Universidad del País Vasco Euskal Herriko Unibertsitatea

Genial... 🙄: Automatic Irony Detection in Spanish Tweets

Author: Paula Diez Ibarbia

Advisors: Rodrigo Agerri

hap/lap

Hizkuntzaren Azterketa eta Prozesamendua
Language Analysis and Processing

Final Thesis

June 2022

Departments: Computer Systems and Languages, Computational Architectures and Technologies, Computational Science and Artificial Intelligence, Basque Language and Communication, Communications Engineer.

Abstract

Irony is a form of non-literal speech that can alter the meaning of an utterance. Understanding irony may greatly impact Natural Language Processing (NLP) tasks such as sentiment analysis or stance detection. While a growing body of NLP research has started to focus on irony detection, little work has been conducted for other languages such as Spanish. This thesis aims to contribute to research in Spanish irony detection by, taking as a basis an existing dataset (IroSvA) for irony detection in Spanish, revising it and enriching it with annotations for irony types. The improved dataset constitutes the first corpus including labels for types of irony in Spanish. Furthermore, we undertake crosslingual experimentation on irony detection in three different evaluation settings: monolingual, multilingual, and crosslingual. For these experiments, Italian and English datasets were employed in addition to the Spanish ones. Results show that irony does not transfer easily across languages except in the case of Italian to Spanish, for which the results are surprisingly good. Furthermore, training on multiple languages does not help to improve results for irony detection. Results also demonstrate that monolingual language models perform better than multilingual ones. Finally, the thesis offers a detailed and comprehensive analysis and discussion on the difficulties in annotating and learning to detect irony.

Keywords: irony detection, irony category classification, Spanish, multilingual, crosslingual

Contents

List of Figures

List of Tables

1 Introduction

Have you ever needed to explain you were being ironic? Or have you, in turn, failed to recognize irony yourself? Despite its complexity, we, humans, often opt for ironic comments rather than literal ones; even though it can easily lead to misunderstandings between the speakers. We may do it for several reasons such as try to be funny, control the aggressiveness of a statement or elevating our status (Dews et al., 2014). So if we sometimes struggle to identify irony, what about machines? Irony can have a great impact on Natural Language Processing (NLP) tasks, such as Sentiment Analysis. For instance, a sentiment analysis shared-task was organized at SemEval 2014 with English tweets Rosenthal et al. (2014). There were two subtasks: a contextual polarity disambiguation subtask, and a message polarity classification subtask. For the testing phase, in addition to the tweets test set, some out-of-domain sets were prepared: one with LiveJournal sentences, and another one containing sarcastic tweets. Unlike LiveJournal out-of-domain sentences, the sarcastic tweets had a great impact on the results causing a considerable drop in the performance of the systems, especially in the message polarity classification task.

During the last decades, a growing body of research has started tackling the topic of automatic detection of irony as well as sarcasm, especially in written texts (e.g. Carvalho et al. (2009); Hee et al., 2018). Irony and sarcasm are often considered to be "essentially the same thing, with superficial differences" (Attardo, 2014). Indeed, little distinction is made between irony and sarcasm in the field of automatic detection of irony since, often, datasets are constructed by extracting tweets with user-labeled tags such as #irony and #sarcasm (e.g. Ghosh and Muresan, 2018; Hee et al., 2018). Different approaches have been proposed to automatically detect irony, the most popular one being binary classifiers that aim to tell whether a given text contains irony or not (for instance Ghosh and Muresan, 2018; Liebrecht et al., 2013). Nonetheless, to a lesser extent, irony types or categories have been additionally taken into account (Hee et al., 2018, Cignarella et al., 2018b, Cignarella et al., 2018a). Also, in a quite different approach, Ghanem et al. (2020) looked into the transferability of irony devices.

Similar to other NLP tasks, most of the research has been conducted in English (for instance, Hee et al., 2018; Ghosh et al., 2020) although other languages have been considered, such as Italian (Cignarella et al., 2018b), Deutch (Liebrecht et al., 2013) or Portuguese (Corrêa et al., 2021). Regarding Spanish, scarce research has been done in the irony detection field (López and Ruiz, 2016; Ortega-Bueno et al., 2019). Additionally, to our knowledge, Spanish irony detection has only been tackled from an irony/not-irony point of view, only considering Spanish data.

Despite the recent advances in the area of irony detection, gaps still abound. This research seeks to address several issues to fill some of that void. In general terms, the purpose of this project was to contribute to the development of irony detection in Spanish. In particular, our approach centers on:

a) irony detection in three different contexts (monolingual, multilingual, and crosslingual) with three languages (Spanish, English, and Italian);

b) and irony-category classification with Spanish tweets.

To address our goals, an existing corpus of Spanish tweets for irony detection named IroSvA (Ortega-Bueno et al., 2019) was employed. This corpus, however, was re-annotated in order to suit a more particular concept of irony. Once the annotation was done, some experiments were prepared to address the questions related to irony detection. Irony detection consists of telling apart tweets containing irony from tweets without irony. The first set of irony detection experiments were monolingual experiments. In other words, training and testing were conducted with the same corpus. This step helped establish some baselines in order to properly analyze the following experiments. With each dataset, two systems were trained: one with a monolingual transformer, and another one with a multilingual transformer. So apart from setting some baselines, these experiments also helped answer the first research question:

RQ1: Are monolingual transformers more suited than multilingual transformers for irony detection? If so, to what extent?

For the second set of irony detection experiments, multilingual data augmentation was considered. For this, the Spanish, English and Italian training sets were joined and used to fine-tune the models. Then, the models were tested in each language separately. Data augmentation has been employed in Natural Language Processing (NLP) tasks to improve the results obtained for low-resource languages (Wang et al. (2020)). Neither Spanish nor English and Italian can be considered low-resource languages, per se; however, data available for irony detection is still scarce. Thus, we aimed to investigate whether data augmentation could improve the baseline results. Therefore, we can formulate the second research question in the following way:

RQ2: Does multilingual data augmentation help improve the results obtained from the monolingual experiments?

The third and last set of experiments for irony detection aimed to examine the universality of irony cues and structures. For this, in the same vein as Ghanem et al. (2020), we proposed a crosslingual setting, also known as zero-shot experiments. In other words, training in one language and testing in a different one. In this project, only pairs with Spanish were taken into account. These experiments were deemed interesting because of the possible applications to low-resource languages like Basque. If there are factors in common in irony realization regardless of the language, low-resource languages could exploit other language resources to boost irony detection on low-resource languages. Thus, the third research question can be formulated as follows:

RQ3: Are irony cues and structures universal and, thus, transferable across languages?

For the second part of the main research question (related to irony categories), we annotated ironic tweets in terms of irony categories. Then, similar to the monolingual irony detection experiments, we trained some classifiers with a monolingual and a multilingual transformer. These experiments were only conducted in Spanish and aimed to answer the fourth and last research question:

RQ4: How well can classifiers distinguish between the different irony categories in Spanish tweets?

Therefore, the main contributions of this project are the following:

1. We provide a re-annotation work of an existing corpus in terms of irony or non-irony, with a narrower and more defined concept of irony than the one in the original corpus;
2. We publish the first corpus of Spanish tweets annotated in terms of irony categories[1];
3. We offer a detailed comparison between multilingual and monolingual transformers to see which is more suitable for irony detection and to which extent;
4. We perform the first multilingual and crosslingual irony detection experiments that include Spanish;
5. We conduct the first study on irony-types classification in Spanish;
6. A full irony detection and classification pipeline is built, showing that much work remains to be done for irony detection and classification in Spanish tweets.

The rest of the document is organized as follows. First, Chapter 2 will discuss the concept of irony from a linguistic point, and review some of the investigations conducted in the field of irony detection. Next, Chapter 3 will introduce the materials employed during the project. Afterwards, Chapter 4 will discuss the data annotation process. Subsequently, Chapter 5 will explain the details of the experiments conducted. Next, Chapter 6 will present and discuss the results obtained from the experiments. Lastly, Chapter 7 will conclude the paper by summarizing the main points of the paper.

[1] `https://github.com/ixa-ehu/irosva-irony-types`

2 Related Work

This section will review the current state of the art in irony detection. To this aim, first, a linguistic explanation of irony and its features will be offered in order to better understand the phenomenon. Next, several works on irony detection will be introduced. Lastly, special attention will be paid to the works conducted for Spanish irony detection.

2.1 Linguistic framework

Irony has been vastly researched from different areas such as linguistics, philosophy, psychology... And the themes and points explored are as wide. In this section, only the most relevant aspects for the project will be tackled, which include: the concept of irony and how it is different from other figurative speech figures; the differences between sarcasm and irony; and the importance of the domain.

The concept of irony

Irony is constantly present in our everyday life, yet providing a solid definition of it seems to be a fairly difficult task. Throughout history, several types of irony have been considered (e.g. Socratic or dramatic), out of which two are relevant for this project: verbal irony and situational irony. The latter seems to be less controversial than the former, thus let us start from there. Situational irony is generally understood as an event that seems to deviate from our expectations (Lucariello, 1994) such as a fire extinguisher on fire (example from Colston (2017)). In our minds, the extinguisher is labeled as 'to be used to put out fires'; therefore, when the extinguisher itself is the one on fire there seems to be rupture in our schema as one would not expect the two opposing elements to be together.

Regarding verbal irony[2], theories and approaches on what irony is and how it works are vast (Colston and Gibbs, 2014). Two definitions have been very recurrent in the literature that also seem to be popular within the area of irony detection, so we shall discuss them briefly. To begin with, irony has often been understood as: "saying something while meaning something else". This definition distinguishes between the literal meaning and the non-literal one, which would be called 'irony'. Nevertheless, this interpretation of irony has been argued to be too broad, without a clear distinction between irony and the other tropes such as metaphor or hyperbole (Kaufer, 1981 via Attardo (2014); Haverkate, 1990 via Attardo (2014)). In this line, several researchers have pointed out that irony and other tropes are different phenomena that sometimes are used together. For instance, Wilson (2013) remarks that hyperbole, though often used as a cue for irony, it is not inherently ironical. Examples (1) and (2) (from Wilson, 2013, pg. 24-25) are non-ironic hyperbolic instance:

[2]From now on it will be referred to as simply 'irony'

(1) *(To a very tall person)* Wow! You're a giant![3]

(2) *(To a very tall person)* Wow! You're the tallest man I've ever seen!

Similarly, Winner and Gardner (1993) point that irony and metaphor are different in three aspects, that are illustrated in Examples (3) and (4) (extracted from Colston and Gibbs, 2002, pg. 57-58). First, irony and metaphor differ in the relationship between what is said and the implication: metaphor focuses on the similarities while irony highlights opposition. As we can in Example (3), "teacher refers to the student's intellectual abilities using a familiar metaphorical comparison whereby the mind is conceived as a cutting instrument (the sharper the cutting instrument, the more a person is seen as possessing greater intellectual abilities)" (Colston and Gibbs, 2002, pg 58). On the other hand, the ironical statement in Example (4) focuses on the opposition in the use of 'sharp' for some scissors that are actually not sharp. The second feature in which irony and metaphor appear to diverge is the communicative function. Winner and Gardner (1993) argue that, while metaphor seems to be primarily descriptive (e.g. that the child in Example (3) is very smart), irony reveals the speaker's attitude towards something (such as annoyance in Example (4)). Lastly, comprehension demands are argued to be different as understanding metaphor requires knowledge about the topic and the vehicle while interpreting irony requires "the ability to make inferences about the speaker's state of mind" (Winner and Gardner, 1993, pg. 429). In the metaphorical sentence (Example (3)), the topic would be the intelligence of the child while the vehicle would be 'sharp'. Therefore, the listener needs to know that 'sharp' is a good attribute that is assigned to the child. On the other hand, in the ironical statement (Example (4)) the assistant should know that the speaker wanted to use the scissors and is not happy about the state they are in.

(3) You are a teacher at an elementary school. You are discussing a new student with your assistant teacher. The student did extremely well on her entrance examinations. You say to your assistant, "This one's really sharp."

(4) You are a teacher at an elementary school. You are gathering teaching supplies with your assistant teacher. Some of the scissors you have are in really bad shape. You find one pair that won't cut anything. You say to your assistant, "This one's really sharp."

The second most frequently used definitions of irony is narrower: "saying something while meaning the opposite". Nonetheless, some researchers find this definition not to be adequate either. For instance, Wilson and Sperber (2014) point out that there are ironic instances in which the opposite meaning is not conveyed as illustrated in Example (5)). The ironic interjection in Example (5), we could argue, does arise from the contrast between the weather at the time of the visit and the one described by the host. However, the speaker does not try to convey the opposite of what they are saying, they are rather

[3]Wilson (2013) notes that this response could arguably be a metaphor rather than hyperbole. Indeed, she points out that hyperbole is often closer to metaphor than irony.

trying to draw the attention to their disappointment as if they were saying "I really do have bad luck" or "I should have come next week". Indeed, if we were to negate the interjection in search of its opposite meaning we would get nonsensical or inaccurate interpretations: "Ah, Tuscany in not May" or "Ah, not Tuscany in May". In a similar vein, Giora (1995) explains that irony may have a 'more than' or 'less than' reading, rather than an 'opposite' one. As we can observe in Example (6), the speaker does not mean the opposite of the whole utterance ("I do not think the washing hasn't dried"[4]), rather they aim to invoke "a stronger interpretation in the form of "I am sure the washing hasn't dried""(Giora, 1995, pg. 246). Additionally, Wilson and Sperber (2014) mention that, if we take irony as "meaning the opposite of what it is said", scenarios like the one in Example (7) should be considered ironic too, yet they are not.

(5) You have invited me to visit you in Tuscany. Tuscany in May, you write, is the most beautiful place on earth. I arrive in a freak cold spell, wind howling, rain lashing down. As you drive me home along flooded roads, I turn to you and exclaim: "Ah, Tuscany in May!" (from Wilson and Sperber, 2014, pg. 37)

(6) It is a rainy day and somebody say "I *think* [italics added] the washing hasn't dried". (from Giora, 1995, pg. 246)

(7) We are out for a stroll and pass a car with a broken window. I turn to you and say the following: "Look, that car has all its windows intact". (from Grice, 1978 via Wilson and Sperber, 2014, pg. 38)

In short, the traditional accounts for irony seem to not be the perfect definitions as they both have been reported to have some inconsistencies. The first definition (irony as non-literal meaning) appears to be too broad, without a proper distinction between irony and other tropes (although such differences have been argued to exist). On the other hand, the second interpretation (irony as opposition) seems to be too narrow and not entirely accurate. Moreover, it should be mentioned that irony definitions in the literature are often discussed with manually chosen examples. In other words, irony definitions in the literature do not aim to offer a wide-coverage theory based on empirical corpus-based approaches.

Irony and Sarcasm

When one starts considering irony, a recurrent question seems to arise: what is the difference between irony and sarcasm? Some scholars consider irony and sarcasm to be two independent phenomena that can co-occur. For instance, Kreuz and Glucksberg (1989) claim that "people can use verbal irony without being sarcastic and can also be sarcastic without being ironic" (pg. 374). They explain that a non-sarcastic comment would not be targeted at anyone but just indicating displeasure, for instance when somebody comments "what a great weather" on a stormy day. On the contrary, they argue, an insincere apology

[4]Which is to say, "I think the washing is dry already"

("thanks a lot") would be an instance of non-ironical sarcastic statement. Basically, the difference would lie in the absence (irony) or presence (sarcasm) of a target (Lee and Katz, 1998). Similarly, Haiman (1998) pinpoints that a) "situations may be ironic, but only people can be sarcastic" (pg. 20), and b) that irony can be unintentional and unconscious whereas sarcasm requires intention. On the other hand, some other scholars suggest that irony and sarcasm are "essentially the same thing, with superficial differences" (Attardo, 2014). Indeed, sarcasm has often been classified as a type of irony (Attardo, 2014, Kreuz and Roberts, 1993) which is characterized by the (usually) negative or aggressive attitude is directed towards somebody (a target) (Attardo, 2014, Kreuz and Roberts, 1993). Also, sarcastic cues tend to be more obvious (Attardo, 2014).

Context and domain

It is common knowledge that context plays an important part on the realization of irony, which includes the modality of the speech (written or oral) (Hancock, 2004). The channel and conditions used to communicate may influence the choice of whether to use or not irony, and how to do it. Although not necessary, irony is often accompanied by irony markers to highlight its presence such as intonation or punctuation marks (Attardo, 2000). Moreover, modality seems not to be the only factor to influence the use of irony. Burgers et al. (2012) examined the realization of irony in six different genres of the written modality. Results suggested that genres had an influence in the use of irony, especially between multimodal genres (like cartoons or advertisements) or purely verbal genres (such as letters or reviews).

In terms of domain, in the area of automatic irony detection, Twitter has been the most popular source when gathering the data for the corpus (e.g. Cignarella et al., 2018b; Hee et al., 2018; López and Ruiz, 2016; Ortega-Bueno et al., 2019). This project too made use of tweets, therefore it is worth noting several features. Tweets are short texts with a limit of 240 characters posted on a social media called Twitter[5]. Some key elements can be found, like the use of hashtags to make reference to topics (#topic) and ats to refer to other users (@username). Additionally, text can also be accompanied by images or videos. The speech register used in this platform tends to be rather informal and frequently spelling errors as well as shortenings or 'chat-talk' ('xoxo' for 'kisses') can be encountered. Additionally, texts tend to often include emoticons or emojis, for instance smiling faces.

So far, irony has been presented from a linguistic point of view: some popular definitions; the difference between irony, sarcasm, and other tropes like hyperbole; and, the effect of genre on the realization of irony. Next, we will examine irony from a computational point of view.

[5] `https://twitter.com/`

2.2 Computational Approaches

Different approaches have been taken for irony and sarcasm[6] detection. For instance, Carvalho et al. (2009) focused on identifying patterns that could indicate the presence of irony. Eight patterns were considered: diminutive forms, demonstrative determiners, interjections, verb morphology, cross-constructions, heavy punctuation, quotation marks, and laughter expression. In order to analyze irony in discourse, online comments from a Portuguese newspaper were extracted. The posts were subjected to name-entity recognition (NER) with a name-entity lexicon. Sentences were labeled as ironic, non-ironic, undecided (for the ones lacking enough context), and ambiguous (for those in which the context could favor both an ironic reading and a non-ironic one). Results showed that the most favored patterns involved emoticons, expressions of laughter, heavy punctuation marks, quotations marks and positive interjections. Similar results were obtained in later research by Vanin et al. (2013). Bazilian Portuguese tweets were analyzed in search of fifteen different patterns that were categorized into seven groups: lists, exact expressions, part of speech, part of speech + exact expressions, part of speech + name entities, demonstrative pronouns + name entities, and symbols. Not all the patterns were found in the analysis, and the most productive out of the present ones were emoticons, expressions of laughter, heavy punctuation and the expression "só que (não)".

Karoui et al.'s (2017) approach focused on data annotation, proposing a multi-layered annotation that took into account irony activators, irony categories and irony markers. According to Karoui et al. (2017), in most cases, irony arises from two (or more) chunks that contradict at some level. These elements are the activators and can be both explicit (they can be found in the text) or implicit (at least one of the elements has to be deduced). Regarding categories, eight were considered: comparison, hyperbole, euphemism, rhetorical question, context shift, false assertion, paradox, and other. Most of them can be either implicit or explicit, except for false assertion (which is always implicit) or context shift (which is always explicit). Lastly, irony markers refer to a set of tokens that may trigger the irony, such as words, symbols or propositions. The annotation was conducted in tweets from three different languages: French, English and Italian. Slight differences were observed in the trends for each language. French and English seemed to be more fond of implicit irony while Italian seemed to favor explicit one; however, these variations could be due to the lack of user-generated ironic hashtags in the Italian corpus, as reported by Karoui et al. (2017). Moreover, regarding categories, the three languages seemed to prefer paradox for explicit irony. Regarding implicit irony, English and French showed a preference for false assertion and 'other'; whereas, Italian relied on false assertion, analogy and 'other'. When it comes to markers, intensifiers, punctuation marks and interjections were observed to be more present in French ironic tweets than in non-ironic ones. With reference to English, in ironic tweets the occurrence of discourse connectors, quotations, comparison words and reporting speech verbs seemed to be higher. Finally, in Italian, all the markers seem to be more present in ironic tweets than in non-ironic ones, with the exception of quotations and

[6]Please, note that, since the line between irony and sarcasm is very blurry and subjective, some sarcasm papers will be also included in this review

URLs.

Later, the TWITTIRÒ-UD corpus was annotated following Karoui et al. (2017) schema (Cignarella et al., 2019; Cignarella et al., 2020). The corpus, composed of 1,724 Italian tweets, was annotated taking into account irony activation type (implicit/explicit) and irony category. They exploited syntactical information by means of Universal Dependencies (UD). A preliminary analysis suggested that, in implicit irony, half of the categories favored a verb as a trigger. Moreover, all the categories seemed to prefer the parataxis[7] dependency relation.

Tang and Chen (2014) also focused on corpus construction as well as in pattern mining. Several patterns were analyzed in Chinese Plurk microblogs (similar to Twitter) and Yahoo blogs. Some of the patterns were more customary (as called by Tang and Chen (2014)); that is two say, they were more generic such as the use of positive nouns with high intensity. Other patterns, however, were more fixed (non-customary patterns) as they were short phrases such as 'it's ok to be worse'. Results showed that there were differences in the frequency of use of the patterns in the Plurk and Yahoo blogs. Additionally, Tang and Chen (2014) report that hyperbole was frequently present.

Apart from corpus construction and pattern mining, building and improving automatic irony detection systems have been a popular approach. For instance, Tsur et al. (2010) developed a Semi-supervised Algorithm for Sarcasm Identification (SASI). The algorithm used feature vectors derived from syntactic and pattern-based features (punctuation, patterns, patterns+puctuation, enrich punctuation, enrich patterns). The algorithm was tested with English Amazon reviews. A number of experiments were conducted in which the patterns included in SASI also studied individually. In the end, the best F1-score was obtained by SASI (i.e. the combination of all features) with an F1-score of 0.827, closely followed by the patterns+punctuation pattern (F1-score of 0.812).

In another study, González-Ibáñez et al. (2011) explored how accurately could systems distinguish between the different polarities in English tweets: positive, negative, and sarcasm (in which the polarity of the utterance and the polarity of the intended meaning are inverted). The performance of two classifiers were examined: a logistic regression and a Support Vector Machine (SVM) with sequential minimal optimization. Three features were used: unigrams together with the presence and frequency of lexical and pragmatic factors. The pragmatic factors considered were: positive emoticons, negative emoticons, and ToUser marks (which implies that the analyzed tweet was a response to another one). The two classifiers were tested with five different combinations: sarcastic/positive/negative, sarcastic/non-sarcastic, sarcastic/positive, sarcastic/negative, and positive/negative. The SVM outperformed the logistic regression in all combinations except for positive/negative. Regarding results, the best one was obtained by the positive/negative scenario with an accuracy of 0.7589, and the worst result was the one obtained by the sarcastic/positive/negative case (an accuracy of 0.5722).

Focusing on the distribution of labels in the datasets, Liebrecht et al. (2013) studied sarcasm detection in two contexts: a balanced one (50% sarcastic vs 50% non-sarcastic)

[7] `https://universaldependencies.org/u/dep/parataxis.html`

and an unbalanced one (25% ironic vs 75% non-ironic). Both datasets were composed of Dutch tweets that were tokenized and stripped of punctuation. Using unigrams, bigrams and trigrams as features and a Balanced Winnow classifier, they observed a drop in the area under the curve (AUC) score from the balanced[8] dataset to the unbalanced[9] one. Further analysis revealed that explicit markers (e.g. the word 'sarcasme'), intensifiers and exclamations could be good indicators of sarcasm detection.

Additionally, the first ones to tackle sarcasm detection in Czech were Ptácek et al. (2014), although English was also considered. They opted for a supervised learning comparing two classifiers: Maximum Entropy, and SVM. Moreover, language-independent features were considered, for instance n-grams, patterns with high frequency words, Part of Speech (POS) characteristics, emoticons or punctuation. The best result for Czech was obtained with the SVM (F1-score of 58,2) whereas for English the Maximum Entropy worked better (an F1-score of 0.94 for the balanced dataset, and an F1-score of 0.92 for the imbalanced dataset).

In an attempt to analyze which irony markers had an impact on irony detection in Tweets and Reddit posts, Ghosh and Muresan (2018) built a binary classifier (ironic vs non-ironic) with a SVM architecture. The irony markers studied were tropes (e.g. hyperbole), morpho-syntactic markers, and typographic markers. Focusing first on tweets, the system obtained an F1-score of 0.7175 for the irony class when all the irony markers were taken into account. The biggest drop when detecting irony came when tropes were eliminated (F1-score of 0.5618), followed by the typographic markers (F1-score of 0. 6605) such as emoticons, emojis, punctuation marks, etc. On the other hand, results for irony class detection in Reddit were, overall, worse than the ones obtained with tweets. Taking into account all the markers, the model obtained an F1-score of 0.58354 for the irony class. The biggest drop in the score came when morpho-syntactic features were discarded (F1-score of 0.5349), and it seemed to slightly improve when tropes were not taken into account (F1-score of 0.5908).

On a more multilingual approach, Swami et al. (2018) created a corpus for sarcasm detection in English-Hindi code-mixed tweets. In other words, tweets in which a writer makes use of English and Hindi at the same time, thus both languages are present in a single sentence[10]. Three approaches were suggested: Random Forest (RF), a linear SVM, and a SVM with a Radial Basis Function kernel. The best score (F1-score of 0.78) was obtained by the RF classifier that used all of the features taken into account, which include: character n-grams, word n-grams, sarcasm indicative tokens, and emoticons.

In the case of Ghanem et al. (2020), they opted for a crosslingual (or zero-shot) approach. In other words, systems were trained in a language and tested on a different one. Three languages were taken into account: Arabic, English and French; and all datasets were composed of tweets. Each dataset was first trained and tested in the same language (monolingual setting) to confirm that the results were according to the state-of-the-art

[8]Irony AUC = 0.79, Non-ironic AUC=0.77

[9]Irony AUC = 0.75, Non-ironic AUC=0.74

[10]While some linguists consider code-switching and code-mixing the same, Swami et al. (2018) specify that "code-switching is generally inter-sentential while code-mixing is intra-sentential"

ones. Then, each pair was tested in the two directions (e.g. English to Arab, and Arab to English), thus creating six scenarios. An additional pair was tested (indoeuropean/non-indoeuropean), in which the English and Italian datasets were joined. In terms of computational approaches, two were taken: a feature-based model with a RF classifier; and a Convolutional Neural Network (CNN). In half of the scenarios, the deep learning approach seemed to be more suitable; whereas in the other half, the feature-based model appeared to work better. The best score (an F1-score of 0.74) of all was achieved by Arab-to-(English+French) pair with the RF classifier; whereas the worst score belonged to the Arab-to-English pair with an F1-score of 0.50.

Moreover, the interest of irony and sarcasm detection has also been boosted through shared-tasks. For instance, at SemEval2018, Hee et al. (2018) presented a shared task with two subtasks: a) a binary detection of the presence of irony (ironic/non-ironic), and b) a multiclass detection that took the type of irony into account. Four categories were employed in the multiclass subtask. First, the ironic tweets in which the polarity of the message was inverted were labeled under 'verbal irony by means of polarity contrast'; those tweets in which the polarity was not inverted but were ironic nonetheless were considered to be 'other verbal irony'; and, lastly, the third type of irony considered was 'situational irony'. Results showed that, overall, systems performed better at the binary task. Also, the best detected category in the multiclass task was polarity contrast. The best official model for the binary class obtained an F1-score of 0.705 and its architecture consisted of a Long-Short Term Memory (LSTM) that made used of word embeddings as well as sentiment and syntactic features (Wu et al., 2018). Regarding the multiclass subclass, the winner system achieved an F1-score of 0.577 with a model that consisted of a siamese structure with word embeddings (Ghosh and Veale, 2018). The category that this system detected best was non-ironic with an F1-score of 0.843, followed by irony by polarity contrast (F1-score of 0.697).

In the same vein, (Cignarella et al., 2018b) presented a shared task in which they proposed a binary detection subtask as well as a multiclass one. For the multiclass detection task, a special focus was given to sarcasm, which was considered a type of irony. Three labels were considered in the multiclass subtask: 1) sarcasm, 2) irony not categorized as sarcasm, and 3) non-ironic. There was a noticeable drop in the results of the multiclass task compared to the ones in the binary task. The best results for the binary class (F1-score of 0.731) were obtained by a Biditectional LSTM (BiLSTM) architecture that made use of word embeddings and sentiment polarity lexicons (Cimino et al., 2018). Regarding the multiclass subtask, the best system consisted of a concatenation of SVMs that also made use of sentiment polarity lexicons (Santilli et al., 2018). This system scored a F1-macro of 0.520, in which the best class detected was the non-ironic one (F1-score of 0.668) followed by the ironic class (F1-score of 0.447) and sarcasm class (F1-score of 0.446).

Similarly, Ghanem et al. (2019) organized an irony detection shared tasks on Arabic tweets. The corpus was composed of 5,030 tweets out of which 2,614 were ironic and 2,416 were non-ironic. The tweets were annotated by two Arabic native speakers with a Kappa score of 76% (i.e. strong agreement). The best score was obtained by YOLO (Khalifa and Hussein, 2019) with an F1-score of 0.844. The system consisted of an ensemble model of

three classifiers (Gradient Boosting, RF, and Multilayer Perceptron).

Ghosh et al. (2020) organized a sarcasm detection task in which datasets were not only composed of the (micro)blog to be analyzed but also the two previous ones, thus creating some kind of context. The data was extracted from Reddit Corpus (Khodak et al., 2018) as well as from Twitter by using hashtags such as #sarcasm or #sarcastic. The same system won both the Reddit and the Twitter track with an F1-score of 0.834 and 0.931, respectively. The model consisted of a combination of a BiLSTM, NeXtVLAD and BERT (Lee et al., 2020).

Lastly, Corrêa et al. (2021) presented a shared task at IberLEF 2021 with the aim of boosting research on irony in Portuguese. The corpus provided was composed of tweets and news articles, and the task consisted of detecting the presence of irony (ironic/not-ironic) in both domains. The dataset was annotated by three people and their inter-annotator agreement was considered 'fair' in Kappa's value. Competitors were evaluated in terms of Balanced Accuracy (Bacc). Regarding news, the best score (Bacc of 0.52) was achieved by a BERT model (Jiang et al., 2021). On the other hand, a SVM model that made use of superficial features (such as NE, emojis, hashtags..) obtained the best results for tweets with a Bacc of 0.92 (Anchiêta et al., 2021).

So far, this section has reviewed some of the research done in the area of irony (and sarcasm) detection. Having discussed irony from a theoretical and a computational perspective, the final section of this chapter will introduce the research done so far in irony detection for Spanish.

2.3 Automatic Irony Detection in Spanish

Most of the research in irony and sarcasm detection has been conducted on English, although research has been done for other languages such as Arabic (Ghanem et al., 2019), Chinese (Tang and Chen, 2014), Czech (Ptácek et al., 2014), Dutch (Liebrecht et al., 2013), Hindi (Swami et al., 2018), Italian (Cignarella et al., 2018b; Cignarella et al., 2020), or Portuguese (Carvalho et al., 2009; Corrêa et al., 2021).

Regarding irony detection in Spanish, López and Ruiz (2016) aimed to establish some baselines for irony detection in Spanish tweets. The experiments were conducted at character and word level in three different scenarios: a balanced dataset; a 70% non-ironic and 30% ironic dataset; and a 90% non-ironic and 10% ironic dataset. Two classifiers were employed: RF and SVM. Regarding features, at a word level two approaches were taken (n-grams and *word2vec*) whereas at a character level only n-grams were taken into account. At word level, the system that acquired better results was the *word2vec* SVM. This system scored an F1-score of 0.78 in the balanced setting and an F1-score of 0.61 in the 70-30 unbalanced setting[11]. Focusing on to the character level features, in the balanced setting, the RF classifier achieved the best results (F1-score of 0.87). In the 70-30 unbalanced setting both classifiers achieved the same score (F1-score of 0.80), while in the 90-10 unbalanced setting the SVM obtained better results (F1-score of 0.74). In conclusion, character-based

[11]There are no results available for the 90-10 unbalanced setting at a word level.

systems were more consistent than word-based systems, even as the datasets got more unbalanced. Additionally, López and Ruiz (2016) found that emojis and smileys together with expressions of laughter were present among the most common features.

Later, Ortega-Bueno et al. (2019) presented the first shared task on irony detection in Spanish Variants (IroSvA). Three Spanish variants were taken into account: Spanish, Mexican and Cuban. Different datasets were released for each variant, although they were all prepared to perform a binary classification (irony/not-irony) and followed the same distribution: 1/3 of ironic texts, and 2/3 of non-ironic texts. The Spanish and Mexican datasets were composed of tweets, while the Cuban one was built with news comments. The texts extracted for the datasets were related to topics that had been polemical in their respective countries. Two types of analysis were computed: intra-variant and cross-variant. The best intra-variant scores for Spanish and Mexican (F1-scores of 0.7167 and 0.6803, respectively) were scored by a model based on Transformer Encoders and Spanish Twitter embeddings (González et al., 2019b). Regarding Cuban, the best model obtained an F1-score of 0.6596. This model consisted of a SVM that made use of the concatenated representations obtained from words embeddings, an LSTM, and n-grams (Miranda-Belmonte and López-Monroy, 2019). In terms of cross-variant results, the numbers dropped. Looking at the best scores for each pair, the pair that seemed to work best was the Cuban-to-Spain pair with an F1-score of 0.6106; whereas the lowest score came from the Spain-to-Cuban pair with an F1-score of 0.5225.

This chapter introducing irony from a linguistic point of view by assessing the difficulty of providing a definition of irony; comparing irony with sarcasm, hyperbole, and metaphor; and explaining the importance of domain in the realization of irony. Afterwards, we presented previous work on irony detection, task that has been done mostly addressed in English, although some works have started to appear in other languages. In the case of Spanish, we have seen that while there have been some approaches to irony detection, none has addressed the task of irony type classification for this specific language. Similarly, Spanish, so far, has not been taken into account in multilingual and crosslingual irony detection contexts.

3 Materials and Methods

This project is composed of an annotation process and some experiments (which will be explained in detail in Chapters 4 and 6, respectively). The purpose of this section is to introduce the materials and sources employed for those tasks. To this aim, first, the datasets used will be presented. Afterwards, the systems use to perform the experiments will be introduced. Lastly, we will describe the evaluation metrics used to assess the performance of the systems.

3.1 Datasets

Throughout this project, three datasets were used. While the Spanish one was used for annotation work as well as experiments, the English one and the Italian one were employed only during the experimental procedures.

IroSvA dataset (Spanish)

At IberLEF 2019, Ortega-Bueno et al. presented a shared task on irony detection in Spanish Variants (IroSvA). Three variants were taken into account: Spain Spanish, Mexican and Cuban. Different datasets were released for each of the varieties, although there were all prepared to perform the same task: binary detection of the presence of irony. In other words, detecting whether a given text contained irony or not. Moreover, in order to give the texts some kind of context, they were related to topics that caused some kind of controversy. The texts were manually labeled by two annotators and the Kappa Score was not computed as they only took into account those in which both annotators agreed on the classification. Annotators of each dataset were native speakers of their respective Spanish variant and they all used their own concept of irony (no standard guidelines were provided). Moreover, no distinction was made between the different types of irony (including sarcasm). Regarding licenses, the use of these datasets is limited to research purposes.

Focusing on the Spain Spanish dataset (the one that was further annotated in this project), the corpus was composed of tweets regarding ten topics. Out of those topics, 8 were somehow related to political issues or icons, another one was about flat-earthers ('Tierraplanistas') and another one about a reality show's episode ('Venacenar'). Table 1 displays the tweet distribution per topic in train and test datasets.

12 teams took part in the shared task and the best system for Spain Spanish was ELiRF-UPV with an F1-score of 0.7167. This model made use of the encoder part of a Transformer Model as well as embeddings (González et al., 2019a). In terms of qualitative results, further analysis revealed that systems detected irony best within the topic of 'Relator' while worst on the topic of 'Venacenar'. Additionally, a crossvariant evaluation was conducted. Overall, the performance of the systems slightly dropped to an F1-score between 0.5225 and 0.6106.

Topic	Train dataset		Test dataset	
	Ironic	Not-ironic	Ironic	Not-ironic
Tardà	32	240	8	64
Relator	112	75	19	15
Librosánchez	162	90	19	15
Franco	52	240	10	86
Grezzi	54	182	20	36
SemáforosA5	48	215	12	54
Tierraplanistas	86	191	31	40
Venacenar	91	113	19	29
Yoconalbert	55	150	12	38
PañalesIglesias	108	104	50	26
Total	800	1600	200	400

Table 1: IroSvA dataset distribution (Ortega-Bueno et al., 2019)

SemEval 2018, Task 3 (English)

In 2018, Hee et al. organized a shared task at SemEval regarding the detection of irony in English tweets. There were two tasks: a binary one that aimed to detect the presence of irony or the lack of it; and a multi-class one that tried not only to detect irony's presence but also the type. Four categories were taken into account in the second task:

1. Verbal irony by means of a polarity contrast: instances in which the polarity of the utterance is contrary to the one intended. For example, "I love waking up with migraines" (Hee et al., 2018)

2. Other verbal irony: instances in which there is no reversal of polarity but the message is ironic anyway. For instance, "Human brains disappear everyday. Some of them have never even appeared." (Hee et al., 2018)

3. Situational irony: instances describing the irony of a situation. E.g: a fireman truck on fire.

4. Non-ironic

Regarding the corpus[12], there were a total of 4,618 tweets, out of which 2,396 were ironic and 2,222 were non-ironic. Regarding the distribution for the multi-class task, the predominant class was non-ironic followed by irony by clash, situational irony and other types of verbal irony. The ironic tweets were extracted by using auto-tagged hashtags such as #irony, #sarcasm and #not. The tweets were manually labeled by three linguistics students whose second language was English. The three of them were provided with Hee et al. (2016) guidelines, in which irony is stated to arise "from a clash between two

[12] `https://github.com/Cyvhee/SemEval2018-Task3`

evaluation polarities" (Hee et al., 2016, pg. 6). The inter-annotator agreement study was composed of two rounds and, in the last round, the Kappa score obtained was 0.72 for both binary class and multi-class. Regarding the license of the corpus, it is limited to academic purposes.

43 groups took part in the task. In the binary task, the best system managed to get an F1-score of 0.705. The architecture of the system was an LSTM that made use of pre-trained word embeddings as well as sentiment and syntactic features. Regarding the multiclass detection, the best system got an F1-score of 0.507 and it was developed with a siamese architecture that made use of word embeddings.

IronITA (Italian)

IronITA shared task was presented at EVALITA 2018 by Cignarella et al. (2018b). There were two tasks: a) to detect the presence of irony, and b) to detect the presence of sarcasm (as a type of irony). The corpus provided was built with Italian tweets from other databases such as Hate Speech Corpus (Sanguinetti et al., 2018) and the TWITTIRÒ (Cignarella et al., 2018a). The presence of irony or the absence of it was already annotated in the source datasets. Additionally, four Italian native speakers annotated the presence of sarcasm following some provided guidelines. These guidelines defined sarcasm as "a kind of sharp, explicit and sometimes aggressive irony, aimed at hitting a specific target to hurt or criticize without excluding the possibility of having fun" (Cignarella et al., 2018b, pg. 3). The guidelines also highlighted the need of three factors: a target, a harmful intention, and negativity. The annotators obtained a kappa score within the range of moderate. The whole corpus is composed of 2,390 non-ironic tweets and 2,390 ironic tweets (out of which 1,289 were sarcastic). The corpus counts with a Creative Commons Non-Commercial copyright.

7 teams took part in the shared tasks. For the detection of irony, the best system achieved a F1-score of 0.731. The architecture of this system was a LSTM that took into account some features like n-grams, word-embeddings and affective lexicons. Regarding the sarcasm detection task, the best system obtained a F1-score of 0.520. This system's architecture was a SVM that made use of features like word-embeddings as well as affective lexicons.

Having introduced three of the four datasets employed during this project (the remaining derives from the annotation work, see Section 4), we will now look at the transformers that were used during the experimental phase.

3.2 Transformers

The state of art approaches seem to have moved from machine learning techniques to deep learning ones. In particular, transformers have become widely employed tools in the NLP area. Transformers are self-attention based deep learning architectures that work, mainly, with linear data such as text or speech. Figure 1 displays the architecture of a transformer. In this project, transformers were employed to train classifiers that a) would

determine whether a given tweet was ironic or not, and b) would assign an irony category to ironic tweets. Two types of transformers were used: monolingual and multilingual. In other words, transformers that were pre-trained for one language only (monolingual) and transformers that were pre-trained using several languages (multilingual).

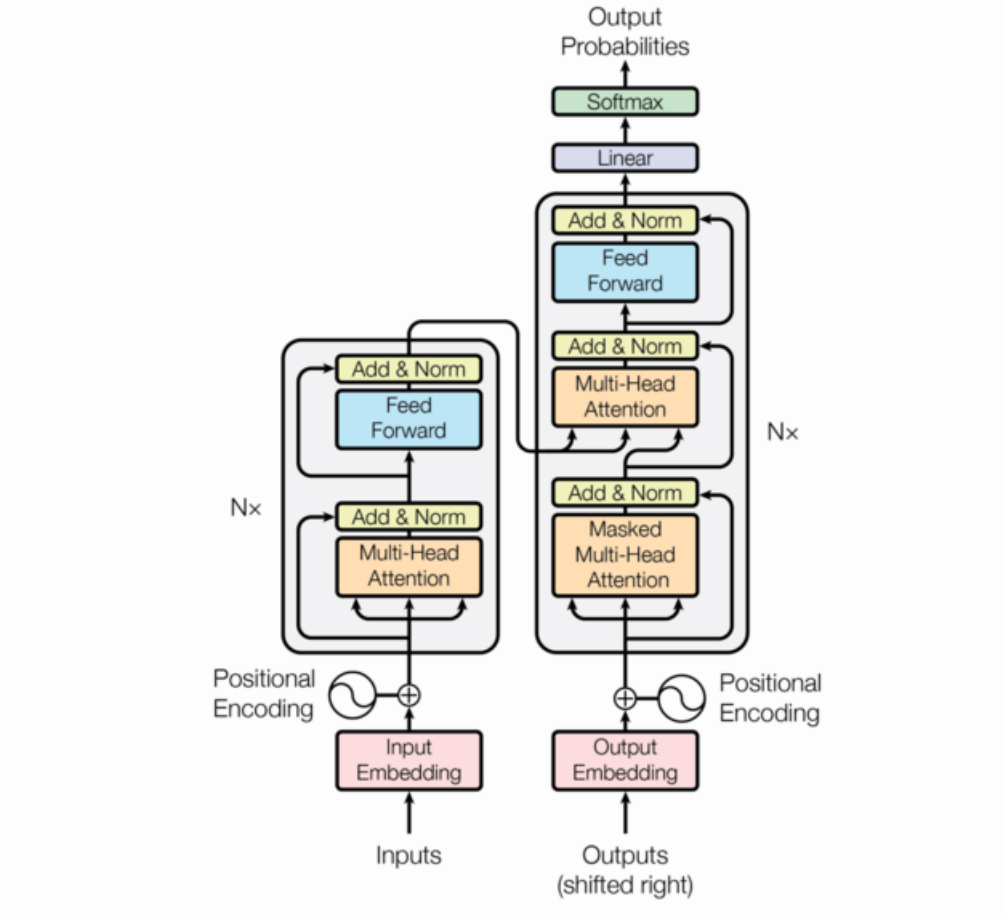

Figure 1: Transformer architecture

To begin with, three monolingual transformers were used during the experiments, one for each language. BETO[13] is a monolingual transformer for Spanish that has a BERT-type of architecture (Cañete et al., 2020). To train this model, they used data extracted from Wikipedia as well as the sources include in OPUS Project (Tjong Kim Sang, 2002), which include TED Talks, United Nations and Government journals, Subtitles… Focusing on the features of BETO, it was trained using Dynamic Masking as well as Whole Word Masking techniques with around 3 billion words. In particular, the BETO-cased version was employed in this project. Focusing on to English, the monolingual RoBERTa-base[14] (Liu et al., 2019) was used. This case-sensitive transformer was trained by applying Masked

[13] `https://huggingface.co/dccuchile/bert-base-spanish-wwm-cased`
[14] `https://huggingface.co/roberta-base`

Language Model and Next Sentence Prediction techniques. A total of 160GB of text was used, extracted from different corpora such as BookCorpus or CC-News. Lastly, UmBERTo Commoncrawl Cased[15] (Parisi et al., 2020) was employed for Italian. This transformer has a RoBERTa-type architecture and was trained using SentencePiece and Whole Word Masking techniques. Moreover, it was trained with 70GB of plain text extracted from a OSCAR[16] subcorpus.

	BETO	RoBERTa-base	UmBERTo
Language	Spanish	English	Italian
Corpora	TED talks, political journals, subtitles, etc.	BookCorpus, news, etc.	CommonCrawl
Size	3 billion words	160GB	70 GB
Type	BERT	RoBERTa	BERT
Features	Dynamic Masking, Whole Word Masking	Masked Language Model, Next Sentence Prediction	SentencePiece, Whole Word Masking

Table 2: Monolingual transformers' features

Regarding the multilingual transformer, XLM-RoBERTa[17] (Conneau et al., 2019) was employed. This transformer is the multilingual version of RoBERTa-base, trained with 2.5TB of data in 100 languages, including Spanish, English and Italian. The data size varies according to the language: almost 323GB for English, around 57GB for Spanish, and nearly 32.5GB for Italian.

These transformers were fine-tuned to perform the tasks proposed in this project by playing with four hyper-parameters: epochs, the maximum sequence length, learning rate, and batch size. The epochs indicate the number of full passes of the entire training dataset through the algorithm's training. The learning rate is used to govern the pace at which an algorithm updates the values of a parameter estimate. In other words, the learning rate regulates the weights of our neural network concerning the loss gradient. The batch size indicates the number of samples that will be passed through to the network at one time, and the maximum sequence length indicates the maximum number of tokens that those samples will have. Also, the train sets were randomly divided into train and deviation files using sklearn tool with a test size of 0.2

The details of the proceedings for each of the experimental frameworks have been slightly different. For monolingual experiments, in which systems were trained and tested with the same corpus, 12 classifiers were trained for each language: six with their respective monolingual transformer and the same six with the multilingual transformer (XLM-RoBERTa). The six configurations were obtained by combining two different numbers of epochs (5 or 10), maximum sequence length (128 or 256) and batch size (16 or 32). However, the combinations involving a maximum length of 256 and a batch size of 32 were discarded

[15] `https://huggingface.co/Musixmatch/umberto-commoncrawl-cased-v1`
[16] `https://oscar-corpus.com/`
[17] `https://huggingface.co/xlm-roberta-base`

as they run out of CUDA memory with some transformers such as XLM-RoBERTa. Lastly, the learning rate was set at 2e-5.

For multilingual experiments, all the training sets were joined and then each language was tested with the same system. Since the training set for these experiments was trilingual, the systems were only trained with XLM-RoBERTa. Moreover, due to the bigger size of the training set, the processing time took longer, thus the number of configurations tried was reduced to four. Since within the monolingual XLM-RoBERTa systems that performed best the parameter maximum length 256 was not present, the two of the previous configurations involving that feature were discarded. In short, four systems XLM-RoBERTa systems with a learning rate of 2e-5 and a maximum length of 128 were trained by changing the number of epochs (5 or 10) and the batch size (16 or 32).

For crosslingual experiments, systems were trained in one language and tested in a different one; therefore, systems were only trained with XLM-RoBERTa too. Moreover, only pairs including Spanish were taken into account, which were eight in total (four with Spanish_orig and four with Spanish_strict). Regarding the parameters used, the same combinations applied in the multilingual experiments were employed.

Lastly, the detection of irony categories was only conducted with the Spanish_strict dataset. Similar to the monolingual experiments, these experiments were run with the monolingual transformer (BETO) as well as the multilingual one (XLM-RoBERTa). At first, the experiments were conducted taking into account all the six categories. Since the processing time was considerably shorter and the results poorer, the number of combinations tried increased by not only widening the range of the epochs' number but also by changing the learning rate in a range of 2e-5 to 5e-5. Later, due to the low F1-scores for some classes, a second round of experiments was organized taking only into account the three main categories and the three best combinations of parameters obtained in the previous round for each transformer.

This section has introduced the transformers employed during the experiments of this project. The following section will describe the metrics used to evaluate the performance of the systems we trained.

3.3 Evaluation Metrics

There are several metrics to evaluate a classifier's performance. Since the datasets used in the project are unbalanced, the accuracy was not taken into account, instead the F1-score macro was used as it gives all the classes the same weight. To compute the F1-score macro, three other measures are needed: precision, recall, and F1-score. Precision indicates how reliable the predictions of a given class are and it is computed like this:

$$precision = \frac{TP}{TP + FP}$$

Recall indicates how many of the predicted instances of a given class were actually from that class. It is calculated the following way:

$$recall = \frac{TP}{TP + FN}$$

Lastly, the F1 score measure is defined as the harmonic mean of precision and recall, and it is computed this way:

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

These three measures are employed at a class level. In order to get an overall performance measure, the F1-macro (i.e. the mean of the F1-score of the different classes) was calculated with the following formula, where i is the class index and N the number of classes:

$$F1 = \frac{1}{N} \sum_{i=0}^{N} F1_i$$

Recall indicates how many of the predicted instances of a given class were actually from that class. It is calculated the following way:

$$recall = \frac{TP}{TP + FN}$$

Lastly, the F1 score measure is defined as the harmonic mean of precision and recall, and it is computed this way:

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

These three measures are employed at a class level. In order to get an overall performance measure, the F1-macro (i.e. the mean of the F1-score of the different classes) was calculated with the following formula, where i is the class index and N the number of classes:

$$F1 = \frac{1}{N} \sum_{i=0}^{N} F1_i$$

This chapter has presented the materials employed in this project. To begin with, we presented three datasets that will be used for irony detection experiments. As we have seen, in terms of irony/non-irony, the three of them follow a different distribution of labels. The Spanish dataset consists of 1/3 ironic tweets and 2/3 non-ironic tweets. On the other hand, the English dataset has a few more ironic instances than non-ironic ones. Lastly, the Italian corpus has a balanced distribution. We have also observed that the corpora for English and Italian are further labeled in terms of irony types, although not the same classes were employed. In the case of Spanish, the corpus is only annotated in terms of irony presence, for this reason we decided to further annotated the dataset in terms of categories. The next chapter explains how the annotation work was conducted.

4 Corpus Annotation

In this section the annotation work will be explained. To this aim, first how irony is defined in this project will be clarified. Afterwards, the schemas followed for irony detection annotation and irony categories annotation will be described. Lastly, some of the problematic cases encountered during the annotation work will be presented.

4.1 Defining irony

By employing irony, a speaker tries to convey a message that does not concord with the utterance (the words spoken). The real message gets somehow hidden between the lines and it is up for the listener to decipher it. However, this definition is rather abstract to use it as annotation guideline. As explained in Section 2.1, there is no consensus about what irony consists of or what its limits are. In this project, the perception of Giora (1995) was used as inspiration, in which irony is not necessarily the opposition between the utterance and the implication. As she explains, "irony differs from non-irony in that it makes use of a highly improbable message (conforming to the marked informativeness requirement) to evoke a less marked, more probable interpretation" (Giora, 1995, pg. 245). She introduces three conditions for irony well-formedness. First, the speaker should introduce information about an accessible discourse topic. In other words, the topic should be shared between the speakers. Second, the ironic utterance should flout the graded informativeness requirement which involves each proposition to be "more (or at least not less) informative than the one that precedes it" (Giora, 1995, pg. 244). Note that, in social media, texts are often independent, expressing an opinion or thought to any person who may read it, there is no necessity of contributing a exiting conversation. Thus, the interpretation of this condition was modified slightly: the speaker is trying to inform about something, such as a thought. Lastly, the third condition explains that irony evokes a non-explicit yet perceivable interpretation that somehow contrasts with the utterance.

In order to illustrate these conditions, we can have have a look into Example (8). The first requirement involves having shared information between the speaker and the addressee, which in this case is the huge traffic jams that were formed during the rush hours when traffic lights were installed at the entrance of Madrid through the A5 highway. The second point requires the message to be more informative than the previous one. Since this is an expression of a thought, the tweet does not follow any previous comments, it is rather isolated. However, we can still considerate it to be informative as we learn something about the user (i.e. the person is annoyed because of the A5 street lights). Lastly, even if the words are those of gratitude, the hidden message is rather an opposing one as she is not grateful at all.

(8) Gracias @MADRID por los #semaforosA5 después de levantarme a las 5 de la mañana para ir a trabajar lo que mas me apetece es llegar a mi casa 45 minutos más tarde de los normal!

Thanks @MADRID for the #A5streetlights after waking up at 5AM to go to work the most desirable thing is to arrive home 45 minutes later than usual!

Having defined irony, the following section will explain how the annotation work was conducted and the criteria taken into account.

4.2 Annotation Schema

The tweets were annotated in two different ways, according to the task they were prepared for. First, the presence of irony in a tweet (or the lack of it) was labeled. Then, within the ironic tweets, the irony categories were annotated according to the way irony was triggered.

4.2.1 Irony Detection

The original IroSvA dataset was already annotated to perform irony detection (i.e. telling apart ironic and non-ironic tweets). According to Ortega-Bueno et al. (2019), the IroSvA annotators were not given any guideline, they were free to use their own concept of irony. A look into the original annotations revealed a clash between the irony perception of this project and the original one. For instance, Example (9) is a tweet that was annotated as ironic by the IroSvA annotators but that was considered non-ironic in this project. The tweet is a joke or word-play triggered by the surname of the politician Joan Tardà[18]. Although this tweet was considered ironic by the IroSvA annotators, it does not follow the guidelines established in this project. In particular, this tweet clashes with the second requirement of irony well-formedness that requires the utterance to be informative. The joke has a clear intention of making the reader laugh; however, when we ask ourselves what do we learn from the speaker with that comment, we find nothing. Does the speaker support or like this politician? Do they agree with their political stance? Are they even invested in politics? We do not know as there is no message, thus no information.

(9) Joan Tardà, Joan Prontà

In the same vein, Examples (10) and (11) were also considered ironic by the IroSvA annotators, yet they do not suit the criteria of irony for this project. These tweets flout the third condition of irony well-formedness that claims the appearance of an unmarked interpretation that contrasts with the words spoken. In these tweets, the is no such contrasting implicature evoked, but rather a non-ironic metaphor that focuses on the similarity between the items ('encipledia' and 'big') or a non-ironic hyperbole (you do not get embolisms because of anger).

(10) Tiene el ego tan Grande que dudo que le quepa en un libro,va a necesitar una enciclopedia.
His ego is so big, I doubt it will fit in a book, he'll need an encyclopedia.

[18]'Tarda' as in 'tarde' (*late*). Its antonym is 'pronto' (*early*).

(11) Entre el documental de la tierra plana y el día que he tenido hoy en el trabajo de verdad que voy a acabar con una embolia fruto del enfado.

Due to the discordances in irony perception between the original annotations and the idea of irony in this project, a re-annotation work of the ironic tweets was conducted. Three scenarios were encountered: 1) tweets that were also considered ironic, 2) tweets that were moved into the non-ironic class, and 3) tweets that were rather ambiguous and required more information about the speaker or the context to make a final decision. Texts belonging to the latter scenario were labeled under the label NV (which stands for 'no valorable'; i.e. can not be assessed). Examples (12) and (13) illustrate some of the NV cases encountered. In the first example (Example (12)), the presence or not of irony would depend on wether the speaker is a flateather themselves or not. If they were, then the message would be literal; on the contrary, if they were not, this tweet would be a form of ironic mockery, trying to sound like one. Similarly, in Example (13), the speaker could be asking a genuine question if they did not know what the program is about or the people that appear in it. However, if the speaker was already familiar with all those details about the program, then the question would have been ironic hinting something like "it is too obvious there is a script in this reality show".

(12) Por supuesto que la Tierra es Plana y no Gira. Punto.
Of course, Earth is flat and it does not spin. Period.

(13) #venacenar289 Esta gente son actores, verdad?
#venacenar289 These people are actors, right?

After the annotation work was finished, the corpus was cleaned and preprocessed. In other words, the NV tweets were filtered out of the corpus. Additionally, few multimodal tweets were also extracted out of the corpus. In other words, tweets in which the irony was triggered by something else that was not text and emojis (e.g. pictures, articles of a magazine, gifs...) were discarded. The final numbers of the corpus are available in Table 3.

A quick scan into the numbers of the re-annotated corpus (hereafter, Spanish_strict to avoid confusions) and the original corpus (from now on Spanish_orig) reveals that one of the most affected categories by the re-annotation work was 'Venacenar'. The topic 'Venacenar' counted with a total of 110 ironic tweets in the Spanish_orig dataset, which were reduced to 44 in the Spanish_strict. The numbers within this topic were affected mostly because of the lack of context. This topic is composed by tweets about a program episode that was aired in TV. Therefore, many tweets were 'live' tweets, that is, you needed to be watching the program at the same time as the speaker did in order to fully understand the utterance. Coincidentally, 'Venacenar' was reported to be the most problematic category when detecting irony (Ortega-Bueno et al., 2019).

Another topic greatly affected by the re-annotation work was 'Relator' as it went from the 131 ironic tweets in Spanish_orig to the 51 ironic tweets in Spanish_strict. Relator was a figure that tried to ease the conflict between the Spanish government and the Catalan

Topic	Train dataset		Test dataset	
	Ironic	Not-ironic	Ironic	Not-ironic
Tardà	14	252	7	65
Relator	41	124	10	19
Librosánchez	95	127	12	18
Franco	24	263	4	91
Grezzi	32	194	12	42
SemáforosA5	40	221	13	50
Tierraplanistas	59	203	23	42
Venacenar	37	131	7	34
Yoconalbert	40	155	10	39
PañalesIglesias	91	116	43	30
Total	473	1786	141	430

Table 3: Dataset after re-annotation for irony detection

one. Among these tweets, we can find a lot of jokes regarding the name as it sounded similar to 'Terminator' or some other words. These cases were considered to be non-ironic as there was no message found in the utterance, that is, they clashed with the informativeness requirement. For instance, Examples (14) and (15) are a tweets that were ironic in Spanish_orig but non-ironic in Spanish_strict.

(14) Relator, relator, que es un relator, pues una campaña del sector del automóvil: Seat relator, Ford relator, Audi relator
Relator, relator, what is it a relator, well relator is a automovilistic campaign: Seat relator, Ford relator, Audi relator

(15) He lanzado un relator al aire y no vuelan...
I've thrown a relator to the air and they don't fly...

4.2.2 Irony Category Classification

Together with the irony's presence annotation work, irony types were labeled. The categories were mainly inspired by the work of Karoui et al. (2017), which were also used by Cignarella et al. (2020). However, some changes were done. First, euphemism (or understatement) was put together with hyperbole. In some languages, like English, we can find two separated phenomena: hyperbole for exaggeration, and understatement for the minimization of an event's importance. However, in Spanish, hyperbole tends to be used for both exaggeration and understatement. Moreover, sometimes it is hard to distinguish between them because they may co-occur at different levels. For instance, the tweet in Example (16) uses understatement at a lexical level, but an exaggeration at a semantic level.

(16) Es fácil: vas hasta el límite de la tierra plana, te asomas al borde y lo preguntas a la tortuga que hay debajo, sosteniendo al mundo.
It's easy: go to Earth's limit, lean out and ask the tortoise below, the one holding Earth

Additionally, similar to Hee et al. (2018), situational irony was considered to be a class on itself whereas in Karoui et al.'s (2017) work situational irony was included in the 'other' category. Also, the label 'context shift' was not taken into account. This class was used for text with 'a sudden change of topic/frame'; however, tweets tend to be rather informal and the ones gathered in IroSvA, in particular, had an already established topic, so it was not deemed a relevant class for this project. At the end, six different categories were taken into account when analyzing the realization of irony in discourse. It should be mentioned that categories are not mutually exclusive; however, only the one that was considered predominant was annotated.

1. Hyperbole: the exaggeration or understatement of something, often used to create a ridicule situation. The realization of this category can be more pragmatic (like Example (17)) or more language-related (such as Example (18)).

   (17) A Joan Ribó: Envío de Grezzi de vuelta a Italia vía @change_es
   To Joan Ribó: Grezzi shipment back to Italy via @change_es

   (18) La justicia suspende cautelarmente la exhumación de Franco. SOR PRE SA.
   Justice provisionally suspends Franco's exhumation. SUR PRISE.

2. Rhetorical question: a question that does not really expect an answer and leaves rather clear the stance or opinion of the speaker in a given topic, as we can see in Example (19). The realization of the questions could be both direct or indirect.

   (19) ¿Pan salvaje? ¿Que lo cazan por el monte? La madre que lo parió #VenACenar289
   Wild bread? Why, is it haunted on the mountains? Seriously #VenACenar289

3. Contrast: instances in which the expectations of the reader are somehow shaken as there are elements that do not fit together. There are some patterns in this category that tend to repeat like giving thanks for something inconvenient or damaging. For instance, (20) was aanotated as contrast because of the opposition of the elements #WithAlbert (a hashtag intended to support the politician Albert Rivera's application for presidency) and the topic of getting drugs (which is socially negatively viewed and brings up the popular belief that he used to sniff).

   (20) #YoConAlbert iria a pillar tema... Fijo que le Pasan buena farla
   #WithAlbert I would go to get some drugs... They probably pass him some good coke

(21) Si os gusta la literatura y la ciencia ficción hoy tengo dos recomendaciones: - El libro de @sanchezcastejon - Los Presupuestos Generales del Estado.
If you are into literature and science fictione, I've got two suggestions: - The book of @sanchezcastejon - The General State Budget.

(22) Ah, nada como respirar el aire puro de un atasco de kilómetros gracias a los #semaforosA5. Otro gran avance a favor del medio ambiente gracias a Carmena. Esto sólo lo mejora poniendo barreras de neumáticos ardiendo en las entradas de Madrid cada mañana.
Ah, nothing like breathing the pure air of a huge traffic jam thanks to the #A5streetlights. Another great measure in favor of the environment thanks to Carmena. This can only be improved by putting barriers of tires in fire every morning at the entrance of Madrid.

4. Comparison: two different items are compared in such a way that one of the elements is ridiculed or given a particular connotation. For instance:

(23) Gente en 1900: En el año 2000 habrá coches voladores Gente en 2019: La tierra es plana
People in 1900: in the year 2000 there will be flying cars People in 2019: Earth is flat.

(24) Un libro muy Maduro[19]

5. False assertion: There is something in the tone that makes the reader understand that the sentence intends to say something else than the utterance per se. However, there is nothing that can be pinpointed.

(25) Joan Tardà responderá a Vox en catalán....si, si
Joan Tardà will answer to Vox in Catalan....yes, yes

6. Situational irony: A situation that turns to be ironical as in Example (26).

(26) Miro el hashtag #YoConAlbert y no hay nadie con Albert
I look the hashtag #I'mWithAbelt and there is no one with Albert

Figure 2 displays the distribution of the categories in the corpus. As it can be observed, the data is unbalanced as almost half of the instances are hyperbolic. Moreover, only three examples of situational irony can be found in the Spanish_strict corpus. In view of these numbers, we could already tell that, most likely, the irony type detectors would a) struggle with situational irony and false assertion, and b) favor the hyperbole class.

[19]Untranslatable tweet due to the joke inside. Maduro means 'mature' but, since its written with capital letter, it refers to the Venezuelan ex-president Nicolás Maduro.

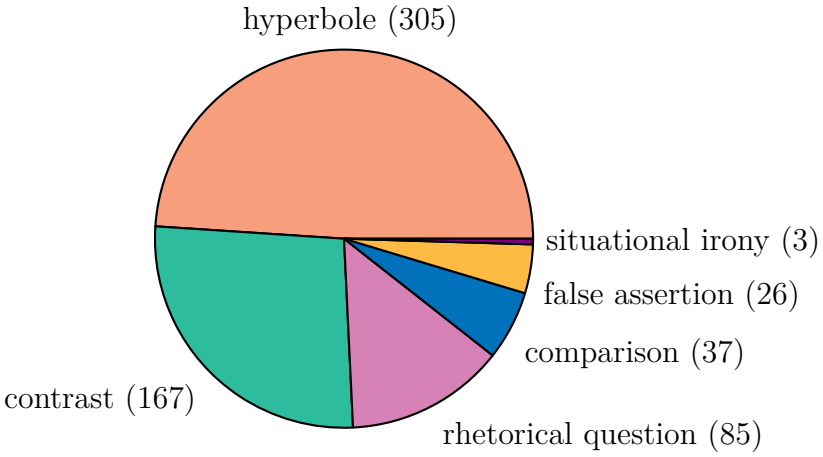

Figure 2: Category distribution in the corpus (Spanish_strict)

4.3 Annotation Problems

Data annotation is not an easy task, especially with abstract topics such as irony. In this subsection we review some of the struggles when annotating tweets for irony detection and irony type detection. Lets first focus on the annotation work for irony detection. Since the guidelines followed contained three main points (the conditions of irony well-formedness), it follows that the problems encountered could have three origins. To begin with, the first condition established the need of a shared topic. The tweets in the set were extracted from six different topics, therefore the theme of the messages was always clear to certain extent. However, sometimes a deeper understanding of the topic was needed to correctly decode the implicature. For instance, to fully understand Example (27) one needs to know that El Valle de los Caídos is a monument built under dictator Franco's order to honor his fallen supporters in the revolt, and that this monument was constructed by the political prisoners. Similarly, in order to correctly decode Example (28), the reader must know that *The Office* is an American Sitcom with absurd situations. Therefore, the speaker in this tweet is trying to convey that, while watching the documentary about the flat Earth, they laughed or wonder what was going on. This way, we can understand that the user is not a flatearther. It should be taken into account that the references in these examples were explicit as they are mentioned in the tweet. In these cases, a little research helped making a decision. Nonetheless, more subtle references can also be found, and some of those non-marked references may have gone undetected which could have led to an annotation error.

(27) El juez dice que puede ser peligroso para los operarios la exhumacion de Franco. No dice nada de lo peligroso que fue para los que construyeron el valle de los caidos
The judge says that Franco's exhumation can put the workers at risk. He does not say anything about the dangers suffered of those who built it.

(28) Me estoy viendo la la tierra es plana" y estoy como cuando vi el primer capítulo de The Office
I'm watching Earth is flat and I feel similar to when I saw the first episode of The Office.

Focusing on to the second requirement of irony well-formedness, the main problem with this point was the lack of continuation of conversations, as explained before. Since tweets do not necessarily follow previous conversations, they may not be able to be more informative than the previous one, as required by Giora (1995). Because of that reason the guidelines were adapted to 'a text that contains a message, not only an intention'.

Lastly, the third requirement was found to be the most problematic one. As mentioned previously, the third condition establishes the presence of an implicature that somehow contrasts with the utterance. The contrast, however, does not need to be opposition as it can also be a change of intensity degree. The hardships of this point are: identifying the elements being contrasted and identifying the relationship between those elements. Lets take, for instance, Example (29). In this tweet, the first and second requirements of irony well-formedness are covered as we have a topic (flatearthers) and a message (the speaker's point of view). The remaining point to be analyzed is the last one: the relationship between the utterance and the implicature. However, one of the issues is the presence of several implicatures, mainly: 'they only say such things because of the show' or 'I do not take seriously what these people say'. If we take the former, the connection between the utterance and the implicature seems to be closer to similarity rather than contrast as we could argue that the utterance and the first implicature could be synonyms. In other words, the first implicature would be a less exaggerated non-ironic reading of the utterance. On the other hand, if we consider the second implicature, the relationship between the words in the tweet and the hinted message is rather one of contrast because of the disagreement in the believes. In this case, the tweet could be considered ironic, if ever so slightly. Due to the difficulty in the label for this tweet, additional opinions were asked for. However, these were also quite divided. Finally, the tweet was labeled as non-ironic since out of the two interpretations the most frequent first interpretation was the first implicature ('they only say such things because of the show').

(29) Coño, si se gana dinero en youtube diciendo que la tierra es plana
Hell, you can make money saying that Earth is flat on Youtube

Regarding the difficulties of annotating irony categories, the main struggle was choosing the predominant category for each tweet as, often, there were several in a single text, sometimes embedded. For instance, Example (30) displays a tweet that was labeled under comparison, although it could have been considered hyperbole too. This sentence, which looks like an assertion, holds two elements being compared and one of them is undeniably wrong/unreal as there are no green chimpanzee. This way, the speaker cancels the assertion of a flat Earth. Nonetheless, the unrealistic element is indeed a hyperbole and without it the power of the irony would probably not be as strong and could be confused with an

actual assertion. So, we could argue that the hyperbole is almost a requirement in this comparison. In the end, the comparison structure was decided to prevail as it is the main structure of the sentence.

(30) Claro. La tierra es plana y lis chimpancés verdes
Of course. Earth is flat and chimpanzees are green

This chapter has discussed the annotation work conducted in this project. First, a definition of irony was provided. Our re-annotation work has evidenced strong differences with respect to the annotation criteria used in the original IroSvA corpus. Furthermore, we have provided an annotation of irony categories and discussed the main difficulties encountered in the annotation process. The next chapters will be leveraging both the original annotations as well as the annotations developed within this thesis in order to experiment with irony detection and irony category classification.

5 Experimental Settings

This section will explain how the transformers introduced in Section 3.2 as well as the datasets introduced in Sections 3.1 and Chapter 4 were employed in the experiments. Two types of classification tasks were proposed. To begin with, we have irony detection tasks that focus on identifying the presence of irony. For these subtasks, the datasets employed were the ones annotated at irony/not-irony level (binary class). These experiments were conducted in three languages: Spanish, English, and Italian. Note that although three languages were taken into account, four datasets were used as there were two Spanish datasets (Spanish_orig and Spanish_strict). Three irony detection subtasks were considered. The three irony detection tasks were the following:

Monolingual Irony Detection Experiments. The first subtask aimed to answer the first research question, which wondered whether a multilingual or a monolingual pretrained transformer would be more suitable for irony detection with a single language or monolingual context. Monolingual transformers have been reported to perform as well as multilingual ones. For instance, de Vargas Feijó and Moreira (2020) trained two monolingual models, BertPT and AlbertPT, which were compared to the multilingual transformer Multilingual BERT (hereafter, mBERT) across 7 different NLP tasks such as Emotion Classification or Semantic Textual Similarity. Results showed that, for most of the tasks, monolingual transformers were able to obtain similar results to mBERT. Similarly, Cañete et al. (2020) trained a Spanish BERT model (BETO) that obtained better results than the best mBERT results at that moment in tasks such NER in Spanish. On this basis, we wondered what kind of transformers would be more suitable for irony detection. To answer this question, the monolingual and multilingual transformers presented in Section 3.2 were tested in a monolingual irony detection task. In other words, systems were fine-tuned and tested in the same language. The results obtained from these experiments were used as baselines for the following irony detection experiments.

Multilingual Experiments. The second set of irony detection experiments tried to answer the question: does multilingual data augmentation improve the results previously obtained by the systems trained in the monolingual experiments? Data augmentation has become a widely used technique, especially with low-resource languages Feng et al. (2021). There are several ways of augmenting data, such as multilingual augmentation. This technique consists of increasing the training data by including data from other languages. This technique has been implemented in other NLP tasks such as Sentence Classification Wang et al. (2020). Although Spanish, Italian and English are not low-resource languages per se, data for irony detection is still limited. Therefore, the purpose of these experiments was to see whether multilingual data augmentation could improve the results obtained in the monolingual experiments. To this aim, a multilingual context was considered. In other words, the training data was increased by joining the Spanish, English and Italian datasets. Then, each language was tested individually. Since there were two Spanish datasets, two

augmented training sets were created: one containing the Spanish_orig, English, and Italian sets; and another one containing the Spanish_rev, English, and Italian training sets.

Crosslingual Irony Detection Experiments. The last type of irony detection experiments sought to research the universality of irony devices. In other words, to see if there are any devices that irony may use that are similar in different languages. If such devices were common across languages, low-resource languages could benefit from them. To this aim, systems were trained in a given language and tested in a different one; in other words, a crosslingual or zero-shot context was considered. In previous work, Ghanem et al. (2020) explored crosslingual irony detection with English, French and Arab. The worst results were obtained by the language pair Arabic to English with an F1-macro of 0.5; on the other hand, the best results were obtained by the English to French pair with an F1-macro of 0.58. Additionally, following a multicultural perspective, Ghanem et al. (2020) joined the Indo-European language sets (English and French), thus creating two other pairs (Indo-European vs non-Indo-European). These new pairs managed to obtain better results: an F1-macro of 0.624 for the (En/Fr) to Arab pair; and an F1-macro of 0.746 for the Arab to (En/Fr) pair. For this project, only Spanish pairs were considered; which is to say, four pairs were tested with each Spanish dataset, eight combinations in total. The pairs are displayed in Table 4.

	Spanish_orig	Spanish_strict
English	English →Spanish_orig Spanish_orig →English	English →Spanish_strict Spanish_strict →English
Italian	Italian →Spanish_orig Spanish_orig →Italian	English →Spanish_strict Spanish_strict →Italian

Table 4: Language pairs for crosslingual experiments

In addition to irony detection tasks, an irony category classification task was proposed. This task aims to distinguish between the irony types listed in 4.2.2. These experiments were only conducted with the Spanish_strict corpus developed in this master thesis. The same procedure used during the monolingual experiments was applied for this task. In other words, experiments were conducted using the monolingual Spanish transformer and the multilingual transformer.

In this chapter, we explained how the experiments were conducted as well as their purpose. In the following chapter, we will present and discuss the results obtained from the experiments.

6 Empirical Results

In this section the results obtained from the experiments described in Chapter 5 will be presented. As previously mentioned, two types of experiments were performed: irony detection to tell apart irony from non-irony, and irony-type classification to distinguish between the different types of categories presented in Section 4.2.2. First, the results obtained from the irony detection experiments in the monolingual, the multilingual, and the crosslingual settings will be reviewed. Lastly, the results for the multiclass detection will be discussed.

6.1 Irony Detection

Irony detection consists of distinguishing tweets that contain irony from those that do not. Three types of settings were considered. First, a monolingual setting that helps establishing some baselines. Then, a multilingual setting that explores the effect of multilingual data augmentation. And, lastly, a crosslingual setting was proposed to investigate the universality of irony structures and devices. All these experiments were conducted with the Spanish_orig, English, and Italian datasets presented in Section 3.1 and the Spanish_strict dataset presented in Section 4.2.1.

6.1.1 Monolingual Setting

The monolingual setting aimed to establish some baselines as well as analyze which kind of transformer (monolingual or multilingual) is more suitable for irony detection and to which extent. First the results for Spanish will be provided, both Spanish_orig and Spanish_strict. Next, English findings will be discussed. Afterwards, Italian results will be presented. Lastly, a summary of the findings will be offered together with an answer for the first research question (*Are monolingual transformers more suited than multilingual transformers for irony detection? If so, to what extent?*).

Spanish

Table 5 displays the results obtained by XLM-RoBERTa and BETO when trained with Spanish_orig and Spanish_strict. Table 5 also includes the results obtained by the winner system (ELiRF-UPV) at the IroSvA shared-task (Ortega-Bueno et al., 2019). As it can be observed, in terms of F1-score, XLM-RoBERTa and BETO obtained a very close result when trained with Spanish_orig; whereas, BETO worked better than XLM-RoBERTa with Spanish_strict. Moreover,BETO appears to have adjusted better to the data change as its F1-score is similar with both datasets, while there is a drop with XLM-RoBERTa when training with Spanish_strict. It can also be observed that both BETO and XLM-RoBERTa improved the non-ironic class detection and worsen the ironic class detection when training with Spanish_strict. This effect is possibly caused by the unbalanced numbers in Spanish_strict. For starters, with the elevated number of non-ironic instances the

system may get biased and favor this class. Additionally, the number of tweets (total and per class) are different in each dataset which may affect the statistical numbers as the errors may have a bigger or lesser impact.

Dataset	Transformer	Non-Ironic	Ironic	F1-macro
Spanish_orig	XLM-RoBERTa	0.8306	0.6650	0.7478
	BETO	0.8416	0.6488	**0.7452**
	ELiRF-UPV			0.7167
Spanish_strict	XLM-RoBERTa	0.8721	0.5267	0.6994
	BETO	0.8821	0.6000	**0.7410**

Table 5: Spanish monolingual results

For a more detailed analysis, the intersection tweets were extracted. In other words, the tweets that had the same truth label in both datasets. Figures 3 and 4 display the confusion matrices for BETO's predictions with Spanish_orig and Spanish_strict, respectively. Similarly, Figures 5 and 6 show the confusion matrices for XLM-RoBERTa's prediction with Spanish_orig and Spanish_strict, respectively.

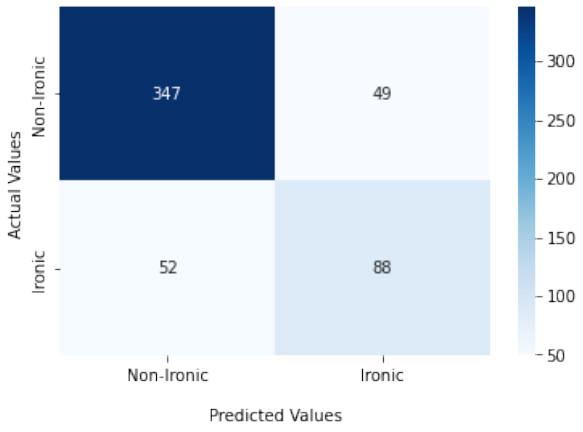

Figure 3: Confusion matrix for BETO with Spanish_orig, monolingual setting

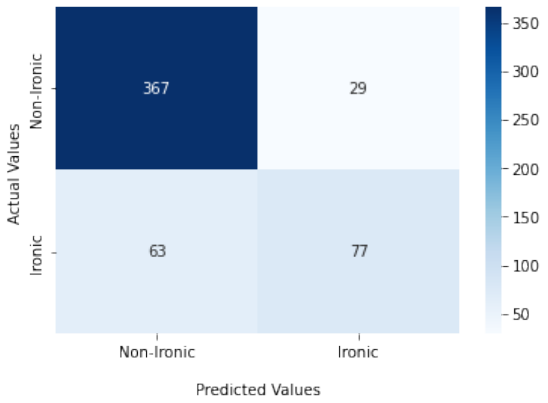

Figure 4: Confusion matrix for BETO with Spanish_strict, monolingual setting

As it can be observed, the number of ironic tweet correctly predicted by BETO and XLM-RoBERTa is higher with Spanish_orig than with Spanish_strict. This follows the results obtained in terms of F1-score (see Table 5). On the other hand, non-ironic tweets appear to be better detected by the systems trained with the Spanish_strict dataset. Focusing on the analysis of the errors, in three of the four models trained, there were more ironic tweets wrongly classified as non-ironic than the other way around. The only exception was the XLM-RoBERTa model trained with Spanish_orig which seems to be more prone to misclasify non-ironic tweets as ironic.

For a better understanding of the errors, a manual examination of the incorrect predictions of ironic tweets was conducted. In other words, ironic tweets that were considered

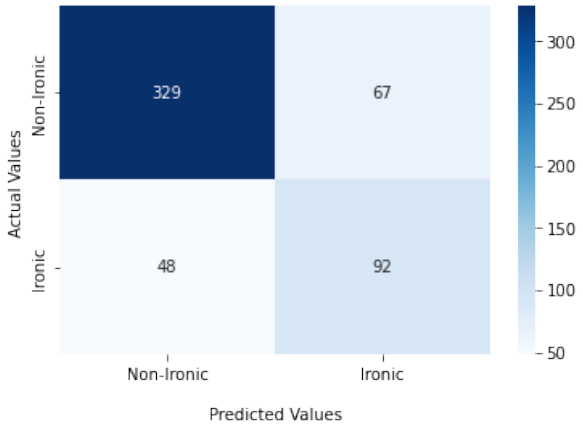

Figure 5: Confusion matrix for XLM-RoBERTa with Spanish_orig, monolingual setting

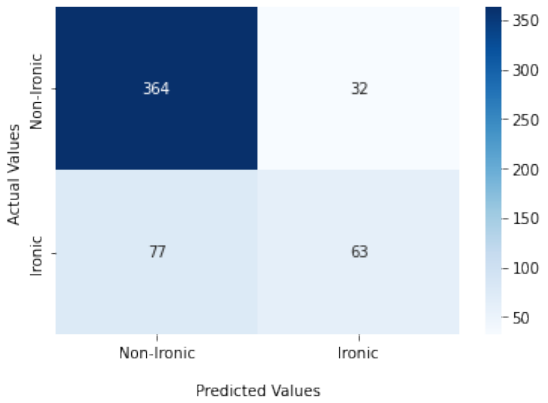

Figure 6: Confusion matrix for XLM-RoBERTa with Spanish_strict, monolingual setting

to be non-ironic by the models. Among those tweets we can find some ironic instances that have no explicit cues such as Example (33). However, there are also instances in which irony is accompanied by cues, and yet are considered to be non-ironic. For instance, Examples (31) and (32) are clearly ironic tweets and contain a rather excessive use of exclamation marks, yet they were labeled as non-ironic. Although no clear pattern was found that could have triggered the mislabeling of ironic tweets, it is worth noting that, in few tweets, there were references to people or days, which could have had an impact on the prediction.

(31) Resultado de la modélica" transacción española un juez franquista decide sobre la exhumación de Franco. Chorpecha!!!!"
The results of the exemplary Spanish Transition, a Francois judge decides about Franco's exhumation. Surprise!!

(32) Que es lo peor que puede pasar con la #exhumacion de Franco, que se caiga la losa y lo remate? Recuerda, los dictadores van al orgánico!!!
What is the worst that can happen with Franco's exhumation, that the gravestone falls and he is finished off? Remember, dictators go organic!!!

(33) He hablado con abogados de Nueva York y me han confirmado que la tierra es plana, cuando uno se equivoca lo mejor es pedir disculpas
I have talked with layers from New York and they have confirmed that Earth is flat. When you make a mistake, it's best to apologize.

Apart from analyzing the erroneous predictions of non-ironic tweets, a look was taken into the non-ironic tweets that were predicted as ironic by the systems. This error could be triggered by different features. First, there seem to be many tweets containing capital

letters, punctuation marks (mostly exclamation and question signs) or/and emojis. For instance, Example (34) shows a non-ironic tweet that was predicted to be ironic that contains words in capital letters and an exaggerated use of exclamation marks. In addition to punctuation marks, non-ironic non-literal tweets (in other words, non-ironic metaphors or hyperboles) appear to also be difficult for the systems to detect as non-ironic. For example, the non-ironic tweet in Example (35) was considered to be ironic, most likely because of the metaphor. Lastly, rude-talk may also affect predictions, making the models believe non-ironic tweets to be ironic, such as Examples (36) or (37).

(34) El cambio climático es una mentira sustentada por falsos dogmas insostenibles!!! Entre ellos la forma de la tierra, ya que nuestro planeta NO es una Bola que gira en el espacio sideral! Es PLANA!
Climate change is a lie supported by unsustainable false dogmas!!! Among them the shape of the earth, since our planet is NOT a ball that rotates in outer space! It's FLAT!

(35) Qué horror, que después de tantos años haya jueces que les gustaría sumergirse en el río para enganchar el salmón para goce del tirano.
What a nightmare, after so many years there are still judges that would dive in the river to catch some salmon for the tyrant's delight.

(36) Cocinar con agua de mar, vamos con la chorrada del mes... #VenACenar289
Cooking with sea water, the goes the stupidity of the month... #VenACenar289

(37) Demasiada gilipollez y pijerío en este menú #VenACenar289
A way too stupid and posh menu #VenACenar289

As previously mentioned, a possible cause of the improvement of non-ironic class with the Spanish_strict dataset could be the elevated number of tweets in this class. For this reason, an additional experiment was conducted in order to see how much of a bias could have caused the elevated number of non-ironic instances, the training dataset of the Spanish_strict was downsampled. Following the original distribution in Spanish_orig, the training set was downsampled so that the corpus was composed of a third of ironic tweets and two thirds of non-ironic ones. Results are displayed in Table 6. As it can be observed, XLM-RoBERTa's performance did not vary much, whereas BETO's results were considerably lower. Moreover, the detection of the ironic class seems to improve slightly with the downsampled dataset with the multilingual transformer (XLM_RoBERTa). However, the best ironic detection score belonged to the non-downsampled BETO system. These results may suggest that the narrowing of the irony concept have had a great impact, greater than the class distribution. Although, having as many ironic instances to train as the original IroSvA would be helpful to confirm this hypothesis.

Transformer	Non-Ironic	Ironic	F1-macro
XLM-RoBERTa	0.8343	0.5413	0.6878
BETO	0.8343	0.5651	**0.6997**

Table 6: Spanish_strict monolingual results, downsampled

English

Table 7 displays the results obtained for English together with the ones achieved by the best model at SemEval2018 shared task (UCDCC) (Hee et al., 2018). As it can be observed, similar to Spanish results, the monolingual transformer (RoBERTa-base) seems to achieve a better score. The classification mistakes seem to be triggered by punctuation marks (such as excessive use of exclamation marks), capital letter and capital letters. However, it should be mentioned that, in some cases such as Examples (38) and (39), the lack of context may also have influenced the misclassification (despite the use of hashtags).

(38) I wonder what triggered the anxiety? #sarcasm

(39) If we win against Poland and Scotland, does that put us 65th in the world rankings. @YouBoysInGreen @FAIreland #Sarcasm

Transformer	Non-Ironic	Ironic	F1-macro
XLM-RoBERTa	0.7244	0.6893	0.7069
RoBERTa-base	0.7910	0.6929	**0.7420**
UCDCC			0.724

Table 7: English monolingual results

Italian

In the Italian case, a slightly bigger difference can be observed between the results of the monolingual transformer from the multilingual one, in general. Also, it is worth noting that systems seem to deal better with ironic tweets than non-ironic ones. Table 8 displays the results of the best XLM-RoBERTa and UmBERTo models as well as the results of the best system of the IronITA shared task (ItaliaNLP) Cignarella et al., 2018b. A look into the predictions revealed that the system tends to classify tweets containing emojis and capital letters as ironic, similar to Spanish models.

Monolingual experiments: summary

Monolingual experiments were conducted to establish some baselines so that the following experiments can be properly analyzed. They also intended to answer the first research question of this project:

Transformer	Non-Ironic	Ironic	F1-macro
XLM-RoBERTa	0.7288	0.7497	0.7393
UmBERTo	0.7534	0.7836	**0.7685**
ItaliaNLP	0.707	0.754	0.731

Table 8: Italian monolingual results

RQ1: Are monolingual transformers more suited than multilingual transformers for irony detection? If so, to what extent?

As we have seen, monolingual transformers seem to be more suitable for irony detection than multilingual ones, or at least equally as good. In most of the cases, monolingual transformers surpased the F1-macro obtained by the multilingual experiments. The only exception being the Spanish monolingual transformer when trained with Spanish_orig, which obtained a similar F1-macro when trained with the multilingual one. Similarly, when detecting ironic tweets, monolingual transformers performed better, except for the case of Spanish_orig, in which the multilingual transformer obtained a higher F1-micro score for the ironic class than the monolingual transformer.

In terms of the possible causes of the errors produced in the predictions, several factors can be named. First, in the same vein as previous researches (for example, Carvalho et al., 2009, López and Ruiz, 2016), capital letters, emojis, and heavy punctuation marks (e.g. an exaggerated use of exclamation marks (!!!)) could be considered to be an indicative of irony. Nonetheless, they are also very frequently used in other contexts such as jokes, expressions of anger, etc. This may be one of the reasons why some non-ironic tweets were mislabeled as ironic. Furthermore, other figures of speech (for example, metaphor or hyperbole), rude-talk and insults may also have an impact on the predictions, leading the systems to believe non-ironic tweets to be ironic. Lastly, some references may also difficult the correct detection of irony.

6.1.2 Multilingual Setting

As previously explained, multilingual experiments aimed to answer the second research question: *Does multilingual data augmentation help improve the results obtained from the monolingual experiments?* This question will be answered at the end of the section, after reviewing the results obtained. Before going any further, it should be mentioned that Table 9 displays the monolingual F1-scores obtained in the previous experiments in order to ease the comparison between the results.

Table 10 displays the results obtained from the multilingual setting that include the Spanish_orig corpus. A drop on the performance can be appreciated in Spanish, compared to the monolingual results. Regarding English and Italian, the systems seem to work better than the monolingual XLM-RoBERTa one, but not as good as the RoBERTa-base and UmBERTo ones (respectively).

Table 11 shows the results obtained for each language when training in multilingual setting with the Spanish_strict dataset. Comparing the results from these experiments with

Language	XLM-RoBERTa	Monolingual transformer
Spanish_orig	0.7478	**0.7452**
Spanish_strict	0.6994	**0.7410**
English	0.7069	**0.7420**
Italian	0.7393	**0.7685**

Table 9: Overview of monolingual irony detection results in terms of F1-macro score

Language	Non-Ironic	Ironic	F1-macro
Spanish	0.8222	0.6059	0.7141
English	0.7620	0.6835	0.7228
Italian	0.7599	0.7515	**0.7557**

Table 10: Multilingual context results, trained with Spanish_orig

the monolingual ones, a clear drop can be appreciated in the F1-score for Spanish_strict. With regard to English, the results seem to be similar to the ones obtained by model trained with the monolingual transformer (RoBERTa-base), which is the best model out of the two models in the monolingual experimental setting. Lastly, the Italian results for these experiments are better than the ones obtained with XLM-RoBERTa in the monolingual setting, but they are still slightly worse than the ones obtained by UmBERTo (the Italian monolingual transformer).

Language	Non-Ironic	Ironic	F1-macro
Spanish	0.8757	0.5150	0.6954
English	0.7838	0.7002	0.7420
Italian	0.7644	0.7488	**0.7566**

Table 11: Multilingual context results, trained with Spanish_strict

Additionally, similar to the Spanish monolingual results, there seems to be a trend by which, a general improvement of the non-ironic class detection and the worsening of the ironic class detection can be appreciated when training with Spanish_strict. Again, this may be due to the difference in the number of examples of each class in each Spanish dataset, which could have affected the models' training. Also, if we compare the results for each language in the two different multilingual trainings (one with Spanish_orig and the other one with Spanish_strict), we can notice that English results for the multilingual Spanish_strict setting are better than the ones obtained for the multilingual Spanish_orig setting. Regarding Italian, very similar results were obtained from both contexts. In terms of Spanish, the results obtained for Spanish_strict are worse than the ones for Spanish_orig.

Lastly, in order to understand why the numbers considerably change from the use of one version to the other, a look into the predictions was taken. It appears that, when Spanish_strict is involved in the training, the system deals slightly better with contrast

(such as Example (40)). However, the English dataset makes use of hashtags, which could have had an impact too.

(40) the cop also had prior misconduct charges. But, ya know, that's ok #sarcasm

Multilingual experiments: summary

Multilingual experiments sought to answer the second research question of this project:

> RQ2: Does multilingual data augmentation help improve the results obtained from the monolingual experiments?

Since the training involved several languages, these experiments were conducted only with the multilingual transformer (XLM-RoBERTa). Data augmentation, in some cases, improved the results obtained from the multilingual transformer; however they were still worse or similar to the ones obtained by the monolingual transformer. Thus, we can conclude that it is not worth training in a multilingual context in these cases since better results can be obtained with the monolingual transformers in a monolingual setting. In other words, with less data, thus quicker processing.

6.1.3 Crosslingual Setting

The crosslingual context, or zero-shot, refers to the experiments in which a system was trained in one language and tested in another one. This experiments intended to answer the third research question: *Are irony cues and structures universal and, thus, transferable across languages?* This question will review the results obtained from these experiments and will answer the research question at the end of the section.

Table 12 displays the results for the crosslingual experiments conducted with the Spanish_orig dataset; while Table 13 shows the results for the models involving Spanish_strict.

Language pair	Non-Ironic	Ironic	F1-macro
Spanish_orig →Italian	0.6170	0.5386	0.5778
Italian →Spanish_orig	0.7885	0.7656	**0.7770**
Spanish_orig →English	0.5497	0.4600	0.5049
English →Spanish_orig	0.7112	0.3315	0.5214

Table 12: Crosslingual context results with Spanish_orig

Several points can be highlighted from the results. First, it looks like pairs involving Spanish and Italian achieve better results than the ones obtained from pairs that include English. This outcome is are rather expected since Italian and Spanish are both Romance languages while English has Germanic roots. In other words, Italian and Spanish are closer in language-related aspects such as syntax, vocabulary or morphology. Additionally, it looks like, regardless of the Spanish dataset, the pairs worked better when models were

Language pair	Non-Ironic	Ironic	F1-macro
Spanish_strict →Italian	0.6297	0.3933	0.5115
Italian →Spanish_strict	0.7629	0.5375	**0.6502**
Spanish_strict →English	0.6279	0.3503	0.4891
English →Spanish_strict	0.7607	0.3221	0.5414

Table 13: Crosslingual context results with Spanish_strict

trained in either English or Italian and tested in Spanish than in the opposite scenario. These results could be related to the use of topics in the Spanish dataset. In order to offer some kind of context, topics were used in the Spanish dataset which could have made tweets more dependable on pragmatics (or knowledge of the world), making the abstraction of structures in the training process more difficult. In terms of compatibility, it seems like Spanish_orig is more similar to the Italian dataset as it obtained better results in both directions than the ones obtained with Spanish_strict. Regarding English, it is hard to tell as in one direction seems to work better with Spanish_orig while the other direction benefits from Spanish_strict. Lastly, the most striking results emerge from the Italian to Spanish_orig pair since the F1-macro of this pair outperform the Spanish monolingual ones (0.7770). Also, the detection of each class seems to be more balanced than in the monolingual results (see Table 5).

A manual analysis of the models' predictions was conducted in order to get a better grasp of the systems' behavior. The aim was to find some patterns as well as to answer two main questions:

1. Focusing on Italian to Spanish pair, there seems to be a huge gap between the results obtained with Spanish_orig and Spanish_strict. Why? Even more, how come the crosslingual results for Spanish_orig are even better than the monolingual ones?

2. Why does English to Spanish seem to be the only combination in which Spanish_strict appears to work better than Spanish_orig?

To begin with, it seems like the definition of irony in the Italian dataset is closer to the one in Spanish_orig than the one in Spanish_strict. This may explain the huge gap between the numbers obtained in Spanish_orig and Spanish_strict in the detection of the ironic class; and, why the Italian system works better when testing with Spanish_orig, especially with the ironic class. For instance, several jokes (see Example (41)) that were considered ironic in Spanish_orig (but not in Spanish_strict), were also considered ironic by the system trained in Italian. Additionally, there were some tweets that were correctly classified by the Italian system that the Spanish models could not handle; which could explain why the Italian system outperformed the Spanish_orig monolingual ones. Examples (42) and (43) show a couple of tweets that were better classified by the Italian model than the Spanish ones. A possible explanation for the better performance of the Italian classifier could be sentence structure. The Italian system appears to have a tendency of classifying

tweets as ironic when they end with some short sentences after some kind of punctuation mark. For example, Example (44) is a non-ironic instance that was considered ironic by the Italian system.

(41) La tierra es plana por eso es planeta, sino sería redondéate
Earth is plane for that reason we call it planet, otherwise it would have been called roundet

(42) Peluqueroterapia" dice.
Heairdress-therapy" he/she says.

(43) "Cosa sencilla pero sofisticada". El concepto. #VenACenar289
'A simple thing but sophisticated'. The concept. #VenACenar289

(44) Que puta locura, dios mio gente creyendo que la Tierra es plana, ahora esto. Vivimos tiempos locos, ultralocos.
What a mess, people thinking Earth is flat, now this. We live at crezy times, super crazy ones.

Focusing on to pairs with English, it seems like the system trained with English struggles with the ironic tweets, and often labels them as non-ironic. This benefits the revised Spanish dataset as the number of non-ironic tweets is higher. A closer look reveals that the English system deals poorly with 'conversational' type of tweets (Example (45)). This may be due to the differences in dialog representation between languages. While English uses quote marks (""), Spanish (and Italian) employ hyphens (-).

(45) - ¿Has leído que @sanchezcastejon publica un libro? - ¿De quién?
"Have you read @sanchezcastejon 's book?", "Whose?"

Lastly, some common trends can be highlighted. Smileys, capital letters (especially when only a couple of words are highlighted) and some punctuation marks (e.g. double exclamation mark (!!) or ellipses (…)) appear to mislead the system into classifying the text as ironic, which align with previous research (for example, Carvalho et al., 2009; López and Ruiz, 2016).

Crosslingual experiments: summary

The last irony detection experiments set out to answer the third research question:

RQ3: Are irony cues and structures universal and, thus, transferable across languages?

The performance of the systems showed that some contexts worked better than others. Spanish_strict-to-English pair only got an F1-score of 0.49; on the other hand, the Italian-to-Spanish_orig pair obtained an F1-score of 0.78, which improved the best score obtained

in the monolingual setting for Spanish_orig. A possible cause of this phenomena could be a combination of: 1) a high compatibility between the two datasets, and 2) a better abstraction of particular structures such as short-sentence endings. Also, the closer the languages are, the better crosslinguality worked. Additionally, regardless od the Spanish dataset, better results were obtained from the pairs that were trained in English or French. We suggested that this could be related to the use of topics in the Spanish datasets, which could have made irony more context-dependant. Moreover, again, the use of excessive punctuation marks, capital letters and emojis could mislead the models.

So far, the results obtained for irony detection have been presented and discussed, answering three of the four research questions of the project. Next, the results obtained for the experiments of irony category classification will be discussed, and the last research question will be answered.

6.2 Irony Category Classification

Multiclass detection experiments were conducted using the irony categories annotated during this project in the IroSvA dataset. Two types of experiments were considered. To begin with, BETO and XLM-RoBERTa were used to train and evaluate two different systems that were fed with the ironical tweets which were annotated in terms of categories. Then, we built a full irony detection and classification pipeline for Spanish by using BETO to the predict the class of the ironical predictions obtained in the binary classification (some of which were not annotated as they were originally labeled as non-ironic).

Focusing first on the annotated ironic tweets, the distribution of the classes within the corpus is very unbalanced (as illustrated in Figure 2), thus it was expected the models to struggle with, at least, two of the classes: false assertion, and situational irony. As shown in Table 14, such predictions were correct since neither BETO nor XLM-RoBERTa were able to detect those two classes. When it comes to comparison, although not as bad as the previous two classes, the results were rather poor too. Plus, a look into the predictions revealed that the comparison structure that systems seemed to know best is the one introduced by 'como' (*like*). Moreover, often tweets were formed by several irony types, which could have difficult the detection of comparison. For instance, the tweet in Example (46)[20] was annotated under comparison since it was considered that the irony arises from the comparison of the two elements (what the politician said and the life of the user). However, the system predicted it to be hyperbole, which is not incorrect as the element that is compared to the 'fact' is indeed hyperbolic. Focusing on to the other three categories (contrast, hyperbole, rhetorical question), the results were more decent and it is worth noting that, similar to the monolingual results, BETO appeared to perform better than XLM-RoBERTa.

[20]This tweet is related to the popular thought that the politician Pablo Iglesias, the vice-president at that time, said on an interview that he felt more prepared to be president after taking care of his children and changing their diapers. Please note that the veracity of the facts is not taken into account for irony detection, only whether the speaker *believes* it to be true.

(46) Pues yo digo q si ese hombre, por decir algo, poe cambiar pañales es presidente. Yo con lo que llevo con Martina y Estrella, creo que me merezco ser ministro, aunque sea de hacienda
Well, I say that if that man, so to speak, can be president because of changing diappers. Looking at how long I've been with Martina and Estrella, I think I should be minister, at least of the tax office

Transformer	Comp.	Contr.	FA	Hyp.	RQ	Situational	F1-macro
XLM-RoBERTa	0.2105	0.4595	0.0000	0.5816	0.5789	0.0000	0.3051
BETO	0.1818	0.5333	0.0000	0.7320	0.7755	0.0000	**0.3704**

Table 14: Categories' detection

Given that the systems were not able to handle three of the classes, the experiment was repeated but, this time, only the three main categories were taken into account (contrast, hyperbole, rhetorical question). This experiment aimed to see if the absence of these categories had any impact on the results of the remaining ones. Similar or better results were expected since the possibilities were more constrained. As displayed in Table 15, such guesses were correct except for contrast.

Transformer	Contrast	Hyperbole	Rhetorical Question	F1-macro
XLM-RoBERTa	0.4103	0.6364	0.6667	0.5711
BETO	0.5217	0.7571	0.7907	**0.6899**

Table 15: Categories' detection (main three categories)

For a deeper understanding of the numbers obtained from both settings, confusion matrices were plotted. Figures 7 and 8 correspond to the ones obtained from XLM-RoBERTa while Figures 9 and 10 belong to BETO. As it can be observed, one of the possible factors in the worsening of contrast's detection could be the misclassified hyperbolic instances. With the absence of the other categories, contrast seems to be harboring the misclassified hyperbolic examples that were previously scattered within the rest of the categories. One of the reasons behind the error in detecting hyperbole may be the absence of exclamation marks or capital letters as in Example (47), which is mostly based on semantic and pragmatic comprehension[21]. Plus the lack of any question mark would probably contribute as the system would reject them to be rhetorical questions.

(47) La fluidez en el tráfico está sobrevalorada. Además, se generan más lugares de trabajo para los limpiacristales, vender pañuelos, etc...
Traffic flow is overrated. Also, more working spaces are created for window washers, tissue sellers, etc...

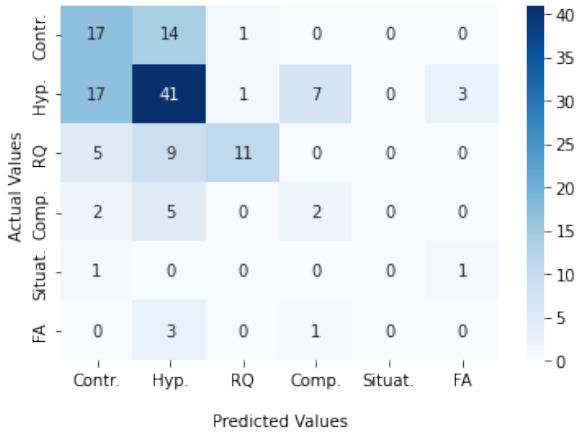

Figure 7: Confusion matrix for XLM-RoBERTa with all the categories

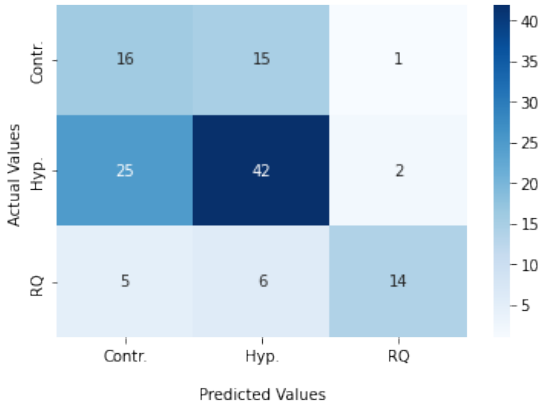

Figure 8: Confusion matrix for XLM-RoBERTa with half of the categories

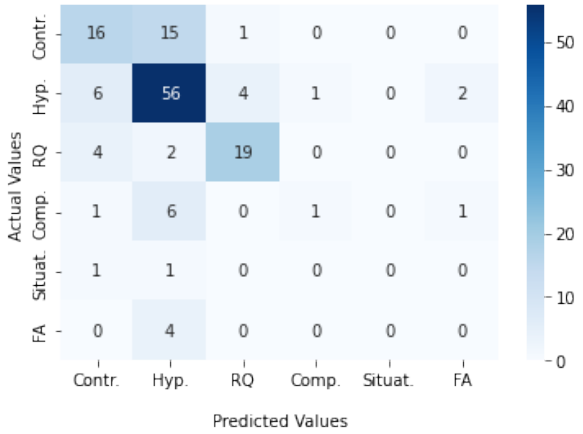

Figure 9: Confusion matrix for BETO with all the categories

Figure 10: Confusion matrix for BETO with half of the categories

Regarding hyperbole, with the reduction of classes, both systems presented a reduction of the errors derived from the assignment of hyperbole label to those instances that were actually not hyperbole (a.k.a. false positives). In other words, the systems improved distinguishing what is **not** hyperbole. This had an impact on the results of the precision metric, which increased, and consequently in the F1-macro. Additionally, XLM-RoBERTa's ability to correctly detect hyperbolic instances improves, whereas BETO's slightly decreases. However, BETO still detects hyperbole instances better. A comparison between the two systems' predictions suggested that BETO dealt better than XLM-RoBERTa with hyperbole examples without cues such as exclamation marks or capital letters, such as the one dis-

[21] It requires the ability to know that the message is not about the creation of those new jobs but rather about *why* those jobs would be created (the traffic jam).

played in Example (49). Regarding rhetorical question, both BETO and XLM-RoBERTa struggle with indirect questions like the one in Example (48). Moreover, XLM-RoBERTa seems to prioritize cues other than question marks (e.g. emojis or exclamation marks); consequently, those examples were often sent to other categories (mostly to contrast as mentioned above).

(48) Son mas importantes los semáforos paritarios o los banquitos de colorines. O la exhumacion de Franco
Which is more important the parity traffic lights or the colorful benches. Or Franco's exhumation.

(49) He llamado a la @NASA y me confirman qué la tierra no es plana
I've called the NASA and they've confirmed that Earth is not flat

The last multiclass experiment consisted of taking the ironic predictions obtained in the binary classification and testing what categories they would have been assigned to. As BETO yielded better results than XLM-RoBERTa in both scenarios binary (monolingual) and muticlass, this experiment was only conducted with BETO. Moreover, it should be mentioned that some of the tweets were actually non-ironical tweets, therefore they lacked of a truth category label. To handle this situation, an extra empty label was used (NaN). Figure 11 shows the confusion matrix for the ironic predictions' multiclass classification. Several observations can be made. First, it seems like BETO not only handled poorly the categorization of false assertion and situational irony but also failed to detect the irony on those tweets as there are no instances of those categories in the binary predictions. Regarding the rest of the classes, a look into the missed examples revealed that, except for rhetorical question, ironic instances that lacked of any marker where harder for the system to detect, as seen in Examples (50) and (51). However, to a lesser extent, there were also examples with emojis and punctuation marks (e.g. Example (52)). With regard to rhetorical question, most of the missed tweets did actually have question marks present; however, for some reason, the system understood them to be literal/genuine questions rather than ironic (for instance, Example (53)).

(50) He hablado con abogados de Nueva York y me han confirmado que la tierra es plana, cuando uno se equivoca lo mejor espedir disculpas
I've talked with New York's layers and they have confirmed that Earth is flat, when you make a mistake, it's best to apologize

(51) Porque además tengo el título de medicina y estoy encantada con la privatización y con que nos bajen el sueldo como en Andalucía. #YoconAlbert
Because I have a degree in medicine and I'm delighted with the privatization and the pay cut like in Andalucía. #I'mWithAlbert

(52) Resultado de la modélica" transacción española un juez franquista decide sobre la exhumación de Franco. Chorpecha!!!!"

The result of the exemplary Spanish transaction a francoist judge decides on Franco's exhumation. Surprise!!!!

(53) Cuando pedirán perdón Alemania e Italia a los Valencianos , por mandarnos a la Oltra y Grezzi ?
When will Germany and Italy apologize to Valencian people for sending us Oltra and Grezzi ?

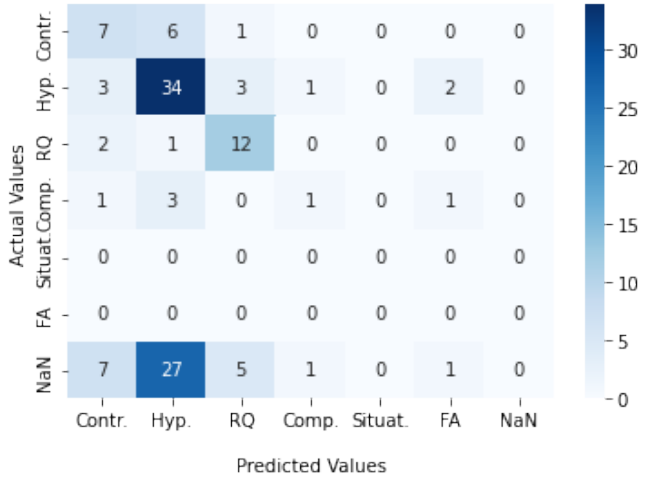

Figure 11: Categories' prediction for the ironic predictions in binary round (BETO)

Regarding the errors within the correctly detected irony tweets, they follow the tendencies previously mentioned (e.g. hyperbole and RQ easiest to detect, the influence of the different cues…). On the other hand, there are some curious errors when looking at non-ironical instances that were labeled as ironic in the binary round. For instance, Example (54) was labeled under false assertion despite the use of capital letters or exclamation marks (which often are related to hyperbole). Or Example (55) which was considered to be a rhetorical question although it has no question marks. Additionally, in occasions the system seemed to prioritize lexical words rather than other cues such as in Example (56) that was categorized as comparison despite the 'ugh' and the exclamation and question marks, most likely because of the use of 'parece' (*seems/looks like*) which can be used, and it is often used, in comparison structures

(54) El cambio climático es una mentira sustentada por falsos dogmas insostenibles!!! Entre ellos la forma de la tierra, ya que nuestro planeta NO es una Bola que gira en el espacio sideral! Es PLANA!
Climate change is a lie supported by unsustainable false dogmas!!! Among them the shape of the earth, since our planet is NOT a ball that rotates in outer space! It's flat!

(55) Si, busca. También encontrarás que el 2012 acababa el mundo, que la tierra es plana o que la magnetoteràpia cura.
Yes, search. You'll also find that the world ended in 2012, that the earth is flat or that magnetotherapy heals.

(56) Son ganas de tirar balones fuera...Nos estamos empezando a parecer a la "santa madre iglesia" que hasta que reconoció que la tierra no es plana y pidió perdón a Galileo.....uuuuuuhhhhhh!!! ¿Tan difícil es reconocer que fue a espada y cruz?!
They want to throw balls out... We are starting to look like the "holy mother church" until it recognized that the earth is not flat and apologized to Galileo.....uuuuuuhhhhhh!!! Is it so difficult to recognize that he went to sword and cross?!

Multiclass approach: summary

A part from irony detection experiments, we conducted some experiments in terms of irony category detection on the Spanish_orig dataset. These experiments aimed to answer the last research question of this project:

RQ4: How well can classifiers distinguish between the different irony categories in Spanish tweets?

As we have seen, three out of the six classes considered were hard to deal with for the systems, mainly because of the scarce data. Two systems were trained, one with the Spanish monolingual transformer (BETO) and another one with the multilingual transformer (XLM-RoBERTa). Hyperbole was the type that XLM-RoBERTa detected better. Hyperbole being the predominant class out of the six considered, the system may have learned easier about it. Regarding BETO, rhetorical question was the category that the model detected, most likely because of the use of question marks. Indeed, indirect questions were often incorrectly labeled.

Additionally, BETO's binary predictions (the ones obtained in the monolingual experiments) were also tested in terms of categories. Results showed that, during the monolingual experiments, the ironic tweets with the category of false assertion and situational irony were considered to be non-ironic. Also, some of curious cases could be found within the categories assigned to non-ironic tweets such as the categorization of a tweet as false assertion despite the use of capital letters which are often related to hyperbole.

In this chapter, we have examined the results obtained from the different experimental settings we proposed in Chapter 5. These experiments helped answer the research questions we aimed to investigate in this project. The following chapter will conclude the paper by summarizing the main points and suggesting some future work.

7 Conclusion and Future Work

This project set out to investigate irony detection in Spanish tweets. To this aim, two lines of work were considered: annotation work and experimental work. The annotation work was conducted in Spanish, and an existing corpus named IroSvA (Ortega-Bueno et al., 2019) was employed. This dataset was already annotated in terms of irony presence; in other words, whether a given tweet contained irony (thus, it was labeled as ironic) or not (labeled as non-ironic). However, irony lacks a single established definition and, as it turned out, the concept of irony employed in the IroSvA dataset deviated from the one used in this project. For this reason, a revision of the ironic class on the IroSvA dataset was conducted, and some ironic tweets were either discarded or moved to the non-irony class. For clarity reasons, the original dataset was named Spanish_orig, and the one obtained after the re-annotation work was called Spanish_strict. After the revision of ironic and non-ironic, the ironic tweets in Spanish_strict were annotated in terms of irony category. Six categories were considered: hyperbole, rhetorical question, contrast, comparison, situational irony, and false assertion. A quick scan into the class distribution suggested that irony-category classifiers would struggle with at least two of the categories (false assertion and situational irony) due to the unbalanced data.

Moving on to the experimental side of this project, two main lines of research were proposed. First, several irony detection tasks were considered to see how classifiers dealt with irony detection. In other words, how well could models tell apart tweets containing irony from tweets that do not. During the irony detection experiments, four datasets were employed: Spanish_orig, Spanish_strict, an English dataset, and an Italian dataset. In particular, three irony detection experimental contexts were considered: monolingual, multilingual, and crosslingual.

To begin with, irony detection in a monolingual setting was proposed in order to set some baselines for the following experiments as well as to answer the first research question: *Are monolingual transformers more suited than multilingual transformers for irony detection? If so, to what extent?* Monolingual transformers appeared to be more suitable for irony detection in monolingual contexts. Regarding the errors in the predictions, several factors that could have influenced the misclassification of ironic tweets were discussed; namely, explicit cues (such as punctuation marks, capital letters or emojis), non-ironic figures of speech (e.g. metaphor or hyperbole), and references to events and people.

After the monolingual experiments, some multilingual experiments were considered in order to answer the second research question: *Does multilingual data augmentation help improve the results obtained from the monolingual experiments?* Results showed that, even though in some cases data augmentation improved the results obtained by the multilingual transformer in the monolingual context, they never surpassed the results obtained by the monolingual transformer during monolingual experiments. For this reason, we concluded that, in these cases, multilingual data augmentation was not worth applying as there was no improvement and the training process takes longer.

The last irony detection experiments were conducted in a crosslingual or zero-shot

setting. Similar to Karoui et al. (2017), we aimed to research the universality of irony structures by answering the third research question: *Are irony cues and structures universal and, thus, transferable across languages?* Results showed that Italian/Spanish pairs obtained a higher F1-macro than the English/Spanish pairs, probably because of the similarity between languages. Focusing on the results, they were very varied, and some contexts seemed to work better than others. The worst result was obtained by the Spanish_strict to English pair with an F1-macro of 0.49. On the other hand, the best result was obtained by the Italian to Spanish_orig pair with an F1-macro of 0.78.

Finally, some irony classifiers were trained with the Spanish_strict set in order to detect the different irony categories. These experiments aimed to answer the last research question of this project: *How well can classifiers distinguish between the different irony categories in Spanish tweets?* Similar to monolingual irony detection experiments, the monolingual transformer worked better than the multilingual one. Regarding the results, systems dealt poorly with false assertion, situational irony, and comparison, probably due to the scarce data in those classes. Apart from experimenting with the ironic tweets labeled, the models were tested with the predictions obtained in the monolingual setting in order to see what categories were assigned. An analysis of the errors revealed that false assertion and situational irony were not detected as ironic in the irony detection round. Also, some curious cases were observed in which, for example, ironical and marked rhetorical questions were labeled as non-ironic.

Taken together, the present project makes several contributions to the area of irony detection in Spanish tweets. To begin with, we provided the re-annotation of the IroSvA corpus (Ortega-Bueno et al., 2019) in terms of irony/non-irony with a stricter concept of irony. We also presented, to our knowledge, the first dataset annotated in terms of irony categories for Spanish. Regarding the experiments, we provided a comparison between monolingual and multilingual transformers in a monolingual experimental setting. We also conducted the first experiments in multilingual and crosslingual settings that include Spanish. Lastly, we conducted the first study on irony categories detection in Spanish.

Irony detection is an interesting problem that could have several applications. For instance, Twitter has become one of the most widely used social media. Misunderstandings can easily occur and automatic irony detection could help prevent unnecessary uproars in the media. Additionally, there are people who struggle a lot detecting irony that could benefit from automatic irony detection, such as autistic patients (Deliens et al., 2018). With reference to future work, a natural progression of this work would be trying to detect irony in lengthier and more challenging genres such as political discourse. Also, more research into the use of irony in figures of speech would be interesting to conduct. Furthermore, developing a more fine-grained detection could be considered so that systems are able to tell:

- whether irony is concentrated in a particular word or it can be found all over the text;
- and, whether the irony segment changes the whole meaning of the text or just a piece of it.

Lastly, as a closing remark, we would like to draw attention to the importance of providing the definition employed when constructing the datasets. Throughout the experiments, we have seen how results changed depending on which Spanish dataset was involved. Although irony and its limits are rather subjective, providing a definition and clear annotation guidelines could help assess compatibility between datasets. This would probably help get more accurate results from the data augmentation experiments, for instance.

References

Rafael Torres Anchiêta, Francisco Assis Ricarte Neto, Jeziel C. Marinho, Kauan V. do Nascimento, and Raimundo Santos Moura. Piln idpt 2021: Irony detection in portuguese texts with superficial features and embeddings. In *IberLEF@SEPLN*, 2021.

Salvatore Attardo. Irony markers and functions: Towards a goal-oriented theory of irony and its processing. *Rask*, 12:3–20, 2000.

Salvatore Attardo. Irony as relevant inappropriateness. In Raymond W. Gibbs and Herbert L. Colston, editors, *Irony In Language and Thought. A Cognitive Science Reader*, pages 135–170. Psychology Press, New York, 2014.

Christian Burgers, Margot van Mulken, and Peter Jan Schellens. Verbal irony: Differences in usage across written genres. *Journal of Language and Social Psychology*, 31:290 – 310, 2012.

Paula Carvalho, Luís Sarmento, Mário J. Silva, and Eugénio de Oliveira. Clues for detecting irony in user-generated contents: oh...!! it's "so easy" ;-). In *CIKM 2009*, 2009.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*, 2020.

Alessandra Teresa Cignarella, Cristina Bosco, Viviana Patti, and Mirko Lai. Application and analysis of a multi-layered scheme for irony on the italian twitter corpus twittirò. In *LREC*, 2018a.

Alessandra Teresa Cignarella, Simona Frenda, Valerio Basile, Cristina Bosco, Viviana Patti, and Paolo Rosso. Overview of the evalita 2018 task on irony detection in italian tweets (ironita). In *EVALITA@CLiC-it*, 2018b.

Alessandra Teresa Cignarella, Cristina Bosco, and Paolo Rosso. Presenting twittirò-ud: An italian twitter treebank in universal dependencies. *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, 2019.

Alessandra Teresa Cignarella, Manuela Sanguinetti, Cristina Bosco, and Paolo Rosso. Marking irony activators in a universal dependencies treebank: The case of an italian twitter corpus. In *LREC*, 2020.

Andrea Cimino, Lorenzo De Mattei, and Felice Dell'Orletta. Multi-task learning in deep neural networks at evalita 2018. In *EVALITA@CLiC-it*, 2018.

Herbert L. Colston. Irony performance and perception: What underlies verbal, situational and other ironies? In Angeliki Athanadiadou and Herbert L. Colston, editors, *Irony in Language Use and Communication*, pages 19–41. John Benjamins Publishing Company, Amsterdam, 2017.

Herbert L. Colston and Raymond W. Gibbs. A brief history of irony. In Raymond W. Gibbs and Herbert L. Colston, editors, *Irony In Language and Thought. A Cognitive Science Reader*, pages 3–21. Psychology Press, New York, 2014.

Herbert L. Colston and Raymond W. Gibbs. Are irony and metaphor understood differently? *Metaphor and Symbol*, 17:57 – 80, 2002.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116, 2019. URL `http://arxiv.org/abs/1911.02116`.

Ulisses Brisolara Corrêa, Leonardo Coelho, Leonardo Pereira dos Santos, and Larissa Astrogildo de Freitas. Overview of the idpt task on irony detection in portuguese at iberlef 2021. *Proces. del Leng. Natural*, 67:269–276, 2021.

Diego de Vargas Feijó and Viviane Pereira Moreira. Mono vs multilingual transformer-based models: a comparison across several language tasks. *ArXiv*, abs/2007.09757, 2020.

Gaétane Deliens, Fanny Papastamou, Nicolas Ruytenbeek, Philippine Geelhand, and Mikhail Kissine. Selective pragmatic impairment in autism spectrum disorder: Indirect requests versus irony. *Journal of Autism and Developmental Disorders*, 48:2938–2952, 2018.

Shelly Dews, Joan Kaplan, and Ellen Winner. Why not say it directly? the social functions of irony. In Raymond W. Gibbs and Herbert L. Colston, editors, *Irony In Language and Thought. A Cognitive Science Reader*, pages 297–317. Psychology Press, New York, 2014.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard H. Hovy. A survey of data augmentation approaches for nlp. *ArXiv*, abs/2105.03075, 2021.

Bilal Ghanem, Jihen Karoui, Farah Benamara, Véronique Moriceau, and Paolo Rosso. Idat at fire2019: Overview of the track on irony detection in arabic tweets. *Proceedings of the 11th Forum for Information Retrieval Evaluation*, 2019.

Bilal Ghanem, Jihen Karoui, Farah Benamara, Paolo Rosso, and Véronique Moriceau. Irony detection in a multilingual context. *Advances in Information Retrieval*, 12036:141 – 149, 2020.

Aniruddha Ghosh and Tony Veale. Ironymagnet at semeval-2018 task 3: A siamese network for irony detection in social media. In **SEMEVAL*, 2018.

Debanjan Ghosh and Smaranda Muresan. "with 1 follower i must be awesome : P". exploring the role of irony markers in irony recognition. In *ICWSM*, 2018.

Debanjan Ghosh, Avijit Vajpayee, and Smaranda Muresan. A report on the 2020 sarcasm detection shared task. *ArXiv*, abs/2005.05814, 2020.

Rachel Giora. On irony and negation. *Discourse Processes*, 19:239–264, 1995.

José-Ángel González, Lluís F. Hurtado, and Ferran Plà. Elirf-upv at irosva: Transformer encoders for spanish irony detection. In *IberLEF@SEPLN*, 2019a.

José-Ángel González, Lluís F. Hurtado, and Ferran Plà. Elirf-upv at irosva: Transformer encoders for spanish irony detection. In *IberLEF@SEPLN*, 2019b.

Roberto I. González-Ibáñez, Smaranda Muresan, and Nina Wacholder. Identifying sarcasm in twitter: A closer look. In *ACL*, 2011.

H. Paul Grice. Further notes on logic and conversation. In Peter Cole, editor, *Syntax and Semantics. Vol. 9: Pragmatics*, pages 41–58. New York: Academic, 1978.

John Haiman. *Talk Is Cheap: Sarcasm, Alienation, and the Evolution of Language*, chapter Sarcasm and Its Neightbours. Oxford University Press, 1998.

Jeffrey T. Hancock. Verbal irony use in face-to-face and computer-mediated conversations. *Journal of Language and Social Psychology*, 23:447 – 463, 2004.

Henk Haverkate. A speech act analysis of irony. *Journal of Pragmatics*, 14:77–109, 1990.

Cynthia Van Hee, Els Lefever, and Veronique Hoste. Guidelines for annotating irony in social media text, version 2.0. 2016.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. Semeval-2018 task 3: Irony detection in english tweets. In **SEMEVAL*, 2018.

Shengyi Jiang, Chuwei Chen, Nankai Lin, Zhuolin Chen, and Jinyi Chen. Irony detection in the portuguese language using bert. In *IberLEF@SEPLN*, 2021.

Jihen Karoui, Farah Benamara, Véronique Moriceau, Viviana Patti, Cristina Bosco, and Nathalie Aussenac-Gilles. Exploring the impact of pragmatic phenomena on irony detection in tweets: A multilingual corpus study. In *EACL*, 2017.

David Kaufer. Understanding ironic communication. *Journal of Pragmatics*, 5:495–510, 1981.

Muhammad Khalifa and Noura Hussein. Ensemble learning for irony detection in arabic tweets. In *FIRE*, 2019.

Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. A large self-annotated corpus for sarcasm. *ArXiv*, abs/1704.05579, 2018.

Roger J. Kreuz and Sam Glucksberg. How to be sarcastic: The echoic reminder theory of verbal irony. *Journal of Experimental Psychology: General*, 118:374–386, 1989.

Roger J. Kreuz and Richard M. Roberts. On satire and parody: The importance of being ironic. *Metaphor and Symbol*, 8:97–109, 1993.

Christopher J. Lee and Albert N. Katz. The differential role of ridicule in sarcasm and irony. *Metaphor and Symbol*, 13:1–15, 1998.

Hankyol Lee, Youngjae Yu, and Gunhee Kim. Augmenting data for sarcasm detection with unlabeled conversation context. *ArXiv*, abs/2006.06259, 2020.

Christine Liebrecht, Florian Kunneman, and Antal van den Bosch. The perfect solution for detecting sarcasm in tweets #not. In *WASSA@NAACL-HLT*, 2013.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL `http://arxiv.org/abs/1907.11692`.

Gabriela Jasso López and Ivan Vladimir Meza Ruiz. Character and word baselines systems for irony detection in spanish short texts. *Procesamiento Del Lenguaje Natural*, 56:41–48, 2016.

Joan M. Lucariello. Situational irony: A concept of events gone awry. *Journal of Experimental Psychology: General*, 123:129–145, 1994.

Hairo Ulises Miranda-Belmonte and Adrián Pastor López-Monroy. Early fusion of traditional and deep features for irony detection in twitter. In *IberLEF@SEPLN*, 2019.

Reynier Ortega-Bueno, Francisco Manuel Rangel Pardo, D. I. H. Farías, Paolo Rosso, Manuel Montes y Gómez, and José Eladio Medina-Pagola. Overview of the task on irony detection in spanish variants. In *IberLEF@SEPLN*, 2019.

Loreto Parisi, Simone Francia, and Paolo Magnani. Umberto: an italian language model trained with whole word masking. *GitHub*, 2020. URL `https://github.com/musixmatchresearch/umberto`.

Tomás Ptácek, Ivan Habernal, and Jun Hong. Sarcasm detection on czech and english twitter. In *COLING*, 2014.

Sara Rosenthal, Alan Ritter, Preslav Nakov, and Veselin Stoyanov. Semeval-2014 task 9: Sentiment analysis in twitter. *ArXiv*, abs/1912.02990, 2014.

Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Antonio Stranisci. An italian twitter corpus of hate speech against immigrants. In *LREC*, 2018.

Andrea Santilli, Danilo Croce, and Roberto Basili. A kernel-based approach for irony and sarcasm detection in italian. In *EVALITA@CLiC-it*, 2018.

Sahil Swami, Ankush Khandelwal, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. A corpus of english-hindi code-mixed tweets for sarcasm detection. *ArXiv*, abs/1805.11869, 2018.

Yi-Jie Tang and Hsin-Hsi Chen. Chinese irony corpus construction and ironic structure analysis. In *COLING*, 2014.

Erik F. Tjong Kim Sang. Introduction to the conll-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*, 2002. URL `https://aclanthology.org/W02-2024`.

Oren Tsur, Dmitry Davidov, and Ari Rappoport. Icwsm - a great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews. In *ICWSM*, 2010.

Aline Aver Vanin, Larissa Astrogildo de Freitas, Renata Vieira, and Marco N. Bochernitsan. Some clues on irony detection in tweets. In *WWW '13 Companion*, 2013.

Yanfei Wang, Yang-De Chen, and Yuejie Zhang. Improving sentence classification by multilingual data augmentation and consensus learning. In *CCL*, 2020.

Deirdre Wilson. Irony comprehension: A developmental perspective. *Journal of Pragmatics*, 59:40–56, 2013.

Dreirdre Wilson and Dan Sperber. On verbal irony. In Raymond W. Gibbs and Herbert L. Colston, editors, *Irony In Language and Thought. A Cognitive Science Reader*, pages 35–95. Psychology Press, New York, 2014.

Ellen Winner and Howard Gardner. Metaphor and irony: Two levels of understanding. In Andrew Ortony, editor, *Metaphor and Thought*, pages 425–443. Cambridge University Press, 1993.

Chuhan Wu, Fangzhao Wu, Sixing Wu, Junxin Liu, Zhigang Yuan, and Yongfeng Huang. Thu_ngn at semeval-2018 task 3: Tweet irony detection with densely connected lstm and multi-task learning. In **SEMEVAL*, 2018.