Ortega-Bueno, R.; Rosso, P.; Medina-Pagola, JE. (2022). Multi-view informed attention-based model for Irony and Satire detection in Spanish variants. Knowledge-Based Systems. 235:1-24. https://doi.org/10.1016/j.knosys.2021.107597

Multi-view informed attention-based model for Irony and Satire detection in Spanish variants

Reynier Ortega-Bueno

PRHLT Research Center, Universitat Politècnica de València, Valencia Spain

Paolo Rosso

PRHLT Research Center, Universitat Politècnica de València, Valencia Spain

José E. Medina Pagola

Universidad de Ciencias Informáticas, Havana, Cuba

Abstract

Making machines understand language and reasoning on it has been one of the most challenging problems addressed by Artificial Intelligent researchers. This challenge increases when figurative language is used for communicating complex meanings, intentions, emotions and attitudes in creative and funny ways. In fact, sentiment analysis approaches struggle when facing irony, satire and other figurative languages, particularly those where the explanation of a prediction might arguably be as necessary as the prediction itself. This paper describes a new model MvAttLSTM based on deep learning for irony and satire detection in tweets written in distinct Spanish variants. The proposed model is based on an attentive-LSTM informed with three additional views learned from distinct perspectives. We investigate two strategies to pass these views into MvAttLSTM. We perform an extensive evaluation on three corpora, one for irony detection and two for satire detection. Moreover, in order to study the robustness of our proposed model, we investigate its performance on humor recognition. Experiments confirm that the proposed views help our model to improve its performance. Moreover, they show that affective information benefits our model to detect irony and satire. In particular, a first analysis of the results highlights the discriminating power of emotional features obtained from SenticNet and SEL lexicon. Overall, our system achieves the state-of-the-art performance in irony and satire detection in Spanish variants and competitive results in humor recognition.

Keywords: Irony and satire, Attention mechanism, Linguistic features, Contextualized pre-trained embedding, Fusing representation, Spanish variants, Figurative language

1. Introduction

Language itself is a perfect illustration of human creativity, and it achieves its splendor when some semantics rules and maxims of human communication (Grice, 1975, 1978) are disrespected to create expressions whose real meaning diverges from what it is apparently said. This peculiar usage of language with creative and funny purposes has been coined with the term *Figurative Language* (Raymond W. Gibbs & Colston, 2012; Dancygier, 2014; Colston, 2015). Irony, satire, sarcasm, humor, puns, simile, hyperbole, metonym and metaphor are forms (or devices) of figurative language. While it is true that all these forms are used to communicate complex meanings, not all of them are used by common people. Some forms are relegated only to literary and poetry usages (Reyes, 2012).

*reynier.ortega@gmail.com

Irony and satire are pervasive and popular in everyday communication. As human beings, we appeal to these devices as effective ways through which literal meanings are intentionally deviated in favor of secondary interpretations. Particularly, both devices are acknowledged to express an attitude that is generally negative and implicit behind an apparent positive message. Thus, they are frequently used to criticize, complain, ridicule or mock. Even when there are commonalities between both phenomena, irony seem to be more primitive and universal than satire.

The most common types of irony used in social media are situational and verbal irony. On the one hand, situational irony refers to specific events that fail to meet expectations (Lucariello, 1994) e.g. *"The fire station burns down while the firemen are out on a call"*. On the other hand, verbal irony has been traditionally identified as figurative device where enunciated words imply something other than their literal meanings. In other words, their real meaning is opposite to the literal one and it needs to be inferred through interpretation e.g. *"A burned tongue is a lovely way to start the day"*. Sarcasm is often considered a specific type of verbal irony which has a more aggressive tone (Attardo, 2000), is directed toward an individual or a group (Kreuz & Glucksberg, 1989; Kreuz & Roberts, 1993; Kreuz & Link, 2002; Sperber & Wilson, 1981), and is used intentionally (Gibbs et al., 1995; John Haiman, 1998). An example of sarcasm would be the exclamation *"You're really brilliant!"* about someone who has done a foolish act.

Satire is an interesting concept which is strongly related with irony and humor. It takes advantage of indirect speech and negative attitude implicit in irony. This device also appeals to features of humor such as: parody, exaggeration, juxtaposition, comparison, analogy, and double entendres with censoring purpose. Satirical messages may be aggressive and offensive, but they always have a deeper meaning and a social signification beyond that of the humour (Colletta, 2009). Satire does not make sense when the reader does not understand the real intent hidden in the ironic/funny dimension; like in irony, the real meaning of a satirical message lays in the figurative interpretation of the content. Satire can be separated in two distinct directions: *Juvenalian* or *Horatian* styles (Condren, 2014). On the one hand, the *Juvenalian* style of satire is based on ridicule and sarcasm. On the other hand, the *Horatian* style contains tease and humor.

Irony and satire have been studied from many disciplines such as Linguistics, Psychology, Rhetoric, Pragmatics, Semantics, etc., however, they are not only enclosed to these theoretical studies. Nowadays, both devices are typically used in social media platforms to favor social interactions, evoking humor (Wilson & Sperber, 1992), diminishing or enhancing criticism (Brown & Levinson, 1987; Simpson, 2003), and getting the attention of the readers by means of the creativity (Veale & Hao, 2009). These forms of figurative language have great impact on several other Natural Language Processing (NLP) tasks that aiming at monitoring social media content. In some cases the presence of ironic message plays a specific role: *"implicit polarity reversal"*. This means, that a message seems to be positive but its real meaning is negative (or vice-versa). Due to this peculiarity of ironic and sarcastic expressions, the sentiment analysis approaches decline when facing irony in social media texts (Maynard & Greenwood, 2014; Ghosh et al., 2015; Basile et al., 2014; Barbieri et al., 2016b; Hee, 2017; Farias & Rosso, 2017). In fact, this problem become more challenging in sentiment analysis approaches where the explanation of the results is more important than the decision itself (Zucco et al., 2019; Bodria et al., 2020). Sarcasm as a specific sub-type of verbal irony has implications to cope with the bad phenomenon of miscommunication, particularly: hate, aggressiveness, and nastiness speech (Justo et al., 2018). Ignoring the presence of sarcasm causes that the implicit meaning, generally hurtful and offensive, be misunderstood with the results of exposing people to toxic information. In this context also it results crucial to understand what pieces of message are relevant. Recently, interesting evidence about the use of satire to disguise fake news has been discussed in (Rubin et al., 2016; Golbeck et al., 2018). For people, understanding satire as fake messages may deprive them of desirable entertainment content, while recognizing fake information as legitimate satire may expose them to disinformation or misinformation.

Following the timeline of computational methods for irony and satire detection, it is possible to envisage two distinguishable approaches namely, classical features-based machine learning approach, and deep-leaning approach. In the literature many works explored several linguistic, stylistic, content, affective, and contextual features to address the problem in a shallow supervised way (Wallace, 2015). The handcraft features derived from this approach have proved to be feasible when small dataset are provided.

In the last five years, deep learning techniques became very popular in NLP, and applied to irony and satire detection. These methods show a better performance than classical feature-based machine learning

models. In this scenario, a vast number of works are based on attentive-Recurrent Neural Networks (att-RNN), particularly by using of Long Short Term Memory (LSTM)(Hochreiter & Schmidhuber, 1997) and Gated Recurrent Unit (GRU) (Chung et al., 2014) with attention mechanisms, which have been effective to capture complex dependencies among words within the text and pay more attention to those words that increase the effectiveness of these networks in several tasks of NLP (Luong et al., 2015; Wang et al., 2016; Yang et al., 2016, 2017b).

Recently, a groundbreaking advance in NLP has been marked by using transformer-based network architectures (Vaswani et al., 2017) which opened a new avenue for training robust and contextual-aware word embeddings in unsupervised manner. The use of these pre-trained contextualized models has been widely spread by means of Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) and BERT's family architectures (Conneau et al., 2020; Liu et al., 2019; Sanh et al., 2019; Lan et al., 2020). Clearly, this spreading has reached figurative language processing, and cause that these new computational methods outperformed the state of the art by a substantial margin (González et al., 2020; Potamias et al., 2020; Ghosh et al., 2020).

The nested non-linear functions of deep learning algorithms provoke these models are usually applied in a black-box manner, that is, no interpretable knowledge is provided about what exactly causes them to arrive at their predictions. In this sense, the attention mechanisms became a (albeit narrow) way for dealing with the problem of model interpretability. Recently, attention as a way of explainable deep learning method became a very popular and controversial topic. Some works claimed that attention weights do not provide meaningful "explanations" for supporting the final predictions (Serrano & Smith, 2020; Jain & Wallace, 2019), however other works state that they are able for discovering how neural models capture several linguistic notions of syntax, semantic and coreference (Vig & Belinkov, 2019; Clark et al., 2019; Tenney et al., 2020). Despite diverse views on the matter, empirical results on the task of binary irony classification show that attention mechanisms are able for capturing ironic cues, word polarity and explicit and implicit sentiment incongruity (González et al., 2020; Zhang et al., 2019).

Even when, remarkable advances have been observed in irony and satire detection, these advances have showed an asymmetric development with respect to languages other than English. This is the case of Spanish and its variants, where few researchers have addressed the problem. In the case of satire, only the works (Barbieri et al., 2015b; Salas-Zárate et al., 2017) have studied the phenomenon in the Spanish language by using a machine learning-based approach. Irony detection in Spanish variants has been surveyed in (Ortega et al., 2019), but few methods relied on deep learning approaches (Seda Mut Altin et al., 2019; González et al., 2019; Miranda-Belmonte & López-Monroy, 2019; García et al., 2019). According to our knowledge only the works (González et al., 2020; García et al., 2019) take advantages of contextualized pre-trained word embedding. For that, more efforts must be paid to study irony and satire in Spanish language. In this work we propose an attentive-based deep learning method in order to investigate further the detection of irony and satire in Spanish:

- Irony and satire are pragmatic phenomena, hence both are contextual-dependents. Additional knowledge such as: language and its variety, sociolinguistics and cultural background are crucial for precisely recognize and understand these forms of figurative language.

- Many studies on irony and satire detection have been conducted from three directions: linguistics features with machine learning approaches, deep learning techniques based on att-RNN, and recently by means of contextualized pre-trained word embeddings with a single-modality representation of the texts. However, there are no works that pay attention to explore irony and satire in Spanish from a fusion information perspective where these three approaches are fused aiming to outperform the state of the art.

To overcome these challenges, we aim at addressing the following research questions:

RQ1. Could irony and satire detection methods take advantage of combining multiple representations (views) of text in terms of linguistic-based representation, universal sentence encoder-based representation, and contextualized pre-trained embeddings?

RQ2. How the proposed views can be effectively combined to proper inform an attentive recurrent model?

RQ3. Are multiple heads of attention (multi-head) more feasible than single attention (self) for capturing multiples and complex relations among words in ironic and satirical texts?

With the aim to answer the formulated research questions, in this work we propose a new model ($MvAttLSTM$) which relies on a multi-view informed attentive-LSTM neural network. We consider an attentive recurrent model due to the attention mechanisms allow the model to focus and place more "attention" on the relevant parts of the text sequence in order to capture complex syntactic and semantic properties used in ironic and satirical messages. Specifically, we learn three independent views for each text, and we pass them to our MvAttLSTM. The first one (*Linguistic-view*) is based on several linguistics features which have proved to be strong cues for discriminating both irony and satire. The second one considers a deep dense encoding of the text by means of Multilingual Universal Sentence Encoder (*MUSE-view*). And, finally the last one (*BERT-view*) considers a contextualized pre-trained embedding obtained after a tuning of the BERT model. We evaluate the effectiveness of our method on one corpus for irony detection and on two distinct corpora for satire detection. For irony detection, the corpus is the one proposed for the *IroSvA'19* shared task: *Irony detection in Spanish Variants* (Ortega et al., 2019) was used, whereas, in the case of satire detection task the corpora introduced in (Salas-Zárate et al., 2017; Barbieri et al., 2015b) were employed. Our proposal outperformed both previous systems participating in the *IroSvA'19* shared task and recent methods (González et al., 2020; Calvo et al., 2020). Also, for satire detection our proposal outperformed previous methods by a substantial margin in both corpora. Additionally, we provide several analyses in order to evaluate two strategies for fusing the learned views into MvAttLSTM and investigate the impact of each view on the performance of MvAttLSTM. Finally, an interesting analysis is carried out on the attention mechanism to observe how our proposal takes into account some features related with irony and satire such affective content. In short, the major contributions of this paper are summarized below:

- To investigate the problem of computational irony and satire detection in Spanish variants in three widely used corpora. Moreover, taking into account the closed relation among irony, satire and humor, we evaluate the robustness of the proposed model on humor recognition.

- To propose a novel approach (MvAttLSTM) based on representation fusion. Particularly, efficient representations from three distinct perspectives are computed and combined to inform an attentive-LSTM model. The proposed method outperforms the state of the art approaches in satire and irony detection in Spanish and obtains competitive results in humor recognition.

- To investigate distinct forms of combining the proposed views, also to evaluate the impact of each view on the proposed model MvAttLSTM.

- To study the impact of two kinds of attention mechanisms, self attention vs. multi-head attention, on the proposed model.

The rest of this paper is structured as follows. Section 2 introduces the state of the art for both irony and satire detection, with special interest in those approaches proposed for Spanish. Section 3 formalizes our proposal based on a multi-view attention based model. Particularly, we describe each one of the representation used and two distinct ways of fusing these views for irony and satire detection. In Section 4 a detailed description of the corpora, resources, preprocessing and the experimental setup is introduced. Also, an exhaustive evaluation of our proposal and a comparison with other approaches is presented. Moreover, considering the closed relation among irony, satire and humor, we evaluate the robustness of our model to recognize humor. Finally, we draw some conclusions and discuss future work.

2. State of the Art

There is a considerable amount of literature on computationally irony and satire detection (see del Pilar Salas-Zárate et al., 2020; Abulaish et al., 2020; Karoui et al., 2019; Joshi et al., 2018). In general,

approaches to deal with these forms of figurative language can be classified into: features-based machine learning approach and deep leaning-based approach. Initially, machine learning method combined with feature-based representations (lexical, contextual, stylistic, affective, discursive, etc.) received the most attention. But, recently, deep learning-based approaches are gaining interest due to the capacity of these models to automatically learn feature representations that are omitted in hand-craft extraction or simply have abstraction levels beyond of human bounds. In this line, the next two subsections survey the relevant works for irony and satire detection. Also, considering the imbalanced number of studies in other languages than English, a third subsection discusses irony and satire in a multilingual setting, with special focus on Spanish.

2.1. Machine learning based approaches

Irony. Computational irony detection has been addressed by the NLP community from different perspectives. In preliminary works, the role of textual-based features obtained from the text (such as n-grams, punctuation marks, part-of-speech tags, among others) has been widely explored for its detection (Carvalho et al., 2009; Davidov et al., 2010; González-Ibáñez et al., 2011; Kunneman et al., 2015; Ptáček et al., 2014). Other works drew attention to theoretical aspects of irony such as incongruity and opposition. Based on these aspects, features derived from semantic ambiguity, synonyms, antonyms and polarity contrast have been studied in (Riloff et al., 2013; Barbieri & Saggion, 2014a,b; Hee, 2017). Many theories seem to agree that an implicit attitude is expressed when being ironic. Aiming to capture the relation between irony and subjectivity in language, several approaches have focused on affective information for improving irony detection (Agrawal & An, 2018; Hernández Farías et al., 2016; Hernández Farías et al., 2015; Barbieri et al., 2014; Reyes et al., 2013). Verbal irony is without doubt a pragmatic phenomenon, hence, contextual and extra-linguistic information result crucial for its detection and comprehension. In this sense, information regarding the context surrounding a given text has been exploited in order to determine whether a text has an ironic or sarcastic intention (Bamman & Smith, 2015; Khattri et al., 2015; Wallace et al., 2015; Ghosh et al., 2018). Discovering new features with discriminative power and topic-independency have been the most active directions of machine learning approaches. Regarding machine learning algorithms, the most used have been Random Forest (RF), Decision Trees (DT), Naïve Bayes (NB) and Support Vector Machines (SVM). Recently, in (Hernández Farías et al., 2020) the impact of the imbalanced distribution of classes in irony and sarcasms detection has been studied from a machine learning perspective.

Satire. Machine learning has been the most used approach for satire detection (Burfoot & Baldwin, 2009; Ahmad et al., 2014; Barbieri et al., 2015a,b; Salas-Zárate et al., 2017). In the seminal paper of Burfoot & Baldwin (2009), the problem of detecting satire was explored with simple bag of words features (BoW) using two feature-weighting methods: i) binary feature weighting and ii) bi-normal separation (BNS) features scaling. Further, lexical (headlines, profanity, slang) and semantic features were added to enrich text representation. To compute the semantic feature they identify the named entities in a given document and query the web for the conjunction of those entities. In this direction, (Ahmad et al., 2014) proposed to extent the BNS features scaling method with the *tf-idf* weighting schema to improve satire detection in news genre.

In (Barbieri et al., 2015b) studied the problem of satire detection in tweets. Linguistic differences between satirical and factual content were explored by mean of frequency, ambiguity, synonyms, part of speech (PoS) tags, sentiments, characters, and slang words as features. Experiments showed that some linguistic features are topic-independent and hence useful clues to address the problem. In a same fashion, a psycho-linguistics approach was introduced in (Salas-Zárate et al., 2017) to identify satirical tweets. A wide variety of psychological and linguistic features from Linguistic Inquiry and Word Count lexicon (LIWC) (Chung & Pennebaker, 2011) were evaluated. Results confirmed the usefulness of emotional, social, and psychological dimension for satire detection. In (Rubin et al., 2016) were considered five predictive features: absurdity, humor, grammar, negative affect, and punctuation, and applied an SVM method. After, combining three out of five features (absurdity, grammar, and punctuation), the authors observed that the BNS feature scaling is suitable for satire detection and the model obtains good results.

Sensibility of lexical, linguistic and n-gram based features across three textual genres was reported in (Reganti & Maheshwari, 2016). Specifically, the impact of features associated on affective words, acts of the

speech, sensorial words, and shallow clues of figurative device (alliteration, grammatical inversion, hyperbole, onomatopoeia and imaginary) was evaluated. Results showed that n-grams and features related with the act of the speech were good as genre-independent and hence they resulted suitable for satire detection in multiples genres. In a similar fashion, an emotions and sentiments based representation was proposed in (Thu & Aung, 2018) for satire detection in newswires, Amazon product reviews and an in-house Twitter corpus. Experiments were performed using the SVM and RF methods. Results concluded the usefulness of the proposed features for satire detection. In (Thu & Nwe, 2017) the impact of emotions on recognizing satirical texts from other figurative forms (humor, irony, sarcasm) and factual language was analyzed.

Recently, satire has received more attention due to the commonalities with the undesirable phenomenon of misinformation in social media, and particularly with fake news spreading (Golbeck et al., 2018; Levi et al., 2019; Guibon et al., 2019). In order to reduce the exposure to misinformation in social media, publishers of fake news have begun to masquerade as satire sites to avoid being demoted. For users, incorrectly recognizing satire as fake news may deprive them of desirable entertainment content, while identifying a fake news story as legitimate satire may expose them to misinformation.

2.2. Deep learning-based approaches

Irony. Recently, many deep learning-based approaches for addressing irony detection have been proposed. Word embeddings, Convolutional Neural Networks (CNNs), att-RNNs, and Transformers-based models have been exploited for capturing the presence of irony in social media content (Ghosh & Veale, 2016; Ghosh et al., 2017; Huang et al., 2017; Joshi et al., 2016; Nozza et al., 2016; Poria et al., 2016; Hazarika et al., 2018; Wu et al., 2018; Baziotis et al., 2018; Zhang & Abdul-Mageed, 2019; Altin et al., 2019; Potamias et al., 2020; González et al., 2020). The semantic and syntactic properties of pre-trained word embeddings have been highlighted in several studies (Joshi et al., 2016; Nozza et al., 2016; Ravi & Ravi, 2018). For instance, word embeddings have been explored to capture incongruity in text with non-affective words (Joshi et al., 2016). In the study (Ravi & Ravi, 2018), irony detection was addressed using a multi-faceted representation which fuses psycho-linguistic features with word embedding vectors that were obtained by using Doc2Vec (Le & Mikolov, 2014a). In (Nozza et al., 2016) the generalization capabilities of an unsupervised topic model trained for irony detection showed a substantial increasing when the word embedding information was incorporated.

Several methods exploited the advantages of CNN for discovering local features that result useful for irony detection. An interesting idea was proposed in (Poria et al., 2016), which introduced a framework for learning irony features from a corpus using CNN. This approach investigated whether features extracted using pre-trained sentiment CNN, emotion CNN and personality CNN models can improve the overall performance. In another direction, the role of the content and contextual information for sarcasm detection taking advantages of a multi-view model were presented in (Hazarika et al., 2018). For that purpose, two CNN models were trained to generate stylometric and personality embeddings for each user's comments. Later, both embedding were fused in a multi-view setting using Canonical Correlation Analysis (CCA) (Hotelling, 1936). A content-based sentence representation was extracted using another CNN and appended with context vectors to obtain the final decision. In another study (Ghosh & Veale, 2016), a model that combines dense neural networks (DNNs) with time-convolution and LSTM (CNN-LSTM) was proposed for detecting sarcasm in tweets. These existing studies use the convolutional network to automatically derive deep features from texts for irony detection. Results of these deep learning-based models are generally better than those obtained with classical feature based machine learning methods.

RNNs have been used for addressing irony detection due to their abilities for capturing long and short dependencies among words within texts. In Ghosh et al. (2017) studied the role played by the conversational context in a sarcasm reply. Particularly, results proved that LSTMs that can model both the context and the sarcastic reply achieve better performance than LSTMs that read only the reply. In another direction, many approaches studied the impact of attentive-based representation with linguistics features (Wu et al., 2018; Kumar et al., 2020). Experiments have concluded that considering hand-crafted features help models to increase their effectiveness. From another point of view, the model introduced in (Zhang et al., 2019), proposed strategies to improve irony detection by transferring knowledge from sentiment resources. This work proposed three different attentive-LSTM approaches that differ in the way of including the sentiment

resources, either injecting the sentiment directly to the attention mechanisms or merging the output of different networks specialized on sentiment analysis and irony detection. In a similar fashion, in (Majumder et al., 2019) a multi-task learning approach was proposed to leverage the knowledge in sarcasm detection and sentiment analysis task. Experiments showed that these two tasks are correlated, and training a deep neural network that models this correlation in a multi-task learning setting improves the performance of both tasks. Moreover, in (Chauhan et al., 2020) a multi-task learning framework for multi-modal sarcasm, sentiment, and emotion analysis was proposed. The authors take advantage of the sentiment and emotions of the speaker to predict sarcasm. In the multi-task framework, sarcasm was considered as the main task, whereas emotion and sentiment detection were used as secondary tasks. Results confirmed that the multi-task framework achieves better performance for the primary task, i.e. sarcasm detection, with the help of emotion and sentiment analysis tasks.

The use of transformer-based models (Vaswani et al., 2017) has changed the way of modeling and working with textual data in an unprecedented way. In fact, these models have been widely spread by means of BERT (Devlin et al., 2019) and other BERT's related architectures (Conneau et al., 2020; Liu et al., 2019; Sanh et al., 2019; Lan et al., 2020). Clearly, this spreading has reached very fast also FL processing: new methods based on transformer models outperformed the state of the art in irony detection by a substantial margin (González et al., 2020; Potamias et al., 2020; Ghosh et al., 2020). In this line, in (Potamias et al., 2020) the RoBERTa model (Liu et al., 2019) was used to encode the sentences, that was further contextualized by means of a Recurrent Convolutional Neural Network to address irony and sarcasm detection. This model outperformed state of the art on four benchmark datasets for irony and sarcasm detection in English. In (González et al., 2020) a simplification of the BERT architecture was proposed to contextualize pre-trained word embeddings. Specifically, this work contextualized Word2Vec word embeddings, trained with several millions of tweets both for English and Spanish. This strategy, opposite to the use of pre-trained BERT, aimed to train the proposed model from in-domain data using the same powerful backbone architecture as BERT. This model outperforms previous models for irony detection in Spanish short texts.

Satire. Notwithstanding a vast amount of deep learning-based methods have been proposed for irony detection, and the commonalities between irony and satire, few methods have addressed the problem of satire detection from a deep learning perspective. Recent works in this direction have been presented in (Yang et al., 2017a; Sarkar et al., 2018; Dutta & Chakraborty, 2019). In (Yang et al., 2017a) a four-level hierarchical network with attention mechanism was presented to differentiate satirical news from true ones. Psycholinguistics, writing stylistic, structural and readability-based features were included to the model at both paragraph and document level. The evaluation suggested that readability features supported the overall classification while psycholinguistic features, writing stylistic features, and structural features are beneficial at paragraph level. The analysis of individual features reveals that satirical news tend to be emotional and imaginative. Another idea was explored in (Sarkar et al., 2018) which proposed to use CNN, LSTM, and GRU to detect satire at both sentence and document levels. They concluded that fine-grained sentence-level analysis provides an in-depth insight into the phenomenon of satire.

2.3. Multilingual approaches

Most of the works on irony and satire detection have investigated the problem in English. Notwithstanding, there have been some efforts to investigate it in other languages such as: Chinese (Tang & Chen, 2014), Czech (Ptáček et al., 2014), Dutch (Kunneman et al., 2015), French (Karoui et al., 2015; Benamara et al., 2017), Italian (Bosco et al., 2013; Barbieri et al., 2016b; Cignarella et al., 2018), Portuguese (Carvalho et al., 2009), Spanish (Rangel et al., 2014; Jasso López & Meza Ruiz, 2016; Ortega et al., 2019), and Arabic (Karouia et al., 2017; Ghanem et al., 2019). Even when in closely related tasks like sentiment analysis have emerged an increasing number of works addressing the multilinguality issue (Singh et al., 2021; Esuli et al., 2020; Galeshchuk et al., 2019; Lo et al., 2017; Abdalla & Hirst, 2017; Dashtipour et al., 2016; Balahur & Turchi, 2012), where few works explored this in the context of irony and satire detection. Taking advantage of the finding achieved for multilingual sentiment analysis would be an interesting direction to improve satire a irony in this scenario.

From a multilingual perspective, the approaches for irony and satire detection can be analyzed in two main directions: i) multilingual setting, where the model is trained and evaluated separately on each language, ii) cross-lingual setting, where the models is trained in one or more language and evaluated on another different one. Multilingual setting has been the most investigated. Prior works where presented in (Ptáček et al., 2014) and (Tang & Chen, 2014) for Czech-English and Chinese-English languages respectively. In (Karoui et al., 2017) a novel fine-grained annotation schema was proposed to annotate irony categories, activators and markers in French, English and Italian language. The role played by dependency-based syntactic features on irony detection from a multilingual perspective (English, Spanish, French and Italian) was investigated in (Cignarella et al., 2020). In the case of satire, in (Salas-Zárate et al., 2017) the authors investigated the impact of psycho-linguistic features in two distinct variants of the Spanish (Mexican and Castilian). Irony detection from a cross-lingual perspective (Arabic, French and English) was investigated in (Ghanem et al., 2020). Results showed that, although irony is contextual, language and cultural-depended pragmatic phenomenon, several features are universal and can be useful for addressing irony detection in languages which lack of annotated data. In the same line, in (Barbieri et al., 2015a) the authors presented a set of language independent features that describe lexical, semantic and usage-related properties of the words in the tweets. The proposed features were evaluated in a cross-lingual setting. Results highlighted the complexity of modeling satirical texts in a cross-lingual setting, due to satire aims at criticizing social and moral behaviors which often are social and cultural-dependent.

3. Multiview informed attention-based models

In this section we introduce MvAttLSTM, our multi-view informed attentive LSTM model for irony and satire detection in Spanish. We addressed both tasks as binary classification problems applying a model based on LSTMs endowed with an attention mechanism. LSTM is an RNN that uses gating mechanisms to overcome the problem of the vanishing gradient. This type of neural networks can capture long-range relationships and hidden patterns in sequential data. In terms of architecture, the Bidirectional LSTM (BiLSTM) (Zhou et al., 2016) is widely used, which has two LSTM units processing sequences forward and backward respectively. This property of BiLSTM is useful for language processing because the meaning of the words in texts can be inferred not only by previous words, but also considering other words after them can help to determine their meanings. Moreover, attention mechanisms have endowed the RNNs with a powerful strategy to enhance their performance and achieve better results. Our model considers multiple representations learned from three distinct perspectives: linguistic-based representation, universal sentence encoder-based and contextualized pre-trained embeddings. We introduce additional knowledge into MvAttLSTM model aiming at reinforcing linguistics and semantics properties which can result beneficial for detecting irony and satire. Concretely, our model is compounded by an embedding layer which is fed into a BiLSTM layer. Later, the hidden states sequence returned by the BiLSTM is fed into an attention layer. Next, on the output of this layer are staked two LSTM layers. Finally, we incorporate a feed forward neural network for final prediction. As explained before, we inform the model with three additional views. Particularly, we investigate two different strategies for fusing these views into our MvAttLSTM. In the next subsections we present in detail the prepossessing carried out on the datasets, the additional representations, the main parts of the MvAttLSTM's architecture and the strategies for informing the model.

3.1. Preprocessing

Social media texts, particularly those from Twitter are informal and noisy. The length constraints, and the free writing style present in this form of online communication provoke that texts have plenty of grammar and spelling mistakes. Particularly, length constrains caused that users use shortenings, abbreviation, homophonic encoding to save characters, and grammar and spelling misuses such as: character flooding, word repetition and wrong use of uppercase letters to denote emphasis. Twitter also offers to the users reserved symbols to mark explicitly important concepts in tweets (*# hashtag*), to refer o mention other users (*@ mention*), to reply the message of other users (*RT retweets*) or simply to mark texts as favorite (*FAV*). Aiming to inject emotional states tone and body language into tweets, emoticons and emojis became

very popular. These symbols are an ultra-concise way to enrich writing language with visual information. All these issues impact sentence structure, content, word forms and increase the difficulty of their automatic processing and comprehension. In order to mitigate the effects of these problems, in our model we applied a basic preprocessing phase for cleaning the texts. Firstly, we applied a tokenization process on the tweets by using the TokTokTokenizer from NLTK (Perkins, 2014). Later, emoticons, emojis, URLs, hashtags, mentions are recognized and replaced by a corresponding wildcard which encodes the meaning of these special words. In the case of hashtag, we replaced the reserved symbol (#) by the word *topic_* and retain the remaining characters. Emoticons and emojis were replaced by the word *emoji_* concatenated with an integer value associated to each emoji. We have included these changes in order to reduce the impact of noisy and inconsistent writing on the processing of the text with the Freeling tool (Padró & Stanilovsky, 2012). Moreover, we replaced each mention and URL by the words *author_token* and *url_token* respectively. Finally, Twitter-reserved words like RT (for retweet) and FAV (for favorite) were removed. It is worth to notice that emoticons and emojis are a valuable source of information to take into account in social media content analysis (Barbieri et al., 2016a,c,d, 2017a,b; Pota et al., 2021a,b). Nevertheless, in this work we used emojis and emoticons to create features for capturing the frequency of positive emojis, negative emojis and neutral emojis as well as detecting polarity contradiction between the words and emojis in the text. In the second stage, and used only to obtain some linguistic features that were considered in the *Linguistic-view*, a more complex language analysis was carried out. For that, flooding tokens were normalized allowing the same character to appear only twice consecutively in a token (e.g. *hooolaaa* becomes *hoolaa*). Afterwards, tweets were morphologically analyzed with the FreeLing tool. In this way, for each resulting token, its lemma and part-of-speech were considered.

3.2. Addition knowledge to inform the model

Our MvAttLSTM relies on fusing multiple representations which are learned from distinct perspective. Specifically, we learn three independent views for each text which are introduced to the model. The first one (*Linguistics-view*) consists in several linguistics features which have proved to be strong cues for discriminating both irony and satire. The second one considers a deep dense encoding of the message using Multilingual Universal Sentence Encoding (*MUSE-view*) (Cer et al., 2018; Yang et al., 2020). And, finally the last one (*BERT-view*) considers a contextualized pre-trained embedding obtained after a tuning of the BERT model (Devlin et al., 2019). Next, we describe the three ways in which each view was learned.

3.2.1. Linguistics-view

Hand-crafted features, often linguistic-based, have proved to be effective for processing figurative language, particularly in case of irony, satire and humor (Wallace, 2015; del Pilar Salas-Zárate et al., 2020; Abulaish et al., 2020; Karoui et al., 2019; Joshi et al., 2018). From our perspective, the linguistics-based representation is able to capture certain types of irony, satire and other figurative devices disregarding textual genres and topics, and it makes this representation content independent and genre-unbiased. To represent each text, we use different group of features: stylistic and structural, semantics, affective, incongruity and psycho-linguistic. Many features are extracted to identify stylistic patterns in the structure of the ironic or satirical texts (e.g., type of punctuation, length, emoticons, distribution of nouns, adjectives, adverbs and verbs). Other features are extracted to consider affective information (e.g., polarity, sentiments, emotions, attitudes, etc.) by using several word-based lexicons resources. Moreover, features are extracted for considering semantics properties of texts (e.g., co-occurence of synonyms and antonyms, maximum, minimum and mean of synsets, etc). Finally, some features are designated to capture contrast and opposition in texts (e.g., polarity contrast, semantic incongruity, etc.). Specifically we use the features proposed in (Ortega-Bueno et al., 2018b, 2019), the incongruity features used in (Ortega-Bueno & Medina Pagola, 2018) but using BabelSenticNet (Vilares et al., 2018) as default polarity lexicon and including the emotional dimensions and the polarity feature in this resource as other affective features. BabelSenticNet is an extension of SenticNet (Cambria et al., 2020) to 40 other languages, including Spanish (henceforth we refer the Spanish version of BabelSenticNet as SenticNet). For more details about the linguistic features considered in this work please see Appendix A.

3.2.2. Task-independent embedding view

Our second representation aims at encoding the whole meaning of the text into a single dense vector based on deep learning models. Particularly, in vectors which capture rich semantic information that can be useful for recognizing semantics proprieties of the ironical and satirical texts. A contextual approach for creating the embedding vectors is proposed in (Cer et al., 2018), where complete sentences, instead of words, are mapped into a latent vector space. The approach provided two variations of Universal Sentence Encoder (USE) with some trade-offs in computation and accuracy. The first one consists of a computationally intensive transformer that resembles a transformer network (Vaswani et al., 2017), proved to achieve a higher performance. In contrast, the second one provides a lightweight model that averages input embedding weights for words and bi-grams by utilizing of a Deep Average Network (DAN) (Iyyer et al., 2015). The output of DAN is passed through a feed-forward neural network in order to produce the sentence embedding. Both approaches take as input lower-cased strings and output a 512-dimensional sentence embedding. Although there are several methods like Doc2Vec (Le & Mikolov, 2014b), Sent2Vec (Pagliardini et al., 2018), FastText (Bojanowski et al., 2017) and InferSent (Conneau et al., 2017) to generate sentence embeddings we used Multilingual Universal Sentences Encoding (MUSE)[1] (Yang et al., 2020) which is an extension of USE trained for 16 languages including Spanish. The most salience characteristic of this model is that it was trained using multi-task learning to integrate semantic information. Particularly, sentence embeddings are learned across several languages and using multiple semantic tasks like sentiment analysis, semantic textual similarity, etc. This enables the learning process to dynamically accommodate a wide variety of knowledge in a single vector which is interesting to transfer to related tasks like irony and satire. Based on the MUSE model we transform the texts of the training dataset into dense vectors of 512 dimension (henceforth, H_{MUSE}). It is important to highlight that H_{MUSE} is a completely task-independent representation, because we do not apply any parameters tuning of the model on the training data.

3.2.3. Task-dependent embedding view

The transformer-based neural network architectures (Vaswani et al., 2017) paved the way for training robust and contextual-aware language models in an unsupervised manner. The use of these pre-trained contextualized models have been widely spread through BERT (Devlin et al., 2019) and BERT's family architectures (Conneau et al., 2020; Liu et al., 2019; Sanh et al., 2019; Lan et al., 2020). To study the linguistic and semantic nuances of irony and satire in Spanish, we decide to incorporate BERT as another representation (view). BERT relies on bidirectional representation from transformers and achieves the state of the art for contextual language modelling and contextual pre-trained embeddings. This model is trained on a large text corpus and then used for downstream NLP tasks. While other word embedding like Word2Vec (Mikolov et al., 2013), Glove (Pennington et al., 2014) and FastText (Bojanowski et al., 2017) are context-free models that produce a single word embedding for each word in the vocabulary, BERT computes a representation of each word that is based on the other words in the context. It was built upon recent works in pre-training contextual representations, such as ELMo (Peters et al., 2018) and Universal Language Model Fine-tuning (ULMFiT) (Howard & Ruder, 2018), and is deeply bidirectional. BERT represents each word using both its left and right contexts. Moreover, it is possible to fine-tune BERT for many downstream NLP tasks, including the tasks we are interested in. This goal can be achieved by removing the language modelling output layer (masked word prediction) and replacing it with a new layer appropriate for the target task (in our case, binary classification). Particularly, in this work we use the pre-trained multilingual versions of BERT[2] (mBERT, henceforth) and carried out a fine-tuned on it, using the training datasets of irony and satire. Our idea is not to use this model for as a classification method; instead, we considered it for the representation purpose.

For fine-tuning mBERT, we add a layer that receives as the input the vector in the first position (the *CLS* token). On this layer, we stacked an output layer that makes the final prediction for the targeted task. For that purpose, we follow the strategy proposed in ULMFiT (Howard & Ruder, 2018). For each layer of

[1] https://tfhub.dev/google/universal-sentence-encoder-multilingual/3
[2] https://huggingface.co/bert-base-multilingual-cased

mBERT, a different learning rate is set up, increasing it using a multiplier while the neural network gets deeper. This multiplier increases 0.1 points from a layer L_i to another L_{i+1}. We use this dynamic learning rate to keep most information from the pre-training at shallow layers and biasing the deeper ones to learn about the specific tasks. For all corpora, the same hyperparameters were used. Concretely, we defined the $batch\_size = 32$ and the sequence length was limited to 50 tokens. The optimizer used is $Adam$ (Kingma & Ba, 2015) with an initial learning rate of 0.001, $\beta_1 = 0.9$ and $\beta_2 = 0.999$ and a $weight\_decay = 0.01$. The model was trained during 15 epochs and using the ModelCheckpoint callback for obtaining the model that has achieved the best performance on the validation subset.

After tuning mBERT, we pass again the training dataset, but this time, we get the deep representation of each text of the training dataset (henceforth, H_{BERT}). This view is task-dependent because we refine the parameters learned by mBERT in order to capture semantics and pragmatics characteristics, which results crucial for understanding and recognizing ironic and satirical intents.

3.3. MvAttLSTM model

Let us describe the architecture of the MvAttLSTM model. We give details about each layer, starting from the embedding layer to the loss function.

3.3.1. Embbeding Layer

In this layer each word w_i is map into a highly dimensional feature space for capturing the meaningful semantic and syntactic information. Given an input text T which consists of at most N words w_i, where $i \in [1, N]$. For each word into T, we examine the embedding matrix $E \in R_{B \times d}$, where B is the length of the vocabulary, and d is the dimension of word embedding vectors. The matrix E can be initialized randomly or by means of a word embedding matrix. In this work we decided to initialize the embedding layer with context-free pre-trained word representations. For that, we learned the embedding matrix E by using the FastText model trained on the Spanish Billion Words Corpus[3] and an in-house background corpus of 9 millions of Spanish tweets. We aim to join both corpora for obtaining robust word representations taking advantage of the peculiar writing style used in Twitter. For training the FastText model we used the setting reported in (Bojanowski et al., 2017), except for the value of the vector size, which was defined as a 300-dimensional. In this layer, each word w_i is transformed into a vector $x_i \in \mathbb{R}^d$:

$$H^0 = Embbeding(E, M) \tag{1}$$

Thus, every text T can be converted in a sequence of vectors, in the form of a 2d-matrix $H^0 = [x_1, x_2, x_3, ..., x_N]^T$ with shape $N \times d$. The matrix H^0 is given as input to the next layer. It is worth noting that in the model the weights in E are fixed. We aim at making the model to be trained faster and mitigate the impact of the overfitting due to the reduction of parameters that must be learned. Moreover, we consider that the recurrent and attention layers are feasible to take advantage of the semantic and syntactic properties of the vectors in E for classifying irony and satire.

3.3.2. BiLSTM Layer

After passing the sequence of word T to the Embedding layer, each word w_i is encoded by a vector x_i which captures the semantic and syntactic properties of w_i out of context. In other words, the representation of each w_i is independent of the other words in the text T. In this layer, a new representation for each word is learned by summarizing the contextual information, previous and after to the word in the text. For achieving this goal we use a BiLSTM layer. This, type of neural network consists of two LSTM units which process the sequential input in both directions forward and backward simultaneously.

$$H^1 = BiLSTM(H^0) \tag{2}$$

[3] https://crscardellino.github.io/SBWCE/

The output of this layer is a sequence of hidden states $H^1 = [h_1^1, h_2^1, ..., h_N^1]$ where each $h_i^1 \in \mathbb{R}^{2\times d_h}$ is the concatenation of the hidden state of each LSTM (right and left), specifically, the $h_i = [\overrightarrow{h_i^1}, \overleftarrow{h_i^1}]$, and d_h is the number of hidden neuron into the LSTM unit. Standard LSTM receives sequentially (in a left to right order) at each time step a word vector x_i and produces a hidden state h_i. For that, this neural network relies on a cell of memory and a gating mechanism consisting of an input gate, forget gate, and output gate. These gates help to determine whether the information in the previous state should be retained or forgotten in the current state. Hence, the gating mechanism helps the LSTM to cope with long-term information preservation. Each hidden state h_i is determined as follows:

$$I_t = \sigma(W^i x_t + U^i h_{t-1} + b^i) \tag{3}$$

$$F_t = \sigma(W^f x_t + U^f h_{t-1} + b^f) \tag{4}$$

$$O_t = \sigma(W^o x_t + U^o h_{t-1} + b^o) \tag{5}$$

$$\tilde{C}_t = \sigma(W^u x_t + U^u h_{t-1} + b^u) \tag{6}$$

$$C_t = i_t \odot \tilde{C}_t + F_t \odot C_{t-1} \tag{7}$$

$$h_t = O_t \odot tanh(C_t) \tag{8}$$

Where all W^*, U^* and b^* are parameters of the recurrent layer which are learned during the training phase and the x_t is the pre-trained vector of the word in the time-step t and it is not trained in the model. The operator σ is the sigmoid function and the operator $\odot$ stands for element-wise vector multiplication. The I_t, F_t and O_t are the input, forget and output gates in the time step $t-1$ whereas $\tilde{C}_t$, C_t and h_t are the new cell, the updated cell memory and the final hidden state in the time step t. Notice that the BiLSTM initial hidden states and cells memory are set to 0 in both directions $\overrightarrow{c_0^1} = \overleftarrow{c_0^1} = \vec{0}$ and $\overrightarrow{h_0^1} = \overleftarrow{h_0^1} = \vec{0}$. We highlight this detail because we use these states as the way to incorporate additional information into the MvAttLSTM model.

3.3.3. Attention Layer

The BiLSTM Layer has two major problems. Since the meaning of the message cannot be encoded in one fixed-size vector, there is some information loss. Hence, the performance of this type of models for representation learning decreases when the length of inputs become large. Another concern is that LSTMs aggregate information word-by-word in sequential order, but there is no explicit mechanism to make inferences over the structure and modeling relations among tokens. To overcome these limitations, the output of the BiLSTM Layer $H^1 \in \mathbb{R}^{2\times d_h \times N}$ is fed into an Attention Layer. This layer helps BiLSTM in deciding which parts of the sequence pay more interest. In this work, we investigate the performance of two attention mechanisms in our model: self-attention and multi-head attention (Vaswani et al., 2017).

Self-attention mechanisms can capture the explicit and latent relations among words beyond their sequential order. While attention mechanism (Bahdanau et al., 2015) allows the outputs for attending some parts of the inputs. Self-attention also allows the inputs for interacting each other, hence amplifying the importance of each one plays in determining the meaning of others. Moreover, is it beneficial for discovering word relations which can be crucial for understanding ironic and satirical texts such as, oppositions and incongruities. Given the matrices A, B and C, mathematically self-attention is formulated as follows:

$$Att(A, B, C) = Attention(AW^Q, BW^K, CW^V) \tag{9}$$

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{10}$$

Where $W^Q \in \mathbb{R}^d$, $W^Q \in \mathbb{R}^d$, $W^Q \in \mathbb{R}^d$ are the projection matrices for the query Q, key K and value V. In the case of self-attention, the matrices A,B and C are the same. Thus, given the output H^1 of the BiLSTM layer, the new sequence of weighted hidden states $H^2 \in \mathbb{R}^{2\times d_h \times N}$ which is the output of the Attention layer is computed by Eq. 11 as:

$$H^2 = Att(H^1, H^1, H^1) \tag{11}$$

In (Vaswani et al., 2017) another attention mechanism was introduced. The multi-head attention mechanism uses multiple individual attention functions (heads) for obtaining different contexts and paying attention simultaneous to distinct aspects in the sequences. This can jointly pay attention to information from different representation sub-spaces at different positions. Like in self-attention, the attention function takes as input a matrix for the query A, a matrix for the keys B and a matrix for values C. The multi-head attention model first transforms A, B and C into $\mathbb{C}$ sub-spaces, with different, trainable linear projections:

$$MultiHead(A, B, C) = [head_1, head_2, ..., head_r] * W^0 \tag{12}$$

$$head_c = Attention(AW_c^Q, BW_c^K, CW_c^V) \tag{13}$$

Where $W_c^Q \in \mathbb{R}^{d \times d_k}, W_c^K \in \mathbb{R}^{d \times d_k}, W_c^V \in \mathbb{R}^{d \times d_v}$ are projection matrices for the inputs A, B, C with respect to $head_c$, and $W^0 \in \mathbb{R}^{r \times d_k \times d}$. The parameter r is the number of heads for the multi-head attention mechanism; and $head_c \in \mathbb{R}^{N \times d_k}$ is the output of the c^{th} head. Notice that, for each $head_c$, the weights of W_c^Q, W_c^K, W_c^V are independently learned during the training phase. The attention for each head c (see Eq. 13), like in self-attention, is computed by the formula in Eq. 10. Thus, given the output of the BiLSTM layer H^1, the output of the multi-head attention is computed as follows:

$$H^2 = MultiHead(H^1, H^1, H^1) \tag{14}$$

3.3.4. LSTM Layers

Even when, it is not theoretically clear what is the additional power gained by the deeper recurrent architectures, it was observed empirically that a deep LSTM works better than shallower ones in some tasks (Irsoy & Cardie, 2014; Wu et al., 2018). Taking this into account, on the output H^2 of the Attention Layer we stacked two layers of LSTMs to deep contextualize the previously learned representation. This means that the output of the first LSTM layer is given as input to the second one. The two LSTM layers are defined as follows:

$$H^3 = LSTM(H^2) \tag{15}$$

$$H^4 = LSTM(H^3) \tag{16}$$

Where $H^3 \in \mathbb{R}^{d_{k1}N}$, $H^4 \in \mathbb{R}^{d_{k1}N}$ are the outputs of the first and second LSTM layer and d_{k1} is the number of hidden neurons into LSTM cells. The last LSTM layer output the hidden representation of the text. Particularly, we only consider the last hidden state (h_N^4) into the matrix H^4. Let us redefine it as h_{last}^4 henceforth. Like in the BiLSTM layer, the initial hidden state and cell memory of both LSTMs are set to 0, $h_0^3 = c_0^3 = \vec{0}$, and $h_0^4 = c_0^4 = \vec{0}$.

3.3.5. Fusion Strategies

Our model aiming at improving irony and satire detection in Spanish by incorporating multiple views into the MvAttLSTM. For this purpose, we investigate two strategies for passing the views to the model. The first one, *Early Fusion* method, aiming at enriching the representation learned by the LSTMs with additional knowledge using the last dense layers. The second one, *Contextual Fusion* method, which aims to condition the learning process of the LSTMs with prior knowledge injected in the initial cell memory. Next, we give details about both strategies.

Early Fusion

The main idea behind this strategy is to separately learn different features spaces from the training data. These feature spaces (views) capture distinct characteristics of the same texts. We aim to jointly use

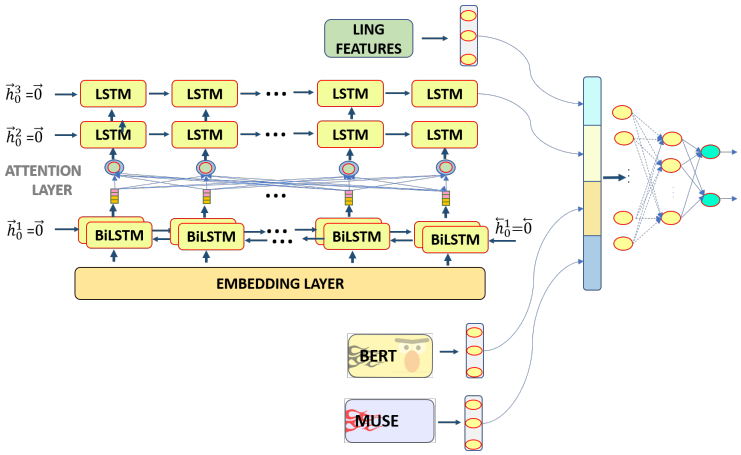

Figure 1: MvAttLSTM: Multi-view informed attentive LSTM deep neural network using an early fusion strategy

these representations to retain discriminant information while reducing the redundant one. The overall architecture of the MvAttLSTM with *Early fusion* is showed in Fig. 1.

Firstly, let us define H_{Ling}, H_{MUSE} and H_{BERT} as the Linguistic-view, MUSE-view and BERT-view respectively (see Section 3.2). These views differ from each other in the way by which were learned and the number of features used for encoding the text. Thus, we pass each view to a dense layer in order to reduce and unify the views' dimensionality using the Eq. 17, 18, 19:

$$g_l(H_{LING}) = \sigma(W^l H_{LING} + b^l) \tag{17}$$

$$g_m(H_{MUSE}) = \sigma(W^m H_{MUSE} + b^m) \tag{18}$$

$$g_b(H_{BERT}) = \sigma(W^b H_{BERT} + b^b) \tag{19}$$

Where W^l, W^m, W^b and b^l, b, b^b are parameters of the model to be learned during the training process, and σ is the sigmoid function. After having reduced representations for each view (g_l, g_m, g_b), then we concatenate them with a deep representation learned by the attentive LSTM based architecture h^4_{last} (Eq. 20). Later, the merged representation denoted as F_0 is fed into a dense layer with sigmoid activation for fusing all views into a new non-linear space using Eq 21. Finally, the output of this layer denoted as F_1 is a multi-view encoding of the texts, and it is fed into a feed-forward neural network for the final classification of the texts in ironic vs. non-ironic or satirical vs. non-satirical.

$$F_0 = Concat(g_l, g_m, g_b, h^4_{last}) \tag{20}$$

$$F_1 = \sigma(W^0 F_0 + b^0) \tag{21}$$

Contextual Fusion

In this strategy, we enrich our MvAttLSTM with additional external knowledge to take advantage of the initial memory cell in the LTSMs. We experiment with a strategy similar to the conditional encoding model introduced in (Rocktäschel et al., 2016) for the task of recognizing textual entailment and applied later in (Ghosh et al., 2018) for modeling conversation context for improving sarcasm detection. Conversely, to the approach presented by Ghosh et al. (2018), we do not learn contextual information by using LSTMs, instead, we learn independently three distinct views with the aim to capture syntactic, semantic, and pragmatics aspects used in ironical and satirical texts. In Fig. 2 is showed the overall architecture of our MvAttLSTM

model using the *Contextual Fusion* strategy. As can be observed, the learning process of each LSTM is conditioned to prior information passed to the initial memory cell.

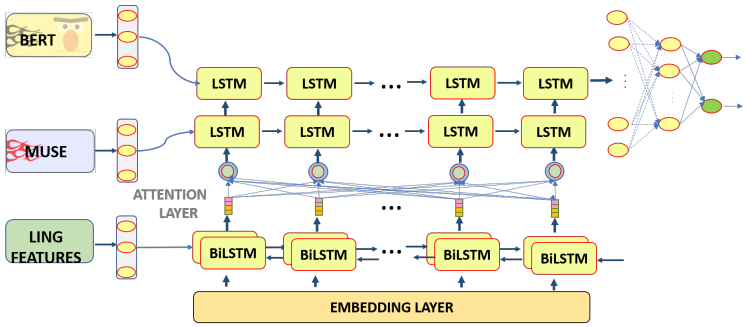

Figure 2: MvAttLSTM: Multi-view informed deep attentive LSTM neural network using contextual fusion strategy

Like in the *Early fusion* strategy, firstly each view is fed into a dense layer with non-linear activation, specifically using the Eq. 17, 18, 19. Later, each reduced representation (g_l, g_m, g_b) is used to inform the LSTM in the MvAttLSTM. The initial memory cell and hidden states of the LSTMs are used as input to pass the prior knowledge of each view as defined in Eq. 22, 23, 24. Where c_0^1 and h_0^1 are the initial memory cells and hidden states of the BiLSTM. And, h_0^3, c_0^3, h_0^4 and c_0^4 are the initial memory cell and hidden states of the second and last LSTM respectively. The order in which each view is assigned to the LSTMs was empirically defined. We decide to introduce low-level linguistic features for reinforcing the BiLSTM layer which aims at capturing language generalization. In the second LSTM, we propose to introduce the MUSE-view for incorporating a high-level semantic representation to encode the global meaning of the text. Finally, in the last LSTM, we introduce the BERT-view to incorporate semantics and pragmatics abstractions useful for the task to solve, considering that in this view the BERT model is tuned using the same training data available for the task. This introduces a task-dependent bias in the language representation learned by the original model:

$$c_0^1 = h_0^1 = g_l(H_{LING}) \tag{22}$$

$$h_0^2 = c_0^2 = g_m(H_{MUSE}) \tag{23}$$

$$h_0^3 = c_0^3 = g_b(H_{BERT}) \tag{24}$$

Notice that in this case, the final multi-view representation of the text F_1 is the same that the last hidden state of the last LSTM layer $h_l^4 ast$ in our MvAttLSTM, hence $F_1 = h_l^4 ast$. And, we pass this representation into the feed forward neural network for the classification of the texts in ironic vs. non-ironic or satirical vs. non-satirical.

3.3.6. Feed-Forward layer for final classification

For achieving the final classification we fed the multi-view encoding of the text F_1 into a Dropout layer to prevent the model's over-fitting. Subsequently, the output of the Dropout layer F_2 is passed to a dense layer with ReLU activation, and finally, the output of this layer F_3 is given as input to another dense layer with two neurons, but this time with the *softmax* function for the prediction:

$$F_2 = \text{Dropout}(F_1) \tag{25}$$

$$F_3 = max(0, W^2 F_2 + b_2) \tag{26}$$

$$O = softmax(W^3 F_3 + b_3) \tag{27}$$

The MvAttLSTM model can be trained in an end-to-end way by the back-propagation method, and we use categorical cross-entropy as the loss function. This function can be observed in Eq. 28, where $\mathfrak{D}$ is the dataset, $\mathfrak{L}$ is the loss function, f is our model parameterized by θ and $\mathbb{G} = \{1, 0\}$ is the set of labels in the task.

$$\mathfrak{L}(\theta) = \mathbb{E}_{\mathfrak{D}}[\mathfrak{L}(f(x, \theta), y)] = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{|\mathbb{G}|} y_{ij} * \log(f(x_i, \theta)_j)w_j \tag{28}$$

4. Experiments and Results

4.1. Datasets Description

In order to validate our proposed model for irony and satire detection in short texts written in distinct Spanish variants, we used three corpora, one for irony detection, and two for satire detection. Moreover, we also tested the robustness of our model on humor recognition on another corpus. They have been extensively used with the aim of training and evaluating state-of-the-art systems for irony, satire and humor detection in Spanish.

Irony Corpus

For what concerns irony detection, we decided to use the corpus proposed in the *IroSvA'19* shared task (Ortega et al., 2019). This is the first public available corpus for irony detection in Spanish. The *IroSvA'19* shared task, framed in the Iberian Languages Evaluation Forum *(IberLEF'19)*[4] and co-located within *SEPLN 2019*[5] aimed at investigating whether a short message, written in Spanish, is ironic or not within a given context. For that, three corpora with short texts from Spain, Mexico and Cuba were proposed with the purpose of exploring the way irony changes in Spanish variants. In particular, the Castilian and Mexican corpora consist of ironic tweets about 10 controversial topics for Spanish and Mexican users. In the case of the Cuban corpus, it consists of ironic news comments which were extracted from 113 controversial news about social, economic, and political issues concerning the Cuban people. It is worthy to notice that, for each text a context is provided, consisting of a short description about the topic, which defines its scope. The distribution of the texts is showed in Table 1.

Table 1: IroSva'19 distribution for ironic and non-ironic classes

Corpus	**Variant**	**Training**			**Testing**		
		Non-Ironic	Ironic	Total	Non-Ironic	Ironic	Total
IroSvA'19	Castilian (es)	1600	800	2400	400	200	600
	Mexican (mx)	1600	800	2400	401	199	600
	Cuban (cu)	1600	800	2400	400	200	600

As can be observed in Table 1 all subcorpora are composed of 3000 texts split into 2400 and 600 texts for training and testing respectively. The training set is separated into 800 ironic and 1600 non-ironic texts, whereas the testing partition is divided into 200 ironic and 400 non-ironic texts. Notice that both training and testing sets maintain the ratio of *2/3 vs. 1/3* between non-ironic and ironic text. In order to assess the performance of the systems, the evaluation metrics used by the organizers were precision (P), recall (R), and F1 score. These metrics were calculated per class and macro-averaged. Due to the imbalance between the non-ironic and ironic classes, the macro-averaged F1 score was used as the overall metric to rank the participating systems.

[4] https://sites.google.com/view/iberlef-2019
[5] http://www.hitz.eus/sepln2019/?language=es

Satire Corpora

For satire detection, we used the corpora proposed in (Salas-Zárate et al., 2017) and (Barbieri et al., 2015a). Both corpora were created using a self-annotation strategy. Specifically, Barbieri et al. (2015a) retrieved tweets from popular satirical news accounts and from legitimate news sources in three languages: Spanish, English, and Italian. In this work, we were interested in the Spanish subset of this corpus, and we refer to this as *Barbieri'15-es* henceforth. The Spanish tweets (Castilian variant) were gathered from two satirical Twitter's accounts *El Mundo Today* and *El Jueves* whereas non-satirical tweets were retrieved from the legitimate newspaper Twitter's accounts *El Mundo* and *El Pais*. Later, a shallow cleaning process was carried out on data for filtering those tweets that were not relevant to satire analysis. As can be observed in Table 2, the corpus is composed of 10888 uniformly distributed in 5444 satirical tweets 5444 non-satirical ones.

The corpus introduced in (Salas-Zárate et al., 2017) was guided by the same methodology presented in (Barbieri et al., 2015a). The most salience difference relies on the study of satirical tweets in two variants of the Spanish. Particularity, tweets from Mexican and Castilian Twitter's accounts were retrieved. For investigating how satire is realized in Mexican tweets, data from four Mexican Twitter accounts were retrieved. The satirical tweets were obtained from *El Deforma* and *El Dizque* satirical accounts whereas the non-satirical tweets were gathered from legitimate newspaper accounts *El Universal* and *Excelsior*. We refer to this subset of data as *Salas'17-mx* henceforth. The tweets in the Castilian corpus of (Salas-Zárate et al., 2017), *Salas'17-es* henceforth, were retrieved using the four Twitter's accounts proposed in (Barbieri et al., 2015a). It is important to note that even when the Twitter's accounts used to obtain the tweets were the same, the tweets in each collection are different.

An automatic cleaning process was carried out on the data. Specifically, retweets, duplicates, tweets only with URLs, and tweets written in a language other than Spanish were removed. Moreover, a manual inspection was performed in order to ensure that the tweets obtained were relevant for satire detection. In Table 2 can be observed that both corpora contain 5000 tweets, which are uniformly distributed in 2500 satirical and 2500 non-satirical ones. The different characteristics of the *Barbieri'15-es* and *Salas'17-es* will allow us to validate the robustness of our *MvAttLSTM* model.

Table 2: Distribution for satirical and non-satirical classes in *Salas'17* and *Barbieri'15* datasets

Corpus	**Variant**	**Data**		
		Non-Satirical	Satirical	Total
Salas'17	Castilian (es)	2500	2500	5000
	Mexican (mx)	2500	2500	5000
Barbieri'15	Castilian (es)	5444	5444	10888

Humor Corpus

For further investigating the robustness of our model we decided to evaluate it on humor recognition in Spanish. We considered the corpus proposed in the *HAHA'19* shared task (Chiruzzo et al., 2020, 2019) organized at *IberLEF'2019* and co-located within *SEPLN 2019*. Two subtasks were proposed, one for humor binary classification (*Humor Recognition*) and another for predicting how funny is a tweet into 5-star ranking (*Funniness Score Prediction*), considering that the tweets present humorous content. The organizers provided a human-annotated corpus of 30000 Spanish tweets separated into 24000 for training and 6000 for testing. The training subset consists of 9253 humorous and 14747 non-funny tweets, whereas the testing subset consists of 2342 humorous tweets and 3658 non-funny ones. In Table 3 we summarize the distribution of the tweets within HAHA'19. Taking into account the scope of this work, we are only interested in the first subtask. As can be noted, in both training and testing subsets the distribution of the classes are slightly unbalanced, hence a difficulty is added to the learning algorithm. The performance metrics used to rank the participated systems in the *Humor recognition* subtask were F1 score for the *humorous* class and accuracy *(Acc)*.

Table 3: HAHA'19 distribution for humorous and non-humorous classes

Corpus	language	Training			Testing		
		Non-Humorous	Humorous	Total	Non-Humorous	Humorous	Total
HAHA'19	Spanish	14747	9253	24000	3658	2342	6000

4.2. Experimental Setting

We use the same architecture for all tasks, but we calibrated the hyper-parameters independently for each corpus. Specifically, we defined the number of hidden neurons in the BiLSTM layer to 64, the number of hidden neurons in the last two LSTM layer to 128, the maximum number of epochs and length of the sequence $ep = 50$ and $N = 50$ respectively. For the remainder of hyperparameters, we experimented with distinct values. Specifically, we defined the search space as follows: batch size $batch \in [32, 64, 128, 265]$, dropout $dp \in [0.25, 0.30, 0.35, 0.4]$, optimizer update rules $op \in (adam, rmsprop)$, learning rate $lr \in [1 \times 10^{-2}, 1 \times 10^{-3}, 1 \times^{1} 0{-}4]$. When the model uses multi-head attention we evaluated distinct number of heads $h \in [2, 4, 8, 16]$. In an intent to prevent the overfitting in the training step, an early stopping with the patience of 10 epochs was used as a stopping criterion. We explored the search space by means of the Grid Search strategy. Analysing the best hyperparameters obtained for each corpus and model, we observed that our models are sensitive to the hyperparameter setting. Particularly, we noted that the learning rate $lr = 1 \times 10^{-2}$ achieved the best performance across all corpora. However, the remainder of the hyperparameters has a distinct behaviour. Roughly speaking, we appreciated that the $MvAttLSTM_{self}Contextual$ and $MvAttLSTM_{multi}Contextual$ are the most sensitives models. One possible reason for that is the small number of ironic examples in each corpus which makes the model generalization more complex. In the case of the $MvAttLSTM_{self}Early$ model, it was observed that it performs well on large corpora using short batches ($batch_{size} = 32$) whereas $MvAttLSTM_{multi}Early$ requires longer batches ($batch_{size} = 128$). Also, we observed that 4 heads of attention were enough to achieve good results. Concerning the optimizer, generally, the *Adam* rule obtained the best result in 9 settings out of 16. The best hyperparameters for each corpus and model are summarized in Appendix B.

In order to evaluate the performance of the distinct settings of our model, we define a baseline method *(Bert-baseline)*. Concretely, we use the mBERT method fine-tuned on each corpus separately. For that, we adopt the same hyperparameter and tuning strategy proposed in Section 3.2.3.

Regarding the *Linguistic-view*, we experimented with distinct views to investigate whether some groups of features are more feasible than others to detect irony cues. In this sense, we defined three views for considering affective information: *Aff_All, Aff_Emo, Aff_App*. In *Aff_All* we considered all features related to polarity, emotions (categorical, and dimensional), and attitudes. Whereas in *Aff_Emo* we only used those features related to emotions, and in *Aff_App* we only use the attitude words. The features that capture polarity oppositions were included in the group *Contrast*. We evaluated two groups (*LIWC, Sverb*) based on the psycho-linguistic dimensions in the LIWC dictionary and the semantic classes of verbs in the ADDESE lexicon[6] respectively. Moreover, other groups of features were obtained by using a feature selection method, specifically the Wilcoxon rank-sum test (Haynes, 2013) were explored. With this statistical test, the features were ranked considering their $p - value$, and three groups were defined. In the groups *W_64*, *W_128* and *W_All* we considered the subsets of 64 and 128 best-ranked features and all features with $p - value \leq 0.05$. Finally, a group with all the linguistic features *LingAll* was considered.

4.3. Results in Irony detection in Spanish variants

In this section, we present an exhaustive evaluation of distinct settings of the MvAttLSTM model in the task of irony detection. Our first experiment aimed at investigating the impact of different types of linguistic views on the model. For that, we analyzed what subsets of features are most relevant to irony detection. The second aspect that we considered relevant to explore was the impact of the fusion strategies to inform

[6] http://adesse.uvigo.es/data/clases.php

the model (*Early vs. Contextual*) and the attention mechanism used by the MvAttLSTM model (*self vs. multi-head*). Lastly, we investigated the impact of each proposed view. For that, we ignored one view and fed the other two into the MvAttLSTM model.

To evaluate the effectiveness of our proposal in each experiment we computed the F1 score for the two classes ($F1_{iro}$ and $F1_{no-iro}$), along with their macro-averaged and micro-averaged versions of F1 ($F1_{Micro}$ and $F1_{Macro}$). We split the training data into 80% and 20% for training and validation purposes and evaluated the generalization of our model on the official test provided by the organizers. The results obtained for *IroSvA'19* corpus on the test dataset in the three Spanish variants are shown in Table 4. We only included the results using the *Contextual fusion* strategy for the Castilian (es), the Mexican (mx) and the Cuban (cu) variants due to another fusion strategy achieved worse results in the three variants.

Table 4: MvAttLSTM for irony detection in *IroSvA'19*

Views	$F1_{iro}$	$F1_{no-iro}$	$F1_{Micro}$	$F1_{Macro}$	$F1_{iro}$	$F1_{no-iro}$	$F1_{Micro}$	$F1_{Macro}$
	$MvAttLSTM^{Contextual}_{self}$ *IroSvA'19-es*				$MvAttLSTM^{Contextual}_{multihead}$ *IroSvA'19-es*			
Bert+Muse+Aff_All	0.487	0.821	0.734	0.654	0.596	0.835	0.766	0.716
Bert+Muse+Aff_Emo	0.619	0.815	0.751	0.717	**0.668**	**0.842**	**0.786**	**0.755**
Bert+Muse+Aff_App	0.538	0.82	0.741	0.679	**0.651**	**0.835**	**0.776**	**0.743**
Bert+Muse+Contrast	0.59	0.833	0.762	0.712	0.587	0.83	0.759	0.708
Bert+Muse+LIWC	0.549	0.801	0.724	0.675	0.626	0.831	0.767	0.728
Bert+Muse+SVerb	0.576	0.798	0.726	0.687	0.627	0.824	0.761	0.726
Bert+Muse+W_64	**0.629**	**0.832**	**0.769**	**0.731**	0.583	0.834	0.762	0.708
Bert+Muse+W_128	**0.656**	**0.835**	**0.777**	**0.746**	0.642	0.82	0.761	0.731
Bert+Muse+W_All	0.595	0.811	0.742	0.703	0.55	0.808	0.731	0.679
Bert+Muse+LingAll	0.553	0.802	0.726	0.678	0.589	0.831	0.761	0.710
Bert-baseline	0.302	0.741	0.594	0.521	0.302	0.741	0.594	0.521
	$MvAttLSTM^{Contextual}_{self}$ *IroSvA'19-mx*				$MvAttLSTM^{Contextual}_{multihead}$ *IroSvA'19-mx*			
Bert+Muse+Aff_All	0.516	0.79	0.708	0.654	**0.647**	**0.817**	**0.759**	**0.732**
Bert+Muse+Aff_Emo	**0.642**	**0.821**	**0.761**	**0.732**	0.508	0.785	0.701	0.647
Bert+Muse+Aff_App	**0.644**	**0.794**	**0.739**	**0.719**	0.585	0.774	0.708	0.68
Bert+Muse+Contrast	0.601	0.812	0.744	0.706	0.506	0.792	0.708	0.649
Bert+Muse+LIWC	0.506	0.792	0.708	0.649	**0.611**	**0.826**	**0.759**	**0.718**
Bert+Muse+SVerb	0.564	0.765	0.694	0.664	0.596	0.828	0.759	0.712
Bert+Muse+W_64	0.529	0.788	0.708	0.659	0.515	0.786	0.703	0.65
Bert+Muse+W_128	0.567	0.777	0.706	0.672	0.591	0.591	0.689	0.670
Bert+Muse+W_All	0.512	0.79	0.706	0.651	0.599	0.790	0.724	0.695
Bert+Muse+LingAll	0.606	0.809	0.743	0.707	0.469	0.808	0.718	0.638
Bert-baseline	0.293	0.771	0.611	0.532	0.293	0.770	0.611	0.532
	$MvAttLSTM^{Contextual}_{self}$ *IroSvA'19-cu*				$MvAttLSTM^{Contextual}_{multihead}$ *IroSvA'19-cu*			
Bert+Muse+Aff_All	**0.604**	**0.807**	**0.741**	**0.706**	0.534	0.796	0.716	0.665
Bert+Muse+Aff_Emo	0.574	0.809	0.736	0.691	0.56	0.796	0.721	0.678
Bert+Muse+Aff_App	0.53	0.803	0.723	0.666	0.563	0.793	0.719	0.678
Bert+Muse+Contrast	0.556	0.827	0.751	0.692	0.557	0.793	0.718	0.675
Bert+Muse+LIWC	0.536	0.8	0.721	0.668	0.577	0.788	0.718	0.683
Bert+Muse+SVerb	0.546	0.819	0.741	0.683	0.568	0.79	0.718	0.679
Bert+Muse+W_64	0.554	0.792	0.716	0.673	**0.596**	**0.816**	**0.748**	**0.706**
Bert+Muse+W_128	0.556	0.791	0.716	0.674	0.529	0.8	0.719	0.665
Bert+Muse+W_All	**0.582**	**0.819**	**0.748**	**0.701**	0.473	0.825	0.738	0.649
Bert+Muse+LingAll	0.517	0.804	0.721	0.661	**0.602**	**0.817**	**0.749**	**0.709**
Bert-baseline	0.472	0.803	0.693	0.638	0.472	0.803	0.693	0.638

It can be observed in Table 4 that the model $MvAttLSTM^{Contextual}_{multi}$ obtained slightly better results than $MvAttLSTM^{Contextual}_{self}$ in the three variants (*es, mx* and *cu*) for all the evaluation metrics. Concretely, for the Castilian variant, the best results were obtained by $MvAttLSTM^{Contextual}_{multi}$ using all views, but in the case of *Linguistic-view* only considering emotional *Aff_Emo* or attitudinal features *Aff_App*. Moreover, the model $MvAttLSTM^{Contextual}_{self}$ achieves competitive results but considering the views *W_128* or *W_64*. Regarding the Mexican variant, both models $MvAttLSTM^{Contextual}_{multi}$ and $MvAttLSTM^{Contextual}_{self}$ obtained the best results when the *Aff_All* and *Aff_Emo* views are used respectively. Also, it is important to notice that the second better results for each model are achieved when the views *LIWC* and *Aff_App* are considered. In the case of the Cuban variant, both models obtain similar results. However, $MvAttLSTM^{Contextual}_{multi}$ using all linguistic features *LingAll* slightly outperform $MvAttLSTM^{Contextual}_{multi}$ with the linguistic view *Aff_All*.

To sum up, we found that some subsets of features are the most relevant in our model for irony detection in the three variants; particularly, those related to affective information such as *Aff_Emo* and *Aff_App*. This fact indicates the discriminatory property of the emotional dimensions in SenticNet and the emotional categories in Spanish Emotions Lexicon (SEL) (Sidorov et al., 2013). Also, the attitude-based features obtained from Appraisal Lexicon (LAM) Hernández et al. (2011) were relevant. This result is in line with the findings presented in (Hernández Farías et al., 2016) which investigated the role of affective information in irony detection using machine learning models. Furthermore, we found that the *Bert-baseline* method performs significantly worse than the MvAttLSTM model in the three variants. One possible explanation for that is the small number of ironic examples in the training dataset that make more complex the learning process.

In a second direction, we investigated the importance of each proposed view (*Bert-view, Muse-view and Linguistic-view)* on the performance of MvAttLSTM. For that, we evaluated the model ignoring one view and including the remaining two. In this experiment, the Linguistic-view (*Ling*) represents the subsets of features that achieved the best $F1_{Macro}$ (see Table 4). Notice that *Ling* is different for each MvAttLSM setting and dataset. The results obtained are summarized in Table 5.

Table 5: The impact of the views on MvAttLSTM for irony detection. The ignored view is denoted by (×) symbol whereas the included views are denoted by (√) symbol.

Model	**Ling**	**Muse**	**Bert**	$F1_{iro}$	$F1_{no-iro}$	$F1_{Micro}$	$F1_{Macro}$
				IroSvA'19-es			
Contextual-self	×	√	√	0.627	0.835	0.771	0.730
	√	×	√	0.593	0.804	**0.736**	**0.699**
	√	√	×	0.611	0.819	0.753	0.715
Contextual-multi	×	√	√	0.611	0.827	0.761	0.719
	√	×	√	0.607	0.788	0.724	0.697
	√	√	×	0.592	0.780	**0.714**	**0.686**
				IroSvA'19-mx			
Contextual-self	×	√	√	0.546	0.824	0.746	0.685
	√	×	√	0.511	0.788	**0.704**	**0.650**
	√	√	×	0.620	0.804	0.741	0.712
Contextual-multi	×	√	√	0.530	0.776	0.696	0.653
	√	×	√	0.503	0.793	**0.708**	**0.648**
	√	√	×	0.565	0.800	0.726	0.682
				IroSvA'19-cu			
Contextual-self	×	√	√	0.557	0.793	**0.718**	**0.675**
	√	×	√	0.559	0.803	0.728	0.681
	√	√	×	0.576	0.824	0.751	0.700
Contextual-multi	×	√	√	0.528	0.805	**0.724**	**0.667**
	√	×	√	0.570	0.786	0.714	0.678
	√	√	×	0.520	0.827	0.746	0.674

As can be shown in Table 5, the best $F1_{Macro}$ in all corpora was obtained when all views were used together. This fact confirms that informing our model with the proposed views helps the model to detect irony. However, we observed that for the Castilian variant, ignoring *Muse-view* caused the most significant drop in performance of $MvAttLSTM^{Contextual}_{self}$ whereas omitting the *Bert-view* produced the worse performance in $MvAttLSTM^{Contextual}_{multi}$. In the case of the Mexican variant, both settings of $MvAttLSTM$ achieved the worse performance when the *Muse-view* was removed from the model. However, for the Cuban variant, we found that the model drops its performance when the *Linguistic-view* was omitted. Analysing the results in Table 4 and 5 together, the linguistic view (LingAll) was found to produce a lower F1-macro than when no linguistic features are introduced in the model. In this sense, we considered that a deeper analysis would be necessary to explain the reasons for the negative result achieved when including all the linguistic features whether it is due to noisy features or the fusion strategy used to feed this view into the model. Further efforts need to be made for investigating why the attention mechanisms (*self vs. multi*) attend different linguistic views for obtaining better effectiveness.

Following, we present a comparison of our best results on the three corpora with other state-of-the-art systems. In Table 6 we show how the results of the participating systems in the *IroSVA'19* shared task ranked according to the official evaluation measure $F1_{Macro}$ average. It is important to highlight that the participants were not restricted to submit the same system for each corpus. Thus, the F1-AVG means the average of the results of the team instead of evaluating the performance of one model on the three corpora. Our best model for the Castilian variant $esMvAttLSTM^{Contextual}_{multi}$ outperforms the results achieved by ELiRF_UPV (González et al., 2020; González et al., 2019) and CIMAT (Miranda-Belmonte & López-Monroy, 2019) on the Castilian and Cuban corpora. However, our model drops its performance on the Mexican corpus. Regarding our best model for the Mexican variant $mxMvAttLSTM^{Contextual}_{multi}$, it outperforms the results obtained by ELiRF_UPV and CIMAT on all corpora. The $cuMvAttLSTM^{Contextual}_{multi}$model achieved better results than ELiRF_UPV and CIMAT on the Cuban corpus but drooped its effectiveness on the Castilian and Mexican variants. It is important to remark that ELiRF_UPV is based on a deep learning model; particularly it proposed a simplification of the BERT model, and CIMAT proposed a combination of deep learning-based representations with n-gram features. From an overall point of view, our proposed models are placed in the first positions in the ranking. This fact shows the effectiveness of our model in addressing the problem of irony detection in multiple variants of Spanish.

Table 6: Comparison with state-of-the-art methods for irony detection in Spanish variants (IroSvA'19).

		IroSvA'19-es	*IroSvA'19-mx*	*IroSvA'19-cu*	*IroSvA'19*
Ranking	Team	$F1_{Macro}$	$F1_{Macro}$	$F1_{Macro}$	$F1_{AVG}$
(*)	$mxMvAttLSTM^{Contextual}_{multi}$	**0.716**	**0.732**	**0.665**	**0.704**
(**)	$esMvAttLSTM^{Contextual}_{multi}$	**0.755**	0.674	**0.678**	**0.702**
(***)	$cuMvAttLSTM^{Contextual}_{multi}$	0.710	0.638	**0.709**	**0.685**
1^{st}	*ELiRF-UPV*	0.717	0.680	0.653	0.683
2^{nd}	*CIMAT*	0.645	0.671	0.660	0.659
3^{th}	*JZaragoza*	0.661	0.67	0.616	0.649
4^{th}	*ATC*	0.651	0.645	0.594	0.630
...	...	...	...	...	...
14^{th}	*UO*	0.511	0.489	0.499	0.499

4.4. Results in Satire detection in Spanish variants

Irony and satire are both indirect forms of communication that are strongly related to each other. These forms aim at communicating in implicit ways complex meanings which often aim at criticizing, offending or hurting a victim. The major differences between them are based on the intention of the author and the linguistic resources used to effectively communicate the real meaning. In this section, we present an evaluation of our model on two corpora of satirical tweets (*Salas'17 and Barbieri'15*) for analyzing the feasibility of our model for satire detection in two Spanish variants (Castilian, and Mexican). Conversely, to the *IroSvA'19* corpus, these corpora are not explicitly divided into train and test, then we use 5-fold cross-validation to compute the generalization capability of our model in each corpus. In each iteration 80% of the data was used for training meanwhile the remainder 20% was considered for testing purpose. Also, to calibrate the hyperparameters of the model, the training set was split into two subsets 90% to train the model and 10% for validation purpose. For each corpus, the hyperparameters were tuned independently.

The results of our model on *Salas'17* and *Barbieri'15* are summarized in Table 7. In this table only the results using the *Early fusion* strategy are reported, due to the other fusion method obtained relatively worse results. At a first glance, in Table 7 can be observed that both settings of the model $MvAttLSTM^{Early}_{self}$ and $MvAttLSTM^{Early}_{multi}$ achieved similar results, even when the model which uses multi-head attention showed a slight improvement at the expense of more trainable parameters.

Concretely, for the Castilian variant in both corpora *Barbieri'15* and *Salas'17-es* the model with self-attention $MvAttLSTM^{Early}_{self}$ obtained good results when the *Linguistic-view* is used to inform the model. Particularly, appraisal features (*Aff_App*) was the most relevant for satire detection in *Salas'17-es* and the second better in *Barbieri'15*. Moreover, the features obtained by using the Wilcoxon test showed a good

Table 7: MvAttLSTM for satire detection in Spanish variants *Salas'17 and Barbieri'15*

Views	$F1_{sat}$	$F1_{no-sat}$	$F1_{Micro}$	$F1_{Macro}$	$F1_{sat}$	$F1_{no-sat}$	$F1_{Micro}$	$F1_{Macro}$
	$MvAttLSTM^{Early}_{self}$ *Salas'17-es*				$MvAttLSTM^{Early}_{multihead}$ *Salas'17-es*			
Bert+Muse+Aff_All	0.958	0.958	0.958	0.958	0.956	0.956	0.956	0.956
Bert+Muse+Aff_Emo	0.957	0.958	0.958	0.958	0.959	0.96	0.959	0.959
Bert+Muse+Aff_App	**0.96**	**0.961**	**0.961**	**0.961**	0.958	0.959	0.958	0.958
Bert+Muse+Contrast	0.956	0.958	0.957	0.957	0.959	0.96	0.959	0.959
Bert+Muse+LIWC	0.941	0.944	0.943	0.943	**0.959**	**0.96**	**0.96**	**0.96**
Bert+Muse+SVerb	0.959	0.959	0.959	0.959	0.959	0.959	0.959	0.959
Bert+Muse+W_64	**0.959**	**0.96**	**0.96**	**0.96**	**0.964**	**0.965**	**0.964**	**0.964**
Bert+Muse+W_128	0.955	0.956	0.955	0.955	0.953	0.954	0.954	0.954
Bert+Muse+W_All	**0.959**	**0.96**	**0.96**	**0.96**	**0.96**	**0.96**	**0.96**	**0.96**
Bert+Muse+LingAll	0.957	0.958	0.958	0.958	0.953	0.954	0.954	0.954
Bert-baseline	0.938	0.925	0.924	0.924	0.938	0.925	0.924	0.924
	$MvAttLSTM^{Early}_{self}$ *Salas'17-mx*				$MvAttLSTM^{Early}_{multihead}$ *Salas'17-mx*			
Bert+Muse+Aff-All	0.964	0.964	0.964	0.964	0.956	0.955	0.956	0.956
Bert+Muse+Aff-Emo	0.948	0.947	0.948	0.948	0.956	0.955	0.956	0.955
Bert+Muse+Aff-App	0.947	0.946	0.947	0.947	0.952	0.951	0.952	0.952
Bert+Muse+Contrast	**0.97**	**0.969**	**0.969**	**0.969**	0.956	0.954	0.955	0.955
Bert+Muse+LIWC	0.965	0.964	0.965	0.965	**0.969**	**0.969**	**0.969**	**0.969**
Bert+Muse+SVerb	**0.967**	**0.967**	**0.967**	**0.967**	**0.967**	**0.966**	**0.966**	**0.966**
Bert+Muse+W-64	0.961	0.96	0.96	0.96	0.96	0.959	0.959	0.959
Bert+Muse+W-128	**0.967**	**0.966**	**0.967**	**0.967**	0.957	0.957	0.957	0.957
Bert+Muse+W_All	0.966	0.966	0.966	0.966	0.964	0.963	0.963	0.963
Bert+Muse+Ling-All	0.961	0.960	0.961	0.961	0.966	0.965	0.966	0.965
Bert-baseline	0.941	0.950	0.951	0.951	0.941	0.950	0.951	0.951
	$MvAttLSTM^{Early}_{self}$ *Barbieri'15-es*				$MvAttLSTM^{Early}_{multihead}$ *Barbieri'15-es*			
Bert+Muse+Aff_All	0.953	0.952	0.953	0.953	**0.955**	**0.954**	**0.955**	**0.955**
Bert+Muse+Aff_Emo	0.954	0.953	0.954	0.954	0.953	0.952	0.952	0.952
Bert+Muse+Aff_App	**0.956**	**0.955**	**0.955**	**0.955**	0.952	0.95	0.951	0.951
Bert+Muse+Contrast	0.951	0.95	0.951	0.951	0.953	0.952	0.952	0.952
Bert+Muse+LIWC	0.954	0.953	0.954	0.954	0.954	0.953	0.954	0.954
Bert+Muse+SVerb	0.948	0.948	0.948	0.948	0.954	0.954	0.954	0.954
Bert+Muse+W_64	0.954	0.954	0.954	0.954	**0.958**	**0.957**	**0.957**	**0.957**
Bert+Muse+W_128	**0.956**	**0.956**	**0.956**	**0.956**	0.954	0.952	0.953	0.953
Bert+Muse+W_All	0.95	0.949	0.949	0.949	0.948	0.946	0.947	0.947
Bert+Muse+LingAll	0.953	0.952	0.952	0.952	0.954	0.954	0.954	0.954
Bert-baseline	0.942	0.947	0.947	0.947	0.942	0.947	0.947	0.947

performance, resulting the second more relevant *W_64* and *W_All* in *Salas'17-es*, and *W_128* the most relevant in *Barbieri'15-es*. In the case of $MvAttLSTM^{Early}_{multi}$ the best results for both Castilian corpora were achieved using the 64 best-ranked features *W_64* according to the Wilcoxon test. Regarding the results on the Mexican variant, they were different to those achieved on the Castilian tweets. Particularly, the model $MvAttLSTM^{Early}_{multi}$ perform better when the features related to polarity opposition *Contrast* were used whereas $MvAttLSTM^{Early}_{multi}$ obtained the best performance when psycho-linguistic *LIWC* features were considered. Moreover, it can be observed that *Bert-baseline* obtained very competitive results on the three corpora. This fact, confirmed that the representations leaned by the BERT model are good enough to discriminate between satirical and non-satirical tweets.

In a second direction, we aim at exploring the role of the three views proposed to inform MvAttLSTM. For that, we evaluated the model ignoring one view and including the remaining two. In this experiment, the linguistic view (*Ling*) represents the subsets of features that achieved the best F1-macro (see Table 7). As can be shown in Table 5 the best $F1_{Macro}$ in all corpora was obtained when all views were used together. In general, we observed that for the three variants, ignoring *Bert-view* caused the most significant drop in performance of both settings of *MvAttLSTM*.

In order to have a comparison with the performance obtained by other methods proposed in the literature, we compare our models with the results presented in (Salas-Zárate et al., 2017; Barbieri et al., 2015b). According to our knowledge, these works are the only two that addressing the problem of satire detection in Spanish. In Table 9 we compare our model with three methods (*SMO+LIWC-ALL*, *BayesNet+LIWC-ALL*, *J48+LIWC-ALL*) proposed in (Salas-Zárate et al., 2017) and three methods (SVM+W-B, SVM+Intrinsic,

Table 8: The impact of the views on MvAttLSTM for satire detection. The ignored view is denoted by (×) symbol whereas the included views are denoted by (√) symbol.

Model	Ling	Muse	Bert	$F1_{sat}$	$F1_{no-sat}$	$F1_{Micro}$	$F1_{Macro}$
				Salas'17-es			
Early-self	×	√	√	0.958	0.959	0.959	0.959
	√	×	√	0.955	0.956	0.955	0.955
	√	√	×	0.710	0.720	**0.721**	**0.715**
Early-multi	×	√	√	0.952	0.954	0.953	0.953
	√	×	√	0.952	0.954	0.953	0.953
	√	√	×	0.603	0.685	**0.654**	**0.644**
				Salas'17-mx			
Early-self	×	√	√	0.958	0.959	0.959	0.959
	√	×	√	0.958	0.957	0.958	0.958
	√	√	×	0.719	0.858	**0.822**	**0.788**
Early-multi	×	√	√	0.956	0.956	0.956	0.956
	√	×	√	0.936	0.931	0.934	0.934
	√	√	×	0.649	0.683	**0.681**	**0.667**
				Barbieri'15-es			
Early-self	×	√	√	0.95	0.949	0.95	0.95
	√	×	√	0.952	0.950	0.951	0.951
	√	√	×	0.743	0.728	**0.736**	**0.735**
Early-multi	×	√	√	0.939	0.935	0.937	0.937
	√	×	√	0.951	0.950	0.951	0.951
	√	√	×	0.743	0.702	**0.725**	**0.722**

SVM+ALL) introduced in (Barbieri et al., 2015b) for satire detection. The methods proposed in (Salas-Zárate et al., 2017) are based on machine learning combined with hand-crafted features, particularly features derived from LIWC. The major difference among these methods is the machine learning algorithm used. The methods were evaluated using precision (P_{sat}) recall (R_{sat}) and F1 score ($F1_{sat}$) on the positive class (satirical tweets).

Table 9: Comparison with state-of-the-art methods for satire detection in Spanish variants.

	Salas'17-mx			*Salas'17-es*			*Barbieri'15-es*
Method	P_{sat}	R_{sat}	$F1_{sat}$	P_{sat}	R_{sat}	$F1_{sat}$	$F1_{Macro}$
SMO+LIWC-ALL	0.855	0.855	0.855	0.846	0.84	0.84	-
BayesNet+LIWC-ALL	0.757	0.756	0.756	0.734	0.734	0.734	-
J48+LIWC-ALL	0.752	0.752	0.752	0.774	0.774	0.774	-
SVM+W-B	-	-	-	-	-	-	0.738
SVM+Intrinsic	-	-	-	-	-	-	0.816
SVM+ALL	-	-	-	-	-	-	0.852
$salasEsMvAttLSTM^{Early}_{multi}$	0.966	0.973	**0.969**	0.969	0.950	**0.959**	**0.954**
$salasMxMvAttLSTM^{Early}_{multi}$	0.951	0.969	**0.960**	0.973	0.955	**0.964**	**0.957**
$barbEsMvAttLSTM^{Early}_{multi}$	0.951	0.969	**0.960**	0.973	0.955	**0.964**	**0.957**

In the same fashion, the methods proposed in (Barbieri et al., 2015b) differ from each other in the features employed to describe the satirical texts: *SVM+W-B* considers features based on word n-grams whereas *SVM+Intrinsic* employs linguistic features which are topic-independent, and finally *SVM+ALL* combines both subgroups of features. In this case, the methods were evaluated using $F1_{Macro}$. To establish a fair comparison with the previous works, we evaluated the performance of our models that achieved the best results on each corpus independently ($salasEsMvAttLSTM^{Early}_{multi}$, $salasMxMvAttLSTM^{Early}_{multi}$ and $barbEsMxMvAttLSTM^{Early}_{multi}$) and we reevaluated the models on the two remaining corpora. As can be observed in Table, 9 the three settings of our model $MvAttLSTM^{Early}_{multi}$ outperformed by almost 10% points the results achieved by the best methods reported in (Salas-Zárate et al., 2017; Barbieri et al., 2015b).

4.5. Discriminating between Irony and Satire

Some figurative languages devices like irony and satire are difficult to distinguish from each other due to they share several characteristics, even can be nested. For instance, satire can appeal to irony for communicating indirect and complex meanings often aiming at censoring or criticizing peoples, things, social and moral norms in an ironical way. Furthermore, to evaluate our model beyond *irony vs. non-irony* and *satire vs. non-satire* scenarios we evaluated the capability of our model for discriminating between both phenomena.

Table 10: MvAttLSTM for irony vs. satire detection in Castilian and Mexican tweets

Views	$F1_{iro}$	$F1_{sat}$	$F1_{Micro}$	$F1_{Macro}$	$F1_{iro}$	$F1_{sat}$	$F1_{Micro}$	$F1_{Macro}$
	$sat-MvAttLSTM^{Early}_{multihead}$ *es*				$sat-MvAttLSTM^{Early}_{multihead}$ *mx*			
Bert+Muse+Aff_All	0.950	0.980	0.972	0.965	0.953	0.981	0.973	0.967
Bert+Muse+Aff_Emo	**0.979**	**0.992**	**0.988**	**0.985**	**0.968**	**0.987**	**0.982**	**0.978**
Bert+Muse+Aff_App	**0.981**	**0.992**	**0.989**	**0.987**	0.955	0.982	0.975	0.969
Bert+Muse+Contrast	0.957	0.983	0.976	0.970	**0.964**	**0.986**	**0.979**	**0.975**
Bert+Muse+LIWC	0.976	0.990	0.986	0.983	0.952	0.981	0.972	0.966
Bert+Muse+SVerb	0.886	0.957	0.938	0.921	0.948	0.980	0.971	0.964
Bert+Muse+W_64	0.953	0.982	0.974	0.968	0.948	0.979	0.970	0.964
Bert+Muse+W_128	0.953	0.982	0.974	0.967	0.950	0.980	0.971	0.965
Bert+Muse+W_All	0.956	0.983	0.975	0.970	0.888	0.963	0.945	0.926
Bert+Muse+LingAll	0.920	0.966	0.952	0.943	0.952	0.981	0.973	0.966
Bert-baseline	0.983	0.991	0.987	0.984	0.963	0.975	0.964	0.956
	$iro-MvAttLSTM^{Contextual}_{multihead}$ *es*				$iro-MvAttLSTM^{Contextual}_{multihead}$ *mx*			
Bert+Muse+Aff_All	**0.966**	**0.987**	**0.981**	**0.976**	0.962	0.985	0.979	0.974
Bert+Muse+Aff_Emo	0.944	0.975	0.966	0.959	**0.968**	**0.987**	**0.982**	**0.978**
Bert+Muse+Aff_App	**0.966**	**0.986**	**0.981**	**0.976**	0.963	0.985	0.979	0.974
Bert+Muse+Contrast	0.963	0.986	0.980	0.975	0.963	0.985	0.979	0.974
Bert+Muse+LIWC	0.965	0.986	0.980	0.975	0.955	0.982	0.974	0.968
Bert+Muse+SVerb	0.963	0.986	0.979	0.974	0.964	0.985	0.979	0.975
Bert+Muse+W_64	0.947	0.980	0.971	0.963	0.963	0.985	0.979	0.974
Bert+Muse+W_128	0.964	0.986	0.980	0.975	0.964	0.986	0.979	0.975
Bert+Muse+W_All	0.929	0.965	0.953	0.947	0.964	0.986	0.980	0.975
Bert+Muse+LingAll	**0.966**	**0.987**	**0.981**	**0.976**	0.962	0.985	0.979	0.974
Bert-baseline	0.983	0.991	0.987	0.984	0.963	0.975	0.964	0.956

In this sense, the first step was building the satire-irony corpus. For that purpose, we merged the 1000 ironic tweets of the *IroSvA'19* corpus with the 2500 satirical tweets of *Salas'17* for the Mexican and Castilian variants of the Spanish independently. After that, we reevaluated the best model for irony detection ($iroMvAttLSTM^{Contextual}_{multi}$) and the best model for satire detection ($satMvAttLSTM^{Early}_{multi}$) in each variant (see Section 4.4). In Table 10, we present the results obtained. As can be observed, the results show that our model is able to effectively discriminate satire from irony in both variants *(es and mx)* with an effectiveness $F1_{Macro} = 0.987$ for Castilian tweets and $F1_{Macro} = 0.978$ for Mexican tweets. Also, *Bert-baseline* showed very high results on both corpora.

Concretely, the model $satMvAttLSTM^{Early}_{multi}$ achieves, in general, the best performance in both variants. All these results make evident that those views such as *Linguistic-view* and *Muse-views* have a low impact on the model. A possible reason is that these views were learned without any supervision related to the specific task. Conversely, *Bert-view*, which is a task-dependent view, has a major impact on the model effectiveness, particularly due to this view was learned in a supervised way and it is strongly related to the specific task dataset. Regarding the linguistics views, the best results of $satMvAttLSTM^{Early}_{multi}$ in the Mexican and Castilian variants were *Aff_Emo* and *Aff_App* respectively. According to these results, we could appreciate that affective information was, in general, the most relevant to inform our model for capturing useful information to detect irony and satire in Spanish variants.

Table 11 shows the impact of the views to inform our model. The obtained results are aligned with the results achieved in the task of satire detection. As can be observed, ignoring *Bert-view* caused the most significant drop in the performance of the $MvAttLSTM$ model whereas the model is less sensitive to exclude *Ling-view* and *Muse-view*.

Table 11: The impact of the views on MvAttLSTM for irony and satire distinguishing. The ignored view is denoted by (×) symbol whereas the included views are denoted by (√) symbol.

Model	Ling	Muse	Bert	$F1_{iro}$	$F1_{sat}$	$F1_{Micro}$	$F1_{Macro}$
				Castilian variant			
Early-multi (sat)	×	√	√	0.957	0.983	0.976	0.970
	√	×	√	0.966	0.987	0.981	0.977
	√	√	×	0.942	0.977	**0.967**	**0.959**
Contextual-multi(iro)	×	√	√	0.961	0.985	0.978	0.973
	√	×	√	0.967	0.987	0.981	0.976
	√	√	×	0.873	0.930	**0.911**	**0.901**
				Mexican variant			
Early-multi(sat)	×	√	√	0.951	0.981	0.973	0.966
	√	×	√	0.948	0.979	0.970	0.964
	√	√	×	0.407	0.896	**0.824**	**0.651**
Contextual-multi(iro)	×	√	√	0.964	0.986	0.979	0.975
	√	×	√	0.966	0.986	0.981	0.976
	√	√	×	0.532	0.730	**0.740**	**0.631**

Analyzing the results presented in Table 7 and Table 10, we could appreciate that our model is better discriminating irony from satire than irony from no-irony and satire from no-satire. This behavior is caused by the nature of the dataset. Satirical tweets were retrieved from different topics than ironic tweets. This fact introduces a bias with respect to the topics discussed in the ironic and satirical tweets.

4.6. Validating the robustness of the models in Humor recognition

Our final experiment aims at investigating the robustness of our model for recognizing humorous tweets written in Spanish. We are intrigued by the fact that irony and satire are two phenomena that are strongly related to humor. Particularly, some theoretical works comments about the relation between humor-irony (Gurillo et al., 2013; Garmendia, 2018) and humor-satire (Simpson, 2003).

Table 12: MvAttLSTM for humour recognition in Spanish tweets (*HAHA'19*)

Views	$F1_{hum}$	$F1_{no-hum}$	$F1_{Micro}$	$F1_{Macro}$	$F1_{hum}$	$F1_{no-hum}$	$F1_{Micro}$	$F1_{Macro}$
	$MvAttLSTM^{Early}_{self}$				$MvAttLSTM^{Early}_{multihead}$			
Bert+Muse+Aff_All	0.802	0.876	0.848	0.839	0.794	0.879	0.848	0.837
Bert+Muse+Aff_Emo	**0.804**	**0.881**	**0.852**	**0.842**	0.799	0.877	0.847	0.838
Bert+Muse+Aff_App	**0.804**	**0.878**	**0.85**	**0.841**	0.798	0.879	0.849	0.838
Bert+Muse+LIWC	0.796	0.881	0.85	0.839	0.798	0.88	0.849	0.839
Bert+Muse+SVerb	0.798	0.88	0.849	0.839	0.797	0.877	0.847	0.837
Bert+Muse+W_64	0.803	0.881	0.852	0.842	0.801	0.879	0.85	0.84
Bert+Muse+W_128	0.798	0.881	0.85	0.839	**0.806**	**0.879**	**0.851**	**0.842**
Bert+Muse+W_All	0.795	0.88	0.849	0.838	**0.804**	**0.882**	**0.853**	**0.843**
Bert+Muse+LingAll	0.802	0.872	0.845	0.837	0.796	0.88	0.849	0.838
Bert-baseline	0.802	0.864	0.84	0.833	0.802	0.864	0.84	0.833

In this sense, we evaluated our model with the corpus *HAHA'19* and the results are shown in Table 12. In this case, only the results achieved by our model using the *Early fusion* strategy are presented due to the *Contextual fusion* method obtained worse results. At a first glance, we can appreciate that our model achieves very similar results for both attention mechanisms, although the model $MvAttLSTM^{Early}_{multi}$ shows a slight improvement in terms of $F1_{humor}$. Another important aspect to notice is regarding the linguistic views, particularly the model $MvAttLSTM^{Early}_{self}$ that performs better when affective features such as *Aff_Emo* and *Aff_App* are used. However, the model $MvAttLSTM^{Early}_{multi}$ obtains its best results when more features are considered, particularly those best-ranked according to the Wilcoxon test *W_128* and *W_All*. This behavior is aligned with the results presented in (Ortega-Bueno et al., 2019). Also, in this corpus *Bert-baseline* achieved competitive results in comparison with our proposed models. With respect

to the impact of the views in our models, we observed in Table 13 that *Bert-view* and *Muse-view* are more important than *Linguistic-view*.

Table 13: The impact of the views on MvAttLSTM for humor recognition. The ignored view is denoted by (×) symbol whereas the included views are denoted by (√) symbol.

Model	Ling	Muse	Bert	$F1_{hum}$	$F1_{no-hum}$	$F1_{Micro}$	$F1_{Macro}$
				HAHA'19			
Early-self	×	√	√	0.792	0.88	0.848	0.836
	√	×	√	0.79	0.874	0.843	0.832
	√	√	×	0.783	0.876	**0.842**	**0.829**
Early-multi	×	√	√	0.795	0.88	0.848	0.838
	√	×	√	0.784	0.878	0.844	0.831
	√	√	×	0.781	0.881	**0.845**	**0.831**

We compare the results of $MvAttLSTM^{Early}_{multi}$ with those of the participating systems in the shared task at *HAHA'19* organized in the framework of *IberLEF'19*. In this task, the systems were ranked according to the official measure F1 score in the humor class, although also and *Acc* was reported. As can be observed in Table 14, the results obtained by our model are very competitive, obtaining the fourth position of the ranking according to $F1_{humor}$ and the third position in terms of *Acc* out of 18 systems. The performance of our model is similar to the best-ranked system *Adilism* in terms of F1. However, the difference in terms of precision and recall shows that the *Adilism* system is better at detecting a major number of humorous tweets whereas our model is better at detecting the humorous tweets. As future work, a deeper study is required to analyze the low recall achieved by our model compared to the *Adilism* system.

Table 14: Comparison with state of the art systems for humor recognition in Spanish *(HAHA'2019)*.

Ranking	*Team*	P_{hum}	R_{hum}	$F1_{hum}$	*Acc*
1^{st}	*Adilism*	0.791	0.852	0.821	0.855
2^{sd}	*Kevin & Hiromi*	0.802	0.831	0.816	0.854
3^{th}	*Bfarzin*	0.782	0.839	0.810	0.846
**	$MvAttLSTM^{Early}_{multi}$	**0.819**	0.792	0.806	**0.851**
4^{th}	*Jamestjw*	0.793	0.804	0.798	0.842
5^{th}	*INGEOTEC*	0.758	0.819	0.788	0.828
6^{th}	*BLAIR GMU*	0.745	0.827	0.784	0.822
7^{th}	*UO_UPV2*	0.78	0.765	0.773	0.824
...	...	...	...	...	...
18^{th}	*Amrita CEN*	0.478	0.514	0.495	0.591

5. Conclusions and Future Work

In this work, we have presented MvAttLSTM, a deep learning-based method for irony and satire detection in Spanish variants. It is based on an Attentive-LSTM model informed with additional knowledge learned from three distinct perspectives: *Linguistic-view*, *MUSE-based view*, and *BERT-based view*. We observed that our model achieved better performance when it is enriched with the three proposed views. We have evaluated our model on the corpus *IroSvA'19* for irony and on the corpora *Salas'17* and *Barbieri'15* for satire detection in Spanish variants. In both tasks, the model outperforms the state-of-the-art results. Furthermore, we have evaluated our model on humour recognition using the corpus *HAHA'19* showing a very competitive behaviour. Particularly, linguistic information and deep sentence encoding were more feasible for irony detection whereas BERT views increased the performance of satire detection and satire vs. irony detection *(RQ1)*. Interestingly, the results revealed that affective information helps in detecting irony and satire. Particularly, those related to emotions (*Aff_Emo*) which are based on the resources SenticNet and SEL; and those related to attitude words (*Aff_App*) based on the LAM lexicon. Experiments also confirmed that both fusion strategies are feasible. However *Contextual fusion* achieved better performance

in relative small corpus like *IroSvA'19*, whereas the *Early fusion* takes advantage of large and self-annotated corpora *(RQ2)*. Unexpectedly we found no strong differences in the effectiveness of our model when self-attention or multi-head attention was considered. However, we appreciated that each attention type attends distinct linguistic features *(RQ3)*. We are aware that our model has an important limitation which lies in the lack of explainability about why the model used some features from one setting to another and from one Spanish variant to the others. As future work, we will aim at investigating other methods for fusing the additional knowledge into our model. Moreover, we plan to carry out a fine-grained analysis on the impact of linguistic features joined with the information captured by the attention mechanism for irony and satire interpretability. Finally, we are interested in exploring our model in multilingual and cross-lingual settings.

6. Acknowledgements

The work of the first two authors was in the framework of the research project MISMIS-FAKEnHATE on MISinformation and MIScommunication in social media: FAKE news and HATE speech (PGC2018-096212-B-C31), funded by Spanish Ministry of Science and Innovation, and DeepPattern (PROMETEO/2019/121), funded by the Generalitat Valenciana.

References

Abdalla, M., & Hirst, G. (2017). Cross-lingual sentiment analysis without (good) translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 506–515). Taipei, Taiwan: Asian Federation of Natural Language Processing. URL: https://www.aclweb.org/anthology/I17-1051.

Abulaish, M., Kamal, A., & Zaki, M. J. (2020). A survey of figurative language and its computational detection in online social networks. *ACM Transactions on the Web*, *14*, 1–52. doi:10.1145/3375547.

Agrawal, A., & An, A. (2018). Affective representations for sarcasm detection. In *41st International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2018* (pp. 1029–1032). Ann Arbor, MI, USA: Association for Computing Machinery. ACM. doi:10.1145/3209978.3210148.

Ahmad, T., Akhtar, H., Chopra, A., & Akhtar, M. W. (2014). Satire Detection from Web Documents using machine Learning Methods. In *International Conference on Soft Computing & Machine Intelligence Satire* (pp. 102–105). IEEE. doi:10.1109/ISCMI.2014.34.

Altin, L. S. M., Bravo, À., & Saggion, H. (2019). LaTUS/TALN at IroSvA: Irony detection in Spanish variants. In *Iberian Languages Evaluation Forum (IberLEF 2019), co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019)* (pp. 291–296). Bilbao, Spain: CEUR Workshop Proceedings. CEUR-WS.org volume 2421.

Attardo, S. (2000). Irony as relevant inappropriateness. *Journal of Pragmatics*, *32*, 793–826. doi:10.1016/S0378-2166(99)00070-3.

Bahdanau, D., Cho, K. H., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. *3rd International Conference on Learning Representations, ICLR 2015*, (pp. 1–15). arXiv:1409.0473.

Balahur, A., & Turchi, M. (2012). Multilingual sentiment analysis using machine translation? In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis. WASSA'13* (pp. 52–60).

Bamman, D., & Smith, N. A. (2015). Contextualized sarcasm detection on Twitter. *Proceedings of the 9th International Conference on Web and Social Media, ICWSM 2015*, (pp. 574–577).

Barbieri, F., Anke, L. E., & Saggion, H. (2016a). Revealing patterns of twitter emoji usage in barcelona and madrid. In *International Conference of the Catalan Association for Artificial Intelligence*.

Barbieri, F., Ballesteros, M., & Saggion, H. (2017a). Are emojis predictable? In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers* (pp. 105–111). Valencia, Spain: Association for Computational Linguistics. URL: https://www.aclweb.org/anthology/E17-2017.

Barbieri, F., Basile, V., Croce, D., Nissim, M., Novielli, N., & Patti, V. (2016b). Overview of the Evalita 2016 SENTIment POLarity Classification Task. In *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016), Napoli, Italy, December 5-7, 2016.*. CEUR-WS.org volume 1749 of *CEUR Workshop Proceedings*.

Barbieri, F., Espinosa-Anke, L., Ballesteros, M., Soler-Company, J., & Saggion, H. (2017b). Towards the understanding of gaming audiences by modeling twitch emotes. In *Proceedings of the 3rd Workshop on Noisy User-generated Text* (pp. 11–20). Copenhagen, Denmark: Association for Computational Linguistics. URL: https://www.aclweb.org/anthology/W17-4402. doi:10.18653/v1/W17-4402.

Barbieri, F., Kruszewski, G., Ronzano, F., & Saggion, H. (2016c). How cosmopolitan are emojis? exploring emojis usage and meaning over different languages with distributional semantics. In *Proceedings of the 24th ACM International Conference on Multimedia* MM '16 (p. 531–535). New York, NY, USA: Association for Computing Machinery. URL: https://doi.org/10.1145/2964284.2967278. doi:10.1145/2964284.2967278.

Barbieri, F., Ronzano, F., & Saggion, H. (2015a). Do we criticise (and Laugh) in the same way? Automatic detection of multi-lingual satirical news in twitter. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)* (pp. 1215–1221).

Barbieri, F., Ronzano, F., & Saggion, H. (2015b). Is this tweet satirical? A computational approach for satire detection in Spanish. *Procesamiento de Lenguaje Natural*, *55*, 135–142.

Barbieri, F., Ronzano, F., & Saggion, H. (2016d). What does this emoji mean? a vector space skip-gram model for Twitter emojis. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 3967–3972). Portorož, Slovenia: European Language Resources Association (ELRA). URL: https://www.aclweb.org/anthology/L16-1626.

Barbieri, F., & Saggion, H. (2014a). Automatic Detection of Irony and Humour in Twitter. *Proceedings of the Fifth International Conference on Computational Creativity*, (pp. 155–162).

Barbieri, F., & Saggion, H. (2014b). Modelling irony in Twitter: Feature analysis and evaluation. *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014*, (pp. 4258–4264).

Barbieri, F., Saggion, H., & Ronzano, F. (2014). Modelling Sarcasm in Twitter, a Novel Approach. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (pp. 50–58). Baltimore, Maryland, USA: Association for Computational Linguistics.

Basile, V., Bolioli, A., Nissim, M., Patti, V., & Rosso, P. (2014). Overview of the Evalita 2014 SENTIment POLarity classification task. In *First Italian Conference on Computational Linguistics (CLiC-it 2014) & the Fourth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian EVALITA 2014* (pp. 50–57).

Baziotis, C., Nikolaos, A., Papalampidi, P., Kolovou, A., Paraskevopoulos, G., Ellinas, N., & Potamianos, A. (2018). NTUA-SLP at SemEval-2018 Task 3: Tracking Ironic Tweets using Ensembles of Word and Character Level Attentive RNNs. (pp. 613–621). doi:10.18653/v1/s18-1100. arXiv:1804.06659.

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 *Actes de l'atelier DEFT2017 Associé à la Conférence TALN*. Orléans, France.

Bodria, F., Panisson, A., Perotti, A., & Piaggesi, S. (2020). Explainability Methods for Natural Language Processing: Applications to Sentiment Analysis. In *CEUR Workshop Proceedings* (pp. 100–107). volume 2646.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the ACL.*, *5*, 135–146.

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.

Brown, P., & Levinson, S. C. (1987). *Politeness: Some Universals in Language Usage*. (2nd ed.). Cambridge University Press. doi:10.2307/3587263.

Burfoot, C., & Baldwin, T. (2009). Automatic Satire Detection: Are You Having a Laugh? In *Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing* (pp. 161–164). Suntec, Singapore: ACL and AFNLP.

Calvo, H., Gambino, O. J., & García, C. V. (2020). Irony Detection using Emotion Cues. *Computación y Sistemas*, *24*, 1281–1287. doi:10.13053/CyS-24-3-3487.

Cambria, E., Li, Y., Xing, F. Z., Poria, S., & Kwok, K. (2020). Senticnet 6: Ensemble application of symbolic and subsymbolic ai for sentiment analysis. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (p. 105–114). New York, NY, USA: Association for Computing Machinery. URL: https://doi.org/10.1145/3340531.3412003.

Carvalho, P., Sarmento, L., Silva, M. J., & de Oliveira, E. (2009). Clues for Detecting Irony in User-generated Contents: Oh...!! it's "so easy" ;-). In *Proceedings of the 1st International Conference on Information Knowledge Management Workshop on Topic-Sentiment Analysis for Mass Opinion* (pp. 53–56).

Cer, D., Yang, Y., yi Kong, S., Hua, N., Limtiaco, N., St. John, R., Constant, N., Guajardo-Céspedes, M., Yuan, S., Tar, C., Sung, Y. H., Strope, B., & Kurzweil, R. (2018). Universal sentence encoder for English. In *Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Proceedings* (pp. 169–174). doi:10.18653/v1/d18-2029. arXiv:arXiv:1803.11175v2.

Chauhan, D. S., S R, D., 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). Online: Association for Computational Linguistics. URL: https://www.aclweb.org/anthology/2020.acl-main.401. doi:10.18653/v1/2020.acl-main.401.

Chiruzzo, L., Castro, S., Etcheverry, M., Garat, D., Prada, J. J., & Rosá, A. (2019). Overview of HAHA at IberLEF 2019: Humor Analysis based on Human Annotation. In *Iberian Languages Evaluation Forum (IberLEF 2019), co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019)* (pp. 132–144). Bilbao, Spain: CEUR Workshop Proceedings. CEUR-WS.org.

Chiruzzo, L., Castro, S., & Rosá, A. (2020). HAHA 2019 Dataset: A Corpus for Humor Analysis in Spanish. In *12th Conference on Language Resources and Evaluation (LREC 2020)* May (pp. 5106–5112). Marseille, France: European Language Resources Association (ELRA).

Chung, C. K., & Pennebaker, J. W. (2011). Linguistic inquiry and word count (LIWC): Pronounced "Luke,". and other useful facts. In *Applied Natural Language Processing: Identification, Investigation and Resolution* (pp. 206–229). doi:10.4018/978-1-60960-741-8.ch012.

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Gated Recurrent Neural Networks on Sequence Modeling. In *NIPS'2014*

Deep Learning workshop (pp. 1–9). arXiv. arXiv:arXiv:1412.3555v1.

Cignarella, A. C., Frenda, S., Basile, V., Bosco, C., Patti, V., Rosso, P., Frenda, S., Basile, V., Bosco, C., Patti, V., & Rosso, P. (2018). Overview of the Evalita 2018 Task on Irony Detection in Italian Tweets (IronITA). In *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18)*. Turin, Italy: CEUR.org.

Cignarella, A. T., Basile, V., Sanguinetti, M., Bosco, C., Rosso, P., & Benamara, F. (2020). Multilingual Irony Detection with Dependency Syntax and Neural Models. In *28th International Conference on Computational Linguistics* (pp. 1346—-1358). Barcelona, Spain (Online): Association for Computational Linguistics (ACL). URL: 10.18653/v1/2020.coling-main.116. arXiv:2011.05706.

Clark, K., Khandelwal, U., Levy, O., & Manning, C. D. (2019). What does BERT look at? An analysis of BERT's attention. doi:10.18653/v1/w19-4828. arXiv:1906.04341.

Colletta, L. (2009). Political satire and postmodern irony in the age of Stephen Colbert and Jon Stewart. *Journal of Popular Culture*, *42*, 856–874. doi:10.1111/j.1540-5931.2009.00711.x.

Colston, H. L. (2015). *Using Figurative Language*. Cambridge University Press.

Condren, C. (2014). Satire. In S. Attardo (Ed.), *Encyclopedia of Humor Studies* chapter Satire. (pp. 661–664). USA: SAGE Publications, Inc.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Guzmán, F., Wenzek, G., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. In *58th Annual Meeting of the Association for Computational Linguistics* (pp. 8440–8451). Association for Computational Linguistics. URL: https://www.aclweb.org/anthology/2020.acl-main.747. arXiv:arXiv:1911.02116v2.

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., & Bordes, A. (2017). Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 670–680). Copenhagen, Denmark: Association for Computational Linguistics. URL: https://www.aclweb.org/anthology/D17-1070. doi:10.18653/v1/D17-1070.

Dancygier, B. (2014). *Figurative Language*. Cambridge University Press.

Dashtipour, K., Poria, S., Hussain, A., Cambria, E., Hawalah, A. Y., Gelbukh, A., & Zhou, Q. (2016). Multilingual sentiment analysis: state of the art and independent comparison of techniques. *Cognitive Computation*, *8*, 757–771.

Davidov, D., Tsur, O., & Rappoport, A. (2010). Semi-supervised Recognition of Sarcastic Sentences in Twitter and Amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning* CoNLL '10 (pp. 107–116). Uppsala, Sweden: Association for Computational Linguistics.

del Pilar Salas-Zárate, M., Alor-Hernández, G., Sánchez-Cervantes, J. L., Paredes-Valverde, M. A., García-Alcaraz, J. L., & Valencia-García, R. (2020). Review of English literature on figurative language applied to social networks. *Knowledge and Information Systems*, *62*, 2105–2137. URL: https://doi.org/10.1007/s10115-019-01425-3. doi:10.1007/s10115-019-01425-3.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *2019 Conference of the North {A}merican Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)"* (pp. 4171—-4186). Minneapolis, Minnesota, USA: Association for Computational Linguistics (ACL). URL: https://www.aclweb.org/anthology/N19-1423. arXiv:v1/N19-1423.

Dutta, S., & Chakraborty, A. (2019). A Deep Learning-Inspired Method for Social Media Satire Detection. In *Soft Computing and Signal Processing , Advances in Intelligent Systems and Computing* (pp. 243–251). Springer Singapore. URL: http://dx.doi.org/10.1007/978-981-13-3393-4{_}25. doi:10.1007/978-981-13-3393-4.

Esuli, A., Moreo, A., & Sebastiani, F. (2020). Cross-lingual sentiment quantification. *IEEE Intelligent Systems*, *35*, 106–114. doi:10.1109/MIS.2020.2979203.

Farias, D. I., & Rosso, P. (2017). Irony, Sarcasm, and Sentiment Analysis. In *Sentiment Analysis in Social Networks* (pp. 113–128). doi:10.1016/B978-0-12-804412-4.00007-3.

Galeshchuk, S., Qiu, J., & Jourdan, J. (2019). Sentiment analysis for multilingual corpora. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing* (pp. 120–125). Florence, Italy: Association for Computational Linguistics. URL: https://www.aclweb.org/anthology/W19-3717. doi:10.18653/v1/W19-3717.

García, L., Moctezuma, D., & Muñiz, V. (2019). A Contextualized Word Representation Approach for Irony Detection. 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)*.. Bilbao, Spain: CEUR Workshop Proceedings. CEUR-WS.org.

Garmendia, J. (2018). *Irony*. (1st ed.). New York, USA: Cambridge University Press. doi:10.1017/9781316136218.

Ghanem, B., Karoui, J., Benamara, F., Moriceau, V., & Rosso, P. (2019). IDAT@FIRE2019: Overview of the track on Irony Detection in Arabic Tweets. In *11th Forum for Information Retrieval Evaluation* (pp. 1–11). CEURS. doi:10.1145/3368567.3368585.

Ghanem, B., Karoui, J., Benamara, F., Rosso, P., & Moriceau, V. (2020). Irony Detection in a Multilingual Context. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020* (pp. 141–149). URL: http://arxiv.org/abs/2003.13924. doi:10.1007/978-3-030-45442-5. arXiv:2003.13924.

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'15)* (pp. 470–478). Denver, Colorado: Association for Computational Linguistics. doi:10.18653/v1/s15-2080.

Ghosh, A., & Veale, T. (2016). Fracking Sarcasm using Neural Network. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (pp. 161–169). San Diego, California: Association for Computational Linguistics. URL: http://www.aclweb.org/anthology/W16-0425.

Ghosh, D., Fabbri, A. R., & Muresan, S. (2018). Sarcasm Analysis Using Conversation Context. *Computational Linguistics*, *44*, 755–792. doi:10.1162/coli.

Ghosh, D., Richard Fabbri, A., & Muresan, S. (2017). The Role of Conversation Context for Sarcasm Detection in Online Interactions. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue* (pp. 186–196). Saarbrücken, Germany: Association for Computational Linguistics. doi:"10.18653/v1/W17-5523".

Ghosh, D., Vajpayee, A., & Muresan, S. (2020). A Report on the 2020 Sarcasm Detection Shared Task. In *Second Workshop on Figurative Language Processing 2020* (pp. 1–11). Association for Computational Linguistics.

Gibbs, R. W., O'Brien, J. E., & Doolittle, S. (1995). Inferring Meanings That Are Not Intended: Speakers' Intentions and Irony Comprehension. *Discourse Processes*, *20*, 187–203. doi:10.1080/01638539509544937.

Golbeck, J., Mauriello, M., Auxier, B., Bhanushali, K. H., Bonk, C., Bouzaghrane, M. A., Buntain, C., Chanduka, R., Cheakalos, P., Everett, J. B., Falak, W., Gieringer, C., Graney, J., Hoffman, K. M., Huth, L., Ma, Z., Jha, M., Khan, M., Kori, V., Lewis, E., Mirano, G., Mohn, W. T., Mussenden, S., Nelson, T. M., Mcwillie, S., Pant, A., Shetye, P., Shrestha, R., Steinheimer, A., Subramanian, A., & Visnansky, G. (2018). Fake News vs Satire : A Dataset and Analysis. In *10th ACM Conference on Web Science (WebSci 2018)* (pp. 17–21). Amsterdam, Netherlands.

González, J. A., Hurtado, L.-F., & Pla, F. (2019). ELiRF-UPV at IroSvA: Transformer Encoders for Spanish Irony Detection. 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)*.. Bilbao, Spain: CEUR Workshop Proceedings. CEUR-WS.org.

González, J. Á., Hurtado, L. F., & Pla, F. (2020). Transformer based contextualization of pre-trained word embeddings for irony detection in Twitter. *Information Processing and Management*, *57*, 1–15. URL: https://doi.org/10.1016/j.ipm.2020.102262. doi:10.1016/j.ipm.2020.102262.

González, M. D. M., Cámara, E. M., & Valdivia, M. T. M. (2015). CRiSOL:Base de conocimiento de opiniones para el español. *Procesamiento del Lenguaje Natural*, (pp. 143–150).

Gonzalez-Agirre, A., Laparra, E., & Rigau, G. (2012). Multilingual central repository version 3.0. *LREC*, (pp. 2525–2529).

González-Ibáñez, R., Muresan, S., & Wacholder, N. (2011). Identifying Sarcasm in Twitter: A Closer Look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* HLT '11 (pp. 581–586). Portland, Oregon: Association for Computational Linguistics.

Grice, H. P. (1975). Logic and Conversation. In P. Cole, & J. Morgan. (Eds.), *Syntax and Semantics 3: Speech Acts* (pp. 41–58). New York: Academic Press.

Grice, H. P. (1978). Further notes on logic and conversation. In P. Cole (Ed.), *Syntax and Semantics 9: Pragmatics* (pp. 113–127). New York: Academic Press.

Guibon, G., Ermakova, L., Seffih, H., Firsov, A., & Noé-bienvenu, G. L. (2019). Multilingual Fake News Detection with Satire To cite this version : HAL Id : halshs-02391141. In *International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2019)*. La Rochelle, France.

Gurillo, L. R., Ortega, M. B. A., Rosique, S. R., Attardo, S., Paredes, E. M.-G., Padilla-García, X. A., Muñoz-Basols, J., Adrjan, P., David, M., Viana, A., Feyaerts, K., & Yus, F. (2013). *Irony and humor: From pragmatics to discourse* volume 231. Amsterdam / Philadelphia: John Benjamins Publishing Company.

Haynes, W. (2013). Wilcoxon Rank Sum Test. In W. Dubitzky, O. Wolkenhauer, K.-H. Cho, & H. Yokota (Eds.), *Encyclopedia of Systems Biology* (pp. 2354–2355). New York, NY: Springer New York. URL: https://doi.org/10.1007/978-1-4419-9863-7{_}1185. doi:10.1007/978-1-4419-9863-7_1185.

Hazarika, D., Poria, S., Gorantla, S., Cambria, E., Zimmermann, R., & Mihalcea, R. (2018). CASCADE: Contextual Sarcasm Detection in Online Discussion Forums. In *27th International Conference on Computational Linguistics* (pp. 1837–1848). Santa Fe, New Mexico, USA: Association for Computational Linguistics (ACL). URL: https://www.aclweb.org/anthology/C18-1156. arXiv:1805.06413.

Hee, C. V. (2017). *Can machines sense irony ?*. Ph.D. thesis Universiteit Gent.

Hernández, L., López-Lopez, A., & Medina-Pagola, J. E. (2011). Classification of Attitude Words for Opinions Mining. *International Journal of Computational Linguistics and Applications*, *2*, 267–283.

Hernández Farías, D. I., Benedí, J.-M., & Rosso, P. (2015). Applying Basic Features from Sentiment Analysis for Automatic Irony Detection. In R. Paredes, J. S. Cardoso, & X. M. Pardo (Eds.), *Pattern Recognition and Image Analysis* (pp. 337–344). Santiago de Compostela, Spain: Springer International Publishing volume 9117 of *Lecture Notes in Computer Science*. doi:10.1007/978-3-319-19390-8_38.

Hernández Farías, D. I., Patti, V., & Rosso, P. (2016). Irony detection in Twitter: The role of affective content. *ACM Transactions on Internet Technology*, *16*, 1–24. doi:10.1145/2930663.

Hernández Farías, D. I., Prati, R., Herrera, F., & Rosso, P. (2020). Irony detection in Twitter with imbalanced class distributions. *Journal of Intelligent & Fuzzy Systems*, (pp. 1–17). doi:10.3233/jifs-179880.

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*, 1735–1780. doi:10.1162/neco.1997.9.8.1735.

Hogenboom, A., Bal, D., Frasincar, F., Bal, M., de Jong, F., & Kaymak, U. (2013). Exploiting Emoticons in Sentiment Analysis. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing* SAC '13 (pp. 703–710). New York, NY, USA: ACM. URL: http://doi.acm.org/10.1145/2480362.2480498. doi:10.1145/2480362.2480498.

Hotelling, H. (1936). Relations Between Two Sets of Variates. *Biometrika*, *28*, 321–377. doi:10.2307/2333955.

Howard, J., & Ruder, S. (2018). Universal Language Model Fine-tuning for Text Classification. In *56th Annual Meeting ofthe Association for Computational Linguistics* (pp. 328–339). Melbourne, Australia,: Association for Computational Linguistics. doi:10.3760/cma.j.issn.04124081.2010.02.006.

Huang, Y.-H., Huang, H.-H., & Chen, H.-H. (2017). Irony Detection with Attentive Recurrent Neural Networks. In J. M. Jose, C. Hauff, I. S. Altıngovde, D. Song, D. Albakour, S. Watt, & J. Tait (Eds.), *Advances in Information Retrieval* (pp. 534–540). Cham: Springer International Publishing.

Irsoy, O., & Cardie, C. (2014). Deep recursive neural networks for compositionality in language. In *Advances in Neural*

Information Processing Systems (pp. 2096–2104). volume 3. URL: http://www.scopus.com/inward/record.url?eid=2-s2.0-84937828128{&}partnerID=tZOtx3y1.

Iyyer, M., Manjunatha, V., Boyd-Graber, J., & Daumé, H. (2015). Deep Unordered Composition Rivals Syntactic Methods for Text Classification. In *53rd Annual Meeting ofthe Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (pp. 1681–1691). Beijing, China: Association for Computational Linguistics.

Jain, S., & Wallace, B. C. (2019). Attention is not Explanation. In *17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)* (pp. 3543–3556). Minneapolis, Minnesota, USA: Association for Computational Linguistics (ACL).

Jasso López, G., & Meza Ruiz, I. (2016). Character and Word Baselines Systems for Irony Detection in Spanish Short Texts. *Procesamiento del Lenguaje Natural*, *56*, 41–48. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/5285.

John Haiman (1998). *Talk is Cheap: Sarcasm, Alienation, and Evolution of Language*. New York, USA: Oxford University Press. doi:10.1017/s0047404500211032.

Joshi, A., Bhattacharyya, P., & Carman, M. J. (2018). *Investigations in computational sarcasm* volume 37. Springer Nature.

Joshi, A., Tripathi, V., Patel, K., Bhattacharyya, P., & Carman, M. J. (2016). Are Word Embedding-based Features Useful for Sarcasm Detection? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November, 2016* (pp. 1006–1011).

Justo, R., Alcaide, J. M., Torres, M. I., & Walker, M. (2018). Detection of Sarcasm and Nastiness: New Resources for Spanish Language. *Cognitive Computation*, *10*, 1135–1151. doi:10.1007/s12559-018-9578-5.

Karoui, J., Benamara, F., & Moriceau, V. (2019). *Automatic Detection of Irony*. (1st ed.). John Wiley & Sons, Inc. doi:10.1002/9781119671183.

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 Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)* (pp. 644–650). Association for Computational Linguistics.

Karoui, J., Benamara, F., 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 *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference* (pp. 262–272). doi:10.18653/v1/e17-1025.

Karouia, J., Zitoune, F. B., & Veronique Moriceau (2017). SOUKHRIA: Towards an Irony Detection System for Arabic in Social Media. In *3rd International Conference on Arabic Computational Linguistics, ACLing 2017* (pp. 161–168). Dubai, United Arab Emirates: Association for Computacional Linguistic (ACL).

Khattri, A., Joshi, A., Bhattacharyya, P., & Carman, M. (2015). Your Sentiment Precedes You: Using an Author's Historical Tweets to Predict Sarcasm. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (pp. 25–30). Lisboa, Portugal: Association for Computational Linguistics.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In Y. Bengio, & Y. LeCun (Eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. URL: http://arxiv.org/abs/1412.6980.

Kreuz, R. J., & Glucksberg, S. (1989). How to Be Sarcastic: The Echoic Reminder Theory of Verbal Irony. *Journal of Experimental Psychology: General*, *118*, 374–386. doi:10.1037/0096-3445.118.4.374.

Kreuz, R. J., & Link, K. E. (2002). Asymmetries in the use of verbal irony. *Journal of Language and Social Psychology*, *21*, 127–143. doi:10.1177/02627X02021002002.

Kreuz, R. J., & Roberts, R. M. (1993). On Satire and Parody: The Importance of Being Ironic. *Metaphor and Symbolic Activity*, *8*, 97–109. doi:10.1207/s15327868ms0802_2.

Kumar, A., Narapareddy, V. T., Srikanth, V. A., Malapati, A., & Neti, L. B. M. (2020). Sarcasm Detection Using Multi-Head Attention Based Bidirectional LSTM. *IEEE Access*, *8*, 6388–6397. doi:10.1109/ACCESS.2019.2963630.

Kunneman, F., Liebrecht, C., van Mulken, M., & van den Bosch, A. (2015). Signaling Sarcasm: From Hyperbole to Hashtag . *Information Processing & Management*, *51*, 500 – 509.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). ALBERT: A lite BERT for self-supervised learning of language representations. In *International Conference on Learning Representation (ICLR 2020)* (pp. 1–17). arXiv:arXiv:1909.11942v6.

Le, Q., & Mikolov, T. (2014a). Distributed representations of sentences and documents. In *31st International Conference on International Conference on Machine Learning - Volume 32* (pp. II–1188–II–1196). Beijing, China volume 4. arXiv:1405.4053.

Le, Q., & Mikolov, T. (2014b). Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32* ICML'14 (p. II–1188–II–1196). JMLR.org.

Levi, O., Hosseini, P., Diab, M., & Broniatowski, D. A. (2019). Identifying Nuances in Fake News vs. Satire: Using Semantic and Linguistic Cues. In *arXiv*. doi:10.18653/v1/d19-5004. arXiv:1910.01160.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv*, . arXiv:1907.11692.

Lo, S. L., Cambria, E., Chiong, R., & Cornforth, D. (2017). Multilingual sentiment analysis: from formal to informal and scarce resource languages. *Artificial Intelligence Review*, *48*, 499–527. URL: https://doi.org/10.1007/s10462-016-9508-4. doi:10.1007/s10462-016-9508-4.

Lucariello, J. (1994). Situational Irony: A Concept of Events Gone Awry. *Journal of Experimental Psychology: General*, *123*, 129–145. doi:10.1037/0096-3445.123.2.129.

Luong, T., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In *Proceed-*

ings of the 2015 Conference on Empirical Methods in Natural Language Processing (pp. 1412–1421). Lisbon, Portugal: Association for Computational Linguistics. URL: https://www.aclweb.org/anthology/D15-1166. doi:10.18653/v1/D15-1166.

Majumder, N., Poria, S., Peng, H., Chhaya, N., Cambria, E., & Gelbukh, A. (2019). Sentiment and sarcasm classification with multitask learning. *IEEE Intelligent Systems*, *34*, 38–43. doi:10.1109/MIS.2019.2904691.

Maynard, D., & Greenwood, M. A. (2014). Who cares about sarcastic tweets? Investigating the impact of sarcasm on sentiment analysis. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014* (pp. 4238–4243). Reykjavik, Iceland: European Language Resources Association (ELRA).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems pp. 3111–3119*.

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, *38*, 39–41.

Miranda-Belmonte, H. U., & López-Monroy, A. P. (2019). Early Fusion of Traditional and Deep Features for Irony Detection in Twitter. 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)*.. Bilbao, Spain: CEUR Workshop Proceedings. CEUR-WS.org.

Nozza, D., Fersini, E., & Messina, E. (2016). Unsupervised Irony Detection: A Probabilistic Model with Word Embeddings. In *8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management* (pp. 68–76). volume 1. doi:10.5220/0006052000680076.

Ortega, R., Rangel, Francisco Hernández Farías, D. I., Rosso, P., Montes, M., & Medina, J. E. (2019). Overview of the Task on Irony Detection in Spanish Variants. In *Iberian Languages Evaluation Forum (IberLEF 2019), co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019)*. (pp. 229–256). Bilbao, Spain: CEUR Workshop Proceedings. CEUR-WS.org.

Ortega-Bueno, R., & Medina Pagola, J. E. (2018). UO_IRO: Linguistic informed deep-learning model for irony detection. In *CEUR Workshop Proceedings* (pp. 1–6). CEUR Workshop Proceedings. CEUR-WS.org volume 2263. doi:10.4000/books.aaccademia.4638.

Ortega-Bueno, R., Medina-Pagola, J. E., Muñiz-Cuza, C. E., & Rosso, P. (2018a). Improving attitude words classification for opinion mining using word embedding. In R. Vera-Rodríguez, J. Fiérrez, & A. Morales (Eds.), *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications - 23rd Iberoamerican Congress, CIARP 2018, Madrid, Spain, November 19-22, 2018, Proceedings* (pp. 971–982). Springer volume 11401 of *Lecture Notes in Computer Science*. URL: https://doi.org/10.1007/978-3-030-13469-3_112. doi:10.1007/978-3-030-13469-3_112.

Ortega-Bueno, R., Muñiz, C. E., Rosso, P., & Medina-Pagola, J. E. (2018b). UO_UPV : Deep Linguistic Humor Detection in Spanish Social Media. In P. Rosso, J. Gonzalo, R. Martínez, S. Montalvo, & J. Carrillo-de Albornoz (Eds.), *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018)* (pp. 203–213). Sevilla, Spain: CEUR-WS.org.

Ortega-Bueno, R., Rosso, P., & Medina Pagola, J. E. (2019). UO_UPV2 at HAHA 2019: BiGRU neural network informed with linguistic features for humor recognition. In *CEUR Workshop Proceedings*. Bilbao, Spain: CEUR Workshop Proceedings. CEUR-WS.org.

Padró, L., & Stanilovsky, E. (2012). FreeLing 3.0: Towards Wider Multilinguality. In *Proceedings of the (LREC 2012)*.

Pagliardini, M., Gupta, P., & Jaggi, M. (2018). Unsupervised learning of sentence embeddings using compositional n-gram features. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 528–540). New Orleans, Louisiana: Association for Computational Linguistics. URL: https://www.aclweb.org/anthology/N18-1049. doi:10.18653/v1/N18-1049.

Peña, A. S., García, L. A., & Dosina, A. R. (2018). Detección de ironía en textos cortos enfocada a la minería de opinión. In *IV Conferencia Internacional en Ciencias Computacionales e Informáticas (CICCI' 2018)* 1-10. Havana, Cuba.

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. In *2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543). Doha, Qatar: Association for Computational Linguistics.

Perkins, J. (2014). *Python 3 Text Processing With NLTK 3 Cookbook*. Packt Publishing. arXiv:arXiv:1011.1669v3.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference* (pp. 2227–2237). Association for Computational Linguistics volume 1. doi:10.18653/v1/n18-1202. arXiv:1802.05365.

Poria, S., Cambria, E., Hazarika, D., & Vij, P. (2016). A Deeper Look into Sarcastic Tweets Using Deep Convolutional Neural Networks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 1601–1612). Osaka, Japan: Association for Computational Linguistics. URL: https://www.aclweb.org/anthology/C16-1151.

Pota, M., Ventura, M., Catelli, R., & Esposito, M. (2021a). An effective bert-based pipeline for twitter sentiment analysis: A case study in italian. *Sensors*, *21*. URL: https://www.mdpi.com/1424-8220/21/1/133. doi:10.3390/s21010133.

Pota, M., Ventura, M., Fujita, H., & Esposito, M. (2021b). Multilingual evaluation of pre-processing for bert-based sentiment analysis of tweets. *Expert Systems with Applications*, *181*, 115119. URL: https://www.sciencedirect.com/science/article/pii/S0957417421005601. doi:https://doi.org/10.1016/j.eswa.2021.115119.

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. URL: https://doi.org/10.1007/s00521-020-05102-3. doi:10.1007/s00521-020-05102-3. arXiv:1911.10401.

Ptáček, T., Habernal, I., & Hong, J. (2014). Sarcasm Detection on Czech and English Twitter. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics* (pp. 213–223). Dublin, Ireland: Dublin City University and Association for Computational Linguistics.

Rangel, F., Hernández Farías, D. I., & Rosso, P. (2014). Emotions and Irony per Gender in Facebook. In *Proc. Workshop on Emotion, Social Signals, Sentiment & Linked Open Data (ES3LOD), LREC-2014* (pp. 68–73). Reykjavík, Iceland.

Ravi, K., & Ravi, V. (2018). Irony Detection Using Neural Network Language Model, Psycholinguistic Features and Text Mining. In *IEEE 17th International Conference on Cognitive Informatics and Cognitive Computing, ICCI*CC 2018* (pp. 254–260). doi:10.1109/ICCI-CC.2018.8482094.

Raymond W. Gibbs, J., & Colston, H. L. (2012). *Interpreting Figurative Meaning*. Cambridge University Press. doi:10.1080/10926488.2018.1407996.

Reganti, A. N., & Maheshwari, T. (2016). Modeling Satire in English Text for Automatic Detection. In *IEEE 16th International Conference on Data Mining Workshops* (pp. 970–977). IEEE. doi:10.1109/ICDMW.2016.146.

Reyes, A. (2012). *Linguistic-based Patterns for Figurative Language Processing : The Case of Humor Recognition and Irony Detection Antonio Reyes P ´*. Phd Universitat Politècnica de València.

Reyes, A., Rosso, P., & Veale, T. (2013). A Multidimensional Approach for Detecting Irony in Twitter. *Language Resources and Evaluation*, *47*, 239–268.

Riloff, E., Qadir, A., Surve, P., Silva, L. D., Gilbert, N., & Huang, R. (2013). Sarcasm as Contrast between a Positive Sentiment and Negative Situation. *Emnlp*, (pp. 704–714).

Rocktäschel, T., Grefenstette, E., Hermann, K. M., Kočiský, T., & Blunsom, P. (2016). Reasoning about entailment with neural attention. *4th International Conference on Learning Representations, ICLR 2016*, (pp. 1–9). arXiv:1509.06664.

Rubin, V. L., Conroy, N. J., Chen, Y., & Cornwell, S. (2016). Fake News or Truth ? Using Satirical Cues to Detect Potentially Misleading News. In *Workshop on Computational Approaches to Deception Detection at the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-CADD2016)* (pp. 7–17). California, USA: Association for Computational Linguistics. doi:10.18653/v1/W16-0802.

Salas-Zárate, M. d. P., Paredes-Valverde, M. A., Rodriguez-García, M. Á., Valencia-García, R., & Alor-Hernández, G. (2017). Automatic detection of satire in Twitter: A psycholinguistic-based approach. *Knowledge-Based Systems*, *128*, 20–33. URL: http://dx.doi.org/10.1016/j.knosys.2017.04.009. doi:10.1016/j.knosys.2017.04.009.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. In *arXiv* (pp. 1–5). arXiv:1910.01108.

Saralegi, X., & Vicente, I. S. (2013). Elhuyar at TASS 2013. In *XXIX Congreso de la Sociedad Española de Procesamiento de Lenguaje Natural". Workshop on Sentiment Analysis at SEPLN (TASS2013)* (pp. 143–150).

Sarkar, S. D., Yang, F., & Mukherjee, A. (2018). Attending Sentences to detect Satirical Fake News. In *27th International Conference on Computational Linguistics (COLING'18)* (pp. 3371–3380).

Seda Mut Altin, L., Bravo, A., & Saggion, H. (2019). LaSTUS/TALN at IroSvA: 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)*.. Bilbao, Spain: CEUR Workshop Proceedings. CEUR-WS.org.

Serrano, S., & Smith, N. A. (2020). Is attention interpretable? In *57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference* (pp. 2931–2951). doi:10.18653/v1/p19-1282. arXiv:1906.03731.

Sidorov, G., Miranda-Jiménez, S., Viveros-Jiménez, F., Gelbukh, A., Castro-Sánchez, N., Velásquez, F., D\'\iaz-Rangel, I., Suárez-Guerra, S., Treviño, A., & Gordon, J. (2013). Empirical Study of Machine Learning Based Approach for Opinion Mining in Tweets. In *Proceedings of the 11th Mexican International Conference on Advances in Artificial Intelligence - Volume Part I* MICAI'12 (pp. 1–14). Berlin, Heidelberg: Springer-Verlag. URL: http://dx.doi.org/10.1007/978-3-642-37807-2{_}1. doi:10.1007/978-3-642-37807-2_1.

Simpson, P. (2003). *On the Discourse of Satire: Towards a Stylistic Model of Satirical Humour* volume 2. John Benjamins Publishing Company. doi:10.1177/0963947006060558.

Singh, R. K., Sachan, M. K., & Patel, R. (2021). 360 degree view of cross-domain opinion classification: a survey. *Artificial Intelligence Review*, *54*, 1385–1506.

Sperber, D., & Wilson, D. (1981). Irony and the use-mention distinction. In P. Cole (Ed.), *Radical Pragmatics* (pp. 295–318.). , New York: Academic Press.

Susanto, Y., Livingstone, A. G., Ng, B. C., & Cambria, E. (2020). The hourglass model revisited. *IEEE Intelligent Systems*, *35*, 96–102. doi:10.1109/MIS.2020.2992799.

Tang, Y.-j., & Chen, H.-H. (2014). Chinese Irony Corpus Construction and Ironic Structure Analysis. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics* (pp. 1269–1278). Dublin, Ireland: Association for Computational Linguistics.

Tenney, I., Das, D., & Pavlick, E. (2020). BERT rediscovers the classical NLP pipeline. In *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics* (pp. 4593–4601). Association for Computational Linguistics (ACL). doi:10.18653/v1/p19-1452. arXiv:1905.05950.

Thu, P. P., & Aung, T. N. (2018). Implementation of Emotional Features on Satire Detection. *International Journal of Networked and Distributed Computing*, *6*, 78–87.

Thu, P. P., & Nwe, N. (2017). Impact Analysis of Emotion in Figurative Language. In *16th IEEE/ACIS International Conference on Computer and Information Science (ICIS'17)* (pp. 209–214).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *31st Conference on Neural Information Processing Systems (NIPS 2017)* (pp. 1—-11). Long Beach, CA, USA.

Veale, T., & Hao, Y. (2009). Support Structures for Linguistic Creativity : A Computational Analysis of Creative Irony in

Similes. In *Proceedings of CogSci 2009, the 31st Annual Meeting of the Cognitive Science Society* (pp. 1376–1381).
Vig, J., & Belinkov, Y. (2019). Analyzing the Structure of Attention in a Transformer Language Model. In *2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (pp. 63—-76). Florence, Italy: Association for Computational Linguistics (ACL). doi:10.18653/v1/w19-4808. arXiv:1906.04284.
Vilares, D., Peng, H., Satapathy, R., & Cambria, E. (2018). Babelsenticnet: a commonsense reasoning framework for multilingual sentiment analysis. In *2018 IEEE Symposium Series on Computational Intelligence (SSCI)* (pp. 1292–1298). IEEE.
W., J., & Pennebarker (2011). *James w.* volume 1890. (Firtst ed.). Bloomsbury Press.
Wallace, B. C. (2015). Computational Irony: A Survey and New Perspectives. *Artificial Intelligence Review*, *43*, 467–483.
Wallace, B. C., Choe, D. K., & Charniak, E. (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). Beijing, China: Association for Computational Linguistics.
Wang, Y., Huang, M., Zhao, L., & Others (2016). Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 606–615).
Wilson, D., & Sperber, D. (1992). On verbal irony. *Lingua*, *87*, 53–76. doi:10.1016/0024-3841(92)90025-E.
Wu, C., Wu, F., Wu, S., Liu, J., Yuan, Z., & Huang, Y. (2018). THU_NGN at SemEval-2018 Task 3: Tweet Irony Detection with Densely connected LSTM and Multi-task Learning. In *12th International Workshop on Semantic Evaluation* (pp. 51–56). New Orleans, Louisiana: Association for Computational Linguistics (ACL). doi:10.18653/v1/s18-1006.
Yang, F., Mukherjee, A., & Gragut, E. (2017a). Satirical News Detection and Analysis using Attention Mechanism and Linguistic Features. In *Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)* (pp. 1979–1989). Copenhagen, Denmark: Association for Computational Linguistics.
Yang, M., Tu, W., Wang, J., Xu, F., & Chen, X. (2017b). Attention Based LSTM for Target Dependent Sentiment Classification. In *AAAI* (pp. 5013–5014).
Yang, Y., Cer, D., Ahmad, A., Guo, M., Law, J., Constant, N., Hernandez Abrego, G., Yuan, S., Tar, C., Sung, Y.-h., Strope, B., & Kurzweil, R. (2020). Multilingual Universal Sentence Encoder for Semantic Retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 87–94). Online: Association for Computational Linguistics. URL: https://www.aclweb.org/anthology/2020.acl-demos.12. doi:10.18653/v1/2020.acl-demos.12.
Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1480–1489).
Zhang, C., & Abdul-Mageed, M. (2019). Multi-task bidirectional transformer representations for irony detection. In *CEUR Workshop Proceedings* (pp. 391–400). volume 2517. arXiv:1909.03526.
Zhang, S., Zhang, X., Chan, J., & Rosso, P. (2019). Irony detection via sentiment-based transfer learning. *Information Processing and Management*, *56*, 1633–1644. URL: https://doi.org/10.1016/j.ipm.2019.04.006. doi:10.1016/j.ipm.2019.04.006.
Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., & Xu, B. (2016). Attention-based bidirectional long short-term memory networks for relation classification. In *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Short Papers* (pp. 207–212). Berlin, Germany: Association for Computational Linguistics (ACL). doi:10.18653/v1/p16-2034.
Zucco, C., Liang, H., Fatta, G. D., & Cannataro, M. (2019). Explainable Sentiment Analysis with Applications in Medicine. In *IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2018* (pp. 1740–1747). doi:10.1109/BIBM.2018.8621359.

,

Appendix A. Linguistic Features

Group	Feature	Description
Stylistics-based	multiLines	It takes into account whether the tweet is composed of multiple lines or not (one vs. many lines).
	lenghtW lengthC meanLengthW	Three different features are considered; i) the number of words, ii) the number of characters, and iii) the means of words' length in the tweet.
	isDialog nDialogMark	Two distinct features are considered; i) the tweet contains any line that starts with a long dash (dialogue marker), ii) the number of lines that start with long dashes.
	hashtagsFreq urlsFreq emojisFreq	These count the number of hashtags, URLs, and emojis in the tweet, respectively.
	exclMarkFreq	It counts the exclamation marks in the tweet.
	wordRep wordUpper wordCharRep wordWithExcl	Four distinct features are considered: i) the number of words emphasized by word's repetition, ii) the number of words with emphasis by uppercase, iii) the number of words emphasized by character flooding, and iv) the number of words emphasized by continues exclamation marks.

Group	Feature	Description
	alliter	It captures the occurrence of simple alliteration in the tweet. For that, we considered a fixed-length sequence of phonetic prefixes with size=3.
	quotation	It quantifies the phrases enclosed in a double quote.
	Q?A	It quantifies the question and answer structures in the tweet.
	person_p[7]	It quantifies the number of verbs conjugated in the first, second, third persons and the nouns and adjectives which agree with such verbal conjugations.
	tense_t[8]	It quantifies the usage of different verbal tenses in the tweet.
	posN posV posA posR	These count the nouns, verbs, adverbs and adjectives in the tweet.
	Punctuation	It counts the occurrence of dots, commas, semicolons, and question marks in the tweet.
Content-based	Animal centred-words	It counts the words that occur in a lexicon of animal names.
	Toponym words	It counts the words that occur in a lexicon of country's names, capital's names, city's names and nationalities.
	ObsceneSexual words	It counts the words that occur in an in-house lexicon of sexual and obscene words.
Semantic-based	Antonyms	It quantifies the pairs of antonyms that occurs in the tweet. This feature is based on the antonym' relations provided by WordNet Miller (1995), particularly, for the Spanish language we used the Multilingual Central Repository (MCR) (Gonzalez-Agirre et al., 2012).
	LexAmbiguity	Three different features are considered; i) the average of the meanings associated with each word in the tweet, ii) the number of meanings for the most ambiguous word in the tweet, iii) the gap between the value of two previous features.
	DomAmbiguity	Conversely, to consider the meanings of the words, in these features we consider the number of domains assigned to the words. Particularly, three distinct features are considered; i) the average of domains associated with each word in the tweet, ii) the greatest number of domains that a single word has in the tweet, iii) the gap between the value of the two previous features. For obtaining the domains of the words we used the WordNet Domains[9] and SUMO[10] each separately.
	SVerb_classes	These features capture distinct semantic frames of the verbs in the tweet based on ADDESE[11].
	Negation	It counts the negation words in the tweet.
Affective-based	SSL_polarity ESL_polarity CriSol_polarity LAM11_polarity[12] SenticNet_polarity	These features count positive and negative words in many sentiment resources. Notice that, for each resource two features are computed. Particularly, we explore four distinct dictionaries: Spanish Sentiment Lexicon (SSL) González et al. (2015), Elhuyar Sentiment Lexicon (ESL) (Saralegi & Vicente, 2013), CriSol lexicon (González et al., 2015), and the lexicon LAM11 introduced in (Hernández et al., 2011). Moreover, the polarity score associated with the words and concepts in SenticNet was considered.
	emojiPol_pos emojiPol_neg	The number of positive and negative emoticons and emojis considering the resource Emoticons Sentiment Hogenboom et al. (2013).
	LAM11_attitude eCrisol_attitude[13]	These features count the number of words according to the three distinct attitude categories (affect, judgment, and appreciation) proposed in (Hernández et al., 2011). For that, we considered two lexicons, i) the LAM11 lexicon introduced in (Hernández et al., 2011) and an extended version of the CriSol lexicon, where all words were automatically annotated with attitudes (eCrisol) by using the method proposed in (Ortega-Bueno et al., 2018a).
	EmoCat	These features count the number of words according to the six basic emotions provided by the resource SEL (Sidorov et al., 2013).

[7] p is parametric to the three persons used in Spanish grammar.
[8] t is parametric to the various tense in Spanish grammar i.e., present, past, future, etc.
[9] http://wndomains.fbk.eu/hierarchy.html
[10] http://www.adampease.org/OP/
[11] http://adesse.uvigo.es/data/clases.php
[12] *polarity* is parametric to the type of sentiment, positive and negative
[13] *attitude* is parametric to the type of attitudes affect, judgement, and appreciation

Group	Feature	Description
	EmoDim	These features are based on the four affective dimensions in SenticNet of the Cambria's hourglass of emotions model (Susanto et al., 2020; Cambria et al., 2020): introspection, temper, attitude and sensitivity [14].
Contrast-based[15]	wordPolCont	It computes the gap between the most positive and the most negative word in the tweet. This feature, consider the distance, in terms of tokens, between the words.
	emoTextPolCont	It computes the polarity difference between emoticons and words in the tweet.
	antConsPolCont	It considers the polarity contrast between two parts of the tweet when the tweet is split by a delimiter. In this work we consider as delimiter some adverbs and punctuation marks.
	meanPolPhrase	It is the mean of the polarities of the words that belong to phrases enclosed by quotes.
	polStandDev	It is the standard deviation of the polarities of the words that belong to phrases enclosed by quotes.
	prePastPolCont	It computes the polarity difference between the parts of the tweet written in present and past tenses.
	skipGPolRate	It computes the rate among skip-grams with polarity opposition on the total of candidate skip-grams. The candidate skip-grams are those composed of two words (nouns, adjectives, verbs, adverbs) with skip=1. The skip-grams with polarity opposition are those that match with the patterns positive-negative, positive-neutral, negative-neutral, and vise-versa.
	upperTextPolCont	It computes the polarity difference between capitalized words and the remainder words in the tweets.
Psycolinguistic-based	LIWC_cat[16]	These features count the frequency of words in each category provided by the resource Linguistic Inquiry and Word Count [17] dictionary (W. & Pennebarker, 2011).

Appendix B. Best MvAttLSTM Hyperparameters for Each Corpus

Table B.16: Hypermapameters for the Contextual MvAttLSTM on the IroSvA'19 corpora

Dataset	Model	Hyperparameters	Model	Hyperparameters
IroSvA'19-es	Contextual Self	*batch=256* *att=self* *h=1* *dp=0.25* *op=adam* *lr=0.01*	Contextual Multi	*batch=256* *att=multihead* *h=2* *dp=0.3* *op=adam* *lr=0.01*
IroSvA'19-mx	Contextual Self	*batch=32* *att=self* *h=1* *dp=0.3* *op=adam* *lr=0.001*	Contextual Multi	*batch=128* *att=multihead* *h=8* *dp=0.25* *op=adam* *lr=0.01*
IroSvA'19-cu	Contextual Self	*batch=128* *att=self* *h=1* *dp=0.4* *op=rmsprop* *lr=0.01*	Contextual Multi	*batch=256* *att=multihead* *h=8* *dp=0.3* *op=rmsprop* *lr=0.01*

[14] It is worthy to note that for the Spanish language, we used BabelSenticNet (Vilares et al., 2018). In this extension of SenticNet, the affective dimensions are sensitivity, attention, aptitude and pleasantness.

[15] With the aim of capturing some types of explicit polarity opposition, we included the features proposed in Peña et al. (2018). The Spanish version of SenticNet was used to determine the polarity contrast between different parts of the text.

[16] *cat* is parametric to the 68 categories in the LIWC 2001 Spanish dictionary.

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

Table B.17: Hyperparameters for the Early MvAttLSTM on the satire and humor corpora

Dataset	**Model**	**Hyperparameters**	**Model**	**Hyperparameters**
Salas'17-es	Early Self	*batch=32* *att=self* *h=1* *dp=0.25* *op=adam* *lr=0.01*	Early Multi	*batch=128* *att=multihead* *h=4* *dp=0.25* *op=adam* *lr=0.01*
Salas'17-mx	Early Self	*batch=32* *att=self* *h=1* *dp=0.4* *op=adam* *lr=0.01*	Early Multi	*batch=128* *att=multihead* *h=2* *dp=0.4* *op=rmsprop* *lr=0.01*
Barbieri'15-es	Early Self	*batch=32* *att=self* *h=1* *dp=0.25* *op=rmsprop* *lr=0.01*	Early Multi	*batch=128* *att=multihead* *h=4* *dp=0.3* *op=adam* *lr=0.01*
HAHA'19	Early Self	*batch=32* *att=self* *h=1* *dp=0.4* *op=adam* *lr=0.01*	Early Multi	*batch=128* *att=multihead* *h=4* *dp=0.25* *op=rmsprop* *lr=0.01*