Tweester at SemEval-2016 Task 4: Sentiment Analysis in Twitter using Semantic-Affective Model Adaptation

Elisavet Palogiannidi[2,3,1], **Athanasia Kolovou**[4,1], **Fenia Christopoulou**[1], **Filippos Kokkinos**[1], **Elias Iosif**[1,3], **Nikolaos Malandrakis**[5], **Harris Papageorgiou**[3], **Shrikanth Narayanan**[5,6], **Alexandros Potamianos**[1,3,6]

[1] School of ECE, National Technical University of Athens, Zografou 15780, Athens, Greece
[2] School of ECE, Technical University of Crete, Chania 73100, Crete, Greece
[3] "Athena" Research and Innovation Center, Maroussi 15125, Athens, Greece
[4] Department of Informatics, University of Athens, Athens, Greece
[5] Signal Analysis and Interpretation Laboratory (SAIL), USC, Los Angeles, CA 90089, USA
[6] Behavioral Informatix, Los Angeles, CA 90089, USA

epalogiannidi@isc.tuc.gr, akolovou@di.uoa.gr, el10136@mail.ntua.gr
el11142@mail.ntua.gr, iosife@central.ntua.gr, malandra@usc.edu
xaris@ilsp.athena-innovation.gr, shri@sipi.usc.edu, potam@central.ntua.gr

Abstract

We describe our submission to SemEval2016 Task 4: Sentiment Analysis in Twitter. The proposed system ranked first for the subtask B. Our system comprises of multiple independent models such as neural networks, semantic-affective models and topic modeling that are combined in a probabilistic way. The novelty of the system is the employment of a topic modeling approach in order to adapt the semantic-affective space for each tweet. In addition, significant enhancements were made in the main system dealing with the data preprocessing and feature extraction including the employment of word embeddings. Each model is used to predict a tweet's sentiment (positive, negative or neutral) and a late fusion scheme is adopted for the final decision.

1 Introduction

Nowadays, the usage of social networks such as Twitter dominates the daily communication of hundreds of millions of people around the world. People often share opinions and express their feelings about various topics through social networks. Tasks such as opinion mining and sentiment analysis (Pang and Lee, 2008) have become very popular, since they can capture a large portion of the public opinion.

The sentiment analysis of tweets was applied in various domains, such as commerce (Jansen et al., 2009), disaster management (Verma et al., 2011) or health (Chew and Eysenbach, 2010). This task is especially challenging due to the terse and informal writing style, the semantic diversity of content, as well as the often "unconventional" grammar and orthography. Many computational systems like those submitted to SemEval 2015 Task 10 (Rosenthal et al., 2015) incorporate bag-of-words models with Twitter-specific features like hashtags and emoticons (Davidov et al., 2010; Büchner and Stein, 2015). Word embeddings obtained from large amounts of tweets are used under the scope of an unsupervised approach for sentiment analysis (Astudillo et al., 2015). Additionally, deep learning models have recently become very popular for Twitter sentiment analysis (Severyn and Moschitti, 2015). Topic modeling approaches for sentiment analysis can also be found in literature, e.g., (Mei et al., 2007; Lin and He, 2009; Lu et al., 2011; Alam et al., 2016; Rao, 2016)

In this paper, we present systems submitted to the SemEval 2016 Task 4 (Nakov et al., 2016) that deal with the sentiment analysis of tweets on the sentence level. The submitted systems are based on the fusion of the different classifiers. Specifically, 1) we enhanced the system submitted by Malandrakis et al. (2014) to the SemEval 2014 Task 9 (Rosenthal et al., 2014), 2) we used the open-source system submitted by Büchner and Stein (2015) to the SemEval 2015 Task 10 (Rosenthal et al., 2015), 3) we trained a convolutional neural network on a large amount of unlabelled Twitter data, 4) we developed a system

Proceedings of SemEval-2016, pages 155–163,
San Diego, California, June 16-17, 2016.

based on topic modeling and 5) we trained a classifier using word embeddings as features. Our system was submitted on two subtasks, namely subtask A (message polarity classification) and subtask B (tweet classification according to a two-point scale) and ranked in the fifth[1] and the first place, respectively.

2 System Description

2.1 Baseline

The baseline system is an enhanced version of the system submitted by Malandrakis et al. (2014) to the SemEval 2014 Task 9 (Rosenthal et al., 2014). The major changes include the different manipulation of hasgatags, muliword expressions and the affective features as well as the incorporation of new features. A plethora of features is extracted, the majority of which are affective. The feature extraction is performed at the tweet, suffix and prefix level. Assume the following tweet: *"Lol Red Sox just slid through 3rd base #out"* (tweet level). A window applied at the beginning estimates the prefix e.g., *"Lol Red Sox"* (prefix level) and a window at the end of the tweet estimates the suffix *"3rd base #out"* (suffix level).

2.1.1 Affective features

The baseline is based on affective features derived from the semantic-affective model proposed by Malandrakis et al. (2013). The semantic-affective model relies on the assumption that semantic similarity implies affective similarity (Malandrakis et al., 2013). First, a semantic model is built and then affective ratings are estimated for unknown tokens exploiting the affective ratings of semantically similar words. It is applicable both to single words or short multiword expression as shown below:

$$\hat{v}(t_j) = \alpha_0 + \sum_{i=1}^{N} \alpha_i v(w_i) S(t_j, w_i), \qquad (1)$$

where t_j is the unknown lexical token, $w_{1..N}$ are the seed words, $v(w_i)$, α_i are the affective rating and the weight corresponding to the word w_i and $S(\cdot)$ is the semantic similarity metric between two tokens. The semantic model was built as shown in (Palogiannidi et al., 2015) using word-level contextual feature vectors and adopting a scheme based on mutual information for feature weighting. The dimensionality of the affective features was reduced by retaining only the polarity features (instead of using additional affective dimensions, like arousal and dominance). Affective lexica were created using a generic corpus (116M sentences) (Iosif et al., 2016) and a Twitter corpus (115M tweets) created for the purpose of our submission. In the former case task-independent affective ratings are estimated. Task-dependent affective ratings can be estimated by keeping domain-specific sentences in the generic corpus. A language model was built using domain relevant sentences and then the top 30% of the most relevant entries of the generic corpus were selected[2].

Affective features were also derived through semantic models that were built from corpora that were collected based on the topic modeling approach described in Section 2.2. **Third party affective lexica** were also used. Affin (Nielsen, 2011) contains discrete polarity ratings in the range $[-5, 5]$, nrc, nrctag (Mohammad et al., 2013) contain continuous polarity ratings for tokens, generated from a collection of tweets that include a positive or a negative word hashtag.

2.1.2 Tokenization

Based on the assumption that **hashtags** in different positions in a tweet may have different semantic interpretation, the tweets are transformed as follows: if a hashtag occurs at the end of the tweet it is assumed to convey semantic information. Otherwise, the hashtag is treated as a word or possibly a union of words. In the latter case, only the corresponding word is kept (e.g., "#moon is so beautiful tonight" $\rightarrow$ "moon is so beautiful tonight", but "What a beautiful night under the moonlight #romantic" is preserved as is). A hashtag that contains a union of words is expanded, e.g., #Hockeyisback $\rightarrow$ Hockey is back. Hashtags were expanded using the Viterbi algorithm (Forney and David, 1973) exploiting n-gram datasets. The n-gram dataset

[1]The actual performance of the system is different from the official results due to an erroneous submission.

[2]Perplexity filtering was used in order to estimate the relevance of a sentence to the language model while 30% was selected to be the most appropriate percentage over other percentages that were examined.

we used is a concatenation of the Google n-gram corpus that contains 1 trillion tokens from publicly accessible web pages (Brants and Franz, 2006) and an n-gram corpus based on 75 million English tweets (Herdağdelen, 2013). The absolute and relative frequencies of hashtags to be expanded were also used as features, as well as the indicators (binary features) that a tweet contains hashtags that require expansion. **POS-tagging / Tokenization** was then performed using the ARK NLP tweeter tagger (Owoputi et al., 2013), a Twitter-specific tagger. A tweet contains single words or emoticons and punctuations or multiword expressions. **Multiword expressions (MWE)** are non-compositional expressions that are processed as a single token. They were detected using the Gensim library (Řehůřek and Sojka, 2010) and they were added to the affective lexica.

Some parts of the tweets may be crucial for the correct understanding of its affective meaning. We assume that such parts, are the prefix, the suffix and the negated parts. **Negations** were detected using the list proposed by Potts (2011). When a negation token is detected, the tokens that follow are marked as negated until a punctuation mark is reached. Then, feature extraction is applied on the negated part of the tweet. **Windows** are used for splitting a tweet into prefix and suffix attempting to estimate the cognitive dissonance phenomenon that is associated with sarcasm, irony and humour (Reyes et al., 2012). The suffix is extracted by applying windows that keep the 20%, 50% and 70% of tokens that occur at the end of the tweet. Feature extraction is performed for both suffixes and prefixes.

2.1.3 Word2vec features

In addition to the use of semantic similarity features as in (Malandrakis et al., 2014), we use semantic representations, i.e., word embeddings that are utilized for the semantic similarity estimation, i.e., the $S(\cdot)$ in (1). **Word embedding features** were derived using word2vec (Mikolov et al., 2013), representing each word as a 300-D vector. Since tweet-level features are required, for each tweet a 300-D vector is generated by averaging the corresponding vectors of the constituent words.

2.1.4 Additional features

Additional features based on characters and subjectivity lexica were used. Character features include the absolute and relative frequencies of selected characters. The selected characters are the **capitalized** letters, the **punctuation** marks, the **emoticons** as well as characters **repetitions** , i.e., at least three same successive characters in a word. **Subjectivity** features were also extracted based on the subjectivity lexicon of (Wilson et al., 2005). Specifically, the absolute and the relative frequencies of the strong positive/negative and weak positive/negative words were used as features.

2.1.5 Statistics extraction

The statistics of the token-level polarity features were estimated in order to extract tweet-level features. The following statistics were computed: length (cardinality), min, max, max amplitude, sum, average, range (max minus min), standard deviation and variance. Normalized versions of the same statistics were also calculated.

2.2 Topic Modeling

Topic modeling is a method for discovering "topics" that occur in collections of documents. Typically, multiple topics are present in a document. The most widely used topic modeling approach is the Latent Dirichlet Allocation (LDA) (Blei et al., 2003) which is based on Latent Semantic Analysis (LSA) (Deerwester, 1988) and probabilistic Latent Semantic Analysis (pLSA) (Hofmann, 2000). Here, we used topic modeling in order to adapt the semantic space on each tweet. In essence, the corpus was split in a probabilistic way based on the detected topics and then a semantic model was built for each subcorpus. Since this technique is probabilistic, each corpus sentence can belong to each revealed topic with a given probability. Then, clusters of sentences per topic were created by classifying each sentence to the most probable topic. A semantic model was built for each cluster, using word2vec (Mikolov et al., 2013). A mixture of the semantic models, weighted by the topic posteriors is used for the estimation of tweet's semantic similarities, e.g., $S(\cdot)$ in (1).

2.3 Convolutional Neural Network

A deep convolutional neural network (CNN) was also developed. The neural network's architecture is inspired by sentence classification tasks (Collobert et al., 2011; Kalchbrenner et al., 2014; Kim, 2014). Each tweet is represented by a "sentence" matrix $\mathbf{S}$ that is created as follows. First, each word is represented as a 300-D vector using word2vec, and then, the word vectors are concatenated as follows: $\mathbf{S} = W_1 \oplus W_2 \oplus W_3 \oplus \cdots \oplus W_n$., $S \in \mathbb{R}^{d \times n}$, where $\oplus$ indicates the concatenation operation. Each column i of $\mathbf{S}$ is a vector $W \in \mathbb{R}^d$ that corresponds to the i^{th} word of the tweet. This way the sequence of the words in the tweet is kept. In order to preserve the same length for all tweets, zero padding was applied concatenating zero word vectors until the length n of the longest tweet is reached. The size of $\mathbf{S}$ is $d \times n$, where d is the dimension of the word embedding and n is the length of the longest tweet. The matrix $\mathbf{S}$ is the input to the network, where a convolution operation is applied between $\mathbf{S}$ and a kernel $F \in \mathbb{R}^{d \times m}$. The width m was set to 5 and the parameters of the model, i.e., the values of the kernel, the size of the sliding window h are learned. The result of the convolutional layer is a vector $c \in \mathbb{R}^{n-m+1}$ (Kim, 2014). In fact, the convolution network uses multiple kernels with varying sliding windows and generates multiple features. These features are the inputs to the next layer which selects the maximum value of each feature by applying a max-over-time pooling operation (max-pooling layer) (Collobert et al., 2011). Next, the output of the max-pooling layer is passed to a dropout layer (Srivastava et al., 2014). A softmax layer that classifies each test instance to one of the possible classes is the final step.

2.4 Word2vec System

Based on the assumption that similarity of meaning implies affective similarity (Malandrakis et al., 2013) we build a system that relies exclusively on tweets' semantic representation. Specifically, word2vec was applied over large text corpora in order to compute the semantic representations of words (formulated as vectors). Then, the vectors of each tweet's constituent words were averaged in order to create a single vector. These vectors were used for training a random forest classifier.

2.5 Webis

The Webis open-source system (Büchner and Stein, 2015) that was submitted on (Rosenthal et al., 2015) is the ensemble of different subsystems that ranked at the top of Semeval 2013, Semeval 2014 Sentiment Analysis tasks (Nakov et al., 2013; Rosenthal et al., 2014). The following systems were combined in a late fusion scheme, based on the mean posterior probabilities: 1) NRC-Canada (Mohammad et al., 2013) is based on morphological, linguistic and polarity features, 2) GU-MLT-LT (Günther and Furrer, 2013) trains a linear classifier by stochastic gradient descent which uses social media specific text preprocessing and linguistic features, 3) KLUE (Proisl et al., 2013) employs a Maximum Entropy classifier with bag-of-words models, sentiment, emoticons and internet slang abbreviations features, 4) TeamX (Miura et al., 2014) is similar to NRC-Canada but uses more lexicon-based features and handles the unbalance distribution of tweets' sentiment by adopting a weighting scheme to bias the output.

2.6 Fusion of the systems

The motivation behind the development of various systems for sentiment classification is that different systems may capture different aspects of the sentiment, and by combining them we can predict more accurately the sentiment of tweets. The system architecture including the fusion scheme is summarized in the figure below.

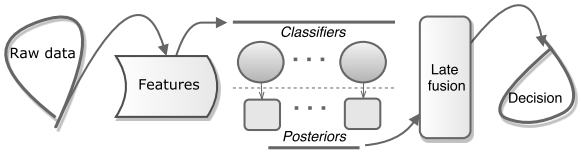

Figure 1: System Architecture.

As shown in Figure 1, the various classifiers were trained with features extracted from Twitter datasets and their posterior probabilities were combined in a late fusion scheme. Various techniques can be used for the late fusion of classifiers, e.g., voting-based (Tulyakov et al., 2008) methods, multi-classifiers fusion that use posteriors as features for training a new classifier (Kittler et al., 1998; Gutierrez et al., 2016)

	Datasets						
	2013		2014			2015	2016
System	SMS	Twitter	LiveJour.	Twitter	TwitterSarcasm	Twitter	Twitter
W-CNN-B-W2V	0.547	0.704	0.687	0.715	0.455	0.658	0.608
W-TM-B⋆-W2V	0.610	0.704	0.705	0.718	0.473	0.638	0.611
W-CNN-B⋆-W2V	0.593	0.716	0.707	0.721	0.485	0.643	0.614
W-B⋆-CNN	0.635	0.717	0.725	0.720	0.508	0.645	**0.618**
W-B⋆-W2V	0.632	0.711	0.732	0.727	0.468	0.640	**0.624**

Table 1: Macro-averaged F-score for subtask A and various system combinations. The submitted system is highlighted in grey.

NRC	GU-MLT-LT	TeamX	Klue	Baseline	Topic Modeling	CNN	AvgR	AvgF1
AvgR								
0.709	0.712	0.712	0.737	**0.821**	0.753	0.752		
√	√	√	√	√	√	√	**0.803**	**0.808**
×	×	×	×	√	√	√	0.818	0.793
√	√	√	√	×	√	√	0.765	0.781
√	√	√	√	√	×	√	0.780	0.790
√	√	√	√	√	√	×	0.798	0.801
√	√	√	×	√	√	×	0.797	0.799
×	×	×	×	√	×	√	**0.827**	0.789
×	√	×	√	√	√	√	**0.824**	**0.805**

Table 2: Macro-averaged recall and F1 for subtask B, for a 2016 test twitter dataset. The submitted system is highlighted in grey.

or algebraic combinations (Kittler et al., 1998). Algebraic combinations are based on operations such as mean, median, product, max or min.

3 Experimental Procedure and Results

3.1 Data

We trained our systems using both general purpose and Twitter data. We used a large generic corpus that contains 116M sentences (G-116M). The dataset was created by posing queries on a web search engine and aggregating the resulting snippets. A Twitter specific dataset was also created collecting 115M tweets (T-115M). We also used the training data provided by SemEval 2016 for subtasks A and B (Sem/A-2016, Sem/B-2016), as well as training data from the corresponding task of SemEval 2013 (Sem-2013). We also used ANEW (Bradley and Lang, 1999) for bootstrapping the affective lexicon expansion process.

3.2 Systems

The following subsystems were combined for the subtask A: Baseline (B), CNN, Word2vec system (W2V) and Webis (W). Baseline and Webis were trained on the concatenation of Sem/A-2016 and Sem-2013. A two-stage feature selection was applied, the first one took place on each feature set separately and the second to the combined feature set of the first stage. Finally, a Naive Bayes tree classifier was trained. Affective features derived by the topic modeling approach were also incorporated. The word vectors that are required for the CNN are derived from different corpora, i.e., the combination of a Google News dataset and the T-115M corpus. Specifically, for each word in the tweet, the word vector extracted from the latter corpus is used only if the word doesn't exist in the former corpus. The vectors of OOV words are initialized randomly from a uniform distribution. For the word2vec-based system we trained a random forest classifier with 100 trees using 300-D feature vectors on the concatenation of Sem/A-2016 and Sem-2013. We experimented with various fusion schemes and system combinations. The mean algebraic fusion scheme was selected for the reported results.

The submission for the subtask B includes the following systems: Baseline (B⋆), Topic Model-

ing (TM) and three Webis subsystems, i.e., NRC-Canada, GU-MLT-LT and Team X. The Baseline ($B^\star$) and the selected Webis systems were trained on the Sem/B-2016 dataset. Both feature selection and the classification of $B^\star$ are similar to the Baseline that was used in subtask A. The difference between B and $B^\star$ is that the former, in contrast to the second, includes affective features extracted through the approach based on topic modeling. The topic modeling system applies the LDA algorithm on the G-116M, using 16 topics[3]. Then, affective ratings are estimated and they are used in order to train a Naive Bayes tree classifier. Similarly to subtask A, a mean algebraic fusion scheme was selected for combining the systems.

3.3 Results

The developed systems for subtask A classify each tweet in one out of three sentiment classes (positive vs. negative vs. neutral), however the performance of the model is measured taking into consideration only the performance on the two polarity classes, i.e., positive and negative. The macro-averaged F-score of the positive and negative classes $F_1^{PN} = \frac{F_1^{Pos}+F_1^{Neg}}{2}$ is reported in Table 1 for various datasets. In the first column the integrated systems are presented, while the submitted system is highlighted with grey color. Our system ranked fifth, while experimenting with different system combinations we can climb up to the third position.

Subtask B is a binary classification task (positive vs. negative). Macro-averaged recall $p^{PN} = \frac{p^{Pos}+p^{Neg}}{2}$, $p^{PN} \in [0, 1]$ and macro-averaged F-score are reported in Table 2 for various system combinations. The systems and their macro-averaged recall are listed in the first and second row of the table, respectively. The reported scores were derived by the test tweets of 2016 that belong to specific topics. Each row that follows indicates a unique combination and the submitted system is highlighted with grey color. In the case that all the available systems are combined, the marco-averaged recall (AvgR) is 0.803 and macro-averaged F-score (AvgF1) is 0.808. The combinations that follow contain a subset of the systems (the selected systems are denoted with $\surd$ and the omitted systems with $\times$). The baseline was proved to be the most robust system achieving the highest performance among the others and higher performance than the submitted system (0.797). When using all the subsystems but one, performance decreases except in the case that the Webis system is omitted, then AvgR increased to 0.818. The highest performance drop is achieved when the baseline system is omitted. Investigating more combination schemes AvgR rises to 0.827. The combination of CNN, TM and $B^\star$ with two of the best Webis systems yields robust performance with AvgR and AvgF1 0.824 and 0.805, respectively.

[3]The number of topics derived after the conduction of experiments on a news headlines dataset.

4 Conclusions

We presented a system for the sentiment classification of tweets for the SemEval 2016 Task 4: Sentiment analysis in Twitter. We participated in subtasks A and B and we won subtask B. We developed various systems including a CNN and a topic modeling approach for the adaptation of the semantic space to each tweet. Regarding task A, the submitted system was ranked between the fifth and third position. The performance for subtask B can be higher (up to +3% compared to the submitted system) for various system combinations. Future work will focus on the fusion of the systems as well as their enhancement in order to achieve higher performance to the three-point scale sentiment classification.

Acknowledgements: Elisavet Palogiannidi, Elias Iosif and Alexandros Potamianos were partially funded by the SpeDial project supported by the EU Seventh Framework Programme (FP7), grant number 611396 and the BabyRobot project supported by the EU Horizon 2020 Programme, grant number: 687831.

References

Md. H. Alam, W. Ryu, and S. Lee. 2016. Joint Multi-grain Topic Sentiment: Modeling Semantic Aspects for Online Reviews. *Information Sciences*, 339:206–223.

R. Astudillo, S. Amir, W. Ling, B. Martins, M. J. Silva, and I. Trancoso. 2015. INESC-ID: Sentiment Analysis without Hand-Coded Features

or Linguistic Resources using Embedding Subspaces. In *Proc. of the 9th International Workshop on Semantic Evaluation (SemEval)*, pages 652–656.

D. M. Blei, A. Y. Ng, and M. I. Jordan. 2003. Latent Dirichlet Allocation. *The Journal of Machine Learning Reasearch*, 3:993–1022.

M. Bradley and P. Lang. 1999. Affective norms for English words (ANEW): Instruction Manual and Affective Ratings. Technical report.

T. Brants and A. Franz. 2006. Web 1T 5-gram corpus Version 1. Technical report, Google Research.

M. Büchner and B. Stein. 2015. Webis: An Ensemble for Twitter Sentiment Detection. In *Proc. of the 9th International Workshop on Semantic Evaluation (SemEval)*, pages 582–589.

C. Chew and G. Eysenbach. 2010. Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. *PloS ONE*, 5(11):1–13.

R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. 2011. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537.

D. Davidov, O. Tsur, and A. Rappoport. 2010. Enhanced sentiment learning using twitter hashtags and smileys. In *Proc. of Conference on Computational Linguistics (Coling)*, pages 241–249.

S. Deerwester. 1988. Improving Information Retrieval with Latent Semantic Indexing. In *Proc. of the 51st Annual Meeting of the American Society for Information Science and Technology (ASIS&T)*, pages 36–40.

J. Forney and G. David. 1973. The Viterbi algorithm. *In Proceedings of the IEEE*, 61(3):268–278.

T. Günther and L. Furrer. 2013. GU-MLT-LT: Sentiment analysis of short messages using linguistic features and stochastic gradient descent. In *Proc. of the 7th International Workshop on Semantic Evaluation (SemEval)*, pages 328–332.

P. A. Gutierrez, M. Perez-Ortiz, J. Sanchez-Monedero, F. Fernandez-Navarro, and C. Hervas-Martinez. 2016. Ordinal regression methods: survey and experimental study. *IEEE Transactions on Knowledge and Data Engineering*, 28(1):127–146.

A. Herdağdelen. 2013. Twitter n-gram corpus with demographic metadata. *Language Resources and Evaluation*, 47(4):1127–1147.

T. Hofmann. 2000. Learning the similarity of documents: An information-geometric approach to document retrieval and categorization. In *Proc. of Conference on Advances in Neural Information Processing Systems (NIPS)*, pages 914–920.

E. Iosif, S. Georgiladakis, and A. Potamianos. 2016. Cognitively Motivated Distributional Representations of Meaning. In *Proc. of the Language Resources and Evaluation Conference (LREC)*.

B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury. 2009. Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology*, 60(11):2169–2188.

N. Kalchbrenner, E. Grefenstette, and P. Blunsom. 2014. A Convolutional Neural Network for Modelling Sentences. In *Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 655–665.

Y. Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

J. Kittler, M. Hatef, R. PW Duin, and J. Matas. 1998. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239.

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

B. Lu, M. Ott, C. Cardie, and B. K. Tsou. 2011. Multi-aspect sentiment analysis with topic models. In *Proc. of International Conference on Data Mining Workshops (ICDMW)*, pages 81–88. IEEE.

N. Malandrakis, A. Potamianos, E. Iosif, and S. Narayanan. 2013. Distributional semantic models for affective text analysis. *IEEE Transactions on Audio, Speech and Language Processing*, 21(11):2379–2392.

N. Malandrakis, M. Falcone, C. Vaz, J. Bisogni, A. Potamianos, and S. Narayanan. 2014. SAIL: Sentiment Analysis using Semantic Similarity and Contrast Features. In *Proc. of the 8th International Workshop on Semantic Evaluation (SemEval)*, pages 512–516.

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

T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proc of. Advances in Neural Information Processing systems (NIPS)*, pages 3111–3119.

Y. Miura, S. Sakaki, K. Hattori, and T. Ohkuma. 2014. TeamX: A sentiment analyzer with enhanced lexicon mapping and weighting scheme for unbalanced data. In *Proc. of the 8th International Workshop on Semantic Evaluation (SemEval)*, pages 628–632.

S. M. Mohammad, S. Kiritchenko, and X. Zhu. 2013. NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets. In *Proc. of the 7th International Workshop on Semantic Evaluation (SemEval)*, pages 321–327.

P. Nakov, Z. Kozareva, A. Ritter, S. Rosenthal, V. Stoyanov, and T. Wilson. 2013. Semeval-2013 Task 2: Sentiment Analysis in Twitter. In *Proc. of the 7th International Workshop on Semantic Evaluation (SemEval)*, pages 312–320.

P. Nakov, A. Ritter, S. Rosenthal, V. Stoyanovand, and F. Sebastiani. 2016. SemEval-2016 task 4: Sentiment Analysis in Twitter. In *Proc. of the 10th International Workshop on Semantic Evaluation (SemEval)*.

F. Å. Nielsen. 2011. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. In *Proc. of the ESWC Workshop on Making Sense of Microposts*, pages 93–98.

O. Owoputi, C. Dyer, K. Gimpel, N. Schneider, and N. A. Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proc. of North American Chapter of the Association for Computational Linguistics (NAACL)*.

E. Palogiannidi, E. Iosif, P. Koutsakis, and A. Potamianos. 2015. Valence, Arousal and Dominance Estimation for English, German, Greek, Portuguese and Spanish Lexica using Semantic Models. In *Proc. of Interspeech*, pages 1527–1531.

B. Pang and L. Lee. 2008. Opinion mining and Sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135.

C. Potts. 2011. Sentiment symposium tutorial. In *Proc. of Sentiment Symposium Tutorial*.

T. Proisl, P. Greiner, S. Evert, and B. Kabashi. 2013. KLUE: Simple and Robust methods for polarity classification. In *Proc. of the 7th International Workshop on Semantic Evaluation (SemEval)*, pages 395–401.

Y. Rao. 2016. Contextual Sentiment Topic Model for Adaptive Social Emotion Classification. *IEEE Intelligent Systems*, 31(1):41–47.

R. Řehůřek and P. Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proc. of the Language Resources and Evaluation Conference (LREC) Workshop on New Challenges for NLP Frameworks*, pages 45–50.

A. Reyes, P. Rosso, and D. Buscaldi. 2012. From humor recognition to irony detection: The figurative language of social media. *Data & Knowledge Engineering*, 74:1–12.

S. Rosenthal, A. Ritter, P. Nakov, and V. Stoyanov. 2014. Semeval-2014 Task 9: Sentiment Analysis in Twitter. In *Proc. of the 8th International Workshop on Semantic Evaluation (SemEval)*, pages 73–80.

S. Rosenthal, P. Nakov, S. Kiritchenko, S. M. Mohammad, A. Ritter, and V. Stoyanov. 2015. Semeval-2015 Task 10: Sentiment Analysis in Twitter. In *Proc. of the 9th International Workshop on Semantic Evaluation (SemEval)*, pages 451–463.

A. Severyn and A. Moschitti. 2015. UNITN: Training deep convolutional neural network for Twitter sentiment classification. In *Proc. of the 9th International Workshop on Semantic Evaluation (SemEval)*, pages 464–469.

N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

S. Tulyakov, S. Jaeger, V. Govindaraju, and D. Doermann. 2008. Review of classifier combination methods. In *Machine Learning in Document Analysis and Recognition*, pages 361–386.

S. Verma, S. Vieweg, W. J Corvey, L. Palen, J. H Martin, M. Palmer, A. Schram, and K. M. Anderson. 2011. Natural Language Processing to the Rescue? Extracting "Situational Awareness" Tweets During Mass Emergency. In *Proc. of the 5th International Conference on Web and Social Media (ICWSM)*.

T. Wilson, J. Wiebe, and P. Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proc. of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP)*, pages 347–354.