in a tweet was considered.

4.5 Experiments

We carried out a set of experiments in order to evaluate and compare the effectiveness of both models IDM and emotIDM in automatically distinguishing between ironic and non-ironic tweets over the set of corpora described in Section 4.3. Using the IDM model a tweet is represented as a vector composed by 16 features, while in emotIDM the vector has 78 features. As we mentioned before, in this work irony is considered as an umbrella term that covers sarcasm. Both IDM and emotIDM were designed to identify ironic content in this general sense. However, some authors developing the datasets used in our experiments used the term 'sarcasm' to refer to their irony-laden textual samples [18, 120, 127], depending on the hashtags used for collecting the samples (see Table 4.1, 4th column). Therefore, in order to be consistent with the original terminology, in the following we describe the experiments using the labels 'ironic' or 'sarcastic' depending on the term used by the authors during the corpora development. But let us remark that we will use the same model to identify both the phenomena in tweets.

Different experimental setting were evaluated:

1. *TwReyes2013*. Three binary classifications: irony-vs-education, irony-vs-humor and irony-vs-politics. Each combination is balanced with 10,000 ironic and 10,000 non-ironic samples (balanced distribution).

2. *TwIronyBarbieri2014*. Four binary classifications: irony-vs-education, irony-vs-humor, irony-vs-newspaper and irony-vs-politics Each combination is balanced with 10,000 ironic and 10,000 non-ironic samples (balanced distribution). Let us remark again that here the non-ironic samples are the same that are used in the previous item, whereas the ironic samples are the new ones introduced in [18].

3. *TwMohammad2015*. Binary classification: ironic-vs-non-ironic (imbalanced distribution)

4. *TwSarcasmBarbieri2014*. Four binary classifications: sarcasm-vs-education, sarcasm-vs-humor, sarcasm-vs-newspaper and sarcasm-vs-politics Each combination is balanced with 10,000 sarcastic and 10,000 non-sarcastic samples (balanced distribution).

5. *TwRiloff2013*. Binary classification: sarcastic-vs-non-sarcastic (imbalanced distribution).

6. *TwPtáček2014*. Binary classification: sarcastic-vs-non-sarcastic (imbalanced distribution).

TABLE 4.7. Results in F-measure obtained by applying both IDM and emotIDM

Corpus	F-Measure					
	IDM			emotIDM		
	NB	DT	SVM	NB	DT	SVM
TwReyes2013						
Irony-vs-Education	0.70	0.83	0.85	0.74	0.90	0.89
Irony-vs-Humor	0.71	0.81	0.83	0.76	0.90	0.90
Irony-vs-Politics	0.71	0.84	0.86	0.74	0.92	0.91
TwIronyBarbieri2014						
Irony-vs-Education	0.67	0.84	0.85	0.75	0.90	0.89
Irony-vs-Humor	0.74	0.84	0.85	0.77	0.91	0.90
Irony-vs-Politics	0.74	0.85	0.86	0.80	0.92	0.91
Irony-vs-Newspaper	0.76	0.85	0.87	0.82	0.91	0.93
TwMohammad2015						
	0.65	0.64	0.62	0.66	0.64	0.60
TwSarcasmBarbieri2014						
Sarcasm-vs-Education	0.75	0.84	0.85	0.81	0.90	0.90
Sarcasm-vs-Humor	0.74	0.83	0.85	0.80	0.92	0.90
Sarcasm-vs-Politics	0.78	0.86	0.88	0.86	0.94	0.93
Sarcasm-vs-Newspaper	0.8	0.88	0.90	0.88	0.96	0.96
TwRiloff2013						
	0.73	0.75	0.71	0.74	0.75	0.73
TwPtáček2014						
	0.68	0.74	0.75	0.70	0.78	0.82

The underlined values are not statistically significant
(t-test with 95% of confidence value).

Three of six sets of experiments used corpora with an imbalanced distribution, as can be seen by observing Table 4.1. Because of the perishability of Twitter data in some cases we could rely only on a subset of the tweets originally collected.

For what concerns classifiers, irony detection mainly comprises traditional supervised methods. The two most widely applied has been the Support Vector Machine (SVM) and Decision Tree (DT) [65, 126, 127, 18, 120, 29, 74]. We evaluated our models by applying Weka[22] implementations of three standard classifiers: Naïve Bayes (NB), Decision Tree and Support Vector Machine[23]. We believe that at this stage the most important issue to address for irony detection as a classification problem is the feature engineering one, not the one related to the optimization of the performance of the classifier [120, 145, 18], which can be an issue to address in a second stage. All experiments were conducted in a 10-fold cross-validation setting. Results obtained are shown in Table 4.7.

[22] `http://www.cs.waikato.ac.nz/ml/index.html`

[23] We used default values of Weka as parameters for each classifier.

4.5.1 Discussion

As a preliminary remark, let us notice that in case of the TwIronyBarbieri2014 and TwMohammad2015 corpora it is not possible to compare our results with results achieved in related work. In fact, this is the first time TwMohammad2015 is used in the context of the irony detection task, whereas the set of ironic samples in TwIronyBarbieri2014 (collected relying on the #irony hashtag) was not used by [18] for evaluating their irony detection model, but it has been created and exploited only in a pilot attempt to distinguish sarcasm from irony, which is a different task. IDM improves the state-of-the-art over the TwReyes2013 corpus, as already highlighted in [74][24]. For what concerns the other corpora which were already used for the evaluation of irony detection models, by observing Table4.8 we can see that IDM outperforms the state-of-the-art in TwRiloff2013, whereas results regarding TwSarcasmBarbieri2014 are not higher than those reported in [18]. It is interesting to note that in general results obtained over the "Self-tagged" corpora (TwReyes2013, TwIronyBarbieri2014, TwSarcasmBarbieri2014 and TwPtáček2014) are higher than those from "Crowd-sourced" ones (TwMohammad2015 and TwRiloff2013). This can be an aspect to be further investigated, reflecting on the differences that exist in corpora construction. In terms of performance over "Crowd-sourced" corpora, there is much less difference between IDM and emotIDM than in "Self-tagged" corpora.

Overall, emotIDM outperforms IDM. The results show that emotional information helps to achieve higher F-measure rates in order to distinguish irony-laden tweets. emotIDM seems to be able to capture relevant features from this kind of tweets. This may confirm our hypothesis about the important role of emotional information for irony detection. Both IDM and emotIDM show a consistent performance even working with different size corpora. The higher results are achieved in balanced distribution (TwReyes2013, TwIronyBarbieri2014, TwSarcasmBarbieri2014). The NB classifier presents the worst performance as in other approaches to irony detection [126, 29, 47]. The SVM classifier obtains slightly better results than DT using IDM, while for emotIDM the DT achieves the best performance.

We compare the performance in terms of F-measure of IDM and emotIDM against the reported results for each corpus (see Table 4.8). For what concerns the state-of-the-art, together with the F-measure we mention the classifier used, and we report our results, both for IDM and emotIDM, by using the same classifier.

Overall, emotIDM outperforms the state-of-the-art (values in bold). All experiments except two were improved. Let us comment on such cases. For what concerns the results achieved on Sarcasm-vs-Newspaper, it is the only outcome where our approach does not improve the state-of-the-art on the TwSarcasmBarbieri2014. However, notice that our set of features does not consider the presence of URL, unlike the proposal in [18], where the authors themselves report that nine of ten tweets in the Newspaper category contain an URL.

[24] As a main difference with the partial results reported in [74], we use of a normalized version of two resources: AFINN and DAL.

TABLE 4.8. Comparison of results with the state-of-the-art

Corpus	State-of-the-art Reference	F-measure	Our Results IDM	Our Results emotIDM
TwReyes2013				
			DT	
Irony-vs-Education	Reyes et al. [126]	0.70	0.83	**0.90**
	Barbieri et al. [18]	0.73		
	Hernández Farías et al. [74]	0.78		
			DT	
Irony-vs-Humor	Reyes et al. [126]	0.76	0.81	**0.90**
	Barbieri et al. [18]	0.75		
	Hernández Farías et al. [74]	0.79		
			DT	
Irony-vs-Politics	Reyes et al. [126]	0.73	0.84	**0.92**
	Barbieri et al. [18]	0.75		
	Hernández Farías et al. [74]	0.79		
TwSarcasmBarbieri2014				
			DT	
Sarcasm-vs-Education	Barbieri et al. [18]	0.88	0.84	**0.90**
Sarcasm-vs-Humor		0.88	0.83	**0.92**
Sarcasm-vs-Politics		0.90	0.86	**0.94**
Sarcasm-vs-Newspaper		0.97	0.88	0.96
TwRiloff2013				
			SVM	
	Riloff et al. [127]	0.51	0.71	**0.73**
	Joshi et al. [79]	0.61		
TwPtáček2014[a]				
			SVM	
	Ptáček et al. [120]	0.90	0.75	0.82

[a] We have selected the imbalanced distribution for evaluation.

The comparison with the results of [120] over the TwPtáček2014 corpus deserves further investigations. Ptáček et al. propose a model to identify sarcastic tweets that include as features information referring to a bag-of-words (BOW) representation of text, whereas our system does not. Their result by using only BOW (0.90 in F-measure) is almost the same that using the whole set of features (including bag-of-words). It is difficult to compare the performance of our system with the one proposed by [120] due to two main reasons: (i) TwPtáček2014 contains sufficient data to train a successful bag-of-words classifier, but the same approach could be not adequate to irony detection across different datasets; (ii) no results without bag-of-words were reported, whereas our system shows consistent results even without the presence of bag-of-words features. Furthermore, more importantly, as explained also in [144, 18], the risk for BOW approaches is to be topic-dependent, since they work a topic-based classifier and not as an irony detection procedure. Instead, the advantage of approaches which are not relying on bag-of-words, like the one we propose, is that they are able to capture ironic style disregarding domain, as it is proved by our evaluation across different datasets which cover different topics.

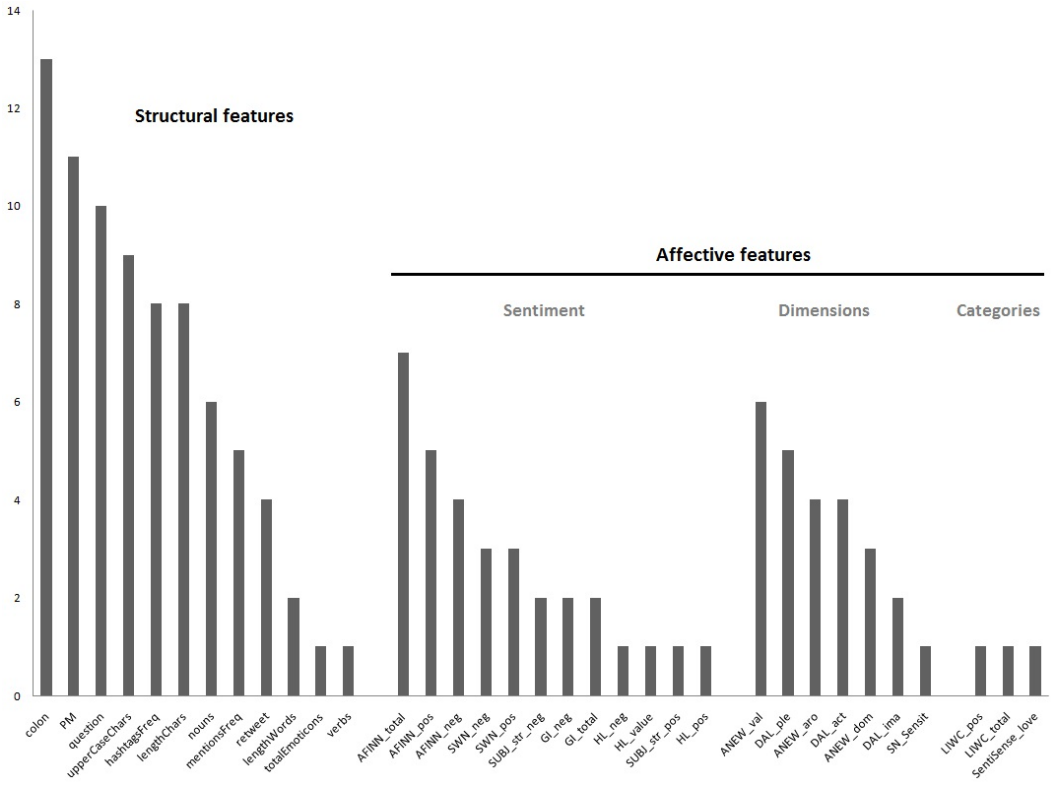
4.5.2 Feature Analysis: Information Gain

We used many features to detect ironic utterances. An Information Gain analysis of features was carried out in order to identify which features are useful in emotIDM. The ten-best ranked features for each binary classification can be seen in the Appendix (Table 4.9). In order to have an overall view, we computed the frequency of each best-ranked feature for all the binary classifications, with the aim to evaluate which features were ranked as the best. A total of 34 features emerged as the most frequent. Figure 4.1 shows the results obtained. For sake of readability, structural features are grouped on the left. The following three groups are related to affective features, and refer to sentiment features, emotional dimensions features and emotional categories features, respectively.

We observe that features derived from the structural group rank high. This validates once again the importance of lexical markers in Twitter ironic contents [86, 35, 41, 126, 18]. Both sentiment features and the ones related to emotional dimensions captured by ANEW, DAL and SenticNet appears to be useful to identify ironic tweets. In particular, AFINN emerges as an efficient sentiment resource for irony detection, but also SentiWordNet, General Inquirer, Hu&Liu and SUBJ play a role. All the dimensions in ANEW and DAL have a relevant discriminative power, whereas, for what concerns SenticNet, the 'Sensitivity' dimension seems to be the most useful. Nevertheless, features related to emotional categories also help in the classification performance, even if they are not among the best ranked features. In this group, we can see coarser-grained features related to LIWC, but also the feature related to words expressing the emotion 'Love' (SentiSense).

Additionally, in order to investigate if some differences could emerge by keeping separate corpora where users were marking the intention to be 'ironic' and the ones where they marked the intention to be 'sarcastic' (see Table 4.1), we calculated the same frequency on the best ranked features according to Information Gain considering

FIGURE 4.1. Best ranked features according to Information Gain

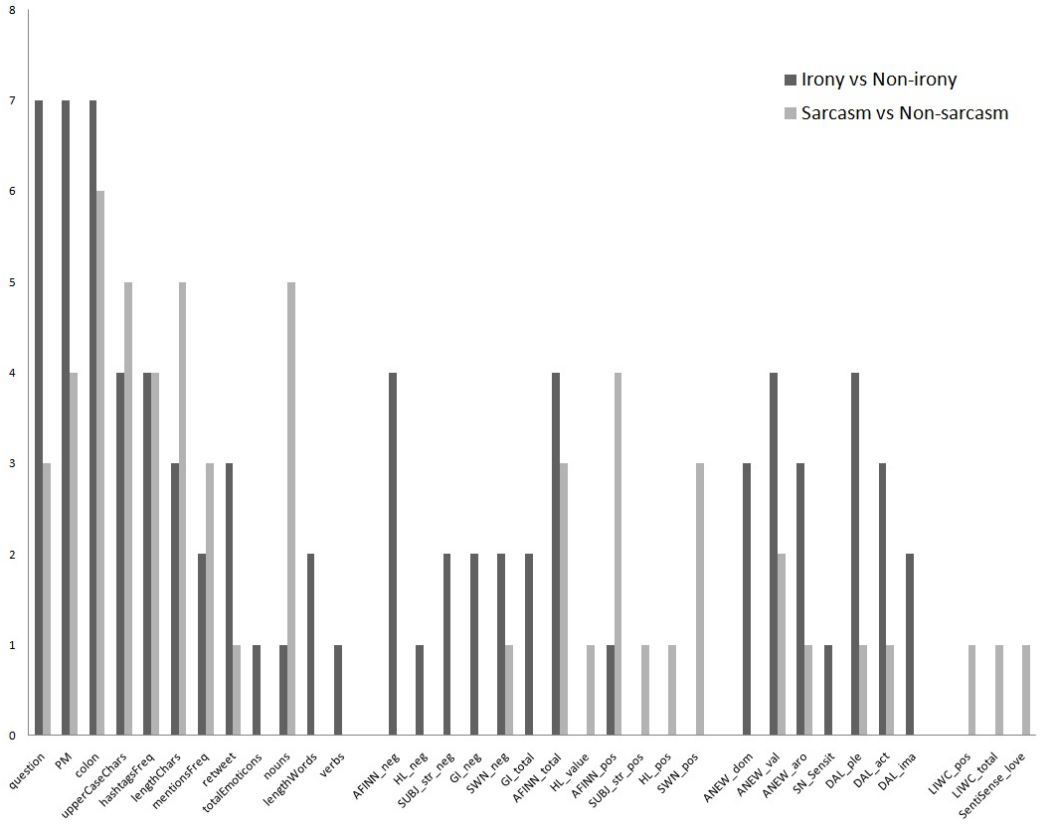


on the one hand ironic-vs-non-ironic tweets, on the other hand sarcastic-vs-non-sarcastic tweets. The outcome, shown in Figure 4.2, is interesting and introduces new data-driven arguments for a possible separation between irony and sarcasm[25].

Information from dimensional models of emotions (in particular from DAL and ANEW) is very important to distinguish tweets belonging to the ironic class. In both tasks features related to sentiment are in the top 10. Some authors consider that one of the main differences between irony and sarcasm is based on the evaluation they express [5]. Irony may be positive (i.e. non-critical) while sarcasm is not [58]. Sarcasm is considered more aggressive and offensive than irony. According to Wang [147], the tweet with more aggressive intention should be sugar coated with more positive words. Such hypothesis seems to be well supported here. Indeed, it can be clearly noticed as the discriminative powers of the sentiment features related to the positive and negative polarity values of words vary in the two cases (positive words are more relevant for identifying sarcasm and vice versa). These could be indicators about the fact that such features could help in differentiate sarcasm from irony. Moreover, it is worth to be noticed that features related to emotional categories seem to be more discriminative in corpora self-tagged with #sarcasm and #sarcastic. In particular, a preliminary analysis for what concerns the feature related to words expressing 'Love', suggests that it could be related to the higher frequencies of

[25]Features are grouped as in the previous figure.

FIGURE 4.2. Best ranked features according to Information Gain, differentiating between tweets tagged as ironic and sarcastic



constructions such as 'I just love...', 'I love when...', 'I love being...' in tweets tagged with #sarcasm. This will be a further data-driven element to investigate to address the finer-grained task of distinguishing different types of irony.

For what concerns the structural features, interestingly, the feature related to frequency of nouns seems to be particularly relevant in tweets containing the #sarcasm hashtag. Besides, the mentionsFreq is also relevant for sarcastic tweets; one possible explanation is that this kind of feature can be considered as the way to point out the target by a specific Twitter marker, i.e. the mention. This is in line with [90]: "Sarcasm conveys ridicule of a specific victim whereas irony does not". In this sense, sarcastic utterances may contain a noun or a mention to refer the target. Finally, also the lenghtChars feature seems to be especially relevant in sarcastic tweets. A possible hypothesis is that sarcastic tweets are sharper, and then shorter.

4.6 Conclusion and future work

In this article, we presented emotIDM a novel model for irony detection in Twitter that includes information on affect encompassing different aspects of this multi-faceted phenomenon. We have performed several experiments over a set of corpora already used in the same task, outperforming previous results both for what concerns

IDM, the previous model we used as a starting point, and results obtained on the same datasets by previous authors, in almost all cases.

To the best of our knowledge this is the first work in irony detection where the robustness of the model is evaluated on a set of representative Twitter corpora including samples of ironic and not ironic messages, which were different along various dimensions: size, balanced vs imbalance distribution, collection methodology and criteria (i.e. self-tagging vs crowd-sourcing, hashtags used for collecting samples, etc.). Dealing also with imbalanced distributions is, indeed, important, since, as highlighted also in [126, 120], real word do not resemble the balanced distribution. Results show that our model achieves good performances in classification terms across all these dimensions. It performs better in cases of datasets with balanced distribution, where a self-tagging methodology has been applied, but it has to be noticed that it achieves good results, improving the state-of-the-art, also with the TwRiloff2013 dataset, with fewer data and imbalanced distribution. A more detailed reflection on the better performances related to corpora developed by using self-tagging is matter of future work.

Overall, results confirm that affective information helps in distinguishing among ironic and non-ironic tweets. In particular, a first analysis of the affective features via information gain highlights the discriminating power, on the one hand, of sentiment-related features based on resources such as AFINN, SentiWordNet, General Inquirer and Subjectivity Lexicons, and, on the other hand, of features related to resources such as ANEW, DAL and SenticNet, which refer to dimentionals models of emotions. For what concerns features related to emotion words such as joy, anger, and so on, they seems to have a minor role, with the exception of the one related to the emotion 'Love'.

Comparative results on corpora collected by using different self-tagging criteria, i.e. on the one hand hashtags such as #irony, and on the other hand hashtags such as #sarcasm and #sarcastic, introduce new data-driven arguments for a possible separation between irony and sarcasm. The issue of distinguishing between such devices is very challenging, still poorly understood and only rarely addressed from computational linguistics [147, 18], deserving further investigations.

A cross-language study of our model could be an interesting line of future research, even if some of the features could be language dependent. Moreover, it could be interesting to apply this model to other languages apart from English also to see if it would assist the state-of-the-art in going beyond irony detection, leading to an improvement of emotion forecast. Finally, it will be interesting to investigate also the effect of using word embeddings as features (extracted from a selected large corpus, e.g., a large corpus of tweets) in the classification system, in order to evaluate their effectiveness and to test if the features extracted from the lexical resources still play a positive role.

APPENDIX

In Table 4.9 the rank for each binary classification mentioned in Section 4.5 is shown.

TABLE 4.9. Ten best ranked features according to Information Gain

TwReyes2013			**TwIronyBarbieri2014**			
Iro-vs-Edu	*Iro-vs-Hum*	*Iro-vs-Pol*	*Iro-vs-Edu*	*Iro-vs-Hum*	*Iro-vs-Pol*	*Iro-vs-News*
question	PM	PM	PM	PM	PM	colon
PM	question	question	question	colon	colon	PM
colon	colon	colon	colon	question	question	upperCaseChars
AFINN_neg	AFINN_neg	ANEW_dom	hashtagsFreq	hashtagsFreq	hashtagFreq	mentionsFreq
HL_neg	GI_total	upperCaseChars	lenghtChar	rt	upperCaseChars	hashtagFreq
SUBJ_str_neg	GI_neg	DAL_ple	upperCaseChars	AFINN_pos	noun	lenghtChar
GI_neg	ANEW_val	ANEW_val	mentionsFreq	AFINN_total	DAL_ple	lenghtWords
AFINN_total	ANEW_aro	ANEW_aro	rt	ANEW_val	rt	DAL_act
ANEW_dom	AFINN_total	AFINN_neg	SWN_neg	emoticons	ANEW_val	DAL_ple
GI_total	SN_Sensit	SUBJ_str_neg	DAL_act	ANEW_arousal	ANEW_dom	DAL_ima

TwSarcasmBarbieri2014				**TwRiloff2013**	**TwPtáček2014**	**TwMohammad2015**
Sar-vs-Edu	*Sar-vs-Hum*	*Sar-vs-Pol*	*Sar-vs-News*			
colon	colon	colon	colon	HL_pos	colon	DAL_ple
PM	PM	PM	upperCaseChars	AFINN_pos	lenghtChar	DAL_act
question	question	question	PM	mentions_Freq	DAL_ple	verbs
hashtagsFreq	hashtagsFreq	lenghtChar	lenghtChar	LIWC_pos	DAL_act	lenghtWords
upperCaseChars	upperCaseChars	upperCaseChars	mentionsFreq	colon	PM	question
lenghtChar	rt	hashtagsFreq	hashtagsFreq	LIWC_total	DAL_ima	DAL_ima
nouns	lenghtChar	nouns	AFINN_pos	HL_value	SWN_pos	AFINN_neg
AFINN_total	ANEW_val	AFINN_pos	SWN_pos	SUBJ_str_pos	AFINN_total	SWN_neg
AFINN_pos	nouns	ANEW_val	AFINN_total	upperCaseChars	nouns	SWN_neg
mentionsFreq	ANEW_aro	SWN_pos	nouns	SentiSense_love	SWN_neg	DAL_ima

Chapter 5

Figurative Messages and Affect in Twitter: Differences between #irony, #sarcasm and #not

In this chapter we describe an analysis carried out in a set of tweets rich in figurative language. We focused on the relationship between figurative language and affective phenomena in tweets labeled with different hasthags. We identify a set of promising features for discriminating between ironic and sarcastic tweets. The obtained results outperform the state of the art. Furthermore, we analyze the role of polarity reversal in tweets labeled with ironic hashtags.

Published in:

Sulis, E., **Hernández Farías, D.I.**, Rosso, P., Patti, V., and Ruffo, G. (2016). Figurative messages and affect in Twitter: Differences between #irony, #sarcasm and #not. *Knowledge-Based Systems* 108:132-143.
DOI: `http://dx.doi.org/10.1016/j.knosys.2016.05.035`

Abstract

The use of irony and sarcasm has been proven to be a pervasive phenomenon in social media posing a challenge to sentiment analysis systems. Such devices, in fact, can influence and twist the polarity of an utterance in different ways. A new dataset of over 10,000 tweets including a high variety of figurative language types, manually annotated with sentiment scores, has been released in the context of the task 11 of SemEval-2015. In this paper, we propose an analysis of the tweets in the dataset to investigate the open research issue of how separated figurative linguistic phenomena irony and sarcasm are, with a special focus on the role of features related to the multi-faceted affective information expressed in such texts. We considered for our analysis tweets tagged with #irony and #sarcasm, and also the tag #not, which has not been studied in depth before. A distribution and correlation analysis over a set of features, including a wide variety of psycholinguistic and emotional features, suggests arguments for the separation between irony and sarcasm. The outcome is a novel set of sentiment, structural and psycholinguistic features evaluated in binary classification experiments. We report about classification experiments carried out on a previously used corpus for #irony vs #sarcasm. We outperform in terms of F-measure the state-of-the-art results on this dataset. Overall, our results confirm the difficulty of the task, but introduce new data-driven arguments for the separation between #irony and #sarcasm. Interestingly, #not emerges as a distinct phenomenon.

Keywords: Figurative Language, Affective Knowledge, Irony, Sarcasm, Twitter

5.1 Introduction

The use of figurative devices such as irony and sarcasm has been proven to be a pervasive phenomenon on social media platforms such as Twitter and poses a significant challenge to sentiment analysis systems, since irony-laden expressions can play the role of polarity reversers [51]. Irony and sarcasm can influence and twist the affect of an utterance in complex and different ways. They can elicit various affective reactions, and can behave differently with respect to the polarity reversal phenomenon, as shown in [25]. However, the issue of distinguishing between such devices is still poorly understood. In particular, the question of whether irony and sarcasm are separated or similar linguistic phenomena is a controversial issue in literature and no clear consensus has already been reached. Although some researchers consider them strongly related figurative devices, other authors proposed

a separation: sarcasm is offensive, more aggressive than irony [90, 10] and delivered with a cutting tone (rarely ambiguous), whereas irony often exhibits great subtlety and has been considered more similar to mocking in a sharp and non-offensive manner [5]. Furthermore, there is a consistent body of work on computational models for sarcasm detection [127] and irony detection [126] in social media, but only preliminary studies addressed the task to distinguish sarcasm and irony[147, 18].

In this paper we contribute to the debate of whether irony and sarcasm are similar or distinct phenomena by investigating how hashtags marking a figurative intent are used in Twitter. Our experiments concern a rich corpus of figurative messages. We considered tweets marked with the user-generated tags #irony and #sarcasm, as such tags reflect a tacit belief about what constitutes irony and sarcasm, respectively [126]. We extend our analysis also to tweets tagged with hashtag #not, previously used to retrieve sarcastic tweets [127, 92], in order to investigate further their figurative meaning.

Samples of tweets marked with different hashtags follow:

(tw1) *Fun fact of the day: No one knows who invented the fire hydrant because its patent was destroyed in a fire. #irony*

(tw2) *I just love it when I speak to folk and they totally ignore me!!! #Sarcasm!*

(tw3) *So I just colored with Ava for an hour. Yeah my summer so far has been so fun [smiling face emoji] #not*

Our methodology comprehends two steps. First, we performed a distribution and correlation analysis relying on the dataset of SemEval2015-Task11 [51], which includes samples of the kinds of figurative messages under consideration here (step 1). We explored the use of the three hashtags including structural as well as psycholinguistic and affective features concerning emotional information.

The affective information expressed in the dataset is multi-faceted. Both sentiment and emotion lexicons, as well as psycholinguistic resources available for English, refer to various affective models and capture different facets of affect, such as *sentiment polarity*, *emotional categories* and *emotional dimensions*. Some of such resources, i.e., SenticNet [33] and EmoSenticNet [118], are not flat vocabularies of affective words, but include and model semantic, conceptual and affective information associated with multi-word natural language expressions, by enabling concept-level analysis of sentiment and emotions conveyed in texts. In our view, all such resources represent a rich and varied lexical knowledge about affect, under different perspectives, therefore we propose here a comprehensive study of their use in the context of our analysis, in order to test if they convey relevant knowledge to distinguishing different kinds of figurative messages.

The analysis provided valuable insights on three kinds of figurative messages, including different ways to influence and twist the affective content. The outcome is a novel set of features evaluated in binary classification experiments (step 2). To

better understand the impact of each feature, we evaluated our model performing experiments with different subset combinations, proceeding also by feature ablation, i.e. removing one feature at time in order to evaluate its contribution on the results.

To sum up, our experiments address the following research questions:

1. Is it possible to distinguish irony from sarcasm?
2. What is the role of the #not hashtag as a figurative language device? Is it a synonym of irony, of sarcasm, or something in between?
3. Does information about sentiment and psycholinguistics features help in distinguishing among #irony, #sarcasm and #not tweets?
4. What is the role of the polarity reversal in the three kinds of figurative messages?

Overall, results confirm the difficulty of the task, but introduce new data-driven arguments for the separation between #irony and #sarcasm. As shown in the next sections, we outperform the state-of-the-art results in #irony vs #sarcasm classification from 0.62 [18] to 0.70, in terms of F-measure.

As for the separation of #irony vs #not and #sarcasm vs #not, interestingly, #not emerges as a distinct phenomenon. Analysis of the relevance of each feature in the model confirms the significance of sentiment and psycholinguistics features. Finally, an interesting finding about polarity reversal is given by correlation study presented in Section 5.4.2: the polarity reversal phenomenon seems to be relevant in messages marked with #sarcasm and #not, while it is less relevant for messages tagged with #irony.

The paper is structured as follows. Section 5.2 surveys main issues in literature about irony and the like. In Section 5.3 we describe the corpus and the resources exploited in our approach. Section 5.4 presents the feature analysis and Section 5.5 describes our experiments. Section 5.6 concludes the paper.

5.2 Irony, Sarcasm *et Similia*

Many authors embrace an overall view on irony. Broadly speaking, under the umbrella term of irony one can find distinct phenomena such as *situational irony* or *verbal irony* [93, 9, 44]. Situational irony (or "irony of fate") refers to the state of affairs or events which is the reverse of what has been expected, while the term verbal irony is applied to refer to a figure of speech, characterized by the possibility of distinguishing between a literal and an intended/implied meaning. In particular, according to many theoretical accounts in ironic utterances the speaker intends to communicate the opposite of what is literally said [56, 134], but since such definition does not allow to account for many samples of utterances which are considered ironic, we prefer to refer to a more general position, on which different authors

in literature would tacitly agree: "Regardless of the type, or absence, of meaning negation/reversal, the literal import of an ironic utterance differs from the implicit meaning the speaker intends to communicate" [44]. Moreover, we can have an ironic statement, meant as utterance of a speaker which refers to certain aspects of an ironic situation [93].

In linguistics, verbal irony is sometimes used as a synonym of sarcasm [63, 62, 59]. According to the literature, boundaries in meaning between irony, sarcasm *et similia* are fuzzy. While some authors consider irony as an umbrella term covering also sarcasm [28, 87, 56], others provide insights for a separation. Sarcasm has been recognized in [26] with a specific target to attack [10, 44], more offensive [90] and "intimately associated with particular negative affective states" [100]. According to [90] hearers perceive aggressiveness as the feature that distinguishes sarcasm. Instead, irony has been considered more similar to mocking in a sharp and non-offensive manner [5].

The presence of irony-related figurative devices is becoming one of the most interesting aspects to check in social media corpora since it can play the role of polarity reverser with respect to the words used in the text unit [124]. However, a variety of typologies of figurative messages can be recognized in tweets: from irony to sarcastic posts, and to facetious tweets that can be playful, aimed at amusing or at strengthening ties with other users. Ironic and sarcastic devices can express different interpersonal meaning, elicit different affective reactions, and can behave differently with respect to the polarity reversal phenomenon [25]. Therefore, to distinguish between them can be important for improving the performances of systems in sentiment analysis.

For computational linguistics purposes irony and sarcasm are often viewed as the same figurative language device. Computational models for sarcasm detection [41, 65, 127, 18, 99] and irony detection [126, 48, 74] in social media has been proposed, mostly focussed on Twitter. Only a few preliminary studies addressed the task to investigate the differences between irony and sarcasm [147, 18]. The current work aims to further contribute to this subject.

Furthermore, a rarely investigated form of irony that can be interesting to study in social media is self-mockery. Self-mockery seems to be different from other forms of irony, also from sarcasm, because it does not involve contempt for others, but the speaker wishes to dissociate from the content of the utterance. According to some theoretical accounts: "Self-mockery usually involves a speaker making an utterance and then immediately denying or invalidating its consequence, often by saying something like *'No, I was just kidding'*" [55]. Moreover, the analysis of complex forms of self-mockery in spontaneous conversations in [85] highlighted interesting practices related to *narrative self-mockery*, where people, in particular women, jokingly tell a story about a personal experience, only apparently offering themselves as object of laughing. The same study shows that, in the conversational contexts analyzed, makings jokes about their own (sometime negative) experience provided the narrator with a way to share the experience and jointly create a dis-

tance through the mocking: "The narrators are not laughed at and do not invite others to do. [...] They seem to be saying, 'I had such an awful experience', or 'I was so dumb', but it is all done with a narrative strategy which prevents regret, pity or even laughter at their expense." [...] "in the episodes there is no invitation to laugh about the teller, but rather with her" [85]. Investigations on the role of the #not hashtag as a figurative language device could maybe provide insights into this phenomenon by relying on social data, where such data, when connected with information about genre and age, could be also be an interesting new research line for studying the relationship between gender and different forms of irony.

People often use specific markers for communication purposes. Research on the use of different hashtags (particularly #irony, #sarcasm and #not) could be useful in order to investigate if they can be low-salience cues [64], i.e. if Twitter users may use these kinds of markers in order to highlight their non-literal intention. This could be the case especially in short texts (such as tweets), where the lack of context could provoke misunderstanding.

5.3 Dataset and Lexical Resources

In this section we describe the resources used in our work. First, the corpus of tweet messages in English developed for Task 11 of SemEval-2015[1] has been studied extensively [51]. It consists in a set of tweets containing creative language that are rich in metaphor and irony. This is the only available corpus where a high variety of figurative language tweets has been annotated in a fine-grained sentiment polarity from -5 to +5. We finally rely on a dataset of 12,532 tweets[2]. Among the 5,114 different hashtags in the corpus, the most used ones are #not (3,247 tweets), #sarcasm (2,260) and #irony (1,737). Table 5.1 shows some introductory statistics over the dataset. The whole distribution of the polarity has a mean value of -1.73, a standard deviation of 1.59 and a median of -2.02. We consider the median as it is less affected by extreme values, instead of mean values. These results confirm that messages using figurative language mostly express a negative sentiment [124].

Table 5.1. Corpus description: Number of tweets (N), Mean (MP) and Standard Deviation (SD) of the Polarity, Median of the Length (ML)

Description	N	MP	SD	ML
With #irony	1,737	-1.77	1.41	83
With #sarcasm	2,260	-2.33	0.77	66
With #not	3,247	-2.16	1.04	71

To cope with emotions and psycholinguistic information expressed in tweets, we explore different lexical resources developed for English. Finally, these can be

[1] We consider the training, the trial and the test set: `http://alt.qcri.org/semeval2015/task11`

[2] Due to the perishability of the tweets we were not able to collect all the 13,000 messages of the corpus.

grouped into three main categories related to "Sentiment polarity", to "Emotional categories" or to "Dimensional models of emotions".

Sentiment polarity. In order to gather information about sentiment polarity expressed in the corpus, we exploited lexicons including positive and negative values associated to terms.

(i) AFINN: This affective dictionary has been collected by Finn Årup Nielsen starting from most frequent words used in a corpus of tweets [107]. Each one has been manually labelled with a sentiment strength in a range of polarity from -5 up to $+5$. The list includes a number of words frequently used on the Internet, like obscene words and Internet slang acronyms such as LOL (laughing out loud). The most recent available version of the dictionary contains 2,477 English words[3]. A bias towards negative words (1,598, corresponding to 65%) compared to positive ones (878) has been observed.

(ii) HL: The Hu-Liu's lexicon is a well-known resource originally developed for opinion mining [76]. The final version of the dictionary includes an amount of 6,789 words divided in 4,783 negative (HL_neg) and 2,006 positive (HL_pos)[4].

(iii) GI: The Harward General Inquirer is a resource for content analysis of textual data originally developed in the 1960s by Philip Stone [136]. The lexicon attaches syntactic, semantic, and pragmatic information to 11,788 part-of-speech tagged words. It is based on the Harvard IV-4 dictionary and Lasswell dictionary content analysis categories. Words are labelled with a total of 182 dictionary categories and subcategories[5]. The positive words (GI_pos) are 1,915, while the negative ones are 2,291 (GI_neg).

(iv) SWN: SentiWordNet [12] is a lexical resource based on WordNet 3.0. Each entry is described by the corresponding part-of-speech tag and associated to three numerical scores which indicate how positive, negative, and "objective" (i.e., neutral) the terms contained in the synset are. Each of the three scores ranges in the interval [0,1] and their sum is 1. Synsets may have different scores for all the three categories: it means the terms have each of the three opinion-related properties to a certain degree. In SentiWordNet 3.0[6] all the entries are classified as belonging to these three sentiment scores including a random-walk step for refining the scores in addition to a semi-supervised learning step. The first two categories (SWN_pos and SWN_neg) will be considered in our analysis.

(v) SN: SenticNet is a recent semantic resource for concept-level sentiment analysis [117, 119, 33]. The current version (SenticNet 3) contains 30,000 words, mainly unambiguous adjectives as stand-alone entries, plus multi-word expressions. The dictionary exploits an energy-based knowledge representation (EBKR) formalism to provide the affective semantics of expressions. Each concept is associated with the four dimensions of the hourglass of emotions [31]: Pleasantness, Attention, Sensitivity and Aptitude. We refer to these four values as SN_dim in our experiments in Section 5.5. A value of polarity is provided directly by the resource (SN_polarity henceforth). Moreover, since polarity is strongly connected to attitude and feelings, a further

[3] `https://github.com/abromberg/sentiment_analysis/blob/master/AFINN/AFINN-111.txt`
[4] `http://www.cs.uic.edu/~liub/FBS/`
[5] `http://www.wjh.harvard.edu/~inquirer/homecat.htm`
[6] `http://sentiwordnet.isti.cnr.it/download.php`

polarity measure is proposed, which can can be defined in terms of the four affective dimensions, according to the formula:

$$p = \sum_{i=1}^{n} \frac{Pl(c_i) + |At(c_i)| - |Sn(c_i)| + Ap(c_i)}{3N}$$

where c_i is an input concept, N is the total number of concepts of the tweet, 3 is a normalization factor. We will also consider such polarity measure in our study. In the following we will use 'SN_formula' to refer to the value p obtained by using the equation above.

(vi) EWN: The EffectWordNet lexicon has been recently developed by Choi [37] as a sense-level lexicon created on the basis of WordNet. The main idea is that the expressions of sentiment are often related to states and events which have positive or negative (or null) effects on entities. This lexicon includes more than 11k events in three groups: positive, negative and null. By exploiting the corresponding synset in WordNet, it is possible to collect a larger list of 3,298 positive, 2,427 negative and 5,296 null events[7].

(vii) SO: Semantic Orientation is a list of adjectives annotated with semantic-orientation values by Taboada and Grieve [138]. The resource is made of 1,720 adjectives and their "near bad" and "near good" values according to the Pointwise Mutual Information - Information Retrieval measure (PMI-IR) as proposed by Turney [140]. In this analysis, the values of Semantic Orientation for each term is obtained by the difference between the corresponding "near good" and "near bad" values.

(viii) SUBJ: The subjectivity lexicon includes 8,222 clues collected by Wilson and colleagues [152] from a number of sources. Some were culled from manually developed resources and others were identified automatically. Each clue can be strongly or weakly subjective, or positive and negative. A clue that is subjective in most contexts is considered strongly subjective, while those that may only have certain subjective usages are considered weakly subjective. This resource is part of the Multi-Perspective Question-Answering lexicons[8].

Emotional categories. In order to gather information about the emotions expressed by referring to a finer-grained categorization (beyond the polarity valence), we considered the following resources which rely on categorical approaches to emotion modeling:

(ix) LIWC: Linguistic Inquiry and Word Counts dictionary[9] contains about 4,500 entries distributed in categories that can further be used to analyse psycholinguistic features in texts. We selected two categories for positive and negative emotions: LIWC_PosEmo, with 405 entries, and LIWC_NegEmo, with 500 entries [113].

(x) EmoLex: The resource EmoLex is a word-emotion association lexicon[10] developed at the National Research Council of Canada by Saif Mohammad [103]. The dictionary contains 14,182 words labelled according to the eight Plutchik's primary

[7] http://mpqa.cs.pitt.edu/

[8] http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/

[9] http://www.liwc.net

[10] http://www.saifmohammad.com/WebPages/lexicons.html

emotions [114]: sadness, joy, disgust, anger, fear, surprise, trust, anticipation.

(xi) EmoSN: EmoSenticNet is a lexical resource developed by Poria and colleagues [116] [118] that assigns WordNet Affect emotion labels to SN concepts. The whole list includes 13,189 entries for the six Ekman's emotions: joy, sadness, anger, fear, surprise and disgust[11].

(xii) SS: SentiSense[12] is a concept-based affective lexicon that has been developed by Carrillo de Albornoz [34]. It attaches emotional meanings to concepts from the WordNet lexical database and consists of 5,496 words and 2,190 synsets labelled with an emotion from a set of 14 emotional categories, which are related by an antonym relationship.

Dimensional models of emotions. To provide some additional measures of the emotional disclosure in the corpus, according to different theoretical perspectives on emotions, we exploited the following resources which refer to dimensional approaches to emotion modelling:

(xiii) ANEW: Affective Norms for English Words is a set of normative emotional rating [27]. Each word in the dictionary is rated from 1 to 9 in terms of the Valence-Arousal-Dominance (VAD) model [131]. The first dimension concerns the valence (or pleasantness) of the emotions invoked by the word, going from unhappy to happy. The second one addresses the degree of arousal evoked by the word, whereas the third one refers to the dominance/power of the word, the extent to which the word denotes something that is weak/submissive or strong/dominant. This work considers the three dimensions separately.

(xiv) DAL: Dictionary of Affective Language developed by Whissell [150] contains 8,742 English words rated in a three-point scale[13]. We employed the following three dimensions: Activation (degree of response that humans have under an emotional state); Imagery (how difficult is to form a mental picture of a given word); Pleasantness (degree of pleasure produced by words).

Finally, we include among the *dimensional models of emotions* also the measures related to the Pleasantness, Attention, Sensitivity and Aptitude dimensions from SenticNet.

5.4 Features: A quantitative Analysis

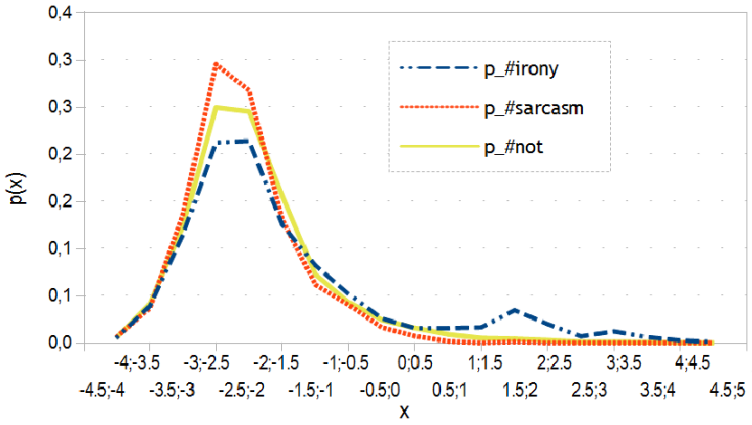
In this section, we identify the main characteristics of the tweets tagged with #irony, #sarcasm and #not from the SemEval 2015-Task 11 corpus. Our main interest is to find differentiating traits among these three kinds of figurative messages.

First, we focus our attention on polarity value which clearly shows a first regularity: the distribution of sarcastic tweets is more positively skewed, as the long "tail" shows, than the ironic ones (Figure 5.1). Moreover, the mean value of tweets marked with #irony is -1.73 instead of -2.33 for the #sarcasm ones.

[11] `http://www.gelbukh.com/emosenticnet/`
[12] `http://nlp.uned.es/~jcalbornoz/SentiSense.html`
[13] `ftp://perceptmx.com/wdalman.pdf`

The differences between the means are statistically significant according to one-way ANOVA (p-value of $3.24e^{-97}$).

FIGURE 5.1. Distribution of tweets by polarity, p(x) is the probability that a tweet has polarity x



These differences show that sarcasm is perceived as more negative than irony by the hashtag adopters in our corpus. On the contrary, ironic messages are more positive as suggested by the above mentioned mean values as well as the little "hill" in the slope. This is a signal that #irony is also used positively (as in positive evaluative irony, i.e. ironic praise), whereas #not and #sarcasm are usually not.

A first hypothesis coming from these results is that Twitter users consider irony as a more nuanced and varied phenomenon in terms of the associated sentiment (see Sec 5.4.2 for further remarks on this issue).

These distributions also signal initially that messages tagged with #not can be considered somehow different from #sarcasm and #irony ones.

In the following, we will perform a distribution analysis in each subgroup for every feature, as well as a correlation study taking into account the fine-grained polarity of the messages. Structural and affective features are considered.

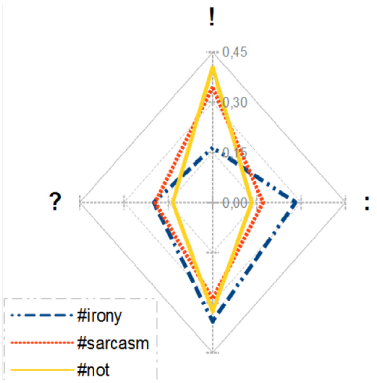
5.4.1 Structural and Tweet Features

Investigating the distributions of most traditional features is our first step. In addition to the analysis of the frequency of the part-of-speech (POS), emoticons, capital letters, URLs, hashtags, re-tweets and mentions, we report here two features showing interesting differences in the three subgroups: tweet length and punctuation marks.

Tweet length. The relation between the length of the tweets and the value of their polarity shows a Pearson's correlation of 0.13, with a statistically significant p-value $p<0.001$. We observe also that shorter messages (5% of tweets with less than

50 characters) are mostly negative with an average value of −2.1 and a standard deviation of 1.2. On the contrary, longer messages (5% of tweets with at least 138 characters) have a mean of −1.6 and a larger standard deviation of 1.7. This suggests that the length could play a role on the polarity of tweets when figurative language is employed. Tweets tagged with #sarcasm are shorter (mean of 66 characters), less than #not (71 char.) and #irony (83 char.). To sum up, it seems that sarcasm expresses in just a few words its negative content (see tweet *tw2* in the Introduction).

FIGURE 5.2. Distribution of punctuation marks in the corpus: colons are most used in #irony tweets, exclamation marks in #sarcasm and #not ones, question marks are less used in #not tweets

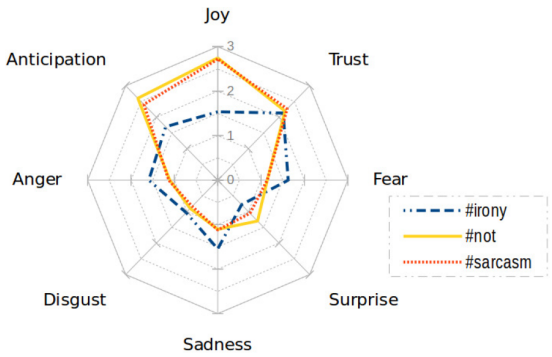


Punctuation marks. Figure 5.2 summarizes the frequency of commas, colons, exclamation and question marks in the three groups of tweets. Given the observed difference in the length of messages, counts are normalized by the length of tweets. While the use of colons is most frequent in #irony tweets and exclamation marks in #sarcasm and #not ones, the frequency of question marks is lower in #not tweets (e.g. tweets *tw1* and *tw2*). This can be linked to the typical grammatical construction of this kind of messages: first a statement, and then the reversal of this statement by the marker #not. Obviously, questions are not easily reversed.

5.4.2 Affective Features

Some important regularities can be detected by analyzing the use of affective words. First, in order to investigate differences in the use of emotions among the three figurative language groups, EmoLex has been used to compute the frequency of words related to emotions, normalized by the number of words. As the distribution in Figure 5.3 shows, tweets marked with #irony contain fewer words related to joy and anticipation than tweets marked with #sarcasm or #not. The same is for surprise, although to a lesser extent. On the other hand, in #irony words related to anger, sadness and fear (and to less extent disgust) are more frequent. Interestingly, tweets tagged with #not and #sarcasm overlap quite perfectly with respect to the use of emotion words, while #irony shows a different behaviour.

FIGURE 5.3. Distribution of emotion words (EmoLex [103]) in the SemEval Task 11 corpus: #not and #sarcasm tweets overlap, while #irony shows a different behaviour.



To further investigate the affective content, we extended the quantitative analysis to all the affective resources mentioned in Section 5.3: ANEW, DAL and the SenticNet's four singular dimensions (dimensional models of emotions); EmoSN, EmoLex, SS and LIWC (emotional categories); AFINN, HL, GI, SWN, EWN, SO, SUBJ and both the SenticNet sentiment polarity values mentioned above.

The values of these resources have been previously normalized in the range from 0 to 1. For each group of tagged messages we compute two kinds of measures, depending on the kind of resource. When the lexicon is a list of terms (i.e., HL, GI, LIWC, EmoLex), we computed the mean value of the occurrences in each group. Instead, for lexicons containing a list of annotated entries (i.e., SN, AFINN, SWN, SO, DAL and ANEW), we calculated the sum of the corresponding values over all the terms, averaged by the total number of words in tweets. Formally, given a group T of n tagged messages where each single tweet t $\in T$ is composed by up to m words, and a lexical resource L assigns to each word w for every tweet in T a corresponding value $L(w)$, we calculated the value $a(T, L)$ according the following equation:

$$a(T, L) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} L(w_{i,j})}{n} \tag{5.1}$$

Results of this analysis are shown from Table 5.2 to Table 5.4, where final values are multiplied by 100 to improve the readability. To investigate the statistical significance on the difference between the mean scores, we performed an ANOVA on our three distributions for each individual resource. Moreover, we computed a Z-test on each pair of distributions [130]. Tables contain in bold, for each lexical resource the highest values which are also statistically significant. In some cases the uncertainty is due to the high variance.

Sentiment polarity features (Table 5.2) seem to be promising. While #sarcasm and #not messages contain more positive words, ironic messages are generally characterized by the use of more words with negative polarity. In fact, we can observe that *all* the lexical resources concerning the polarity of terms we considered (HL, AFINN, GI, SWN, SUBJ, SN and SO) confirm that sarcastic and #not messages contain

TABLE 5.2. Normalized counts for *sentiment polarity* features: values for resources with * are based on scores according to Equation 5.1. For each resource, higher scores are in bold if they are statistically significant.

	Resource	#irony	#sarcasm	#not
Sentiment Polarity	AFINN*	33.63	47.89	47.14
	SN_polarity*	51.28	55.54	56.59
	SN_formula*	26.11	37.31	**41.05**
	SO*	39.53	**45.32**	**45.54**
	GI_pos	1.68	**2.65**	**2.53**
	HL_pos	2.33	**4.97**	4.62
	SWN_pos*	11.52	**15.43**	14.12
	SUBJ_weak_pos	2.18	**2.69**	**2.62**
	SUBJ_strong_pos	2.46	**4.83**	4.44
	GI_neg	**1.26**	1.00	0.91
	HL_neg	**3.15**	2.53	2.31
	SWN_neg*	**11.98**	10.49	10.20
	SUBJ_weak_neg	**1.78**	1.51	1.49
	SUBJ_strong_neg	**1.77**	**1.70**	1.34
	SWN_obj*	87.97	84.64	87.05
	EWN_pos	7.61	8.54	**9.61**
	EWN_neg	4.34	4.20	**4.89**
	EWN_null	8.40	9.21	**10.26**

more positive terms than ironic ones; on the other hand, ironic messages contain more negative terms. Furthermore, also if we consider the polarity of terms related to *events*, detected by EWN, we obtain similar findings for what concerns irony and sarcasm. In fact, as shown in the last rows of Table 5.2, #not messages always contain more terms related to events (both positive, negative and null ones), but positive events are more frequent in sarcastic messages than in ironic ones, whereas negative events are more frequent in ironic than in sarcastic messages. Finally, the objectivity measure from SWN highlights that messages tagged with #irony and #not contain more objective terms than sarcastic messages.

Lexicons related to *dimensional models of emotions* (Table 5.3) also introduce interesting patterns: messages marked with #irony almost always contain a smaller amount of words belonging to these resources. In contrast, #not messages always have a large number of words belonging to these dimensions, i.e. Arousal, Dominance from ANEW or Imagery from DAL. We can also notice a larger frequency of terms related to Imagery in #irony than in #sarcasm, whereas we observe a higher use of words related to Dominance (DAL) in #sarcasm than in #irony. These findings support the idea that irony is more creative than sarcasm (see Sec. 5.4.2 for a deeper discussion on this issue). Results related to the degree of pleasantness produced by words (DAL and SN) and valence of words (ANEW) are higher in sarcastic and #not messages than in ironic ones. This is in tune with the *sentiment polarity* values, confirming what we already noticed before.

TABLE 5.3. Normalized counts for *dimensional models of emotions*: values for resources with * are based on scores according to Equation 5.1. For each resource, higher scores are in bold if they are statistically significant.

	Resource	#irony	#sarcasm	#not
Dimensional Models of Emotions	ANEW_val*	51.24	54.81	**60.03**
	ANEW_arousal*	44.84	45.44	48.63
	ANEW_dominance*	46.14	47.59	**52.07**
	DAL_pleasantness*	61.72	63.46	64.09
	DAL_activation*	56.25	56.55	57.22
	DAL_imagery*	51.81	50.21	**52.12**
	SN_pleasantness*	50.61	**55.54**	**56.70**
	SN_attention*	50.83	**52.10**	**52.24**
	SN_sensitivity*	51.11	49.56	51.19
	SN_aptitude*	52.44	56.82	57.80

Lexicons related to *emotional categories* (Table 5.4) allow to detect further regularities. Terms related to positive emotions (joy, love, like) are nearly always more frequent in #sarcasm and #not messages, whereas negative emotions terms (anger, fear, disgust, sadness) in EmoLex and LIWC are more frequent in #irony ones. This confirms, at a finer granularity level (i.e. the one of *emotional categories*), our findings at the sentiment polarity level, e.g., ironic tweets contain more negative words than the sarcastic ones.

TABLE 5.4. Normalized counts for *emotional categories*. For each resource, higher scores are in bold if they are statistically significant.

	Resource	#irony	#sarcasm	#not
Emotional Categories	EmoLex_anger	**1.59**	1.13	1.10
	EmoLex_anticipation	1.70	2.41	**2.60**
	EmoLex_disgust	**1.03**	0.83	0.90
	EmoLex_fear	**1.62**	1.14	1.14
	EmoLex_surprise	0.78	1.05	**1.30**
	EmoLex_joy	1.54	2.72	**2.75**
	EmoLex_sadness	**1.55**	1.12	1.10
	LIWC_PosEmo	1.71	**3.71**	3.59
	LIWC_NegEmo	1.25	1.13	1.08
	EmoSN_joy	21.63	20.5	**21.99**
	EmoSN_sadness	2.30	2.21	2.21
	EmoSN_surprise	1.61	1.38	1.45
	SS_anticipation	0.84	0.91	**1.06**
	SS_joy	0.40	**0.89**	0.72
	SS_disgust	1.56	1.67	**1.81**
	SS_like	1.73	**2.91**	2.65
	SS_love	0.33	0.89	**0.94**

To sum up, the quantitative analysis carried out above suggests the following considerations concerning the distinction between irony and sarcasm, the role of the #not hashtag and the polarity reversal phenomenon.

Irony is more creative and implicit than sarcasm

Analysis over affective content suggests that irony is more creative than sarcasm, and it is used to convey implicit emotions, whereas sarcasm messages are far more explicit. For what concerns the first aspect, we observed traces of it in the values of *dimensional models of emotion* lexica. In particular, we observe higher values for the dimension Imagery of DAL. Such dimension gives a measure of how difficult is to form a mental picture of a given term. In other words, it provides an estimate for a lexical items efficacy in activating mental images associated with the concept. We think that these results can be interpreted as indicating that irony is more creative than sarcasm. Focusing on sarcasm, we observe not only lower values of Imagery but also higher values of Dominance. Let us recall that the latter dimension from ANEW gives a measure about the fact that the word denotes something that is weak/submissive or strong/dominant. Higher values of Dominance are signals of the fact that words making people feel in control are more frequent in #sarcasm messages than in #irony messages.

For what concerns the second aspect, i.e. the use of the different hashtags #irony and #sarcasm for conveying explicit or implicit figurative messages by Twitter users, when we look at those resources which include information about emotions (see for instance the distribution in Fig. 5.3) we can observe that words related to negative emotions (fear, anger, and sadness) are more frequent in #irony than in #sarcasm, but, more importantly, #sarcasm is usually accompanied by emotions with higher intensity than irony. For instance, the intensity of some emotions such as joy and anticipation in #sarcasm messages is clearly higher. This could be also meant as a signal of the fact that ironic messages are used to convey implicit messages, whereas sarcasm is more explicit.

Finally, focusing on sentiment lexica, we observe that sarcasm tends to involve more positive words than irony. However, as shown by Fig. 5.1, #irony messages are also used positively, when we look at the figurative, intended meaning, whereas #sarcasm messages are usually not. A first hypothesis is that Twitter users consider irony as a more nuanced and varied phenomenon in terms of the associated sentiment. Another interesting hypothesis could be that Twitter users exploit the hashtag #irony for marking situational irony. In fact, in such cases normally speakers humorously lament a situation, without intending to negate the literal meaning of the utterance, in other words without disengaging from what is said. This would be in tune with the lower frequency of negative polarity terms and lower values for intensity of emotions observed in messages marked with #irony. In fact, ironic utterances referring to certain aspects of an ironic situation can also come without evaluative remarks, but only with the observation that something in a situation is ironical.

Is #not a category on its own? Comparison with # irony and #sarcasm

Values related to affect and polarity suggest that tweets tagged with #not could be considered as a category on their own. On the one hand, #not is used quite often with a figurative meaning closer to sarcasm from a perspective of sentiment polarity and finer-grained emotional contents. Tweets marked by #sarcasm and #not are usually accompanied by explicit emotions with higher intensity. Moreover, sentiment polarity values are very similar to sarcasm ones and tend to involve words with positive sentiment and emotions, intending the opposite of what they mean. These results are consistent with findings showing that sarcasm is easier to derive with positive than with negative concepts, and with the idea that people tend to use positive terms to express indirectly that something is negative [39, 50], think for instance to the verbal politeness issue: asserting directly that a particular person has an unfavourable quality is not polite.

However, the #not messages show some peculiarities. By using the tag #not the speaker explicitly manifests the intention of dissociating herself from the literal content of the post, as in certain forms of self-mockery. The impression is that such explicit dissociation introduces an attenuation with respect to the aggressiveness which apparently characterize messages marked with #sarcasm (e.g. tweet *tw3* in the Introduction). Moreover, #not messages differ from #sarcasm messages in that they use negation to invite a sarcastic interpretation of the message. Overall, this seems to be in line with the findings in [63, 62, 59], where the role of negation, as low-salience marker that can affect sarcastic non-literal interpretations is studied, and the role of negation as a "mitigator, retaining in memory the concept within its scope while slightly attenuating it" [63] is highlighted. Referring to this theoretical framework, we can hypothesize to consider the #not hashtag as a negation marker used to achieve a non-literal interpretation of the messages, which characterize, in Twitter negative constructions, expressions of more implicit form of sarcasm or self-mockery. Let us also observe that, although #not is used quite often with a figurative meaning closer to sarcasm, when we look at the information related to resources such as DAL, which include dimensions referring to cognitive processes, such as Imagery, it shows a certain similarity with irony. For instance, the values obtained in terms of Imagery, Valence, and Dominance are higher than in the case of #sarcasm[14].

Polarity reversal

Sentiment polarity values and the use of emotion words related to positive emotions discussed above show that sarcastic and #not messages contain more positive words than the ironic ones. This finding is in line with what was empirically shown also in [147], where the following hypothesis has been tested: "Given the fact that sarcasm

[14]For what concerns higher values of Imagery in words occurring in #not messages than in #sarcasm posts, since such factor is commonly known to affect brain activity, and it is generally accepted, as regards linguistic competence, that visual load facilitates cognitive performance, we can hypothesize that from a cognitive point of view the lexical processing of #not and #sarcasm messages will be different.

TABLE 5.5. Correlation (p-value < 0.001) between scores from lexical resources (RES) and polarity of the annotation in the Corpus (C), forcing the reversal for #irony (revI), #sarcasm (revS), #not (revN), and both #sarcasm and #not (revSN). Darker\lighter shades indicate higher\lower values.

RES	C	revI	revS	revN	revSN
AFINN	0.032	0.018	0.096	0.096	0.160
GI	0.116	0.109	0.168	0.175	0.228
HL	0.128	0.118	0.188	0.172	0.236
SN_pol	0.006	0.001	0.158	0.145	0.268
SN	0.058	0.049	0.179	0.180	0.297
SWN	0.062	0.065	0.115	0.115	0.168

is being identified as more aggressive than irony, the sentiment score in it should be more positive".

In this section, we further investigate the role of the polarity reversal in the three kinds of figurative messages, also in order to understand when the expressed sentiment is only superficially positive. A correlational study is presented in Table 5.5. The results offer further interesting suggestions related to the polarity reversal phenomenon. No relation exists between the polarity values calculated by lexical resources (RES) and the annotation, considering the whole Corpus (C). Our experiment consists in forcing the reversal of RES polarity values for one kind of tweets at a time. Then, we calculate the correlations between these groups and the annotated values. Thus, in *revI* group we only forced the reversal of the RES values for messages tagged with #irony. The same is for #sarcasm (*revS*), #not (*revN*), and both #sarcasm and #not (*revSN*). This clearly states how the correlation improves with the reversal of #sarcasm and #not, while the polarity reversal phenomena is less relevant for ironic messages.

A preliminary manual analysis of the corpus has been performed by two human evaluators with the aim to explore the direction of the polarity reversal phenomenon in sarcastic tweets (i.e., from the positive literal polarity to the negative intended one, or *vice versa*). Such analysis shown that sarcasm is very often used in conjunction with a seemingly positive statement, to reflect a negative one, but very rarely the other way around. In fact, tweets marked with the hashtag #sarcasm and tagged with a positive polarity score were very few in the Semeval2015-Task11 corpus (only 18). Among them, human evaluators could detect only three tweets expressing a literally negative statement, that finally reverted to an intended positive one, as for instance:"RT GregCooper: These annoying home buyers want to purchase my listings before the sign actually goes up. How inconvenient. #sarcasm #grate". This is in accordance with theoretical accounts stating that expressing positive attitudes in a negative mode are rare and harder to process for humans [10]. On the contrary, our evaluators have found many tweets expressing a literally positive statement, that was finally reverted to an intended negative one, as for instance: "There is nothing better than Pitbull singing 'playoffs' as Timber plays in the background. #sarcasm" or "YAY A TEST AND A BUNCH OF HOMEWORK DUE TOMORROW! I LOVE SCHOOL! #sarcasm".

5.5 Classification Experiments

On the basis of the results obtained in identifying differences among the three kinds of figurative messages, we formulate an experimental setting in terms of a classification task. A novel set of structural and affective features is proposed to perform binary classification experiments: #irony-vs-#sarcasm (Iro - Sar), #irony-vs-#not (Iro - Not) and #sarcasm-vs-#not (Sar - Not). The best distinguishing features have been grouped in three sets, including common patterns in the structure of the messages (*Str*), sentiment analysis (*SA*), emotional (*Emot*) features. Structural features include: length, count of colons, question and exclamation marks (*PM*), part-of-speech tags (*POS*). Tweet features (*TwFeat*) refer to the frequency of hashtags, mentions and a binary indicator of retweet. Emotional features belong to two kinds of groups: "Emotional Categories" (*EC*) and "Dimensional Models" (*DM*) of emotions. The first group includes LIWC (positive and negative emotions), EmoSN (surprise, joy, sadness), EmoLex (joy, fear, anger, trust) and SS (anticipation, disgust, joy, like, love). The second group includes ANEW (Valence, Arousal, Dominance), DAL (Pleasantness, Activation and Imagery) and SenticNet four dimensions (Pleasantness, Attention, Sensitivity and Aptitude). In addition, the Sentiment Analysis set is composed by features extracted from SN (SN_polarity and SN_formula), referred as SN_pol in the following tables, as well as positive, negative and polarity values[15] from AFINN, HL, GI, SWN, SUBJ, SO and EWN. Finally, our tweet representation is composed of 59 features (*AllFeatures* henceforth) that have been evaluated over a corpus of 30,000 tweets equally distributed in three categories: 10,000 tweets labeled with #irony and 10,000 with #sarcasm retrieved by [18]. In addition, a novel dataset of 10,000 tweets with the #not hashtag has been retrieved. The criteria adopted to automatically select only samples of figurative use of #not were: having the #not in the last position (without considering urls and mentions) or having the hashtag followed by a dot or an exclamation mark. Only a small percentage of tweets selected according to such criteria resulted to be unrelated to a figurative use of #not[16].

The classification algorithms used are: Naïve Bayes (NB), Decision Tree (DT), Random Forest (RF), Logistic Regression (LR) and Support Vector Machine (SVM)[17]. We performed a 10-fold cross-validation for each binary classification task. F-measure values are reported in Table 5.6. Generally, our model is able to distinguish among the three kinds of figurative messages. The best result is achieved in #irony vs #not classification using Random Forest (0.75). In the #irony vs #sarcasm task, we improve in terms of F-measure the state-of-the-art results (same dataset of [18]) from 0.62 to 0.70 approximately.

[15] We consider polarity values as the difference between the positive and the negative scores.
[16] The dataset with the IDs of the #not tweets is available upon request.
[17] We used the Weka toolkit: `http://www.cs.waikato.ac.nz/ml/weka/`

TABLE 5.6. F-measure values (multiplied by 100) for each binary classification with all features. The underlined values are not statistically significant (t-test with 95% of confidence value)

F-1	Iro - Sar	Iro - Not	Sar - Not
Naïve Bayes	65.4	67.5	57.7
Decision Tree J48	63.4	69.0	62.0
Random Forest	**69.8**	**75.2**	**68.4**
SVM	68.6	74.5	66.9
LogReg	68.7	72.4	64.6

5.5.1 Analysis of Features

To investigate the contribution of the different features further experiments were performed. We divided features into the four main sets already mentioned. Table 5.7 shows the results for ten different configurations. The first experiment involves the use of each set individually (1st row in Table 5.7). From the results, we clearly observe that using only one category of features is not enough. At the same time, we state which group of features are more interesting. Let us comment each subtask. In the *#irony vs #sarcasm* subtask, while the most relevant subsets are *Sentiment Analysis* (0.68 with Logistic Regression) and *emotional categories* (0.634), the worst are the *structural* and *dimensional model of emotions* ones. These results clearly confirm the usefulness of adopting affective resources in the distinction of irony and sarcasm. This is not so evident in the *#irony vs #not* subtask. Notice also that the *structural* set is the most relevant in the *#sarcasm vs #not* subtask. This is coherent with the findings of our preliminary analysis, where "structural" differences in messages have been identified looking at length or punctuation marks.

TABLE 5.7. Comparison of classification methods using ten different feature sets. The underlined values of F-measure (multiplied by 100) are not statistically significant (t-test with 95% of confidence value)

	#irony-vs-#sarcasm					**#irony-vs-#not**					**#sarcasm-vs-#not**				
Conf.	**NB**	**DT**	**RF**	**SVM**	**LR**	**NB**	**DT**	**RF**	**SVM**	**LR**	**NB**	**DT**	**RF**	**SVM**	**LR**
Each set individually															
Str	59.6	60.3	60.9	61.2	61.3	**66.0**	**68.0**	68.6	**69.6**	67.2	**58.9**	**66.2**	**64.5**	**66.1**	**62.6**
SA	**64.1**	**64.4**	**66.2**	**65.1**	**68.0**	63.8	64.4	**70.2**	68.7	**68.0**	54.0	55.5	58.2	57.9	57.4
EC	61.6	62.1	61.7	52.9	63.4	65.0	65.8	64.4	66.2	66.1	54.1	55.3	54.7	56.9	56.4
DM	54.0	57.7	59.9	60.0	59.5	56.9	60.8	63.3	62.6	62.2	53.5	55.1	54.2	56.1	55.5
Combination between sets															
SA+EC	64.4	62.2	67.9	66.1	66.0	67.0	65.3	70.1	68.8	68.5	54.5	54.7	59.7	58.8	58.0
SA+DM	63.5	60.4	66.6	65.7	65.3	64.1	66.6	69.9	67.7	67.6	54.4	54.7	58.8	58.3	58.6
SA+Str	**64.7**	63.2	**69.3**	**67.3**	**67.6**	**67.9**	**69.8**	**75.2**	**73.4**	**71.7**	**58.9**	62.7	**68.3**	66.5	**64.3**
Str+EC	**64.7**	**63.6**	67.5	65.9	66.8	**67.9**	69.7	74.0	72.6	70.3	**58.9**	63.7	67.8	65.5	63.1
DM+EC	62.6	60.7	64.8	64.9	64.5	63.0	63.7	68.1	67.7	66.8	54.5	54.1	56.6	57.5	56.8
DM+Str	59.4	59.6	64.9	64.0	64.6	64.9	67.1	72.7	71.9	69.7	58.2	**64.0**	67.7	**66.9**	63.7

A second experiment presents all possible pair combinations constructed from the four sets (i.e., six different pairs). One of the best results, very similar to those reached by *AllFeatures* (see Table 5.6), is achieved using the "*Sentiment Analysis + Structural*" pair for the *#irony vs #sarcasm* task. In this task, it can be noticed that, while *structural* features alone are not important as detailed in the previous experiment, the result increases just adding features from *emotional categories* or *sentiment analysis*. Furthermore, the *emotional categories* set, combined both with *sentiment analysis* and with *structural* features, obtains relevant results in all the three subtasks.

To further investigate the obtained results from the perspective of the importance of the affective resources, we took into consideration the contribution of individual features. A third experiment includes all pair combinations between the *structural* features (which seems to be a strong indicator in all the binary classification tasks at issue) and each one of the *Sentiment Analysis* and *Emotional* resources (Table 5.8).

TABLE 5.8. Comparison of classification methods using different feature sets. The underlined F-measure values (multiplied by 100) are not statistically significant (t-test with 95% of confidence)

	#irony-vs-#sarcasm					**#irony-vs-#not**					**#sarcasm-vs-#not**				
Conf.	**NB**	**DT**	**RF**	**SVM**	**LR**	**NB**	**DT**	**RF**	**SVM**	**LR**	**NB**	**DT**	**RF**	**SVM**	**LR**
Structural + each resource from SA and Emotional															
Str+AFINN	**63.7**	64.8	66.4	65.6	65.7	**67.3**	**70.8**	72.7	**71.8**	**70.1**	**58.8**	**65.7**	66.4	**66.5**	62.8
Str+HL	63.3	**64.9**	66.3	**66.0**	**66.1**	66.7	70.4	71.6	71.7	68.9	58.6	65.0	65.3	66.1	62.5
Str+GI	59.5	60.5	60.8	61.4	62.2	65.0	67.0	68.2	68.7	66.4	58.6	64.9	64.4	66.0	62.5
Str+SWN	60.0	61.4	**65.1**	62.2	64.5	66.3	69.1	**73.0**	70.8	69.8	58.7	64.7	66.9	66.1	63.1
Str+SN_dim	59.1	58.6	62.9	61.4	62.1	65.0	65.9	70.1	69.8	67.3	58.5	64.6	66.1	65.9	62.9
Str+EWN	57.8	58.1	61.1	60.5	61.4	64.5	65.9	68.8	68.2	65.7	**58.8**	64.3	66.0	65.0	62.6
Str+SO	58.0	60.2	61.6	61.4	60.6	63.7	67.3	69.1	69.0	65.6	56.7	65.4	65.3	66.1	62.5
Str+LIWC	62.7	63.7	64.2	64.8	64.9	66.6	69.6	70.8	70.9	68.6	58.4	64.7	65.1	66.2	62.5
Str+EmoLex	58.6	59.5	61.8	61.2	61.9	65.0	67.5	69.5	69.5	66.5	58.5	64.6	65.3	66.1	62.5
Str+EmoSN	58.3	58.2	60.7	60.2	60.9	66.0	67.1	70.2	68.9	67.2	**58.8**	63.7	65.7	64.9	62.5
Str+SS	61.6	62.4	63.8	63.1	64.1	65.7	68.3	70.1	69.9	67.6	**58.8**	64.4	65.8	66.3	62.6
Str+ANEW	58.1	59.1	62.2	60.9	61.1	64.7	66.6	69.3	68.8	66.2	58.3	65.4	66.2	66.1	62.5
Str+DAL	57.6	58.7	63.1	62.5	63.3	64.7	66.7	70.6	70.0	68.1	58.6	65.0	**67.0**	66.4	**63.2**
Str+SUBJ	60.5	61.7	64.6	63.6	64.0	65.7	68.7	71.3	70.3	67.8	58.6	63.6	66.4	65.8	62.5

First, it is important to note that in many cases, an improvement with respect to the results in [18] is achieved for *#irony vs #sarcasm*. The higher contribution is given by resources AFINN, HL, LIWC, SS and SUBJ. In *#irony vs #not*, the F-measure is higher when the *structural* set is applied together with AFINN, HL, SWN, and LIWC, including also SUBJ, SN, SS, DAL, and EmoSN. In the *#sarcasm vs #not* task, where only DAL slightly improves the results for each classifier, measures are not as clear.

Further experiments are specifically related to *Sentiment Analysis* and *Emotional* sets. Each resource in the *Emotional* set is combined with the *Sentiment Analysis*

one and *vice versa* (Table 5.9). Generally, adding an *Emotional* resource to the *Sentiment Analysis* set in *#irony vs #not* and *#sarcasm vs #not* tasks, most of the times allows to obtain better results than adding a *Sentiment Analysis* feature to the *Emotional* one. This does not happen in *#irony vs #sarcasm* task.

TABLE 5.9. Comparison of classification methods using different feature sets. Best performances for each classifier are in bold. The underlined F-measure values (multiplied by 100) are not statistically significant (t-test with 95% of confidence)

	#irony-vs-#sarcasm					**#irony-vs-#not**					**#sarcasm-vs-#not**				
Conf.	**NB**	**DT**	**RF**	**SVM**	**LR**	**NB**	**DT**	**RF**	**SVM**	**LR**	**NB**	**DT**	**RF**	**SVM**	**LR**
SA + each resource from Emotional															
SA+LIWC	64.2	**61.3**	66.7	65.5	65.2	**65.0**	64.5	**70.7**	69.3	68.1	53.8	**55.2**	58.3	58.3	57.5
SA+EmoLex	64.2	60.6	66.7	65.2	65.2	63.3	64.3	70.3	68.9	67.9	52.3	54.2	57.8	56.4	56.9
SA+EmoSN	64.0	60.0	**66.8**	65.2	65.0	64.2	**64.8**	70.6	**69.0**	68.2	54.9	54.4	58.8	58.2	58.2
SA+SS	64.2	61.2	66.7	65.2	65.4	64.6	64.6	70.4	69.0	**68.2**	**55.0**	**55.2**	**59.3**	**58.5**	58.2
SA+ANEW	64.2	60.6	66.5	65.3	65.0	63.6	64.5	70.6	68.8	68.0	53.9	**55.2**	58.7	58.3	57.4
SA+DAL	63.8	60.2	66.6	**65.7**	**65.5**	63.9	64.4	70.2	69.0	68.0	54.6	**55.2**	58.6	58.1	**58.5**
SA+SN_dim	**64.3**	60.6	66.5	65.1	65.0	63.4	64.4	70.6	68.8	68.0	53.8	54.9	58.5	58.0	57.7
Emotional (EC+DM) + each one of the resources from SA															
Emot+AFINN	63.8	**61.8**	65.8	65.3	64.9	**64.4**	64.1	68.9	67.8	67.3	54.4	54.4	57.0	57.7	57.3
Emot+HL	**64.1**	**61.8**	66.2	**65.6**	**65.7**	**64.4**	**65.1**	69.1	**68.6**	67.6	54.5	**54.6**	56.7	57.7	57.0
Emot+GI	62.6	60.9	65.2	64.7	64.8	63.1	63.4	68.0	67.7	67.0	54.5	54.3	56.6	57.8	57.1
Emot+SWN	63.2	60.7	66.0	**65.6**	65.4	63.3	63.7	68.9	68.3	67.6	54.9	53.8	57.1	57.7	56.9
Emot+SN_pol	62.4	61.3	64.7	64.5	64.6	64.1	63.5	69.1	67.8	**67.7**	**55.1**	54.4	**57.8**	57.8	**58.6**
Emot+EWN	62.1	60.5	65.4	64.6	64.6	63.0	63.5	67.7	67.4	66.4	55.0	53.9	57.5	**58.6**	57.4
Emot+SO	62.4	61.1	65.8	64.8	64.5	61.8	64.9	68.3	67.6	66.5	53.1	54.1	56.4	57.6	56.8
Emot+SUBJ	63.4	61.1	**66.5**	**65.6**	65.6	63.5	63.7	**69.5**	68.1	67.3	54.5	54.0	56.9	57.9	56.9

In a last experiment, we performed feature ablation by removing one feature or one group of features (i.e. all the features belonging to a particular resource) at a time in order to evaluate the impact on the results. First, we investigated the effects of each structural features, in Table 5.10, where bold values highlight the most important results. A drop in performance for each subtask can be observed when Punctuation Marks (PM) are removed. Furthermore, removing the length features also significantly affects the overall performance for *#irony vs #not* and *#sarcasm vs #not* tasks. These results confirm the role of punctuation marks and length, as described by Figures 5.1 and 5.2 in Section 5.4.

Moreover, to measure the contribution of each resource in the *Sentiment Analysis* and *Emotional* sets, we proceeded by feature ablation in Table 5.11. The most relevant resources are HL in *#irony vs #sarcasm* and *#irony vs #not* tasks, and EWN in *#sarcasm vs #not* task. The most relevant emotional resources are LIWC in *#irony vs #sarcasm* and EmoSN in *#sarcasm vs #not* task. Both of them are relevant in the *#irony vs #not* task. As we have already noted, the Dictionary of Affective Language is the most relevant among the *dimensional model of emotions* ones, in the three tasks.

TABLE 5.10. Comparison of classification methods with feature ablation. Worst performances for each classifier are in bold, to underline the more relevant role of the feature removed. The underlined values are not statistically significant (t-test with 95% of confidence value)

Structural - one of the resources each time

	#irony-vs-#sarcasm					**#irony-vs-#not**					**#sarcasm-vs-#not**				
Conf.	**NB**	**DT**	**RF**	**SVM**	**LR**	**NB**	**DT**	**RF**	**SVM**	**LR**	**NB**	**DT**	**RF**	**SVM**	**LR**
Str	59.6	60.3	60.9	61.2	61.3	66.0	68.0	68.6	69.6	67.2	58.9	66.2	64.5	66.1	62.6
Str-lenght	59.2	59.9	58.0	61.1	60.6	**62.8**	66.9	**64.8**	68.0	66.9	**55.7**	63.6	62.0	64.0	61.7
Str-PM	**57.9**	**58.1**	**57.8**	**59.3**	**59.9**	64.8	**66.1**	66.0	**67.7**	**65.2**	58.2	**62.3**	**59.6**	**62.1**	**58.9**
Str-POS	59.2	60.5	58.2	60.7	60.5	65.1	70.0	67.4	69.9	67.1	56.7	66.9	64.8	66.8	62.4
Str-TwFeat	59.8	60.5	58.8	59.9	60.8	66.2	69.0	67.3	69.4	67.0	58.6	65.7	62.7	64.7	60.7

TABLE 5.11. Comparison of classification methods with feature ablation. Lowest performances for each classifier are in bold, indicating the greater contribution of the feature removed. The underlined values are not statistically significant (t-test with 95% of confidence value).

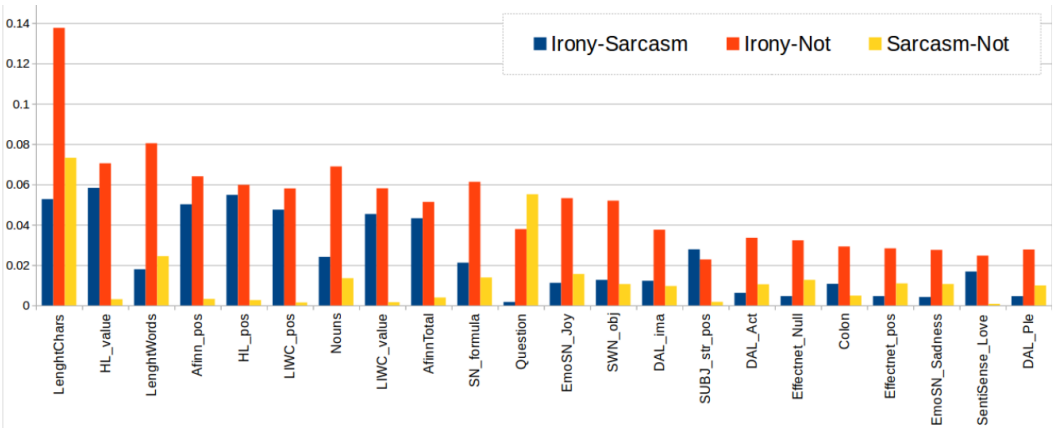
	#irony-vs-#sarcasm					**#irony-vs-#not**					**#sarcasm-vs-#not**				
Conf.	NB	DT	RF	SVM	LR	NB	DT	RF	SVM	LR	NB	DT	RF	SVM	LR
SA - one of the resources each time															
SA	64.1	64.4	66.2	65.1	68.0	63.8	64.4	70.2	68.7	68.0	54.0	55.5	58.2	57.9	57.4
SA-AFINN	63.0	60.9	65.8	64.8	64.8	62.9	64.3	**69.4**	68.4	67.8	53.9	54.6	58.6	57.7	57.2
SA-HL	**62.7**	60.9	**65.2**	**63.8**	**63.8**	62.7	**63.5**	69.8	**67.5**	**66.9**	54.4	54.1	58.2	57.6	57.3
SA-GI	64.2	61.1	66.2	65.2	65.0	64.0	65.3	69.9	68.9	68.0	54.2	55.4	58.5	57.9	57.4
SA-SWN	63.8	61.2	65.6	64.8	64.6	63.4	64.4	69.8	68.3	67.6	53.4	55.0	57.3	57.4	57.2
SA-SN	64.1	**60.7**	66.2	65.3	65.1	**62.6**	64.5	69.5	68.5	67.5	53.1	54.7	57.6	57.9	**55.8**
SA-EWN	63.8	62.1	66.5	64.8	65.0	63.7	65.4	**69.4**	68.5	67.8	**52.5**	**53.3**	**57.1**	**56.2**	57.0
SA-SO	64.1	61.0	66.1	64.4	65.0	64.2	66.0	69.6	68.0	67.5	55.5	55.3	58.2	58.0	57.4
SA-SUBJ	64.0	61.8	65.5	65.1	64.5	64.2	64.8	70.0	68.7	67.9	53.9	55.3	58.0	57.7	57.4
EC - one of the resources each time															
EC	61.6	62.1	61.7	52.9	63.4	65.0	65.8	64.4	66.2	66.1	54.1	55.3	54.7	56.9	56.4
EC-LIWC	**60.0**	**60.0**	**59.3**	**61.4**	**60.9**	**62.1**	**64.6**	**62.9**	64.6	64.6	54.5	55.4	54.9	57.7	56.5
EC-EmoLex	61.6	62.0	60.2	65.1	63.1	65.2	66.2	64.1	65.8	65.8	54.9	56.3	53.7	57.0	56.6
EC-EmoSN	61.5	62.1	61.5	62.2	62.2	63.1	63.9	63.4	**64.0**	**63.8**	**50.1**	**52.3**	**52.2**	**53.4**	**52.7**
EC-SS	61.7	61.9	59.7	62.5	62.8	64.0	66.1	63.6	66.1	65.7	54.1	56.5	54.3	56.8	56.4
DM - one of the resources each time															
DM	54.0	57.7	59.9	60.0	59.5	56.9	60.8	63.3	62.6	62.2	53.5	55.1	54.2	56.1	55.5
DM-ANEW	54.4	57.6	59.0	59.4	59.3	57.7	60.5	62.7	62.2	61.6	53.9	55.3	54.2	55.6	55.3
DM-DAL	**51.9**	**54.3**	**58.2**	**54.9**	**54.9**	**53.3**	**57.2**	**60.8**	**57.2**	**57.1**	**51.6**	**53.6**	**52.8**	**53.7**	**53.3**
DM-SN_dim	53.7	57.4	58.9	59.4	59.0	57.5	60.7	61.8	62.0	61.8	53.7	55.1	55.0	56.2	55.4

5.5.2 Information Gain

In order to measure the relevance that a single feature provides in our classification model, we calculated the Information Gain for each binary experiment. According to Figure 5.4, most features among the best ranked ones (17 over 22) are related to sentiment and emotion resources (e.g. HL, AFINN, SN, LIWC, DAL, SWN).

This clearly confirms the importance of this kind of features in figurative language processing.

FIGURE 5.4. Information Gain values for the 22 best ranked features in binary experiments.



Sentiment and affective features are more relevant in the *#irony vs #sarcasm* task, including terms with positive valence from different lexicons. In particular, 6 over the first 7 features are related to the HL, AFINN and LIWC lexicons.

Structural features are more relevant in the *#irony vs #not* task, together with the Sentiment Analysis ones. In particular, the length of messages both in characters and in words plays an important role. Interestingly, besides the structural features, the three emotional dimensions of DAL are useful to discriminate between figurative messages. Imagery is the most relevant dimension in this task. A special mention is reserved for Objectivity terms from SWN and neutral events from EWN: we think that their relevance could be related to the larger presence of events in #not, detected thanks to the quantity analysis related to EWN reported in Table 5.2.

In the *#sarcasm vs #not* subtask, the structural features play a relevant role, outperforming the other subsets. This is true also for *#irony vs #not*, coherently with previous analysis (i.e., punctuation marks play an important role, as observed also in Figure 5.2). The relevance of question marks is notable. This is coherent with our preliminary analysis and with the idea that a sort of self-mockery is expressed by this kind of messages.

The three subtasks clearly indicate the usefulness of adopting lexical resources that linked to semantic information, such as the one encoded in emotional categories and dimensional models of emotion.

5.6 Conclusions

In this paper, we investigated the use of figurative language in Twitter. Messages explicitly tagged by users as #irony, #sarcasm and #not were analysed in order to test the hypothesis to deal with different linguistic phenomena. In our experiments we took into account emotional and affective lexical resources, in addition to structural features, with the aim of exploring the relationship between figurativity, sentiment and emotions at a finer level of granularity. Classification results obtained confirm the important role of affective content. In particular, when sentiment analysis and emotional resources are used as features, for #irony vs #sarcasm an improvement w.r.t. state-of-the-art results is achieved in terms of F-measure.

As for the separation of #irony vs #not and #sarcasm vs #not, our results contribute to shed light on the figurative meaning of the #not hashtagging, which emerges as a distinct phenomenon. They can be considered as a baseline for future research on this topic. We also created a dataset to study #not as a category on its own[18].

An assumption underlying our proposal concerns the reliability of the user-generated hashtags #irony and #sarcasm as labels exploited by Twitter users in English speaking countries to really mark *distinguished* phenomena. Let us notice that the use of hashtags marking irony and sarcasm can be language-specific. It can vary in different languages and cultures, and similar markers in different languages could have different distributions. For what concerns English tweets, in order to get an idea about the distribution of the three hashtags investigated in our study, we collected a sample of English tweets posted on a single day[19]. After some pre-processing steps inspired by [89], mainly devoted to discard re-tweets and to filter out tweets where the hashtags were not used to invite an ironic or sarcastic interpretation of the post, we counted 1,461 tweets: 411 marked with #irony, 698 with #sarcasm and 352 with #not. We can observe that the distribution in case of English tweets seems to be not very imbalanced. This is in favor of the hypothesis that users, in this linguistic context, really exploit the three hashtags in order to mark different phenomena. Different findings have been reported about Dutch tweets in [89], where a similar experiment shown that irony-tweets (i.e., tweets marked with #ironie, the Dutch equivalent of #irony) were very rare; in such a scenario it would be hard to state that irony tweets are really exploited by Dutch users in order to mark a phenomenon which is different from sarcasm. A cross-language study of markers for irony and sarcasm could be an interesting strand of future research.

Another interesting direction to further investigate is the educational and socio-demographic background of irony-users and sarcasm-users. Unfortunately, in Twitter explicit meta data about age and gender of users are not provided, thus extracting such information is a further issue that needs to be addressed. Nevertheless, for some authors it is possible to manually inspect the information that they may have

[18] Available under request.

[19] We retrieved from Twitter Streaming API all tweets in English language (lang: 'en') from 2016-02-01 12:00:00 to 2016-02-02 12:00:00.

published in other social media, e.g. LinkedIn[20], on their user's profile. For what concerns age, in case the information is not published in the user's profile, it could be approximated taking into account, if present, the information included in the education section, for instance, the degree starting date. For what concerns the information about gender, it could be inferred form the user's photography and name, by following a methodology similar to the one exploited in [123].

In this work we focused on the new task of differentiating between tweets tagged with #irony, #sarcasm and #not, in order to provide some useful insights on the use of these hashtags to label what users consider as ironic or sarcastic content in a social media platform such as Twitter. Investigating the application of our approach in distinguishing between ironic and sarcastic tweets in absence of the explicit hashtags could be also an interesting matter of future work. Moreover, since our analysis shows that different kinds of figurative messages behave differently with respect to the polarity reversal phenomenon (see Table 5.5, Section 5.4.2), in future work we will further experiment the impact of our findings on the sentiment analysis task, investigating if our classification outcome can be a useful precursor to the analysis. Some of the results reported here about the *polarity reversal* phenomenon in tweets tagged as #sarcasm and #not have been already exploited in a sentiment analysis task by the ValenTo system, obtaining promising results [73].

Acknowledgments

The National Council for Science and Technology (CONACyT Mexico) has funded the research work of Delia Irazú Hernández Farias (Grant No. 218109/313683 CVU-369616). Paolo Rosso has been partially funded by *SomEMBED* MINECO research project (TIN2015-71147-C2-1-P) and by the Generalitat Valenciana under the grant ALMAMATER (PrometeoII/2014/030). The work of Viviana Patti was partially carried out at the Universitat Politècnica de València within the framework of a fellowship of the University of Turin co-funded by Fondazione CRT (WWS Program 2).

[20] `http://www.linkedin.com`

Chapter 6

Sentiment Polarity Classification of Figurative Language: Exploring the Role of Irony-Aware and Multifaceted Affect Features

In this chapter, we describe a pipeline system for sentiment analysis. Our model consists of two phases: first, we incorporated emotIDM to identify irony; then, the polarity degree of a tweet is calculated taking into account a broad range of features related to affect phenomena together with features indicating the presence of ironic intention. For the sake of comparison, we evaluate our approach on a benchmark dataset and contrast our results against the ones of the first shared task on sentiment analysis in figurative language. Our pipeline has a comparable performance with the best-ranked systems validating the relevance of affect-related and irony-aware features for sentiment analysis.

Published in:

Hernández Farías, D.I., Bosco, C., Patti, V., and Rosso, P. (2017). Sentiment Polarity Classification of Figurative Language: Exploring the Role of Irony-Aware and Multifaceted Affect Features. In: *Proceedings of the 18th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2017)*, In Press.

Abstract

The presence of figurative language represents a big challenge for sentiment analysis. In this work, we address the task of assigning sentiment polarity to Twitter texts when figurative language is employed, with a special focus on the presence of ironic devices. We introduce a pipeline model which aims to assign a polarity value exploiting, on the one hand, irony-aware features, which rely on the outcome of a state-of-the-art irony detection model, on the other hand a wide range of affective features that cover different facets of affect exploiting information from various sentiment and emotion lexical resources for English available to the community, possibly referring to different psychological models of affect. The proposed method has been evaluated on a set of tweets especially rich in figurative language devices proposed as a benchmark in the shared task on "Sentiment Analysis of Figurative Language" at SemEval-2015. Experiments and results of feature ablation show the usefulness of irony-aware features and the impact of using different affective lexicons for the task.

6.1 Introduction

Twitter has provided a huge volume of data containing judgments, attitudes, and beliefs of people. Opinions and their related concepts such as sentiments and emotions are the subjects addressed by Sentiment Analysis (SA) [94]. Figurative language devices, such as, for instance, irony and sarcasm, represent one of the main challenges for SA [102]. The presence of these kinds of expressions could indeed undermine the accuracy of SA systems [25]. Therefore, identifying irony and sarcasm become crucial for a SA system.

Among the definitions proposed in theoretical pragmatics for irony there is that of Grice [66], which refers to the speaker's intention to express the opposite meaning of what it is literally said. When irony becomes offensive with a specific target to attack it is considered as a form of sarcasm [26, 90]. Irony detection in social media has become a hot research topic and many research works have been carried out recently on this topic, with a special focus on Twitter data [126, 127, 18, 120, 81, 115]. Most of the current approaches consider irony as an umbrella term that covers also sarcasm.

A SA system may fail when applied to inferring the polarity in sentences like:

(1) *Thanks for this birthday card.*
I'm really glad you didn't put any money in it.[1]
(2) *My level of annoyance is at an all time high right now. Thanks to this wonderful @Starbucks experience. #wah*[2].
(3) *RT GregCooper: These annoying home buyers want to purchase my listings before the sign actually goes up. How inconvenient. #sarcasm #grate.*[3]

In (1) the overall polarity of the tweet is negative but the presence of three positive terms ("Thanks", "birthday" and "glad") could be misinterpreted by sentiment analysis systems, which often rely on information included in sentiment lexicons. Also in sarcastic posts like the one reported in (2) [127], where there is a contrast between a positive sentiment expressed and a situation which is typically negative, the presence of two positive terms ("thanks" and "wonderful") and of a negative one ("annoyance") could cause a problem to SA system in assigning the correct polarity. In such cases, indeed, tweets could be identified as positive under a basic approach for SA which simply considers the presence and frequency of positive and negative terms to assign polarity. However, both tweets convey a meaning far from being positive: the authors use irony/sarcasm to express their evaluation towards a target, by using a literally positive sentence to point out their real negative opinion on the specific target.
While the use of sarcasm to convey a negative sentiment is the most common, it is very rare to use it the other way around. Theoretical accounts state that expressing positive attitudes in a negative mode is rare and harder to process for humans [10]. This seems to be confirmed by an analysis of the SemEval-2015 Task 11 corpus, where tweets marked with the hashtag #sarcasm and tagged with a positive polarity score were very few (only 18 out of 2,260 posts). Among them, only three tweets expressing a literally negative statement, that finally reverted to an intended positive one, were identified by manual analysis [137], see for instance post (3) reported above.

Currently, even if SA systems are able to understand the most salient polarity of words, they do not have a well-established methodology to deal with the presence of figurative language expressions [99]. In this sense, in order to develop an irony-aware system which correctly identifies the sentiment behind a text, it is needed to recognize whether the sentence contains some figurative device, such as irony, before deciding on sentiment polarity. In general, irony detection and SA have been addressed individually. However, there are some efforts devoted to integrate both tasks in the framework of evaluation campaigns, where the main objective is to perform Twitter SA considering the presence of irony [51, 20, 14].

In order to investigate whether the performance of a SA system improves or not when it takes into account the presence of ironic content, we propose an approach based on a pipeline that incorporates two modules: irony detection and SA for polarity assignment. To the best of our knowledge, exploiting an irony detection module in a sentiment analysis pipeline has not been investigated in depth before. In our approach, the irony detection module was trained by using a set of tweets labeled as ironic. Whereas the sentiment analysis one was trained by using tweets with figurative language manually annotated with their polarity degree.

[1]This tweet is part of the dataset used in the SemEval-2015 Task 11: Sentiment Analysis of Figurative Language in Twitter [106]. It was labeled as having negative polarity (-1.8).
[2]This tweet is part of the sarcastic tweets in the dateset of Riloff et al. [127].
[3]This tweet is part of the dateset used in the SemEval-2015 Task 11: Sentiment Analysis of Figurative Language in Twitter [106]. It was labeled as having positive polarity (+0.63).

The paper is organized as follows. Section 6.2 introduces the SA task on figurative language. Section 6.3 describes our method to perform irony-aware SA. Section 6.4 describes the evaluation and results. Finally, Section 6.5 draws some conclusions.

6.2 Sentiment Analysis and Figurative Language

The *SemEval-2015 Task 11: Sentiment Analysis of Figurative Language in Twitter*[4] was the first SA task attempting to identify the sentiment score in texts featured by the occurrence of figurative language devices. The goal of the task was to determine the degree in which a sentiment was communicated in a fine-grained scale ranging from -5 (very negative) to +5 (very positive) over a set of tweets rich in metaphorical, sarcastic and ironical content. Overall, the dataset included more than 13,000 tweets (SE15-Task11 dataset, henceforth).

Fifteen teams participated in the task on SA of figurative language [51]. Their systems were evaluated using the cosine similarity (cosSim) measure. The best ranked system, called ClaC [109], exploited n-grams, some SA resources as well as linguistic features such as negations and modality. ClaC achieved a cosSim measure of 0.758. The UPF-taln [16] system considered a set of features to detect the style and the unexpectedness in tweets combined with textual features such as bigrams, skipgrams and other word patterns. It achieved the second place in the ranking with a 0.711 in cosSim terms.

LLT_PolyU [155] and EliRF, [57], ranked as third (0.687) and fourth (0.658), respectively, considering features such as n-grams, negations as well as some SA resources. LT3 [142] ranked as fifth (0.687) and ValenTo [73], ranked as sixth (0.634), systems included in their sets of features the presence of punctuation marks, emoticons and hashtags. LT3 took advantage of features such as contrasting, contradictory and polysemic words. ValenTo system exploited SA resources as well as some emotional and psycholinguistic information. Besides, it considered the presence of sarcastic content in a tweet by exploiting specific hashtags.

6.3 Our Proposal

The aim of our approach is to perform irony-aware SA by exploiting different facets of affective information. Our irony-aware SA attempts to incorporate two strongly related tasks: irony detection and sentiment analysis. The importance of considering the presence of ironic content before performing sentiment analysis has been recognized by several authors [51, 72]. Our main objective is not only identifying the presence of ironic content but rather assigning a polarity value consistent with its detected presence. The overall process in our irony-aware SA system can be briefly summarized as follows:

Given a tweet, we first identify the presence of ironic content. Then, both the tweet and its irony-aware features are processed by a sentiment analysis model in order to calculate a polarity value for the post.

Unlike the best ranked systems at SemEval-2015 Task 11, our approach does not exploit n-grams as features. Instead our irony-aware SA system mainly relies on affective information for both identifying irony and calculating a polarity degree.

[4] http://alt.qcri.org/semeval2015/task11/

We propose a pipeline involving two main phases:

I. *Irony-aware features.* As output of this step we have two irony-aware features: the first one depends on the possible presence of explicit irony-related hashtags, whereas the second one is obtained by using the irony detection model described in Section 6.3.1. In Twitter messages hashtags such as "#irony", "#sarcasm", and "#not" can indeed be recognized as labels used to point out user's ironical intention [147, 137]. However, this is not enough to identify irony. When a SA system is dealing with a tweet as the one mentioned in Section 6.1, in which no explicit hashtag indicating the user ironic intention is present, it is needed to apply a model able to identify irony without considering potentially ironic hashtags. Thus, we exploit an irony detection model (see Section 6.3.1) by using a set of 10,000 ironic tweets retrieved by Reyes et al. [126], 10,000 sarcastic and 20,000 non-ironic tweets retrieved by Ptáček et al. [120]. For this purpose, we trained the Weka [67] implementation of a set of classifiers (Naïve Bayes, Decision Tree and Support Vector Machine) with default parameters. Then, the "ironic" or "non-ironic" label was determined by a majority vote between these classifiers.

II. *Polarity assignment.* The polarity degree of a tweet is the output of this step. It is determined by a SA model that exploits a set of features that covers not only textual markers but also affective information as well as the irony-aware features obtained from the previous step. Since the irony detection model we exploited here does not distinguish between different types of figurative language, such as irony and sarcasm, we decided to use the presence of ironic content only as a feature for assigning polarity rather than for reverting the polarity of a given tweet. This is motivated by the results of the analysis in Sulis et al. [137], which highlights that tweets tagged with #irony and #sarcasm behave differently with respect to the polarity reversal phenomenon. In fact, with respect to the twist of the polarity in tweets tagged with #irony and #sarcasm, it has been observed that, when the #sarcasm hashtag is used, it is common to have a full polarity reversal (from a polarity to its opposite, mostly from positive to negative polarity), while, when #irony is used, there is often just an attenuation of the polarity (mostly from negative to neutral). See also [25] for a similar study about this issue on the Italian corpus Senti-TUT.

6.3.1 Irony Detection Model

Irony has been recognized as a linguistic device strongly connected with the expression of feelings, emotions, attitudes and evaluations [66, 151, 5]. We relied on a state-of-the-art irony detection model: *emotIDM* described in [71]. emotIDM detects irony in Twitter taking advantage of different facets of affective content by exploiting a wide range of resources available for English. Such facets include sentiment and emotional aspects in a finer-grained way by capturing information from both categorical and dimensional models of emotions. Besides, it considers textual markers (such as punctuation marks, part-of-speech labels, emoticons, and specific Twitter's markers) that have been recognized as reliable clues for identifying ironic intent in social media. *emotIDM* considers irony as an umbrella term that covers also sarcasm. It outperforms the state-of-the-art results validating the importance of affect-related information for detecting ironic content in tweets.

6.3.2 Sentiment Analysis Model

The SA model takes ValenTo system [73] as starting point and improve its performance by adding lexical resources with the aim to capture affective information[5]. We chose to use ValenTo system for two reasons: 1) It does not include bag-of-words (BOW) as features to perform SA. Such features can be highly topic and domain dependent. We are instead

[5] `https://github.com/ironyAware-SA/sentimentAnalysisFeatures`

interested in proposing a model exploiting mainly affective information, and therefore it considers features able to capture this kind of information disregarding domain. 2) It includes a feature to identify ironic content by exploiting the presence of hashtags.

The SA module in our pipeline is then composed by seven groups of features:

i) *Structural*: punctuation marks, POS labels, uppercase chars, URL, and emoticons.

ii) *Twitter markers*: hashtags, mentions and retweets.

iii) *Sentiment modifiers*: elongated words, interjections and negations.

iv) *Sentiment Analysis lexica*: AFINN [107], Hu&Liu (HL) [76], SentiWordNet (SWN) [12], SenticNet polarity (SNpol) [33], Emolex polarity (EmoLexPol) [103], General Inquirer (GI) [136], Sentiment140 (S140) and NRC Hashtag Sentiment Lexicon (NRC-Hash) [101], MPQA [152] and Sentiment-Pattern[6] (sPat).

v) *Categorical models of emotions*: Emolex emotions (EmoLexEmot), EmoSenticNet (EmoSN) [118], Linguistic Inquiry and Word Count (LIWC) [113], and DepecheMood (DM) [135].

vi) *Dimensional models of emotions*: ANEW [27], Dictionary of Affect in Language (DAL) [150], and SenticNet (SNemot).

vii) *Irony-aware*: two binary features are also considered in order to take into account the presence of ironic intent (*ironyIDM*) as well the presence of an ironic hashtag (*ironyHashtag*). These features are obtained in the first phase of our pipeline.

The polarity assignment is carried out by building a regression model. We used the Weka implementation of M5P, a decision tree regressor. We experimented with other algorithms, and found that the results were worst than those obtained using M5P.

6.4 Evaluation

We experimented with the SE15-Task11 dataset; it is distributed in *training* (8,000 tweets), *trial* (1,000 tweets), and *test* (4,000 tweets). The organizers of the task retrieved tweets rich on figurative language by considering either the presence of specific hashtags (such as #irony, #sarcasm, and #not) or words commonly associated with the use of metaphor (such as “literally" and “virtually"). We present experimental results for the *test* set used in SE15-Task11. For the training phase, we used the remaining tweets. As evaluation measures we used the cosine similarity (cosSim) and the Mean-Squared-Error (MSE) as were defined in [51]. cosSim is calculated as the cosine between the vector containing the golden labels in the *test* set and the vector with the results obtained by our pipeline. A score of 1 is achieved when a given system provides all the same scores than in the *test* set. For what concerns to MSE, lower measures of it indicates better performance.

In order to evaluate the effectiveness of our method, we trained the SA module in the pipeline by using each group of features described in Section 6.3.2 individually as well as different combinations among of them. It is important to highlight that we applied the same irony detection model in all the experiments. To further investigate the importance of the different lexica considered in our model, we evaluated the sentiment analysis, categorical models of emotions, and dimensional models of emotions groups of features by removing an affective resource each time.

[6] `http://www.clips.ua.ac.be/pages/pattern-en#sentiment`

Finally, we also are interested in to find how well different groups of features performed when bag-of-words are also exploited (unigrams with binary representation were used as BOW features). Our experimental setting was two-fold: i) To demonstrate the robustness of our method in assigning polarity, by exploiting high-level features comprising mainly affective information from different aspects; and ii) To compare the performance of our model when n-grams are combined with the set of features described in Section 6.3.2.

6.4.1 Results

Table 6.1 shows the results of our system in cosSim and MSE terms. *All features* label in the first row of the table refers to all the features described in Section 6.3.2 (composed by a total of 140 features). The second row shows the performance of our sentiment analysis module when the irony-aware features are removed from *All features*. As can be noticed, there is a drop, although small, in the performance of our system. This result could provide an insight useful to validate our hypothesis about the usefulness of recognizing irony before performing SA. Therefore, in the rest of the experiments the irony-aware features were always considered.

TABLE 6.1. Comparison of the performance of our approach when it is evaluated with and without irony-aware features. Both results are statistically significant.

Features	**cosSim**	**MSE**
All features	0.689	2.640
All features without irony-aware features	0.673	2.836

Table 6.2 shows the performance of the pipeline when the SA module is trained with different sets of features. From Table 6.2 it can be appreciated that, in general, our model outperforms the official result before achieved by ValenTo (0.634 in cosSim). As can be noticed, the result obtained in the experiment involving saLex group of features together with irony-aware features is the best one with respect to all the groups of features in the sentiment analysis module. Our best result in conSim terms slightly outperforms the one obtained by the second place in the official ranking in the SemEval-2015 Task 11 (0.710 in cosSim). It uses irony-aware features in addition to some widely known features for sentiment analysis related tasks.

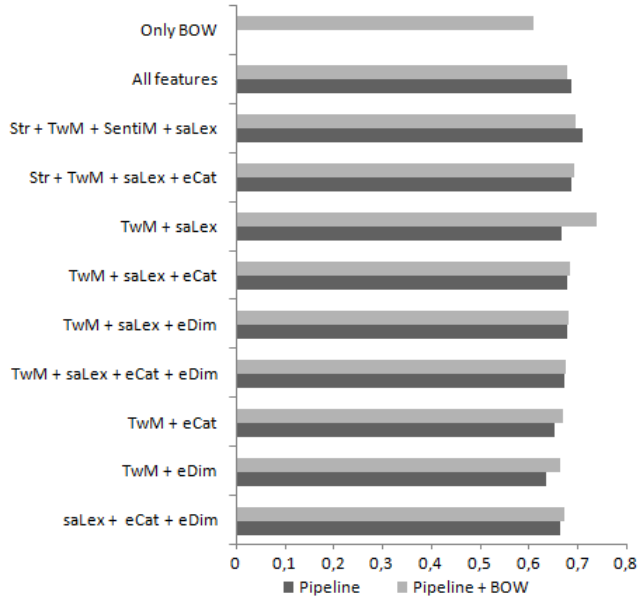
In order to investigate the performance of the resources exploited in our approach, we performed feature ablation by removing one resource in the sentiment analysis, categorical models of emotions and dimensional models of emotions groups of features. Figure 6.1 shows the results of this experiment. As can be noticed the performance of our irony-aware system when it exploits each group of features individually is still competitive. It seems that removing the resource *S140* from the saLex group provokes the biggest drop in the performance of our pipeline. *LIWC* could be considered as one of the most important resource in the eCat group. Furthermore, when *SenticNet* is not considered, the performance of our pipeline decreases with respect to using all the features in the eDim group. On the other hand, there are some resources that when are removed allow us for a slight improvement of the performance in terms of cosSim.

Additionally, we carried out experiments by adding bag-of-words (more than 10,000 features composed the set of those coming from n-grams) together with our set of features. Figure 6.2 shows the obtained results. When we experimented by using BOW combined only with irony-aware features, the cosSim achieved was 0.61. Our evaluation shows that the proposed pipeline achieves comparative performance at assigning polarity degree even without exploiting BOW. Besides, the dimensionality of the feature space in our model is noticeably lower when compared with BOW. This means that by using our set of features it

TABLE 6.2. Performance of the proposed pipeline in cosine similarity and MSE terms by using different features in the sentiment analysis module. All the experiments use also the features belonging to the irony-aware group.

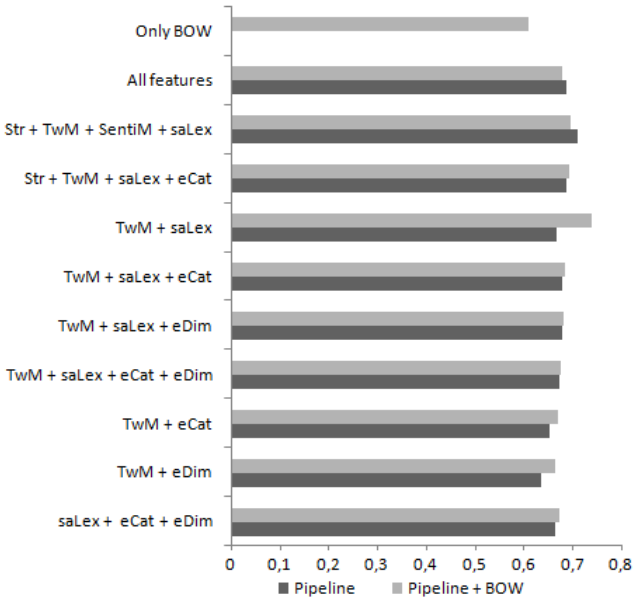
Features	cosSim	MSE
Structural (Str)	0.588	3.381
Sentiment modifiers (SentiM)	0.558	3.498
Twitter markers (TwM)	0.589	3.458
Sentiment analysis lexica (saLex)	**0.67**	**2.836**
Categorical models of emotions (eCat)	0.63	3.070
Dimensional models of emotions (eDim)	0.60	3.296
Str + TwM + SentiM + saLex (group1)	**0.711**	**2.504**
Str + TwM + saLex + eCat (group2)	0.687	2.663
TwM + saLex (group3)	0.669	2.797
TwM + saLex + eCat (group4)	0.68	2.705
TwM + saLex + eDim (group5)	0.678	2.737
TwM + saLex + eCat + eDim (group6)	0.674	2.771
TwM + eCat (group7)	0.653	2.879
TwM + eDim (group8)	0.635	3.070
saLex + eCat + eDim (group9)	0.665	2.856

FIGURE 6.1. Ablation experiment results in cosine similarity terms for the (a) saLex, (b) eCat, and (c) eDim groups of features.



is possible to obtain a lower computational cost with a set of relevant features for assigning polarity in tweets with figurative language. Our two best results 0.71 (by using group1) and 0.74 (by using group3 + BOW) are not higher than the one of the best ranked system in the task (0.758 in cosSim), but they are still competitive and reach the second position in the official ranking, showing that affective information helps.

FIGURE 6.2. Cosine similarity results of applying different groups of features together with bag-of-words.



6.5 Conclusions

In this paper we have shown that including irony detection is a relevant step for sentiment analysis. The experiments described were conducted on a Twitter dataset including a rich variety of figurative language devices, labeled with sentiment at a fine-grained level. We obtained comparable results to the best systems in the SE15-Task11 and show that features related to affective content play an important role. As future work, it would be interesting to distinguish between irony and sarcasm [84, 137] in order to reason on the possibility to apply different polarity reversal criteria. Moreover, we consider also to employ other kinds of resources for improving the coverage of our approach such as the one described in [148] that is based on ANEW. Besides, we are planning to improve the irony detection module by exploiting not only affective information but also aspects related to pragmatic context [82]. Furthermore, we are interested in evaluating our system using datasets coming from different social media as well as in other languages.

Acknowledgments

The National Council for Science and Technology (CONACyT Mexico) has funded the research work of the first author (Grant No. 218109/313683 CVU-369616). The work of Paolo Rosso has been partially funded by the SomEMBED TIN2015-71147-C2-1-P MINECO research project and by the Generalitat Valenciana under the grant ALMAMATER (PrometeoII/2014/030).

Chapter 7

Discussion of the results

Aiming to provide a more comprehensive evaluation of our emotIDM model, we carried out further experiments. We study the performance of the features in emotIDM by experimenting with different configurations of them. Besides, we also assess the classification results with our model when it is used under data with various imbalance degrees. Finally, we performed a qualitative analisys on some of the affective resources in emotIDM.

7.1 Introduction

In this thesis, we have presented the findings of our proposal for irony detection. We explored the role of different facets of affective information on the use of ironic utterances in Twitter. The main contribution of our research lies in exploiting a wide range of aspects covering sentiment and finer-grained emotions for detecting irony.

In Chapter 3, we experimented with a set of textual markers enriched with features based on the sentiment expressed in a tweet. Our results outperformed those from the state of the art. According to this, and taking into consideration previous studies on the importance of affect and emotions in irony, we decided to investigate further the role of affect-related information for identifying irony.

In Chapter 4, we proposed emotIDM, an irony detection model that exploits mainly an extensive set of features encompassing several facets of affect distributed in sentiment and emotions. Unlike in most of the related work, we decided to evaluate the performance of our approach over a set of Twitter corpora composed by: TwReyes2013, TwRiloff2013, TwBarbieri2014, TwMohammad2015, and TwPtáček2014[1]; these corpora have different propierties covering from collection mode, size, and imbalance degree. Most of these corpora were used previously to evaluate other irony detection models. By exploiting emotIDM, we obtained encouraging results that allow us to validate the usefulness of affect-related features for detecting ironic utterances in Twitter. The experiments carried out with emotIDM were dedicated only to identify the presence of irony. As in the majority of the research in irony detection, our approach does not distinguish among different kinds of irony.

Research on the differences between irony and sarcasm has rarely been addressed in computational linguistics. There is a set of hashtags that have been widely exploited to collect ironic instances from Twitter: #irony, #sarcasm, and #not. Aiming to determine whether there are differences in the use of these hashtags from a figurative language perspective, in Chapter 5 we performed an analysis of the affective content over a benchmark dataset developed in Task 11 ("Sentiment Analysis of Figurative Language in Twitter") at SemEval-2015, that is particularly rich in figurative language devices. As a starting point, we considered the same group of features and resources used in emotIDM. A set of tweets labeled with #irony, #sarcasm, and #not was analyzed attempting to identify a possible separation in the use of these hashtags to highlight different figurative meanings as well as the role of polarity reversal among such kinds of tweets.

The analysis carried out provides interesting insights. First, we identify a set of promising features for discriminating between tweets labeled with #irony and #sarcasm. We carried out a set of binary classification tasks outperforming the state-of-the-art results. Besides, we also experimented by classifying between tweets labeled with #not and #irony, and also with #not versus #sarcasm. The obtained results could be considered as a baseline for future research on this particular topic. From the classification outcomes, it is possible to validate the idea concerning these hashtags that are, in fact, used to express different kinds of figurative language devices. We identify that tweets containing hashtags #sarcasm and #not often contain more positive terms. However they tend to be more negative, the

[1]The TwPtáček2014 dataset was used to analyze emotional content from an additional perspective. In Appendix C an emotional-graph representation of its sarcastic tweets is illustrated.

literal meaning in these tweets is used for expressing the opposite intention. The #irony hashtag is commonly used for denoting some ironical situation. Furthermore, we investigated an attractive property of ironic utterances: polarity reversal. According to the literature, due to the polarity reversal effect of ironic and sarcastic expressions, the performance of sentiment analysis systems could be undermined [25]. We carried out an analysis of the role of polarity reversal over the data mentioned above. Our results show that the effect of reversing polarity is not the same among these kinds of figurative tweets.

The polarity shifting in ironic sentences represents one of the biggest challenges for sentiment analysis. In Chapter 6, we described an approach for sentiment analysis where the presence of irony was considered before to determine the polarity degree of a given text. For doing so, we propose a pipeline that incorporates two phases. First, by exploiting emotIDM, we determine whether a tweet is ironic or not. We trained an irony detection model with a set of tweets labeled with #irony and #sarcasm as the "ironic" class. Then, an ensemble of classifiers was used to identify the presence of irony. In the second phase, we performed the sentiment analysis by taking advantage of the model proposed in [73]. This model was enhanced by adding features related to affect. We took advantage of an extensive set of sentiment resources in English. We evaluated our proposal on the dataset of Task 11 at SemEval 2015. The obtained results are not higher than the one from the best-ranked system, but they are still competitive among the best-ranked ones. Our method covers two main objectives. First, we incorporated our irony detection model in a pipeline for sentiment analysis. Second, the approach we propose allows evaluating a model for sentiment analysis that relies mainly on affective-related features. Furthermore, we outperformed our results on the shared task at SemEval-2015. This allowed us to validate the importance of considering ironic and multifaceted-affective information for sentiment analysis.

In the next section of this chapter, we present additional experiments carried out with the objective of having a broader perspective of the emotIDM model. This chapter is structured as follows. In Section 7.2, we summarize and discuss the results presented in Chapter 4. An ablation experiment on the features included in emotIDM is presented in Section 7.3. In Section 7.4, we describe a set of experiments carried out when emotIDM is evaluated considering different imbalanced scenarios. Section 7.5 presents a qualitative analysis on the affective resources included in emotIDM. Finally, we draw some conclusions in Section 7.6.

7.2 emotIDM: A model for Detecting Irony and Sarcasm in Twitter

In Chapter 4 we described emotIDM, a novel irony detection model that exploits information on different facets of affect. emotIDM covers one of the main objectives of this thesis: *building an irony detection model taking advantage of affective information.* We collected a broad range of resources developed for diverse tasks dedicated to studying affect phenomena in texts. These resources were separated into three groups[2]:

(i) Those defining some sort of polarity degree in a set of words (as in Chapter 6 *Sentiment*).

(ii) Those containing a set of words distributed among a given *categorical* model of emotions (as in Chapter 6 *eCat*).

(iii) Those where a set of words have associated a degree in a certain aspect of the *dimensional* model of emotions (as in Chapter 6 *eDim*).

[2]Besides, we added a set of features including some lexical markers such as punctuation marks, emoticons, among others.

Being a complex communication mode, identifying irony deserves to be addressed considering features capturing the ironic style even from non-salient factors such as sentiment and emotions. One of the main advantages of our model lies in its capacity to catch ironic intention exploiting high-level features and disregarding the domain. From a computational linguistics perspective, irony and sarcasm are usually considered as synonyms. The most salient difference between the data used for experimental purposes among the related work is the use of specific labels to retrieve ironic utterances. The model we propose does not make the distinction between irony and sarcasm: the former is considered as an umbrella term covering the latter. Our main aim with emotIDM is to identify ironic sentences; differentiate between these figurative languages devices was addressed as a further step.

The performance of emotIDM was assessed considering a distinctive experimental setting: instead of collecting our own dataset, we took advantage of benchmark corpora used in the state of the art. A total of five different corpora were considered for experimental purposes. Three of these corpora were retrieved by taking advantage of specific Twitter labels: #irony and #sarcasm. This collection mode has been widely exploited under the assumption that users self-annotate their ironic or sarcastic intention reflecting the real use of irony on Twitter. A less explored method for collecting ironic data is crowd-sourcing. Asking people to determine whether a given text is ironic or not is per se a big challenge. We also experimented with two corpora collected by crowd-sourcing. Having a broad set of data allowed us to evaluate and to compare the performance of our model across several aspects concerning to the corpora. Attributes such as size, collection methodology, and differences on class imbalance were covered in our experiments.

Overall, the obtained results served to validate the robustness of the model we propose. We experimented with three classifiers widely applied in text classification tasks. Decision tree emerges as the classifier with the best performance. By using it, the majority of our results outperformed those of the state of the art. One significant benefit of our model concerns to the dimensional feature space. We obtained competitive performance by exploiting a vector representation for each tweet composed by 78 features. It seems that our model is able to capture relevant information for discriminating between ironic and non-ironic tweets.

emotIDM shows better performance on balanced datasets: TwReyes2013 and TwBarbieri2014. Experiments on these corpora were performed by considering different binary classifications between the ironic and non-ironic classes. The non-ironic classes were collected by exploiting specific hashtags. The obtained results by applying emotIDM outperformed those of the state of the art in TwReyes2013 corpus. For what concerns to TwBarbieri2014, there is only one experiment where our model did not achieve the best result. A set of tweets containing the #newspaper hashtag comprises one of the non-ironic classes in TwBarbieri2014. The authors in [18] pointed out that their results on this experiment were high due tweets in the #newspaper class are written in a formal language. Besides, the majority of them contained an Internet link and emotIDM does not consider any feature to capture the presence of links. Regarding the rest of the binary classifications of these corpora, seems that our model was able to discriminate ironic tweets from others coming from different domains.Considering the fact that our model has a similar performance among the different binary classifications, we could say that disregarding the domain of non-ironic class, emotIDM identifies ironic tweets quite well.

We carried out an analysis of the relevance of the features in emotIDM. Results confirm the importance of lexical markers for identifying irony pointed out by [86]. For what concerns to features related to sentiment, it seems that they play a crucial role to characterize the subjective property of ironic utterances. Features arising from both categorical and dimensional models of emotions seem to provide a considerable amount of information in

emotIDM. The features related to emotions in our model allow us to cover different aspects to detect both ironic and sarcastic tweets. The dimensions defined in DAL and ANEW are very relevant for detecting ironic tweets. While for sarcastic ones, the category "'Love" in SentiSense emerges as one of the most relevant features. A vast amount of tweets expressing a sarcastic intention consists in the use of the verb "love" to point out a big disagreement with a certain situation, e.g. "I *love* when ... ", "I just *love* ...".

Furthermore, we analyzed the relevance of the features in emotIDM by differentiating when the task was recognizing tweets labeled with #irony and #sarcasm. We observed some interesting findings on this subject. Three features in the *Structural*[3] group are especially relevant in tweets labeled with #sarcasm: (i) the length in characters of sarcastic tweets tends to be shorter than non-sarcastic and even ironic ones, that it could be considered as a signal of the intention of the speaker to be sharper; and, (ii) the frequency of nouns and Twitter mentions: both features emerge among the most discriminating ones; this could be related to the particular characteristic of sarcastic utterances: the presence of a specific target to attack [90]. The higher amount of words referring to people, animals, places, and so on, could be an attractive cue for determining the target behind a sarcastic intention in a given utterance. We found data-driven arguments to validate the intuitions about the critical role of the overall sentiment expressed in an ironic sentence to differentiate among different kinds of irony [5].

Very recently fashionable deep learning techniques have been also applied for irony and sarcasm detection. Nozza et al. [108], Ghosh and Veale [52] and Poria et al. [115] evaluated novel techniques such as word-embeddings and convolutional neural networks that have demonstrated be useful for several natural language processing tasks. Ghosh and Veale in [52] investigated the use of a semantic model that uses deep neural networks to recognize irony. A convolutional neural network model for detecting irony was proposed by Poria et al. [115]; it uses features belonging to three groups of affective phenomena: sentiment, emotion, and for the first time in irony detection personality-related features. Ghosh et al. [53] and Nozza et al. [108] experimented with word-embeddings to identify irony. Joshi et al. [80] and Van Hee et al. [70] explored the use of word-embeddings together with text-based features such as n-grams and punctuation marks, among others.

Some of the benchmark corpora of this thesis have been used to evaluate the deep learning models for detecting ironic content in Twitter. Nozza et al. [108] experimented with the TwReyes2013 dataset. They also evaluated binary classification between tweets labeled with #irony versus those tagged with #education, #humor, and #politics obtaining 0.84, 0.83, and 0.88 in F-measure terms. The dataset collected by Riloff et al. [127] was used by Ghosh and Veale [52]; the model achieved an F-measure of 0.88. Poria et al. [115] evaluated the performance of their approach in the TwPtáček2014 dataset obtaining 0.89 in F-measure. Deep learning techniques have been used for detecting irony on tweets containing both #irony and #sarcasm hashtags. At the present time, the best results have been achieved on corpora where the #sarcasm hashtag was used for collecting data. Our results obtained on the TwReyes2013 dataset are still the state of the art. For what concerns to the TwRiloff2013 and TwPtáček2014 corpora, although our results with emotIDM over these datasets are the highest (0.75 vs. 0.88 and 0.82 vs 0.89, respectively), they are still competitive against powerful techniques, demonstrating the importance of affect-related information.

[3]The *Structural* group of features in emotIDM comprises lexical markers (such as punctuation marks and emoticons), frequency of POS tags, uppercase characters, length in words and characters, Twitter markers (such as hashtags, mentions and retweet), among others.

7.3 emotIDM: An Ablation Test on Affect-related Features

emotIDM is composed of a wide range of affect-related features covering various facets of affective information. All these features together have demonstrated to be useful for identifying ironic (or sarcastic) tweets from non-ironic (or non-sarcastic) ones. One of the main objectives of this thesis consists in validating the role of affect-related information for detecting irony in Twitter. With the aim to investigate the performance of the different features and resources we employed, we carried out further experiments.

As in the rest of the experiments described in this thesis, we conducted binary classification for detecting ironic tweets by exploiting the features in emotIDM under several conditions. We performed a set of experiments comprising different combination among the groups of features: *Structural*, *Sentiment*, *eCat*, and *eDim* (see 4.4.2 for a detailed description of the features.). With respect to the classifiers, as in the Chapter 4, we used the Weka version of three classification algorithms: Naïve Bayes, Decision Tree and Support Vector Machine.

We experimented with various subsets of the features of emotIDM. In total ten different combinations of features were evaluated:

(i) *Each set individually.* It involves the use of each group of the affective related features. This experiment means that in each experiment we attempted to identify irony using only information related to sentiment or emotions.

(ii) *Combination of affective-related groups of features.* The second experiment was proposed with the aim to investigate the performance of combining two groups of features related to affective information. Three different combinations were evaluated: *Sentiment + eCat*, *Sentiment + eDim*, and *eCat + eDim*.

(iii) *Structural + each group of affective-related features.* According to the Information Gain analysis carried out in [71] (see Section 4.5.2), the features belonging to the *Structural* group emerged among the most relevant ones in emotIDM. Given that we are more interested in evaluating the role of affect in irony detection, we decided to carry out an experiment consisting in combining the *Structural* group of features together with *Sentiment*, *eCat*, and *eDim*.

(iv) *Structural + each one of the resources in Sentiment.* The main objective of this experimental setting is to investigate which of the sentiment resources performs better for identifying irony in Twitter when it is combined with structural features.

(v) *Structural + each one of the resources in eCat.* We also experimented combining *Structural* features with each of the resources in *eCat*. Each resource in this group of features was developed taking into account different theories defining emotions. With these experiments, we are attempting to investigate if there is some emotional approach that represents better the use of irony on Twitter.

(vi) *Structural + each one of the resources in eDim.* We experimented with the resources in the *eDim* group of features together with those in the *Structural* one. According to the Information Gain results in Chapter 4, the majority of the features of *eDim* emerged among the most relevant in emotIDM.

(vii) *Sentiment + each one of the resources in eCat.* In most of the experiments involving features in the *Sentiment* group, they seem to be relevant in irony detection. We decided to investigate this also by combining *Sentiment* features with each one of the resources in *eCat*. This could be useful for obtaining insights on the importance of

using different theoretical perspectives in emotional categories together with the overall sentiment expressed in a given tweet.

(viii) *Sentiment + each one of the resources in eDim.* We also evaluate the contribution of each one of the resources in *eDim* when they are combined with the *Sentiment* group of features.

(ix) *Emotions + each one of the resources in Sentiment.* Finally in a similar manner, we performed a set of experiments that concatenate the features of *eCat* and *eDim* into a single group denoted as *Emotions*. Then the *Emotions* group is combined with each of the resources in *Sentiment*. The main objective of this experiment is to evaluate the role of emotional features when the overall sentiment of a given tweet is calculated under different perspectives.

(x) *Ablation in each group of affect-related features.* Each group of the affect-related groups of features is composed of a set of resources describing different facets of affective information; we are interested in investigating the importance of each affective resource. We performed a set of ablation experiments where from each group of features a resource was removed each time.

Next we present the obtained results of applying the aforementioned configuration on the corpora where emotIDM shows the best performance: TwReyes2013, TwIronyBarbieri2014, and TwPtáček2014. For each dataset, a table containing the results of the first eight configurations is presented. Regarding the crowdsourcing datasets, the ablation results are presented in Appendix A; the outcomes in both TwMohammad2015 and TwRiloff2013 considering different feature combinations are very similar. This could be due to the small amount of data available for learning a model with a reduced set of features. Therefore, it is not possible to identify significative differences on the obtained results. The results concerning to experiment (x) for all the corpora are shown in Appendix B; a summary of the obtained results is presented in at the end of this subsection.

Ablation experiments in TwReyes2013

Table 7.1 shows the obtained results in F-measure terms when different features combination are used in TwReyes2013. According to the results obtained, none of the features combinations in the *Irony-vs-Politics* outperforms the result using all the features of emotIDM (0.92 in F-measure terms). On the other hand, there are some experiments where exploiting *Structural* together with *Sentiment* features it is possible to achieve the same F-measure than using emotIDM (in bold). Comparing the results when only sentiment or emotional related features are used against those in [126] allows us to validate the usefulness of this kind of information. It seems that by taking advantage only of affective-related features is possible to achieve similar results. In other words, our reduced vector space representation can capture relevant information for discriminating ironic tweets.

The best performance on the "*Sentiment + each one of the resources in eCat*" is achieved when Sentiment features are combined with LIWC; whereas in the case of eDim the highest result occurs when Sentiment is used together with DAL. AFINN used with all the features belonging to the Emotions group obtains the best results.

TABLE 7.1. Results of different combination of features on TwReyes2013. The underlined values are not statistically significant (t-test with 95% of confidence value)

	TwReyes2013								
	Irony-vs-Education			*Irony-vs-Humor*			*Irony-vs-Politics*		
	NB	DT	SVM	NB	DT	SVM	NB	DT	SVM
Each set individually									
Sentiment	0.72	0.74	0.72	0.68	0.76	0.72	0.65	0.73	0.70
eCat	0.59	0.66	0.66	0.63	0.70	0.67	0.62	0.71	0.67
eDim	0.63	0.71	0.67	0.62	0.70	0.64	0.63	0.72	0.69
Combination of affective-related groups of features									
Sentiment + eCat	0.69	0.76	0.77	0.69	0.78	0.76	0.66	0.79	0.77
Sentiment + eDim	0.71	0.75	0.75	0.69	0.76	0.73	0.66	0.76	0.75
eCat + eDim	0.70	0.76	0.77	0.69	0.78	0.76	0.66	0.79	0.77
Structural + each group of affective-related features									
Structural + Sentiment	0.79	**0.90**	0.88	0.78	**0.90**	0.87	0.78	0.90	0.88
Structural + eCat	0.77	0.88	0.86	0.76	0.88	0.87	0.83	0.91	0.88
Structural + eDim	0.78	0.88	0.86	0.76	0.88	0.87	0.80	0.90	0.87
Structural + each one of the resources in Sentiment									
Structural + AFINN	0.83	0.89	0.88	0.79	**0.90**	0.88	0.83	0.90	0.87
Structural + HL	0.83	0.89	0.88	0.77	0.89	0.87	0.84	0.90	0.87
Structural + GI	0.84	0.89	0.88	0.79	**0.90**	0.88	0.84	0.90	0.87
Structural + EmoLex	0.82	0.88	0.85	0.77	0.88	0.86	0.83	0.89	0.86
Structural + SWN	0.81	0.88	0.85	0.77	0.88	0.86	0.82	0.89	0.86
Structural + SN	0.82	0.88	0.85	0.77	0.88	0.86	0.83	0.89	0.86
Structural + SO	0.82	0.89	0.85	0.77	0.88	0.86	0.83	0.89	0.86
Structural + EWN	0.81	0.88	0.85	0.77	0.88	0.86	0.82	0.89	0.86
Structural + SUBJ	0.83	0.89	0.88	0.78	0.89	0.87	0.83	0.90	0.87
Structural + each one of the resources in eCat									
Structural + LIWC	0.78	0.88	0.85	0.76	0.88	0.87	0.83	0.91	0.88
Structural + EmoLex	0.80	0.88	0.85	0.76	0.88	0.86	0.83	0.90	0.87
Structural + EmoSN	0.78	0.88	0.85	0.76	0.89	0.87	0.83	0.90	0.87
Structural + SentiSense	0.77	0.88	0.85	0.76	0.89	0.87	0.83	0.91	0.88
Structural + each one of the resources in eDim									
Structural + DAL	0.81	0.88	0.85	0.76	0.88	0.86	0.82	0.90	0.87
Structural + SN	0.77	0.88	0.85	0.76	0.88	0.86	0.80	0.89	0.87
Structural + ANEW	0.80	0.88	0.85	0.76	0.88	0.87	0.81	0.89	0.87
Sentiment + each one of the resources in eCat									
Sentiment + LIWC	0.70	0.76	0.76	0.69	0.78	0.76	0.67	0.79	0.76
Sentiment + EmoLex	0.71	0.75	0.73	0.68	0.77	0.73	0.66	0.76	0.72
Sentiment + EmoSN	0.72	0.75	0.73	0.69	0.77	0.73	0.66	0.77	0.72
Sentiment + SentiSense	0.72	0.75	0.73	0.69	0.76	0.72	0.65	0.74	0.70
Sentiment + each one of the resources in eDim									
Sentiment + DAL	0.73	0.75	0.74	0.69	0.76	0.73	0.66	0.76	0.74
Sentiment + SN	0.72	0.74	0.73	0.68	0.76	0.72	0.65	0.74	0.70
Sentiment + ANEW	0.72	0.74	0.72	0.68	0.76	0.72	0.64	0.74	0.72
Emotions + each one of the resources in Sentiment									
Emotions + AFINN	0.67	0.75	0.75	0.69	0.77	0.74	0.66	0.78	0.75
Emotions + HL	0.66	0.75	0.74	0.65	0.74	0.72	0.65	0.78	0.74
Emotions + GI	0.66	0.74	0.74	0.68	0.76	0.73	0.66	0.77	0.75
Emotions + EmoLex	0.64	0.72	0.72	0.64	0.72	0.70	0.64	0.76	0.72
...									

	NB	DT	SVM	NB	DT	SVM	NB	DT	SVM
Emotions + SWN	0.64	0.72	0.71	0.63	0.73	0.70	0.64	0.76	0.72
Emotions + SN	0.64	0.71	0.71	0.63	0.72	0.70	0.63	0.75	0.70
Emotions + SO	0.64	0.72	0.71	0.64	0.73	0.70	0.64	0.76	0.71
Emotions + EWN	0.64	0.71	0.71	0.64	0.73	0.71	0.64	0.76	0.72
Emotions + SUBJ	0.66	0.75	0.74	0.66	0.76	0.73	0.65	0.78	0.75

Ablation experiments in TwIronyBarbieri2014

Tables 7.2 and 7.3 show the obtained results in F-measure terms when different features combination are used on TwIronyBarbieri2014. The Sentiment group has the best performance among all the affective features. When features from emotional categories are combined with sentiment ones, the results are better than using other combination among affective features. In the case of *Irony-vs-Humor* and *Irony-vs-Politics*, there are feature combinations involving sentiment and *eCat* features improving the results obtained by emotIDM (in bold). Once more, using Sentiment + LIWC allows to obtain the best performance on the "*Sentiment + each one of the resources in eCat*". It is even more evident in the *Irony-vs-Politics* classification.

Table 7.2. Results of different combination of features on TwIronyBarbieri2014. The underlined values are not statistically significant (t-test with 95% of confidence value)

TwIronyBarbieri2014						
	Irony-vs-Education			*Irony-vs-Humor*		
	NB	DT	SVM	NB	DT	SVM
Each set individually						
Sentiment	0.61	0.66	0.64	0.64	0.72	0.67
eCat	0.56	0.64	0.63	0.63	0.70	0.65
eDim	0.59	0.64	0.62	0.63	0.68	0.66
Combination of affective-related groups of features						
Sentiment + eCat	0.62	0.69	0.71	0.62	0.67	0.68
Sentiment + eDim	0.62	0.67	0.68	0.65	0.73	0.70
eCat + eDim	0.62	0.67	0.68	0.65	0.73	0.70
Structural + each group of affective-related features						
Structural + Sentiment	0.78	0.90	0.83	0.79	0.91	0.85
Structural + eCat	0.79	0.89	0.83	0.79	0.91	0.85
Structural + eDim	0.80	0.89	0.83	0.80	0.91	0.84
Structural + each one of the resources in Sentiment						
Structural + AFINN	0.81	0.90	0.82	0.80	0.91	0.85
Structural + HL	0.81	0.90	0.83	0.79	0.91	0.85
Structural + GI	0.81	0.89	0.82	0.79	0.92	0.85
Structural + EmoLex	0.81	0.90	0.82	0.80	0.91	0.85
Structural + SWN	0.81	0.90	0.83	0.79	0.91	0.85
Structural + SN	0.81	0.89	0.82	0.79	0.92	0.85
Structural + SO	0.81	0.90	0.82	0.79	0.91	0.85
Structural + EWN	0.81	0.89	0.82	0.79	0.91	0.85
Structural + SUBJ	0.81	0.89	0.82	0.79	0.91	0.84
...						

	NB	DT	SVM	NB	DT	SVM
Structural + each one of the resources in eCat						
Structural + LIWC	0.79	0.89	0.83	0.79	0.92	0.85
Structural + EmoLex	0.81	0.89	0.83	0.79	0.91	0.84
Structural + EmoSN	0.79	0.89	0.83	0.79	0.92	0.85
Structural + SentiS	0.79	0.89	0.82	0.79	0.92	0.85
Structural + each one of the resources in eDim						
Structural + DAL	0.81	0.89	0.82	0.79	0.91	0.85
Structural + SN	0.80	0.89	0.82	0.79	0.91	0.85
Structural + ANEW	0.81	0.89	0.82	0.80	0.91	0.84
Sentiment + each one of the resources in eCat						
Sentiment + LIWC	0.61	0.70	0.71	0.65	0.79	0.73
Sentiment + EmoLex	0.60	0.67	0.65	0.64	0.73	0.69
Sentiment + EmoSN	0.62	0.67	0.66	0.65	0.73	0.69
Sentiment + SentiS	0.62	0.68	0.66	0.64	0.72	0.68
Sentiment + each one of the resources in eDim						
Sentiment + DAL	0.63	0.67	0.68	0.65	0.72	0.69
Sentiment + SN	0.61	0.67	0.65	0.64	0.71	0.67
Sentiment + ANEW	0.61	0.67	0.64	0.64	0.72	0.68
Emotions + each one of the resources in Sentiment						
Emotions + AFINN	0.61	0.68	0.68	0.66	0.74	0.71
Emotions + HL	0.61	0.67	0.68	0.64	0.73	0.71
Emotions + GI	0.61	0.67	0.67	0.64	0.73	0.71
Emotions + EmoLex	0.61	0.68	0.69	0.64	0.73	0.71
Emotions + SWN	0.61	0.68	0.68	0.64	0.72	0.71
Emotions + SN	0.60	0.67	0.67	0.63	0.72	0.71
Emotions + SO	0.61	0.67	0.67	0.64	0.73	0.71
Emotions + EWN	0.60	0.67	0.68	0.64	0.72	0.71
Emotions + SUBJ	0.61	0.68	0.67	0.64	0.73	0.71

TABLE 7.3. Results of different combination of features on TwIronyBarbieri2014. The underlined values are not statistically significant (t-test with 95% of confidence value)

TwIronyBarbieri2014						
	Irony-vs-Politics			*Irony-vs-Newspaper*		
	NB	DT	SVM	NB	DT	SVM
Each set individually						
Sentiment	0.56	0.66	0.60	0.60	0.67	0.65
eCat	0.59	0.68	0.64	0.57	0.62	0.61
eDim	0.59	0.68	0.65	0.59	0.64	0.63
Combination of affective-related groups of features						
Sentiment + eCat	0.60	0.74	0.73	0.61	0.69	0.72
Sentiment + eDim	0.59	0.70	0.70	0.60	0.69	0.70
eCat + eDim	0.59	0.70	0.70	0.61	0.69	0.72
...						

	NB	DT	SVM	NB	DT	SVM
Structural + each group of affective-related features						
Structural + Sentiment	0.81	0.91	0.85	0.86	0.92	0.91
Structural + eCat	0.80	0.92	0.86	0.89	0.92	0.91
Structural + eDim	0.82	0.91	0.85	0.87	0.92	0.91
Structural + each one of the resources in Sentiment						
Structural + AFINN	0.80	0.91	0.84	0.90	0.92	0.90
Structural + HL	0.80	0.91	0.84	0.90	0.92	0.90
Structural + GI	0.80	0.91	0.84	0.90	0.92	0.90
Structural + EmoLex	0.80	0.91	0.84	0.90	0.92	0.90
Structural + SWN	0.80	0.91	0.84	0.90	0.92	0.90
Structural + SN	0.80	0.91	0.84	0.90	0.92	0.90
Structural + SO	0.80	0.91	0.84	0.90	0.92	0.90
Structural + EWN	0.80	0.91	0.84	0.89	0.92	0.90
Structural + SUBJ	0.80	0.91	0.84	0.90	0.92	0.90
Structural + each one of the resources in eCat						
Structural + LIWC	0.81	0.92	0.86	0.89	0.92	0.90
Structural + EmoLex	0.81	0.92	0.86	0.89	0.92	0.90
Structural + EmoSN	0.80	0.92	0.86	0.89	0.92	0.91
Structural + SentiS	0.81	0.92	0.86	0.89	0.92	0.91
Structural + each one of the resources in eDim						
Structural + DAL	0.80	0.91	0.85	0.89	0.92	0.90
Structural + SN	0.82	0.92	0.85	0.87	0.92	0.91
Structural + ANEW	0.82	0.91	0.85	0.88	0.92	0.90
Sentiment + each one of the resources in eCat						
Sentiment + LIWC	0.60	0.75	0.73	0.61	0.69	0.71
Sentiment + EmoLex	0.58	0.70	0.65	0.60	0.67	0.66
Sentiment + EmoSN	0.57	0.70	0.65	0.60	0.67	0.66
Sentiment + SentiS	0.57	0.67	0.61	0.60	0.67	0.66
Sentiment + each one of the resources in eDim						
Sentiment + DAL	0.60	0.70	0.69	0.61	0.69	0.70
Sentiment + SN	0.56	0.67	0.61	0.60	0.67	0.66
Sentiment + ANEW	0.56	0.68	0.63	0.60	0.67	0.67
Emotions + each one of the resources in Sentiment						
Emotions + AFINN	0.60	0.73	0.69	0.60	0.66	0.66
Emotions + HL	0.61	0.73	0.69	0.60	0.65	0.66
Emotions + GI	0.61	0.72	0.69	0.60	0.65	0.66
Emotions + EmoLex	0.60	0.73	0.69	0.60	0.65	0.67
Emotions + SWN	0.60	0.73	0.70	0.61	0.67	0.69
Emotions + SN	0.60	0.73	0.67	0.60	0.65	0.65
Emotions + SO	0.60	0.73	0.69	0.60	0.65	0.66
Emotions + EWN	0.60	0.73	0.69	0.60	0.65	0.66
Emotions + SUBJ	0.61	0.73	0.69	0.60	0.66	0.67

Ablation experiments in TwSarcasmBarbieri2014

Tables 7.4 and 7.5 show the obtained results in F-measure terms when different features combination are used on TwSarcasmBarbieri2014. Overall, the results of the experiments carried out on TwSarcasmBarbieri2014 are higher than those on TwIronyBarbieri2014. *Sentiment* obtains the best performance among the affective groups of features. The result of combining *Structural* with *eCat* features is the only one outperforming emotIDM in *Sarcasm-vs-Humour* (0.92 in F-measure). When *Sentiment* is combined with each of the resources in *eDim*, the best performance is obtained by DAL. Combining SentiWordNet with emotional features allows for the best result among all the resources in *Sentiment*.

Table 7.4. Results of different combination of features on TwSarcasmBarbieri2014. The underlined values are not statistically significant (t-test with 95% of confidence value)

TwSarcasmBarbieri2014						
	Sarcasm-vs-Education			*Sarcasm-vs-Humour*		
	NB	DT	SVM	NB	DT	SVM
Each set individually						
Sentiment	0.63	0.72	0.71	0.61	0.73	0.68
eCat	0.61	0.67	0.66	0.60	0.70	0.63
eDim	0.53	0.64	0.61	0.57	0.67	0.64
Combination of affective-related groups of features						
Sentiment + eCat	0.63	0.72	0.76	0.63	0.75	0.75
Sentiment + eDim	0.62	0.71	0.73	0.62	0.74	0.73
eCat + eDim	0.63	0.72	0.76	0.63	0.75	0.75
Structural + each group of affective-related features						
Structural + Sentiment	0.85	0.91	0.88	0.82	0.92	0.87
Structural + eCat	0.86	0.94	0.90	0.92	0.95	0.95
Structural + eDim	0.86	0.90	0.87	0.82	0.92	0.86
Structural + each one of the resources in Sentiment						
Structural + AFINN	0.86	0.91	0.87	0.82	0.92	0.85
Structural + HL	0.86	0.91	0.87	0.82	0.92	0.86
Structural + GI	0.86	0.91	0.86	0.82	0.92	0.85
Structural + EmoLex	0.86	0.91	0.86	0.82	0.92	0.85
Structural + SWN	0.86	0.91	0.87	0.83	0.92	0.86
Structural + SN	0.85	0.90	0.86	0.83	0.92	0.85
Structural + SO	0.86	0.91	0.86	0.83	0.92	0.86
Structural + EWN	0.86	0.91	0.86	0.82	0.92	0.85
Structural + SUBJ	0.86	0.91	0.87	0.83	0.92	0.85
Structural + each one of the resources in eCat						
Structural + LIWC	0.85	0.90	0.87	0.82	0.92	0.87
Structural + EmoLex	0.86	0.91	0.88	0.83	0.92	0.86
Structural + EmoSN	0.86	0.91	0.88	0.82	0.92	0.87
Structural + SentiS	0.86	0.91	0.87	0.82	0.92	0.86
Structural + each one of the resources in eDim						
Structural + DAL	0.86	0.91	0.86	0.82	0.92	0.86
...						

	NB	DT	SVM	NB	DT	SVM
Structural + SN	0.86	0.90	0.86	0.82	0.92	0.86
Structural + ANEW	0.86	0.90	0.86	0.83	0.92	0.86
Sentiment + each one of the resources in eCat						
Sentiment + LIWC	0.63	0.72	0.74	0.64	0.75	0.74
Sentiment + EmoLex	0.62	0.71	0.71	0.62	0.75	0.69
Sentiment + EmoSN	0.63	0.71	0.71	0.63	0.73	0.70
Sentiment + SentiS	0.64	0.72	0.71	0.63	0.74	0.69
Sentiment + each one of the resources in eDim						
Sentiment + DAL	0.63	0.71	0.72	0.62	0.73	0.71
Sentiment + SN	0.63	0.71	0.71	0.61	0.73	0.69
Sentiment + ANEW	0.62	0.72	0.71	0.61	0.73	0.69
Emotions + each one of the resources in Sentiment						
Emot + AFINN	0.62	0.70	0.70	0.60	0.75	0.70
Emot + HL	0.61	0.69	0.70	0.62	0.74	0.71
Emot + GI	0.61	0.69	0.69	0.60	0.74	0.71
Emot + EmoLex	0.61	0.71	0.71	0.61	0.74	0.71
Emot + SWN	0.62	0.71	0.71	0.62	0.74	0.72
Emot + SN	0.62	0.69	0.70	0.61	0.74	0.71
Emot + SO	0.61	0.69	0.70	0.61	0.74	0.71
Emot + EWN	0.61	0.70	0.70	0.60	0.73	0.70
Emot + SUBJ	0.62	0.69	0.71	0.60	0.74	0.71

TABLE 7.5. Results of different combination of features on TwSarcasmBarbieri2014. The underlined values are not statistically significant (t-test with 95% of confidence value)

TwSarcasmBarbieri2014						
	Sarcasm-vs-Politics			*Sarcasm-vs-Newspaper*		
	NB	DT	SVM	NB	DT	SVM
Each set individually						
Sentiment	0.66	0.74	0.71	0.70	0.77	0.76
eCat	0.67	0.72	0.68	0.66	0.70	0.68
eDim	0.56	0.69	0.67	0.58	0.70	0.69
Combination of affective-related groups of features						
Sentiment + eCat	0.67	0.78	0.79	0.70	0.77	0.81
Sentiment + eDim	0.64	0.75	0.77	0.69	0.77	0.79
eCat + eDim	0.67	0.77	0.80	0.70	0.77	0.80
Structural + each group of affective-related features						
Structural + Sentiment	0.87	0.93	0.90	0.91	0.96	0.95
Structural + eCat	0.86	0.94	0.90	0.92	0.95	0.95
Structural + eDim	0.87	0.93	0.90	0.93	0.96	0.95
Structural + each one of the resources in Sentiment						
Structural + AFINN	0.87	0.93	0.90	0.93	0.96	0.95
Structural + HL	0.86	0.93	0.90	0.93	0.95	0.95
Structural + GI	0.86	0.93	0.88	0.93	0.96	0.95
...						

	NB	DT	SVM	NB	DT	SVM
Structural + EmoLex	0.86	0.93	0.88	0.93	0.96	0.95
Structural + SWN	0.86	0.93	0.90	0.93	0.96	0.95
Structural + SN	0.86	0.93	0.88	0.93	0.96	0.95
Structural + SO	0.86	0.93	0.89	0.93	0.96	0.95
Structural + EWN	0.86	0.93	0.88	0.93	0.96	0.95
Structural + SUBJ	0.86	0.93	0.89	0.93	0.96	0.95
Structural + each one of the resources in eCat						
Structural + LIWC	0.86	0.94	0.90	0.93	0.96	0.95
Structural + EmoLex	0.87	0.94	0.90	0.93	0.96	0.95
Structural + EmoSN	0.86	0.94	0.90	0.93	0.96	0.95
Structural + SentiS	0.86	0.94	0.90	0.93	0.96	0.95
Structural + each one of the resources in eDim						
Structural + DAL	0.86	0.93	0.89	0.93	0.95	0.95
Structural + SN	0.86	0.93	0.89	0.93	0.96	0.95
Structural + ANEW	0.86	0.93	0.89	0.93	0.95	0.95
Sentiment + each one of the resources in eCat						
Sentiment + LIWC	0.68	0.78	0.78	0.7	0.77	0.79
Sentiment + EmoLex	0.67	0.77	0.73	0.70	0.77	0.77
Sentiment + EmoSN	0.67	0.76	0.73	0.70	0.77	0.77
Sentiment + SentiS	0.66	0.74	0.72	0.70	0.77	0.77
Sentiment + each one of the resources in eDim						
Sentiment + DAL	0.65	0.75	0.76	0.70	0.77	0.78
Sentiment + SN	0.65	0.74	0.72	0.70	0.76	0.77
Sentiment + ANEW	0.65	0.74	0.73	0.70	0.77	0.77
Emotions + each one of the resources in Sentiment						
Emot + AFINN	0.66	0.76	0.73	0.67	0.73	0.74
Emot + HL	0.67	0.76	0.73	0.67	0.73	0.74
Emot + GI	0.66	0.75	0.73	0.66	0.73	0.73
Emot + EmoLex	0.66	0.76	0.74	0.66	0.73	0.74
Emot + SWN	0.67	0.76	0.76	0.68	0.75	0.77
Emot + SN	0.66	0.76	0.73	0.66	0.72	0.72
Emot + SO	0.66	0.76	0.73	0.66	0.72	0.73
Emot + EWN	0.65	0.76	0.74	0.66	0.73	0.75
Emot + SUBJ	0.66	0.76	0.74	0.67	0.73	0.75

Ablation experiments in TwPtáček2014

Table 7.6 shows the obtained results in F-measure terms when different features combination are used on TwPtáček2014. The best result on the TwPtáček2014 dataset is the one obtained with all the features in emotIDM. None of the combinations of a subset of features evaluated in the ablation experiments outperfom this result. Among all the results, the highest ones (in bold) are achieved by combining *Structural* features with *eCat*, some resources from *Sentiment*, and those lexica defining a certain categorical model of emotions. When the *Structural* features are not considered for classification purposes, the overall performance of all the experiments drops. It is important to highlight that there are some cases where the

SVM does not have enough information to produce a model for classifying sarcastic tweets; we use a "-" symbol to denote the lack of result for a given binary classification.

Table 7.6. Results of different combination of features on TwPtáček2014. The underlined values are not statistically significant (t-test with 95% of confidence value)

TwPtáček2014			
	NB	DT	SVM
Each set individually			
Sentiment	0.66	0.69	-
eCat	0.66	0.69	-
eDim	0.65	0.62	-
Combination of affective-related groups of features			
Sentiment + eCat	0.66	0.70	-
Sentiment + eDim	0.65	0.70	-
eCat + eDim	0.66	0.70	-
Structural + each group of affective-related features			
Structural + Sentiment	0.71	0.78	0.71
Structural + eCat	0.72	0.79	0.72
Structural + eDim	0.70	0.78	0.70
Structural + each one of the resources in Sentiment			
Structural + AFINN	0.72	0.79	-
Structural + HL	0.72	0.79	0.67
Structural + GI	0.72	0.78	-
Structural + EmoLex	0.72	0.78	-
Structural + SWN	0.72	0.78	0.65
Structural + SN	0.72	0.78	-
Structural + SO	0.72	0.79	-
Structural + EWN	0.72	0.78	-
Structural + SUBJ	0.72	0.78	-
Structural + each one of the resources in eCat			
Structural + LIWC	0.72	0.78	0.68
Structural + EmoLex	0.72	0.79	0.70
Structural + EmoSN	0.72	0.79	0.70
Structural + SentiSense	0.72	0.79	0.66
Structural + each one of the resources in eDim			
Structural + DAL	0.72	0.78	-
Structural + SN	0.71	0.78	0.68
Structural + ANEW	0.71	0.78	0.70
Sentiment + each one of the resources in eCat			
Sentiment + LIWC	0.66	0.70	-
Sentiment + EmoLex	0.66	0.70	-
Sentiment + EmoSN	0.66	0.69	-
Sentiment + SentiSense	0.67	0.70	-
Sentiment + each one of the resources in eDim			
Sentiment + DAL	0.66	0.70	-
...			

	NB	DT	SVM
Sentiment + SN	0.66	0.69	-
Sentiment + ANEW	0.66	0.69	-
Emotions + each one of the resources in Sentiment			
Emotions + AFINN	0.66	0.70	-
Emotions + HL	0.66	0.70	-
Emotions + GI	0.66	0.70	-
Emotions + EmoLex	0.66	0.70	-
Emotions + SWN	0.66	0.70	-
Emotions + SN	0.66	0.70	-
Emotions + SO	0.66	0.70	-
Emotions + EWN	0.66	0.69	-
Emotions + SUBJ	0.66	0.70	-

Ablation in each group of affect-related features

Figure 7.1 shows the average results in F-measure obtained among all the binary classifications. The average was calculated considering the result of each experiment with respect to each set of features.

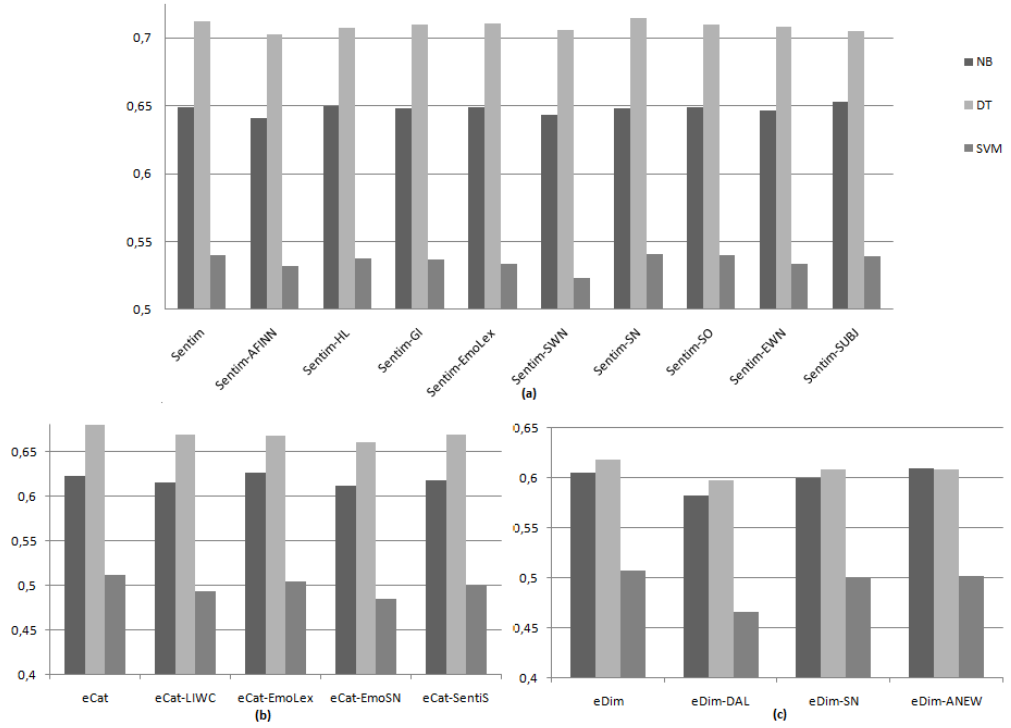


FIGURE 7.1. Ablation experiment of each group of affect-related information

The *Sentiment* group of features achieves the best results demonstrating the importance of considering the subjective value in ironic instances. For what concerns to *eCat* and *eDim*, it seems that the properties they captured are enough for discriminating among ironic and non-ironic utterances.

The highest performance is achieved by the decision tree classifier. The *Sentiment* group is the one that obtains the best results in comparison with *eCat* and *eDim*. When AFINN and SentiWordNet (SWN) are removed from the *Sentiment* group, there is a drop in the performance of all the classifiers. The slight differences on the *eCat* group are not enough for determining which resource(s) provides more information to this set of features. For what concerns to *eDim*, the most significant drop in the achieved results occurs when the Dictionary of Affect in Language (DAL) is not considered as part of this group of features.

Overall comments on the ablation test

Figure 7.1 shows the average results in F-measure obtained among all the binary classifications. The average was calculated considering the result of each experiment with respect to each set of features. In this section we have presented the results on several combinations of the features in emotIDM. Decision tree achieves the best results on all corpora, especially on the TwReyes2013 and TwBarbieri2014. As can be noticed, there is a drop in the performance of the classifiers when *Structural* features are not considered. When combining *Structural* with some of the affect-related groups of features, we achieve competitive results in comparison with the ones in the state of the art. By combining AFINN, HL and GI together with features in *Structural* it is possible to obtain the highest results in most of the binary classification. Among the best performing features are those of the *Sentiment* group. However, there are some cases where features in the *eCat* and *eDim* achieve competitive results. The findings of the ablation experiments point out to the importance of affective information to characterize ironic utterances in Twitter. We observed that using the features in *eDim* with *Structural* ones it is possible to obtain similar results than those showed when *Structural* features are used together with *Sentiment*. It could be considered as an insight on the importance of information related to dimensional models of emotions for detecting irony in Twitter. Overall, the best results were achieved when *Sentiment* features are combined with LIWC. When features from emotions are combined with each of the resources included in *Sentiment*, the best performances are those obtained with AFINN, SUBJ, and SWN.

7.4 emotIDM: Evaluating Imbalanced Scenarios in Twitter

Irony is widely used in social media. However, when compared the total amount of non-ironic comments with the ironic ones the difference is enormous. In other words, a more realistic distribution of the use of irony in social media involves this variance in the ratio between ironic and non-ironic instances. Therefore, an important issue to address is data skew, i.e. the imbalance between ironic vs. non-ironic samples. Despite its importance, only in few research works irony detection as a class imbalanced problem has been carried out.

Joshi et al. [78] have recognized data skew as a critical issue for detecting irony. Liu et al. [95] proposed a multi-strategy ensemble learning approach (MSELA) to deal with imbalanced datasets of ironic comments written in English and Chinese. The MSELA integrates sample-ensemble strategy, classifier-ensemble strategy, and weighted voting strategy. Features such as punctuation marks, n-grams and POS tags were used for English, whereas extreme positive and negative nouns, adjectives, adverbs of degree and proverbs were exploited for Chinese. Results on different settings exploiting MSELA were reported. The best performance was obtained by MSELA across all imbalanced datasets. Abercrombie and Hovy [3] experimented with a corpus of manually annotated Twitter conversations. The authors compare the performance of recognizing irony of both human and machine learning algorithms. Two different settings on the corpus were evaluated: balanced and imbalanced. A logistic regression classifier with n-gram, POS tags, author and audience[4] features was employed. Regarding the results obtained by the classification task, the authors mentioned that the performance of their approach suffers from significant drops on the imbalanced data when compared with those from the balanced one.

Our model performs better in a balanced scenario. To further evaluate emotIDM, we carried out an additional classification experiment on the TwReyes2013 and TwBarbieri2014 datasets. It considers all the tweets in "Education, Humor, Politics" and "Education, Humor, Politics, Newspaper" as a single "non-ironic" class, respectively. The distribution of instances in each dataset is as shown in Table 7.7. This configuration had been already evaluated on the TwReyes2013 dataset by Reyes et al. [126] (as binary classification), Fersini et al. [47] (applying an ensemble of classifiers), and Nozza et al. [108] (taking advantage of word embeddings).

Table 7.7. Imbalance distribution in TwReyes2013 and TwBarbieri2014

Corpus	Ironic	Non-ironic
TwReyes2013	10,000	30,000
TwIronyBarbieri2014	10,000	40,000
TwSarcasmBarbieri2014	10,000	40,000

Table 7.8 shows the obtained results on the imbalanced version of TwReyes2013 and TwBarbieri2014 datasets. We evaluated the performance of our irony detection model with the same set of classifiers than in the previous experiments, also adding Random Forest (RF). The results presented in [47, 108] as well as the ones obtained with emotIDM are higher than those of [126]. As can be noticed, by employing RF we outperform the results obtained by the other approaches on TwReyes2013. With respect to TwIronyBarbieri2014 and TwSarcasmBarbieri2014, the results are in line with the ones obtained on the TwReyes2013 corpus. These results allow confirming the robustness of our model.

[4]The authors used the features described in Bamman and Smith [13].

Table 7.8. Results in F-measure terms when emotIDM is evaluated in an imbalanced scenario.

Corpus	State of the art		emotIDM			
		Result	NB	DT	SVM	RF
TwReyes2013	Reyes et al. [126] Fersini et al. [47] Nozza et al. [108]	0.53 0.83 0.85	0.57	0.83	0.80	**0.93**
TwIronyBarbieri2014			0.60	0.80	0.78	**0.93**
TwSarcasmBarbieri2014			0.70	0.84	0.83	**0.93**

We also experimented by modifying the distribution of each of the benchmark corpora. We generated a balanced distribution for TwMohammad2015, TwRiloff2013, and TwPtáček2014; on these corpora we randomly selected the same amount of ironic and non-ironic tweets, in other words we forced a balanced distribution. For each of the binary classifications of TwReyes2013 and TwBarbieri2014, we randomly selected 3,000 tweets for the ironic class, and 7,000 tweets for the non-ironic one. This distribution was chosed considering the imbalance degree shown in the TwMohammad2015 and TwPtáček2014 datasets. Table 7.9 shows the original distribution and the modified one (in bold) for each corpus.

Table 7.9. Balanced and imbalanced distribution for our set of corpora.

Dataset	Balanced				Imbalanced			
	Ironic		Non-ironic		Ironic		Non-ironic	
	Tweets	%	Tweets	%	Tweets	%	Tweets	%
TwReyes2013								
Irony-vs-Education	10,000	50	10,000	50	**3,000**	**30**	**7,000**	**70**
Irony-vs-Humor	10,000	50	10,000	50	**3,000**	**30**	**7,000**	**70**
Irony-vs-Politics	10,000	50	10,000	50	**3,000**	**30**	**7,000**	**70**
TwBarbieri2014								
Irony-vs-Education	10,000	50	10,000	50	**3,000**	**30**	**7,000**	**70**
Irony-vs-Humor	10,000	50	10,000	50	**3,000**	**30**	**7,000**	**70**
Irony-vs-Politics	10,000	50	10,000	50	**3,000**	**30**	**7,000**	**70**
Irony-vs-Newspaper	10,000	50	10,000	50	**3,000**	**30**	**7,000**	**70**
Sarcasm-vs-Education	10,000	50	10,000	50	**3,000**	**30**	**7,000**	**70**
Sarcasm-vs-Humor	10,000	50	10,000	50	**3,000**	**30**	**7,000**	**70**
Sarcasm-vs-Politics	10,000	50	10,000	50	**3,000**	**30**	**7,000**	**70**
Sarcasm-vs-Newspaper	10,000	50	10,000	50	**3,000**	**30**	**7,000**	**70**
TwMohammad2015								
	532	**50**	**532**	**50**	532	27	1,397	73
TwRiloff2013								
	474	**50**	**474**	**50**	474	22	1,689	78
TwPtáček2014								
	19,026	**50**	**19,026**	**50**	19,026	27	51,860	73

Figures 7.2, 7.3, and 7.4 show the performance in ROC terms[5] obtained by applying emotIDM in both balanced and imbalanced distributions on each one of the corpora. As expected, we observed changes in the performance of our model when it is evaluated on balanced and imbalanced scenarios. The most remarkable differences occur in the results obtained on TwRiloff2013, TwMohammad2015, and TwPtáček2014. When we experimented with DT and SVM on the balanced version of these corpora, the overall performance of emotIDM improves. On the contrary, when binary classifications on the imbalance versions of TwReyes2013 and TwBarbieri2014 are considered the performance drops. Results on TwRiloff2013 and TwMohammad2015 are lower in comparison with the rest even in a balanced scenario. This confirms the difficulty in recognizing crowdsourced ironic tweets and the self-tagged ones.

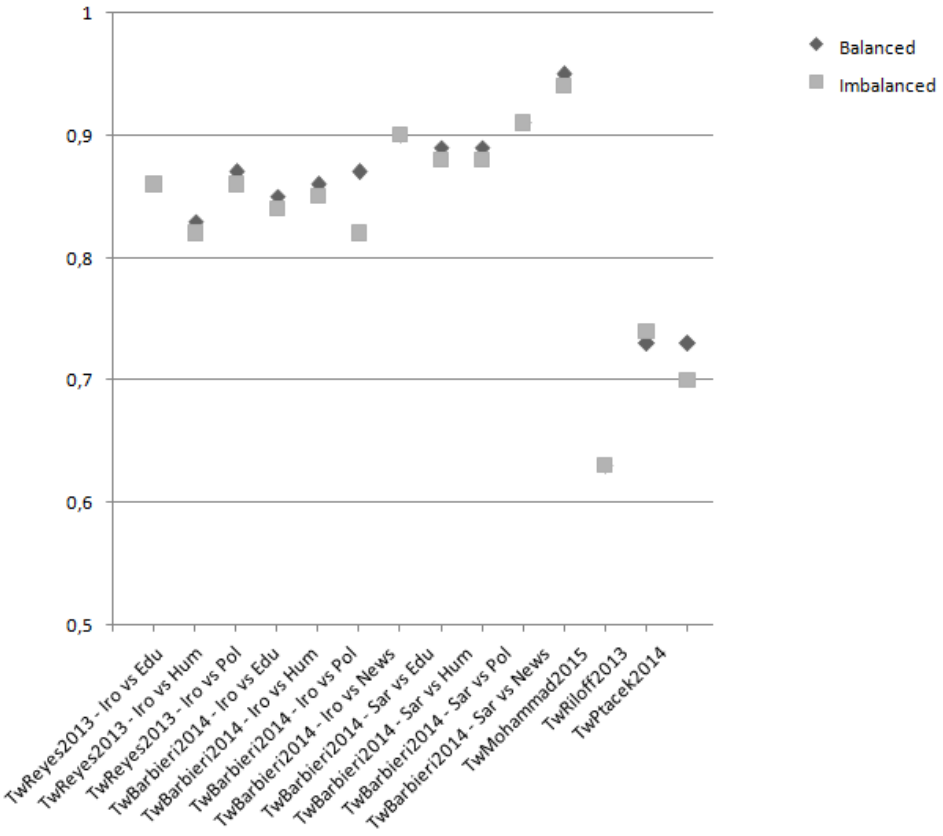


FIGURE 7.2. Performance of emotIDM using NB in ROC measure terms.

[5]ROC score is widely used to measure the performance of classifiers in imbalanced domains.

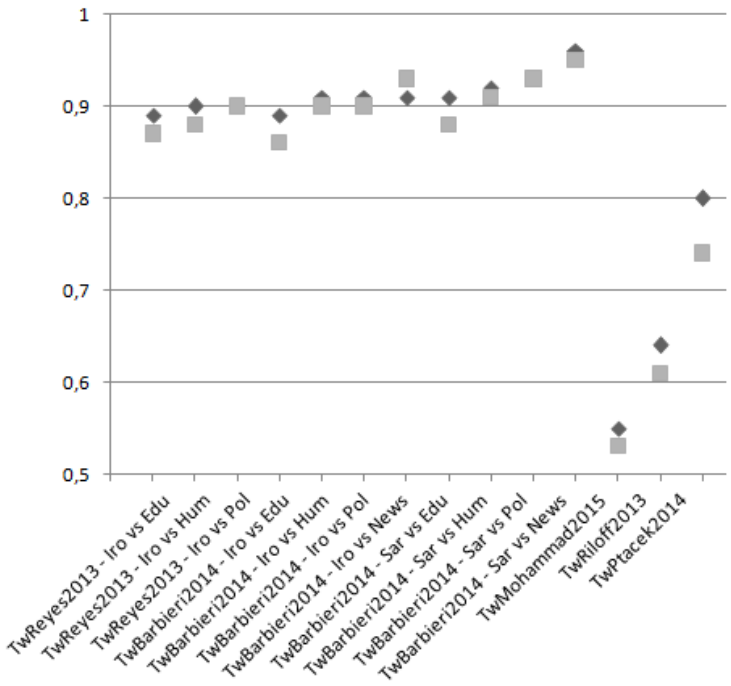


FIGURE 7.3. Performance of emotIDM using DT in ROC measure terms.

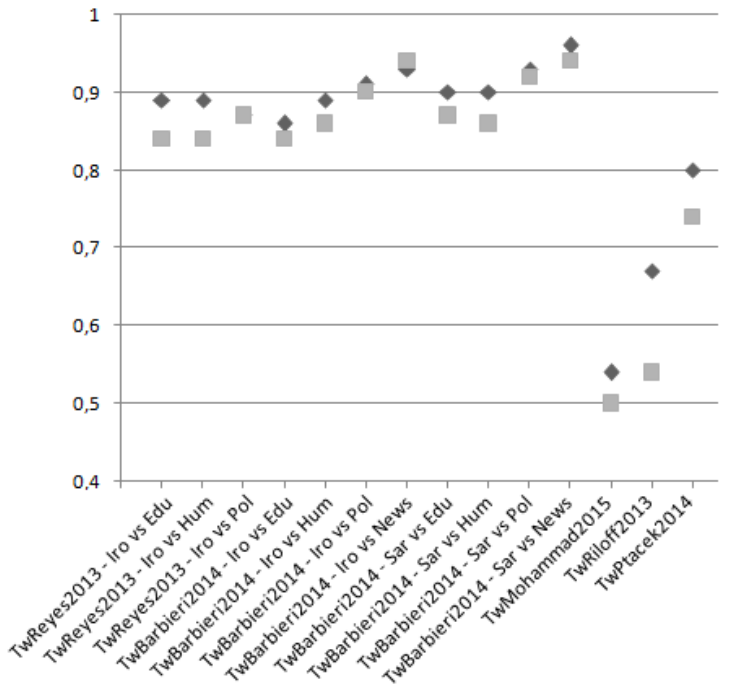


FIGURE 7.4. Performance of emotIDM using SVM in ROC measure terms.

7.5 A Qualitative Analysis of Affective Resources

As mentioned before, emotIDM makes use of several resources related to affective information. These lexica have been divided into different groups: *Sentiment* and *Emotions (Categorical and Dimensional)*. Some of these resources are similar regarding the nature of the information they provide. For example, HL and GI define two lists of words: positives and negatives. Therefore, an analysis of the content of these lexica is needed in order to determine the degree of overlapping among lexica containing the same information. The lower overlapping between the resources in emotIDM is, the higher coverage of our irony detection model will be.

In order to calculate the degree of overlap between two lexica we consider the *Matching coefficient* as defined in [98]:

$$Matching\ coefficient = X \cap Y \tag{7.1}$$

where X and Y are two different resources.

We evaluated the *Matching coefficient* over two groups of affective-related resources: Sentiment and Emotional Categories. For the sake of comparison, we selected only those lexica containing lists of words being: (i) positive and negative[6]; and (ii) categorical models of emotions considering only those categories that are present in at least two resorces[7].

7.5.1 Sentiment Resources

Six resources containing information related to Sentiment were selected. Table 7.10 shows the distribution of positive and negative words in each lexicon.

Table 7.10. Distribution of positive and negative words in sentiment resources

Resource	Polarity	
	Positive	Negative
AFINN	876	1,586
HL	2,006	4,783
GI	1,915	2,291
EWN	3,288	2,427
SUBJ_ST	1,717	3,619
SUBJ_WK	1,001	1,292
EmoLex	2,312	3,324

We calculated the Matching coefficient among each possible pair combination of sentiment resources. Tables 7.11 and 7.12 show the obtained results for positive and negative lists of words, respectively.

[6]We also included AFINN by applying two criteria: if a word has a polarity value lower than 0 it is considered as negative; if the polarity is higher than 0 is considered as positive.

[7]The information in Table 4.5 was used for selecting the categories.

Table 7.11. Matching coefficient among positive words in sentiment resources.

	HL	GI	EWN	SUBJ_ST	SUBJ_WK	EmoLex
AFINN	433	255	79	308	163	325
HL		723	123	1,093	475	681
GI			138	613	395	696
EWN				108	102	232
SUBJ_ST					20	537
SUBJ_WK						349

Table 7.12. Matching coefficient among negative words in sentiment resources.

	HL	GI	EWN	SUBJ_ST	SUBJ_WK	EmoLex
AFINN	865	351	138	536	214	647
HL		1,162	376	2,916	934	1,675
GI			166	823	391	903
EWN				236	162	332
SUBJ_ST					23	1,104
SUBJ_WK						535

In overall, the matching coefficient is lower among the positive words than the negative ones. The obtained results in both cases show that the overlap of redundant information is low. There are several factors that need to be considered for evaluating the overlap among different resources. For instance, the data used for building each lexicon: while AFINN was collected from words frequently used in Internet, EWN comes from information described in WordNet. As shown in Tables 7.11 and 7.12, EWN is the resource having on average the lower matching coefficient with respect to all resources considered. Therefore, the content in each lexicon is in some way dissimilar. By exploiting a wide range of sentiment resources, emotIDM should be able to cover several words expressing sentiment for characterizing irony.

We also evaluated the matching coefficient between each of the ironic and sarcastic set of tweets in the Twitter corpora used for experimental purposes. The results are shown in Appendix D.

7.5.2 Emotion Categories Resources

The emotional categories that EmoLex, EmoSN, and SentiSense have in common were used for calculating the degree of overlap among the resources related to emotions. A total of seven emotional categories were considered. Only one of them (i.e., 'Anticipation') is present only in two resources. Table 7.13 shows the distribution of words annotated with a particular emotional category. As it can be noticed EmoSN allows to cover the highest amount of words related to 'Joy' in comparison with the other two lexica.

Table 7.13 shows that the overlap degree concerning emotional categories is very low. For instance, with respect to the matching coefficient for 'Anticipation', it can be noticed that there are only two words in common between EmoLex and

Table 7.13. Distribution of words according to emotional categories.

Emotion	Resources		
	EmoLex	EmoSenticNet	SentiSense
Anger	1,246	829	55
Anticipation	1,190	-	152
Disgust	1,058	1,159	547
Fear	1,476	1,199	159
Joy	689	9,389	132
Sadness	1,191	1,536	134
Surprise	534	905	29

SentiSense. Therefore the coverage of using them for capturing information on this particular emotion is not redundant. In the case of 'Joy', although EmoSenticNet includes more than 9,000 words related to this emotional category, the coverage of emotIDM is improved by adding two more resources containing words labeled with the same affective category.

Table 7.14. Matching coefficient among different emotional categories.

Emotion	Resources	Matching coefficient
Anger	EmoLex and EmoSN	85
	EmoLex and SentiSense	12
	EmoSN and SentiSense	10
Anticipation	EmoLex and SentiSense	2
Disgust	EmoLex and EmoSN	85
	EmoLex and SentiSense	89
	EmoSN and SentiSense	25
Fear	EmoLex and EmoSN	132
	EmoLex and SentiSense	76
	EmoSN and SentiSense	28
Joy	EmoLex and EmoSN	248
	EmoLex and SentiSense	41
	EmoSN and SentiSense	22
Sadness	EmoLex and EmoSN	190
	EmoLex and SentiSense	60
	EmoSN and SentiSense	24
Surprise	EmoLex and EmoSN	24
	EmoLex and SentiSense	7
	EmoSN and SentiSense	1

7.6 Conclusions

In this chapter we have investigated further the robustness of emotIDM. We evaluated our irony detection model under several experimental settings obtaining competitive results. We carried out ablation experiments over the features comprised in emotIDM. Such experiments allow us to assess the performance of different affect features for identifying irony in Twitter. Different features emerged as relevant for identifying ironic and sarcastic tweets. We provided data-driven arguments to confirm the importance of considering features related to affective information for irony detection.

It has been recognized that irony is widely used in social media, however, is also well known that the difference between ironic and non-ironic data is enormous. Therefore, we investigated emotIDM in a more realistic scenario resembling the real use of irony in social media. According to the obtained results, our irony detection model seems to have a robust behaviour considering different degrees of imbalance.

Finally, we analized the content of the affective resources exploited in emotIDM. The obtained results allow us to confirm the importance of using a wide range of lexical resources for covering as much affective words as possible.

Chapter 8

Conclusions and Future Work

8.1 Conclusions

In this thesis, we have approached the problem of irony detection in Twitter. The goal of the present research was to determine whether affective information serves for detecting the presence of this figurative language device.

Attempting to take advantage of the subjective intrinsic value enclosed in ironic expressions, we presented a novel model, called emotIDM, for detecting irony relying mainly on affective information. For characterizing an ironic utterance, we used an extensive set of resources covering different facets of affect from sentiment to finer-grained emotions. We addressed irony detection by casting it as a binary classification problem. To evaluate our model, we collected a set of Twitter corpora already used in previous research. We decided to consider benchmark data based on two-fold purposes: first to evaluate our model covering several aspects related to the corpora such as collection mode, size, imbalance degree, etc.; and to compare the performance of our model against other approaches of the state of the art. Several experiments were carried out to validate the effectiveness of the proposed approach. We assessed the contribution of each of the affect-related resources comprised in emotIDM (cf. Chapter 7). A broad number of ablation experiments were carried out over the benchmark corpora. Our experimental setting demonstrate and confirm that our method is able to recognize the presence of irony in Twitter. A qualitative analysis of the content of the lexical resources exploited has been performed in order to determine the degree of overlapping among lexica containing the same information. This confirmed the importance of including several affective resources for having a wider coverage of words related to affect phenomena. To sum up, the obtained results serve to validate the importance of capturing affect-related phenomena from texts for irony detection.

Furthermore, we investigated the differences among tweets labeled with #irony and #sarcasm. We also studied a less explored hashtag, #not, that has been used for collecting sarcastic intention. We found that, in fact, these hashtags are used to refer to different figurative language devices. We contributed to the less investigated topic in computational linguistics on the separation between irony

and sarcasm in social media. Our results outperformed the ones of the state of the art. We identified promising features based on affect-related phenomena for discriminating among different kinds of figurative language devices. We analyzed the role of polarity reversal in tweets containing ironic hashtags: it seems that the impact of such phenomenon is not the same among them. Finally, we introduced an irony-aware sentiment analysis system. We incorporated emotIDM into a pipeline for determining the polarity of a given text. To calculate the sentiment expressed we took advantage of several facets of affect. We compared our results with the state of the art demonstrating the relevance of considering affective information together with features alerting on the presence of irony for perform sentiment analysis.

The results of this thesis allows us to address the research questions described in the Chapter 1:

I. ***Could information about different facets of affect be useful for detecting irony in social media?***

> We introduced emotIDM (see Section 4.4), a novel model for identifying irony in Twitter. emotIDM comprises a wide range of affect-related features ranging from sentiment to finer-grained emotions. Our model takes advantage of several resources covering different facets of affective information. To the best of our knowledge, emotIDM is the first irony detection model broadly exploiting such kind of information for characterizing ironic utterances. We addressed irony detection following the widely accepted assumption in computational linguistics of considering irony as an umbrella term that also covers sarcasm. Therefore, emotIDM does not distinguish among different kinds of irony. We evaluated the robustness of our model over Twitter corpora with different properties. Results show that emotIDM performs quite well across the experiments carried out. In the majority of the cases, our results outperform those from the related work confirming that affective information helps in distinguishing between ironic and non-ironic tweets.

II. ***Are the #irony and #sarcasm hashtags used to label different ironic intentions***

> With the aim to further investigate tweets labeled with the #irony and #sarcasm hashtags, in Chapter 5 we analyzed a set of tweets rich on figurative language. Our study was focused on the role of different facets of affect in such kinds of tweets. We found data-driven arguments on the differences among tweets containing these hashtags. They are used for labeling different figurative phenomena. We also analyze the role of the #not hashtag, often used to collect sarcastic tweets. Tweets labeled with #sarcasm and #not are perceived as more negative. They usually contain words involving positive sentiment and emotions for communicating the opposite of their meaning.

The hashtag #irony is used in tweets expressing a positive praise. Moreover, very often such hashtag is used for marking situational irony, where the intention is not precisely to said the contrary of the literal meaning of the words. Interestingly, tweets labeled with #not use negation as a low-salient marker to achieve a sarcastic intention. Besides, our findings also show distinctions for what concerns to the polarity reversal among these kinds of tweets (see Section 5.4.2). When the #sarcasm hashtag is used there is usually a full polarity reversal (mostly from positive to negative); instead, the #irony hashtag tends to have an attenuation effect on the polarity (often from negative to neutral).

III. ***Is it possible to improve the performance of sentiment analysis systems being aware of the presence of ironic content?***

Irony is widely used in social media. It represents one of the biggest challenges for determining the sentiment expressed in a given piece of text. We analyzed the impact of these figurative language devices on some shared tasks dedicated to sentiment analysis (see Section 2.3). The performance of sentiment analysis systems exploiting different approaches is, broadly speaking, good when they need to deal with literal content. However, when ironic content is involved the performance drops significatively. Such results demonstrated the importance of considering the irony as a fundamental part of developing accurate sentiment analysis systems. Correctly identifying ironic content could help to avoid misleading sentiment evaluation. In Section 6.3, we proposed a sentiment analysis system based on a pipeline. In a first phase, emotIDM is used for recognizing irony. Then, being aware of the presence of ironic content, we calculated and assigned a polarity degree. We compared our results against those of the shared task at SemEval 2015 Task 11 ranking among the best systems.

To summarize, we demonstrated the usefulness of exploiting different facets of affective information for dealing with the presence of irony on Twitter. The findings of this research were presented in different publications that are introduced in the following section.

8.2 Research Contributions

Below, we outline the publications derived from this research by grouping them into three main groups:

1. Irony and Sarcasm Detection

We describe our approach for addressing irony detection by means of affective information in two journals and one conference papers:

- *Journal papers*
  - **Hernández Farías, D.I.**, Patti, V., and Rosso, P. (2016). Irony Detection in Twitter: The Role of Affective Content. *ACM Transactions on Internet Technology* 16(3), pp. 1-24.
    DOI: `http://dx.doi.org/10.1145/2930663`
  - Sulis, E., **Hernández Farías, D.I.**, Rosso, P., Patti, V., and Ruffo, G. (2016). Figurative messages and affect in Twitter: Differences between #irony, #sarcasm and #not. *Knowledge-Based Systems* 108, pp. 132-143.
    DOI: `http://dx.doi.org/10.1016/j.knosys.2016.05.035`
- *Conference paper*
  - **Hernández Farías, D.I.**, Benedí, J.M., and Rosso, P. (2015). Applying Basic Features from Sentiment Analysis for Automatic Irony Detection. In: *Proceedings of the Pattern Recognition and Image Analysis: 7th Iberian Conference, IbPRIA 2015, Santiago de Compostela, Spain, June 17-19, 2015.* Springer International Publishing, pages 337-344.
    DOI: `http://dx.doi.org/10.1007/978-3-319-19390-8_38`

We presented both a literature review on irony and sarcasm detection, and an analysis of various sentiment analysis shared tasks where the presence of irony was considered for evaluation purposes in the following chapter:

- **Hernández Farías, D.I.** and Rosso, P. (2017). Irony, Sarcasm and Sentiment Analysis. Chapter 7 in Pozzi, F.A., Fersini, E., Messina, E., and Liu, B., eds.: *Sentiment Analysis in Social Networks.* ISBN: 978-0-12-804412-4. Morgan Kaufmann, pages 113-128.
  DOI: `http://dx.doi.org/10.1016/B978-0-12-804412-4.00007-3`

We incorporated emotIDM into a sentiment analysis system attempting to validate the importance of detecting irony before assigning polarity in one conference paper:

- **Hernández Farías, D.I.**, Bosco, C., Patti, V., and Rosso, P. (2017). Sentiment Polarity Classification of Figurative Language: Exploring the Role of Irony-Aware and Multifaceted Affect Features. In: *Proceedings of the 18th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2017)*, In Press.

2. Sentiment Analysis and Irony and Sarcasm Detection

We participated in different shared tasks on sentiment analysis in both English and Italian. We describe our participating systems that also exploit different kinds of affective information in three conference papers:

- Buscaldi, D., **Hernández Farías, D.I.** (2016). IRADABE2: Lexicon Merging and Positional Features for Sentiment Analysis in Italian. In: *Proceedings*

of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016), Napoli, Italy. `http://ceur-ws.org/Vol-1749/`.

- **Hernández Farías, D.I.**, Sulis, E., Patti, V., Ruffo, G., and Bosco, C. (2015). ValenTo: Sentiment Analysis of Figurative Language Tweets with Irony and Sarcasm. In: *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, Denver, Colorado, Association for Computational Linguistics, pages 694-698.

- **Hernández Farías, D.I.**, Buscaldi, D., and Priego-Sánchez, B. (2014). IRADABE: Adapting English Lexicons to the Italian Sentiment Polarity Classification task. In: *Proceedings of the First Italian Conference on Computational Linguistics (CLiC-it 2014) and the fourth International Workshop EVALITA2014, Pisa, Italy.*

We collaborated in a research involving sentiment analysis in Spanish; the results were published in two papers: one in journal and one in a conference:

- *Journal paper*

  - Baca Gomez, Y.R., Martinez, A., Rosso, P., Estrada, H., and **Hernández Farías, D.I.** (2016). Web Service SWePT: A Hybrid Opinion Mining Approach. *Journal of Universal Computer Science* 22(5):671-690.

- *Conference paper*

  - Baca Gómez, Y., Castro Sánchez, N., Martinez, A., **Hernández Farías, D.I.**, and Rosso, P. (2014). Impacto de la ironía en la minería de opiniones basada en un léxico afectivo. *Actas de las V Jornadas de la Red en Tratamiento de la Información Multilingüe y Multimodal, Cazalla de la Sierra, España,* CEUR Workshop Proceedings. 1199:41-44.

3. Other contributions

Below, a list of additional research works carried out is presented. These papers are partially related to the main objectives of this thesis.

- Developing corpora for sentiment analysis and irony detection in Italian and Spanish.

  - Applying sentiment analysis on a specific domain: The project *Subjective Well Being and Fertility*[1].

[1] `https://swellfer.wordpress.com/`

    - ∗ Sulis, E., Bosco, C., Patti, V., Lai, M., **Hernández Farías, D.I.**, Mencarini, L., Mozzachiodi, M., and Vignoli, D. (2016). Subjective Well-Being and Social Media. A Semantically Annotated Twitter Corpus on Fertility and Parenthood. In: *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016), Napoli, Italy.* `http://ceur-ws.org/Vol-1749/`.
  - Stranisci M., Bosco C., **Hernández Farías, D.I.**, and Patti, V. (2016). Annotating Sentiment and Irony in the Online Italian Political Debate on #labuonascuola. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia*, European Language Resources Association (ELRA).
  - Rosso, P., **Hernández Farías, D.I.**, and Rangel, Francisco. (2016). Universality and Creativity: The Usage of Language in Gender and Irony. Chapter in Degli Esposti, M., Altmann, E.G., and Pachet, F. eds.: *Creativity and Universality in Language.* ISBN: 978-3-319-24403-7. Springer International Publishing, pages 177-186. DOI: `http://dx.doi.org/10.1007/978-3-319-24403-7_11`
  - Stranisci M., Bosco C., Patti, V., and **Hernández Farías, D.I.** (2015). Analyzing and Annotating for Sentiment Analysis the Socio-political Debate on #labuonascuola. In: *Proceedings of the Second Italian Conference on Computational Linguistics CLiC-it 2015*, Trento, Italy.
  - Buscaldi D. and **Hernández Farías, D.I.** (2015). Sentiment Analysis on Microblogs for Natural Disasters Management: A Study on the 2014 Genoa Floodings. In: *Proceedings of the 24th International Conference on World Wide Web (WWW '15 Companion).* ACM, pages: 1185-1188. DOI: `http://dx.doi.org/10.1145/2740908.2741727`
  - Rangel F., **Hernández Farías, D.I.**, Rosso, P., and Reyes, A. (2014). Emotions and irony per gender in Facebook. In: *Proceedings of Workshop on Emotion, Social Signals, Sentiment & Linked Open Data (ES3LOD), LREC-2014*, Reykjavík, Iceland.

- Evaluating the relevance of affective information in stance detection.
  - Lai, M., **Hernández Farías, D.I.**, Patti, V., and Rosso, P., Friends and Enemies of Clinton and Trump: Using Context for Detecting Stance in Political Tweets. (2016) In: *Proceedings of the 15th Mexican International Conference on Artificial Intelligence.* `https://arxiv.org/abs/1702.08021`

8.3 Future Work

From a computational linguistics perspective, understanding the ironic intention of a given message is an ongoing task. Despite the fact that several approaches have

been proposed for addressing irony and sarcasm detection, there still leaves room for future study. We identify some potential future research areas:

- *Corpora.* Currently, there are some available corpora for experimenting with irony detection. However, there is a lack of a well-defined baseline to evaluate and compare the performance of the approaches related to irony detection. Furthermore, concerning to crowdsourcing data, it could be useful to improve the annotation process to cover finer-grained properties of irony. A proposal considering a multi-layered annotation schema on pragmatic aspects of irony is described in [82].

- *Context.* Commonly, without an overall understanding of the context surrounding an utterance it is not possible to infer its real intention. A less explored area in irony detection conveys how to incorporate and exploit common-sense and world-knowledge concepts for identifying incongruities and inconsistencies, and then determine whether a given sentence is ironic or not.

- *Markers of ironic praise and ironic situation.* According to the results in [137], ironic tweets tend to express different degrees of sentiment when they contain specific markers. In this context, a further studying of the positive or 'genteel' forms of irony is a promising research direction, think for instance to the use of idiomatic expressions such as "just kidding" or to tweets marked with hashtag #irony, which in some cases can be regarded as samples of ironic praise and in many cases as expressions related to situational irony. Such expressions of positive irony, rather than invoking a negative attitude, as in the case of sarcasm, tend to be playful and produce a comic or parodic effect, often to strengthen ties with others chatting online. In particular, concerning the use of the hashtag #irony, it could be interesting to investigate when and how it is used for marking an ironic praise or as a way to highlight an ironic situation, also in connection with the theoretical framework recently proposed in [61] where the authors investigated the pleasantness degree of default and nondefault irony and sarcasm.

- *Multimodal-irony detection.* Social media users tend to use different kinds of content (such as text, images, video, etc.) to create their messages. Often, the text in a post is not enough for capturing the ironic intention. Therefore, being able to manage all (or the majority of) the information in social media content represents an attractive challenge for determining the presence of figurative language devices. Schifanella et al. [132] proposed the first approach considering multimodal information for identifying sarcasm in social media.

- *Well-being studies.* On social media users freely express what is on their mind at any moment in time, at any location, and about virtually anything. As we have studied in this thesis these large amounts of spontaneously produced texts open up a unique opportunity to learn more about such users' emotions and opinion expressions, possibly involving the use of irony, but also about their traits, both demographics (e.g. age, gender) and psychographics (e.g. personality). Therefore, an interesting direction of future work is to study

how to combine irony detection with the recognition of such authors' traits. Being able to recognize when the texts have an ironical intention could help to identify potential focal points, where people prefer to communicate their opinions subtly[2].

[2] http://pan.webis.de/clef17/pan17-web/index.html

Appendix A

Ablation Test in TwMohammad2015 and TwRiloff2013

Table A.1 shows the obtained results in F-measure terms when different features combination are used on two different corpora: TwMohammad2015 and TwRiloff2013.

TABLE A.1. Results of different combination of features on TwMohammad2015 and TwRiloff2013. The underlined values are not statistically significant (t-test with 95% of confidence value)

	TwMohammad2015			**TwRiloff2013**		
	NB	DT	SVM	NB	DT	SVM
Each set individually						
Sentiment	0.65	0.64	-	0.70	0.74	-
eCat	0.66	0.64	-	0.73	0.74	-
eDim	0.65	-	-	0.70	-	-
Combination of affective-related groups of features						
Sentiment + eCat	0.65	0.64	-	0.71	0.72	-
Sentiment + eDim	0.64	0.63	-	0.70	0.73	-
eCat + eDim	0.65	0.64	-	0.71	0.72	-
Structural + each group of affective-related features						
Structural + Sentiment	**0.67**	0.64	0.62	0.75	0.77	-
Structural + eCat	0.66	0.64	0.61	0.75	0.76	-
Structural + eDim	0.64	0.65	0.61	0.74	0.76	-
Structural + each one of the resources in Sentiment						
Structural + AFINN	0.66	0.64	-	0.73	0.76	-
Structural + HL	**0.67**	0.64	-	0.73	**0.78**	-
Structural + GI	0.66	0.64	-	0.73	0.76	-
Structural + EmoLex	**0.67**	0.64	-	0.72	0.75	-
Structural + SWN	0.66	0.63	-	0.73	0.74	-
Structural + SN	0.66	0.64	-	0.72	0.75	-
Structural + SO	0.66	0.63	-	0.72	0.74	-
...						

	NB	DT	SVM	NB	DT	SVM
Structural + EWN	0.66	0.63	-	0.71	0.75	-
Structural + SUBJ	0.66	0.65	-	0.73	0.76	-
Structural + each one of the resources in eCat						
Structural + LIWC	0.66	0.64	-	0.74	0.77	-
Structural + EmoLex	0.65	0.65	-	0.74	0.76	-
Structural + EmoSN	**0.67**	0.66	-	0.75	0.77	-
Structural + SentiSense	**0.67**	0.64	-	0.73	0.76	-
Structural + each one of the resources in eDim						
Structural + DAL	0.66	0.65	-	0.74	0.74	-
Structural + SN	0.64	0.63	-	0.73	0.76	-
Structural + ANEW	0.64	0.66	-	0.73	0.74	-
Sentiment + each one of the resources in eCat						
Sentiment + LIWC	0.66	0.63	-	0.71	0.72	-
Sentiment + EmoLex	0.65	0.64	-	0.72	0.73	-
Sentiment + EmoSN	0.65	0.64	-	0.71	0.73	-
Sentiment + SentiSense	0.66	0.64	-	0.71	0.72	-
Sentiment + each one of the resources in eDim						
Sentiment + DAL	0.66	0.65	-	0.71	0.73	-
Sentiment + SN	0.65	0.63	-	0.71	0.73	-
Sentiment + ANEW	0.65	0.63	-	0.71	0.74	-
Emotions + each one of the resources in Sentiment						
Emotions + AFINN	0.66	0.62	-	0.71	0.74	-
Emotions + HL	0.66	0.64	-	0.71	0.74	-
Emotions + GI	0.65	0.63	-	0.71	0.74	-
Emotions + EmoLex	0.66	0.64	-	0.71	0.73	-
Emotions + SWN	0.65	0.63	-	0.71	0.74	-
Emotions + SN	0.66	0.63	-	0.71	0.73	-
Emotions + SO	0.65	0.62	-	0.70	0.73	-
Emotions + EWN	0.65	0.63	-	0.71	0.73	-
Emotions + SUBJ	0.66	0.64	-	0.71	0.74	-

Appendix B

Ablation Test by Group of Affective Features

TwReyes2013

Figures B.1, B.2, and B.3 show the results of the ablation experiments on the TwReyes2013 dataset.

FIGURE B.1. Ablation results in the *Irony-vs-Education* on TwReyes2013.

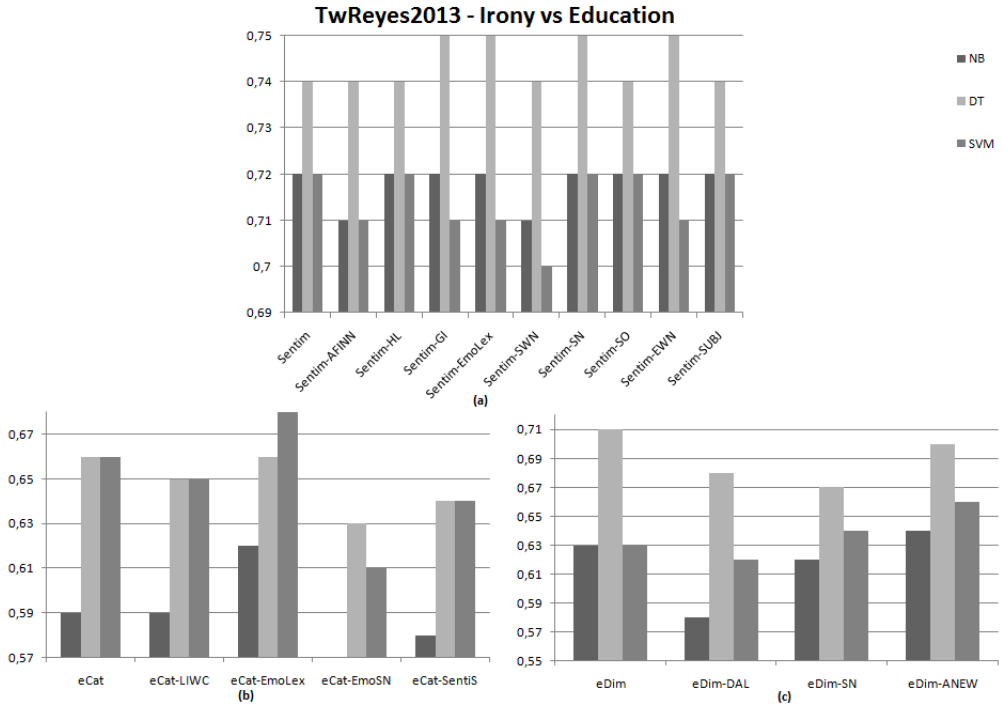


FIGURE B.2. Ablation results in the *Irony-vs-Humor* on TwReyes2013.

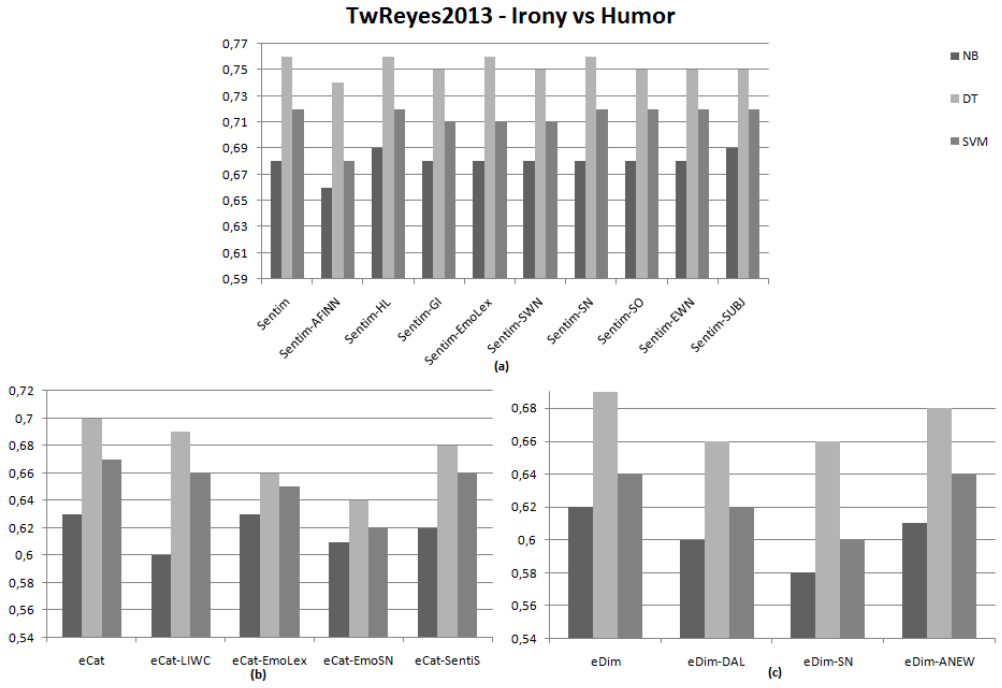
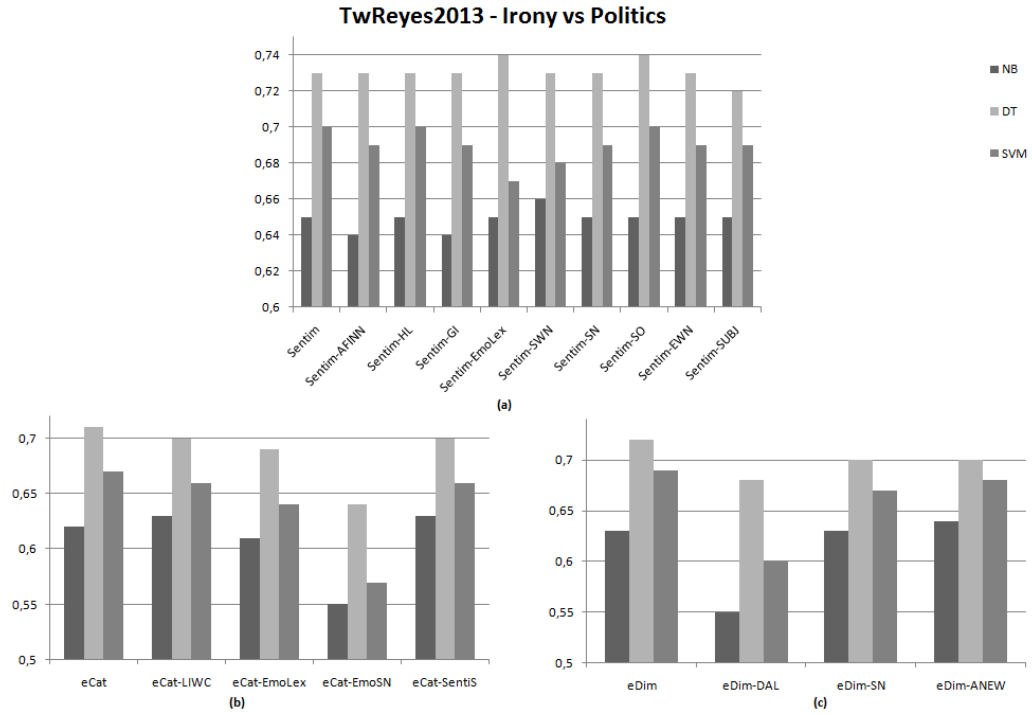


FIGURE B.3. Ablation results in the *Irony-vs-Politics* on TwReyes2013.



TwIronyBarbieri2014

Figures B.4, B.5, B.6, and B.7 show the results of the ablation experiments on the TwIronyBarbieri2014 dataset.

FIGURE B.4. Ablation results in the *Irony-vs-Education* on TwIronyBarbieri2014.

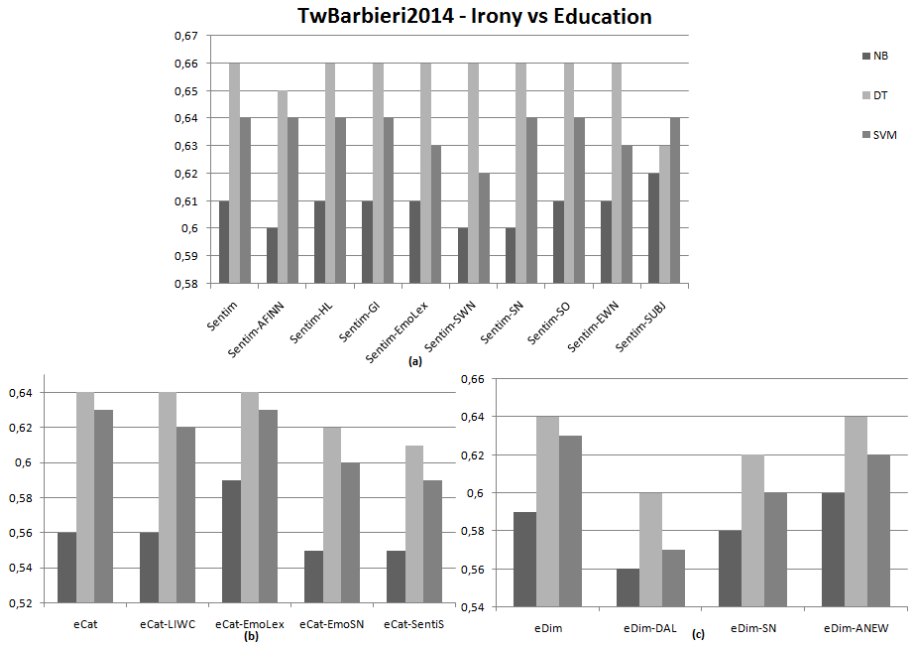


FIGURE B.5. Ablation results in the *Irony-vs-Politics* on TwIronyBarbieri2014.

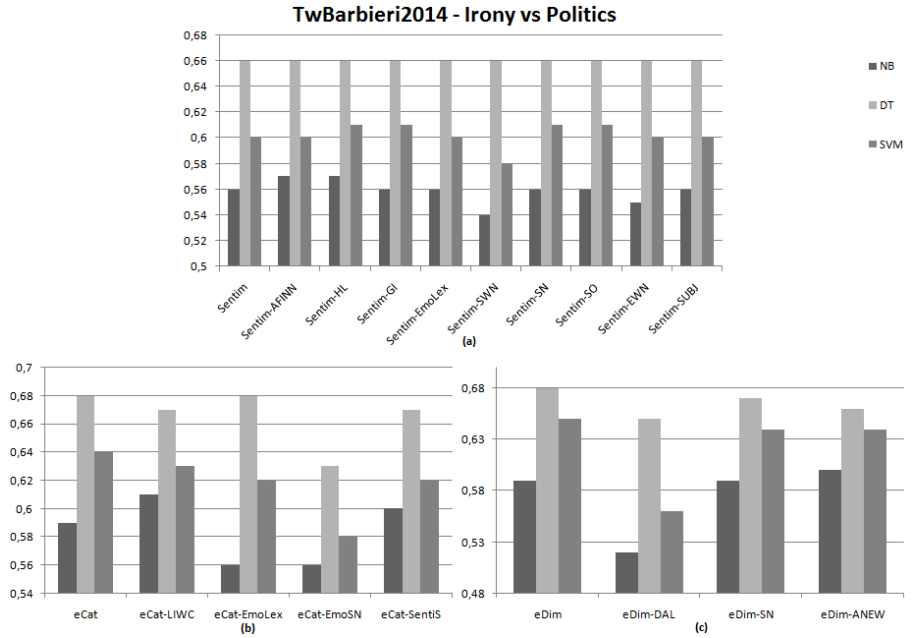


FIGURE B.6. Ablation results in the *Irony-vs-Humour* on TwIronyBarbieri2014.

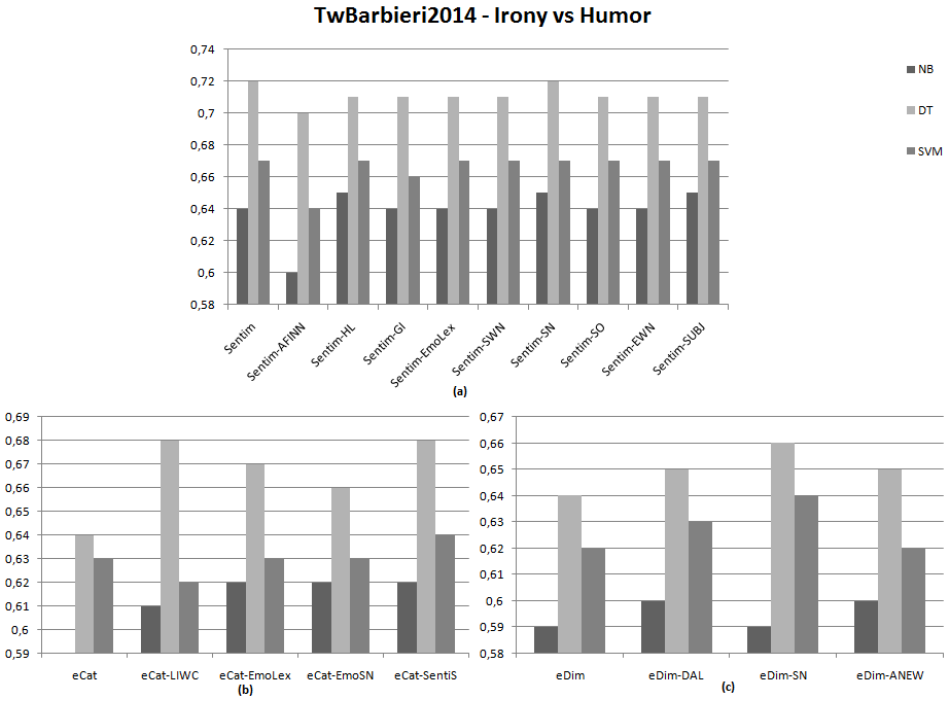
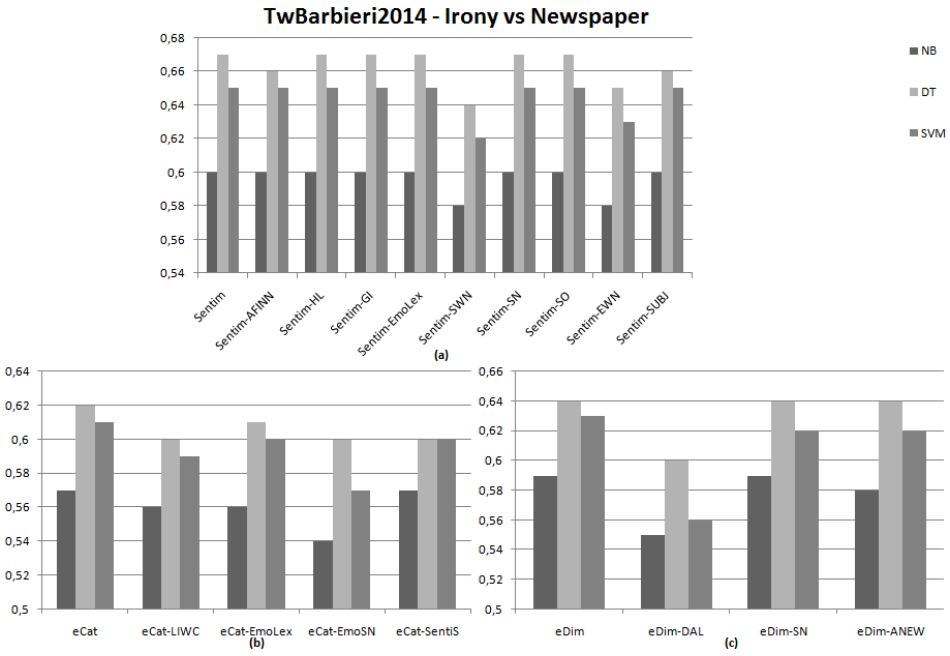


FIGURE B.7. Ablation results in the *Irony-vs-Newspaper* on TwIronyBarbieri2014.



TwSarcasmBarbieri2014

Figures B.8, B.9, B.10, and B.11 show the results of the ablation experiments on the TwSarcasmBarbieri2014 dataset.

FIGURE B.8. Ablation results in the *Sarcasm-vs-Education* on TwSarcasmBarbieri2014.

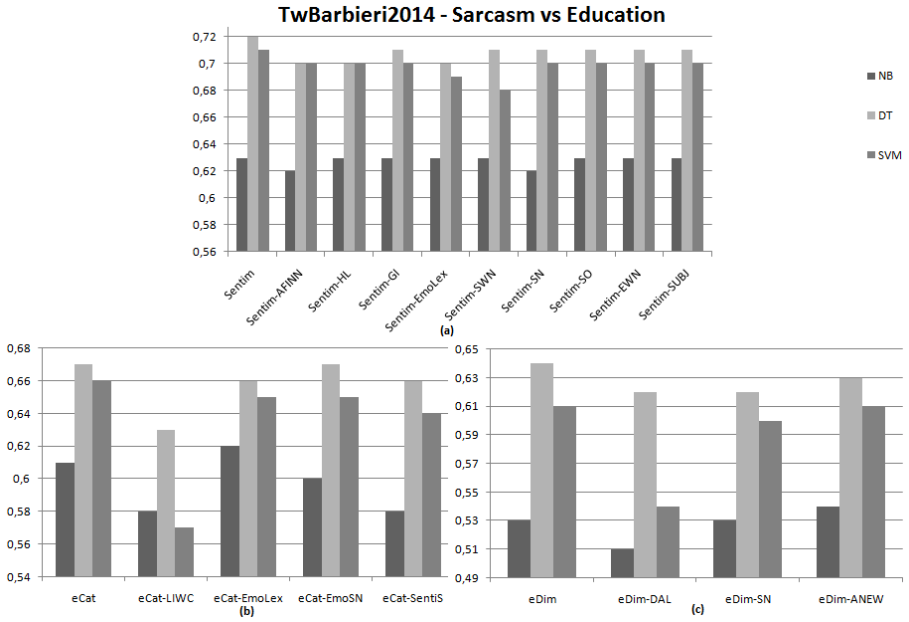


FIGURE B.9. Ablation results in the *Sarcasm-vs-Politics* on TwSarcasmBarbieri2014.

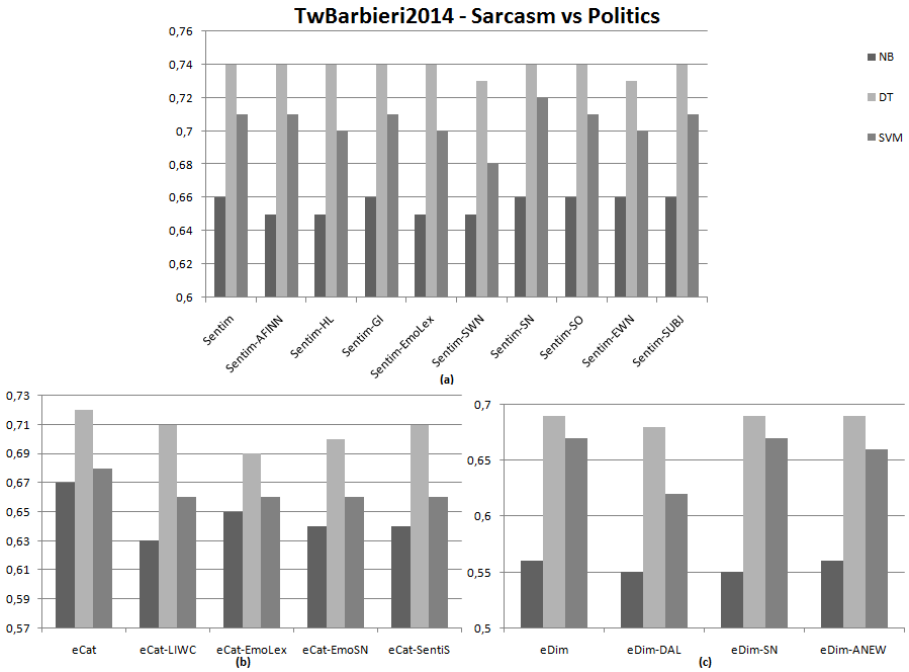


FIGURE B.10. Ablation results in the *Sarcasm-vs-Humour* on TwSarcasmBarbieri2014.

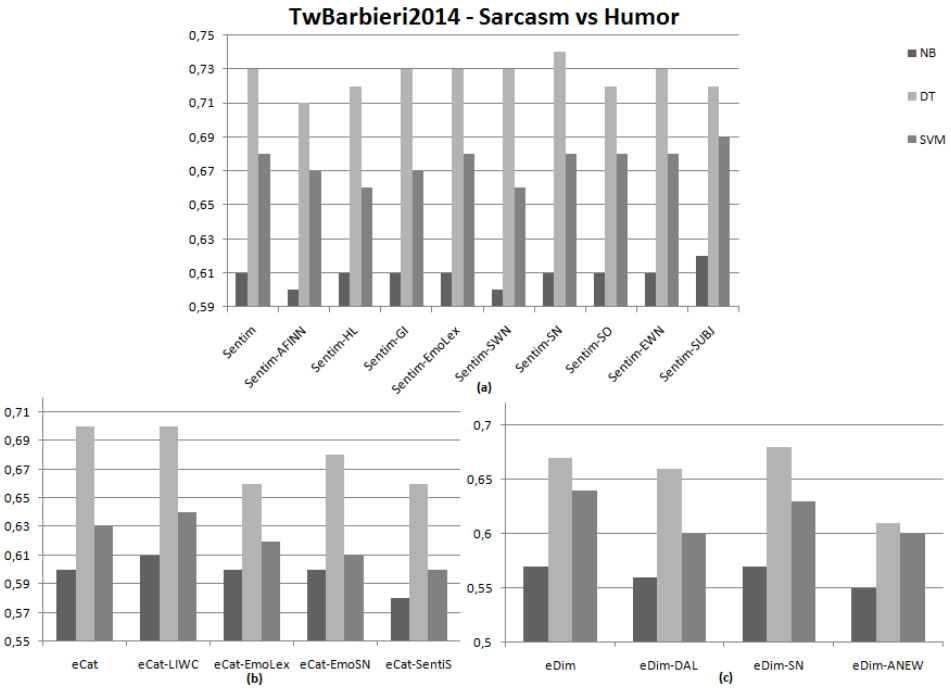
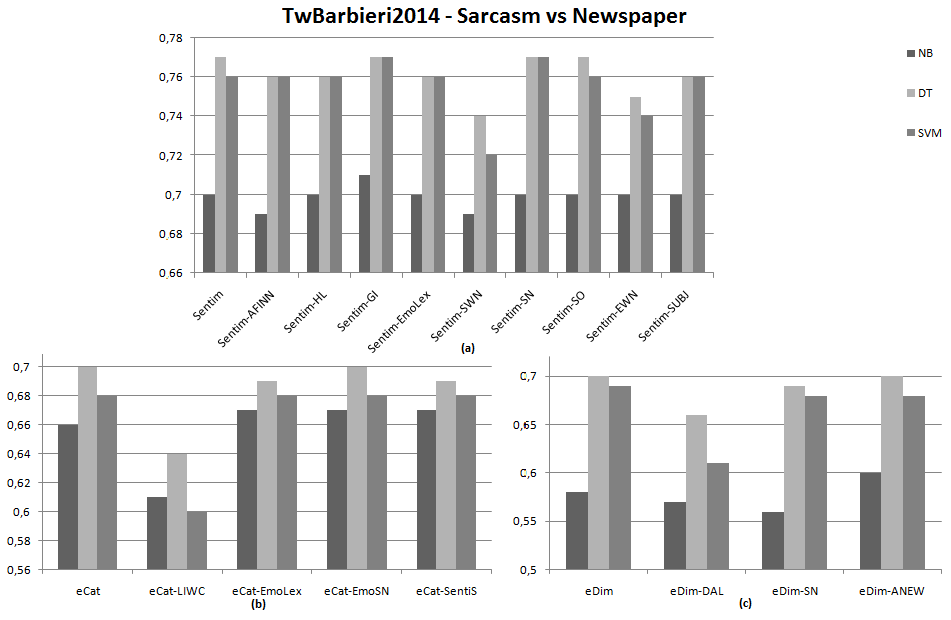


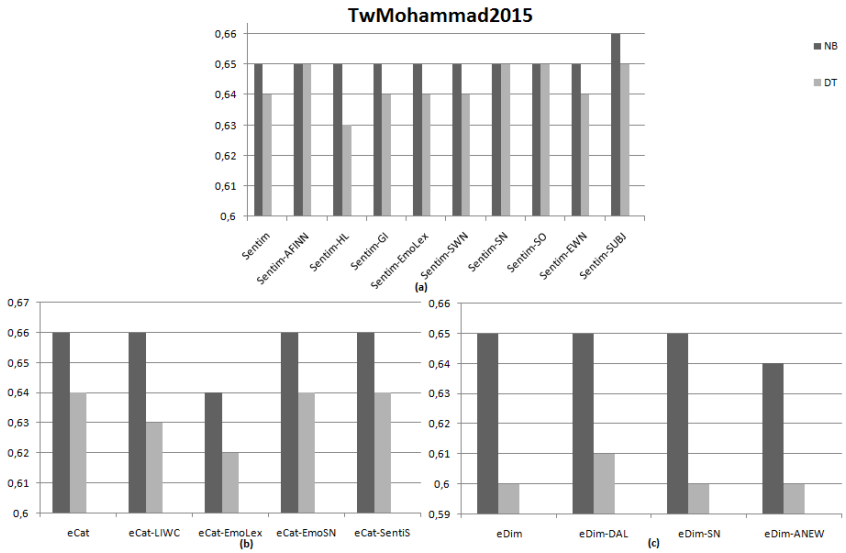
FIGURE B.11. Ablation results in the *Sarcasm-vs-Newspaper* on TwSarcasmBarbieri2014.



TwMohammad2015

Figure B.12 shows the results of the ablation experiments on the TwMohammad2015 dataset.

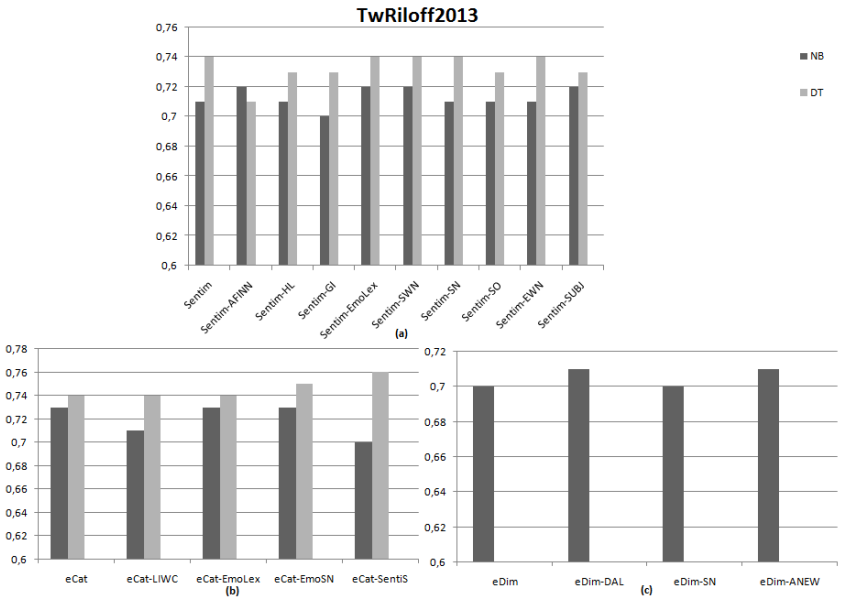
FIGURE B.12. Ablation results on the TwMohammad2015 dataset.



TwRiloff2013

Figure B.13 shows the results of the ablation experiments on the TwRiloff2013 dataset.

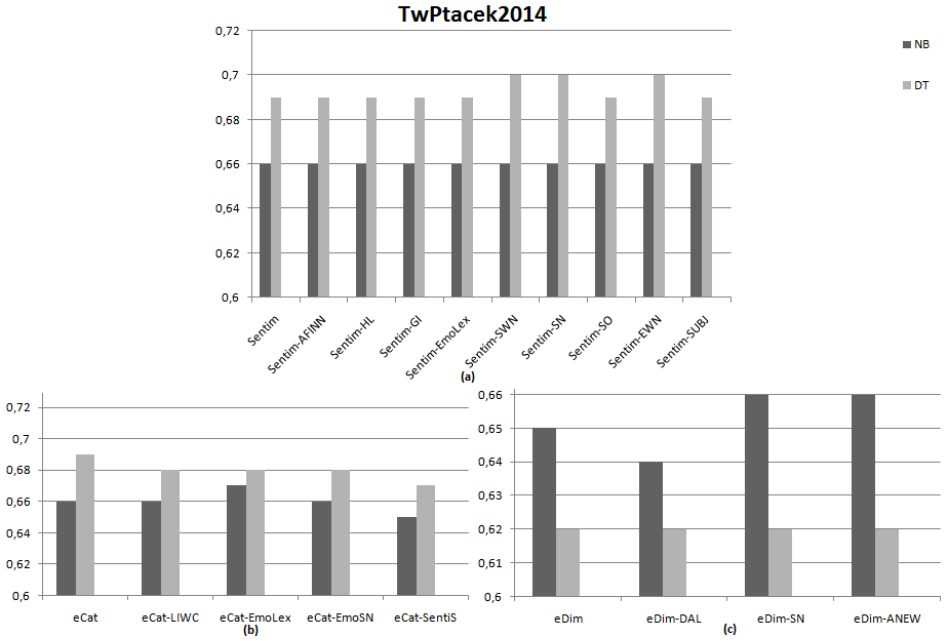
FIGURE B.13. Ablation results on the TwRiloff2013 dataset.



TwPtáček2014

Figure B.14 shows the results of the ablation experiments on the TwPtáček2014 dataset.

FIGURE B.14. Ablation results on the TwPtáček2014 dataset.



Appendix C

An Emotional-graph Representation of Sarcastic Tweets

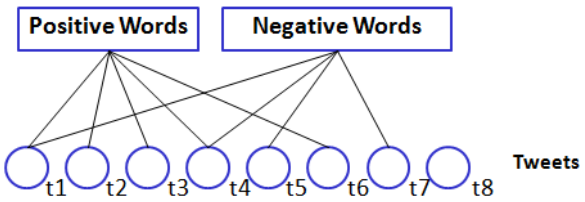
The aim of the present work is to explore the role of affective content under a graph approach. The TwPtáček2014 dataset was used for experimental purposes. The open-source network analysis and visualization software Gephi [23] was used[1]. This tool allows visualizing graphs and networks. In this work, the ForceAtlas2 algorithm [77] in Gephi was applied to produce the representation of the network structure. The default values defined in Gephi were used during the experimental setting.

C.1 Sentiment Words Distribution

The main objective of this experiment is to identify the relationship between sarcastic and non-sarcastic tweets, and words labeled as "positive" and "negative". According to [147, 137], in sarcastic tweets, the positive words are used to emphasize the contrast between what is literally said and the real intention of the speaker.

The problem was modeled as shown in Figure C.1, where the "positive" and "negative" are the words defined in AFINN.

FIGURE C.1. Graph representation of sentiment words distribution

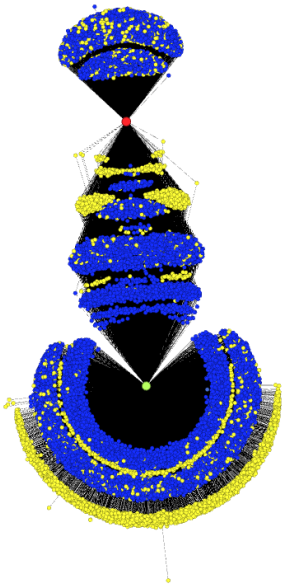


[1] https://gephi.org/

AFINN

In Figure C.2, it is possible to observe that this resource seems to cover a high percentage of sarcastic tweets (in blue). The difference among the "positive" and "negative" nodes is very noticeable; it allows confirming in some way the theoretical statement on the use of positive words in sarcastic tweets. Besides, it is also interesting the high amount of tweets containing more or less the same amount of "positive" and "negative" words; this could be considered as an insight on the "contradiction" that sometimes occurs in sarcastic sentences.

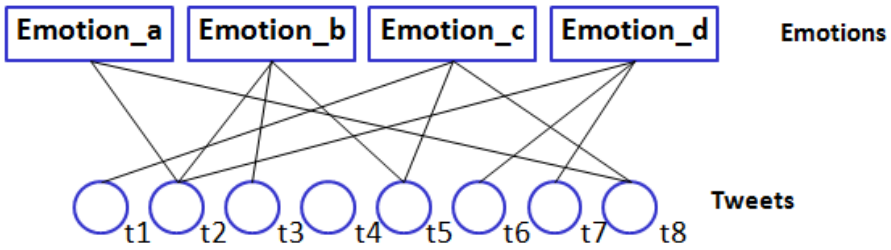
FIGURE C.2. Positive (green node) and negative (red node) words distribution of the TwPtáček2014 dataset by using AFINN. The sarcastic tweets are represented in blue, while the non-sarcatic ones are in yellow.



C.2 Emotional Words Distribution

In a similar fashion than in Section C.1, an experiment involving the presence of words related with emotions was carried out. As mentioned before, the emotional content in sarcastic utterances could provide interesting insights to characterize this kind of figurative language device. Figure C.3 shows a schema to represent this problem, where "Emotion_letter" means an emotional category (for instance joy or anger).

FIGURE C.3. Graph representation of emotion-related words distribution



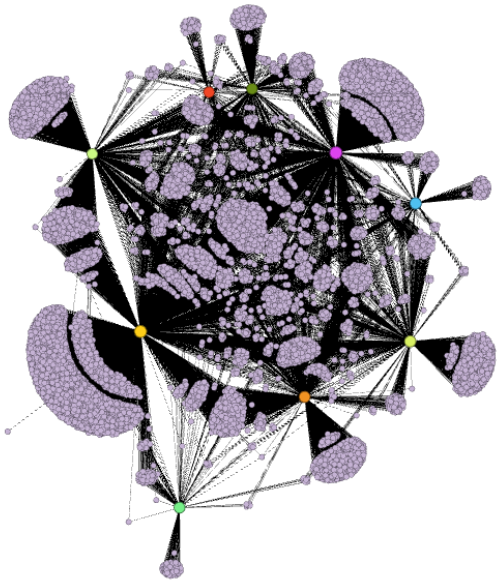
SentiSense

Figure C.4 shows the distribution of the nine emotions (*joy*, *fear*, *surprise*, *anger*, *disgust*, *love*, *anticipation*, *sadness*, and *like*) from SentiSense and the sarcastic tweets on the TwPtáček2014 dataset. The words defined in *like* (light orange node) seem to be highly frequent in the sarcastic tweets.

SentiSense captures in an interesting way some relations between emotional categories. According to the information gain analysis carried out in [71], *love* (pistachio green node) is one of the most relevant emotional categories to distinguish between sarcastic and non-sarcastic tweets, this is in some way confirmed by Figure C.4. The words in the *disgust* (purple node) category seem to be frequent in the sarcastic tweets (some of most frequent ones are: annoying, blame, break, cold, fake, hard, hate, jealous, little, problem, wait, and wrong).

There is an interesting relation between three emotional categories: *anticipation* (orange node), *like* (light orange node), and *surprise* (light green node). It seems that a considerable amount of sarcastic tweets share words that come from these emotional categories. Two of them, i.e., *anticipation* and *suprise* are taken into account as "opposite" emotions; therefore, when these kinds of words are used together with those from *like* this could be considered as an insight on the inherent "contradiction" in sarcastic utterances. Besides, the amount of tweets sharing words from *like* and *love* is quite frequent. This relation can be considered as a kind of "hyperbolic mark". In some cases the sarcastic intention is achieved by using hyperbole.

FIGURE C.4. Sarcastic tweets from the TwPtáček2014 dataset distributed across the basic emotions defined in SentiSense.



C.3 Discovering Communities of Words in Sarcastic Tweets

The objective of this experiment is to go beyond the representation of the polarity or emotional category as nodes in a graph. Instead, the nodes represented each word in a lexicon. For this experiment, two lexica were exploited: AFINN and EmoLex. Figure C.5 shows the representation as a graph of this experiment.

FIGURE C.5. Graph representation of the words in a lexicon and a set of tweets.

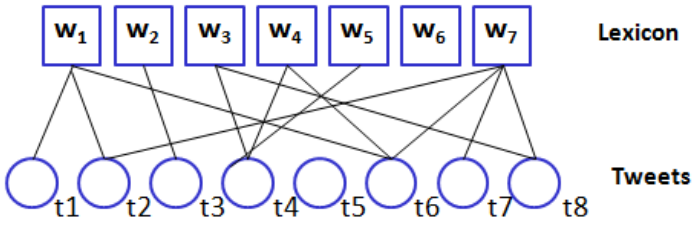


Figure C.6 shows the most salient communities found by using AFINN. As expected, the majority of them are formed by positive words, i.e., there are some frequent positive terms in sarcastic utterances [147]. Moreover, as can be noticed there are more positive than negatives communities. The word *love* emerges as one of the most evident communities. Regarding the negative terms, *no* is the one forming the most salient community.

FIGURE C.6. Graph representation of the relations between the sentiment words in AFINN and the sarcastic tweets in TwPtáček2014 dataset

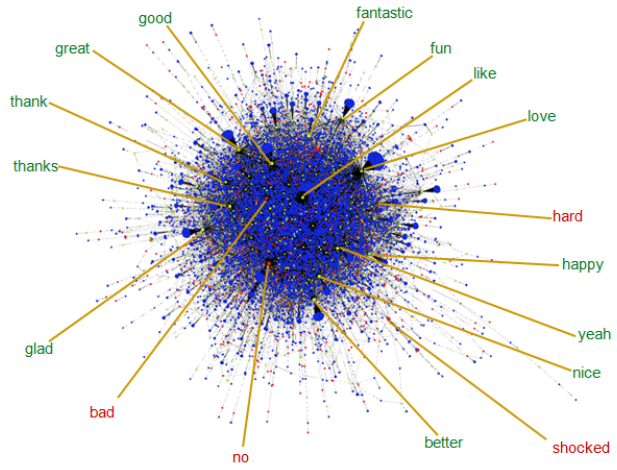
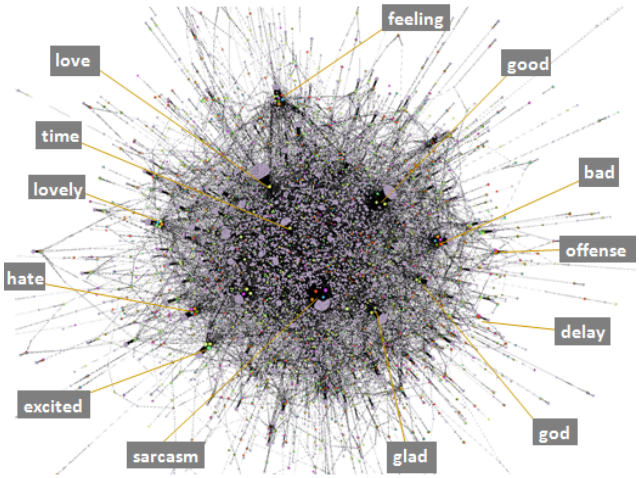


Figure C.7 shows the graph representation of the relations between the words in EmoLex and the sarcastic tweets. As can be noticed, the communities in this chart are smaller and are harder to be identified than those from sentiment words. The most salient community is the one of the word *love* (defined in the *joy* category). Besides, some of the communities are formed by more than one emotional category simultaneously; this is due to some words in EmoLex are included in various emotional categories. For what concerns to the "polarity" of the communities formed by words, the majority of them could be classified as positive, in a similar way than in the experiment with AFINN.

FIGURE C.7. Graph representation of the relations between the emotion-related words in EmoLex and the sarcastic tweets in TwPtáček2014 dataset



Appendix D

A Qualitative Analysis of Affective Resources

TABLE D.1. Matching coefficient results between sentiment resources and ironic tweets.

		TwReyes2013	**TwIronyBarbieri2014**	**TwMohammad2015**
AFINN	POS	486	479	111
	NEG	719	760	183
HL	POS	634	597	107
	NEG	1,200	1,223	225
GI	POS	502	485	68
	NEG	482	480	92
EWN	POS	740	725	213
	NEG	525	529	138
SUBJ_STRONG	POS	446	428	81
	NEG	672	683	121
SUBJ_WEAK	POS	359	360	82
	NEG	405	399	93
EmoLex	POS	973	971	172
	NEG	1,098	1,118	231

TABLE D.2. Matching coefficient results between sentiment resources and sarcastic tweets.

		TwSarcasmBarbieri2014	**TwRiloff2013**	**TwPtáček2014**
AFINN	POS	446	112	562
	NEG	594	105	800
HL	POS	585	100	764
	NEG	907	125	1,309
GI	POS	428	52	600
	NEG	366	53	537
EWN	POS	622	159	797
	NEG	448	105	582
SUBJ_STRONG	POS	391	73	523
	NEG	482	63	739
SUBJ_WEAK	POS	333	60	420
	NEG	330	56	433
EmoLex	POS	830	138	1,082
	NEG	852	123	1,198

Bibliography

[1] Mohammad-Ali Abbasi, Sun-Ki Chai, Huan Liu, and Kiran Sagoo. Real-World Behavior Analysis Through a Social Media Lens. In *Proceedings of the 5th International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction*, SBP'12, pages 18–26, Berlin, Heidelberg, 2012. Springer-Verlag.

[2] Rob Abbott, Marilyn Walker, Pranav Anand, Jean E. Fox Tree, Robeson Bowmani, and Joseph King. How Can You Say Such Things?!?: Recognizing Disagreement in Informal Political Argument. In *Proceedings of the Workshop on Languages in Social Media*, LSM '11, pages 2–11, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[3] Gavin Abercrombie and Dirk Hovy. Putting Sarcasm Detection into Context: The Effects of Class Imbalance and Manual Labelling on Supervised Machine Classification of Twitter Conversations. In *Proceedings of the ACL 2016 Student Research Workshop*, pages 107–113, Berlin, Germany, August 2016. Association for Computational Linguistics.

[4] Laura Alba-Juez. Irony and the Other Off-record Strategies within Politeness Theory. *A Journal of English and American Studies*, 16:13–24, 1995.

[5] Laura Alba-Juez and Salvatore Attardo. The Evaluative Palette of Verbal Irony. In Geoff Thompson and Laura Alba-Juez, editors, *Evaluation in Context*, pages 93–116. John Benjamins Publishing Company, Amsterdam/Philadelphia, 2014.

[6] Arseni Anisimovich. Self-evaluating Workflow for Language-independent Sentiment Analysis. In *Proceedings of the First Italian Conference on Computational Linguistics (CLiC-it 2014) & the Fourth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian EVALITA 2014*, pages 108–111, 2014.

[7] Magda B. Arnold. *Emotion and Personality*, volume 1. Columbia University Press, New York, US, 1960.

[8] Giuseppe Attardi, Valerio Basile, Cristina Bosco, Tommaso Caselli, Felice Dell'Orletta, Simonetta Montemagni, Viviana Patti, Maria Simi, and Rachele Sprugnoli. State of the Art Language Technologies for Italian: The EVALITA 2014 Perspective. *Journal of Intelligenza Artificiale*, 9(1):43–61, 2015.

[9] Salvatore Attardo. Irony. *Encyclopedia of Language & Linguistics*, 6:26–28, 2006.

[10] Salvatore Attardo. Irony as relevant inappropriateness. In H. Colston and R. Gibbs, editors, *Irony in language and thought: A cognitive science reader*, pages 135–172. Lawrence Erlbaum, 2007.

[11] Salvatore Attardo, Jodi Eisterhold, Jennifer Hay, and Isabella Poggi. Multimodal Markers of Irony and Sarcasm. *Humor*, 16(2):243–260, 2003.

[12] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 2200–2204, Valletta, Malta, May 2010. European Language Resources Association (ELRA).

[13] David Bamman and Noah A. Smith. Contextualized Sarcasm Detection on Twitter. In *Proceedings of the Ninth International Conference on Web and Social Media, ICWSM 2015*, pages 574–577, Oxford, UK, 2015. AAAI.

[14] Francesco Barbieri, Valerio Basile, Danilo Croce, Malvina Nissim, Nicole Novielli, and Viviana Patti. Overview of the Evalita 2016 SENTIment POLarity Classification Task. In *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016), Napoli, Italy, December 5-7, 2016.*, volume 1749 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2016.

[15] Francesco Barbieri, Ronzano Francesco, and Saggion Horacio. Relying on Intrinsic Word Features to Characterize Subjectivity, Polarity and Irony of Tweets. In *Proceedings of the First Italian Conference on Computational Linguistics (CLiC-it 2014) & the Fourth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian EVALITA 2014*, pages 104–107, 2014.

[16] Francesco Barbieri, Francesco Ronzano, and Horacio Saggion. UPF-taln: SemEval 2015 tasks 10 and 11. Sentiment analysis of literal and figurative language in Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, pages 704–708, Denver, Colorado, 2015. Association for Computational Linguistics.

[17] Francesco Barbieri and Horacio Saggion. Modelling Irony in Twitter. In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, Gothenburg, Sweden, 2014.

[18] Francesco Barbieri, Horacio Saggion, and Francesco Ronzano. Modelling Sarcasm in Twitter, a Novel Approach. In *Proceedings of the 5th Workshop*

on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 50–58, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics.

[19] Pierpaolo Basile and Nicole Novielli. UNIBA at EVALITA2014-SENTIPOLC Task: Predicting tweet sentiment polarity combining micro-blogging, lexicon and semantic features. In *Proceedings of the First Italian Conference on Computational Linguistics (CLiC-it 2014) & the Fourth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian EVALITA 2014*, pages 58–63, 2014.

[20] Valerio Basile, Andrea Bolioli, Malvina Nissim, Viviana Patti, and Paolo Rosso. Overview of the Evalita 2014 SENTIment POLarity classification task. In *Proceedings of the First Italian Conference on Computational Linguistics (CLiC-it 2014) & the Fourth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian EVALITA 2014*, pages 50–57, 2014.

[21] Valerio Basile and Malvina Nissim. Sentiment Analysis on Italian Tweets. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 100–107. Association for Computational Linguistics, June 2013.

[22] Roberto Basili, Alessandro Lenci, and Bernardo Magnini, editors. *Proceedings of the First Italian Conference on Computational Linguistics & the Fourth International Workshop EVALITA 2014*. Pisa University Press, Pisa, Italia, December 2014.

[23] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. Gephi: An Open Source Software for Exploring and Manipulating Networks. In *International AAAI Conference on Weblogs and Social Media*, 2009.

[24] Johan Bollen, Huina Mao, and Xiao-Jun Zeng. Twitter Mood Predicts the Stock Market. *CoRR*, abs/1010.3003, 2010.

[25] Cristina Bosco, Viviana Patti, and Andrea Bolioli. Developing Corpora for Sentiment Analysis: The Case of Irony and Senti-TUT. *IEEE Intelligent Systems*, 28(2):55–63, 2013.

[26] Andrea Bowes and Albert Katz. When Sarcasm Stings. *Discourse Processes: A Multidisciplinary Journal*, 48(4):215–236, 2011.

[27] Margaret M Bradley and Peter J Lang. Affective Norms for English Words (ANEW): Instruction manual and affective ratings. Technical report, Center for Research in Psychophysiology, University of Florida, Gainesville, Florida, 1999.

[28] Robert L Brown. The Pragmatics of Verbal Irony. *Language use and the uses of language*, pages 111–127, 1980.

[29] Konstantin Buschmeier, Philipp Cimiano, and Roman Klinger. An Impact Analysis of Features in a Classification Approach to Irony Detection in Product

Reviews. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 42–49, June 2014.

[30] Rafael A. Calvo and Sunghwan Mac Kim. Emotions in text: Dimensional and Categorical Models. *Computational Intelligence*, 29(3):527–543, 2013.

[31] Erik Cambria and Amir Hussain. *Sentic Computing: A Common-sense-based Framework for Concept-level Sentiment Analysis*, volume 1. Springer, 2015.

[32] Erik Cambria, Andrew Livingstone, and Amir Hussain. The Hourglass of Emotions. In *Cognitive Behavioural Systems*, volume 7403 of *Lecture Notes in Computer Science*, pages 144–157. Springer, Berlin Heidelberg, 2012.

[33] Erik Cambria, Daniel Olsher, and Dheeraj Rajagopal. SenticNet 3: A Common and Common-Sense Knowledge Base for Cognition-Driven Sentiment Analysis. In *Proceedings of AAAI Conference on Artificial Intelligence*, pages 1515–1521, Québec, Canada, 2014. AAAI.

[34] Jorge Carrillo de Albornoz, Laura Plaza, and Pablo Gervás. SentiSense: An Easily Scalable Concept-based Affective Lexicon for Sentiment Analysis. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3562–3567, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA).

[35] Paula Carvalho, Luís Sarmento, Mário J. Silva, and Eugénio de Oliveira. Clues for Detecting Irony in User-generated Contents: Oh...!! it's "so easy" ;-). In *Proceedings of the 1st International Conference on Information Knowledge Management Workshop on Topic-sentiment Analysis for Mass Opinion*, pages 53–56, 2009.

[36] Giuseppe Castellucci, Danilo Croce, Diego De Cao, and Roberto Basili. A Multiple Kernel Approach for Twitter Sentiment Analysis in Italian. In *Proceedings of the First Italian Conference on Computational Linguistics (CLiC-it 2014) & the Fourth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian EVALITA 2014*, pages 98–103, 2014.

[37] Yoonjung Choi and Janyce Wiebe. +/-EffectWordNet: Sense-level Lexicon Acquisition for Opinion Inference. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1181–1191, Doha, Qatar, October 2014. Association for Computational Linguistics.

[38] Joanne Cohen-Katz, Susan D Wiley, Terry Capuano, Debra M Baker, and Shauna Shapiro. The Effects of Mindfulness-based Stress Reduction on Nurse Stress and Burnout, Part II: A Quantitative and Qualitative Study. *Holistic nursing practice*, 19(1):26–35, 2005.

[39] Herbert L. Colston. "not good" is "bad", but "not bad" is not "good": An analysis of three accounts of negation asymmetry. *Discourse Processes*, 28(3):237–256, 1999.

[40] Carmen Curcó. Irony: Negation, Echo and Metarepresentation. *Lingua*, 110(4):257 – 280, 2000.

[41] Dmitry Davidov, Oren Tsur, and Ari Rappoport. Semi-supervised Recognition of Sarcastic Sentences in Twitter and Amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, CoNLL '10, pages 107–116, Uppsala, Sweden, 2010. Association for Computational Linguistics.

[42] Rodolfo Delmonte. ITGETARUNS a Linguistic Rule-based System for Pragmatic Text Processing. In *Proceedings of the First Italian Conference on Computational Linguistics (CLiC-it 2014) & the Fourth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian EVALITA 2014*, pages 64–69, 2014.

[43] Shelly Dews, Joan Kaplan, and Ellen Winner. Why Not Say it Directly? The Social Functions of Irony. *Discourse Processes*, 19(3):347–367, 1995.

[44] Marta Dynel. Linguistic Approaches to (Non) Humorous Irony. *Humor - International Journal of Humor Research*, 27(6):537–550, 2014.

[45] Paul Ekman. An Argument for Basic Emotions. *Cognition and Emotion*, 6(3-4):169–200, 1992.

[46] Elisabetta Fersini, Enza Messina, and Federico Alberto Pozzi. Subjectivity, Polarity and Irony Detection: A Multi-layer Approach. In *Proceedings of the First Italian Conference on Computational Linguistics (CLiC-it 2014) & the Fourth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian EVALITA 2014*, pages 70–74, 2014.

[47] Elisabetta Fersini, Federico Alberto Pozzi, and Enza Messina. Detecting Irony and Sarcasm in Microblogs: The Role of Expressive Signals and Ensemble Classifiers. In *2015 IEEE International Conference on Data Science and Advanced Analytics, (DSAA 2015)*, pages 1–8, Paris, France, 2015. IEEE Xplore Digital Library.

[48] Elena Filatova. Irony and Sarcasm: Corpus Generation and Analysis Using Crowdsourcing. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 392–398, Istanbul, May 2012. European Language Resources Association (ELRA).

[49] Ruth Filik, Christian Mark Hunter, and Hartmut Leuthold. When language gets emotional: Irony and the embodiment of affect in discourse. *Acta Psychologica*, 156:114 – 125, 2015.

[50] Tamar Fraenkel and Yaacov Schul. The Meaning of Negated Adjectives. *Intercultural Pragmatics*, 5(4):517–540, 2008.

[51] Aniruddha Ghosh, Guofu Li, Tony Veale, Paolo Rosso, Ekaterina Shutova, John Barnden, and Antonio Reyes. SemEval-2015 Task 11: Sentiment Analysis of Figurative Language in Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, pages 470–478, Denver, Colorado, 2015. Association for Computational Linguistics.

[52] Aniruddha Ghosh and Tony Veale. Fracking Sarcasm using Neural Network. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 161–169, San Diego, California, June 2016. Association for Computational Linguistics.

[53] Debanjan Ghosh, Weiwei Guo, and Smaranda Muresan. Sarcastic or Not: Word Embeddings to Predict the Literal or Sarcastic Meaning of Words. In Lluís Márquez, Chris Callison-Burch, Jian Su, Daniele Pighin, and Yuval Marton, editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, (EMNLP 2015)*, pages 1003–1012. Association for Computational Linguistics, 2015.

[54] Raymond W. Gibbs. Irony in Talk Among Friends. *Metaphor and Symbol*, 15(1-2):5–27, 2000.

[55] Raymond W. Gibbs and Herbert L. Colston, editors. *Irony in Language and Thought*. Routledge (Taylor and Francis), New York, 2007.

[56] Raymond W. Gibbs Jr and Jennifer O'Brien. Psychological Aspects of Irony Understanding. *Journal of Pragmatics*, 16(6):523–530, 1991.

[57] Mayte Giménez, Ferran Pla, and Lluís-F. Hurtado. ELiRF: A SVM Approach for SA tasks in Twitter at SemEval-2015. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, pages 574–581, Denver, Colorado, 2015. Association for Computational Linguistics.

[58] Rachel Giora and Salvatore Attardo. Irony. *Encyclopedia of Humor Studies, Thousand Oaks, CA: SAGE*, pages 397–401, 2014.

[59] Rachel Giora, Ari Drucker, Ofer Fein, and Itamar Mendelson. Default Sarcastic Interpretations: On the Priority of Nonsalient Interpretations. *Discourse Processes*, 52(3):173–200, 2015.

[60] Rachel Giora and Ofer Fein. Irony: Context and Salience. *Metaphor and Symbol*, 14(4):241–257, 1999.

[61] Rachel Giora, Shir Givoni, Vered Heruti, and Ofer Fein. The role of defaultness in affecting pleasure: The optimal innovation hypothesis revisited. *Metaphor and Symbol*, 32(1):1–18, 2017.

[62] Rachel Giora, Shir Givonin, and Ofer Feian. Defaultness Reigns: The Case of Sarcasm. *Metaphor and Symbol*, 30(4):290–313, 2015.

[63] Rachel Giora, Elad Livnat, Ofer Fein, Anat Barnea, Rakefet Zeiman, and Iddo Berger. Negation Generates Nonliteral Interpretations by Default. *Metaphor and Symbol*, 28(2):89–115, 2013.

[64] Shir Givoni, Rachel Giora, and Dafna Berberbest. How Speakers Alert Addresses to Multiple Meanings. *Journal of Pragmatics*, 48:29–40, 2013.

[65] Roberto González-Ibáñez, Smaranda Muresan, and Nina Wacholder. Identifying Sarcasm in Twitter: A Closer Look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, HLT '11, pages 581–586, Portland, Oregon, 2011. Association for Computational Linguistics.

[66] H. P. Grice. Logic and Conversation. In P. Cole and J. L. Morgan, editors, *Syntax and Semantics: Vol. 3: Speech Acts*, pages 41–58. Academic Press, San Diego, CA, 1975.

[67] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA Data Mining Software: An Update. *SIGKDD Explor. Newsl.*, 11(1):10–18, 2009.

[68] Melanie Harris and Penny M. Pexman. Children's Perceptions of the Social Functions of Verbal Irony. *Discourse Processes*, 36(3):147–165, 2003.

[69] Shigehiro Haruki. Jocularity in Irony and Humor: A Cognitive-to-affective Process. *Osaka Literary Review*, 39:17–34, 2000.

[70] Cynthia Van Hee, Els Lefever, and Véronique Hoste. Monday mornings are my fave :) #not Exploring the Automatic Recognition of Irony in English tweets. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 2730–2739, 2016.

[71] Delia Irazú Hernández Farías, Viviana Patti, and Paolo Rosso. Irony Detection in Twitter: The Role of Affective Content. *ACM Trans. Internet Technol.*, 16(3):19:1–19:24, 2016.

[72] Delia Irazú Hernández Farías and Paolo Rosso. Irony, Sarcasm, and Sentiment Analysis. Chapter 7. In Federico A. Pozzi, Elisabetta Fersini, Enza Messina, and Bing Liu, editors, *Sentiment Analysis in Social Networks*, pages 113–127. Morgan Kaufmann, 2016.

[73] Delia Irazú Hernández Farías, Emilio Sulis, Viviana Patti, Giancarlo Ruffo, and Cristina Bosco. ValenTo: Sentiment analysis of figurative language tweets with irony and sarcasm. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, pages 694–698, Denver, Colorado, 2015. Association for Computational Linguistics.

[74] Irazú Hernández Farías, José-Miguel Benedí, and Paolo Rosso. Applying Basic Features from Sentiment Analysis for Automatic Irony Detection. In Roberto Paredes, Jaime S. Cardoso, and Xosé M. Pardo, editors, *Pattern Recognition and Image Analysis*, volume 9117 of *Lecture Notes in Computer Science*, pages 337–344. Springer International Publishing, Santiago de Compostela, Spain, 2015.

[75] Irazú Hernandez-Farias, Davide Buscaldi, and Belém Priego-Sánchez. IRADABE: Adapting English Lexicons to the Italian Sentiment Polarity Classification Task. In *Proceedings of the First Italian Conference on Computational Linguistics (CLiC-it 2014) & the Fourth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian EVALITA 2014*, pages 75–81, 2014.

[76] Minqing Hu and Bing Liu. Mining and Summarizing Customer Reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 168–177, Seattle, WA, USA, 2004. ACM.

[77] Mathieu Jacomy, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. *PLoS ONE*, 9(6):1–12, 06 2014.

[78] Aditya Joshi, Pushpak Bhattacharyya, and Mark James Carman. Automatic Sarcasm Detection: A Survey. *CoRR*, abs/1602.03426, 2016.

[79] Aditya Joshi, Vinita Sharma, and Pushpak Bhattacharyya. Harnessing Context Incongruity for Sarcasm Detection. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 757–762, Beijing, China, July 2015. Association for Computational Linguistics.

[80] Aditya Joshi, Vaibhav Tripathi, Kevin Patel, Pushpak Bhattacharyya, and Mark James Carman. Are Word Embedding-based Features Useful for Sarcasm Detection? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November, 2016*, pages 1006–1011, 2016.

[81] Jihen Karoui, Farah Benamara, Véronique Moriceau, Nathalie Aussenac-Gilles, and Lamia Hadrich-Belguith. Towards a Contextual Pragmatic Model to Detect Irony in Tweets. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 644–650. Association for Computational Linguistics, July 2015.

[82] Jihen Karoui, Farah Benamara, Veronique Moriceau, Viviana Patti, Cristina Bosco, and Nathalie Aussenac-Gilles. Exploring the Impact of Pragmatic Phenomena on Irony Detection in Tweets: A Multilingual Corpus Study. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, Valencia, Spain, April 2017. Association for Computational Linguistics.

[83] Anupam Khattri, Aditya Joshi, Pushpak Bhattacharyya, and Mark Carman. Your Sentiment Precedes You: Using an Author's Historical Tweets to Predict Sarcasm. In *Proceedings of the 6th Workshop on Computational Approaches*

to Subjectivity, Sentiment and Social Media Analysis, pages 25–30, Lisboa, Portugal, September 2015. Association for Computational Linguistics.

[84] Maria Khokhlova, Viviana Patti, and Paolo Rosso. Distinguishing between Irony and Sarcasm in Social Media Texts: Linguistic Observations. In *Proc. of ISMW FRUCT*, pages 1–6. IEEE Xplore, 2016.

[85] Helga Kotthoff. Gender and Joking: On the Complexities of Women's Image Politics in Humorous Narratives. *Journal of Pragmatics*, 32(1):55 – 80, 2000.

[86] Roger J. Kreuz and Gina M. Caucci. Lexical Influences on the Perception of Sarcasm. In *Proceedings of the Workshop on Computational Approaches to Figurative Language*, FigLanguages '07, pages 1–4, Rochester, New York, 2007. Association for Computational Linguistics.

[87] Roger J. Kreuz and Richard M. Roberts. The Empirical Study of Figurative Language in Literature. *Poetics*, 22(1):151–169, 1993.

[88] Sachi Kumon-Nakamura and Sam Glucksberg. How About Another Piece of Pie: The Allusional Pretense Theory of Discourse Irony. *Journal of Experimental Psychology: General*, 124(1):3, 1995.

[89] Florian Kunneman, Christine Liebrecht, Margot van Mulken, and Antal van den Bosch. Signaling Sarcasm: From Hyperbole to Hashtag . *Information Processing & Management*, 51(4):500 – 509, 2015.

[90] Christopher Lee and Albert Katz. The Differential Role of Ridicule in Sarcasm and Irony. *Metaphor and Symbol*, 13(1):1–15, 1998.

[91] John S. Leggitt and Raymond W. Gibbs. Emotional Reactions to Verbal Irony. *Discourse Processes*, 29(1):1–24, 2000.

[92] Christine Liebrecht, Florian Kunneman, and Antal Van den Bosch. The Perfect Solution for Detecting Sarcasm in Tweets #not. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 29–37, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

[93] David C. Littman and Jacob L. Mey. The Nature of Irony: Toward a Computational Model of Irony. *Journal of Pragmatics*, 15(2):131 – 151, 1991.

[94] Bing Liu. Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies*, 5:1–167, 2012.

[95] Peng Liu, Wei Chen, Gaoyan Ou, Tengjiao Wang, Dongqing Yang, and Kai Lei. Sarcasm Detection in Social Media Based on Imbalanced Classification. In Feifei Li, Guoliang Li, Seung-won Hwang, Bin Yao, and Zhenjie Zhang, editors, *Proceedings of the Web-Age Information Management: 15th International Conference*, pages 459–471, Macau, China, 2014. Springer International Publishing.

[96] Joan Lucariello. Situational Irony: A Concept of Events Gone Awry. *Journal of Experimental Psychology: General*, 123(2):129–145, 2014.

[97] Stephanie Lukin and Marilyn Walker. Really? Well. Apparently Bootstrapping Improves the Performance of Sarcasm and Nastiness Classifiers for Online Dialogue. In *Proceedings of the Workshop on Language Analysis in Social Media*, pages 30–40, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

[98] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA, 1999.

[99] Diana Maynard and Mark Greenwood. Who cares about Sarcastic Tweets? Investigating the Impact of Sarcasm on Sentiment Analysis. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4238–4243, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA).

[100] Skye McDonald. Neuropsychological Studies of Sarcasm. In H. Colston and R. Gibbs, editors, *Irony in language and thought: A cognitive science reader*, pages 217–230. Lawrence Erlbaum, 2007.

[101] Saif Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets. In *Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013)*, USA, June 2013.

[102] Saif M. Mohammad. Sentiment Analysis: Detecting Valence, Emotions, and Other Affectual States from Text. In Herb Meiselman, editor, *Emotion Measurement*. Elsevier, 2016.

[103] Saif M. Mohammad and Peter D. Turney. Crowdsourcing a Word–Emotion Association Lexicon. *Computational Intelligence*, 29(3):436–465, 2013.

[104] Saif M. Mohammad, Xiaodan Zhu, Svetlana Kiritchenko, and Joel Martin. Sentiment, Emotion, Purpose, and Style in Electoral Tweets. *Information Processing & Management*, 51(4):480 – 499, 2015.

[105] Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. Semeval-2016 Task 4: Sentiment Analysis in Twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1–18, San Diego, California, June 2016. Association for Computational Linguistics.

[106] Preslav Nakov, Torsten Zesch, Daniel Cer, and David Jurgens, editors. *Proceedings of the 9th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, Denver, Colorado, June 2015.

[107] Finn Årup Nielsen. A New ANEW: Evaluation of a Word List for Sentiment Analysis in Microblogs. In *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages*, volume 718 of *CEUR*

Workshop Proceedings, pages 93–98, Heraklion, Crete, Greece, 2011. CEUR-WS.org.

[108] Debora Nozza, Elisabetta Fersini, and Enza Messina. Unsupervised Irony Detection: A Probabilistic Model with Word Embeddings. In *Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, pages 68–76, 2016.

[109] Canberk Özdemir and Sabine Bergler. CLaC-SentiPipe: SemEval2015 Subtasks 10 B,E, and Task 11. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, pages 479–485, Denver, Colorado, 2015. Association for Computational Linguistics.

[110] Canberk Özdemir and Sabine Bergler. A comparative study of different sentiment lexica for sentiment analysis of tweets. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*. INCOMA Ltd. Shoumen, 2015.

[111] W. Gerrod Parrot. *Emotions in Social Psychology: Essential Readings*. Psychology Press., Philadelphia, 2001.

[112] Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. WordNet::Similarity: Measuring the Relatedness of Concepts. In *Proceedings of the 9th National Conference on Artificial Intelligence*, pages 1024–1025. Association for Computational Linguistics, 2004.

[113] James W. Pennebaker, Martha E. Francis, and Roger J. Booth. Linguistic Inquiry and Word Count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, 71, 2001.

[114] Robert Plutchik. The Nature of Emotions. *American Scientist*, 89(4):344–350, 2001.

[115] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, and Prateek Vij. A Deeper Look into Sarcastic Tweets Using Deep Convolutional Neural Networks. *CoRR*, abs/1610.08815, 2016.

[116] Soujanya Poria, Alexander Gelbukh, Erik Cambria, Amir Hussain, and Guang-Bin Huang. EmoSenticSpace: A Novel Framework for Affective Common-sense Reasoning. *Knowledge-Based Systems*, 69:108–123, 2014.

[117] Soujanya Poria, Alexander Gelbukh, Erik Cambria, Peipei Yang, Amir Hussain, and Tariq S Durrani. Merging SenticNet and WordNet-Affect emotion lists for Sentiment Analysis. In *Proceedings of 11th International Conference on Signal Processing (ICSP)*, volume 2, pages 1251–1255. IEEE, 2012.

[118] Soujanya Poria, Alexander Gelbukh, Amir Hussain, Newton Howard, Dipankar Das, and Sivaji Bandyopadhyay. Enhanced SenticNet with Affective Labels for Concept-Based Opinion Mining. *IEEE Intelligent Systems*, 28(2):31–38, 2013.

[119] Soujanya Poria, Alexander F. Gelbukh, Erik Cambria, Dipankar Das, and Sivaji Bandyopadhyay. Enriching SenticNet Polarity Scores through Semi-Supervised Fuzzy Clustering. In Jilles Vreeken, Charles Ling, Mohammed Javeed Zaki, Arno Siebes, Jeffrey Xu Yu, Bart Goethals, Geoffrey I. Webb, and Xindong Wu, editors, *ICDM Workshops*, pages 709–716. IEEE Computer Society, 2012.

[120] Tomáš Ptáček, Ivan Habernal, and Jun Hong. Sarcasm Detection on Czech and English Twitter. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics*, pages 213–223, Dublin, Ireland, August 2014. Dublin City University and Association for Computational Linguistics.

[121] Ashwin Rajadesingan, Reza Zafarani, and Huan Liu. Sarcasm Detection on Twitter: A Behavioral Modeling Approach. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, pages 97–106, 2015.

[122] Francisco Rangel and Paolo Rosso. On the Impact of Emotions on Author Profiling. *Information Processing & Management*, 52(1):73 – 92, 2016. Emotion and Sentiment in Social and Expressive Media.

[123] Francisco Rangel, Paolo Rosso, Irina Chugur, Martin Potthast, Martin Trenkmann, Benno Stein, Ben Verhoeven, and Walter Daelemans. Overview of the 2nd Author Profiling Task at PAN 2014. In L. Cappellato, N. Ferro, M. Halvey, and W. Kraaij, editors, *CLEF 2014 Labs and Workshops, Notebook Papers*, volume 1180, pages 898–927. CEUR-WS.org, 2014.

[124] Antonio Reyes and Paolo Rosso. On the Difficulty of Automatically Detecting Irony: Beyond a Simple Case of Negation. *Knowl. Inf. Syst.*, 40(3):595–614, 2014.

[125] Antonio Reyes, Paolo Rosso, and Davide Buscaldi. From Humor Recognition to Irony Detection: The Figurative Language of Social Media . *Data & Knowledge Engineering*, 74:1 – 12, 2012. Applications of Natural Language to Information Systems.

[126] Antonio Reyes, Paolo Rosso, and Tony Veale. A Multidimensional Approach for Detecting Irony in Twitter. *Language Resources and Evaluation*, 47(1):239–268, 2013.

[127] Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. Sarcasm as Contrast between a Positive Sentiment and Negative Situation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, (EMNLP 2013)*, pages 704–714, Seattle, Washington, USA, 2013. Association for Computational Linguistics.

[128] Sara Rosenthal, Preslav Nakov, Svetlana Kiritchenko, Saif Mohammad, Alan Ritter, and Veselin Stoyanov. SemEval-2015 Task 10: Sentiment Analysis in Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, pages 451–463, Denver, Colorado, 2015. Association for Computational Linguistics.

[129] Sara Rosenthal, Alan Ritter, Preslav Nakov, and Veselin Stoyanov. SemEval-2014 Task 9: Sentiment Analysis in Twitter. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 73–80, Dublin, Ireland, August 2014. Association for Computational Linguistics and Dublin City University.

[130] Sheldon M. Ross. *Introduction to Probability and Statistics for Engineers and Scientists*. Academic Press, 2014.

[131] James A. Russell and Albert Mehrabian. Evidence for a Three-factor Theory of Emotions. *Journal of Research in Personality*, 11(3):273–294, 1977.

[132] Rossano Schifanella, Paloma de Juan, Joel Tetreault, and LiangLiang Cao. Detecting Sarcasm in Multimodal Social Platforms. In *Proceedings of the 2016 ACM on Multimedia Conference*, MM '16, pages 1136–1145, New York, NY, USA, 2016. ACM.

[133] Simone Shamay-Tsoory, Rachel Tomer, B. D. Berger, Dorith Goldsher, and Judith Aharon-Peretz. Impaired "Affective Theory of Mind" is Associated with Right Ventromedial Prefrontal Damage. *Cogn. Behav. Neurol.*, 18(1):55–67, 2005.

[134] Dan Sperber and Deirdre Wilson. *Relevance: Communication and Cognition*. Harvard University Press, Cambridge, MA, USA, 1986.

[135] Jacopo Staiano and Marco Guerini. DepecheMood: A Lexicon for Emotion Analysis from Crowd-Annotated News. *CoRR*, abs/1405.1605, 2014.

[136] Philip J. Stone and Earl B. Hunt. A Computer Approach to Content Analysis: Studies Using the General Inquirer System. In *Proceedings of the May 21-23, 1963, Spring Joint Computer Conference*, AFIPS '63 (Spring), pages 241–256, New York, NY, USA, 1963. ACM.

[137] Emilio Sulis, Delia Irazú Hernández Farías, Paolo Rosso, Viviana Patti, and Giancarlo Ruffo. Figurative messages and affect in Twitter: Differences between #irony, #sarcasm and #not. *Knowledge-Based Systems*, 108:132–143, 2016.

[138] Maite Taboada and Jack Grieve. Analyzing Appraisal Automatically. In *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*, pages 158–161, Stanford, US, 2004. AAAI.

[139] Yi-jie Tang and Hsin-Hsi Chen. Chinese Irony Corpus Construction and Ironic Structure Analysis. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics*, pages 1269–1278, Dublin, Ireland, 2014. Association for Computational Linguistics.

[140] Peter D. Turney. Thumbs up or thumbs down?: Semantic Orientation Applied to Unsupervised Classification of Reviews. In *Proceedings of the 40th annual meeting on Association for Computational Linguistics*, pages 417–424. Association for Computational Linguistics, 2002.

[141] Akira Utsumi. Verbal Irony as Implicit Display of Ironic Environment: Distinguishing Ironic Utterances from Nonirony. *Journal of Pragmatics*, 32(12):1777 – 1806, 2000.

[142] Cynthia Van Hee, Els Lefever, and Veronique Hoste. LT3: Sentiment Analysis of Figurative Tweets: Piece of Cake #notreally. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, pages 684–688, Denver, Colorado, 2015. Association for Computational Linguistics.

[143] Tony Veale and Yanfen Hao. Detecting Ironic Intent in Creative Comparisons. In *Proceedings of the 19th European Conference on Artificial Intelligence*, pages 765–770, Amsterdam, The Netherlands, 2010. IOS Press.

[144] Byron C. Wallace. Computational Irony: A Survey and New Perspectives. *Artificial Intelligence Review*, 43(4):467–483, 2015.

[145] Byron C. Wallace, Do Kook Choe, and Eugene Charniak. Sparse, Contextually Informed Models for Irony Detection: Exploiting User Communities, Entities and Sentiment. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1035–1044, Beijing, China, July 2015. Association for Computational Linguistics.

[146] Byron C. Wallace, Do Kook Choe, Laura Kertz, and Eugene Charniak. Humans Require Context to Infer Ironic Intent (so Computers Probably do, too). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 512–516, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[147] Angela P. Wang. #irony or #sarcasm — A Quantitative and Qualitative Study Based on Twitter. In *Proceedings of the 27th Pacific Asia Conference on Language, Information, and Computation*, pages 349–356. National Chengchi University, 2013.

[148] Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. Norms of Valence, Arousal, and Dominance for 13,915 English lemmas. *Behavior Research Methods*, 45(4):1191–1207, 2013.

[149] Juanita M. Whalen, Penny M. Pexman, J. Alastair Gill, and Scott Nowson. Verbal Irony Use in Personal Blogs. *Behaviour & Information Technology*, 32(6):560–569, 2013.

[150] Cynthia Whissell. Using the Revised Dictionary of Affect in Language to Quantify the Emotional Undertones of Samples of Natural Languages. *Psychological Reports*, 2(105):509–521, 2009.

[151] Deirdre Wilson and Dan Sperber. On Verbal Irony. *Lingua*, 87(1-2):53–76, 1992.

[152] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing Contextual Polarity in Phrase-level Sentiment Analysis. In *Proceedings of the Conference*

on Human Language Technology and Empirical Methods in Natural Language Processing, HLT '05, pages 347–354, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[153] Alecia Wolf. Emotional Expression Online: Gender Differences in Emoticon Use. *CyberPsychology & Behavior*, 3(5):827–833, 2000.

[154] Zhibiao Wu and Martha Palmer. Verbs Semantics and Lexical Selection. In *Proceedings of the 32Nd Annual Meeting on Association for Computational Linguistics*, ACL '94, pages 133–138, Stroudsburg, PA, USA, 1994. Association for Computational Linguistics.

[155] Hongzhi Xu, Enrico Santus, Anna Laszlo, and Chu-Ren Huang. LLT-PolyU: Identifying Sentiment Intensity in Ironic Tweets. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, pages 673–678, Denver, Colorado, 2015. Association for Computational Linguistics.