SiTAKA at SemEval-2017 Task 4: Sentiment Analysis in Twitter Based on a Rich Set of Features

Mohammed Jabreel **Antonio Moreno**
Intelligent Technologies for Advanced Knowledge Acquisition (ITAKA),
Departament d'Enginyeria Informàtica i Matemàtiques,
Universitat Rovira i Virgili,
Av. Països Catalans, 26, 43007 Tarragona, Spain
<first_name>.<last_name>@urv.cat

Abstract

This paper describes SiTAKA, our system that has been used in task 4A, English and Arabic languages, Sentiment Analysis in Twitter of SemEval2017. The system proposes the representation of tweets using a novel set of features, which include a bag of negated words and the information provided by some lexicons. The polarity of tweets is determined by a classifier based on a Support Vector Machine. Our system ranks 2nd among 8 systems in the Arabic language tweets and ranks 8th among 38 systems in the English-language tweets.

1 Introduction

Sentiment analysis in Twitter is the problem of identifying people's opinions expressed in tweets. It normally involves the classification of tweets into categories such as positive, negative and in some cases, neutral. The main challenges in designing a sentiment analysis system for Twitter are the following:

- Twitter limits the length of the message to 140 characters, which leads users to use novel abbreviations and often disregard standard sentence structures.
- The informal language and the numerous spelling errors.

Most of the existing systems are inspired by the work presented in (Pang et al., 2002). Machine Learning techniques have been used to build a classifier from a set of tweets with a manually annotated sentiment polarity. The success of the Machine Learning models is based on two main facts: a large amount of labeled data and the intelligent design of a set of features that can distinguish between the positive, negative and neutral samples.

With this approach, most studies have focused on designing a set of efficient features to obtain a good classification performance (Feldman, 2013; Liu, 2012; Pang and Lee, 2008). For instance, the authors in (Mohammad et al., 2013) used diverse sentiment lexicons and a variety of hand-crafted features.

This paper proposes the representation of tweets using a novel set of features, which include the information provided by seven lexicons and a bag of negated words (BonW). The concatenation of these features with a set of basic features improves the classification performance. The polarity of tweets is determined by a classifier based on a Support Vector Machine.

The system has been evaluated on the Arabic and English language test sets of the Twitter Sentiment Analysis Track in SemEval 2017, subtask A (Message Polarity Classification). Our system (SiTAKA) has been ranked 8th over 38 teams in the English language test set and 2nd out of 8 teams in the Arabic language test set.

The rest of the paper is structured as follows. Section 2 presents the tools and the resources that have been used. In Section 3 we describe the system. The experiments and results are presented and discussed in Section 4. Finally, in the last section the conclusions as well as further work are presented.

2 Resources

This section explains the tools and the resources that have been used in the SiTAKA system. Let us denote its Arabic language and English language versions by *Ar*-SiTAKA and *En*-SiTAKA, respectively.

Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017), pages 694–699,
Vancouver, Canada, August 3 - 4, 2017. ©2017 Association for Computational Linguistics

2.1 Sentiment Lexicons

2.1.1 *En*-SiTAKA Lexicons

We used for *En*-SiTAKA five lexicons in this work, namely: General Inquirer (Stone et al., 1968), Hu-Liu opinion lexicon (HL) (Hu and Liu, 2004), NRC hashtags lexicon (Mohammad et al., 2013), SenticNet (Cambria et al., 2014), and TS-Lex (Tang et al., 2014b). More details about each lexicon, such as how it was created, the polarity score for each term, and the statistical distribution of the lexicon, can be found in (Jabreel and Moreno, 2016).

2.1.2 *Ar*-SiTAKA Lexicons

In this version of the SiTAKA system, we used four lexicons created by (Saif M. Mohammad and Kiritchenko, 2016): Arabic Hashtag Lexicon, Dialectal Arabic Hashtag Lexicon, Arabic Bing Liu Lexicon and Arabic Sentiment140 Lexicon. The first two were created manually, whereas the rest were translated to Arabic from the English version using Google Translator.

2.2 Embeddings

We used two pre-trained embedding models in *En*-SiTAKA. The first one is word2vec which is provided by Google. It is trained on part of the Google News dataset (about 100 billion words) and it contains 300-dimensional vectors for 3M words and phrases (Mikolov et al., 2013b). The second one is SSWEu, which has been trained to capture the sentiment information of sentences as well as the syntactic contexts of words (Tang et al., 2014c). The SSWEu model contains 50-dimensional vectors for 100K words.

In *Ar*-SiTAKA we used the model Arabic-SKIP-G300 provided by (Zahran et al., 2015). Arabic-SKIP-G300 has been trained on a large corpus of Arabic text collected from different sources such as Arabic Wikipedia, Arabic Gigaword Corpus, Ksucorpus, King Saud University Corpus, Microsoft crawled Arabic Corpus, etc. It contains 300-dimensional vectors for 6M words and phrases.

3 System Description

This section explains the main steps of the SiTAKA system, the features used to describe a tweet and the classification method.

3.1 Preprocessing and Normalization

Some standard pre-processing methods are applied on the tweets:

- *Normalization*: Each tweet in English is converted to the lowercase. URLs and usernames are omitted. Non-Arabic letters are removed from each tweet in the Arabic-language sets. Words with repeated letters (i.e. elongated) are corrected.

- *Tokenization and POS tagging*: All English-language tweets are tokenized and tagged using Ark Tweet NLP (Gimpel et al., 2011), while all Arabic-language tweets are tokenized and tagged using Stanford Tagger (Green and Manning, 2010).

- *Negation*: A negated context can be defined as a segment of tweet that starts with a negation word (e.g. *no*, *don't* for English-language, لا و ليس for Arabic-language) and ends with a punctuation mark (Pang et al., 2002). Each tweet is negated by adding a suffix ("_NEG" and "_منفي") to each word in the negated context.

  It is necessary to mention that in *Ar*-SiTAKA we did not use all the Arabic negation words due to the ambiguity of some of them. For example, the first word ما, is a question mark in the following ”ما رأيك في ما حدث؟-What do you think about what happened?” and it means ”which/that” in the following example ”إن ما حدث اليوم سيء جدا- The matter that happened today was very bad”.

As shown in (Saif et al., 2014), stopwords tend to carry sentiment information; thus, note that they were not removed from the tweets.

3.2 Features Extraction

SiTAKA uses five types of features: *basic text*, *syntactic*, *lexicon*, *cluster* and *Word Embeddings*. These features are described in the following subsections:

3.2.1 Basic Features

These basic features are extracted from the text. They are the following:

Bag of Words (BoW): Bag of words or n-grams features introduce some contextual information.

The presence or absence of contiguous sequences of 1, 2, 3, and 4 tokens are used to represent the tweets.

Bag of Negated Words (BonW): Negated contexts are important keys in the sentiment analysis problem. Thus, we used the presence or absence of contiguous sequences of 1, 2, 3 and 4 tokens in the negated contexts as a set of features to represent the tweets.

3.2.2 Syntactic Features

Syntactic features are useful to discriminate between neutral and non-neutral texts.

Part of Speech (POS): Subjective and objective texts have different POS tags (Pak and Paroubek, 2010). According to (Zhou et al., 2014), non-neutral terms are more likely to exhibit the following POS tags in Twitter: nouns, adjectives, adverbs, abbreviations and interjections. The number of occurrences of each part of speech tag is used to represent each tweet.

Bi-tagged: Bi-tagged features are extracted by combining the tokens of the bi-grams with their POS tag e.g. ”*feel*_VBP *good*_JJ” ”جميل_JJ جداً_VBD”. It has been shown in the literature that adjectives and adverbs are subjective in nature and they help to increase the degree of expressiveness (Agarwal et al., 2013; Pang et al., 2002).

3.2.3 Lexicon Features

Opinion lexicons play an important role in sentiment analysis systems, and the majority of the existing systems rely heavily on them (Rosenthal et al., 2014). For each of the chosen lexicons, a tweet is represented by calculating the following features: (1) *tweet polarity*, (2) *the average polarity of the positive terms*, (3) *the average polarity of the negative terms*, (4) *the score of the last positive term*, (5) *the score of the last negative term*, (6) *the maximum positive score* and (7) *the minimum negative score*.

The polarity of a tweet *T* given a lexicon *L* is calculated using the equation (1). First, the tweet is tokenized. Then, the number of positive (*P*) and negative (*N*) tokens found in the lexicon are counted. Finally, the polarity measure is calculated as follows:

$$polarity = \begin{cases} 1 - \frac{N}{P} & ; if P > N \\ 0 & ; if P = N \\ \frac{P}{N} - 1 & ; if P < N \end{cases} \qquad (1)$$

3.2.4 Cluster Features

We used two set of clusters in *En*-SiTAKA to represent the English-language tweets by mapping each tweet to a set of clusters. The first one is the well known set of clusters provided by the Ark Tweet NLP tool which contains 1000 clusters produced with the Brown clustering algorithm from 56M English-language tweets. These 1000 clusters are used to represent each tweet by mapping each word in the tweet to its cluster. The second one is *Word2vec cluster ngrams*, which is provided by (Dong et al., 2015). They used the word2vec tool to learn 40-dimensional word embeddings of 255,657 words from a Twitter dataset and the K-means algorithm to cluster them into 4960 clusters. We were not able to find publicly available semantic clusters to be used in *Ar*-SiTAKA.

3.2.5 Embedding Features

Word embeddings are an approach for distributional semantics which represents words as vectors of real numbers. Such representation has useful clustering properties, since the words that are semantically and syntactically related are represented by similar vectors (Mikolov et al., 2013a). For example, the words ”coffee” and ”tea” will be very close in the created space.

We used *sum*, *standard-deviation*, *min* and *max* pooling functions (Collobert et al., 2011) to obtain the tweet representation in the embedding space. The result is the concatenation of vectors derived from different pooling functions. More formally, let us consider an embedding matrix $E \in \mathbb{R}^{d \times |V|}$ and a tweet $T = w_1, w_2, ..., w_n$, where d is the dimension size, $|V|$ is the length of the vocabulary (i.e. the number of words in the embedding model), w_i is the ith word in the tweet and n is the number of words. First, each word w_i is substituted by the corresponding vector v_i^j in the matrix E where j is the index of the word w_i in the vocabulary. This step ends with the matrix $W \in \mathbb{R}^{d \times n}$. The vector $V_{T,E}$ is computed using the following formula:

$$V_{T,E} = \bigcup_{pool \in \{max, min, sum, std\}} pool_{i=1}^{n} v_i \qquad (2)$$

where $\bigcup$ denotes the concatenation operation. The pooling function is an element-wise function, and it converts texts with various lengths into a fixed-length vector allowing to capture the information throughout the entire text.

3.3 Classifier

Up to now, Support Vector Machines (SVM) (Cortes and Vapnik, 1995) have been used widely and reported as the best classifier in the sentiment analysis problem. Thus, we trained a SVM classifier on the training sets provided by the organizers. For the English-language we combined the training sets of SemEval 13-16 and testing sets of SemEval 13-15, and used them as a training set. Table 1 shows the numerical description of the datasets used in this work. We used the linear kernel with the value 0.5 for the cost parameter C. All the parameters and the set of features have been experimentally chosen based on the development sets.

System	Training set	Dev set
En-SiTAKA	27,700	20,632
Ar-SiTAKA	2684	671

Table 1: Numerical description of the set of tweets

4 Results

The evaluation metrics used by the task organizers were the macroaveraged recall (ρ), the F1 averaged across the positives and the negatives $F1^{PN}$ and the accuracy (Acc) (Rosenthal et al., 2017).

The system has been tested on 12,284 English-language tweets and 6100 Arabic-language tweets provided by the organizers. The golden answers of all the test tweets were omitted by the organizers. The official evaluation results of our system are reported along with the top 10 systems and the baseline results in Table 2 and 3. Our system ranks 8th among 38 systems in the English-language tweets and ranks 2nd among 8 systems in the Arabic language tweets. The baselines 1, 2 and 3 stand for the cases in which the system classifies all the tweets as positive, negative and neutral respectively.

#	System	ρ	$F1^{PN}$	Acc
1	DataStories	$\mathbf{0.681}_1$	0.677_2	0.651_5
	BB_twtr	$\mathbf{0.681}_1$	0.685_1	0.658_3
3	LIA	$\mathbf{0.676}_3$	0.674_3	0.661_2
4	Senti17	$\mathbf{0.674}_4$	0.665_4	0.652_4
5	NNEMBs	$\mathbf{0.669}_5$	0.658_5	0.664_1
6	Tweester	$\mathbf{0.659}_6$	0.648_6	0.648_6
7	INGEOTEC	$\mathbf{0.649}_7$	0.645_7	0.633_{11}
8	***En*-SiTAKA**	$\mathbf{0.645}_8$	0.628_9	0.643_9
9	TSA-INF	$\mathbf{0.643}_9$	0.620_{11}	0.616_{17}
10	UCSC-NLP	$\mathbf{0.642}_{10}$	0.624_{10}	0.565_{30}
	baseline 1	0.333	0.162	0.193
	baseline 2	0.333	0.224	0.323
	baseline 3	0.333	0.00	0.483

Table 2: Results for SemEval-2017 Task 4, subtask A, English.

#	System	ρ	$F1^{PN}$	Acc
1	NileTMRG	$\mathbf{0.583}_1$	0.610_1	0.581_1
2	***Ar*-SiTAKA**	$\mathbf{0.550}_2$	0.571_2	0.563_2
3	ELiRF-UPV	$\mathbf{0.478}_3$	0.467_4	0.508_3
4	INGEOTEC	$\mathbf{0.477}_4$	0.455_5	0.499_4
5	OMAM	$\mathbf{0.438}_5$	0.422_6	0.430_8
	LSIS	$\mathbf{0.438}_5$	0.469_3	0.445_6
7	1w-StAR	$\mathbf{0.431}_7$	0.416_7	0.454_5
8	HLP@UPENN	$\mathbf{0.415}_8$	0.320_8	0.443_7
	baseline 1	0.333	0.199	0.248
	baseline 2	0.333	0.267	0.364
	baseline 3	0.333	0.00	0.388

Table 3: Results for SemEval-2017 Task 4, subtask A, Arabic.

5 Conclusion

We have presented a new set of rich sentimental features for the sentiment analysis of the messages posted on Twitter. A Support Vector Machine classifier has been trained using a set of basic features, information extracted from a set of useful and publicly available opinion lexicons, syntactic features, clusters and embeddings. Deep learning approaches have recently been used to build supervised, unsupervised or even semi-supervised methods to analyze the sentiment of texts and to build efficient opinion lexicons (Severyn and Moschitti, 2015; Tang et al., 2014a,c); thus, the authors are considering the possibility of also using this technique to build a sentiment analysis system.

Acknowledgment

This work was partially supported by URV Research Support Funds (2015PFR-URV-B2-60, 2016PFR-URV-B2-60 and Martí i Franqués PhD grant).

References

Basant Agarwal, Natasha Mittal, and Erik Cambria. 2013. Enhancing sentiment classification performance using bi-tagged phrases. In *Data Mining Workshops (ICDMW), 2013 IEEE 13th International Conference on*. IEEE, pages 892–895.

Erik Cambria, Daniel Olsher, and Dheeraj Rajagopal. 2014. SenticNet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis. In *Twenty-eighth AAAI conference on artificial intelligence*.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research* 12:2461–2505.

Corinna Cortes and Vladimir Vapnik. 1995. Support vector machine. *Machine learning* 20(3):273–297.

Li Dong, Furu Wei, Yichun Yin, Ming Zhou, and Ke Xu. 2015. Splusplus: A Feature-Rich Two-stage Classifier for Sentiment Analysis of Tweets. *SemEval-2015* page 515.

Ronen Feldman. 2013. Techniques and applications for sentiment analysis. *Communications of the ACM* 56(4):82–89.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech Tagging for Twitter: Annotation, Features, and Experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*. Association for Computational Linguistics, Stroudsburg, PA, USA, HLT '11, pages 42–47.

Spence Green and Christopher D Manning. 2010. Better arabic parsing: Baselines, evaluations, and analysis. In *Proceedings of the 23rd International Conference on Computational Linguistics*. Association for Computational Linguistics, pages 394–402.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pages 168–177.

Mohammed Jabreel and Antonio Moreno. 2016. Sentirich: Sentiment analysis of tweets based on a rich set of features. In *Artificial Intelligence Research and Development - Proceedings of the 19th International Conference of the Catalan Association for Artificial Intelligence, Barcelona, Catalonia, Spain, October 19-21, 2016*. pages 137–146.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies* 5(1):1–167.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* .

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. pages 3111–3119.

Saif 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.

Alexander Pak and Patrick Paroubek. 2010. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA), Valletta, Malta.

Bo Pang and Lillian Lee. 2008. Opinion Mining and Sentiment Analysis. *Found. Trends Inf. Retr.* 2(1-2):1–135.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs Up?: Sentiment Classification Using Machine Learning Techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*. Association for Computational Linguistics, Stroudsburg, PA, USA, EMNLP '02, pages 79–86.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, Vancouver, Canada, SemEval '17.

Sara Rosenthal, Alan Ritter, Preslav Nakov, and Veselin Stoyanov. 2014. Semeval-2014 task 9: Sentiment analysis in twitter. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. pages 73–80.

Hassan Saif, Miriam Fernández, Yulan He, and Harith Alani. 2014. On stopwords, filtering and data sparsity for sentiment analysis of Twitter. In *Proceedings of the 9th language resources and evaluation conference (LREC)*. Reykjavik, Iceland.

Mohammad Salameh Saif M. Mohammad and Svetlana Kiritchenko. 2016. Sentiment lexicons for arabic social media. In *Proceedings of 10th edition of the the Language Resources and Evaluation Conference (LREC)*. Portorož, Slovenia.

Aliaksei Severyn and Alessandro Moschitti. 2015. UNITN: Training deep convolutional neural network for Twitter sentiment classification. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Association for Computational Linguistics, Denver, Colorado*. pages 464–469.

Philip Stone, Dexter C Dunphy, Marshall S Smith, and DM Ogilvie. 1968. The general inquirer: A computer approach to content analysis. *Journal of Regional Science* 8(1):113–116.

Duyu Tang, Furu Wei, Bing Qin, Ting Liu, and Ming Zhou. 2014a. Coooolll: A Deep Learning System for Twitter Sentiment Classification. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. Association for Computational Linguistics and Dublin City University, Dublin, Ireland, pages 208–212. http://www.aclweb.org/anthology/S14-2033.

Duyu Tang, Furu Wei, Bing Qin, Ming Zhou, and Ting Liu. 2014b. Building Large-Scale Twitter-Specific Sentiment Lexicon : A Representation Learning Approach. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin City University and Association for Computational Linguistics, Dublin, Ireland, pages 172–182. http://www.aclweb.org/anthology/C14-1018.

Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014c. Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Baltimore, Maryland, pages 1555–1565. http://www.aclweb.org/anthology/P14-1146.

Mohamed A Zahran, Ahmed Magooda, Ashraf Y Mahgoub, Hazem Raafat, Mohsen Rashwan, and Amir Atyia. 2015. Word representations in vector space and their applications for arabic. In *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, pages 430–443.

Zhixin Zhou, Xiuzhen Zhang, and Mark Sanderson. 2014. Sentiment analysis on twitter through topic-based lexicon expansion. In *Databases Theory and Applications*, Springer, pages 98–109.