How Topic Biases Your Results?
A Case Study of Sentiment Analysis and Irony Detection in Italian

Francesco Barbieri, Francesco Ronzano, Horacio Saggion
Universitat Pompeu Fabra, DTIC, Barcelona, Spain
name.surname@upf.edu

Abstract

In this paper we present our approach to automatically identify the subjectivity, polarity and irony of Italian Tweets. Our system which reaches and outperforms the state of the art in Italian is well adapted for different domains since it uses abstract word features instead of bag of words. We also present experiments carried out to study how Italian Sentiment Analysis systems react to domain changes. We show that bag of words approaches commonly used in Sentiment Analysis do not adapt well to domain changes.

1 Introduction

The automatic identification of sentiments and opinions expressed by users online is a significant and challenging research trend. The task becomes even more difficult when dealing with short and informal texts like Tweets and other microblog texts. Sentiment Analysis of Tweets has been already investigated by several research studies (Jansen et al., 2009; Barbosa and Feng, 2010). Moreover, during the last few years, many evaluation campaigns have been organised to discuss and compare Sentiment Analysis systems tailored to Tweets. Among these campaigns, since 2013, in the context of SemEval (Nakov et al., 2013), several tasks targeting Sentiment Analysis of English Short Texts took place. In 2014, SENTIPOLC (Basile et al., 2014), the SENTIment POLarity Classification Task of Italian Tweets, was organized in the context of EVALITA 2014, the fourth evaluation campaign of Natural Language Processing and Speech tools for Italian. SENTIPOLC distributed a dataset of Italian Tweets annotated with respect to subjectivity, polarity and irony. This dataset enabled training, evaluation and comparison of the systems that participated to the three tasks of SENTIPOLC, respectively dealing with Subjectivity, Polarity and Irony detection. In the Subjectivity task participants were asked to recognise whether a Tweet is objective or subjective, in the Polarity Task they were asked to classify Tweets as positive or negative, and finally, in the Irony Task to detect whether the content of a Tweet is ironic. The following Tweets include an example of each SENTIPOLC class:

- **Objective Tweet:**
  RT @user: Fine primo tempo: #Fiorentina-Juve 0-2 (Tevez, Pogba). Quali sono i vostri commenti sui primi 45 minuti?#ForzaJuve
  (RT @user: First half: #FiorentinaJuve 0-2 (Tevez, Pogba). What are your comments on the first 45 minutes? #GOJUVE)

- **Subjective / Positive / Non-Ironic Tweet:**
  io vorrei andare a votare, ma non penso sia il momento di perder altro tempo e soprattutto denaro.Un governo Monti potrebbe andare. E x voi?
  (I would like to vote, but I do not think it is the moment to waste time and money. Monti's government might work. What do you think?)

- **Subjective / Negative / Ironic Tweet:**
  Brunetta sostiene di tornare a fare l'economista, Mario Monti terrorizzato progetta di mollare tutto ed aprire un negozio di pescheria
  (Brunetta states he will work as an economist again, a terrified Mario Monti plans to leave everything and open a fish shop)

The first example is an objective Tweet as the user only asks what are the opinions on the football match Fiorentina against Juventus. The second Tweet is subjective, positive and non-ironic as the user is giving his positive opinion on

Proceedings of Recent Advances in Natural Language Processing, pages 41–47,
Hissar, Bulgaria, Sep 7–9 2015.

the new government ("Monti's government might work"). The last Tweet is subjective, negative and ironic since the user is making fun of the politician Brunetta (who stated he would work as an economist again), saying that the prime minister Monti is so worried that he is considering to open a fish shop instead of working with Brunetta as an economist.

In this paper we introduce an extended version of the system reported in Barbieri et. al (2014) adding new features that improve our previous results and outperform the best systems presented at SENTIPOLC 2014. We explore the combination of domain independent features (like usage frequency in a reference corpus, number of associated synsets, etc.) and word-based features (like lemmas and bigrams). We employed the supervised algorithm Support Vector Machine (Platt, 1999). Additionally we describe the experiments performed in order to analyse the influence of the topic (politic vs non-politic Tweets) on the results.

The paper is structured in six sections. In the next Section we review the state of the art, in Section 3 we describe dataset and tools used to process Tweet contents, while in Section 4 we introduce the features of our model. In Section 5 we describe our experiments and the performances of our model. In the last two Sections we discuss our results and conclude the paper with future work.

2 Literature Review

The area of Sentiment Analysis includes all those studies that aim to automatically mine opinions and sentiments of the people. Sentiment Analysis became recently the subject of several works, many of them focused on short text (Jansen et al., 2009; Barbosa and Feng, 2010; Bifet et al., 2011; Tumasjan et al., 2010). Some of the best systems for Sentiment Analysis in English also participated to the SemEval shared task (Nakov et al., 2013; Rosenthal et al., 2014). The system that obtained the best performance in the Sentiment Analysis at message level task of Semeval 2013 (Nakov et al., 2013) and 2014 (Rosenthal et al., 2014) mined Twitter to build big sentiment (Mohammad et al., 2013) and emotion lexicons (Mohammad, 2012). Regarding Sentiment Analysis in Italian, the best system (Basile and Novielli, 2014) presented at the 2014 SENTIPOLC shared task used Distributional Semantics. This system took advantage of ten million Tweets split into four classes: subjective, objective, positive and negative ones. Word vectors were created by modelling the contents of the Tweets of each class and exploited to support the classification of new Tweets as belonging to one of these classes.

Since 2010 researchers have been proposing several models to detect irony automatically. Veale and Hao (2010) suggested an algorithm for separating ironic from non-ironic similes in English, detecting common terms used in this ironic comparison, Reyes et. al (Reyes et al., 2013) proposed a model to detect irony in English Tweets, pointing out the relevance of skip-grams (word sequences that contain arbitrary gap) to carry out this task. Barbieri and Saggion (2014) designed an irony detection system that avoided the use of word-based features, employing features like frequency imbalance (rare words in a context of common words) and ambiguity (number of senses of a word). However, irony has not been studied intensively in languages other than English. A few researches have been carried out on irony detection on other languages like Portuguese (Carvalho et al., 2009; de Freitas et al., 2014), Dutch (Liebrecht et al., 2013), Spanish (Barbieri et al., 2015), and Italian (Barbieri et al., 2014). Bosco et. al (2013) collected and annotated tweets in Italian for Sentiment Analysis and Irony detection (the corpus was used for EVALITA 2014).

3 Text Analysis and Tools

In order to process the text of Tweets so as to enable the feature extraction process, we used the same methodology and tools as Barbieri et al. (2014), the reader can find all the details on the tools used in the said paper.

In our experiments we used the dataset employed in SENTIPOLC – the combination SENTITUT (Bosco et al., 2013) and TWITA (Basile and Nissim, 2013)). Each Tweet was annotated over four dimensions: subjectivity/objectivity, positivity/negativity, irony/non-irony, and political/non-political topic. SENTIPOLC dataset is made of a collection of Tweet IDs, since the privacy policy of Twitter does not allow to share the text of Tweets. As a consequence we were able to retrieve by the Twitter API the text of only a subset of the Tweets included in the original SENTIPOLC dataset. In particular, our training set included 3998 Tweets (while the original dataset included 4513).

		Our system	Best of SENTIPOLC
Subjectivity	subjective	0.866	0.828
	objective	0.564	0.601
	avg	**0.715**	0.714
Polarity (POS)	positive	0.554	0.823
	other	0.839	0.527
	avg	**0.697**	0.675
Polarity (NEG)	negative	0.619	0.717
	other	0.741	0.641
	avg	**0.680**	0.679
Irony	ironic	0.260	0.355
	non-ironic	0.916	0.796
	avg	**0.588**	0.576

Table 1: Results of our system and best system of SENTIPOLC in the three Tasks subjectivity, polarity, and irony. We show F-Measures scores for each class and the arithmetic average too.

4 The Model

We extract two kind of features from the Tweets: domain dependent (Section 4.1 and 4.2) and domain independent which are the features proposed in Barbieri et al. (2014). The domain dependent group includes Word-Based and Synsets features described in Section 4.1 and 4.2 often used in text classifications and topic recognition tasks. On the other hand, the domain independent features are not strictly related to the topic of the message. These features are five: Synonyms, Ambiguity, Part Of Speech, Sentiments, Characters.

4.1 Word-Based

We designed this group of features to detect common word-patterns. With these features we are able to capture common phrases used in certain type of Tweet and grasp the common topics that are more frequent in certain type of Tweet (positive/negative/ironic). We computed three word-based features: *lemma* (lemmas of the Tweet), *bi-grams* (combination of two lemmas in a sequence) and *skip one gram* (combination of three lemmas in a row, excluding the one in the middle).

4.2 Synsets

This group of features included features related to WordNet Synsets. After removing stop words, we disambiguated each word against Wordnet (UKB), thus obtaining the most likely sense (Synset) associated to the same word.

5 Experiments and Results

In this Section we show the performance of our system with respect to the three Tasks of SENTIPOLC 2014 (see Table 1). In order to compare our system with the best ones of SENTIPOLC, beside using the same dataset, we adopted the same experimental framework. Since each task was a binary decision (e.g. subjective vs objective), SENTIPOLC organisers computed the arithmetic average of the F-measures of the two classes (e.g. mean of F-Measures of subjective and objective).

We carried out a study of the features contribution to the classification process performing six classification experiments. In each experiment we added to the baseline (domain dependent features) one of the feature groups described in the previous Section. Thus we were able to measure the effect that the addition of the features has on the F-measure.

In Section 5.4 we present an experiment useful to check if our classification features are effective across different domains.

5.1 Task 1: Subjectivity Classification

SENTIPOL 2014 Task 1 was as follows: *given a message, decide whether the message is subjective or objective.*

As we can see in Table 1, in the subjectivity Task our system scored a very similar F-Measure score to the best of SENTIPOLC (0.715 vs 0.714). However, the two systems behave in different ways: our system scored less in the detection of the objective class (0.564 vs 0.601), but it is more accurate in subjective detection (0.866 vs 0.828).

		Subjectivity	Polarity (pos)	Polarity (neg)	Irony
BL	class 1	0.842	0.507	0.509	0.2
	class 2	0.335	0.829	0.720	0.913
	avg	0.589	0.668	0.6145	0.5565
BL + Ambig.	class 1	0.843	0.515	0.529	0.196
	class 2	0.327	0.833	0.716	0.914
	avg	0.585	0.674	0.623	0.555
	improvement	-0.004	0.006	0.008	-0.002
BL + Synset	class 1	0.835	0.514	0.520	0.239
	class 2	0.542	0.82	0.716	0.903
	avg	0.689	0.667	0.618	0.571
	improvement	**0.1**	-0.001	0.004	**0.015**
BL + Senti.	class 1	0.847	0.522	0.578	0.192
	class 2	0.520	0.833	0.731	0.911
	avg	0.684	0.678	0.655	0.552
	improvement	**0.095**	0.010	**0.040**	-0.005
BL + POS	class 1	0.847	0.513	0.542	0.192
	class 2	0.447	0.831	0.717	0.911
	avg	0.647	0.672	0.630	0.552
	improvement	**0.059**	0.004	**0.015**	-0.005
BL + Syno.	class 1	0.843	0.506	0.515	0.195
	class 2	0.322	0.828	0.718	0.913
	avg	0.583	0.667	0.617	0.554
	improvement	-0.006	-0.001	0.002	-0.0025
BL + Char.	class 1	0.832	0.532	0.559	0.212
	class 2	0.463	0.834	0.722	0.914
	avg	0.648	0.683	0.641	0.563
	improvement	**0.059**	0.015	**0.026**	0.007

Table 2: Features Analysis of our system. We add to the baseline (BL) one feature group of our domain independent model per time. We do it for all the four SENTIPOLC Tasks (Subj, Pol(pos), Pol(neg) and irony). In each task, class 1 and 2 are respectively: subjective/objective, positive/non-positive, negative/non-negative and ironic/non-ironic.

In Table 2 we can examine the F-Measure improvement of each feature group. We can note that the greatest improvement is given by Synset and Sentiment features (adding respectively 0.1 and 0.95 points to the baseline); POS and Characters produce an increasing of 0.059, hence can be considered rich features as well. The groups Ambiguity and Synonym do not increase the accuracy of the classification.

5.2 Task 2: Polarity Classification

SENTIPOL 2014 task 2 required *given a message, to decide whether the message is of positive, negative, neutral or mixed sentiment (i.e. conveying both a positive and negative sentiment).*
SENTIPOLC annotators tagged each Tweet with four tags related to polarity: positive, negative, mixed polarity, unspecified. As in SENTIPOLC we split up the Polarity classification in two subclassifications. The first one is the binary classification of positive and mixed-polarity Tweets versus negative and unspecified ones. The second one is focused on the recognition of negative Tweets being the binary decision between negative and mixed polarity versus positive and unspecified tags.

In the positive classification, our system reached a F-Measure of 0.697, while the F-Measure of the best SENTIPOLC system was 0.675 (see Table 1). As previously, the systems behaved differently: ours lacked in detection of the Positive + Mixed-polarity class but it was able to achieve a good F1 in the negative + unspecified class. In the negative classification we out-

Subjectivity	Polarity	Irony
monti	**syn (no, non, neanche)**	governo
syn (no, non, neanche)	**grazie**	passera
governo	monti	politico
syn (avere, costituire, rimanere)	grillo	bersani_non
syn (essere, fare, mettere)	governo	monti
mi	**piacere**	se_governo
paese	**syn (avere, costituire, rimanere)**	grillo
prince	**syn (essere, fare, mettere)**	bersani
essere_dire	paese	capello
of_Persia	**syn (migliaio, mille)**	cavallo

Table 3: For each test set topic the Ten Word-based and Synset features with higher information gain are shown. The domain independent words are in bold. "Syn(word1, word2)" is the synset associated to word1 and word2.

performed the SENTIPOLC system with a score of 0.680 (versus a 0.675). Again, the best SENTIPOLC system got a better score in negative + mixed-polarity and ours reached a better F1 in positive + unspecified.

In the feature analysis (Table 2) we can see that the most important groups of features for the negative classification were Sentiments (giving an improvement of 0.040 points), Characters (0.026) and POS (0.015). On the other hand, in the Positive classification, the word-base features seem to be the most important suggesting that word-patterns were very relevant for this task.

5.3 Task 3: Irony Detection

SENTIPOL 2014 Task 1 asked *given a message, to decide whether the message was ironic or not.* Our system scored a F1 of 0.059 (0.26 in the irony class, and 0.916 in non-irony) while best SENTIPOLC system a F1 of 0.5759 (0.3554 in the irony class and 0.7963 in non-irony). In this Task the use of the words and domain dependent features is very relevant. None of the other domain independent features increase the F1. The only feature that gives a F1 increase is Synset, which can be considered domain dependent. With the help of Table 3 we can note that the ten most important textual features in the irony task are related to a specific topic, and 4 out of 10 words are names of politicians (Passera, Bersani, Monti, Grillo) and the 4 are related to politics (with words like "politics" or "government"). Of course a name of a Politician can not be a good feature for irony detection in general.

5.4 Cross-Domain Experiments

In this section we show the results of the cross-domain experiments. We trained our classifier with the Tweets of one topic (politics related Tweets) and tested the same classifier with the Tweets related to the other topic (non-politics related Tweets). In this way, we can examine whether the model is robust with respect to domain-switches. We were able to run these experiments as SENTIPOLC Tweets provided a topic flag that points out if a Tweet is political or not. We obtained two different systems dividing our features in two groups: domain dependent (word-based and synset group) and domain independent (Sentiment, Synonyms, Character, Ambiguity). We run the cross-domain experiments over the Subjectivity and Polarity datasets with these two systems, and also with our model ("all"). Unfortunately, we were not able to run cross-domain experiments on irony as there were not enough data to effectively train a classifier (e.g. non-political ironic Tweets were only 39 in the test set).

We can see in Table 4 that in the cross-domain experiments domain independent features are five out of six times outperforming the domain dependent system. Moreover an interesting result is that in five out of six combinations the domain independent system outperforms the respective "all" features system, suggesting that when the domain changes, domain dependent features introduce noise.

6 Discussion

Our system outperformed the best SENTIPOLC systems in all the tasks. However, as showed in

		political / non-political	non-political / political
Subjectivity	dom. dependent	0.734	0.672
	dom. indepentent	**0.767**	**0.746**
	all	0.747	0.689
Polarity (POS)	dom. dependent	0.555	0.631
	dom. indepentent	0.443	**0.736**
	all	**0.583**	0.728
Polarity (NEG)	dom. dependent	0.614	0.554
	dom. indepentent	**0.671**	**0.624**
	all	0.663	0.567

Table 4: Cross-domain experiments, where "political / non-political" means training in politics dataset and testing in non-political dataset, "non-political / political" vice-versa. For these two domain combinations we report the results of three models: "domain dependent" (word-based + synset), "domain independent" (Sentiment, Synonyms, Character, Ambiguity), and the model "all" with all the features of our model.

the previous section, not all of our features are effective for the SENTIPOLC Tasks. Specifically, in Polarity and Irony Tasks the features with biggest impact on the classification accuracy resulted to be the domain dependent ones. We can identify two possible explanations. The first one is that for these Tasks is very important to model pattern that are representative of the different classes (for example common phrases used in negative Tweets to detect this class). The second hypothesis is that word-based features, that are often used to model a domain, worked well because training and test set of the dataset shared the same topics. Hence, word-based features worked well because there was a topic bias. For example, in the case of the Polarity Task, a word-based system could detect that often the name of a certain politician is present in the negative Tweets, then using this name as feature to model negative Tweets. With cross-domain experiments we confirmed the second hypothesis, showing that word-based features are not robust when the topic of training and test set are different. On the other hand domain independent features do not decrease their performance when training and test do not share the same topics.

However, in the SENTIPOLC task domain dependent features were relevant, and detecting the topic of a specific class was important. We show (Table 3) that the ten best word-based features are often related to a specific topic (politics in this particular case, see Table 3) rather than to typical expression (e.g. "worst", "don't like" to mean something negative), meaning that our word-based features modelled a specific domain. For example, using words like "Monti" and "Grillo" who are two Italian politicians is important to detect negative Tweets. These features may be in some cases important but they narrow the use of the system to the domain of the training set (and eventually to Tweets generated in the same time-frame).

In the light of these results, we suggest that if a Sentiment Analysis system has to recognise polarity cross-domain should avoid word-based features and focus more on features that are not influenced by the content. On the other hand, if the a Sentiment Analysis system is used in a specific domain, words may have an important role to play.

7 Conclusions

We presented a model for the automatic classification of subjectivity, polarity and recognition of irony in Twitter that outperform the best systems of SENTIPOLC, a shared Task of the EVALITA. Our model included two type of features: domain dependent and domain independent features. We showed with cross-domain experiments that the use of domain dependent feature may constrain a system to work only on a specific domain, while using domain independent features achieved domain independence and a greater robustness when the topic of the Tweet changes.

We are planning to combine the model used in this paper with new distributional semantics based approaches such Basile and Novielli (2014), and to explore new classification techniques like cascade classifiers to combine different classes (e.g. detecting if the Tweet is subjective before deciding if it is ironic, as irony implies subjectivity).

References

Francesco Barbieri and Horacio Saggion. 2014. Modelling Irony in Twitter. In *Proceedings of the EACL Student Research Workshop*, pages 56–64, Gothenburg, Sweden, April. ACL.

Francesco Barbieri, Francesco Ronzano, and Horacio Saggion. 2014. Italian Irony Detection in Twitter: a First Approach. *The First Italian Conference on Computational Linguistics CLiC-it 2014*, page 28.

Francesco Barbieri, Francesco Ronzano, and Horacio Saggion. 2015. Is this tweet satirical? a computational approach for satire detection in spanish. In *Spanish Society for Natural Language Processing*. Alicante, SEPLN.

Luciano Barbosa and Junlan Feng. 2010. Robust Sentiment Detection on Twitter from Biased and Noisy Data. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 36–44. Association for Computational Linguistics.

Valerio Basile and Malvina Nissim. 2013. Sentiment Analysis on Italian Tweets. In *Proceedings of the 4th WASSA Workshop*, pages 100–107.

Pierpaolo Basile and Nicole Novielli. 2014. UNIBA at EVALITA 2014-SENTIPOLC Task: Predicting tweet Sentiment Polarity Combining Micro-Blogging, Lexicon and Semantic Features.

Valerio Basile, Andrea Bolioli, Malvina Nissim, Viviana Patti, and Paolo Rosso. 2014. Overview of the Evalita 2014 SENTIment POLarity Classification Task. In *Proceedings of the 4th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'14)*, Pisa, Italy, December.

Albert Bifet, Geoff Holmes, Bernhard Pfahringer, and Ricard Gavalda. 2011. Detecting Sentiment Change in Twitter Streaming Data.

Cristina Bosco, Viviana Patti, and Andrea Bolioli. 2013. Developing Corpora for Sentiment Analysis and Opinion Mining: the Case of Irony and SENTI-TUT. *Intelligent Systems, IEEE*.

Paula Carvalho, Luís Sarmento, Mário J Silva, and Eugénio de Oliveira. 2009. Clues for Detecting Irony in User-Generated Contents: oh...!! it's so easy;-). In *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, pages 53–56. ACM.

Larissa A de Freitas, Aline A Vanin, Denise N Hogetop, Marco N Bochernitsan, and Renata Vieira. 2014. Pathways for Irony Detection in Tweets. In *Proceedings of the 29th Annual ACM Symposium on Applied Computing*, pages 628–633. ACM.

Andrea Gianti, Cristina Bosco, Viviana Patti, Andrea Bolioli, and Luigi Di Caro. 2012. Annotating Irony in a Novel Italian Corpus for Sentiment Analysis. In *Proceedings of the 4th Workshop on Corpora for Research on Emotion Sentiment and Social Signals, Istanbul, Turkey*, pages 1–7.

Bernard J Jansen, Mimi Zhang, Kate Sobel, and Abdur Chowdury. 2009. Twitter power: Tweets as electronic word of mouth. *Journal of the American society for information science and technology*, 60(11):2169–2188.

Christine Liebrecht, Florian Kunneman, and Antal van den Bosch. 2013. The Perfect Solution for Detecting Sarcasm in tweets# not. *WASSA 2013*, page 29.

Saif M. Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets. In *Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013)*, Atlanta, Georgia, USA, June.

Saif Mohammad. 2012. #Emotional Tweets. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 246–255, Montréal, Canada, 7-8 June. Association for Computational Linguistics.

Preslav Nakov, Zornitsa Kozareva, Alan Ritter, Sara Rosenthal, Veselin Stoyanov, and Theresa Wilson. 2013. Semeval-2013 Task 2: Sentiment Analysis in Twitter.

John Platt. 1999. Fast Training of Support Vector Machines Using Sequential Minimal Optimization. *Advances in kernel methodssupport vector learning*, 3.

Antonio Reyes, Paolo Rosso, and Tony Veale. 2013. A multidimensional Approach for Detecting Irony in Twitter. *Language Resources and Evaluation*, pages 1–30.

Sara Rosenthal, Preslav Nakov, Alan Ritter, and Veselin Stoyanov. 2014. Semeval-2014 Task 9: Sentiment Analysis in Twitter. *Proc. SemEval*.

Andranik Tumasjan, Timm Oliver Sprenger, Philipp G Sandner, and Isabell M Welpe. 2010. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. *ICWSM*, 10:178–185.

Tony Veale and Yanfen Hao. 2010. Detecting Ironic Intent in Creative Comparisons. In *ECAI*, volume 215, pages 765–770.