The Unbearable Hurtfulness of Sarcasm

Simona Frenda[a,b] (simona.frenda@unito.it),
Alessandra Teresa Cignarella[a,b] (alessandrateresa.cignarella@unito.it),
Valerio Basile[a] (valerio.basile@unito.it),
Cristina Bosco[a] (cristina.bosco@unito.it),
Viviana Patti[a] (viviana.patti@unito.it)
and Paolo Rosso[b] (prosso@dsic.upv.es)

[a] Department of Computer Science, Università degli Studi di Torino, Italy
[b] PRHLT Research Center, Universitat Politècnica de València, Spain

Corresponding Author:
Simona Frenda
Department of Computer Science, Università degli Studi di Torino, Italy
Email: simona.frenda@unito.it

The Unbearable Hurtfulness of Sarcasm

Simona Frenda[a,b*], Alessandra Teresa Cignarella[a,b], Valerio Basile[a], Cristina Bosco[a], Viviana Patti[a] and Paolo Rosso[b]

[a] *Department of Computer Science, Università degli Studi di Torino, Italy*
[b] *PRHLT Research Center, Universitat Politècnica de València, Spain*

*Corresponding author.

Email addresses: simona.frenda@unito.it (Simona Frenda), alessandrateresa.cignarella@unito.it (Alessandra Teresa Cignarella), valerio.basile@unito.it (Valerio Basile), cristina.bosco@unito.it (Cristina Bosco), viviana.patti@unito.it (Viviana Patti), prosso@dsic.upv.es (Paolo Rosso)

Abstract

In the last decade, the need to detect automatically irony to correctly recognize the sentiment and hate speech involved in online texts increased the investigation on humorous figures of speech in NLP. The slight boundaries among various types of irony lead to think of *irony* as a linguistic phenomenon that covers sarcasm, satire, humor and parody joined by their trend to create a secondary or opposite meaning to the literal one expressed in the message. Although this commonality, in literature *sarcasm* is defined as a type of irony more aggressive with the intent to mock or scorn a victim without excluding the possibility to amuse. The aggressive tone and the intent of contempt suggest that sarcasm involves some peculiarities that make it a suitable type of irony to disguise negative messages. To investigate these peculiarities of sarcasm, we examined the dataset of the IronITA shared task. It consists of Italian tweets about controversial social issues, such as immigration, politics and other more general topics. Each tweet is annotated as *ironic* and *non-ironic*, and, at a deeper level, as *sarcastic* and *non-sarcastic*. Qualitative and quantitative analyses of the dataset showed how sarcasm tends to be expressed with hurtful language revealing the aggressive intention with which the author targets the victim. While irony is characterized by being offensive in hateful context and, in general, moved by negative emotions. For a better understanding of the impact of hurtful and affective language on the detection of irony and sarcasm, we proposed a transformer-based system, called AlBERToIS, combining pre-trained AlBERTo model with linguistic features. This approach obtained the best performances on irony and sarcasm detection on the IronITA dataset.

Keywords: Affective Language, Hurtful Language, Irony Detection, Sarcasm Detection, Linguistic Features, AlBERTo.

1. Introduction

Rhetorical literature converges towards a common definition of irony as semantic inversion, that is to say the opposite of what is believed and what really is (Garavelli, 1997). As figure of speech that overturns the literal meaning of the message, irony is used for various purposes: mocking or making fun of someone or something, underlining the paradox of a situation, or echoing the violation of a norm with dismissive attitude (Wilson & Sperber, 2012). These purposes of irony could manifest in more explicit manner through jocularity, sarcasm, parody and humor.

Especially focusing on *sarcasm*, dictionaries[1] and linguistic literature (Du Marsais et al., 1981; Gibbs, 2000; Attardo, 2007; Dynel, 2014) define it as a type of irony more offensive with the intent to convey scorn or mock a clear victim (Bowes & Katz, 2011). According to Lee & Katz (1998), the hearers perceive the aggressive tone as the feature that perfectly distinguishes this figure of speech, as in: *Non bastano i nostri falsi invalidi! Manteniamo anche falsi invalidi stranieri! https://t.co/WZGgbTP1FR*[2]. The aggressive tone and the intent to scorn a specific target suggest that sarcasm could involve some characteristics of abusive language, especially in delicate contexts such as in the discussion online about sensitive social issues. Therefore, sarcasm detectors need to take into account also the hateful aspects that could be implied in the expression of sarcasm. This peculiarity could make it more suitable to disguise negative messages as well as *hate speech*[3]. Indeed, some works on hate speech detec-

[1] `https://www.merriam-webster.com/dictionary/sarcasm`

[2] "Our fake invalids are not enough! We also support false foreign invalids! https://t.co/WZGgbTP1F"

[3] In accordance with the most common definitions (Nockleby, 2000; Schmidt & Wiegand, 2017; Davidson et al., 2017), with the expression "Hate Speech", we refer to any utterance "*that is abusive, insulting, intimidating, harassing, and/or incites to violence, hatred, or discrimination. It is directed against people on the basis of their race, ethnic origin, religion, gender, age, physical condition, disability, sexual orientation, political conviction, and so forth.*"(Erjavec & Kovačič, 2012).

tion (Nobata et al., 2016; Frenda et al., 2018, 2020) showed how the presence of sarcasm could affect the performance of systems. This intuition about the appropriateness of sarcasm to express contempt and to subtly offend the victim, without excluding the possibility of having fun, leads to three important questions:

RQ1 Is it possible to characterize sarcasm and irony in informal contexts, such as Twitter, in terms of different features on affective and hurtful language use?

RQ2 Can knowledge about hurtful and affective language be helpful in addressing the task of sarcasm and irony detection?

RQ3 Can transformer-based architectures benefit from the addition of linguistic features related to hatred and emotions?

In order to answer these questions, we choose the IronITA dataset as a case study. To the best of our knowledge, this dataset, released in occasion of the IronITA shared task (Cignarella et al., 2018b) organized in 2018 within the framework of the Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA), is the first dataset collecting tweets annotated, firstly, as *ironic*/*non-ironic* and, in a finer-grained layer, only if the tweets are *ironic*, as *sarcastic*/*non-sarcastic*. In addition, the interest in analyzing this dataset lies in its composition. In fact, a part of this dataset, is extracted from a corpus of hate speech against minorities such as roma community, immigrants and Muslims. That allows to investigate properly the role of various dimensions of hate such as hate speech, aggressiveness, offensiveness and stereotype in sarcastic and ironic tweets. This issue is a novelty in the study of the affect involved in sarcastic and ironic expressions. Indeed, previous studies principally focused on the role of emotions in the expression of irony especially in English (Hernández Farías et al., 2016; Sridhar et al., 2017; Kanwar et al., 2019; Babanejad et al., 2020), leaving hostile language almost unexplored. The other part of the dataset is extracted from more general corpora reporting information

about linguistic categories such as rhetorical and pragmatic elements related to irony. This information helps us to identify some linguistic peculiarities that characterizes sarcasm with respect to other types of irony, contributing to the discussion on analogies and differences between irony and sarcasm (Wang, 2013; Sulis et al., 2016).

In the perspective of designing dedicated systems able to correctly recognize irony and sarcasm online and overcome the difficulties encountered by the IronITA's participating systems in the detection of sarcasm, we performed a qualitative and quantitative error analysis on the predictions provided by the three best ranked systems in both the subtasks of the contest. In the framework of IronITA, participants were indeed asked to distinguish, firstly, *ironic* from *non-ironic* tweets (Task A) and, secondly, *sarcastic* tweets from both the *non-ironic* and *ironic non-sarcastic* ones (Task B).

On the basis of these previous analyses, we developed a transformer-based system composed of the AlBERTo model (the BERT language understanding model for the Italian language) (Polignano et al., 2019) pre-trained on tweets and informed by stylistic, syntactic, and semantic features. Specifically, BERT (Devlin et al., 2018) stands for Bidirectional Encoder Representations from Transformers, and it is designed to pre-train deep bidirectional representations on a large dataset of unlabeled texts creating deeper language models. The proposed system, called AlBERToIS, integrates the knowledge of AlBERTo language model with the weights of linguistic features that aim to introduce stylistic, syntactic, and semantic information. A correct identification of irony and sarcasm is, indeed, crucial for the development of systems aware of irony and sarcasm, especially in hate speech detection (Nobata et al., 2016; Frenda, 2018) and sentiment analysis. In sentiment analysis, for example, Farias & Rosso (2017) underlined a significant gap between the performance of sentiment analysis systems on non-figurative content and the performance reached on sarcastic content.

Therefore, the principal contributions of our work could be summarized as below:

- study hatred, emotions and linguistic markers in the expression of irony and, in particular, of sarcasm, delineating its peculiarities;
- investigate what makes irony and sarcasm hard to detect, by examining the misclassified tweets of the best ranked participating teams at IronITA shared task;
- disclose the impact of features related to hatred and emotions on transformer-based architectures to detect these figurative devices;
- obtain the best performances for irony and sarcasm detection in Italian.

The following sections focus on defining the related works (Section 2), the description of IronITA shared task and dataset (Section 3), the analysis of the dataset (Section 4), the error analysis of the best performing systems (Section 5), the proposed system and features (Section 6), the experiments and the obtained results (Section 7), the discussion of findings (Section 8) and, finally, we conclude by defining future work (Section 9).

2. Related Works

The detection of irony and sarcasm is gaining more and more interest in scientific communities and companies. In fact, it proves to be relevant in Sentiment Analysis for recognizing correctly the opinion or orientation of users about a specific subject (product, service, topic, issue, person, organization, or event) (Reyes & Rosso, 2012) as well as on Hate Speech detection (Nobata et al., 2016; Frenda, 2018). Many have been the recent shared tasks on irony/sarcasm detection and figurative language in general: SENTIPOLC 2014 and 2016 subtask *Irony detection in Italian tweets* (Basile et al., 2014; Barbieri et al., 2016), DEFT2017-Task2 *Figurative language detection* in French tweets (Benamara et al., 2017), SemEval2018-Task3 *Irony detection in English tweets* (Van Hee et al., 2018a) that asked participants to distinguish also among four categories of irony (irony by clash, situational irony, other verbal irony and non-irony), IroSvA2019 *Irony Detection in Spanish Variants* (Ortega-Bueno et al., 2019)

where also the context was provided to understand to what ironic comments referred to, ALTA2019 shared task on *Sarcastic Target Identification* (Molla & Joshi, 2019), and, more recently, FigLang2020-Task2 *Sarcasm Detection* (Ghosh et al., 2020) focused on sarcastic texts identification in English conversations on Twitter and Reddit. Differently from the mentioned shared tasks, IronITA shared Task at EVALITA 2018 proposed a deeper analysis of ironic text asking participants to recognize, firstly, if a tweet is ironic or not, and, secondly, to discriminate sarcastic tweets from non-sarcastic ones in Italian language. Its purpose was to investigate the possibility to approach these two different linguistic phenomena, although complementary, and analyze their characteristics in hateful and general context.

Irony and Sarcasm Characteristics The works that investigated the typical characteristics of irony and sarcasm are not that many. From a more linguistic and cognitive perspective, sarcasm could be distinguished from other forms of irony for involving negative evaluation against the victim (Alba-Juez & Attardo, 2014). The negativity of sarcasm covered by apparent positivity is found out in qualitative and quantitative analyses carried out on English self-tagged tweets by Wang (2013). This study reveals that users aware to be sarcastic tend to use more positive words in tweets labeled with #sarcasm to sugar-coat the more aggressive meaning. Similar findings are reported by Sulis et al. (2016). The authors examined qualitatively and quantitatively the dataset released by the organizers of SemEval2015-Task11 (Ghosh et al., 2015) containing English self-annotated tweets that include specific hashtags (e.g. #not, #sarcasm and #irony). In particular, they investigated the impact of sentiment, emotions, various affective lexica, tweets length and punctuation in this dataset, revealing some important differences especially between tweets containing #irony and #sarcasm: tweets with #irony are especially related to negative sentiment and emotions (anger, disgust, fear and sadness), differently from the ones with #sarcasm that contain words expressing mainly joy, anticipation, trust, surprise and positive sentiment; polarity reversal (Bosco et al., 2013) is more relevant in tweets with #sarcasm, showing a particular shift from literal positive to

real negative polarity; tweets with #irony prove to be more creative and implicit than the ones with #sarcasm. These observations are supported also at computational and multilingual level. In various English datasets of tweets, Hernández Farías et al. (2016) demonstrate the discriminating power of negative sentiment in irony detection, and of positive sentiment (and words expressing "love") in sarcasm detection and the relevance of features such as the presence of mentions and the length of tweets especially in sarcasm detection. In Spanish tweets, Frenda & Patti (2019) show that in three different variants of Spanish the most significant emotions for irony detection are principally negative (anger, fear, disgust and sadness). The present work aims to analyze emotions and linguistic characteristics proper of sarcasm and irony also in the Italian language, poorly explored until now.

Irony and Sarcasm Detection As in many NLP tasks, deep learning-based approaches reach very competitive results also in irony and sarcasm detection. Especially transformers models such as BERT and its variants (Potamias et al., 2020), have been largely employed in the last competition in FigLang2020-Task2 confirming the importance for an automatic system of having extended language knowledge. In other works, the authors studied aspects such as a potential incongruity information within ironic or sarcastic messages, as well as language ambiguities (Reyes et al., 2012; Barbieri et al., 2015; Naseem et al., 2020), semantic contrast (Pan et al., 2020), sentiment discordance (Zhang et al., 2019), emotional shift (Agrawal et al., 2020), dissonance between positive sentiment and negative situations (Riloff et al., 2013) and contrast between the orientation of a specific community (e.g. forum) and the published message (Wallace et al., 2015; Joshi et al., 2015). In this work, in line with the computational novelties, we propose an approach that combines language model knowledge and linguistic features in a deep learning architecture.

Emotions and Hatred Another aspect previously investigated in irony and sarcasm detection is the contribution of emotional and sentiment information in various languages (Calvo et al., 2020; Hernández Farías et al., 2016) and in different contexts (Chauhan et al., 2020; Babanejad et al., 2020). With respect

to hate information, the intuition about the use of sarcasm to disguise hateful and offensive utterances was preliminary investigated in Justo et al. (2014); Nobata et al. (2016) and Frenda (2018). In Justo et al. (2014) the authors showed differences and analogies in sarcasm and nastiness detection. In particular, they observed that length and linguistic information are relevant especially for sarcasm detection, whereas semantic information improve results for both tasks. However, specific lexical cues seem to work really well for nastiness detection demonstrating that nasty opinions tend to be expressed by users overtly and without ambiguities. Nobata et al. (2016) and Frenda (2018) showed instead how abusive contents sometimes are disguised by sarcasm making hate speech more subtle and, thus, more difficult to be recognized. Nevertheless, the intuitive correlation between sarcasm and abusive language is poorly discussed and experimented (De Mattei et al., 2018; Frenda & Patti, 2019). In this framework, the analyses and experiments on the IronITA dataset will contribute to reveal the role played by hatred and emotions in ironic and sarcastic tweets.

3. IronITA

The IronITA contest provides a framework especially suitable for investigating our research questions. On the one side, the multi-source composition of the IronITA dataset allows us to perform statistical analyses able to disclose specific characteristics of sarcasm related, firstly, to hostility that move sarcastic expressions, and, secondly, to rhetorical and pragmatic elements that distinguish sarcasm from other types of irony. On another side, the IronITA shared task gives the opportunity to reveal the difficulties of existing systems of irony and sarcasm detection in Italian tweets, and provides a frame where the approach we propose can be tested and compared.

The IronITA shared task, as described by the organizers in Cignarella et al. (2018b), consists of two tasks of detection of ironic and sarcastic texts from Twitter.

- **Task A** is a coarse-grained binary classification where systems have to

predict whether a tweet is ironic or not.

- **Task B** is a multi-class classification where systems have to predict one out of the three following labels: i) sarcasm, ii) irony not categorized as sarcasm (i.e., other kinds of verbal irony or descriptions of situational irony which do not show the characteristics of sarcasm), and iii) non-irony.

irony	sarcasm	text
0	0	*Le critiche al governo monti da parte di chi ci ha portato sull'orlo del fallimento sono intollerabili.* → The criticisms towards Monti's government by those who have brought us to the verge of bankruptcy are just intolerable.
1	0	*@matteorenzi le risorse della scuola pubblica alle private... Questa è la buona scuola!* → @matteorenzi resources of public schools to private ones... This is the good school!
1	1	*@Bisbeticah @NmargheNiki stiamo consegnando l'Italia ai stranieri..... GrazieStato* → @Bisbeticah @NmargheNiki we're handing Italy over to foreigners..... ThankYouState

Table 1: Examples from IronITA

As defined by the majority of rhetorical literature, sarcasm is conceived in the IronITA annotation schema as a type of verbal irony, as the crudest and sharpest form of irony moved by negativity and intended to criticize and hurt the target without excluding the possibility of having fun[4]. Table 1 reports the possible combination of labels annotated on the tweets of the IronITA dataset.

The IronITA dataset is, to the best of our knowledge, the only dataset collecting tweets annotated for *irony* and, in a finer-grain, for *sarcasm*. The tweets of this dataset come from different sources: Hate Speech Corpus (HSC) (Sanguinetti et al., 2018b) and TWITTIRÒ corpus (Cignarella et al., 2018a), composed of tweets from LaBuonaScuola (TW-BS) (Stranisci et al., 2016), Sentipolc (TW-SENTIPOLC), Spinoza (TW-SPINO) (Barbieri et al., 2016). Only in the test set, some tweets have been added from the TWITA collection of tweets (Barbieri

[4]A detailed explanation of the used schema of annotation is presented here: `http://di.unito.it/guidelines`.

et al., 2016). The distribution of tweets according to the various source datasets is shown in Table 2. This multi-source composition allows to bring out statistically significant characteristics of ironic and sarcastic tweets related to hateful and general contexts. To this purpose, we retrieved the original labels of the HSC and TWITTIRÒ corpora, and extended the IronITA annotation:

	training set				**test set**				
	ironic	**non-iro**	**sarc**	**iro non-sarc**	**ironic**	**non-iro**	**sarc**	**iro non-sarc**	**total**
TW-BS	467	646	173	294	111	161	51	60	
TW-SPINO	342	0	126	216	73	0	32	41	
TW-SENTIPOLC	461	625	143	318	0	0	0	0	3,109
TWITA	0	0	0	0	67	156	28	39	
HSC	753	683	471	282	184	120	105	79	1,740
total	**3,977**				**872**				**4,849**

Table 2: Distribution of Tweets According to the Source

- **HSC** annotation, as described in Sanguinetti et al. (2018b), consists of various labels referring to *dimensions of hate*, such as aggressiveness (`agg`), offensiveness (`off`)[5], stereotype (`stereotype`) and hate speech (`hs`);

- **TWITTIRÒ** schema has three levels of annotation as described in Cignarella et al. (2018a). In particular, we applied two levels of annotations related to *linguistic characteristics*:

    1 **Contradiction Type**[6]: If the tweet is ironic, one can individuate the type of contradiction that activates irony (Giora et al., 2015). Actually, irony is often expressed through a contradiction that could occur between two lexicalized clues (such as opposite terms or propositions) within the sentence (`explicit`), or between an internal lexi-

[5]Although the original annotation established a range of strength (no, weak and strong) for aggressiveness and offensiveness, in our work we took into account only the presence of these phenomena.

[6]In accordance with the TWITTIRÒ schema of annotation, the labels of level 2 and 3 are applied only to ironic tweets (see Table 5).

calized cue and an external pragmatic context echoed in the sentence (`implicit`). For example: (1) *Vedo che c'è molta disinformazione sul referendum del 17 maggio. [@MisterDonnie13]*[7] (`ironic` and `non-sarcastic`), and (2) *Trovato l'ispiratore delle ricette del governo Monti: Bisogna prendere il denaro dove si trova. Presso i poveri.... http:t.cosh54bMiN*[8] (`ironic` and `sarcastic`)

2 **Linguistic Categories**: If the tweet is ironic and a type of contradiction has been individuated, the final level of annotation specifies the linguistic elements creating the contradiction, and, therefore, the ironic expression. The figures of speech and pragmatic clues relative to implicit and explicit contradiction are listed in Table 3.

Categories	**Label**	**Definition**
Analogy$^{\text{Both}}$	`an`	Analogy covers figures of speech, such as metaphor, analogy, simile and similarity, used to compare different ontological concepts or domains.
Hyperbole$^{\text{Both}}$	`hyp`	Hyperbole is used to emphasize or exaggerate something.
Euphemism$^{\text{Both}}$	`euph`	Euphemism allows to reduce the duress of an idea or a fact to soften the reality.
Rhetorical Question$^{\text{Both}}$	`r_q`	Rhetorical question is used to make a point about an issue rather than to elicit an answer.
Context Shift$^{\text{Expl}}$	`c_s`	Context shift involves a sudden change of topic or frame, such as the use of exaggerated politeness in an inappropriate situation.
False Assertion$^{\text{Impl}}$	`f_a`	False assertion assumes the assertion of a unreal fact or declaration.
Oxymoron/Paradox$^{\text{Expl}}$	`o/p`	Oxymoron and Paradox concern an explicit lexical (antonyms) and pragmatic contradiction.
Other$^{\text{Both}}$	`other`	"Other" category covers humor and situational irony, where the contradiction involves events and not the use of words.

Table 3: Linguistic Categories

The preexisting annotations of the source corpora HSC and TWITTIRÒ in the tweets of IronITA covered only the data of the training set. In order to perform the analysis on the whole IronITA dataset, we extended these two fine-grained annotations, also, to the tweets of the test set, following the respective

[7] The referendum was indeed on April 17th, 2016: "I see there's a lot of disinformation on the referendum of May 17th. [@MisterDonnie13]"

[8] "Found the inspirer of Monti's government's recipes: One must take money where it lies. From poor people...."

guidelines[9]. At the end of the process of extension, the tweets of the IronITA dataset are labeled with IronITA, HSC and TWITTIRÒ schema of annotation as shown in Tables 4 and 5.

irony	sarcasm	hs	agg	off	stereotype	text
0	0	yes	yes	no	yes	*@repubblicait tutto tempo danaro e sacrificio umano sprecato senza eliminazione fisica dei talebani e dei radicali musulmani è tutto inutile* → @repubblicait all the time money and human sacrifice wasted without purge of talibans and muslim radicals it's all useless
1	0	no	yes	yes	yes	*Gentili proprietari dei resort alle #maldive... accogliete il profugo dall'Italia per dieci giorni. #profughi #esiamonoi #notengodinero* → Respectable owners of the resorts at the #maldives... welcome the refugee from Italy for ten days. #refugees #andit'sus #notengodinero
1	1	yes	yes	no	no	*Dai ragazzi, è Natale! Portiamo un po' di calore al campo nomadi. Io penso alla benzina, voi portate i fiammiferi?* → Come on guys, it's Christmas! Let's bring some warmth to the nomad's camp. I'll take care of the gasoline, you'll bring the matches?

Table 4: Examples from HSC Source for Each Possible Combination of IronITA

irony	sarcasm	level 2	level 3	text
0	0	0	0	*Come fare in modo che gli studenti sperimentino l'entusiasmo della scoperta scientifica? #AmgenTeach http://t.co/fCDpQAIyNB #labuonascuola* → How to do make students experiment the enthusiasm of scientific discovery? #AmgenTeach http://t.co/fCDpQAIyNB #labuonascuola
1	0	explicit	an	*Crolla la borsa di Shanghai. Ora bisogna risollevarla senza muovere le altre. [@blogstark]* → Shanghai's stock market crashes. Now we should pick it up, but without moving the others. [@blogstark]
1	1	implicit	im:f_a	*E comunque @matteorenzi alla lezione di sillabazione de #labuonascuola era assente http://t.co/bEpiecpfx3* → Anyway @matteorenzi was absent at the lesson of the #labuonascuola regarding hyphenation http://t.co/bEpiecpfx3

Table 5: Examples from TWITTIRÒ Source for Each Possible Combination of IronITA

[9]For the tweets coming from HSC, the schema of annotation in `http://di.unito.it/hsc`.; and for the data coming from the other sources related to political or more general topics (TW-BS, TW-SPINO, TW-SENTIPOLC and TWITA) TWITTIRÒ schema of annotation in `http://di.unito.it/twittiro`.

4. Analysis of the Dataset

Taking into account the extended annotation in the IronITA dataset, we applied a statistical analysis to study the association between irony/sarcasm and the dimensions of hate/linguistic characteristics interpreted as nominal variables of a population. In particular, we computed: χ^2 test of independence that, by means of the interpretation of p-value, gives information on the existence or not of significant relations between nominal variables; and Yule's Q to indicate if the association between two binary variables is positive (values close to 1), negative (values close to -1), or null (values close to 0).

		hs	agg	off	stereotype
Task A	irony	0.00/0.22	0.00/0.35	**0.00/0.45**	0.00/0.37
Task B	sarcasm	0.00/0.37	**0.00/0.59**	0.01/0.23	0.02/0.19
	non-sarcastic irony	0.65/-0.05	0.28/-0.11	**0.00/0.32**	0.00/0.26

Table 6: p-Values/Yule's Q Values for Dimensions of Hate

Dimensions of Hate Table 6 shows the p-values for the χ^2 test of independence and the Yule's Q values of the possible associations between *irony*[10]/*non-sarcastic irony*/*sarcasm* and each dimension of hate considered in the HSC. We remember that to reject the null hypothesis (hypothesis that the variables are independent) of the χ^2 test of independence, the p-value should be minor than the significance level set by convention to 0.05. To calculate the p-value, we consider a degree of freedom based on the number of observations.

Looking at Table 6, we notice that: *sarcasm* is related to some degree on all the dimensions of hate and, especially, on aggressiveness, whereas *non-sarcastic irony* and, in general, *irony* are strongly associated with offensiveness, showing that, in presence of specific targets in the discussed issues, irony could be also offensive (*@LaGabbiaTw Mi hanno insegnato che non tutti i musulmani sono*

[10]This label includes all types of irony.

terroristi ma il 99% dei terroristi nel mondo sono musulmani.[11]). These results confirm our initial intuitions: sarcasm appears more aggressive than other types of irony and, considering the high values for hate speech, could perfectly fit to disguise negative messages.

	an	euph	ex:c_s	ex:o/p	im:f_a	hyp	other	r_q
Explicit sarcasm	0.28/0.08	**0.02/0.25**	0.00/-0.28	0.01/0.17		0.28/-0.14	*0.00/-0.30*	0.24/0.09
Implicit sarcasm	0.18/-0.23	0.92/0.03			**0.01/0.31**	*0.23/-0.54*	0.47/-0.11	0.31/-0.24

Table 7: p-Values/Yule's Q Values for Linguistic Characteristics

Linguistic Characteristics Since the TWITTIRÒ schema of annotation is only focused on ironic texts, the set of observations is composed of *sarcastic* and *ironic non-sarcastic* tweets only. In this context, we could calculate statistical values for *sarcasm* and infer possible association for *non-sarcastic irony* by the sign of the Yule's Q values. Therefore, in Table 7, positive Q values refer to associations with *sarcasm* (maximum value in bold) and negative Q values to associations with *non-sarcastic irony* (minimum value in italic); while p-values indicate in general the existence or not of a dependence. Table 7 reports significant signals of association, on the one side, between *non-sarcastic irony* and `other` category (containing, indeed, other types of irony, such as situational irony) in the explicit class, and with hyperbole (`hyp`) in the implicit one; and, on another side, between *sarcasm* and euphemism (`euph`) (maybe used to mask the negativity of messages) in the explicit class, and with false assertion (`f_a`) in the implicit one. Moreover, looking at the distribution of the *sarcastic*/*ironic non-sarcastic* tweets with respect to the explicit/implicit type of contradiction, we noted that sarcastic tweets tend to be more explicit than non-sarcastic ones (tweets 1 and 2 in Section 3 are a clear example of that). A similar trend was

[11] "@LaGabbiaTw They have taught me that not all Muslims are terrorists, but 99 percent of the world's terrorists are Muslims."

observed also in English by Sulis et al. (2016). In general, although the lower distribution of sarcastic texts in the IronITA dataset (see Table 2), the statistical measures helped to delineate some typical features of irony and sarcasm.

5. Error analysis

Correctly detecting irony and sarcasm, especially in social media texts, is a challenging task. First of all, it is difficult to create a ground-truth dataset where to train and test systems because of the subjectivity intrinsically involved in the interpretation of these figurative language devices. Indeed, although irony and sarcasm are well defined in literature, their interpretation may be strongly influenced by cultural background and contextual knowledge Basile (2020). For example, for the annotation of sarcasm in the IronITA dataset, the annotators achieved a moderate final inter-annotator agreement of Fleiss' $\kappa = 0.56$ for the tweets belonging to the TWITTIRÒ corpus and $\kappa = 0.52$ for the data coming from the HSC (Cignarella et al., 2018b) [12].

In addition, as seen in previous sections, ironic and sarcastic texts involve various and complex elements that could be explicit or implicit in the text, or that could concern the intentions or affects of an author, making hard their detection. In IronITA shared task, this difficulty seems to meaningfully increase in sarcasm detection, due probably to the scarcity of sarcastic tweets and to the lack of dedicated systems.

Results in IronITA Shared Task The participants were invited to participate at both tasks (Task A and Task B) or at Task A only, submitting runs constrained or unconstrained (when additional data are used for training phase).

[12] As described in Cignarella et al. (2018b), the annotation was organized in two steps. Firstly, the dataset was split in two halves and two couples of Italian native speakers (specialized in figurative language) annotated sarcasm in each half. Secondly, to solve the disagreement, the couple previously involved in the annotation of the first half of the dataset produced a new annotation for the tweets in disagreement of the second portion of the dataset and vice versa. Then, the cases where the disagreement persisted (131 tweets) have been discarded as too ambiguous to be classified.

In total the participating teams were 7, and only 4 of them submitted runs also to Task B (Cignarella et al., 2018b). No matter the challenging task and the lower amount of linguistic resources available for the Italian language, the systems obtained high results in Task A.

team name	run	rank	F1-score		
			non-iro	iro	macro
ItaliaNLP	1	1	0.707	**0.754**	**0.731**
UNIBA	1	3	0.689	0.730	0.710
X2Check	1	5	**0.708**	0.700	0.704
baseline-random			*0.503*	*0.506*	*0.505*
baseline-mfc			*0.668*	*0.000*	*0.334*

Table 8: Results for Task A

team name	run	rank	F1-score			
			non-iro	iro	sarc	macro
UNITOR	2	1	0.668	**0.447**	0.446	**0.520**
ItaliaNLP	1	3	**0.707**	0.432	0.409	0.516
Aspie96	1	5	0.668	**0.438**	0.289	0.465
baseline-random			*0.503*	*0.266*	*0.242*	*0.337*
baseline-mfc			*0.668*	*0.000*	*0.000*	*0.223*

Table 9: Results for Task B

Looking at Table 8[13], the first ranked system reported the trend to identify correctly ironic messages more than non-ironic ones, and obtained a macro f1-score of 0.731, revealing a performance in line with the results in SemEval2018-Task3 about irony detection in English tweets (Van Hee et al., 2018b). About Task B, we can notice lower f-scores in Table 9 due probably to the difficulty to distinguish sarcasm from other types of irony, and to the scarce amount of

[13]Table 8 and 9 show the results obtained by the three best systems as described in Section 5.1. Unconstrained runs are in grey background.

sarcastic data with respect to the rest (see Table 2). The complete ranking for both tasks is published in Cignarella et al. (2018b). This difficulty of detecting sarcasm makes even more interesting an in-depth error analysis in order to understand whether systems did not detected sarcastic tweets confusing sarcasm with other types of irony, or finding too challenging to recognize it for its peculiar characteristics. To this purpose, in order to study the set of the common predictions (correct and incorrect) of the three best runs for each task, we applied two main types of analyses. Firstly, a qualitative analysis on the common misclassified ironic and sarcastic tweets. Secondly, we deepened the qualitative observations with a quantitative analysis exploiting: the multi-label annotation of the IronITA dataset, and the morphosyntactic information extracted by PoS-tagging and parsing the misclassified ironic/sarcastic tweets with the *UDPipe* pipeline (Straka & Straková, 2017). This analysis helps us to understand which are the difficulties of state-of-the-art systems to detect irony and sarcasm in Italian tweets; it reveals some information about the impact of emotional and hurtful language on the detection of irony and sarcasm; and it leads to define a specific set of features that help to overcome these difficulties.

5.1. Hard and Simple Cases

Since the differences between runs of the same systems are not significant, we considered the predictions of the best run submitted by the teams that obtained the best scores. In particular, we considered:

- for Task A: the first runs of the teams ItaliaNLP, UNIBA and X2Check (unconstrained)
- for Task B: the second run of the team UNITOR (unconstrained) and the first runs of the teams ItaliaNLP and Aspie96.

This choice allowed us to take into account the predictions that were obtained with different approaches (see Table 10). The majority of them used the same system to detect irony and sarcasm, except UNITOR that employed a cas-

cade architecture of classifiers that selected automatically the most distinctive information for each task among a consistent set of features.

Team	Run	Task	Approach
ItaliaNLP (De Mattei et al., 2018)	1	A, B	Multi-task learning approach based on Bidirectional Long Short-Term Memory (biLSTM) networks exploiting the correlation among various related sentiment analysis tasks. They used additional tweets from SENTIPOLC 2016 dataset (Barbieri et al., 2016) (first run) and HaSpeeDe 2018 (Bosco et al., 2018) (second run), in addition to sentiment polarity lexica, semantic and morpho-syntactic features.
UNIBA (Basile & Semeraro, 2018)	1	A	Support Vector Machine (SVM) taking advantage of sentiment information (Basile & Novielli, 2014), unigrams, bigrams, trigrams, microblogging features and word embedding vectors from TWITA (Basile et al., 2018) as semantic representation of tweets and to intercept the usage of words in Twitter context.
X2Check (Di Rosa & Durante, 2018)	1	A	P Principally exploiting n-grams word representation, they built a system based on Multinomial Naive Bayes algorithm trained on additional tweets annotated as ironic from SENTIPOLC 2016.
UNITOR (Santilli et al., 2018)	2	A,B	Cascade of kernel-based SVM classifiers: the first classifier discriminated between *ironic* and *non-ironic* tweets, while the second one distinguished *sarcastic* and *non-sarcastic* tweets. To generalize lexical information of training texts, they created a word embedding using about 10 millions of tweets downloaded in July 2016, and computed the cosine similarity between words and sentence word embedding to capture the unconventional use of a word and PoS-tag. Finally, they used various sizes of characters n-grams, synthetic features, sentiment information for words and PoS-tags extracted by a distributional polarity lexicon built in (Castellucci et al., 2016). Only for the unconstrained run, that reaches the first rank in Task B classification, the team built a specific ironic dataset collecting 6,000 tweets assuming to be ironic on specific hashtags (#irony or #ironia) to get, also, specific words or patterns of ironic texts.
Aspie96 (Giudice, 2018)	1	A, B	Gated Recurrent Units (GRU) exploiting the advantages of character level representation.

Table 10: Best Performimg Systems in IronITA shared task

Collecting the predictions of the best performing systems in the IronITA shared task, we selected the set of *hard cases* (HC henceforth) composed of the common misclassified tweets, and the set of *simple cases* (SC henceforth) composed of the common tweets correctly classified.

Table 11 and 12 show the sizes of HC and SC sets for each task and their

	Hard Cases		Simple Cases	
	ironic	**non-iro**	**ironic**	**non-iro**
NOHSC	18	39	125	153
HSC	10	23	112	48
TOTAL CLASS	28 (6%)	62 (14%)	237 (54%)	201 (46%)
TOTAL CASES	**90**		**438**	

Table 11: Hard and Simple Cases in Task A

	Hard Cases			Simple Cases		
	sarc	**iro non-sarc**	**non-iro**	**sarc**	**iro non-sarc**	**non-iro**
NOHSC	66	0	1	0	91	258
HSC	16	4	1	19	31	83
TOTAL CLASS	82 (38%)	4 (2%)	2 (0.5%)	19 (9%)	122 (56%)	341 (78%)
TOTAL CASES		**88**			**482**	

Table 12: Hard and Simple Cases in Task B

percentage calculated on the total of tweets in the test set for each class. Considering the fact that our interest is in the comprehension of hurtful language that could characterize sarcasm especially in controversial issues, in Tables 11 and 12 we divide the sets of tweets in two principal domains: HSC and NO-HSC. The latter collects tweets coming from TW-BS, TW-SPINO, TW-SENTIPOLC and TWITA and covering general issues not necessarily related to abusive context. Comparing the distribution of HC and SC in Task A and B, we can notice that: ironic tweets are in general correctly identified, whereas sarcastic ones result more difficult to detect; and, looking at the difference between the sets of HSC and NO-HSC in Table 12, sarcastic tweets tend to be identified correctly in hateful context.

Moreover, to measure the impact of the low inter-annotator agreement in the results obtained in the competition on Task B, we observed if the common misclassified tweets by the three best systems in the competition (88 HC in

Table 12) caused also disagreement during the annotation. Among these 88 HC, only 4 tweets were considered hard to interpret even by the annotators. However, during the second phase of the annotation, the disagreement was solved. Considering this low percentage (4.5% of HC), we can state that the low inter-annotator agreement did not affect the results in the competition.

5.2. Qualitative Analysis

Our first step is to examine qualitatively HC carrying out a manual error analysis with the purpose to find stylistic, syntactic and semantic markers that made irony and, especially, sarcasm difficult to identify. Secondly, we deepened these findings with a quantitative analysis. The results of this analysis will lead us to a better feature engineering for the design of our system. It is important to underline that our attention in Task B is focused on understanding if unidentified sarcasm is confused with other types of irony, or is not recognized for its peculiarities. Considering that, our analysis in Task B will concern only *sarcastic* and *ironic non-sarcastic* tweets.

Stylistic Markers refer to those patterns related to the writing style in a social media like Twitter, such as discursive and informal elements. In particular, in ironic/sarcastic HC we noticed a great number of quotation marks, ellipsis and intensifiers (*sempre più*, *150k*, *solo*). Especially sarcastic HC contain also negation markers (*non*, *nemmeno*, *né*) and informal language (such as swear words, dialectal and colloquial expressions).

Syntactic Markers involve phrase types and syntactic coarse-grained classes. In particular, in ironic/sarcastic HC, we noticed a high frequency of: noun phrases that work sometimes as slogan (*Stop profughi*, *città sotto assedio*, *buona scuola o buona propaganda*); adverbial locutions (*altro che*, *bene*, *di certo*) and, especially, discourse connectors with function adversative (*invece*, *ma*), causal (*perché*) or sequential (*prima*, *ora*). A fine-grained morphosyntactic analysis will be described in Section 5.3.1.

Semantic Markers cover elements that could be caught analyzing the meaning of the message. Ironic/sarcastic HC tend to have a surprise effect caused by a

contrast between phrases or sentences within the message (*@MiurSocial "ti aggiorneremo sull'avvio della consultazione" Sto ancora aspettando #labuonascuola*[14]), or by an unexpected answer or solution (*@fattoquotidiano Anche noi abbiamo la nostra via x i rom: quella dei forni della Italsider.*[15]). Another common semantic element is the assertion of false events (*Wojtyla era pronto alle dimissioni. Ma non riusciva a firmarle. [fedgross]*[16]). Sarcastic HC, moreover, involve echoic mentions (*La moglie di Bobo Craxi scippata ad Hammamet. In un commosso ricordo del suocero. [fdecollibus]*[17]) and context shifting (*Frattini pubblica sul sito del ministero le foto delle sue vacanze. La mia preferita è quella dove sta alla scrivania. [stenit]*[18]). All these elements are far from the textual markers and require an extended knowledge of the language, as well as of the world, to be captured. This makes irony and sarcasm detection a real challenging task.

5.3. Quantitative Analysis

At a deeper level, we carried out a more quantitative analysis aimed at identifying specific elements of irony and sarcasm that could make hard their detection. Firstly, we focus on stylistic and syntactic markers examining morphosyntactic information extracted by PoS-tagging and parsing the misclassified ironic and sarcastic tweets. Secondly, we exploit the multi-label annotation of the IronITA dataset to analyze, at a semantic level, the impact of the dimensions of hate on irony and sarcasm detection as well as of rhetorical and pragmatic elements.

[14] "@MiurSocial "we will let you know regarding the start of consultation" I'm still waiting #labuonascuola"

[15] "@fattoquotidiano We too have our own way for romas: the ovens of Italsider"

[16] "Wojtila was ready to write his resignation. But he wasn't able to sign it. [fedgross]"

[17] "The wife of Bobo Craxi mugged in Hammamet. In a moved memory of her father-in-law. [fdecollibus]"

[18] "Frattini posts photos of his vacations on the ministry website. My favorite one is that where he's behind his work-desk. [stenit]"

5.3.1. Morphosyntactic Analysis

We conducted an error analysis investigating the morphosyntactic characteristic of the language used in misclassified tweets, taking advantage of the fact that a portion of the IronITA dataset has been annotated accordingly to the format of *Universal Dependencies*[19] (henceforth UD) (Cignarella et al., 2019). By training the *UDPipe* pipeline on other available Italian treebanks ISDT (Simi et al., 2014), PoSTWITA (Sanguinetti et al., 2018a), and TWITTIRÒ-UD (Cignarella et al., 2019) we easily tokenized, lemmatized, PoS tagged and parsed the remaining tweets that were not released as part of a gold standard in the official UD repository[20] obtaining a full morphosyntactic annotation for the whole IronITA test set.

We proceeded in two steps: firstly we observed the distribution of Part-of-Speech (PoS) tags in the entire test set and compared it with the PoS tags distribution in HC of both tasks, and later we focused only on *ironic* tweets that were wrongly classified as *non-ironic* (28 tweets for Task A) and on *sarcastic* tweets that were wrongly classified as *ironic non-sarcastic* (82 tweets for Task B) (see Table 13). In a following step we applied the same procedure also accordingly to the distribution of dependency relations (see Table 14).

Morphology Observing Table 13, we are able to see how PoS tags are distributed across the test set and examine whether the PoS tags in HC report any significant difference in their distribution. For instance, the high number of `NOUN` PoS tag (3.10% [in red]) in ironic HC suggests that these tweets could contain noun phrases or slogans with ironic meaning not recognized by the systems. On the other hand, it seems that the presence of the `SYM` PoS tag (8.61% [in green]) and of the `X` PoS tag (5.95% and 7.14% [in magenta]) is lower especially in sarcastic HC, suggesting that the tokens with these PoS tags (e.g. foreign words, emojis, hashtags, mentions and URLs) might be good indicators for the detection of sarcasm. Moreover, we can notice a high frequency of `DET` PoS

[19] `https://universaldependencies.org/`.
[20] `http://di.unito.it/uditaliantwittiro`.

	whole test set	all HC				only ironic or sarcastic tweets			
PoS tags	**Entire test set (782 tweets)**	**HC Task A (90 tweet)**	**freq (%)**	**HC Task B (88 tweet)**	**freq (%)**	**HC Task A (28 tweets)**	**freq (%)**	**HC Task B (82 tweets)**	**frequ (%)**
ADJ	816	73	8.95	86	10.54	26	3.19	82	10.05
ADP	1,964	207	10.54	218	11.10	52	2.65	197	10.03
ADV	870	103	11.84	91	10.46	15	1.72	81	9.31
AUX	579	79	13.64	59	10.19	16	2.76	56	9.67
CCONJ	338	41	12.13	34	10.06	6	1.78	28	8.28
DET	1,999	203	10.16	237	11.86	52	2.60	213	**10.66**
INTJ	100	7	7.00	15	15.00	2	2.00	14	**14.00**
NOUN	2,583	288	11.15	275	10.65	80	**3.10**	249	9.64
NUM	172	18	10.47	18	10.47	3	1.74	18	10.47
PRON	900	111	12.33	94	10.44	26	2.89	84	9.33
PROPN	879	56	6.37	92	10.47	17	1.93	81	9.22
PUNCT	2,247	186	8.28	272	12.11	47	2.09	208	9.26
SCONJ	200	19	9.50	22	11.00	2	1.00	17	8.50
SYM	1,557	157	10.08	144	9.25	68	4.37	134	**8.61**
VERB	1,572	185	11.77	166	10.56	48	3.05	148	9.41
X	168	10	**5.95**	12	**7.14**	4	2.38	12	**7.14**
Total	16,944	1,743	10.29	1,835	10.83	464	2.74	1,622	9.57

Table 13: Distribution of PoS Tags in HC

tag (10.66% [in orange]) in sarcastic HC. Accordingly with the UD tagset, DET PoS tag includes quantifiers and various determiners (indefinite, exclamatory, demonstrative and so on). All these elements could be used as intensifiers. Another interesting value is the frequency of INTJ PoS (14.00% [in cyan]), that, as seen in Section 5.2, seems to play an important role in sarcasm detection.

Syntax In the same way, we then calculated the distribution of *dependency relations* (deprels). In Table 14 we illustrate a list of all the dependency relations and their frequency in the three different subsets. With the hyphen "–" we indicate that a dependency relation is not present in a subset. Considering that what we analyze is user-generated content, it is not surprising to see that the most frequent deprel is punct (used 2,245 times [**in bold**], being 13.25% of the total), which stands for punctuation, as its extensive usage in social media platforms is widely attested in literature (Bazzanella, 2011; Sanguinetti et al.,

	whole test set	all HC				only ironic or sarcastic tweets			
Deprels	**Test set (782 tweets)**	**HC Task A (90 tweet)**	**freq (%)**	**HC Task B (88 tweet)**	**freq (%)**	**HC Task A (28 tweets)**	**freq (%)**	**HC Task B (82 tweets)**	**freq (%)**
acl	128	10	7.81	11	8.59	3	2.34	10	7.81
acl:relcl	149	23	**15.44**	14	9.40	5	3.36	13	8.72
advcl	191	24	12.57	16	8.38	4	2.09	13	6.81
advmod	842	96	11.40	87	10.33	14	1.66	77	9.14
amod	682	61	8.94	72	10.56	25	3.67	69	10.12
appos	55	2	**3.64**	1	**1.82**	–	–	1	**1.82**
aux	293	45	15.36	24	8.19	8	2.73	23	7.85
aux:pass	42	6	**14.29**	7	16.67	1	2.38	7	16.67
case	1,760	188	10.68	202	11.48	49	2.78	184	10.45
cc	338	39	11.54	34	10.06	6	1.78	28	8.28
ccomp	114	13	11.40	15	13.16	4	3.51	9	7.89
compound	54	5	9.26	2	3.70	–	–	1	1.85
conj	391	37	9.46	36	9.21	6	1.53	32	8.18
cop	244	28	11.48	28	11.48	7	2.87	26	10.66
csubj	19	2	10.53	1	5.26	–	–	1	5.26
dep	473	34	7.19	19	4.02	21	4.44	18	3.81
det	1,901	194	10.21	224	11.78	51	2.68	201	10.57
det:poss	73	6	8.22	12	16.44	1	1.37	11	15.07
det:predet	22	4	18.18	2	9.09	–	–	1	4.55
discourse	97	9	9.28	14	14.43	2	2.06	13	13.40
discourse:emo	48	8	16.67	8	16.67	2	4.17	9	**18.75**
dislocated	2	–	–	–	–	–	–	–	–
expl	161	11	6.83	23	14.29	3	1.86	18	11.18
expl:impers	17	7	**41.18**	1	5.88	1	5.88	1	5.88
expl:pass	6	1	**16.67**	–	–	–	–	–	–
fixed	38	3	7.89	2	5.26	–	–	2	5.26
flat	18	2	11.11	2	11.11	–	–	2	11.11
flat:foreign	40	1	**2.50**	1	**2.50**	1	**2.50**	1	**2.50**
flat:name	157	8	**5.10**	13	8.28	2	1.27	12	7.64
iobj	110	11	10.00	9	8.18	2	1.82	10	9.09
mark	398	38	9.55	38	9.55	5	1.26	30	7.54
nmod	1,081	99	9.16	102	9.44	34	3.15	96	8.88
nsubj	791	91	11.50	87	11.00	22	2.78	79	9.99
nsubj:pass	48	3	6.25	4	8.33	–	–	4	8.33
nummod	146	16	10.96	14	9.59	3	2.05	14	9.59
obj	791	105	13.27	91	11.50	30	3.79	81	10.24
obl	749	93	12.42	100	13.35	23	3.07	87	11.62
obl:agent	19	–	–	1	5.26	–	–	1	5.26
parataxis	435	49	11.26	74	**17.01**	17	3.91	66	**15.17**
parataxis:appos	1	–	–	–	–	–	–	–	–
parataxis:hashtag	228	26	11.40	22	9.65	15	**6.58**	20	8.77
punct	**2,245**	186	8.29	271	12.07	47	2.09	208	9.27
root	872	90	10.32	88	10.09	28	3.21	82	9.40
vocative	17	2	11.76	1	5.88	–	–	–	–
vocative:mention	487	51	10.47	52	10.68	16	3.29	49	10.06
xcomp	171	16	9.36	10	**5.85**	6	3.51	12	7.02
Total	16,944	1,743	10.29	1,835	10.83	464	2.74	1,622	9.57

Table 14: Distribution of Dependency Relations in HC

2017). For what concerns other deprels in the subset of misclassified tweets of Task A, we notice a distribution that deviates from the standard of the following relations: `acl:relcl` (relative clauses), `aux:pass` (auxiliary verbs in a passive voice construction), `expl:impers` and `expl:pass` (expletive particles), indicating that tweets with these syntactic features tend to be misclassified [in blue]. On the other hand, tweets containing the following deprels, seem to be correctly classified the majority of the times: `appos` (appositional modifiers), `flat:foreign` (foreign words) and `flat:name` (multiword expressions) [in green]. The deprel `discourse:emo` seems to have an unbalanced distribution in Task B, suggesting it might be creating noise and making more difficult the detection of sarcasm (18.75%, $\Delta = 9.18$ deviation from the average distribution) [in red]. The `parataxis` dependency relation has a greater distribution in the misclassified tweets of Task B in both scenarios (all HC: 17.01%, and sarcastic HC: 15.17%), deviating $\Delta = 6.18$ in the first case and $\Delta = 5.6$ in the second [in orange], but presents an average distribution in the two scenarios of Task A. Similarly, the deprel `parataxis:hashtag` presents a $\Delta = 3.84$ with regard to the average distribution in the misclassified tweets of Task A, in the scenario where we look at all the misclassified tweets (6.58%), but then its distribution is around average values in all the other cases [in magenta]. Finally, `xcomp` seems to be less present in the misclassified tweets of Task A (5.85%) [in cyan], presenting a deviation of $\Delta = 4.98$.

5.3.2. Semantic and Pragmatic Analysis

To enrich and reinforce qualitative semantic markers identified in Section 5.2, we examine the percentages of false positive (FP) and negative (FN), and, equally, true positive (TP) and negative (TN) in presence of the dimensions of hate and linguistic characteristics. The percentages are calculated considering the absolute frequency of each dimension of hate/linguistic characteristic in HC and SC and its distribution in test set. Taking into account the low values of HC and SC in both tasks, below we report the most relevant observations.

Hurtful and Affective Language To analyze the impact of hurtful language,

we considered the presence of hate speech, aggressiveness, offensiveness and stereotype in *ironic/non-ironic* and *sarcastic/ironic non-sarcastic* tweets (as shown in Table 15).

	Task A						Task B					
	Test set (304 hsc tweets)						Test set (184 hsc tweets)					
Dimensions of Hate	**ironic (184 tweets)**	**non-ironic (120 tweets)**	**FP (%)**	**FN (%)**	**TP (%)**	**TN (%)**	**sarc (105 tweets)**	**iro non-sarc (79 tweets)**	**FP (%)**	**FN (%)**	**TP (%)**	**TN (%)**
`hs yes`	37	22	27.27	5.40	**70.27**	18.18	26	11	**9.09**	19.23	7.69	36.36
`hs no`	147	98	17.35	5.44	58.50	44.90	79	68	4.41	13.92	**21.52**	39.70
`agg yes`	59	24	**29.17**	6.78	62.71	16.67	44	15	6.67	18.18	15.91	13.34
`agg no`	125	96	16.67	4.80	60.00	45.84	61	64	4.69	13.11	19.67	45.31
`off yes`	61	21	19.05	1.64	65.57	19.04	38	23	8.70	7.89	13.16	26.09
`off no`	123	99	19.19	**7.32**	58.54	44.45	67	56	3.57	**19.40**	20.90	44.64
`stereotype yes`	77	36	19.45	5.19	61.04	25.00	48	29	6.89	10.42	14.58	27.59
`stereotype no`	107	84	19.05	5.61	60.75	**46.43**	57	50	4.00	19.30	21.05	**46.00**

Table 15: Distribution of Dimensions of Hate

In Task A, high percentages of TP in presence of hate speech (70.27%), aggressiveness (62.71%), offensiveness (65.57%) and stereotypes (61.04%) and of TN in non-hateful contexts (respectively 44.90%, 45.84%, 44.45% and 46.43%) suggest that systems tend to correctly classify tweets as ironic when text contains a more hurtful language. Indeed, observing the highest values of FN in both tasks (7.32% in Task A and 19.40% in Task B), we can hypothesize that the lack of offenses could conduct to predict ironic/sarcastic tweet as non-ironic/non-sarcastic, but, conversely, the presence of derogatory speech could increase the FP, as shown in Task A (29.17% and 22.27%) and in Task B (9.09% and 8.70%). Therefore, it appears necessary to balance the information about hateful language given to the system.

In NO-HSC, the highest percentages of false predictions are related to FP cases (12.30%). Analyzing these tweets that the systems tend to predict as ironic, we noticed that are principally characterized by negative emotions as well as rage or frustration. It is clear that negative emotions and a more hurtful language have an impact on the detection of irony and sarcasm.

Rhetorical and Pragmatic Characteristics Since the annotation schema of

TWITTIRÒ focuses only on ironic texts, Table 16 does not report FP and TN values calculated on the negative class for Task A.

	Task A						Task B					
	Test set (568 no-hsc tweets)						Test set (251 no-hsc tweets)					
Linguistic Categories	**ironic (251 tweets)**	**non-ironic (317 tweets)**	**FP (%)**	**FN (%)**	**TP (%)**	**TN (%)**	**sarc (111 tweets)**	**iro non-sarc (140 tweets)**	**FP (%)**	**FN (%)**	**TP (%)**	**TN (%)**
Explicit												
an	14			7.14	50.00		8	6		62.50		50.00
euph	23			**26.09**	34.78		14	9		64.29		**77.78**
ex: c_s	43			2.33	**60.47**		15	28		46.67		50.00
ex: o/p	52			7.69	55.77		30	22		53.33		72.73
hyp	13			7.69	53.85		4	9		75.00		66.67
other	26			3.85	38.46		5	21		**80.00**		76.19
r_q	20			10.00	45.00		13	7		76.92		71.43
Implicit												
euph	3				33.33		1	2		**100.00**		**100.00**
hyp	2				**100.00**			2				**100.00**
im: f_a	25			4.00	68.00		11	14		45.45		35.71
other	21				19.05		9	12		66.67		66.67
r_q	9			**11.11**	55.56		1	8				87.50

Table 16: Distribution of Linguistic Categories

Taking into account the percentages of TP, we can delineate some important linguistic markers in ironic texts that could help irony detection: context shift (60.47%), oxymoron (55.77%) and hyperbole (53.85%). Other more subtle linguistic categories, such as euphemism (*Altro che 'merito', #labuonascuola ha anche profumo di incostituzionalità http://t.co/pfvzeu4T3L #sapevatelo @GildaInsegnanti @ALMCalabria*[21]), and rhetorical question that could be confused as simple question (*Si può fare "buona scuola" senza Geografia? | Orizzonte Scuola http://t.co/cM0ln-O6ceY via @orizzontescuola*[22]) tend to increase the FN values (respectively 26.09% in explicit contradictions and 11.11% in implicit ones).

[21] "What 'merit'? #labuonascuola also stinks as unconstitutional http://t.co/pfvzeu4T3L #sapevatelo @GildaInsegnanti @ALMCalabria"

[22] "Is it possible to have a "good school" without Geography? | Orizzonte Scuola http://t.co/cM0lnO6ceY via @orizzontescuola"

With respect to Task B, since HC are *sarcastic* and in SC are only *ironic non-sarcastic*, Table 16 does not report FP and TP percentages computed respectively on the negative and positive classes. Moreover, in Task B frame, TN represents the *ironic non-sarcastic* texts. In Task B, we can observe that percentages of FN are higher with respect to Task A, probably for the complexity of the task. Examining the FN cases, sarcasm tends to be predicted as *non-sarcastic irony* especially when it contains rhetorical questions (that make difficult the correct identification also in Task A), hyperbole (more related to irony) and situational irony. `other` category, normally observed in *ironic non-sarcastic* texts for its references to specific funny situations, as explained in Wang (2013) could involve also sarcastic situations, even if in a more subtle manner than in ironic ones: *Quando mi dicono: "stai zitta che bevi ancora il latte" io rispondo: "si ma con il cioccolato perché io sono già grande" ahahhahaha*[23] (`non-sarcastic irony`) and *@SteGiannini @davidefaraone @MiurSocial La buona scuola in cui tutti parleranno solo inglese.Come Renzi.Che pena.*[24] (`sarcasm`).

6. Proposed Approach

IronITA shared task suggests a novel computational interpretation of sarcasm detection task as a sub-task of irony detection: if a tweet is ironic it could be sarcastic or not. Therefore, to detect sarcasm we need to recognize before the presence of irony in the text. From this perspective, we adopted a cascade architecture where tweets that were predicted as ironic in Task A are classified as sarcastic and non-sarcastic in Task B. Although we used the same neural network for both tasks, the selected features in each classification task are different. Indeed, computing the χ^2 value for each feature, we are able to observe which feature is more significant for irony and sarcasm detection. We designed

[23] "When they tell me: "shut up since you're still drinking milk" I reply "yes, but with cocoa since I'm already grown up" ahahhahaha"

[24] "@SteGiannini @davidefaraone @MiurSocial The good school in which everyone will speak English.As Renzi.What a shame."

specific stylistic, syntactic and semantic features taking into account the previous observations coming from the error analysis of the IronITA shared tasks and the observation of associations between irony/sarcasm and dimensions of hate/linguistic characteristics.

Our main idea is to converge in an unique system the awareness coming from the learning of a pre-trained language model with the specific knowledge derived from dedicated linguistic features. On the one side, the learning transferred by a language model trained on Italian tweets should help the classifier to be more sensitive to style and semantics of a more informal writing and make the system able to "understand" better unseen cases. On another side, engineered features lead the system to pay attention to specific elements, expressed or unexpressed in the text, that characterize irony and sarcasm. As pre-trained language model specific for Italian on social media texts, we used AlBERTo, the model for Twitter Italian language understanding created by Polignano et al. (2019). This language model was trained on TWITA (Basile et al., 2018), a large dataset collecting Italian tweets from February 2012. The model that we used was trained on 200M tweets published from 2012 to 2015 using 12 hidden layers with size of 768 neurons[25].

In order to evaluate the performance of the proposed approach, called AlBERToIS (AlBERTo for Irony and Sarcasm detection), we compared it with the basic system (using only AlBERTo without linguistic features) and with the results of IronITA shared task.

6.1. System Description

AlBERToIS takes in account two principal sets of inputs: AlBERTo's inputs and the features' vector representation. In accordance with standard BERT input representation (Devlin et al., 2018), the text is represented for AlBERTo as tokens, segments and masked input. In order to load the trainable model of AlBERTo and tokenize the texts for creating tokens-input, we used keras-bert

[25] `https://github.com/marcopoli/AlBERTo-it`

implementation for BERT[26]. Moreover, we used *keras*[27] and *tensorflow*[28] as principal libraries to build our system exploiting the GPU process.

With respect to the creation of the features' vector representation, a data preprocessing phase is performed in accordance with the information that we wanted to extract from the tweets. For the majority of the features, we took into account a dictionary of words weighted with TF-IDF (Term Frequency–Inverse Document Frequency) values. To create this dictionary and the word embedding model used to extract semantic information, we preprocessed the tweets as follows: deleting URLs and symbols like @ and # to maintain the lexical information of hashtags and users' names; tokenizing and lemmatizing words using the TreeTagger tool[29] (Schmid, 2013) implemented for python in the *treetaggerwrapper* library[30]; and removing stopwords[31] to retain lexical significant words. Moreover, to extract PoS tags and syntactic dependencies from texts we used spacy-udpipe library with TWITTIRÒ model for the Italian language in Twitter[32] (Cignarella et al., 2019). Finally, the majority of the features have been standardized using *MinMaxScaler* of *scikit-learn*[33] with default range of scaling. The ensemble of features extracted from tweets is described in the next Section 6.2. Before combining these features with AlBERTo, we applied the batch normalization technique to the input-layer for features to standardize the layer and stabilize the learning process.

In the end, the combination is attained concatenating the final-layer of AlBERTo network with the input-layer of the features' vector representation. In addition, taking into account the considerable size of AlBERTo network, after the concatenation step, we used a dropout layer with a rate of 0.3 to prevent

[26] `https://github.com/CyberZHG/keras-bert`
[27] `https://keras.io/`
[28] `https://www.tensorflow.org/`
[29] Using this tool the numbers are replaced by `@card@` tag.
[30] `https://treetaggerwrapper.readthedocs.io/en/latest/`
[31] For the list of stopwords see: `http://di.unito.it/stopwordsit`
[32] `http://di.unito.it/twittirotreebank`
[33] `https://scikit-learn.org/stable/index.html`

the overfitting. At the end of our neural network, we added a dense-layer with standard ReLU activation with an input of 256 neurons and an output-dense-layer with a sigmoid function for binary classification in Task A (*ironic* and *non-ironic* classes) and in Task B (*sarcastic* and *non-sarcastic* classes). Specifically for the Task B, we adopted also a technique to care about the initial bias calculated taking into account the imbalance between sarcastic and non-sarcastic[34] classes. As optimizer we used Adam with a really low learning rate (0.00001) found by means of a specific callback function[35]. Finally, to minimize the loss function during the training we used the binary cross-entropy function for binary classification provided by *keras*.

6.2. Linguistic Features

Inspired by the errors emerged in HC of both tasks, we decided to design specific features that could improve the identification of these ironic and sarcastic patterns.

Stylistic Features Especially in short and informal texts such as tweets (see Table 14), punctuation helps authors to express better their intention (i.e. quotation marks to underline the opposite of the literal meaning: *"merito"*, *"buona scuola"*). Like punctuation, negation patterns show to play an important role in the process of comprehension of ironic and sarcastic texts (Giora et al., 2015, 2018; Karoui et al., 2015, 2017). Therefore, these patterns and their relevance are caught by system providing as vectorized inputs the sum of TF-IDF weights of punctuation characters (`punct`) and negation elements (`negation`) in the text.

Syntactic Features As shown in other works (Cignarella et al., 2020), syntactic features are proven to be useful to detect irony in social media. In particular, inspired by the error analysis in Section 5.2 and 5.3.1, we helped the system to

[34] We train our model on sarcastic and non-sarcastic tweets, including ironic/non-ironic ones, to ensure that system could recognize specific characteristics of sarcasm.

[35] `http://di.unito.it/lrfinder`

capture syntactic dependencies expressing adverbial locutions (`adv_loc`), intensifiers (`intens`), discourse connections (`disc_conn`), mentions (`mention`) and nominal phrases (and the number of nominal phrases in the tweet) (`nom_phrase` and `num_nom_phrase`).

Semantic Features The previous analysis suggests that specific emotions and semantic incongruities within the text could trigger ironic and sarcastic interpretation. To take into account these aspects, we used a set of lexical resources (Sentix[36], HurtLex[37] and EmoLex[38]) and an ensemble of features aimed to help the system to understand the semantic incongruities and similarities revealed by words and pairs of words used in ironic and sarcastic texts.

Sentiment Lexicon In Sentix (Basile & Nissim, 2013) each entry (for a total of 44715 words) consists of an Italian lemma followed by information as PoS tag, WordNet synset ID, a positive and a negative score from SentiWordNet, a polarity and an intensity score. Using this information, we calculated the average of positive and negative score of words in the tweet (`avg_positive` and `avg_negative`), the standard deviation (σ) of polarity inside the tweet and the intensity score average to indicate whether the tweet expresses an objective or subjective message (`avg_intensity`).

Hurtful Words HurtLex (Bassignana et al., 2018) is a multilingual lexicon of hateful words created from the Italian lexicon "Le Parole per Ferire" by Tullio de Mauro. The entries in the lexicon are categorized in 17 types of offenses (see Table 17) enclosed in two macro-categories: *conservative* (words with literally offensive sense) and *inclusive* (words with not literally offensive sense, but that could be used with negative connotation). To extract features from tweets relative to the 17 categories, we used a specific *featurizer*[39] created specifically for this lexicon. As weight for each category, we computed the sum of TF-IDF of words in the tweet belonging to each category without omitting

[36] `http://valeriobasile.github.io/twita/sentix.html`

[37] `http://hatespeech.di.unito.it/resources.html`

[38] `http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm`

[39] `https://github.com/valeriobasile/hurtlex`

the macro-category of reference.

Category	Length	Description
PS	254	Ethnic Slurs
RCI	36	Location and Demonyms
PA	167	Profession and Occupation
DDP	496	Physical Disabilities and Diversity
DDF	80	Cognitive Disabilities and Diversity
DMC	657	Moral Behavior and Defect
IS	161	Words Related to Social and Economic advantages
OR	144	Words Related to Plants
AN	775	Words Related to Animals
ASM	303	Words Related to Male Genitalia
ASF	191	Words Related to Female Genitalia
PR	138	Words Related to Prostitution
OM	145	Words Related to Homosexuality
QAS	536	Descriptive Words with Potential Negative Connotations
CDS	2042	Derogatory Words
RE	391	Felonies and Words Related to Crime and Immoral Behavior
SVP	424	Words Related to the Seven Deadly Sins of the Christian Tradition

Table 17: HurtLex Categories

Emotional Lexicon EmoLex (Mohammad & Turney, 2013) is a multilingual lexicon containing sentiment and affective information for each entry (for a total of 11360 words). For our purposes, we principally used the annotation relative to the 8 principal emotions of Plutchik (Plutchik & Kellerman, 2013). Inspired by Plutchik (2001), we exploited the wheel of emotions to capture in the message the principal emotions (anger, anticipation, disgust, fear, joy, sadness, surprise, trust), the primary dyads or feelings (aggressiveness, optimism, love, submission, awe, disapproval, remorse, contempt), and the variability of opposite emotions and contrary feelings by means of σ. The weight of emotions and feelings are computed summing the TF-IDF of words belonging to the specific categories.

Incongruity/Similarity Features Inspired by Riloff et al. (2013); Joshi et al. (2015); Tay et al. (2018); Pan et al. (2020) and the error analysis in Section 5.2, we calculated: the variability of the TF-IDF weights of the words inside the tweet by means of σ and the coefficient of variation (`cv`), the average of weights

(avg), and the maximum (max), minimum (min) and median (med) values of list of TF-IDF weights of words (W) and of bigrams of words (B) of a text to take into account the most significant tokens (such as interjections and hashtags) in ironic and sarcastic texts. The values related to bigrams are computed using the weights normalization on maximum and minimum scores (C1) and on standard deviation and average (C2). Additionally, we created a word embedding model starting from a pre-trained model on TWITA (Basile et al., 2018). Firstly, using the *Gensim* library[40], we updated the vocabulary and the word embeddings of the TWITA model with the SENTIPOLC 2016 tweets. Secondly, we extended the updated word embedding model with out of vocabulary words predicting their most probable embedding vectors considering their context. The prediction is based on a language model built on the IronITA dataset using Bi-directional Recurrent Neural Network with Long-Short Term Memory cell[41]. The final word embedding model is used to calculate the similarity ($\cos(\theta)$) between pairs of words (vector of bigram of words) and the sentence context (corresponding to sentence vector) ($\cos(\theta)$_BS), and between the bigrams of words within the sentence ($\cos(\theta)$_BB). To create the feature vector for our system we computed σ, the coefficient of variation, the average, and maximum, minimum and median scores of lists of cosine similarity values.

Figure 1 shows the most relevant features for irony and sarcasm detection calculated by means of χ^2 value[42].

As mentioned, χ^2 test measures the dependence between variables (in this case non-negative features and classes) to see if they are related. In spite of the difference of the distribution of ironic and sarcastic tweets in the training set, looking at Figure 1, we can observe an important lexical trend in ironic and sarcastic tweets. Users tend to use hurtful words especially to express sarcasm, and affective words to express irony. With respect to other features, we can

[40] https://radimrehurek.com/gensim/

[41] This methodology is inspired by Kandi's Master Thesis work presented in http://di.unito.it/oov

[42] The complete list of features is reported in Table 22 in Appendix A.

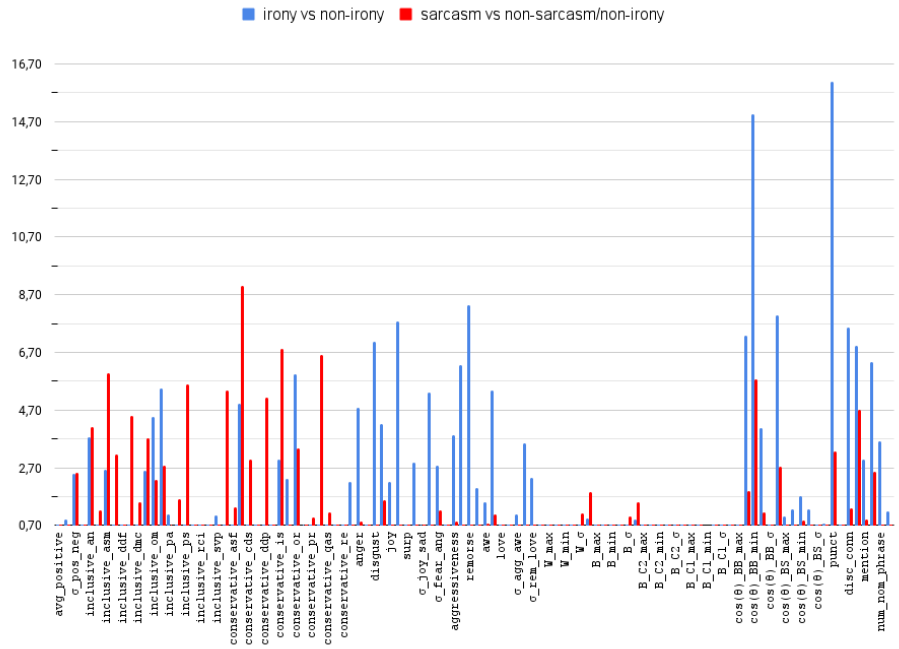


Figure 1: Representation of The Most Relevant Features in the Training Set

notice that: the variability of sentiment polarity in the message is characteristic of ironic and sarcastic statements, the variation of weights of words and pairs of words in the tweet appears more significant in sarcastic expressions, whereas especially ironic messages imply semantic similarities and incongruities disclosed by means of the computation of cosine similarity. About the syntactic features, the graphic shows that, in general, punctuation plays an important role in the expression of irony in short texts. However, also the other syntactic features investigated here show to be involved mainly in ironic utterances.

7. Experiments and Results

The experimental phase focused on the analysis of the contribution of the designed features for irony and sarcasm detection. To perform the experiments, we used 20% of the training set as validation set. The models were trained with

a maximum of 10 epochs for each run and a batch size of 8 for each epoch. To avoid problems of overfitting during the training phase, we employed an early stopping strategy, monitoring the minimum value of the loss curve on the validation set with patience of 3 epochs[43]. Moreover, to obtain reproducible results, we set the seed function from the tensorflow library.

In order to select the features that help the system to generalize better, we carried out various experiments taking into account their χ^2 value, and chose the best model for each task by means of binary accuracy values obtained on validation set. Indeed, binary accuracy metric is typically used for calculating how often predictions match binary labels. In particular, for Task A the best binary accuracy score (0.817) is obtained with the set of 24 features with a χ^2 greater than 3. As shown in Figure 1, the most contributing set of features for this task includes: hurtful words; most of statistical values calculated considering the cosine similarity between bigrams vectors and the sentence/vector context; stylistic features; adverbial locutions, discourse connections, number of nominal phrases among the syntactic features; and, finally, all the negative emotions (such as anger, disgust, fear, sadness, as well as the variability of trust and disgust) and the negative feelings (such as aggressiveness, contempt, remorse and submission, as well as the variability of contempt and submission). Differently from Task A, the best model selected for Task B, with a binary accuracy score of 0.772, involves all the extracted features.

The selected best models for Task A and Task B are, finally, evaluated on the test set used in the shared task. To this purpose, we used the same evaluation metrics used in IronITA: F1 for each class and F1-macro as average score. Specifically for Task B, we adopted a cascade architecture. Therefore, the predictions are obtained only for the tweets that were predicted as ironic in Task A.

For the competition, the organizers provided two straightforward baselines: *baseline-mfc* (Most Frequent Class) that assigns to each instance the majority

[43]Appendix B shows the learning curves.

class of the respective task, namely *non-ironic* for Task A and *non-sarcastic* for Task B; and *baseline-random* that assigns uniformly random values to the instances[44]. To prove the efficiency of our approach, we compared the obtained results with the baselines of IronITA shared task and the results obtained by the best performing systems. Moreover, to demonstrate the contribution of the selected set of features, we added a new baseline using the AlBERTo model without linguistic features.

team name	**id**	**F1-score**		
		non-iro	**iro**	**macro**
AlBERToIS		**0.739**	**0.768**	**0.754**
AlBERTo		*0.722*	*0.747*	*0.735*
ItaliaNLP	1	0.707	0.754	0.731
baseline-random		*0.503*	*0.506*	*0.505*
baseline-mfc		*0.668*	*0.000*	*0.334*

Table 18: Comparison of Results for Task A

team name	**id**	**F1-score**			
		non-iro	**non-sarc iro**	**sarc**	**macro**
AlBERToIS		**0.739**	**0.471**	0.518	**0.576**
AlBERTo		***0.739***	*0.416*	***0.527***	*0.561*
UNITOR	2	0.668	0.447	0.446	0.520
baseline-random		*0.503*	*0.266*	*0.242*	*0.337*
baseline-mfc		*0.668*	*0.000*	*0.000*	*0.223*

Table 19: Comparison of Results for Task B

Tables 18 and 19 report the results obtained respectively in Task A and Task B. As we can notice, in both tasks AlBERToIS performs better in both classes

[44]It is necessary to specify that for Task A a class is assigned randomly to every instance, while for Task B the classes are assigned randomly only to eligible tweets who are marked `ironic`.

overcoming the first ranked system and the provided baselines. In spite of the F-score achieved in the sarcastic class with a simple system using AlBERTo model is slightly higher than the one obtained with AlBERToIS, the proposed model reveals to be more balanced and solid to discriminate between sarcasm and non-sarcastic irony.

8. Discussion

Error Analysis on AlBERToIS's predictions The values of the confusion matrix in Table 20 confirm the increase of sensibility of AlBERToIS respect to the best performing systems in IronITA shared task. In particular, we noticed a reduction of 5% of FP in Task A, and a notable increment of TP of 11% in Task B. The error analysis in Section 5.3.2 revealed, mainly, how the lack of offenses on the one hand, and the presence of derogatory speech on the other hand, tend to improve, respectively, FN and FP in both tasks. Using the selected categories of hurtful words and specific affective features may have allowed AlBERToIS to improve the detection of ironic tweets when they contain or not offensive language. However, in Task B the confusion matrix reports an increase of FP of 8%. Analyzing the set of ironic tweets misclassified as sarcastic, we noted that most of the tweets containing especially stereotypes and offensive expressions (*I rom saranno pure l'etnia più meschina, ladra, bugiarda del globo, ma NON GIUSTIFICA QUESTO. Manco allo zoo dai, a me viene il vomito #lidl*[45]). Nevertheless, looking at the TP and FN cases of AlBERToIS, we notice that in presence of aggressive language, sarcasm is correctly detected (*Ma pensa te! I ladri rampicanti sono rom quelli che portano cultura!! #Roma https://t.co/oPZz8gq0a8*[46]). This matches with the TP and FN values in Table 20.

[45] "The Roma will also be the meanest, thief, liar ethnic group in the world, but DO NOT JUSTIFY THIS. Not even at the zoo, come on, I vomit #lidl"

[46] "Can you believe it? The climbing thieves are Roma who bring culture!! #Roma https://t.co/oPZz8gq0a8"

	team name	id	FP (%)	FN (%)	TP (%)	TN (%)
Task A						
	ItaliaNLP	1	**36**	18	82	64
	AlBERToIS		31	18	82	**69**
Task B						
	UNITOR	2	22	**59**	41	**78**
	AlBERToIS		**30**	48	**52**	70

Table 20: Values of Confusion Matrix for Task A and B

In addition, we can observe in Table 20 a similar trend to the one of Section 5.3.2: the percentage of FP is higher than FN in irony detection. The tweets misclassified as ironic by AlBERToIS contain, especially, questions: rhetorical, such as *@matteorenzi bel programma #labuonascuola ma come è possibile per noi giovani andare a scuola senza avere i soldi per il pane?*[47]; and simple, as *@Frankytrash alla fine t'han messa dentro o no?*[48]. We hypothesize that the questions need to be addressed more specifically at a syntactic level as well as exclamations (*@TeamLodoFlorida tra mezz'ora?! Ok... mi tocca aspettare ancora... ce la posso fare!*[49]). Another typical aspect of irony that makes hard its detection, also with AlBERToIS, is the use of euphemisms (*Messico, uccisa reginetta di bellezza. È quel piccolo difetto che la valorizza. [mukenin]*[50]). However, differently from values in Table 16, AlBERToIS could classify correctly the majority of situational ironic/sarcastic tweets. Examining the TP and FP cases, we noticed, moreover, that the semantic features helped our model to detect correctly sarcastic tweets containing false assertions and oxymoron, whereas texts involving a context shift tend to be misclassified as sarcastic (*Mattarella batte*

[47] "@matteorenzi nice program #labuonascuola but how is it possible for us to go to school without having money for bread?"

[48] "@Frankytrash in the end did they put you in or not?"

[49] @TeamLodoFlorida in half an hour?! Ok... I have still to wait... I can make it!

[50] Mexico, beauty queen killed. It is that small flaw that valorizes her. [mukenin]

le mani al ritmo di Bella ciao. Batterie non incluse. [@sisivabbe][51]). Actually, for sarcasm detection, AlBERToIS takes into account all the engineered features that could, as in this last case, capture some patterns that are more related to irony as shown in Table 7.

Emotions in Irony and Sarcasm In line with previous works in various languages (Hernández Farías et al., 2016; De Mattei et al., 2018; Kanwar et al., 2019; Babanejad et al., 2020; Calvo et al., 2020), our results confirm the relevance of affective features for irony detection. As in English (Sulis et al., 2016) and in Spanish (Frenda & Patti, 2019), also in Italian, the most discriminating emotions for irony detection are all negative (anger, disgust, fear and sadness). Also negative feelings (aggressiveness, contempt, remorse and submission) appear to be significant as well as the variability of contempt and submission. A different trend is visible in Task B. Indeed, as shown in Figure 1 the affective features, in general, report a really low score except for fear, submission and the variability of fear and anger.

Hurtful Language in Irony and Sarcasm The presence of hurtful language in ironic/sarcastic tweets has been investigated by Frenda & Patti (2019) in Spanish, revealing that aggressive language is present in ironic texts. Looking at Figure 1, especially for discriminating *sarcastic* from *non-sarcastic* tweets, hurtful language seems to play an important role. Therefore, we carried out an additional experiment in sarcasm detection using in AlBERToIS the features with a χ^2 greater than 3 like in Task A. This set of 15 features includes the minimum value of cosine similarity calculated between pairs of words and the sentence context, the weight of punctuation in the tweet, adverbial locutions and various hurtful words with a conservative and inclusive negative connotation. These words are mainly related to animals, male genitalia, physical disabilities/diversity, social and economic advantages, ethnicity, plants and general insults. The F1-macro obtained on the test set with this model is really com-

[51] Mattarella claps his hands to the rhythm of Bella ciao. Batteries not included. [@sisivabbe]

petitive (0.573) showing that the contribution of features linked to the hurtful intention of sarcasm is notable. With respect to irony detection, in Section 7 the best selected model uses as features some categories of hurtful words related to plants, animals, male genitalia and homosexuality. These words are especially inclusive.

Ablation Test In order to understand the contribution of each feature in AlBERToIS, we carried out an ablation test. Observing its results in Table 21, we notice that in general the system tends to perform worse when the information about hurtful words is subtracted in both tasks. Moreover, it is interesting to note that knowledge about sentiment, and in particular about the variation of polarity in the message (see Figure 1), proves to be essential for sarcasm detection just as the features used to extract semantic incongruities and similarities are for irony detection.

	Task A	**Task B**
	F1-macro	**F1-macro**
AlBERToIS	**0.754**	**0.576**
Stylistic Features	0.749 (↓0.5%)	0.551 (↓2.5%)
Syntactic Features	0.738 (↓1.6%)	0.556 (↓2%)
Semantic Features		
- *Sentiment Lexicon*	–	0.532 (↓**4.4%**)
- *Hurtful Words*	0.725 (↓**2.9%**)	0.534 (↓**4.2%**)
- *Emotional Lexicon*	0.737 (↓1.7%)	0.551 (↓2.5%)
- *Incongruities and Similarities*	0.727 (↓**2.7%**)	0.545 (↓3.1%)

Table 21: Ablation Test in AlBERToIS for Task A and B

9. Conclusions and Future Work

In this paper we investigated the use of sarcastic figurative devices in Italian Twitter texts, with a special focus on abusive contexts, where such devices can be exploited to disguise hate speech against people from vulnerable categories and to convey hateful messages.

We distinguish sarcasm as a specific type of irony. In order to get insights about the language used to express sarcasm and other forms of verbal irony in Twitter, with a specific focus on hatred and emotions expressed [RQ1], we carried out a battery of statistical analyses on the IronITA Italian benchmark dataset, that consists of data from different sources, namely, the Hate Speech Corpus, including a set of hateful tweets targeting immigrants, and TWITTIRÒ, which includes tweets covering more general issues, not necessarily linked to abusive contexts. The analyses reveal that sarcasm is characterized by a more hurtful and aggressive language than that which characterizes other forms of irony, appearing principally offensive in abusive contexts. Moreover, an extensive error analysis of the predictions of the best performing systems at IronITA shared task confirms a significant impact of negative emotions and aggressive language on the detection of irony and sarcasm, providing useful knowledge about linguistic sarcasm and irony markers pertaining to different layers, ranging from the morphosyntactic to the semantic and pragmatic ones.

On the basis of these findings, we investigated if knowledge about hurtful and affective language could be helpful for irony and sarcasm detection [RQ2]. Extracting these aspects from texts, we noticed an interesting lexical trend on ironic and sarcastic tweets: in line with the findings on other languages (Sulis et al., 2016; Frenda & Patti, 2019), the expression of irony involves very negative emotions, but sarcasm, specifically, tends to be expressed with a more hurtful language, revealing the aggressive intention of the author towards the victim. The emerging of this clear trend led us to propose an experimental setting to investigate if transformer-based architecture could benefit from the addition of linguistic features related to hatred and emotions [RQ3]. To this purpose, we propose an approach that combines language knowledge (from AlBERTo) and linguistic features in a simple neural architecture: AlBERToIS. Its performance in both tasks of irony and sarcasm detection overcomes the best scores of the IronITA competition, showing an optimal increase when we introduce linguistic features.

Looking at Table 18 and 19 we can notice that the results of our model for

sarcasm detection are still lower than those obtained on the course-grained task on irony detection. We hypothesize that the scarcity of sarcastic samples in the IronITA dataset could have impacted such outcome. To address the issue, in future work we would like to experiment techniques of data augmentation to improve the current performance. Moreover, considering the fact that the investigation of the role of hurtful language - characterized in terms of hate speech, aggressiveness, offensiveness and stereotype dimensions - in ironic and sarcastic tweets is a promising novelty proposed in this study, we would like to extend it by covering other languages and contexts. In addition, considering the significant correlation between sarcasm and various dimensions of hate, it could be interesting to focus on how the victim is targeted in sarcastic hateful utterances, and on the viral potential of such implicit expressions of hate. Finally, comparing the error analysis carried out on the predictions produced by the three best systems in IronITA competition and by AlBERToIS, we identified some elements, such as euphemism and rhetorical questions, that make irony more subtle, which also deserve a more in-depth study in the future.

In conclusion, this paper addresses a novel issue on the hurtfulness of sarcasm. Our findings can have an important impact in the context of social media content moderation, contributing to the development of systems able to detect abusive language even if it is disguised by sarcasm.

Acknowledgements

The work of S. Frenda was funded by Ricerca di Ateneo-Compagnia di San Paolo 2016 under the research project "Immigrants, Hate and Prejudice in Social Media". The work of V. Basile, C. Bosco and V. Patti was partially funded by the research projects "STudying European Racial Hoaxes and sterEOTYPES" (STERHEOTYPES, under the call "Challenges for Europe" of VolksWagen Stiftung and Compagnia di San Paolo) and "Be Positive!" (under the 2019 "Google.org Impact Challenge on Safety" call). Under the latter project "Be Positive!" also the work of A.T. Cignarella was funded. Finally, the work of

the last author was partially funded by the Spanish Ministry of Science and Innovation under the research project MISMIS-FAKEnHATE on MISinformation and MIScommunication in social media "FAKE news and HATE speech" (PGC2018-096212-B-C31) and by the Generalitat Valenciana under DeepPattern (PROMETEO/2019/121).

References

Agrawal, A., An, A., & Papagelis, M. (2020). Leveraging transitions of emotions for sarcasm detection. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 1505–1508).

Alba-Juez, L., & Attardo, S. (2014). The evaluative palette of verbal irony. *Evaluation in context*, *242*.

Attardo, S. (2007). Irony as relevant inappropriateness. In H. Colston, & R. Gibbs (Eds.), *Irony in language and thought: A cognitive science reader* (pp. 135–172). Lawrence Erlbaum.

Babanejad, N., Davoudi, H., An, A., & Papagelis, M. (2020). Affective and contextual embedding for sarcasm detection. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 225–243).

Barbieri, F., Basile, V., Croce, D., Nissim, M., Novielli, N., & Patti, V. (2016). Overview of the Evalita 2016 sentiment polarity classification task. In *Proceedings of 3rd Italian Conference on Computational Linguistics (CLiC-it 2016) & 5th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*. Naples, Italy: CEUR.org.

Barbieri, F., Ronzano, F., & Saggion, H. (2015). 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 (SemEval 2015)* (pp. 704–708).

Basile, P., & Novielli, N. (2014). Uniba at evalita 2014-sentipolc task predicting tweet sentiment polarity combining micro-blogging, lexicon and semantic features. *UNIBA at EVALITA 2014-SENTIPOLC Task Predicting tweet sentiment polarity combining micro-blogging, lexicon and semantic features.*, (pp. 58–63).

Basile, P., & Semeraro, G. (2018). UNIBA - Integrating distributional semantics features in a supervised approach for detecting irony in Italian tweets. In *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18)*. Turin, Italy: CEUR.org.

Basile, V. (2020). It's the end of the gold standard as we know it. on the impact of pre-aggregation on the evaluation of highly subjective tasks. In *2020 AIxIA Discussion Papers Workshop, AIxIA 2020 DP* (pp. 31–40). CEUR-WS volume 2776.

Basile, V., Bolioli, A., Nissim, M., Patti, V., & Rosso, P. (2014). Overview of the Evalita 2014 sentiment polarity classification task. In *Proceedings of the 4th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'14)*. Pisa, Italy: Pisa University Press.

Basile, V., Lai, M., & Sanguinetti, M. (2018). Long-term Social Media Data Collection at the University of Turin. In *Proceedings of the 5th Italian Conference on Computational Linguistics (CLiC-it 2018)*. Turin, Italy: CEUR.org.

Basile, V., & Nissim, M. (2013). Sentiment analysis on Italian tweets. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (pp. 100–107).

Bassignana, E., Basile, V., & Patti, V. (2018). Hurtlex: A multilingual lexicon of words to hurt. In *5th Italian Conference on Computational Linguistics, CLiC-it 2018* (pp. 1–6). CEUR-WS volume 2253.

Bazzanella, C. (2011). Oscillazioni di informalità e formalità: scritto, parlato e rete. *Formale e informale. La variazione di registro nella comunicazione elettronica*, (pp. 68–83).

Benamara, F., Grouin, C., Karoui, J., Moriceau, V., & Robba, I. (2017). Analyse d'opinion et langage figuratif dans des tweets : présentation et résultats du Défi Fouille de Textes DEFT2017. In *Atelier TALN 2017 : Défi Fouille de Textes (DEFT 2017)* (pp. pp. 1–12). Orléans, France. URL: `https://hal.archives-ouvertes.fr/hal-01912785`.

Bosco, C., Dell'Orletta, F., Poletto, F., Sanguinetti, M., & Tesconi, M. (2018). Overview of the Evalita 2018 Hate Speech Detection Task. In *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18)*. Turin, Italy: CEUR.org.

Bosco, C., Patti, V., & Bolioli, A. (2013). Developing corpora for sentiment analysis: The case of irony and senti-tut. *IEEE intelligent systems*, *28*, 55–63.

Bowes, A., & Katz, A. (2011). When sarcasm stings. *Discourse Processes: A Multidisciplinary Journal*, *48*, 215–236.

Calvo, H., Gambino, O. J., & García Mendoza, C. V. (2020). Irony detection using emotion cues. *Computación y Sistemas*, *24*.

Castellucci, G., Croce, D., & Basili, R. (2016). A language independent method for generating large scale polarity lexicons. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 38–45). ELRA.

Chauhan, D. S., Dhanush, S., Ekbal, A., & Bhattacharyya, P. (2020). Sentiment and emotion help sarcasm? a multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 4351–4360).

Cignarella, A. T., Basile, V., Sanguinetti, M., Bosco, C., Benamara, F., & Rosso, P. (2020). Multilingual Irony Detection with Dependency Syntax and Neural Models. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 1346–1358). ACL.

Cignarella, A. T., Bosco, C., Patti, V., & Lai, M. (2018a). Application and Analysis of a Multi-layered Scheme for Irony on the Italian Twitter Corpus TWITTIRÒ. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: ELRA.

Cignarella, A. T., Bosco, C., & Rosso, P. (2019). Presenting TWITTIRO-UD: An Italian Twitter Treebank in Universal Dependencies. In *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)* (pp. 190–197).

Cignarella, A. T., Frenda, S., Basile, V., Bosco, C., Patti, V., Rosso, P. et al. (2018b). Overview of the Evalita 2018 task on irony detection in Italian tweets (Ironita). In *Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2018)* (pp. 1–6). CEUR-WS volume 2263.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*. volume 11.

De Mattei, L., Cimino, A., & Dell'Orletta, F. (2018). Multi-task learning in Deep Neural Networks for Irony Detection. In *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18)*. Turin, Italy: CEUR.org.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, .

Di Rosa, E., & Durante, A. (2018). Irony detection in tweets: X2Check at Ironita 2018. In *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18)*. Turin, Italy: CEUR.org.

Du Marsais, C. C., Paulhan, J., & Mouchard, C. (1981). *Traité des tropes*. Le Nouveau Commerce.

Dynel, M. (2014). Linguistic approaches to (non) humorous irony. *Humor - International Journal of Humor Research*, *27*, 537–550.

Erjavec, K., & Kovačič, M. P. (2012). "you don't understand, this is a new war!" analysis of hate speech in news web sites' comments. *Mass Communication and Society*, *15*, 899–920.

Farias, D. H., & Rosso, P. (2017). Chapter 7 - irony, sarcasm, and sentiment analysis. In F. A. Pozzi, E. Fersini, E. Messina, & B. Liu (Eds.), *Sentiment Analysis in Social Networks* (pp. 113 – 128). Boston: Morgan Kaufmann. URL: `http://www.sciencedirect.com/science/article/pii/B9780128044124000073`. doi:`https://doi.org/10.1016/B978-0-12-804412-4.00007-3`.

Frenda, S. (2018). The role of sarcasm in hate speech: A multilingual perspective. In *Proceedings of Doctoral Symposium at SEPLN 2018*. CEUR-WS.

Frenda, S., Banerjee, S., Rosso, P., & Patti, V. (2020). Do Linguistic Features Help Deep Learning? The Case of Aggressiveness in Mexican Tweets. *Computación y Sistemas*, *24*.

Frenda, S., Ghanem, B., Guzmán-Falcón, E., Montes-y Gómez, M., Villasenor-Pineda, L. et al. (2018). Automatic expansion of lexicons for multilingual misogyny detection. In *6th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop, EVALITA 2018* (pp. 1–6). CEUR-WS volume 2263.

Frenda, S., & Patti, V. (2019). Computational models for irony detection in three spanish variants. In *IberLEF@ SEPLN* (pp. 297–309).

Garavelli, B. M. (1997). *Manuale di retorica*. Bompiani Milan.

Ghosh, A., Li, G., Veale, T., Rosso, P., Shutova, E., Barnden, J., & Reyes, A. (2015). Semeval-2015 task 11: Sentiment analysis of figurative language in Twitter. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)* (pp. 470–478).

Ghosh, D., Vajpayee, A., & Muresan, S. (2020). A report on the 2020 sarcasm detection shared task. `arXiv:2005.05814`.

Gibbs, R. W. (2000). Irony in talk among friends. *Metaphor and symbol*, *15*, 5–27.

Giora, R., Givoni, S., & Fein, O. (2015). Defaultness reigns: The case of sarcasm. *Metaphor and Symbol*, *30*, 290–313.

Giora, R., Jaffe, I., Becker, I., & Fein, O. (2018). Strongly attenuating highly positive concepts. *Review of Cognitive Linguistics. Published under the auspices of the Spanish Cognitive Linguistics Association*, *16*, 19–47.

Giudice, V. (2018). Aspie96 at IronITA (EVALITA 2018): Irony Detection in Italian Tweets with Character-Level Convolutional RNN. In *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18)*. Turin, Italy: CEUR.org.

Hernández Farías, D. I., Patti, V., & Rosso, P. (2016). Irony detection in Twitter: The role of affective content. *ACM Transactions on Internet Technology (TOIT)*, *16*, 19.

Joshi, A., Sharma, V., & Bhattacharyya, P. (2015). 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) (pp. 757–762). volume 2.

Justo, R., Corcoran, T., Lukin, S. M., Walker, M., & Torres, M. I. (2014). Extracting relevant knowledge for the detection of sarcasm and nastiness in the social web. *Knowledge-Based Systems*, *69*, 124–133.

Kanwar, N., Mundotiya, R. K., Agarwal, M., & Singh, C. (2019). Emotion based voted classifier for arabic irony tweet identification. In *FIRE (Working Notes)* (pp. 426–432).

Karoui, J., Benamara, F., Moriceau, V., Aussenac-Gilles, N., & Hadrich Belguith, L. (2015). Towards a contextual pragmatic model to detect irony in tweets. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguisticsand the 7th International Joint Conference on Natural Language Processing* (pp. 644–650). Association for Computational Linguistics (ACL).

Karoui, J., Farah, B., Moriceau, V., Patti, V., Bosco, C., & Aussenac-Gilles, N. (2017). 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* (pp. 262–272). volume 1.

Lee, C. J., & Katz, A. N. (1998). The differential role of ridicule in sarcasm and irony. *Metaphor and Symbol*, *13*, 1–15.

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, *29*, 436–465.

Molla, D., & Joshi, A. (2019). Overview of the 2019 ALTA shared task: Sarcasm target identification. In *Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association* (pp. 192–196). Sydney, Australia: Australasian Language Technology Association. URL: `https://www.aclweb.org/anthology/U19-1026`.

Naseem, U., Razzak, I., Eklund, P., & Musial, K. (2020). Towards improved deep contextual embedding for the identification of irony and sarcasm. In *2020 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–7). IEEE.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web* (pp. 145–153).

Nockleby, J. T. (2000). Hate speech. *Encyclopedia of the American constitution*, *3*, 1277–1279.

Ortega-Bueno, R., Rangel, F., Hernández Farıas, D., Rosso, P., Montes-y Gómez, M., & Medina Pagola, J. E. (2019). Overview of the task on irony detection in spanish variants. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019), co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019). CEUR-WS. org*.

Pan, H., Lin, Z., Fu, P., & Wang, W. (2020). Modeling the incongruity between sentence snippets for sarcasm detection. In *24th European Conference on Artificial Intelligence, 29 August–8 September 2020, Santiago de Compostela, Spain – Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020)*.

Plutchik, R. (2001). The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*, *89*, 344–350. URL: `http://www.jstor.org/stable/27857503`.

Plutchik, R., & Kellerman, H. (2013). *Theories of emotion* volume 1. Academic Press.

Polignano, M., Basile, P., de Gemmis, M., Semeraro, G., & Basile, V. (2019).

AlBERTo: Italian BERT Language Understanding Model for NLP Challenging Tasks Based on Tweets. In *CLiC-it*.

Potamias, R. A., Siolas, G., & Stafylopatis, A.-G. (2020). A transformer-based approach to irony and sarcasm detection. *Neural Computing and Applications*, *32*, 17309–17320.

Reyes, A., & Rosso, P. (2012). Making objective decisions from subjective data: Detecting irony in customer reviews. *Decision support systems*, *53*, 754–760.

Reyes, A., Rosso, P., & Buscaldi, D. (2012). From humor recognition to irony detection: The figurative language of social media. *Data & Knowledge Engineering*, *74*, 1–12.

Riloff, E., Qadir, A., Surve, P., De Silva, L., Gilbert, N., & Huang, R. (2013). Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (pp. 704–714).

Sanguinetti, M., Bosco, C., Lavelli, A., Mazzei, A., Antonelli, O., & Tamburini, F. (2018a). PoSTWITA-UD: An Italian Twitter treebank in Universal Dependencies. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).

Sanguinetti, M., Bosco, C., Mazzei, A., Lavelli, A., & Tamburini, F. (2017). Annotating Italian social media texts in Universal Dependencies. In *Proceedings of the 4th International Conference on Dependency Linguistics (Depling 2017)* (pp. 229–239).

Sanguinetti, M., Poletto, F., Bosco, C., Patti, V., & Stranisci, M. (2018b). An Italian Twitter Corpus of Hate Speech against Immigrants. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan.

Santilli, A., Croce, D., & Basili, R. (2018). A Kernel-based Approach for Irony and Sarcasm Detection in Italian. In *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18)*. Turin, Italy: CEUR.org.

Schmid, H. (2013). Probabilistic part-ofispeech tagging using decision trees. In *New methods in language processing* (p. 154).

Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International workshop on natural language processing for social media* (pp. 1–10).

Simi, M., Bosco, C., & Montemagni, S. (2014). Less is More? Towards a Reduced Inventory of Categories for Training a Parser for the Italian Stanford Dependencies. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)* (p. 83–90). European Language Resources Association (ELRA).

Sridhar, R. et al. (2017). Emotion and sarcasm identification of posts from facebook data using a hybrid approach. *ICTACT journal on soft computing*, *7*.

Straka, M., & Straková, J. (2017). Tokenizing, PoS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies* (pp. 88–99).

Stranisci, M., Bosco, C., Farías, D. I. H., & Patti, V. (2016). Annotating Sentiment and Irony in the Online Italian Political Debate on #labuonascuola. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož, Slovenia: ELRA.

Sulis, E., Hernández Farías, D. I., Rosso, P., Patti, V., & 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. New Avenues in Knowledge Bases for Natural Language Processing.

Tay, Y., Tuan, L. A., Hui, S. C., & Su, J. (2018). Reasoning with sarcasm by reading in-between. *arXiv preprint arXiv:1805.02856*, .

Van Hee, C., Lefever, E., & Hoste, V. (2018a). Semeval-2018 task 3: Irony detection in english tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation* (pp. 39–50).

Van Hee, C., Lefever, E., & Hoste, V. (2018b). SemEval-2018 task 3: Irony detection in English tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation* (pp. 39–50). New Orleans, Louisiana: Association for Computational Linguistics.

Wallace, B. C., Charniak, E. et al. (2015). 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)* (pp. 1035–1044). volume 1.

Wang, P.-Y. A. (2013). #irony or #sarcasm — a quantitative and qualitative study based on Twitter. In *Proceedings of the 27th Pacific Asia Conference on Language, Information, and Computation (PACLIC 27)* (pp. 349–356). Taipei, Taiwan: Department of English, National Chengchi University. URL: https://www.aclweb.org/anthology/Y13-1035.

Wilson, D., & Sperber, D. (2012). Explaining irony. *Meaning and relevance*, (pp. 123–145).

Zhang, S., Zhang, X., Chan, J., & Rosso, P. (2019). Irony detection via sentiment-based transfer learning. *Information Processing & Management*, *56*, 1633–1644.

A. Appendix - List of Features

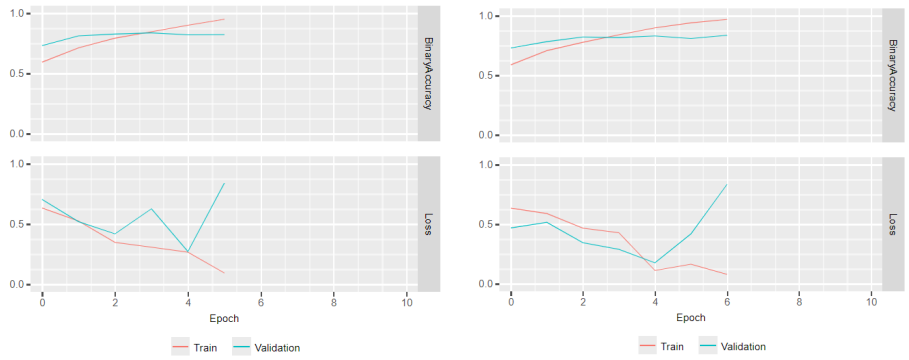
In Table 22, we report the complete list of features developed to linguistically inform AlBERToIS.

Type	Group	Features		
Stylistic		punct	negation	
Syntactic		num_nom_phrase	disc_conn	nom_phrase
		adv_loc	intens	mention
Semantic	Sentiment Lexicon	avg_positive	σ_pos_neg	avg_negative
		avg_intensity		
	Hurtful Words	inclusive_an	conservative_an	inclusive_asf
		conservative_asf	inclusive_asm	conservative_asm
		inclusive_cds	conservative_cds	inclusive_ddf
		conservative_ddf	inclusive_ddp	conservative_ddp
		inclusive_dmc	conservative_dmc	inclusive_is
		conservative_is	inclusive_om	conservative_om
		inclusive_or	conservative_or	inclusive_pa
		conservative_pa	inclusive_pr	conservative_pr
		inclusive_ps	conservative_ps	inclusive_qas
		conservative_qas	inclusive_rci	conservative_rci
		inclusive_re	conservative_re	inclusive_svp
		conservative_svp		
	Emotional Lexicon	anger	aggressiveness	anticipation
		contempt	disgust	remorse
		fear	disapproval	joy
		awe	sadness	submission
		surp	love	trust
		optimism	σ_joy_sad	σ_agg_awe
		σ_trust_disg	σ_cont_sub	σ_fear_ang
		σ_rem_love	σ_surp_ant	σ_dis_opt
	Incongruity/Similarity	W_max	B_C1_max	W_med
		B_C1_med	W_min	B_C1_min
		W_avg	B_C1_avg	W_σ
		B_C1_σ	W_cv	B_C1_cv
		B_max	cos(θ)_BB_max	B_med
		cos(θ)_BB_med	B_min	cos(θ)_BB_min
		B_avg	cos(θ)_BB_avg	B_σ
		cos(θ)_BB_σ	B_cv	cos(θ)_BB_cv
		B_C2_max	cos(θ)_BS_max	B_C2_med
		cos(θ)_BS_med	B_C2_min	cos(θ)_BS_min
		B_C2_avg	cos(θ)_BS_avg	B_C2_σ
		cos(θ)_BS_σ	B_C2_cv	cos(θ)_BS_cv

Table 22: List of Features

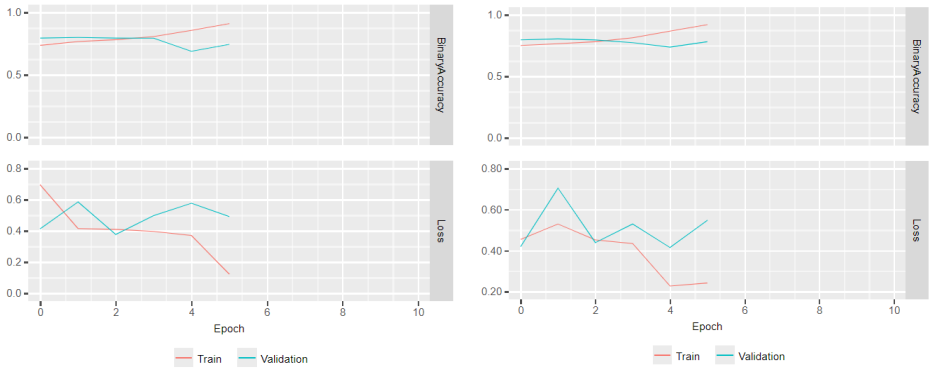
B. Appendix - Learning Curves

Figures 2 and 3 show the learning curves on loss and binary accuracy obtained during the learning of baseline system (AlBERTo-based) and AlBERToIS respectively in Task A and Task B.



(a) AlBERTo Baseline System Training (b) AlBERToIS Training

Figure 2: Learning Curves in Task A



(a) AlBERTo Baseline System Training (b) AlBERToIS Training

Figure 3: Learning Curves in Task B

In these figures, we can observe the contribution of linguistic information that in both tasks reduces the loss and increases the binary accuracy during the training. Moreover, in Figure 3 the scarcity of sarcastic data leads the classifiers

to slightly overfit on training data. However, the early stopping strategy adopted in our set of experiments helps us to stop the learning when generalization error starts increasing.

C. Appendix - List of Acronyms

AlBERTo	BERT language understanding model for the Italian language
AlBERToIS	AlBERTo for Irony and Sarcasm detection
AUC	Area Under the Curve
baseline-mfc	Baseline system based on Most Frequent Class
baseline-random	Baseline system that assigns uniformly random values to the instances
BERT	Bidirectional Encoder Representations from Transformers
EVALITA	Evaluation Campaign of Natural Language Processing and Speech Tools for Italian
FN	False Negative
FP	False Positive
HC	Hard Cases
HSC	Hate Speech Corpus
IronITA	Irony Detection in Italian Tweets
NLP	Natural Language Processing
PoS	Part-of-Speech
SC	Simple Cases
TF-IDF	Term Frequency–Inverse Document Frequency
TN	True Negative
TP	True Positive
UD	Universal Dependencies