Unsupervised Irony Detection: A Probabilistic Model with Word Embeddings

Debora Nozza, Elisabetta Fersini and Enza Messina
DISCo, University of Milano-Bicocca, Viale Sarca, 336, Milan, Italy

Keywords: Irony Detection, Unsupervised Learning, Probabilistic Model, Word Embeddings.

Abstract: The automatic detection of figurative language, such as irony and sarcasm, is one of the most challenging tasks of Natural Language Processing (NLP). This is because machine learning methods can be easily misled by the presence of words that have a strong polarity but are used ironically, which means that the opposite polarity was intended. In this paper, we propose an unsupervised framework for domain-independent irony detection. In particular, to derive an unsupervised Topic-Irony Model (TIM), we built upon an existing probabilistic topic model initially introduced for sentiment analysis purposes. Moreover, in order to improve its generalization abilities, we took advantage of Word Embeddings to obtain domain-aware ironic orientation of words. This is the first work that addresses this task in unsupervised settings and the first study on the topic-irony distribution. Experimental results have shown that TIM is comparable, and sometimes even better with respect to supervised state of the art approaches for irony detection. Moreover, when integrating the probabilistic model with word embeddings (TIM+WE), promising results have been obtained in a more complex and real world scenario.

1 INTRODUCTION

Mining opinions and sentiments from user generated texts expressed in natural language is an extremely difficult task. It requires a deep understanding of explicit and implicit information conveyed by language structures, whether in a single word or an entire document (Bosco et al., 2013). In particular, social media users are inclined to adopt a creative language making use of original devices such as sarcasm and irony (Ghosh et al., 2015a).

These figures of speech are commonly used to intentionally convey an implicit meaning that may be the opposite of the literal one. According to Colston and Gibbs (Colston and Gibbs, 2007) an ironic message typically conveys a negative opinion using only positive words.From the sentiment analysis perspective such utterances represent a challenge as an interfering factor that can revert the message polarity (usually from positive to negative). The detection of ironic expressions is crucial in different application domains, such as marketing and politics, where the users tend to subtly communicate dissatisfaction usually referring to a product or to a political ideology or politician.

Although sarcasm and irony are a well-studied phenomenons in linguistics, psychology and cognitive science, their automatic detection is still a great challenge because of its complexity. Standard dictionary-based methods for sentiment analysis, based on a predefined sentiment-driven lexicon, have often shown to be inadequate in the face of indirect figurative meanings (Ghosh et al., 2015a). Several methods have been proposed to evaluate the abilities of semi-supervised and supervised machine learning approaches to tackle irony detection problem. However, they assume as prerequisite human annotation of texts as training data, which in a real social media context is costly and difficult even for human, so as to make it prohibitive. Moreover, it is commonly known that supervised machine learning classifiers trained on one domain often fail to produce satisfactory results when shifted to another domain, since natural language expressions can be quite different (Blitzer et al., 2007).

In this paper we propose a fully unsupervised framework for domain-independent irony detection. To perform unsupervised topic-irony detection, we built upon an existing *probabilistic topic model*, initially introduced for sentiment analysis purposes. The aim of this model is to discover the hidden thematic structure in large archives of texts. Probabilistic topic models are particularly suitable for two main reasons: first, they are able to discover topics embedded in text

Nozza, D., Fersini, E. and Messina, E.
Unsupervised Irony Detection: A Probabilistic Model with Word Embeddings.
DOI: 10.5220/0006052000680076
In *Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2016) - Volume 1: KDIR*, pages 68-76
ISBN: 978-989-758-203-5

messages in an unsupervised way, and second, they result in a language model that estimates how much a word is related to each topic and to the irony figure of speech.

Moreover, in order to improve the generalization abilities we took advantage of *word embeddings* to obtain domain-aware ironic orientation for words. This is the first work that addresses the problem of irony detection in a fully unsupervised settings. Furthermore, this paper contributes as a first investigation on irony-topic models.

The rest of the paper is organized as follows. Section 2 introduces the related work. In Section 3, the proposed framework grounded on an unsupervised Topic-Irony model and Word Embeddings are presented. In Section 4, the experimental investigation is presented. Finally, we conclude and discuss further research directions in Section 5.

2 RELATED WORK

As defined in (Edward and Connors, 1971), a figure of speech is any artful deviation from the ordinary mode of speaking or writing. Among the most problematic figures of speech in Natural Language Processing (NLP) we focused on sarcasm and irony (Katz et al., 2005), which are commonly used to convey implicit criticism with a particular victim as its target, saying or writing the opposite of what the author means (McDonald, 1999). As mentioned in (Weitzel et al., 2016), language should not be taken literally, especially when addressing a sentiment analysis task. The presence of strongly positive (or negative) words that are used ironically, which means that the opposite polarity was intended, can easily mislead sentiment analysis classification models (Reyes and Rosso, 2014).

In the last year several approaches for irony detection based on different set of features have been investigated. In (Davidov et al., 2010), the authors proposed a semi-supervised technique to detect sarcasm in Amazon product reviews and tweets. They used pattern-based (high frequency words and content words) and punctuation-based features to build the sarcasm detection model. A supervised approach has been proposed in (González-Ibánez et al., 2011), where the irony detection problem is studied for sentiment analysis in Twitter data. The authors used unigrams, word categories, interjections (e.g., ah, yeah), and punctuation as features. Emoticons and ToUser (which marks if a tweet is a reply to another tweet) were also used. In (Riloff et al., 2013), the authors considered a specific type of sarcasm where sarcastic tweets include a positive sentiment (such as “love” or “enjoy”) followed by an expression that describes an undesirable activity or state (e.g., “taking exams” or “being ignored”). In (Reyes et al., 2013) the authors focused on developing classifiers to detect verbal irony based on a set of high-level features: ambiguity, polarity unexpectedness and emotional cues. In (Ptáček et al., 2014) a supervised model has been exploited for document-level irony detection in Czech and English by using n-grams, patterns, POS tags, emoticons, punctuation and word case.

A similar approach, where a novel set of linguistically related features are used, has been presented in (Barbieri and Saggion, 2014). In (Fersini et al., 2015) the authors proposed an ensemble approach, based on a Bayesian Model Averaging paradigm, which makes use of models trained using several linguistic features, such as pragmatic particles and Part-Of-Speech tags. In (Hernández-Farías et al., 2015), the irony detection problem has been addressed by investigating statistical-based and lexicon-based features paired with two semantic similarity measures, i.e. Lesk and Wu-Palmer (Pedersen et al., 2004).

Other recent works (Bamman and Smith, 2015; Rajadesingan et al., 2015) aim to address the sarcasm detection in microblogs by including extra-linguistic information from the context such as properties of the author, the audience, the immediate communicative environment and the user’s past messages. Word embeddings have been used as features in a supervised approach in (Ghosh et al., 2015b), where the authors expressed the sarcasm detection task as a word sense disambiguation problem.

Although the above mentioned studies represent a fundamental step towards the definition of effective irony detection systems, they suffer of three main limitations:

- they assume a labelled corpus for training supervised and semi-supervised models;
- they are tailored for domain-dependent irony detection, restraining their applicability to other domain of interest;
- they disregard the topic subjected to the irony.

In order to overcome these limitations, we investigated an unsupervised topic-irony model enriched with domain-independent word embeddings.

3 PROPOSED FRAMEWORK

3.1 Topic-Irony Model (TIM)

In order to perform unsupervised irony detection, taking into account also the topic-dependency of the words, we focused our investigation on the suite of generative models called *probabilistic topic models*, originally defined for sentiment purposes. We considered three main generative models, which are extensions of the well-known Latent Dirichlet Allocation model (Blei et al., 2003).

The first one is Topic Sentiment Mixture (TSM) (Mei et al., 2007), that jointly models the mixture of topics and sentiment predictions for the entire document. Here, the sentiment language model is considered as separated from the topics ones, that can lead to a language model that is not able to explain the hidden correlation between a topic and sentiment. The second one is Joint Sentiment-Topic (JST) model (Lin and He, 2009), which assumes that topics are dependent on sentiment distributions and words are conditioned on sentiment-topic pairs. The last one is Aspect and Sentiment Unification Model (ASUM) (Jo and Oh, 2011), that slightly differs from JST with respect to the language distribution constraints. While in JST each word may come from different language models, ASUM constrains the words in a single sentence to come from the same language model.

Among these models, we based our Topic-Irony Model on ASUM. This choice is motivated by the fact that (1) the topic-irony model should generate a topic and an ironic/not-ironic orientation for each word (2) this model is particularly suitable for microblog text, where messages have a maximum number of characters and a sentence would be either ironic or not ironic with respect to a specific topic (3) ASUM makes use of a set of seed words explicitly integrated into the generative process, making the model more stable from a statistical point of view.

The proposed Topic-Irony model (TIM) is able to model irony toward different topics in a fully unsupervised paradigm, enabling each word in a sentence to be generated from the same *irony-topic* distribution. More formally, let D be the number of documents, M the number of sentences, N the number of words, T the number of topics, I the number of irony classes {ironic, not ironic} and V the vocabulary size.

The generative process is as follows:

1. For every pair of (i,z) such that $i \in I$ and $z \in T$, draw a word distribution $\phi_{iz} \sim Dirichlet(\beta_i)$.
2. For each document d,
   - **(a)** Draw the document's irony distribution $\pi_d \sim Dirichlet(\gamma)$
   - **(b)** For each $i \in I$, draw a topic distribution $\theta_{di} \sim Dirichlet(\alpha)$
   - **(c)** For each sentence
     - **i.** Choose an irony class $\hat{i} \sim Multinomial(\pi_d)$
     - **ii.** Given $\hat{i}$, choose a topic $\hat{z} \sim Multinomial(\theta_{d\hat{i}})$
     - **iii.** Generate words $w \sim Multinomial(\phi_{\hat{i}\hat{z}})$

Following (Jo and Oh, 2011), β is the parameter that controls the integration of seed words in the models and we used its asymmetric form. Indeed, one can expect that the words "news, bbc, science" are not probable in ironic expressions, and similarly "lol, oh, duh" are probably ironic expressions. This expectation can be encoded in β. The latent variables θ, π, and ϕ are inferred by Gibbs sampling. The graphical representation of TIM is shown in Figure 1.

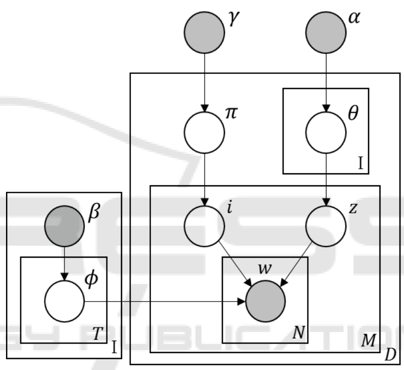

Figure 1: Graphical representation of TIM. Nodes are random variables, edges are dependencies, and plates are replications. Shaded nodes are observable.

3.2 Word Embeddings (WE)

The original ASUM topic model makes use of known general sentiment seed words to derive domain-specific sentiment words (Jijkoun et al., 2010). For sentiment seed words, existing sentiment word lexicons can be used (e.g., SentiWordNet (Esuli and Sebastiani, 2006)) or a new set of words may be obtained by using sentiment propagation techniques (Kaji and Kitsuregawa, 2007; Mohammad et al., 2009; Rao and Ravichandran, 2009; Lu et al., 2011).

For irony detection, a lexicon cannot be a priori defined, but it can be automatically derived in an unsupervised way using huge quantity of text. To this purpose, *word embeddings* can be adopted to derive latent relationships among words (e.g. *irony* is strictly related to *epic fail*) and therefore to automatically create lexicons based on the language model used in online social networks. This representation is derived by various training methods inspired from neural-network models In our investigation the ironic-lexicon, among the available distributed representations (Bengio et al., 2006; Turian et al., 2010; Huang et al., 2012), two model architectures have been used (Mikolov et al., 2013), Continuous Bag of Words (CBOW) and Skip-gram have been chosen because of their efficiency on training and their limited loss of information. The training objective of CBOW is to combine the representations of surrounding words to predict the word in the middle. Similarly, in the Skip-gram model, the training objective is to learn word vector representations that are good at predicting context in the same sentence. The model architectures of these two methods are shown in Figure 2.

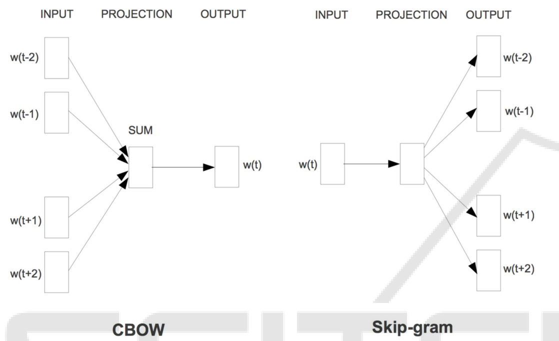

Figure 2: Graphical representation of the CBOW and Skip-gram model. The CBOW architecture predicts the current word based on the context, and the Skip-gram predicts surrounding words given the current word.

In practice, Skip-gram gives better word representations when the monolingual data is small. However, CBOW is faster and more suitable for larger datasets. The Skip-gram and CBOW models are typically trained using stochastic gradient descent. The gradient is computed using back propagation rule (Rumelhart et al., 1986). When trained on a large dataset, these models capture a substantial amount of semantic information. Closely related words will have similar vector representations, e.g., *Italy*, *France*, *Portugal* will be similar to *Spain*. More interestingly, the word vectors can also capture complex analogy patterns. For example, *vector(king)* is to *vector(man)* as *vector(queen)* is to *vector(woman)*.

4 EXPERIMENTS

4.1 Dataset and Evaluation Settings

We evaluated the proposed framework on a benchmark dataset for irony detection (Reyes et al., 2013). The dataset contains 10,000 ironic tweets and 30,000 non-ironic tweets (10,000 for each topic: Education, Humour and Politics). As in the original paper, we performed a series of binary classifications, between Irony vs Education, Irony vs Humour and Irony vs Politics in a **balanced** settings (50% ironic texts and 50% not ironic texts). We also considered the task with **unbalanced** classes, i.e. to learn ironic vs others. In order to deal with a more realistic and complex scenario, where the term irony can not be explicitly available, we evaluated the proposed model according to two experimental conditions:

- Original scenario (O): the dataset has been maintained as it is (where the hashtags have been removed), in order to allow a direct comparison with the state of the art models;
- Simulated scenario (S): the hashtags and the term *irony* have been removed from the data in order to simulate a more realistic and complex scenario where the presence of irony is not explicitly pointed out.

Concerning the proposed model, two hyper-parameters, γ and β, have been tuned. γ is a prior for the irony distribution in texts. Because it is not possible to make assumptions on this distribution, several configurations have been evaluated. The second hyper-parameter, β, is the key elements for integrating the seed words that originate through WE into TIM. β is the prior of the word-irony-topic distribution defined for ironic seed words, not ironic seed words and all the other words.

The construction of the irony lexicon (to be enclosed as seed words) has been performed by training the *word embedding* model on all the tweets in the corpus. The seed words have been obtained by extracting the most similar words to the term “irony”. After a preliminary experimental investigation, we decided to report the results related to the best distributed representation. In particular, the following results are related to the CBOW model thanks to its ability to deal with large corpus.

In the following experimental results, TIM will denote the Topic-Irony Model, while TIM+WE will represent the Topic-Irony Model based on the lexicons induced by CBOW. The experimental investigation is conducted by comparing TIM, TIM+WE and two supervised approaches available in the literature.

We evaluated the performance in terms of Precision (P), Recall (R), F-Measure (F), distinguishing between ironic (+) and not ironic (-). A global performance measure is also reported in terms of Accuracy.

4.2 Irony Detection Results

4.2.1 Balanced Dataset

Original Scenario (O). The results of the proposed framework is compared with (Reyes et al., 2013) in Table 1. Our framework clearly outperforms the supervised method with significant improvements, i.e. (on average) 11% for Precision, 14% for Recall and 13% for F-Measure.

Table 1: Results compared with a supervised state-of-the-art method for each binary problem (O).

		P	R	F
irony	(Reyes et al., 2013)	0.7600	0.6600	0.7000
vs	TIM	0.8225	**0.8746**	**0.8477**
education	TIM + WE	**0.8228**	0.8629	0.8423
irony	(Reyes et al., 2013)	0.7500	0.7100	0.7300
vs	TIM	0.9127	**0.8560**	**0.8834**
politics	TIM + WE	**0.9131**	0.8373	0.8735
irony	(Reyes et al., 2013)	0.7800	0.7400	0.7600
vs	TIM	**0.8414**	**0.8174**	**0.8292**
humour	TIM + WE	0.8142	0.7832	0.7983

A further remark relates to the Precision and Recall obtained by TIM and TIM+WE. It can be easily noted that the two proposed models achieve homogeneous performance on both orientations and in all the binary classification problems, obtaining Precision and Recall performance of similar magnitude. In order to grasp more peculiar behaviours, the performance measures both for ironic (+) and not ironic (-) texts have been reported in Table 3. In this case, we can highlight that Precision and Recall for both classes are well balanced, ensuring good performance also on the most difficult (ironic) target. Concerning accuracy, TIM and TIM+WE are able not only to outperform a trivial classifier that would ensure 50% of accuracy, but they perform differently according to the binary problem that they address. We can note that tackling Irony vs Humor is more difficult than Irony vs Politics and Irony vs Education. In fact, as stated by the authors of the original paper (Reyes et al., 2013), the similarity estimated between pairs of classes is significantly higher in Irony vs Humor than the other binary problems.

Moreover, we can notice that in this scenario the contribution of the ironic-lexicon derived through WE does not generally improve the performances of TIM. This is probably due to the impact that the word *irony* has into the dataset and into the model: the lexicon of TIM only composed of the *irony* term is sufficient to discriminate between the ironic and non-ironic orientations. Although the additional seed words enclosed in TIM+WE allow the model to obtain remarkable results with respect to the supervised settings and similar performance compared to TIM, the only presence of the term *irony* guarantees better performance than richer lexicons. As expected, TIM better fits the original scenario where the ironic statements available into the dataset are strongly characterized by the *irony* term. In order to evaluate the generalization abilities of the proposed models in a real and more complex scenario, where the term *irony* is not explicitly available into the ironic statements, we evaluated the performance in the following simulated scenario.

We report some additional results to compare the proposed approaches with respect to some related works (supervised) on the same dataset used for the experimental investigation. In particular, the benchmark corpus exploited for training and inference TIM and TIM+WE has been previously adopted also in (Barbieri and Saggion, 2014) and (Hernández-Farías et al., 2015) (only in a balanced settings for the original scenario). The results reported in terms of F-Measure by the original authors are shown in Table 2.

Table 2: Results in terms of F-Measure of the proposed models against the state of the art approaches.

	Irony vs.		
	Education	Humour	Politics
(Reyes et al., 2013)	0,70	0,76	0,73
(Barbieri and Saggion, 2014)	0,73	0,75	0,75
(Hernández-Farías et al., 2015)[1]	0,78	0,75	0,79
(Hernández-Farías et al., 2015)[2]	0,78	0,79	0,79
TIM	**0,85**	**0,83**	**0,88**
TIM+WE	**0,84**	**0,80**	**0,87**

This final comparison clearly highlights the contribution that the proposed models are able to provide. Not only TIM and TIM+WE perform significantly better than the state of the art models, but it is even more remarkable that they perform better although their nature is completely unsupervised.

Simulated Scenario (S). We report in the following the computational results on the simulated scenario, where the ironic figurative language is not explicitly marked in the dataset, but embedded in to the sentences. In Table 4 the results in terms of precision, recall and F-measure are reported distinguishing between ironic (+) and not-ironic (-) classes, together with the global Accuracy measure.

As expected, the recognition performance of TIM and TIM+WE decrease, compared to the original scenario (see Table3), once the term *irony* is removed from the corpus. However, in this case where the

[1]In this experiment, the authors used the Lesk similarity measure.

[2]In this experiment, the authors used the Wu-Palmer similarity measure.

Table 3: Results of our framework for each binary problem (O).

		P (+)	R (+)	F (+)	P (-)	R (-)	F (-)	Accuracy
irony *vs* education	TIM	0,8225	**0,8746**	**0,8477**	**0,8664**	0,8116	**0,8380**	**0,8430**
	TIM + WE	**0,8228**	0,8629	0,8423	0,8566	**0,8146**	0,8350	0,8388
irony *vs* politics	TIM	0,9127	**0,8560**	**0,8834**	**0,8644**	0,9183	**0,8905**	**0,8871**
	TIM + WE	**0,9131**	0,8373	0,8735	0,8498	**0,9204**	0,8836	0,8788
irony *vs* humour	TIM	**0,8414**	**0,8174**	**0,8292**	**0,8227**	**0,8461**	**0,8342**	**0,8318**
	TIM + WE	0,8142	0,7832	0,7983	0,7911	0,8214	0,8059	0,8022

Table 4: Results of our framework for each binary problem (S).

		P (+)	R (+)	F (+)	P (-)	R (-)	F (-)	Accuracy
irony *vs* education	TIM	0,7996	0,7934	0,7964	0,7958	0,8022	0,7989	0,7977
	TIM + WE	**0,8050**	**0,8103**	**0,8075**	**0,8098**	**0,8046**	**0,8070**	**0,8073**
irony *vs* politics	TIM	0,8719	0,8358	0,8534	0,8426	0,8775	0,8596	0,8567
	TIM + WE	**0,8780**	**0,8420**	**0,8596**	**0,8485**	**0,8833**	**0,8655**	**0,8627**
irony *vs* humour	TIM	**0,7356**	0,7675	0,7510	0,7574	**0,7247**	**0,7405**	0,7460
	TIM + WE	0,7205	**0,8392**	**0,7752**	**0,8079**	0,6752	0,7354	**0,7570**

presence of irony is not explicitly pointed out, a lexicon able to boost TIM and therefore the recognition performance of ironic messages becomes beneficial. By analysing all the performance measures, it is clear that the introduction of WE derived-lexicon allow the probabilistic model TIM+WE to achieve better results than simple TIM. Also in this experimental settings, we can remark that the two proposed models are able to obtain Precision and Recall performance of similar magnitude, highlighting robust performance in this complex scenario.

4.2.2 Unbalanced Dataset

Original Scenario (O). In order to compare the proposed framework with the state of the art on irony detection, we reported in Figure 3 the results obtained by TIM and TIM+WE with two supervised approaches, i.e. the irony model presented in (Reyes et al., 2013) and the ensemble approach introduced in (Fersini et al., 2015). First of all, TIM and TIM+WE are able perform better than a trivial classifier that would ensure 70% of accuracy. Furthermore, we can point out that both proposed unsupervised models achieve remarkable results compared to the supervised ones. In particular, we can highlight that both

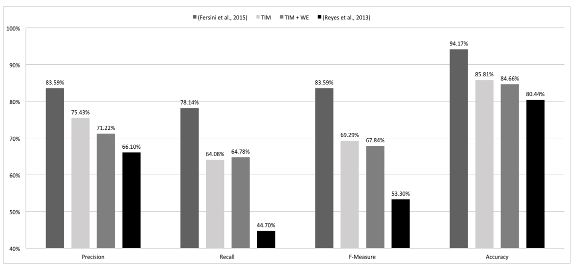

Figure 3: Comparison of TIM and TIM+WE with supervised state of the art methods on the unbalanced dataset.

TIM and TIM+WE are able to obtain higher recognition performance than the supervised irony model introduced in (Reyes et al., 2013). When comparing the proposed models to the ensemble presented in (Fersini et al., 2015), we can point out that TIM and TIM+WE are not so far from the supervised model.

Considering that our method is fully unsupervised, we can state that the proposed models are promising. Extended results of our framework are shown in Table 5. Similar to the balanced experimental settings on the original scenario, the contribution of WE does not generally improve the performance of TIM, still remaining comparable. Again, the better performance obtained by TIM with respect to TIM+WE is related to the dataset composition, where more than 36% of the ironic textual messages contains the word *irony*.

Simulated Scenario (S). In the following, we report the computational results on the simulated scenario, where irony meaning is embedded in the sentences with no reference to the term *irony*. We report in Table 6 the behavior of both proposed models. Similar to the previous balanced case study, the recognition performance of TIM and TIM+WE decrease, compared to the original scenario (see Table 5), once the term irony is removed from the corpus. However, in this context we can grasp even more the contribution of WE. In a more complex and real scenario, where the ratio of ironic and not ironic messages is low and the ironic orientation in a sentence can be derived only by the surrounding context, TIM+WE is able to provide a valuable contribution to bridge the semantic gap. If we analyse in details Precision and Recall of both models, we can derive two main observations:

Table 5: Results of our framework on the unbalanced dataset (O).

	P (+)	R (+)	F (+)	P (-)	R (-)	F (-)	Accuracy
TIM	**0,7543**	0,6408	**0,6929**	0,8861	**0,9305**	**0,9078**	**0,8581**
TIM + WE	0,7122	**0,6478**	0,6784	**0,8862**	0,9128	0,8993	0,8466

Table 6: Results of our framework on the unbalanced dataset (S).

	P (+)	R (+)	F (+)	P (-)	R (-)	F (-)	Accuracy
TIM	0,4406	**0,5902**	0,5044	**0,8464**	0,7507	0,7957	0,7107
TIM + WE	**0,5320**	0,4958	**0,5132**	0,8361	**0,8550**	**0,8455**	**0,7654**

- TIM and TIM+WE, although induced in the worst scenario where the dataset is imbalanced and lacks explicit reference to irony, are able to perform better than a trivial classifier that would ensure 70% of accuracy. This makes the proposed models particularly suitable for real world applications.
- TIM+WE obtains Precision and Recall of the same magnitude both for the ironic class (0,5320 for P(+) and 0,4958 for R(+)) and not ironic class (0,8361 for P(-) and 0,8550 for R(-)), compared to TIM which obtains a poor trade-off between the two performance measures (0,4406 for P(+) and 0,5902 for R(+), and 0,8464 for P(-) and 0,7507 for R(-)). This suggests that TIM+WE has good predictive performance characterized by well proportioned abilities both in terms of precision and recall on both ironic and not-ironic orientations.

4.3 Topic Detection Results

In order to perform a qualitative analysis of the obtained results, we report in the following some examples of discovered ironic and not ironic topics. In particular, Table 7 shows a sample of the topics underlying ironic and not-ironic messages derived in the original scenario and in a balanced settings by TIM. In Table 8, the same output is shown for TIM+WE in the simulated scenario again in a balanced settings.

As general remark, the experimental results suggests that the proposed Topic-Irony Model may not only help the irony classification step, but also the ability to identify the underlying topics. In fact, the considered topics are well distinguished by looking at most relevant keywords identified by the proposed approach, still maintaining a good characterization of ironic and not ironic orientations. For instance, the sentence *@user Deeper irony would be Sarah Palin campaigning for literacy*" is correctly classified as ironic and properly related to the topic Politics.

A further remark concerns TIM+WE, and in particular to its ability to deal with short and noisy text. The fact that social network text is composed of few words poses considerable problems when applying traditional probabilistic topic models. These models typically suffer from data sparsity to estimate robust word co-occurrence statistics when dealing with short and ill- formed text. The proposed model is able to reduce the negative impact of short and noisy text in real and complex scenarios thanks its ability to take advantage of distributed representation derived through word embeddings.

TIM+WE is therefore particularly suitable for dealing with those topic-related ironic sentences where the ironic orientation is not explicitly available.

Table 7: Topic-related words are reported in **bold**, while the irony-related ones are marked as <u>underlined</u> . These results are related to TIM in the original scenario (O) and the balanced settings.

humour(-)	humour(+)	politics(-)	politics(+)	education(-)	education(+)
funny	unions	**tcot**	<u>irony</u>	**technology**	<u>irony</u>
posemoticon	workers	**politics**	<u>oh</u>	**education**	<u>**linux**</u>
shoy	benefit	**obama**	<u>get</u>	**new**	**org**
award	<u>**always**</u>	**news**	<u>lol</u>	**apple**	<u>**microsoft**</u>
nominate	cd	**p**	<u>like</u>	**google**	open
lol	movies	**gop**	<u>u</u>	**school**	<u>**tsunami**</u>
humor	labor	**tlot**	<u>**people**</u>	**news**	attack
jokes	<u>**porn**</u>	**teapay**	day	**ipad**	<u>**creates**</u>
joke	fox	**us**	one	posemoticon	sponsors
q	**tv**	**palin**	<u>love</u>	**twitter**	<u>**openmainframe**</u>
comedy	<u>**news**</u>	**iran**	common	via	<u>**gnu**</u>
quote	**playboy**	**pay**	**got**	ac	<u>**religion**</u>
like	**weed**	**sgp**	time	**iphone**	ban
get	marijuana	**iranelection**	posemoticon	**edtech**	thought
one	cannabis	**hcr**	see	**web**	<u>dilemma</u>

Table 8: Topic-related words are reported in **bold**, while the irony-related ones are marked as underlined . These results are related to TIM+WE in the simulated scenario (S) and the balanced settings.

humour(-)	humour(+)	politics (-)	politics(+)	education (-)	education (+)
funny	**quote**	**tcot**	oh	**technology**	**common**
posemoticon	popular	**obama**	u	**education**	postrank
shoy	love	**politics**	lol	**new**	**education**
award	palin	**news**	get	**apple**	**health**
nominate	blind	**p**	like	**google**	**nowplaying**
lol	lingerie	**gop**	day	**news**	make
humor	vote	**tlot**	posemoticon	**school**	lol
jokes	quickpolls	**us**	one	**twitter**	flaker
joke	anonymous	**teapay**	**people**	**ipad**	**cholesterol**
q	com	**pay**	**got**	via	**video**
comedy	voteglobal	**iran**	common	posemoticon	man
quote	gotpolitics	**sgp**	love	ac	difference
like	politics	hcr	yet	**iphone**	causes
one	friends	**iranelection**	**politics**	one	**sense**
get	**barbie**	health	time	**edtech**	fiction

An instance of its ability can be grasped by the following sentence "*catching up on news... see that Pres. Obama's aunt is in the news again, and that she said she loves Pres. Bush.*", where the model correctly classifies the statement as Politics and recognizes as ironic (even if the ironic orientation is not explicitly marked in the text).

5 CONCLUSION

In this paper, we proposed an unsupervised generative model for topic-irony detection, enriched with a neural language lexicon derived through word embeddings. The proposed model has been shown to achieve remarkable results, significantly outperforming existing supervised models currently available in the state of the art.

Concerning the future work, two main research directions will be investigated to improve the generalization abilities of the proposed generative model. First, we would like to overcome the limitation related to the word independence assumption by introducing latent relationships that could exist among different terms and/or sentences. Second, we would like to model parameter switching when dealing with ironic and not ironic statements, in order to set the different level of importance of seed words according to each modeled class.

REFERENCES

Bamman, D. and Smith, N. A. (2015). Contextualized sarcasm detection on twitter. In *Proceedings of the 9th International AAAI Conference on Web and Social Media*, pages 574–77.

Barbieri, F. and Saggion, H. (2014). Modelling irony in twitter. In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 56–64.

Bengio, Y., Schwenk, H., Senécal, J.-S., Morin, F., and Gauvain, J.-L. (2006). Neural probabilistic language models. In *Innovations in Machine Learning: Theory and Applications*, pages 137–186. Springer.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022.

Blitzer, J., Dredze, M., and Pereira, F. (2007). Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Association for Computational Linguistics*, volume 7, pages 440–447.

Bosco, C., Patti, V., and Bolioli, A. (2013). Developing corpora for sentiment analysis: The case of irony and senti-tut. *IEEE Intelligent Systems*, 28(2):55–63.

Colston, H. and Gibbs, R. (2007). A brief history of irony. In *Irony in language and thought: A cognitive science reader*, pages 3–21. Lawrence Erlbaum Assoc Incorporated.

Davidov, D., Tsur, O., and Rappoport, A. (2010). Semi-supervised recognition of sarcastic sentences in twitter and amazon. In *Proceedings of the 14th Conference on Computational Natural Language Learning*, pages 107–116. Association for Computational Linguistics.

Edward, P. C. and Connors, R. (1971). Classical rhetoric for the modern student.

Esuli, A. and Sebastiani, F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation*, volume 6, pages 417–422. Citeseer.

Fersini, E., Pozzi, F. A., and Messina, E. (2015). Detecting irony and sarcasm in microblogs: The role of expressive signals and ensemble classifiers. In *Proceedings of IEEE International Conference on Data Science and Advanced Analytics*, pages 1–8. IEEE.

Ghosh, A., Li, G., Veale, T., Rosso, P., Shutova, E., Barnden, J., and Reyes, A. (2015a). Semeval-2015 task 11: Sentiment analysis of figurative language in twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, pages 470–478.

Ghosh, D., Guo, W., and Muresan, S. (2015b). Sarcastic or not: Word embeddings to predict the literal or sarcastic meaning of words. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1003–1012.

González-Ibánez, R., Muresan, S., and Wacholder, N. (2011). Identifying sarcasm in twitter: a closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 581–586. Association for Computational Linguistics.

Hernández-Farías, I., Benedí, J.-M., and Rosso, P. (2015). Applying basic features from sentiment analysis for automatic irony detection. In *Pattern Recognition and Image Analysis*, pages 337–344. Springer.

Huang, E. H., Socher, R., Manning, C. D., and Ng, A. Y. (2012). Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 873–882. Association for Computational Linguistics.

Jijkoun, V., de Rijke, M., and Weerkamp, W. (2010). Generating focused topic-specific sentiment lexicons. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 585–594. Association for Computational Linguistics.

Jo, Y. and Oh, A. H. (2011). Aspect and sentiment unification model for online review analysis. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 815–824, New York, NY, USA. ACM.

Kaji, N. and Kitsuregawa, M. (2007). Building lexicon for sentiment analysis from massive collection of html documents. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1075–1083, Prague, Czech Republic. Association for Computational Linguistics.

Katz, A. N., Colston, H., and Katz, A. (2005). Discourse and sociocultural factors in understanding nonliteral language. In *Figurative language comprehension: Social and cultural influences*, pages 183–207. Lawrence Erlbaum Associates, Inc. Mahwah, NJ.

Lin, C. and He, Y. (2009). Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, pages 375–384. ACM.

Lu, Y., Castellanos, M., Dayal, U., and Zhai, C. (2011). Automatic construction of a context-aware sentiment lexicon: An optimization approach. In *Proceedings of the 20th International Conference on World Wide Web*, pages 347–356. ACM.

McDonald, S. (1999). Exploring the process of inference generation in sarcasm: A review of normal and clinical studies. *Brain and Language*, 68(3):486–506.

Mei, Q., Ling, X., Wondra, M., Su, H., and Zhai, C. (2007). Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of the 16th International Conference on World Wide Web*, pages 171–180. ACM.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3:1–12.

Mohammad, S., Dunne, C., and Dorr, B. (2009). Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pages 599–608. Association for Computational Linguistics.

Pedersen, T., Patwardhan, S., and Michelizzi, J. (2004). Wordnet::similarity: Measuring the relatedness of concepts. In *Demonstration Papers at HLT-NAACL 2004*, HLT-NAACL–Demonstrations '04, pages 38–41, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ptáček, T., Habernal, I., and Hong, J. (2014). Sarcasm detection on czech and english twitter. In *Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers*, pages 213–223, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Rajadesingan, A., Zafarani, R., and Liu, H. (2015). Sarcasm detection on twitter: A behavioral modeling approach. In *Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, pages 97–106. ACM.

Rao, D. and Ravichandran, D. (2009). Semi-supervised polarity lexicon induction. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 675–682. Association for Computational Linguistics.

Reyes, A. and Rosso, P. (2014). On the difficulty of automatically detecting irony: beyond a simple case of negation. *Knowledge and Information Systems*, 40(3):595–614.

Reyes, A., Rosso, P., and Veale, T. (2013). A multidimensional approach for detecting irony in twitter. *Language resources and evaluation*, 47(1):239–268.

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

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323:533–536.

Turian, J., Ratinov, L., and Bengio, Y. (2010). Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 384–394. Association for Computational Linguistics.

Weitzel, L., Prati, R. C., and Aguiar, R. F. (2016). *The Comprehension of Figurative Language: What Is the Influence of Irony and Sarcasm on NLP Techniques?*, pages 49–74. Springer International Publishing.