Sentiment Analysis in Twitter for Macedonian

Dame Jovanoski, Veno Pachovski
University American College Skopje
UACS, Macedonia
{jovanoski,pachovski}@uacs.edu.mk

Preslav Nakov
Qatar Computing Research Institute
HBKU, Qatar
pnakov@qf.org.qa

Abstract

We present work on sentiment analysis in Twitter for Macedonian. As this is pioneering work for this combination of language and genre, we created suitable resources for training and evaluating a system for sentiment analysis of Macedonian tweets. In particular, we developed a corpus of tweets annotated with tweet-level sentiment polarity (positive, negative, and neutral), as well as with phrase-level sentiment, which we made freely available for research purposes. We further bootstrapped several large-scale sentiment lexicons for Macedonian, motivated by previous work for English. The impact of several different pre-processing steps as well as of various features is shown in experiments that represent the first attempt to build a system for sentiment analysis in Twitter for the morphologically rich Macedonian language. Overall, our experimental results show an F_1-score of 92.16, which is very strong and is on par with the best results for English, which were achieved in recent SemEval competitions.

1 Introduction

The increasing popularity of social media services such as Facebook, Twitter and Google+, and the advance of Web 2.0 have enabled users to share information and, as a result, to have influence on the content distributed via these services. The ease of sharing, e.g., directly from a laptop, a tablet or a smart phone, have contributed to the tremendous growth of the content that users share on a daily basis, to the extent that nowadays social networks have no choice but to filter part of the information stream even when it comes from our closest friends.

Naturally, soon this unprecedented abundance of data has attracted business and research interest from various fields including marketing, political science, and social studies, among many others, which are interested in questions like these: *Do people like the new Apple Watch? What do they hate about iPhone6? Do Americans support ObamaCare? What do Europeans think of Pope's visit to Palestine? How do we recognize the emergence of health problems such as depression?*

Such questions can be answered by studying the sentiment of the opinions people express in social media. As a result, the interest for sentiment analysis, especially in social media, has grown, further boosted by the needs of various applications such as mining opinions from product reviews, detecting inappropriate content, and many others.

Below we describe the creation of data and the development of a system for sentiment polarity classification in Twitter for Macedonian: positive, negative, neutral. We are inspired by a similar task at SemEval, which is an ongoing series of evaluations of computational semantic analysis systems, composed by multiple challenges such as text similarity, word sense disambiguation, etc. One of the challenges there was on Sentiment Analysis in Twitter, at SemEval 2013-2015 (Nakov et al., 2013; Rosenthal et al., 2014; Rosenthal et al., 2015; Nakov et al., 2015), where over 40 teams participated three years in a row.[1] Here we follow a similar setup, focusing on message-level sentiment analysis of tweets, but for Macedonian instead of English. Moreover, while at SemEval the task organizers used Mechanical Turk to do the annotations, where the control for quality is hard (everybody can pretend to know English), our annotations are done by native speakers of Macedonian.

[1] Other related tasks were the Aspect-Based Sentiment Analysis task (Pontiki et al., 2014; Pontiki et al., 2015), and the task on Sentiment Analysis of Figurative Language in Twitter (Ghosh et al., 2015).

Proceedings of Recent Advances in Natural Language Processing, pages 249–257,
Hissar, Bulgaria, Sep 7–9 2015.

The remainder of the paper is organized as follows: Section 2 presents some related work. Sections 3 and 4 describe the datasets and the various lexicons we created for Macedonian. Section 5 gives detail about our system, including the preprocessing steps and the features used. Section 6 describes our experiments and discusses the results. Section 7 concludes with possible directions for future work.

2 Related Work

Research in sentiment analysis started in the early 2000s. Initially, the problem was regarded as standard document classification into topics, e.g., Pang et al. (2002) experimented with various classifiers such as maximum entropy, Naïve Bayes and SVM, using standard features such as unigram/bigrams, word counts/present, word position and part-of-speech tagging. Around the same time, other researchers realized the importance of external sentiment lexicons, e.g., Turney (2002) proposed an unsupervised approach to learn the sentiment orientation of words/phrases: positive vs. negative. Later work studied the linguistic aspects of expressing opinions, evaluations, and speculations (Wiebe et al., 2004), the role of context in determining the sentiment orientation (Wilson et al., 2005), of deeper linguistic processing such as negation handling (Pang and Lee, 2008), of finer-grained sentiment distinctions (Pang and Lee, 2005), of positional information (Raychev and Nakov, 2009), etc. Moreover, it was recognized that in many cases, it is crucial to know not just the polariy of the sentiment, but also the topic towards which this sentiment is expressed (Stoyanov and Cardie, 2008).

Early sentiment analysis research focused on customer reviews of movies, and later of hotels, phones, laptops, etc. Later, with the emergence of social media, sentiment analysis in Twitter became a hot research topic. The earliest Twitter sentiment datasets were both small and proprietary, such as the i-sieve corpus (Kouloumpis et al., 2011), or relied on noisy labels obtained from emoticons or hashtags. This situation changed with the emergence of the SemEval task on Sentiment Analysis in Twitter, which ran in 2013-2015 (Nakov et al., 2013; Rosenthal et al., 2014; Rosenthal et al., 2015). The task created standard datasets of several thousand tweets annotated for sentiment polarity. Our work here is inspired by that task.

In our experiments below, we focus on Macedonian, for which we only know two publications on sentiment analysis, none of which is about Twitter.

Gajduk and Kocarev (2014) experimented with 800 posts from the Kajgana forum (260 positive, 260 negative, and 280 objective), using SVM and Naïve Bayes classifiers, and features such as bag of words, rules for negation, and stemming.

Uzunova and Kulakov (2015) experimented with 400 movie reviews[2] (200 positive, and 200 negative; no objective/neutral), and a Naïve Bayes classifier, using a small manually annotated sentiment lexicon of unknown size, and various preprocessing techniques such as negation handling and spelling/character translation. Unfortunately, the datasets and the generated lexicons used in the above work are not publicly available, and/or are also from a different domain. As we are interested in sentiment analysis of Macedonian tweets, we had to build our own datasets.

In addition to preparing a dataset of annotated tweets, we further focus on creating sentiment polarity lexicons for Macedonian. This is because lexicons are crucial for sentiment analysis. As we mentioned above, since the very beginning, researchers have realized that sentiment analysis was quite different from standard document classification (Sebastiani, 2002), and that it crucially needed external knowledge in the form of suitable sentiment polarity lexicons. For further detail, see the surveys by Pang and Lee (2008) and Liu and Zhang (2012).

Until recently, such sentiment polarity lexicons have been manually crafted, and were of small to moderate size, e.g., LIWC (Pennebaker et al., 2001), General Inquirer (Stone et al., 1966), Bing Liu's lexicon (Hu and Liu, 2004), and MPQA (Wilson et al., 2005), all have 2000-8000 words. Early efforts in building them automatically also yielded lexicons of moderate sizes (Esuli and Sebastiani, 2006; Baccianella et al., 2010).

However, recent results have shown that automatically extracted large-scale lexicons (e.g., up to a million words and phrases) offer important performance advantages, as confirmed at shared tasks on Sentiment Analysis in Twitter at SemEval 2013-2015 (Nakov et al., 2013; Rosenthal et al., 2014; Rosenthal et al., 2015).

[2]There have been also experiments on movie reviews for the closely related Bulgarian language (Kapukaranov and Nakov, 2015), but there the objective was to predict user rating, which was addressed as an ordinal regression problem.

Similar observations were made in the Aspect-Based Sentiment Analysis task, which ran at SemEval 2014-2015 (Pontiki et al., 2014; Pontiki et al., 2015). In both tasks, the winning systems benefited from building and using massive sentiment polarity lexicons (Mohammad et al., 2013; Zhu et al., 2014). These large-scale automatic lexicons were typically built using bootstrapping, starting with a small seed of, e.g., 50-60 words (Mohammad et al., 2013), and sometimes even using just two emoticons.

3 Data

During a period of six months from November 2014 to April 2015, we collected about half a million tweet messages. In the process, we had to train and use a high-precision Naïve Bayes classifier for detecting the language, because the Twitter API often confused Macedonian tweets with Bulgarian or Russian. From the resulting set of tweets, we created training and testing datasets, which we manually annotated at the tweet level (using *positive*, *negative*, and *neutral/objective* as labels[3]).

The training dataset was annotated by the first author, who is a native speaker of Macedonian. In addition to tweet-level sentiment, we also annotated the sentiment-bearing words and phrases inside the *training* tweets, in order to obtain a sentiment lexicon.

The testing dataset was only annotated at the tweet level, and for it there was one additional annotator, again a native speaker of Macedonian. The value of the Cohen's Kappa statistics (Cohen, 1960) for the inter-annotator agreement between the two annotators was 0.41, which corresponds to *moderate* agreement (Landis and Koch, 1977); this relatively low agreement shows the difficulty of the task. For the final testing dataset, we discarded all tweets on which the annotators disagreed (a total of 474 tweets).

Table 1 shows the statistics about the training and the testing datasets. We can see that the data is somewhat balanced between positive and negative tweets, but has a relatively smaller proportion of neutral tweets.[4]

[3]Following (Nakov et al., 2013), we merged *neutral* and *objective* as they are commonly confused by annotators.

[4]It was previously reported that most tweets are neutral, but this was for English, and for tweets about selected topics (Rosenthal et al., 2014). We have no topic restriction; more importantly, there is a severe ongoing political crisis in Macedonia, and thus Macedonian tweets were full of emotions.

Dataset	Positive	Neutral	Negative	Total
Train	2,610 (30%)	1,280 (15%)	4,693 (55%)	8,583
Test	431 (38%)	200 (18%)	508 (44%)	1,139

Table 1: Statistics about the datasets.

We faced many problems when processing the tweets. For example, it was hard to distinguish advertisements vs. news vs. ordinary user messages, which is important for sentiment annotations. Here is an example tweet by a news agency, which should be annotated as neutral/objective:

Лицето АБВ е убиецот и виновен за убиството на БЦД. [5]

The above message has good grammatical structure, but in our datasets there are many messages with missing characters, missing words, misspellings and with poor grammatical structure; this is in part what makes the task difficult. Here is a sample message with missing words and misspellings:

брао бе, ги утепаа с....!!! [6]

Non-standard language is another problem. This includes not only slang and words written in a funny way on purpose, but also many dialectal words from different regions of Macedonia that are not used in Standard Macedonian. For example, in the Eastern part of the Republic of Macedonia, there are words with Bulgarian influence, while in the Western part, there are words influenced by Albanian; and there is Serbian influence in the North.

Finally, many problems arise due to our using a small dataset for sentiment analysis. This mainly affects the construction of the sentiment lexicons and the reason for this is the distribution of emoticons, hashtags and sentiment words. In particular, if we want to use hashstags or emoticons as seeds to construct sentiment lexicons, we find that very few tweet messages have emoticons or hashtags. Table 2 shows the statistics about the distribution of the emoticons and hashtags in the dataset (half a million tweet messages). That is why, in our experiments below, we do not rely much on hashtags for lexicon construction.

[5]Translation: *The person ABC is the killer, and he is responsible for the murder of BCD.*

[6]Translation: *That's great, they have smashed them with....!!!*

Token type	No. of messages
Without emoticons and hashtags	473,420
With emoticons	3,635
With hashtags	521
Total	477,576

Table 2: Number of tweets in our datasets that contain emoticons and hashtags.

4 Sentiment Lexicons

Sentiment polarity lexicons are key resources for the task of sentiment analysis, and thus we have put special efforts to generate some for Macedonian using various techniques.[7] Typically, a sentiment lexicon is a set of words annotated with positive and negative sentiment. Sometimes there is also a polarity score of that sentiment, e.g., *spectacular* could have positive strength of 0.91, while for *okay* that might be 0.3.

4.1 Manually-Annotated Lexicon

As we mentioned above, in the process of annotation of the training dataset, the annotator also marked the sentiment-bearing words and phrases in each tweet, together with their sentiment polarity in that context: positive or negative.

The phrases for the lexicon were annotated by two annotators, both native speakers of Macedonian. We calculated the Cohen's Kappa statistics (Cohen, 1960) for the inter-annotator agreement, and obtained the score of 0.63, which corresponds to *substantial* agreement (Landis and Koch, 1977).

We discarded all words with disagreement, a total of 122, and we collected the remaining words and phrases in a lexicon. The lexicon contained 1,088 words (459 positive and 629 negative).

4.2 Translated Lexicons

Another way to obtain a sentiment polarity lexicon is by translating a preexisting one from another language. We translated some English manually-crafted lexicons such as Bing Liu's lexicon (2,006 positive and 4,783 negative), and MPQA (2,718 positive and 4,912 negative), and an automatically extracted Bulgarian lexicon (5,016 positive and 2,415 negative), extracted from a movie reviews website (Kapukaranov and Nakov, 2015). For the translation of the lexicons we used Google Translate, and we further manually corrected the results, removing bad or missing translations.

[7] All lexicons presented here are publicly available at `https://github.com/badc0re/sent-lex`

4.3 Automatically-Constructed Lexicons

Sentiment lexicons can also be constructed automatically by using Pointwise Mutual Information as a way to calculate the semantic orientation of a word (Turney, 2002) or a phrase in a message (text). In sentiment analysis, using the orientation of a word, the positive and the negative score of a word/phrase can be calculated. The semantic orientation can be calculated as follows:

$$SO(w) = PMI(w, pos) - PMI(w, neg)$$

where PMI is the pointwise mutual information, and pos and neg are placeholders standing for any of the seed positive and negative terms.

A positive/negative value for $SO(w)$ indicates positive/negative polarity for w, and its magnitude shows the corresponding sentiment strength. In turn, $PMI(w, pos) = \frac{P(w,pos)}{P(w)P(pos)}$, where $P(w, pos)$ is the probability to see w with any of the seed positive words in the same tweet,[8] $P(w)$ is the probability to see w in any tweet, and $P(pos)$ is the probability to see any of the seed positive words in a tweet; $PMI(w, neg)$ is defined similarly.

Turney's PMI-based approach further serves as the basis for two popular large-scale automatic lexicons for English sentiment analysis in Twitter, initially developed by NRC for their participation in SemEval-2013 (Mohammad et al., 2013). The *Hashtag Sentiment Lexicon* uses as seeds hashtags containing 32 positive and 36 negative words, e.g., `#happy` and `#sad`; it then uses PMI and extracts 775,000 sentiment words from 135 million tweets. Similarly, the *Sentiment140* lexicon contains 1.6 million sentiment words and phrases, extracted from the same 135 million tweets, but this time using smileys as seed indicators for positive and negative sentiment, e.g., `:)`, `:-)` and `:))` serve as positive seeds, and `:(` and `:-(` as negative ones.

In our experiments, we used all words from our manually-crafted Macedonian sentiment polarity lexicon above as seeds, and then we mined additional sentiment-bearing words from a set of half a million Macedonian tweets. The number of tweets we used was much smaller in scale compared to that used in the *Hashtag Sentiment Lexicon* and in the *Sentiment140* lexicon, since there are much less Macedonian tweets (compared to English).

[8] Here we explain the method using number of tweets, as this is how we are using it, but Turney (2002) actually used page hits in the AltaVista search engine.

However, we used a much larger seed; as we will see below, this turns out to be a very good idea. We further tried to construct lexicons using words from the translated lexicons as seeds.

5 System Overview

The language of our tweet messages is Macedonian, and thus the text processing is a bit different than for English. As many basic tools that are freely available for English do not exist for Macedonian, we had to implement them in order to improve our model's performance. Our system uses logistic regression for classification, where words are weighted using TF.IDF.

5.1 Preprocessing

For pre-processing, we applied various algorithms, which we combined in order to achieve better performance. We used Christopher Potts' tokenizer,[9] and we had to be careful since we had to extract not only the words but also other tokens such as hashtags, emoticons, user names, etc. The pre-processing of the tweets goes as follows:

1. **URL and username removal**: tokens such as URLs and usernames (i.e., tokens starting with `@`) were removed.
2. **Stopword removal**: stopwords were filtered out based on a word list (146 words).
3. **Repeating characters removal**: consecutive character repetitions in a word were removed; also were removed repetitions of a word in the same token, e.g., 'какоооо' or 'дадада' (translated in English as 'what' and 'yes', respectively).
4. **Negation handling**: negation was addressed using a predefined list of negation tokens, then the prefix `NEG_CONTEXT_` was attached to the following tokens until a clause-level punctuation mark, in order to annotate it as appearing in a negated context, as suggested in (Pang et al., 2002). A list of 45 negative phrases and words was used to signal negation.
5. **Non-standard to standard word mapping**: non-standard words (slang) were mapped to an appropriate form, according to a manualy crafted predefined list of mappings.
6. **PoS tagging**: rule-based, using a dictionary.
7. **Tagging positive/negative words**: positive and negative words were tagged as `POS` and `NEG`, using sentiment lexicons.
8. **Stemming**: rule-based stemming was performed, which removes/replaces some prefixes/suffixes.

In sum, we started the transformation of an input tweet by converting it to lowercase, followed by removal of URLs and user names. We then normalized some words to Standard Macedonian using a dictionary of 173 known word transformations and we further removed stopwords (a list of 146 words). As part of the transformation, we marked the words in a negated context.

We further created a rule-based stemming algorithm with a list of 65 rules for removing/replacing prefixes and suffixes (Porter, 1980). We used two groups of rules: 45 rules for affix removal, and 20 rules for affix replacement. Developing a stemmer for Macedonian was challenging as this is a highly inflective language, rich in both inflectional and derivational forms. For example, here are some of the forms for the word навреда (English noun '*insult, offense*', verb '*offend, insult*'):

навредам	навредел
навредат	навредела
навредата	навределе
навредеа	навредело
навредев	навреден
навредевме	навредена
навредевте	...

In total, this word can generate over 90 inflected forms; in some cases, this involves a change in the last letter of the stem.

We further performed PoS (part-of-speech) tagging with our own tool based on averaged perceptron trained on MULTEXT-East resources (Erjavec, 2012). Here is an annotated tweet:

го/PN даваат/VB Глуп/NN и/CC
Поглуп/NN на/CC Телма/NN[10]

Here are the POS tags used in the above example: (*i*) NN-noun; (*ii*) AV-adverb; (*iii*) VB-verb; (*iv*) AE-adjective; (*v*) PN-pronoun; (*vi*) PN-pronoun; (*vii*) CN-cardinal number; (*viii*) CC-conjunction.

[9]http://sentiment.christopherpotts.net/tokenizing.html

[10]The translation for this message is: *Dump and Dumper is on Telma.*

We also developed a lemmatizer based on *approximate fuzzy string matching*. First, we used the *candidate word* (the one we want to lemmatize) to retrieve word lemmata that are similar to it; we then used *Jaro–Winkler* distance and *Levenshtein* distance to calculate a score that will determine whether the word matches closely enough some of the retrieved words. Such techniques have been used by other authors for *record linkage* (Cohen et al., 2003). Finally, as a last step in the transformation, we weighed the words using TF.IDF.

5.2 Features

In order to evaluate the impact of the sentiment lexicon, we defined features that are fully or partially dependent on the lexicons. When using multiple lexicons at the same time, there are separate instances of these features for each lexicon. Here are the features we used:

(*i*) Unigrams/bigrams: each one is a feature and its value is its TF.IDF score; (*ii*) Number of positive words in the tweet; (*iii*) Number of negative words in the tweet; (*iv*) Ratio of the number of positive words to the total number of sentiment words in the tweet; (*v*) Ratio of negative words to the total number of sentiment words in the tweet; (*vi*) Sum of the sentiment scores for all dictionary entries found in the tweet; (*vii*) Sum of the positive sentiment scores for all dictionary entries found in the tweet; (*viii*) Sum of the negative sentiment scores for all dictionary entries found in the tweet; (*ix-x*) Number of positive and negative emoticons in the tweet.

For classification, we used logistic regression. Our basic features were TF.IDF-weighted unigram and bigrams, and also emoticons. We further included additional features that focus on the positive and negative terms that occur in the tweet together with their scores in the lexicon. In case of two or more lexicons being used together, we had a copy of each feature for each lexicon.

6 Experiments

Our evaluation setup follows that of the SemEval 2013-2015 task on Sentiment Analysis in Twitter (Nakov et al., 2013; Rosenthal et al., 2014; Rosenthal et al., 2015), where the systems were evaluated in terms of an F-score that is the average of the F_1-score for the positive, and the F_1-score for the negative class. Note that, even though implicit, the neutral class still matters in this score.

Features	F-score	Diff.
All	92.16	
All - stop words	86.24	-5.92
All - negation	87.51	-4.65
All - norm. words to STD. Macedonian	90.22	-1.94
All - repeated characters	91.10	-1.06
All - stemming	93.14	0.98
All - PoS	92.01	-0.15

Table 3: The impact of excluding the preprocessing steps one at a time.

Features	F-score	Diff.
All	92.16	
All - automatically-constructed lexicons	72.77	-19.39
All - our manually-crafted lexicon	79.32	-12.84
All - all translated lexicons	91.89	-0.27

Table 4: The impact of excluding the features derived from the sentiment polarity lexicons.

Table 3 shows the impact of each pre-processing step. The first row shows the results when using all pre-processing steps and all sentiment lexicons. The following rows show the impact of excluding each of the preprocessing steps, one at a time. We can see that stopword removal and negation handling are most important: excluding each of them yields a five point absolute from in F-score. Normalization to Standard Macedonian turns out to be very important too as excluding it yields a drop of two points absolute. Handling repeating characters and stemming are also important, each yielding one point drop in F-score. However, the impact of using POS tagging is negligible.

Table 4 shows the impact of excluding some of the lexicons. We can see that our manually-crafted lexicon is quite helpful, contributing 13 points absolute in the overall F-score. Yet, the bootstrapped lexicons are even more important as excluding them yields a drop of 19 points absolute.

7 Conclusion and Future Work

We have presented work on sentiment analysis in Twitter for Macedonian. As this is pioneering work for this combination of language and genre, we created suitable resources for training and evaluating a system for sentiment analysis of Macedonian tweets. In particular, we developed a corpus of tweets annotated with tweet-level sentiment polarity (positive, negative, and neutral), as well as with phrase-level sentiment, which we made freely available for research purposes.

We further bootstrapped several large-scale sentiment lexicons for Macedonian, motivated by previous work for English. The impact of several different pre-processing steps as well as of various features is shown in experiments that represent the first attempt to build a system for sentiment analysis in Twitter for the morphologically rich Macedonian language. Overall, our experimental results show an F_1-score of 92.16, which is very strong and is on par with the best results for English, which were achieved in recent SemEval competitions.

In future work, we are interested in studying the impact of the raw corpus size, e.g., we could only collect half a million tweets for creating lexicons and analyzing/evaluating the system, while Kiritchenko et al. (2014) built their lexicon on million tweets and evaluated their system on 135 million English tweets. Moreover, we are interested not only in quantity but also in quality, i.e., in studying the quality of the individual words and phrases used as seeds. An interesting work in that direction, even though in a different domain and context, is that of Kozareva and Hovy (2010). We are further interested in finding alternative ways for defining the sentiment polarity, including degree of positive or negative sentiment, and in evaluating them by constructing polarity lexicons in new ways (Severyn and Moschitti, 2015).

More ambitiously, we would like to extend our system to detecting sentiment over a period of time for the purpose of finding trends towards a topic (Nakov et al., 2013; Rosenthal et al., 2014; Rosenthal et al., 2015), e.g., predicting whether the sentiment is strongly negative, weakly negative, strongly positive, etc. We further plan application to other social media services, with the idea of analyzing the sentiment of an online conversation. We would like to see the impact of earlier messages on the sentiment of newer messages, e.g., as in (Vanzo et al., 2014; Barrón-Cedeño et al., 2015; Joty et al., 2015). Finally, we are interested in applying our system to help other tasks, e.g., by using sentiment analysis to finding opinion manipulation trolls in Web forums (Mihaylov et al., 2015a; Mihaylov et al., 2015b).

Acknowledgments

We would like to thank the anonymous reviewers for their constructive comments, which have helped us improve the final version of the paper.

References

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the International Conference on Language Resources and Evaluation*, LREC '10, Valletta, Malta.

Alberto Barrón-Cedeño, Simone Filice, Giovanni Da San Martino, Shafiq Joty, Lluís Màrquez, Preslav Nakov, and Alessandro Moschitti. 2015. Thread-level information for comment classification in community question answering. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, ACL-IJCNLP '15, pages 687–693, Beijing, China.

William Cohen, Pradeep Ravikumar, and Stephen Fienberg. 2003. A comparison of string metrics for matching names and records. In *Proceedings of the KDD workshop on data cleaning and object consolidation*, volume 3, pages 73–78, Washington, D.C., USA.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37.

Tomaž Erjavec. 2012. MULTEXT-East: Morphosyntactic resources for Central and Eastern European languages. *Lang. Resour. Eval.*, 46(1):131–142.

Andrea Esuli and Fabrizio Sebastiani. 2006. SENTIWORDNET: A publicly available lexical resource for opinion mining. In *Proceedings of the International Conference on Language Resources and Evaluation*, LREC '06, pages 417–422, Genoa, Italy.

Andrej Gajduk and Ljupco Kocarev. 2014. Opinion mining of text documents written in Macedonian language. *arXiv preprint arXiv:1411.4472*.

Aniruddha Ghosh, Guofu Li, Tony Veale, Paolo Rosso, Ekaterina Shutova, John Barnden, and Antonio Reyes. 2015. SemEval-2015 task 11: Sentiment analysis of figurative language in Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, SemEval '15, pages 470–478, Denver, CO, USA.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 168–177, Seattle, WA, USA.

Shafiq Joty, Alberto Barrón-Cedeño, Giovanni Da San Martino, Simone Filice, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov. 2015. Global thread-level inference for comment classification in community question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '15, Lisbon, Portugal.

Borislav Kapukaranov and Preslav Nakov. 2015. Fine-grained sentiment analysis for movie reviews in Bulgarian. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, RANLP '15, Hissar, Bulgaria.

Svetlana Kiritchenko, Xiaodan Zhu, and Saif M Mohammad. 2014. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, pages 723–762.

Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. 2011. Twitter sentiment analysis: The good the bad and the OMG! In *Proceedings of the International Conference on Weblogs and Social Media*, ICWSM '11, Barcelona, Spain.

Zornitsa Kozareva and Eduard Hovy. 2010. Not all seeds are equal: Measuring the quality of text mining seeds. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL-HLT '10, pages 618–626, Los Angeles, CA, USA.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–74, 3.

Bing Liu and Lei Zhang. 2012. A survey of opinion mining and sentiment analysis. In Charu C. Aggarwal and ChengXiang Zhai, editors, *Mining Text Data*, pages 415–463. Springer.

Todor Mihaylov, Georgi Georgiev, and Preslav Nakov. 2015a. Finding opinion manipulation trolls in news community forums. In *Proceedings of the Conference on Computational Natural Language Learning*, pages 310–314, Beijing, China.

Todor Mihaylov, Ivan Koychev, Georgi Georgiev, and Preslav Nakov. 2015b. Exposing paid opinion manipulation trolls. In *Proceedings of the Conference on Computational Natural Language Learning*, Hissar, Bulgaria.

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 '13, pages 321–327, Atlanta, GA, USA.

Preslav Nakov, Sara Rosenthal, Zornitsa Kozareva, Veselin Stoyanov, Alan Ritter, and Theresa Wilson. 2013. SemEval-2013 task 2: Sentiment analysis in Twitter. In *Proceedings of the Seventh International Workshop on Semantic Evaluation*, SemEval '13, pages 312–320, Atlanta, GA, USA.

Preslav Nakov, Sara Rosenthal, Svetlana Kiritchenko, Saif Mohammad, Zornitsa Kozareva, Alan Ritter, Veselin Stoyanov, and Xiaodan Zhu. 2015. Developing a successful SemEval task in sentiment analysis of Twitter and other social media texts. *Language Resources and Evaluation*.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, ACL '05, pages 115–124, Ann Arbor, MI, USA.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 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 Conference on Empirical Methods in Natural Language Processing*, EMNLP '02, pages 79–86, Philadelphia, PA, USA.

James W. Pennebaker, Martha E. Francis, and Roger J. Booth. 2001. *Linguistic Inquiry and Word Count*. Lawerence Erlbaum Associates, Mahwah, NJ.

Maria Pontiki, Harris Papageorgiou, Dimitrios Galanis, Ion Androutsopoulos, John Pavlopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation*, SemEval '14, pages 27–35, Dublin, Ireland.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. SemEval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, SemEval '2015, pages 486–495, Denver, CO, USA.

Martin F Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.

Veselin Raychev and Preslav Nakov. 2009. Language-independent sentiment analysis using subjectivity and positional information. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, RANLP '09, pages 360–364, Borovets, Bulgaria.

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 '14, pages 73–80, Dublin, Ireland.

Sara Rosenthal, Preslav Nakov, Svetlana Kiritchenko, Saif Mohammad, Alan Ritter, and Veselin Stoyanov. 2015. SemEval-2015 task 10: Sentiment analysis in Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, SemEval '15, pages 450–462, Denver, CO, USA.

Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1):1–47, March.

Aliaksei Severyn and Alessandro Moschitti. 2015. On the automatic learning of sentiment lexicons. In *Proceedings of the 2015 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1397–1402, Denver, CO, USA.

Philip J. Stone, Dexter C. Dunphy, Marshall S. Smith, and Daniel M. Ogilvie. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press.

Veselin Stoyanov and Claire Cardie. 2008. Topic identification for fine-grained opinion analysis. In *Proceedings of the 22nd International Conference on Computational Linguistics*, COLING '08, pages 817–824, Manchester, United Kingdom.

Peter D. Turney. 2002. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, ACL '02, pages 417–424, Philadelphia, PA, USA.

Vasilija Uzunova and Andrea Kulakov. 2015. Sentiment analysis of movie reviews written in Macedonian language. In *ICT Innovations 2014*, pages 279–288. Springer.

Andrea Vanzo, Danilo Croce, and Roberto Basili. 2014. A context-based model for sentiment analysis in twitter. In *Proceedings of the 25th International Conference on Computational Linguistics*, COLING '14, pages 2345–2354, Dublin, Ireland.

Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2004. Learning subjective language. *Comput. Linguist.*, 30(3):277–308, September.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT-EMNLP '05, pages 347–354, Vancouver, BC, Canada.

Xiaodan Zhu, Svetlana Kiritchenko, and Saif M. Mohammad. 2014. NRC-Canada-2014: Detecting aspects and sentiment in customer reviews. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '14, pages 437–442, Dublin, Ireland.