Two-Step Model for Sentiment Lexicon Extraction from Twitter Streams

Ilia Chetviorkin
Lomonosov Moscow State University
Moscow, Leninskiye Gory 1
ilia.chetviorkin@gmail.com

Natalia Loukachevitch
Lomonosov Moscow State University
Moscow, Leninskiye Gory 1
louk_nat@mail.ru

Abstract

In this study we explore a novel technique for creation of polarity lexicons from the Twitter streams in Russian and English. With this aim we make preliminary filtering of subjective tweets using general domain-independent lexicons in each language. Then the subjective tweets are used for extraction of domain-specific sentiment words. Relying on co-occurrence statistics of extracted words in a large unlabeled Twitter collections we utilize the Markov random field framework for the word polarity classification. To evaluate the quality of the obtained sentiment lexicons they are used for tweet sentiment classification and outperformed previous results.

1 Introduction

With growing popularity of microblogging services such as Twitter, the amount of subjective information containing user opinions and sentiments is increasing dramatically. People tend to express their opinions about events in the real life and such opinions contain valuable information for market research, brand monitoring and political polls.

The task of automatic processing of such informal resources is challenging because people use a lot of slang, vulgarity and out-of-vocabulary words to state their opinions about various objects and situations. In particular, it is difficult to achieve the high quality of sentiment analysis on such type of short informal texts as tweets are. Standard domain-independent lexicon-based methods suffer from low coverage, and for machine learning methods it is difficult to prepare a representative collection of labeled data because topics of discussion are changing rapidly.

Thus, special methods for processing social media data streams should be developed. We proposed and evaluated our approach for Russian language, where only a limited number of natural language processing tools and resources are available. Then to demonstrate the robustness of the method and to compare the results with the other approaches we used it for English.

The current research can be separated into two steps. We start with a special supervised model based on statistical and linguistic features of sentiment words, which is trained and evaluated in the movie domain. Then this model is utilized for extraction of sentiment words from unlabeled Twitter datasets, which are preliminary filtered using the domain-independent lexicons: ProductSentiRus (Chetviorkin and Loukachevitch, 2012) for Russian and MPQA (Wilson et al., 2005) for English.

In the second step an algorithm for polarity classification of extracted sentiment words is introduced. It is built using the Markov random field framework and uses only information contained in text collections.

To evaluate the quality of the created lexicons extrinsically, we conduct the experiments on the tweet subjectivity and polarity classification tasks using various lexicons.

The key advantage of the proposed two-step algorithm is that once trained it can be utilized to different domains and languages with minor modifications. To demonstrate the ability of the proposed algorithm to extract sentiment words in various domains we took significantly different collections for training and testing: movie review collection for training and large collections of tweets for testing.

2 Related work

There are two major approaches for creation of a sentiment lexicon in a specific language: dictionary-based methods and corpus-based methods.

Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 67–72,
Baltimore, Maryland, USA. June 27, 2014.

Dictionary-based methods for various languages have received a lot of attention in the literature (Pérez-Rosas et al., 2012; Mohammad et al., 2009; Clematide and Klenner, 2010), but the main problem of such approaches is that it is difficult to apply them to processing social media. The reason is that short informal texts contain a lot of misspellings and out-of-vocabulary words.

Corpus-based methods are more suitable for processing social media data. In such approaches various statistical and linguistic features are used to discriminate opinion words from all other words (He et al., 2008; Jijkoun et al., 2010).

Another important group of approaches, which can be both dictionary-based and corpus-based are graph-based methods. In (Velikovich et al., 2010) a new method for constructing a lexical network was proposed, which aggregates the huge amount of unlabeled data. Then the graph propagation algorithm was used. Several other researchers utilized the graph or label propagation techniques for solving the problem of opinion word extraction (Rao and Ravichandran, 2009; Speriosu et al., 2011).

In (Takamura et al., 2005) authors describe a probabilistic model for assigning polarity to each word in a collection. This model is based on the Ising spin model of magnetism and is built upon Markov random field framework, using various dictionary-based and linguistic features. In our research, unlike (Takamura et al., 2005) we use only information contained in a text collection without any external dictionary resources (due to the lack of necessary resources for Russian). Our advantage is that we use only potential domain-specific sentiment words during the construction of the network.

A large body of research has been focused on Twitter sentiment analysis during the previous several years (Barbosa and Feng, 2010; Bermingham and Smeaton, 2010; Bifet and Frank, 2010; Davidov et al., 2010; Kouloumpis et al., 2011; Jiang et al., 2011; Agarwal et al., 2011; Wang et al., 2011). In (Chen et al., 2012) authors propose an optimization framework for extraction of opinion expressions from tweets. Using extracted lexicons authors were able to improve the tweet sentiment classification quality. Our approach is based on similar assumptions (like consistency relations), but we do not use any syntactic parsers and dictionary resources. In (Volkova et al., 2013) a new multilingual bootstrapping technique for building tweet sentiment lexicons was introduced. This method is used as a baseline in our work.

3 Data

For the experiments in this paper we use several collections in two domains: movie review collection in Russian for training and fine-tuning of the proposed algorithms and Twitter collections for evaluation and demonstration of robustness in Russian and English languages.

Movie domain. The movie review dataset collected from the online service *imhonet.ru*. There are $28,773$ movie reviews of various genres with numeric scores specified by their authors (DOM).

Additionally, special collections with low concentration of sentiment words are utilized: the contrast collection consists of $17,980$ movie plots (DESC) and a collection of two million news documents (NEWS). Such collections are useful for filtering out of domain-specific and general neutral words, which are very frequent in news and object descriptions.

Twitter collections. We use three datasets for each language: 1M+ of unlabeled tweets (UNL) for extraction of sentiment lexicons, 2K labeled tweets for development data (DEV), and 2K labeled tweets for evaluation (TEST). DEV dataset is used to find the best combination of various lexicons for processing Twitter data and TEST for evaluating the quality of constructed lexicons.

The UNL dataset in Russian was collected during one day using Twitter API. These tweets contain various topics without any filtering. Only strict duplicates and retweets were removed from the dataset. The similar collection for English was downloaded using the links from (Volkova et al., 2013).

All tweets in DEV and TEST collections are manually labeled by subjectivity and polarity using the Mechanical Turk with five workers (majority voting). This data was used for development and evaluation in (Volkova et al., 2013).

4 Method for sentiment word extraction

In this section we introduce an algorithm for sentiment lexicon extraction, which is inspired by the method described in (Chetviorkin and Loukachevitch, 2012), but have more robust features, which allow us to apply it to any unlabeled text collection (e.g. tweets collection). The pro-

posed algorithm is applied to text collections in Russian and English and obtained results are evaluated intrinsically for Russian and extrinsically for both languages.

4.1 An extraction model

Our algorithm is based on several text collections: collection with the high concentration of sentiment words (e.g. DOM collection), contrast domain-specific collection (e.g. DESC collection), contrast domain-independent collection (e.g. NEWS collection). Thus, taking into account statistical distributions of words in such collections we are able to distinguish domain-specific sentiment words.

We experimented with various features to create the robust cross-domain feature representation of sentiment words. As a result the eight most valuable features were used in further experiments:

Linguistic features. Adjective binary indicator, noun binary indicator, feature reflecting part-of-speech ambiguity (for lemma), binary feature of predefined list of prefixes (e.g. *un, im*);

Statistical features. Frequency of capitalized words, frequency of co-occurrence with polarity shifters (e.g. *no, very*), TFIDF feature calculated on the basis of various collection pairs, weirdness feature (the ratio of relative frequencies of certain lexical items in special and general collections) calculated using several pairs of collections.

To train supervised machine learning algorithms all words with frequency greater than three in the Russian movie review collection (DOM) were labeled manually by two assessors. If there was a disagreement about the sentiment of a specific word, the collective judgment after the discussion was used as a final ground truth. As a result of the assessment procedure the list of 4079 sentiment words was obtained.

The best quality of classification using labeled data was shown by the ensemble of three classifiers: Logistic Regression, LogitBoost and Random Forest. The quality according to Precision@n measure can be found in Table 1. This trained model was used in further experiments for extraction of sentiment words both in English and in Russian.

4.2 Extraction of subjective words from Twitter data

To verify the robustness of the model on new unlabeled data it was utilized for sentiment word extraction from multi-topic tweet collection UNL in each language. To apply this model we prepared three collections: domain-specific with high concentration of sentiment words, domain-specific with low concentration of sentiment words and one general collection with low concentration of sentiment words. As the general collection we could take the same NEWS collection (see Section 3) for Russian and British National Corpus[1] for English.

Lexicon	$P@100$	$P@1000$
MovieLex	95.0%	78.3%
TwitterLex	95.0%	79.9%

Table 1: Quality of subjective word extraction in Russian

To prepare domain-specific collections we classified the UNL collections by subjectivity using general purpose sentiment lexicons ProductSentiRus and MPQA in accordance with the language. The subjectivity classifier predicted that a tweet was subjective if it contained at least one subjective term from this lexicon. All subjective tweets constituted a collection with the high concentration of sentiment words and all the other tweets constituted the contrast collection.

Finally, using all specially prepared collections and the trained model (in the movie domain), new lexicons of twitter-specific sentiment words were extracted. The quality of extraction in Russian according to manual labeling of two assessors can be found in Table 1. The resulting quality of extracted Russian lexicon is on the same level as in the initial movie domain, what confirms the robustness of the proposed model.

We took 5000 of the most probable sentiment words from each lexicon for further work.

5 Polarity classification using MRF

In the second part of current research we describe an algorithm for polarity classification of extracted sentiment words. The proposed method relies on several assumptions:

- Each word has the prior sentiment score calculated using the review scores where it appears (simple averaging);
- Words with similar polarity tend to co-occur closely to each other;

[1] http://www.natcorp.ox.ac.uk/

- Negation between sentiment words leads to the opposite polarity labels.

5.1 Algorithm description

To formalize all these assumptions we construct an undirected graphical model using extracted sentiment word co-occurrence statistics. Each extracted word is represented by a vertex in a graph and an edge between two vertexes is established in case if they co-occur together more than once in the collection. We drop all the edges where average distance between words is more than 8 words.

Our model by construction is similar to approach based on the Ising spin model described in (Takamura et al., 2005). Ising model is used to describe ferromagnetism in statistical mechanics. In general, the system is composed of N binary variables (spins), where each variable $x_i \in \{-1, +1\}, i = 1, 2, ..., N$. The energy function of the system is the following:

$$E(x) = -\sum_{ij} s_{ij} x_i x_j - \sum_i h_i x_i \qquad (1)$$

where s_{ij} represents the efficacy of interaction between two spins and h_i stands for external field added to x_i. The probability of each system configuration is provided by Boltzmann distribution:

$$P(X) = \frac{\exp^{-\beta E(X)}}{Z} \qquad (2)$$

where Z is a normalizing factor and $\beta = (T^{-1} > 0)$ is inverse temperature, which is parameter of the model. We calculate values of $P(X)$ with several different values of β and try to find the locally polarized state of the network.

To specify the initial polarity of each word, we assume that each text from the collection has its sentiment score. This condition is not very strict, because there are a lot of internet review services where people assign numerical scores to their reviews. Using such scores we can calculate the deviation from the average score for each word in the collection:

$$h(i) = E(c|w_i) - E(c)$$

where c is the review score random variable, $E(c)$ is the expectation of the score in the collection and $E(c|w_i)$ is the expectation of the score for reviews containing word w_i. Thus we assign the initial weight of each vertex i in the MRF to be equal to $h(i)$.

To specify the weight of each edge in the network we made preliminary experiments to detect the dependency between the probability of the word pair to have similar polarity and average distance between them. The result of such experiment for movie reviews can be found on Figure 1. One can see that if the distance between the words

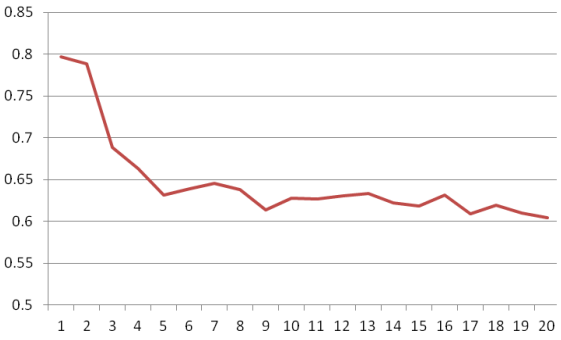

Figure 1: The dependency between the probability to have similar polarity and average distance

is above four, then the probability is remain on the same level which is slightly biased to similar polarity. Relying on this insight and taking into account the frequency of co-occurrence of the words we used the following edge weights:

$$s(i,j) = f(w_i, w_j) \max\Big(0.5 - \frac{d(w_i, w_j)}{d(w_i, w_j) + 4}, 0\Big)$$

where $f(w_i, w_j)$ is the co-occurrence frequency in the collection and $d(w_i, w_j)$ is the average distance between words w_i and w_j.

Finally, we revert the sign of this equation in case of more than half of co-occurrences contains negation (*no, not, but*) between opinion words.

In practice we can find approximate solution using such algorithms as: Loopy Belief Propagation (**BP**), Mean Field (**MF**), Gibbs Sampling (**Gibbs**).

The performance of the methods was evaluated for a lexical network constructed from the first 3000 of the most probable extracted sentiment words in the movie review collection (DOM). We took from them 822 interconnected words with strict polarity labeled by two assessors as a gold standard. Testing was performed by varying β from 0.1 to 1.0. The primary measure in this experiment was accuracy. The best results can be found in Table 2.

The best performance was demonstrated by **MF** algorithm and $\beta = 0.4$. This algorithm and parameter value were used in further experiments on unlabeled tweet collections.

β	**BP**	**MF**	**Gibbs**
0.4	83.8	**85.2**	**83.7**
0.5	83.6	84.5	82.0
0.6	**85.0**	83.1	79.4

Table 2: Dependence between the accuracy of classification and β

5.2 Polarity classification of subjective words from Twitter data

Using the general polarity lexicons we classify all subjective tweets in large UNL collections into positive and negative categories. For the polarity classifier, we predict a tweet to be positive (negative) if it contains at least one positive (negative) term from the lexicon taking into account negation. If a tweet contains both positive and negative terms, we take the majority label. In case if a tweet does not contain any word from the lexicon we predict it to be positive.

These labels ($+1$ for positive and -1 for negative) can be used to compute initial polarity $h(i)$ for all extracted sentiment words from the UNL collections. The weights of the links between words $s(i, j)$ can be also computed using full unlabeled collections.

Thus, we can utilize the algorithm for polarity classification of sentiment words extracted from Twitter. The resulting lexicon for Russian contains 2772 words and 2786 words for English (we take only words that are connected in the network). To evaluate the quality of the obtained lexicons the Russian one was labeled by two assessors. In result of such markup 1734 words with strict positive or negative polarity were taken. The accuracy of the lexicon on the basis of the markup was equal to **72%**, which is 1.5 % better than the simple average score baseline.

6 Lexicon Evaluations

To evaluate all newly created lexicons they were utilized in tweet polarity and subjectivity classification tasks using the TEST collections. The results of the classification for both languages can be found in Table 3 and Table 4.

As one can see, the newly created Twitter-specific sentiment lexicon results outperform the result of (Volkova et al., 2013) in subjectivity classification for Russian but slightly worse than the result for English. On the other hand the results of polarity classification are on par or better

Lexicon	P	R	F_{subj}
Russian			
Volkova, 2013	-	-	61.0
TwitterLex	60.2	79.3	68.5
English			
Volkova, 2013	-	-	75.0
TwitterLex	58.8	95.5	73.0

Table 3: Quality of tweet subjectivity classification

Lexicon	P	R	F_{pol}
Russian			
Volkova, 2013	-	-	73.0
TwitterLex	65.5	82.0	72.8
Combined	65.8	85.5	74.3
English			
Volkova, 2013	-	-	78.0
TwitterLex	72.1	88.1	79.3
Combined	73.2	89.3	80.4

Table 4: Quality of tweet polarity classification

than the results of (Volkova et al., 2013) lexicons bootstrapped from domain-independent sentiment lexicons. Thus, to push the quality of polarity classification forward we combined the domain-independent lexicons and our Twitter-specific lexicons. We experimented with various word counts from general lexicons and found the optimal combination on the DEV collection: all words from TwitterLex and 2000 the most strong sentiment words from ProductSentiRus in Russian and all strong sentiment words from MPQA in English. The lexicon combination outperforms all previous results by F-measure leading to the conclusion that proposed method can capture valuable domain-specific sentiment words.

7 Conclusion

In this paper we proposed a new method for extraction of domain-specific sentiment lexicons and adopted the Ising model for polarity classification of extracted words. This two-stage method was applied to a large unlabeled Twitter dataset and the extracted sentiment lexicons performed on the high level in the tweet sentiment classification task. Our method can be used in a streaming mode for augmentation of sentiment lexicons and supporting the high quality of multilingual sentiment classification.

Acknowledgements This work is partially supported by RFBR grant 14-07-00682.

References

Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. 2011. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Languages in Social Media*, pages 30–38. Association for Computational Linguistics.

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

Adam Bermingham and Alan F Smeaton. 2010. Classifying sentiment in microblogs: is brevity an advantage? In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1833–1836. ACM.

Albert Bifet and Eibe Frank. 2010. Sentiment knowledge discovery in twitter streaming data. In *Discovery Science*, pages 1–15. Springer.

Lu Chen, Wenbo Wang, Meenakshi Nagarajan, Shaojun Wang, and Amit P Sheth. 2012. Extracting diverse sentiment expressions with target-dependent polarity from twitter. In *ICWSM*.

Ilia Chetviorkin and Natalia V Loukachevitch. 2012. Extraction of russian sentiment lexicon for product meta-domain. In *COLING*, pages 593–610.

Simon Clematide and Manfred Klenner. 2010. Evaluation and extension of a polarity lexicon for german. In *Proceedings of the First Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, pages 7–13.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 241–249. Association for Computational Linguistics.

Ben He, Craig Macdonald, Jiyin He, and Iadh Ounis. 2008. An effective statistical approach to blog post opinion retrieval. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 1063–1072. ACM.

Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. 2011. Target-dependent twitter sentiment classification. In *ACL*, pages 151–160.

Valentin Jijkoun, Maarten de Rijke, and Wouter Weerkamp. 2010. Generating focused topic-specific sentiment lexicons. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 585–594. Association for Computational Linguistics.

Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. 2011. Twitter sentiment analysis: The good the bad and the omg! In *ICWSM*.

Saif Mohammad, Cody Dunne, and Bonnie Dorr. 2009. Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2*, pages 599–608. Association for Computational Linguistics.

Verónica Pérez-Rosas, Carmen Banea, and Rada Mihalcea. 2012. Learning sentiment lexicons in spanish. In *LREC*, pages 3077–3081.

Delip Rao and Deepak Ravichandran. 2009. Semi-supervised polarity lexicon induction. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 675–682. Association for Computational Linguistics.

Michael Speriosu, Nikita Sudan, Sid Upadhyay, and Jason Baldridge. 2011. Twitter polarity classification with label propagation over lexical links and the follower graph. In *Proceedings of the First workshop on Unsupervised Learning in NLP*, pages 53–63. Association for Computational Linguistics.

H. Takamura, T. Inui, and M. Okumura. 2005. Extracting semantic orientations of words using spin model. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 133–140.

Leonid Velikovich, Sasha Blair-Goldensohn, Kerry Hannan, and Ryan McDonald. 2010. The viability of web-derived polarity lexicons. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 777–785. Association for Computational Linguistics.

Svitlana Volkova, Theresa Wilson, and David Yarowsky. 2013. Exploring sentiment in social media: Bootstrapping subjectivity clues from multilingual twitter streams. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL13)*, pages 505–510.

Xiaolong Wang, Furu Wei, Xiaohua Liu, Ming Zhou, and Ming Zhang. 2011. Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 1031–1040. ACM.

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*, pages 347–354. Association for Computational Linguistics.