ISSN 2007-9737

Linguistic-based Approach for Recognizing Implicit Language in Hate Speech: Exploratory Insights

Antonio Reyes, Rafael Saldívar

Universidad Autónoma de Baja California,
Facultad de Idiomas,
Mexico

{areyes98, rafaelsaldivar}@uabc.edu.mx

Abstract. Language, in all its forms, is one of the most comprehensive ways to characterize human societies. By means of the analysis of regular components, either at phonetic, morphological, syntactic or semantic level, human language provides valuable information that can be translated into knowledge in order to represent behavioral patterns. For instance, in web texts, such as the ones posted on Twitter or Facebook, it is quite frequent to find linguistic expressions, such as the following one: "Don't come here. If you are afraid for your life and you have no place to go, don't pick this country." This text could denote an explicit description of the current immigration phenomenon and, likewise, it could connote an implicit content of mockery, aggressiveness, or even hate. Both interpretations are possible, but only one of them is more likely according to the author profiling. This fact stresses out the underlying problem that it is faced in this investigation: Many of our utterances entail two communicative dimensions. The explicit dimension (literal use of language), and the implicit dimension (figurative use of language). Both dimensions are supposed to communicate information thought consciously. In this respect, the most challenging issue for this approach relies on the recognition of the correct communicative dimension profiled by the author in a web text. In this context, this article focuses on analyzing textual information, mainly extracted from Twitter, in order to set a computational framework to differentiate between explicit and implicit language. In particular, we are interested in recognizing figurative uses regarding irony and sarcasm, in order to apply the findings to better understand and prevent social problems related to hate speech.

Keywords. Implicit language, figurative language, hate speech, irony, sarcasm.

1 Introduction

Language, in all its forms, is one of the most comprehensive ways to characterize human societies. However, given its social nature, it cannot be only defined in terms of grammatical issues. In this respect, while it is true that grammar regulates language in order to have a non-chaotic system, it is also true that language is dynamic, and accordingly, a live entity. This means that language is not static; rather, it is in constant interaction between the rules of its grammar and its pragmatic use. For instance, the idiom *"all of a sudden"* has a grammatical structure which is not made intelligible only by knowledge of the familiar rules of its grammar [15], but by inferring implicit information. This latter process fills in the gap to properly interpret the idiom.

The previous example shows how our utterances entail two dimensions to decode what it is intended to be communicated: The explicit dimension, which is mainly featured by the use of literal language (*not* means not), and the implicit dimension, in which the use of figurative language is often profiled, for instance, by the use of figurative devices, such as irony, sarcasm, metaphor, among others (*not* could mean yes, perhaps, possibly, or more).

In simple words, it could be argued that the explicit dimension is what any hearer could understand effortlessly, whereas the implicit dimension is the hidden information to be unveiled by the same hearer to fully understand what the speaker is communicating.

This latter dimension is clearly the most challenging one to be recognized (and formalized), for both people and computers.

In this context, this article is focused on analyzing textual information, mainly extracted from Twitter, in order to recognize formal elements for setting a computational framework to differentiate between explicit and implicit language. In particular, the analysis is performed in the scenario of hate speech. To this end, a corpus with hate speech tweets in Spanish was built. It is divided in four classes to better understand how hate speech is verbalized explicitly and implicitly.

The challenge of recognizing whether an utterance conveys implicit content of hate speech or not is faced by analyzing two figurative devices: Irony and sarcasm. According to the specialized literature, one of the most challenging issues regarding hate speech is precisely the presence of devices such as the ones cited [22, 38, 25, 27, 47, 30]. In addition, as mentioned in the previous paragraphs, figurative language is commonly used to communicate information not given literally [33, 34]. This fact can be seen in the following tweet:

> "Don't come here. If you are afraid for your life and you have no place to go, don't pick this country."

Literally, this text is communicating an explicit description of the current immigration phenomenon; therefore, it could be classified as a harmless tweet. However, in the implicit dimension, it is also communicating a veiled threat; therefore, it should be classified as a hate speech tweet. Assuming that the second interpretation is correct, one way to unveil the threat is by recognizing that a figurative device, such as irony, underlies the tweet.

Given this distinction, we are interested in analyzing figurative language (irony and sarcasm, specifically) in order to better understand how hate speech is linguistically expressed.

The rest of the article is organized as follows: In Section 2 the theoretical background concerning figurative language will be introduced. The related work on irony, sarcasm, and hate speech will be described in Section 3. The analysis of the data and the discussion of the findings will be detailed in Section 4. Finally, in Section 5, we will conclude with some final remarks and some pointers to address the future work.

2 Two Dimensions of Language

Modern linguists deem language as a continuum of symbolic structures in which lexicon, morphology, and syntax form a continuum which differs along various parameters, what can be divided into separate components only arbitrarily [24].

Language, thus, is viewed as an entity whose components and levels of analysis cannot be independent nor isolated. On the contrary, they are embedded in a global system which depends on cognitive, experiential, and social contexts, which go far beyond the linguistic system proper [21].

This vision, according to the cognitive linguistics bases, entails a close relation between semantics and conceptualization (cf. [24]), i.e., apart from grammar, the linguistic system is dependent on cognitive domains, in which both referential knowledge (e.g., lexical semantic information) and inferential knowledge (e.g., contextual and pragmatic information) are fundamental to understand what it is communicated.

Based on this integral vision of language, in which its grammatical substance is as important as its social referents, the explicit (literal) and implicit (figurative) dimensions of language will be described below.

2.1 Literal Language (Explicit Dimension)

The simplest definition of literal language is related to the notion of $true$, $exact$ or $real$ meaning, i.e., a word (isolated or within a context) conveys one single meaning (the one conventionally accepted), which cannot be deviated. In this respect, some experts have highlighted certain properties of literalness: It is direct, grammatically specified, sentential, necessary, and context-free [20].

Hence, it is assumed that it must be invariant in all contexts. According to [1], literalness is generated by linguistic knowledge of lexical items, combined with linguistic rules. Therefore, it is determined, **explicit**, and fully compositional. For instance, the word $flower$ can only refer to the

concept of plant, regardless of its use in different communicative acts or discourses (e.g., botany, evolution, poetry, etc.).

2.2 Figurative Language (Implicit Dimension)

In the context of a dichotomous view of language, figurative language could be regarded as the opposite of literal language. Thus, whereas the latter is assumed to communicate a direct and explicit meaning, the former is more related to the notion of conveying veiled or implicit meanings.

For instance, the word $flower$, which literally refers only to the concept of plant, speaking figuratively can refer to several concepts, which not necessarily are linked to plants. Therefore, it can be used instead of concepts such as beauty, peace, purity, life, and so on, in such a way its literal meaning is intentionally deviated in favor of secondary interpretations.

Although, at first glance, this distinction seems to be clear and sufficient on its own, figurative language involves basic cognitive processes rather than only deviant usage [29]. Therefore, it is necessary to go deeper into the mechanisms and processes that differentiate both dimensions of language.

In accordance with classical perspectives, the notions of literalness and figurativity are viewed as pertaining directly to language, i.e., words have literal meanings, and can be used figuratively [20].

Consequently, figurative language could be regarded as a type of language that is based on literal meaning, but is disconnected from what people learn about the world [or about the words] based on it [them] [4].

Thus, by breaking this link, literal meaning loses its primary referent and, accordingly, the interpretation process becomes senseless. Let us consider Chomsky's famous example to explain this issue:

> "Colorless green ideas sleep furiously" [8].

Beyond grammatical aspects, in the previous example it is possible to observe how the decoding process is achieved easily enough. Either phonologically or orthographically, Chomsky's example is fully understandable in terms of its linguistic constituents.

However, when interpreting, its literal meaning is completely nonsensical. For instance, the bigrams [colorless green] or [green ideas] are sufficiently disconnected from their conventional referents for being able to produce a coherent interpretation.

Thus, in order to make the example understandable, secondary interpretations are necessary. If such interpretations are successfully activated, then figurative meaning is triggered and, accordingly, a more coherent interpretation can be achieved.

Based on this explanation, literal meaning could be deemed as denotative, whereas figurative meaning, connotative, i.e., figurative meaning is not given a priori; rather, it must be implicated.

Finally, it is worth stressing out that language on its own provides specific linguistic devices to intentionally express different types of implicit contents: Metaphor, allegory, irony, similes, analogy, sarcasm, and so on.

2.3 Objective

Unlike literal language, figurative language uses linguistic devices such as irony, sarcasm, metaphor, analogy, and so on, in order to communicate implicit content, which is not usually interpretable by simply decoding syntactic or semantic information. Rather, figurative language reflects patterns of thought within a communicative and social framework that turns quite challenging its linguistic representation, as well as its computational processing.

In this respect, our objective is to develop a linguistic-based framework to recognize implicit content about hate speech in web texts. By the analysis of two specific domains of figurative language, it is intended to provide arguments about how people conceptualize hate speech, and how they verbalize such discourse deliberately. In particular, we are interested in developing formal models to recognize ironic and sarcastic texts, in

which people veil consciously the contents of hate speech, in order to prevent and reduce, hopefully, the impact of such behaviors on the society.

3 Related Work on Figurative Language and Hate Speech

In this section, two figurative devices, irony and sarcasm, will be described in terms of their automatic processing. In addition, the related work on hate speech will be referred.

3.1 Irony

Like most figurative devices, irony is difficult to pin down in formal terms, and no single definition ever seems entirely satisfactory. According to various experts, irony is essentially a communicative act that expresses an opposite meaning of what was literally said, i.e., irony is a playful use of language in which a speaker implies the opposite of what is literally said [50, 9].

In terms of its automatic processing, there have been various approaches to automatically detect irony in text. For instance, [42] reported one of the first computational attempts to formalize the phenomenon. His model attempted to represent irony by modeling the interaction between speakers and hearers. [43, 44] analyzed the cognitive processes that underlie verbal irony to separate irony from non-irony in figurative comparisons.

In addition, [7] determined some clues for automatically identifying ironic sentences [34] and [33], in turn, presented a set of linguistic-based features to determine whether a tweet is ironic or not. More recently, [5], as well as [40] have developed ad hoc corpora for the task in languages beyond English. Likewise, some other researchers have addressed the task by setting a social media scenario in which it is quite common to find ironic statements about anything.

For instance, [6] focus their approach on product reviews, [49], on personal blogs, [13, 18, 39], on microblogs such as Twitter. [45], in turn, investigate irony in broader scenarios such as online communities.

3.2 Sarcasm

Although at first glance the terms irony and sarcasm seem to be concepts perfectly distinguishable from each other, when they are used in real communicative scenarios, such distinction is rarely accomplished. In this respect, [19] states that sarcasm, but not irony, involves the ridicule of a specific person or group of people.

It could be argued, for instance, that irony courts ambiguity and often exhibits great subtlety, whereas sarcasm is delivered with a cutting or withering tone that is rarely ambiguous. However, these differences rely indeed on matters of usage, tone, and obviousness, rather than only on theoretical assumptions.

With respect to sarcasm detection, [41], as well as [10], addressed their research to finding sarcastic patterns in online products reviews and tweets, respectively. [16] investigated the impact of lexical and pragmatic features on the task. Some others works have addressed the task by analyzing texts from social media platforms, especially, Twitter: [3, 28, 2, 32, 39] are examples about it. On the other hand, [31] approached sarcasm from a multilingual point of view. Finally, some research works have provided corpora for detecting sarcasm in different types of documents, for instance, [14, 26, 17].

3.3 Hate Speech

As stated previously, the challenge of detecting implicit content in text will be focused on hate speech. First of all, the term hate speech tends to be too general.

According to the United Nations, the term refers to “any kind of communication that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are”[1]. In addition, [27] listed six different definitions of hate speech. On this matter, the authors highlight what hate speech is for both specialists and social media platforms, such as Twitter or Facebook. Some targets of hate speech are related to others’ inherent properties, such as religion, ethnicity, nationality, race, color, descent,

[1] https://www.un.org/

or gender. On the other hand, as noted by various researchers, any formal definition of hate speech is far from being universal. For instance, there is no clear boundary between hate speech and freedom of speech.

Despite these drawbacks, there is an increasing interest in the academic community on this topic due to the social implications of hate speech found all over the internet and mass media. In the context of Natural Language Processing, some approaches to hate speech are totally related to social media.

For instance, [46, 12, 11] analyze hate speech on websites, web comments and Facebook, respectively. [51, 30] focus on hate speech about the migration phenomenon and the immigrants. [23, 36] approach hate speech regarding specific target communities. Likewise, some other investigations have addressed the problem by assessing particular features to detect hate speech automatically in different languages. The works reported by [47, 48, 30, 35] represent fair examples about this concern. It is also worth noting the development of lexical resources, language models, and systems to automatically deal with this phenomenon [22, 38, 25].

4 Analysis

The data for the analysis, as well as the experiments performed to assess our preliminary findings are described below.

4.1 Hate Speech Data

A data set with tweets in Spanish was built in order to analyze how people verbalize explicit and implicit content related to hate speech. The tweets were collected manually by twelve doctoral students in Language Sciences. Because of the manual gathering, no hashtags were considered for collecting the data. Instead, the students were asked to read the most tweets they could in a time interval of three weeks. After reading them, they had to select the ones that they deemed to express hate speech. To this end, they attended some lectures on the topic; therefore, each one had a theoretical background about hate speech, as well as a variety of discussions about the different ways to express it linguistically.

In order to provide the students with a guide to systematize the task, four categories of hate speech were defined a priori: Violence, discrimination, bullying/harassment, and general (this last category is intended to cover tweets that cannot be classified in the previous ones). Each student should classify his/her tweets into one of these categories by identifying the target of the message. Finally, the students should annotate their tweets with two labels: Explicit hate speech or implicit hate speech.

The total amount of tweets collected with these criteria was 10.883. All of them were written in Spanish. Although the data set contains different dialectal variants, the most representative one is the Mexican. General statistics of the data set are provided in Table 1. This data set will be available for academic purposes in the near future.

4.2 Implicit Hate Speech Agreement

In Section 2.2 it was stated that the implicit content is not given straightforwardly. Therefore, to guarantee that the tweets annotated with the implicit label were, in fact, members of this class, a subsequent task was requested of the students. They had to read the tweets annotated with the label implicit hate speech to confirm that the tweet, indeed, belongs to such class.

The total number of tweets annotated with this label was 2.638. Thus, each student assessed 220 tweets, i.e., every tweet in this class was annotated twice.

The final data set with implicit hate speech content was built by selecting only the tweets assessed by two students as belonging to the implicit class. If a tweet was assessed by one student as implicit hate speech, but the second student assessed as explicit hate speech, or vice-versa, then such tweet was disregarded.

By doing this, the total number of tweets in the implicit hate speech class was reduced to 1.973. Such reduction, hypothetically, should ensure a set of fine-grained tweets in which the implicit content could be analyzed with deeper insights. The final distribution per category is depicted in Table 2.

Table 1. Statistics of the Hate Speech (HS) data set

	Violence	Discrimination	Bullying Harassment	General
Explicit HS content	1.928	1.952	1.160	3.205
Implicit HS content	402	616	546	1.074
Total tweets	2.330	2.568	1.706	4.279

Table 2. Final tweets in Implicit HS class

	Original class	Fine-grained class
Violence	402	347
Discrimination	616	461
Bullying/Harassment	546	476
General	1.074	689

4.3 Figurative Language Recognition

In order to examine how often the figurative devices appeared in the tweets with implicit hate speech, a classification task was performed. The underlying assumption, according to the information given in the previous sections, was to verify whether or not this set of tweets were implicitly communicating hate speech content by means of using irony or sarcasm.

In this respect, the remaining 1,973 tweets were classified in three categories: Ironic, sarcastic or literal. A set of some of the most discriminating features described in the specialized literature was used for representing both figurative devices in the texts (see Sections 3.1 and 3.2). In this respect, features such as BoW, polarity, aggressiveness, among other were used to represent the documents.

Finally, the Bayes algorithm was used to classify. The results are summarized in Figure 1.

As noted in the figure, when focusing on both figurative devices, most tweets were classified as ironic for almost the four categories, except for the category Bullying/Harassment, in which the balance between the classes ironic and sarcastic was very similar. However, it is worth noting that several tweets were classified in the third class, i.e., according to the set of features used in the classification, they are neither ironic nor sarcastic.

This outcome highlights two aspects to consider: (i) as described in the previous sections, figurative language is used to convey hate speech in a more sophisticated way, especially, by using irony. This means that more complex models regarding irony detection or sarcasm detection could improve the performance of current systems to detect implicit hate speech automatically in online communities; (ii) it is quite fuzzy to establish a formal boundary between the explicit and the implicit content when analyzing hate speech data. If several tweets were classified as literal hate speech, then the implicit content is being conveyed by means of different communicative strategies, not necessarily related to figurative language.

In the following section, both aspects are approached in linguistic terms in order to set a framework to allow the processing of implicit hate speech based on the observations noted so far.

4.4 Linguistic Features

One of the most challenging issues when manually revising some of the tweets with implicit hate content was related to the distinction between literal and figurative language. Although there are some works related to explain such distinction theoretically, when reading what common users post in social media, it is evident that the problem is much more complex than the functional distinction exposed in sections 2.1 and 2.2. In this respect, one element that we observed in the manual review to differentiate between literal and figurative content is the so-called *intention*.

This extra-linguistic element is useful to explain why figurative language requires much more cognitive effort to correctly interpret its meaning. If we look at any of the tweets from this data set (or any other), it is easy to realize that

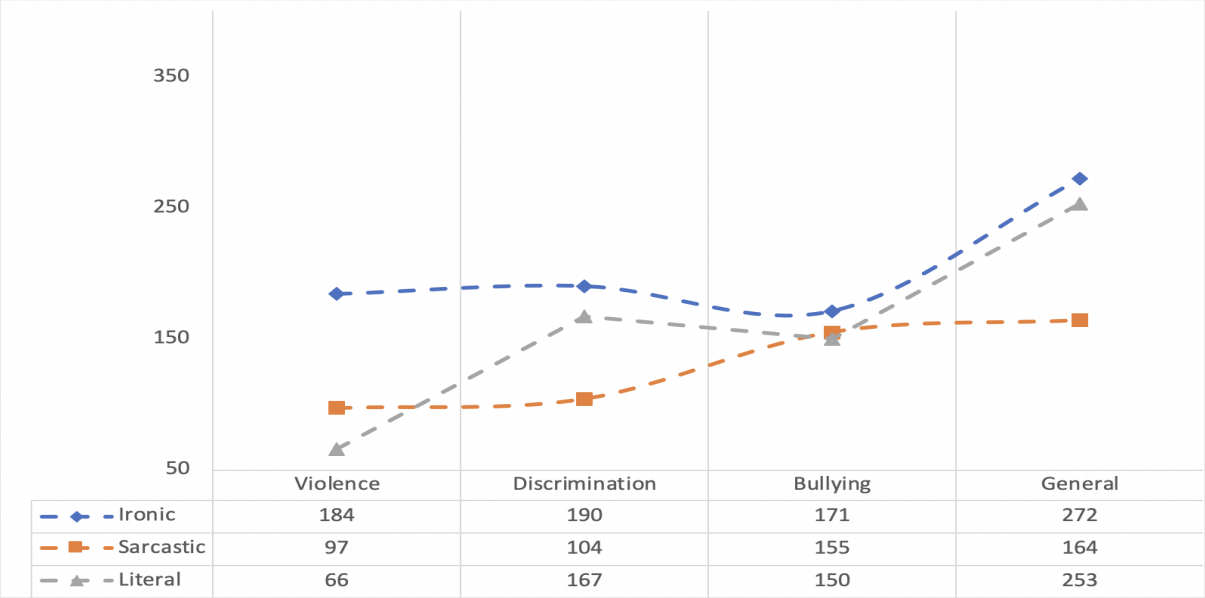

Fig. 1. Tweets classification per category

they are only sequences of words with semantic meaning. Perhaps, such meaning is totally explicit (literalness), or perhaps, it could be senseless (figurativity). This difference could be explained in terms of performance and competence, or even as a matter of correctness.

However, in a more comprehensive conception of language, such difference would be motivated by the need of maximizing a communicative success [37].

This need could be the element that will determine what type of information has to be profiled linguistically. If a literal content is profiled, then certain intention will permeate the statement. This intention will find a linguistic formalization by selecting some words or syntactic structures, for instance, to successfully communicate what it is intended. In contrast, if the figurative content is profiled, then the intention will guide the choice of different linguistic elements to guarantee the right transmission of information. It is likely that such content cannot be accomplished, but in this case, the failure will not rely on the speaker's intention; rather, on the hearer's skills to interpret what is communicated figuratively. Let us observe the following tweets to clarify this point.

(a) "Esta gente solo merece el rechazo y el desprecio". (These people deserve rejection and contempt only).

(b) "Podrán decir lo que quieran de los de Tepito pero son de las pocas personas que se tapan la boca para estornudar o toser, hasta cargan tiner para desinfectarse las manos". (You could say anything about people from Tepito, but they indeed cover their mouths when sneezing or coughing. Actually, they even use thinner to clean their hands).

Whereas in (a) the intention is to express hate speech against a social group, in (b) the intention is to express hate speech implicitly by means of using encrypted elements.

In each statement, the speaker has a communicative need, which is solved by maximizing certain elements. Thus, in the first example, the communicative success is based on making a precise affirmation (note that all the words in this context are very clear in terms of their semantic meaning). In contrast, the second example is based on deliberately selecting elements that entail secondary and non literal relations: Using thinner

to clean hands is a sarcastic way to say that these people are drug addicts. In addition, by naming the place Tepito, the speaker is implicitly communicating that they are poor and, likely, criminal. Now, it is not that simple to identify what the intention is.

As noted above, this is an extra-linguistic element. Therefore, it is quite difficult to be formalized. However, there is a fact that deserves in depth analysis to face this issue: Understanding the intention often involves an interpretive adjustment to individual words, i.e., not all the words in an utterance are triggering an implicit intention; for this reason, the intention tends to be usually triggered by manipulating individual words.

In addition, we explore some linguistic features to go in deep with the recognition of the mechanisms to convey implicit hate speech (beyond figurative issues). It is worth noting that such features are work in progress; thus, their usefulness is preliminary. Finally, in order to be assessed further, some of them are listed below:

1. Senseless and incongruity.
2. Textual entailment.
3. Semantic frames.
4. Entropy.

These features are intended to provide elements to analyze implicit content at different level. For instance, implicit content is supposed to be achieved by processing the linguistic input in secondary paths; then, by analyzing components such as the incongruity produced by the simplest interpretation, or by analyzing the valences in syntactic chunks within a discussion thread, or even, by measuring the entropy among n-grams, we consider that it is linguistically feasible to recognize some patterns to approach implicit language.

To illustrate this, let us consider a 4-gram, such as “mafia del no poder”. This is an atypical sequence in a reference corpus; therefore, its entropy could make evident that something is happening: processed literally is a senseless sequence, but making the right inferences, its violent content unveils.

5 Conclusions and Further work

In this article it has been presented an exploratory approach for facing implicit language in hate speech tweets. To this end, a data set with hate speech content in Spanish was manually built. The tweets were classified in four categories (violence, discrimination, bullying/harassment, and general), and then, they were labeled by human annotators in two classes: Explicit hate speech or implicit hate speech.

The approach relied on first analyzing figurative language, especially regarding irony and sarcasm, in the tweets belonging to the implicit hate speech class. Then, a manual review was carried out for investigating in deep what kind of formal information could be recognized for characterizing implicit language (considering both figurative and literal use of language) in the context of hate speech. In this respect, a core feature was suggested for differentiating figurative from literal language, as well as a set of exploratory linguistic features was introduced to approach implicit language in the near future beyond the presence of figurative devices.

The initial findings are encouraging, although a more robust set of experiments has to be done in order to demonstrate how useful such a set of features could be. The further work consists in assessing the exploratory linguistic features by comparing with some of the data sets used in some competitions, such as HatEval, MeOffendEs, and others.

References

1. **Ariel, M. (2002).** The demise of a unique concept of literal meaning. Journal of Pragmatics, Vol. 34, No. 4, pp. 361–402.
2. **Bamman, D., Smith, N. (2015).** Contextualized sarcasm detection on twitter. Proceedings of the International AAAI Conference on Web and Social Media, Vol. 9, No. 1, pp. 574–577.
3. **Barbieri, F., Saggion, H., Ronzano, F. (2014).** Modelling sarcasm in Twitter, a novel approach. Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Association for Computational

Linguistics, Baltimore, Maryland, pp. 50–58. DOI: 10.3115/v1/W14-2609.

4. **Bergen, B. (2005).** Mental simulation in literal and figurative language understanding. In **Coulson, S.**, editor, The Literal and Nonliteral in Language and Thought. Peter Lang Publishing, pp. 255–280.
5. **Bosco, C., Patti, V., Bolioli, A. (2013).** Developing corpora for sentiment analysis: The case of irony and senti-tut. Intelligent Systems, IEEE, Vol. 28, pp. 55–63. DOI: 10.1109/MIS.2013.28.
6. **Buschmeier, K., Cimiano, P., Klinger, R. (2014).** An impact analysis of features in a classification approach to irony detection in product reviews. Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Association for Computational Linguistics, Baltimore, Maryland, pp. 42–49. DOI: 10.3115/v1/W14-2608.
7. **Carvalho, P., Sarmento, L., Silva, M., de Oliveira, E. (2009).** Clues for detecting irony in user-generated contents: oh...!! It's "so easy" ;-). TSA '09: Proceeding of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion, ACM, Hong Kong, China, pp. 53–56.
8. **Chomsky, N. (1957).** Syntactic Structures. Mouton and Co, The Hague.
9. **Colston, H., Gibbs, R. (2007).** A brief history of irony. In **Gibbs, R., Colston, H.**, editors, Irony in Language and Thought. Taylor and Francis Group, pp. 3–24.
10. **Davidov, D., Tsur, O., Rappoport, A. (2010).** Semi-supervised recognition of sarcastic sentences in Twitter and Amazon. Proceedings of the Fourteenth Conference on Computational Natural Language Learning, CoNLL '10, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 107–116.
11. **Del Vigna, F., Cimino, A., Dell'Orletta, F., Petrocchi, M., Tesconi, M. (2017).** Hate me, hate me not: Hate speech detection on facebook. pp. .
12. **Erjavec, K., Kovačič, M. P. (2012).** "you don't understand, this is a new war!" analysis of hate speech in news web sites' comments. Mass Communication and Society, Vol. 15, No. 6, pp. 899–920. DOI: 10.1080/15205436.2011.619679.
13. **Fersini, E., Pozzi, F., Messina, E. (2015).** Detecting irony and sarcasm in microblogs: The role of expressive signals and ensemble classifiers. 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA), pp. 1–8.
14. **Filatova, E. (2012).** Irony and sarcasm: Corpus generation and analysis using crowdsourcing. Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), European Language Resources Association (ELRA), Istanbul, Turkey, pp. 392–398.
15. **Fillmore, C., Kay, P., O'Connor, M. (1988).** Regularity and idiomaticity in grammatical constructions: The case of let alone. Language, Vol. 64, No. 3, pp. 501–538.
16. **González-Ibáñez, R., Muresan, S., Wacholder, N. (2011).** Identifying sarcasm in Twitter: A closer look. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Vol. 2, The Association for Computer Linguistics, pp. 581–586.
17. **Joshi, A., Sharma, V., Bhattacharyya, P. (2015).** Harnessing context incongruity for sarcasm detection. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Association for Computational Linguistics, Beijing, China, pp. 757–762. DOI: 10.3115/v1/P15-2124.
18. **Karoui, J., Benamara Zitoune, F., Moriceau, V., Aussenac-Gilles, N., Hadrich Belguith, L. (2015).** Towards a contextual pragmatic model to detect irony in tweets. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Association for Computational Linguistics, Beijing, China, pp. 644–650. DOI: 10.3115/v1/P15-2106.
19. **Katz, A. (2005).** Discourse and sociocultural factors in understanding nonliteral language. In **Colston, H., Katz, A.**, editors, Figurative Language Comprehension: Social and Cultural Influences. Lawrence Erlbaum Associates, pp. 183–208.
20. **Katz, J. (1980).** Propositional structure and illocutionary force: A study of the contribution of sentence meaning to speech acts. Harvard University Press.
21. **Kemmer, S. (2010).** About cognitive linguistics: Historical background. online resource.

22. **Klein, G. B. (2018).** Applied linguistics to identify and contrast racist 'hate speech': Cases from the English and Italian language. ALRJournal, Vol. 2, No. 3, pp. 1–16. DOI: 10.14744/alrj.2018.36855. Doi: 10.14744/alrj.2018.36855.

23. **Kwok, I., Wang, Y. (2013).** Locate the hate: Detecting tweets against blacks. Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence, AAAI'13, AAAI Press, pp. 1621–1622.

24. **Langacker, R. (1991).** Concept, Image and Symbol. The Cognitive Basis of Grammar. Mounton de Gruyter.

25. **Lemmens, J., Markov, I., Daelemans, W. (2021).** Improving hate speech type and target detection with hateful metaphor features. Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda, Association for Computational Linguistics, Online, pp. 7–16. DOI: 10.18653/v1/2021.nlp4if-1.2.

26. **Lukin, S., Walker, M. (2013).** Really? well. apparently bootstrapping improves the performance of sarcasm and nastiness classifiers for online dialogue. Proceedings of the Workshop on Language Analysis in Social Media, Association for Computational Linguistics, Atlanta, Georgia, pp. 30–40.

27. **MacAvaney, S., Yao, H.-R., Yang, E., Russell, K., Goharian, N., Frieder, O. (2019).** Hate speech detection: Challenges and solutions. PLoS ONE, Vol. 14, No. 8.

28. **Maynard, D., Greenwood, M. (2014).** Who cares about sarcastic tweets? Investigating the impact of sarcasm on sentiment analysis. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), European Language Resources Association (ELRA), Reykjavik, Iceland, pp. 4238–4243.

29. **Peters, W. (2004).** Detection and Characterization of Figurative Language Use in WordNet. Ph.D. thesis, University of Sheffield, Sheffield, England.

30. **Poletto, F., Stranisci, M., Sanguinetti, M., Patti, V., Bosco, C. (2017).** Hate speech annotation: Analysis of an italian twitter corpus. CLiC-it.

31. **Ptácek, T., Habernal, I., Hong, J. (2014).** Sarcasm detection on Czech and English twitter. COLING.

32. **Rajadesingan, A., Zafarani, R., Liu, H. (2015).** Sarcasm detection on twitter: A behavioral modeling approach. Proceedings of the Eighth ACM International Conference on Web Search and Data Mining.

33. **Reyes, A., Rosso, P. (2012).** Making objective decisions from subjective data: Detecting irony in customers reviews. Decision Support Systems, Vol. 53, No. 4, pp. 754–760.

34. **Reyes, A., Rosso, P., Veale, T. (2013).** A multidimensional approach for detecting irony in Twitter. Language Resources and Evaluation, Vol. 47, No. 1, pp. 239–268.

35. **Sanguinetti, M., Poletto, F., Bosco, C., Patti, V., Stranisci, M. (2018).** An Italian Twitter corpus of hate speech against immigrants. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Miyazaki, Japan.

36. **Silva, L. A., Mondal, M., Correa, D., Benevenuto, F., Weber, I. (2016).** Analyzing the targets of hate in online social media. ICWSM.

37. **Sperber, D., Wilson, D. (2002).** Relevance theory. Handbook of Pragmatics, Vol. 42, No. 5, pp. 607–632.

38. **Srivastava, A., Vajpayee, A., Akhtar, S. S., Jain, N., Singh, V., Shrivastava, M. (2020).** A multi-dimensional view of aggression when voicing opinion. Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, European Language Resources Association (ELRA), Marseille, France, pp. 13–20.

39. **Sulis, E., Irazú Hernández Farías, D., Rosso, P., Patti, V., Ruffo, G. (2016).** Figurative messages and affect in twitter: Differences between #irony, #sarcasm and #not. Knowledge-Based Systems, Vol. 108, pp. 132–143. DOI: https://doi.org/10.1016/j.knosys.2016.05.035. New Avenues in Knowledge Bases for Natural Language Processing.

40. **Tang, Y.-j., Chen, H.-H. (2014).** Chinese irony corpus construction and ironic structure analysis. Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, Dublin City University and Association for Computational Linguistics, Dublin, Ireland, pp. 1269–1278.

41. **Tsur, O., Davidov, D., Rappoport, A. (2010).** ICWSM – A great catchy name: Semi-supervised

recognition of sarcastic sentences in online product reviews. **Cohen, W. W., Gosling, S.**, editors, Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, The AAAI Press, Washington, D.C., pp. 162–169.

42. **Utsumi, A. (1996).** A unified theory of irony and its computational formalization. Proceedings of the 16th conference on Computational Linguistics, Association for Computational Linguistics, Morristown, NJ, USA, pp. 962–967.
43. **Veale, T., Hao, Y. (2009).** Support structures for linguistic creativity: A computational analysis of creative irony in similes. Proceedings of CogSci 2009, the 31st Annual Meeting of the Cognitive Science Society, pp. 1376–1381.
44. **Veale, T., Hao, Y. (2010).** Detecting ironic intent in creative comparisons. Proceedings of 19th European Conference on Artificial Intelligence - ECAI 2010, IOS Press, Amsterdam, The Netherlands, pp. 765–770.
45. **Wallace, B. C., Choe, D. K., Charniak, E. (2015).** Sparse, contextually informed models for irony detection: Exploiting user communities, entities and sentiment. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Beijing, China, pp. 1035–1044. DOI: 10.3115/v1/P15-1100.
46. **Warner, W., Hirschberg, J. (2012).** Detecting hate speech on the world wide web. Proceedings of the Second Workshop on Language in Social Media, LSM '12, Association for Computational Linguistics, USA, pp. 19–26.
47. **Waseem, Z. (2016).** Are you a racist or am I seeing things? annotator influence on hate speech detection on Twitter. Proceedings of the First Workshop on NLP and Computational Social Science, Association for Computational Linguistics, Austin, Texas, pp. 138–142. DOI: 10.18653/v1/W16-5618.
48. **Waseem, Z., Hovy, D. (2016).** Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. Proceedings of the NAACL Student Research Workshop, Association for Computational Linguistics, San Diego, California, pp. 88–93. DOI: 10.18653/v1/N16-2013.
49. **Whalen, J. M., Pexman, P. M., Gill, A. J., Nowson, S. (2013).** Verbal irony use in personal blogs. Behaviour & Information Technology, Vol. 32, No. 6, pp. 560–569. DOI: 10.1080/0144929X.2011.630418.
50. **Wilson, D., Sperber, D. (2007).** On verbal irony. In **Gibbs, R., Colston, H.**, editors, Irony in Language and Thought. Taylor and Francis Group, pp. 35–56.
51. **Zagheni, E., Garimella, V. R. K., Weber, I., State, B. (2014).** Inferring international and internal migration patterns from twitter data. Proceedings of the 23rd International Conference on World Wide Web, WWW '14 Companion, Association for Computing Machinery, New York, NY, USA, pp. 439–444. DOI: 10.1145/2567948.2576930.

Article received on 26/07/2021; accepted on 20/09/2021.
Corresponding author is Antonio Reyes.