MIB at SemEval-2016 Task 4a: Exploiting lexicon-based features for sentiment analysis in Twitter

Vittoria Cozza
IIT-CNR
Pisa Italy
v.cozza@iit.cnr.it

Marinella Petrocchi
IIT-CNR
Pisa Italy
m.petrocchi@iit.cnr.it

Abstract

This work presents our team solution for task 4a (Message Polarity Classification) at the SemEval 2016 challenge. Our experiments have been carried out over the Twitter dataset provided by the challenge. We follow a supervised approach, exploiting a SVM polynomial kernel classifier trained with the challenge data. The classifier takes as input advanced NLP features. This paper details the features and discusses the achieved results.

1 Introduction

Revealing the sentiment behind a text is motivated by several reasons, e.g., to figure out how many opinions on a certain topic are positive or negative. Also, it could be interesting to span positivity and negativity across a n-point scale. As an example, a five-point scale is now widespread in digital scenarios where human ratings are involved: Amazon, TripAdvisor, Yelp, and many others, adopt the scale for letting their users rating products and services.

Under the big hat of sentiment analysis (Liu, 2012), polarity recognition attempts to classify texts into positive or negative, while the rating inference task tries to identify different shades of positivity and negativity, e.g., from strongly-negative, to strongly-positive. There currently exists a number of popular challenges on the matter, as those included in the SemEval series on evaluations of computational semantic analysis systems[1]. Both polarity recognition and rating inference have been applied to recommendation systems. Recently, Academia has been focusing on the feasibility to apply sentiment analysis tasks to very short and informal texts, such as tweets (see, e.g. (Rosenthal et al., 2015)).

This paper shows the description of the system that we have set up for participating into the Semeval 2016 challenge in (Nakov et al., 2016b), task 4a (Message Polarity Classification). We have adopted a supervised approach, a SVM polynomial kernel classifier trained with the data provided by the challenge, after extracting lexical and lexicon features from such data.

The paper is organised as follows. Next section briefly addresses related work in the area. Section 3 describes the features extracted from the training data. In Section 4, we present the results of our attempt to answer to the challenge. Finally, we give concluding remarks.

2 Related work

In the last recent years, the Semeval tasks series challenges the polarity evaluation of tweets. This represents a detachment from the traditional polarity detection task. Tweets usually features the use of an informal language, with mispellings, new words, urls, abbreviations and specific symbols (like RT for "re-tweet" and # for hashtags, which are a type of tagging for Twitter messages). Existing approaches and open issues on how to handle such new challenges are in related work like (Kouloumpis et al., 2011; Barbosa and Feng, 2010).

At the 2015 challenge (Rosenthal et al., 2015), the top scored systems were those using deep learning, i.e., semantic vector spaces for single words, used

[1] https://en.wikipedia.org/wiki/SemEval

Proceedings of SemEval-2016, pages 138–143,
San Diego, California, June 16-17, 2016. ©2016 Association for Computational Linguistics

as features in (Turney and Pantel, 2010). Other approaches, as (Basile and Novielli, 2015), exploited lexical and sentiment lexicon features to classify the sentiment of the tweet through machine learning.

In (Priyanka and Gupta, 2013), the authors also exploited different lexical and lexicon features for evaluating the sentiment of a review corpus. The current work inherits most of such features. While they used the lexicon SentiWordNet (Esuli and Sebastiani, 2006), we rely instead on two different ones, LabMT in (Dodds et al., 2011) and SenticNet3.0 presented in (Cambria et al., 2014).

All the above cited lexicons (Cambria et al., 2014; Esuli and Sebastiani, 2006; Dodds et al., 2011) are popular and extensively adopted lexicons for sentiment analysis tasks.

3 Sentiment analysis

The SemEval 2016 Sentiment Analysis challenge (Nakov et al., 2016b) requires the labelling of a test set of 28,481 tweets. In order to facilitate the application of supervised machine learning approaches, the challenge organisers provide the access to a gold dataset: a set of labeled tweets, where the labels - positive, negative or neutral - were manually assigned. In detail, the labeled dataset is divided in a training set of 4,000 tweets and a development set of 2,000 tweets. 340 tweets in the training data and 169 ones in the development data could have not be accessed, since such tweets were“Not Available" at crawling time. We rely on the provided labeled dataset (train + devel) in order to respectively train and evaluate a Support Vector Machine (SVM) classifier (Chang and Lin, 2011) to learn a model for automatic Sentiment Analysis on Twitter. We investigate four groups of features based on: keyword and micro-blogging characteristics, n-grams, negation, and sentiment lexicon.

After evaluating the feature set, we built a new classifier model for annotating the unlabeled test set provided by the challenge (prediction phase, see Figure 1). In this phase, we used as features the best combination of the features previously extracted (actually, all of them) and as the training corpus the overall labeled tweet data (devel+test). The results were not satisfactory, being our team ranked 30 (over 34 teams). The challenge results are reported and discussed in (Nakov et al., 2016b).

In the following, we will detail the process of test cleaning and feature extractions. Then, we present our evaluation, which has been designed to test the efficacy of our feature set for sentiment analysis. Finally, we provide the results obtained at the challenge.

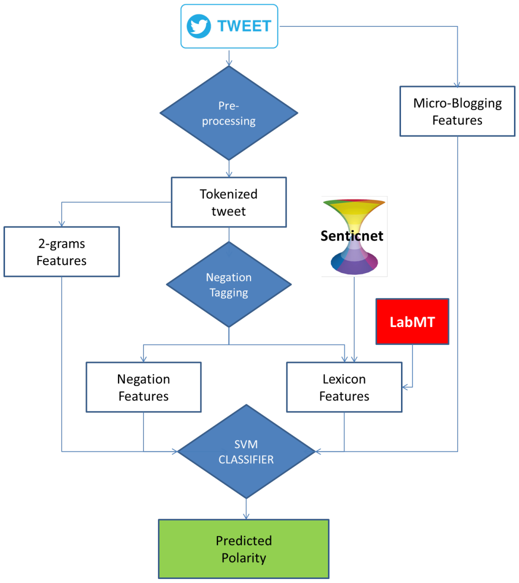


Figure 1: Sentiment analysis: Prediction phase

In the follwoing, we will use the following tweet as a running example.

> Happy hour at @Microsoft #msapc2015 with @sarahvaughan and friends. Good luck for tomorrow’s keynote `http://t.co/emvqoeRS6j`

3.1 Micro-blogging features

The microblogging features have been extracted from the tweet original text, without pre-processing it. We have defined such features with the aim of capturing some typical aspects of micro-blogging. These have been extracted by simply matching regular expressions. First of all, we have cleaned the text from the symbols of mentions “@" and hashtags “#", from urls, from emoticons. Indeed, their presence makes challenging to analyze the text with a traditional linguistic pipeline. Before deleting symbols, emoticons and urls, we have counted them, having as features:

- the number of hashtags;
- the number of mentions;

- the number of urls;
- EmoPos, i.e., the number of positive emoticons;
- EmoNeg, i.e., the number of negative emoticons;

Also, we have also focused on vowels repetitions, exclamations and question marks, introducing the following features:

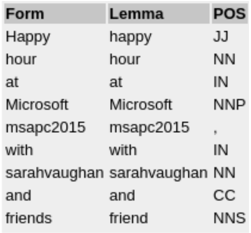
- the number of vowels repetitions;
- the number of question marks and exclamation marks repetitions.

Concerning the marks, we consider a repetition when they are repeated more than once, as in “!!". Instead, we have considered a vowel as repeated when it occurs more than twice, as in “baaad". The positive and negative emoticons we considered are those on the Wikipedia’s page[2].

3.2 Text pre-processing

In order to extract syntactic and semantic features from the text, we pre-processed it with the *Tanl pipeline* (Attardi et al., 2010), a suite of modules for text analytics and natural language processing, based on machine learning. Pre-processing has consisted in first dividing the text in sentences and then into the single word forms composing the sentence. Then, for each form, we have identified the lemma (when available) and the part of speech (POS). As an example, starting from the sentence *Happy hour at Microsoft msapc2015 with sarahvaughan and friends* in the above sample tweet, we obtain the annotation shown in Figure 2. The last column gives the part of speech that a word form yields in a sentence, according to the *Penn Treebank Project*[3].

The last phase of pre-processing is data cleaning. For each sentence, we removed conjunctions, number, determiners, pronouns, and punctuation (still relying on the Penn Treebank POS tags). For the remaining terms, we keep the lemma. Thus, the example sentence results in the following list of lemmas: (*happy, hour, microsoft, msapc2015, sarahvaughan, friend*).



Form	Lemma	POS
Happy	happy	JJ
hour	hour	NN
at	at	IN
Microsoft	Microsoft	NNP
msapc2015	msapc2015	,
with	with	IN
sarahvaughan	sarahvaughan	NN
and	and	CC
friends	friend	NNS

Figure 2: Annotation with Tanl English Pipeline (http://tanl.di.unipi.it/en/)

In the following, we describe the features we have extracted from the pre-processed text. Among others, we inherit some of the features in (Priyanka and Gupta, 2013) and (Basile and Novielli, 2015), which face with sentiment analysis on Twitter.

3.3 n-grams features

Upon pre-processing, we have obtained a words vector representation of each tweet. Then, we have extracted n-grams, i.e., all the pairs of sequencing lemmas in the vector. As an over simplification, we have considered only the case of n=2. We thought this was reasonable, since tweets are short portions of text bounded to 140 characters. In the example sentence, some are (*happy-hour, hour-microsoft*).The 2-grams have been discretised into binary attributes representing their presence or not in the text. There are 1,237 unique 2-grams.

3.4 Negation-based features

Handling negations is an important step in sentiment analysis, as they can reverse the meaning of a sentence. Also, negations can often occur with sarcastic and ironic goals, which are quite difficult to detect. We consider 1-grams and we prefix them with P (N) when they are asserted (negated). To identify if the unigram appears in a negated scope, we have applied a rule-based approach[4]. The approach works as follows. Considering a negative sentiment tweet, like, e.g., “*It might be not nice but it’s the reality.*, the “nice" unigram is in the scope of negation, and, thus, it will be labeled as N_nice. The “but" unigram changes again the scope, thus “reality" will be

[2] `https://it.wikipedia.org/wiki/Emoticon`

[3] `https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html`

[4] `http://sentiment.christopherpotts.net/lingstruc.html#negation`

labeled as P_reality.

We have identified the following features as suitable for handling negations:

- Unigrams with scope;
- Positiveterms: Number of lemmas with positive scope;
- Negativeterms: Number of lemmas with negative scope;

The first feature has been discretised into binary attributes representing the presence (or not) of the 1-gram. The number of unique unigrams (with scope) are 5,110.

3.5 Sentiment lexicon-based features

Several lexicons are available for sentiment analysis. In this work, we consider SenticNet 3.0 (Cambria et al., 2014) and the LabMT (Dodds et al., 2011).

SenticNet 3.0 is a large concept-level base of knowledge, assigning semantics, sentics, and polarity to 30,000 natural language concepts. In particular, polarity is a floating number between -1 (extreme negativity) and +1 (extreme positivity) [5]. We rely on SenticNet 3.0 to compute features based on polarity, according to the SenticNet lexicon:

- Min, max, average and standard deviation polarity of lemmas;
- PA Positive Asserted: number of lemmas with a positive polarity, e.g., “good";
- PN Positive Negated: number of lemmas with a positive polarity, but negated, e.g., “not good";
- NA Negative Asserted: n. lemmas with a negative polarity (e.g., “bad");
- NN Negative Negated: n. lemmas and with a negative polarity, and negated (e.g., “not bad").

To assign the polarity to “not good", we consider the polarity of “good" in the SenticNet lexicon (0.667) and we revert it, assigning -0.667.

Also, SenticNet provides the polarity score to complex expressions. As an example, the popular idiomatic expression “32 teeth" obtains a polarity score of 0.903. Thus, beside unigrams polarity scores, we have also considered 2-grams polarity scores.

Class	Precision	Recall	F1
negative	0.16	0.17	0.16
neutral	0.46	0.27	0.34
positive	0.50	0.75	0.60
avg / total	0.42	0.43	0.51

Table 2: MIB results (Tweets 2016 - dev, all feats)

Since not all the lemmas in the dataset were covered by the SenticNet lexicon, we have enlarged the covering by relying on LabMT (Dodds et al., 2011). LabMT is a list of words, manually labeled with a sentiment score through crowdsourcing. In particular, we considered the happiness score. This value ranges over 1 and 9 (1 is very unhappy, while 9 absolutely happy). We have normalised such values to range over -1 and 1, using the linear function y=(x-5)/4.

4 Evaluation and results

We have preliminarily built a prediction model trained with the 2016 challenge data, in details we have used the train and the devel data, respectively for training and evaluation. The prediction model is based on an SVM linear kernel classifier. For the experiments, the classifier has been implemented through sklearn[6] in Python. We have used a linear classifier suitable for handling unbalanced data: SGDClassifier with default parameters[7]. The model exploits the four groups of features presented in Section 3. Upon extracting the features from the training dataset, we obtained 6,547 features. In the following, we will show some feature ablation experiments, each of them corresponds to remove one category of features from the full set. Results are in terms of Precision and Recall, see Table 1.

The features evaluation shows that we do not have a set of dominant features group, leading to a not satisfying discrimination among positive, negative, and neutral tweets. Ablation tests show that negation-based features are the most relevant ones. Polarity lexicon features are influential to identify the nega-

[5] http://sentic.net/

[6] http://scikit-learn.org/stable/

[7] http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html

System	Negative			Neutral			Positive			Average
	P	R	F1	P	R	F1	P	R	F1	F1
All but n-grams feats	0.33	0.24	0.28	0.45	0.22	0.30	0.41	0.72	0.53	0.38
All but negation-based feats	0.42	0.08	0.13	0.52	0.06	0.10	0.39	0.97	0.56	0.28
All but polarity lexicon feats	0.34	0.10	0.15	0.44	0.42	0.43	0.42	0.58	0.48	0.40

Table 1: Ablation tests

tive class, but, overall, less than we have expected.

In (Cozza et al., 2016), the authors have proposed a similar approach to the one here presented. The aim was to evaluate the sentiment of a large set of online reviews. In online reviews, the textual opinion is usually accompanied by a numerical score, and sentiment analysis could be a valid alley for identifying misalignment between the score and the satisfaction expressed in the text. Work in (Cozza et al., 2016) shows that the features' set was discriminant for evaluating the sentiment of the reviews. In part, this would support the thesis that standard sentiment analysis approaches are more suitable for "literary" texts than for short, informal texts featured by tweets.

It is worth noting that the lexicons we rely on are based on lemmas, while there exist other lexicons that consider also the part of speech, see, e.g., SentiWordNet (Esuli and Sebastiani, 2006). Let the reader consider the following tweet, from the SemEval 2016 training set:

> #OnThisDay1987 CBS records shipped out the largest preorder in the company's history for Michael Jackson's album Bad `http://t.co/v4fkyOx2eW`

In this example, the word "Bad" should not be considered as a negative adjective, since it is an album name. However, in the current work, we have not discriminated between nouns and adjectives with same spelling.

4.1 Challenge results

The results over the challenge test set are available on the SemEval website[8], according to the challenge score system described in (Nakov et al., 2016a). Table 3 shows the comparison of our results with the ones of the winning team. In the submitted result, the classifier has been trained over the training + development dataset, annotated with the best combination of features, as analyzed before.

team	score
SwissCheese	63.301
MIB	40.10

Table 3: SemEval 2016 task 4a results (Tweet 2016)

5 Conclusions

The approach proposed in this work achieved unsatisfactory results. This was in part due to a data preprocessing phase and a feature extraction phase that do not consider characteristics intrinsic to microblogging. Indeed, we mostly dealt with tweets handling them as regular text. The challenge data have been preprocessed by supervised approached, where features have been extracted through a NLP pipeline, trained on newswire domain. Within our proposed features, the sentiment lexicon-based features has proved to work well. However, we believe their extraction could take advantage of the adoption of other lexicons, different to those we have relied on. Specifically, there exist lexicons trained over tweets, such as the NCR emotion lexicon (Mohammad and Turney, 2013). Finally, we expect that a better solution could be achieved by extending the approach to include features extracted by unsupervised approaches (word embeddings), or by adopting a deep learning classifier, instead of a linear one.

References

Giuseppe Attardi, Stefano Dei Rossi, and Maria Simi. 2010. The TANL pipeline. *Web Services and Processing Pipelines in HLT: Tool Evaluation, LR Production and Validation (LREC:WSSP).*

Luciano Barbosa and Junlan Feng. 2010. Robust sentiment detection on Twitter from biased and noisy data. In *23rd International Conference on Computational Linguistics: Posters*, pages 36–44. Association for Computational Linguistics.

P. Basile and N. Novielli. 2015. UNIBA: Sentiment analysis of English tweets combining micro-blogging, lexicon and semantic features. In *9th International Work-*

[8] `http://alt.qcri.org/semeval2016/task4/data/uploads/semeval2016_task4_results.pdf`

shop on Semantic Evaluation (SemEval 2015), pages 595–600. Association for Computational Linguistics.

Erik Cambria, Daniel Olsher, and Dheeraj Rajagopal. 2014. SenticNet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis. In *28th AAAI Conference on Artificial Intelligence*, pages 1515–1521.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: a library for support vector machines. *Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27.

Vittoria Cozza, Marinella Petrocchi, and Angelo Spognardi. 2016. Write a number in letters: A study on text-score disagreement in online reviews.

Peter Sheridan Dodds, Kameron Decker Harris, Isabel M. Kloumann, Catherine A. Bliss, and Christopher M. Danforth. 2011. Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *PLoS ONE*, 6(12):e26752, 12.

Andrea Esuli and Fabrizio Sebastiani. 2006. SENTIWORDNET: A publicly available lexical resource for opinion mining. In *5th Conference on Language Resources and Evaluation*, pages 417–422.

Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. 2011. Twitter sentiment analysis: The good the bad and the omg. In *AAAI Conference on Weblogs and Social*.

Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers, May.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. 2016a. Evaluation measures for the semeval-2016 task 4 sentiment analysis in Twitter.

Preslav Nakov, Alan Ritter, Sara Rosenthal, Veselin Stoyanov, and Fabrizio Sebastiani. 2016b. SemEval-2016 task 4: Sentiment analysis in Twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, California, June. Association for Computational Linguistics.

C. Priyanka and D. Gupta. 2013. Identifying the best feature combination for sentiment analysis of customer reviews. In *Advances in Computing, Communications and Informatics (ICACCI)*, pages 102–108.

Sara Rosenthal, Preslav Nakov, Svetlana Kiritchenko, Saif Mohammad, Alan Ritter, and Veselin Stoyanov. 2015. SemEval-2015 task 10: Sentiment analysis in Twitter. In *9th International Workshop on Semantic Evaluation*, pages 451–463. Association for Computational Linguistics, June.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *CoRR*, abs/1003.1141.