Bull. Nov. Comp. Center, Comp. Science, 37 (2014), 107–116

An approach to construction and analysis of a corpus of short Russian texts intended to train a sentiment classifier

Yu. V. Rubtsova, Yu. A. Zagorulko

Abstract. The paper describes a method for construction and annotation of a corpus of short texts made up on the basis of Russian posts from Twitter. This corpus is intended to train a sentiment classifier that sorts the general topic texts into three classes: "positive", "negative", and "neutral". The corpus is morphologically tagged in order to identify the characteristic features of each of the three classes of short texts. Parts of speech and unigrams and bigrams of terms were selected as the characteristic features. A vocabulary of emotional words was constructed based on the corpus; the weight of each term in the corpus was calculated.

Keywords: corpus linguistics, text classification and categorization, text annotation, morphological tagging, social network data analysis.

1. Introduction

Human speech is constantly changing and evolving: new words are included in the active vocabulary, while the old ones cease to be used. Also, a natural language is constantly evolving together with the spoken language. New words are born every day, and about half of them are slang. Slang quickly reacts to changes in all spheres of our life and allows us to express complex notions in a concise and understandable form. Slang is actively used in conversations and friendly dialogues on social networking sites, especially to express an emotional attitude towards a particular issue. In this regard, it is essential to take slang into consideration when designing sentiment classifiers, particularly when creating a vocabulary of emotional words.

An approach to constructing and analysing a corpus of annotated texts intended to train the sentiment classifiers is considered in the paper. The automatic classification of texts expressing some opinions was carried out by the method proposed by Jonathon Read in [1]. After collecting the texts, the corpus processing included two stages:

1. Morphological tagging of the corpus: all words in each text of the corpus are tagged (labelled) with the part of speech (POS) tags. The TreeTagger tool [11] was used for morphological analysis.
2. Making an emotional vocabulary: a list of the most frequently used words and phrases is extracted from the texts of the corpus in order

to define the evaluative lexicon of general topic opinions. In addition, a set of statistical characteristics is calculated for each word and a list of general topic evaluative words and phrases used to express positive or negative views is composed.

The paper presents a corpus of short texts in modern Russian, its initial linguistic analysis, the analysis of the results of its morphological tagging, the vocabulary of emotional words built on the basis of the corpus, as well as a brief analysis of changes in the use of various word forms over time. The obtained results are used to build and train a sentiment classifier.

2. An overview of corpora for sentiment analysis

One of the ways to organise textual information for its subsequent analysis is to create a text corpus. A text corpus is a collection of categorised texts composed using a certain method and presented electronically. The texts are categorised according to both the integral characteristics of each text (e.g., a certain subject matter) and the specific characteristics of certain terms (e.g., wordform, lemma, and morpheme). Moreover, this collection of texts should be organised as a database to make possible a practical use of the corpus when solving various tasks such as dictionary building, machine learning, or testing a morphological analyzer, parser and text classifiers of various types.

In recent years, a lot of research has been conducted in the field of text sentiment classification. Many works are devoted to the sentiment classification of product and film reviews [2, 3], as well as to the analysis of blogs or news. In other words, rather long passages of texts related to a given domain have been studied. Sometimes, the texts of reviews used in the studies were marked on a five or ten point scale by the author himself [4]. Therefore, the collections available in the Russian language prepared for the automatic classification of reviews into two or three classes are the collections united by one topic, for example, a collection of film reviews with user ratings (ROMIP 2011, [5]). In [2–5], the training corpora with the following parameters were developed:

- reviews with ratings manually entered by the author;
- highly specific reviews (film or book reviews);
- important news (long texts consisting of several paragraphs).

All the available collections are collections of reviews belonging to a certain domain and not general topic collections of short texts (microblogs).

Twitter is a social networking and microblogging service that allows its users to write messages in real time. Tweets are often posted directly from a mobile device and from the location where the event is taking place, which

adds emotions to posts. Because of the platform's limit, the length of a post on Twitter may not exceed 140 characters, so people use abbreviations, shortened words, and emoticons and intentionally misspell words. As Twitter has the features of a social network, its users are able to express their opinion on a variety of issues and actively do so ranging from the quality of multi-cookers to inter-national economic and political developments.

Classification of short phrases and expressions, rather than paragraphs or entire documents, has been carried out by [6]. In their study, the authors showed that it is often important to determine the sentiment (positive or negative) of a single sentence, not of the text in its entirety. In a long document, the author's views on the subject may change from positive to negative and vice versa, as he/she may speak negatively about minor flaws, but on the whole maintain a positive attitude towards the subject. In other words, it is not always possible to classify clearly a long document or review as having a positive or negative sentiment.

3. An approach to making up a corpus based on Russian posts from Twitter

The method described in [1] showed the effectiveness of the use of emoticons (special symbols denoting emotions in written communications) for an automatic classification of texts into positive and negative. The emotion of the post can be detected with high accuracy if the author has inserted an emoticon. For this reason, the vocabularies of characters representing the positive or negative attitude of the author were made first. The Wikipedia resource[1] was used to collect the vocabulary of symbols that designate emotions. For example, the icon ":)" stands for a positive emotion and ":(" represents a negative one. Since the length of a post is limited to 140 characters, it was assumed that an emoticon refers to the whole post, not just to a part of it.

Posts with positive and negative sentiments were searched for based on the written symbols for emotions, and two collections were formed. These collections will be used for further analysis of posts with positive and negative sentiments and for the identification of patterns in positive and negative posts. Posts taken from news microblogging accounts formed a collection of neutral posts.

3.1. Filtering

Filtering was carried out to maintain the experimental integrity:

- Texts containing both positive and negative emotions were deleted from the collection. Such texts cannot be automatically attributed to either collection of posts (positive or negative).

[1] `http://en.wikipedia.org/wiki/List_of_emoticons`

- Non-informative tweets (less than 40 characters long) were also deleted because most of them do not make any sense being a part of a dialogue or clipped reply.
- Texts containing non-UTF-8 emoticons were deleted as the developed tools cannot process non-UTF-8 symbols.
- Texts containing non-Cyrillic symbols were deleted as the developed tools work only with the Russian language.

3.2. Characteristics of the corpus

It took several weeks to create the corpus because of the Twitter API constraint: we can get only 100 posts for one query. As a result, a collection consisting of approximately 15 million short posts was formed using the method from [1]. After filtering this collection, we obtained a corpus consisting of the following collections:

- positive posts 114,911 texts,
- negative posts 111,923 texts,
- neutral posts 107,990 texts.

The number of word forms in the united collection (positive, negative and neutral) is 5,867,082; the number of unique word forms in this collection is 454,603.

A database was constructed from the microblog texts to provide further processing and analysis of the corpus. Each text in the corpus has the following attributes: the date of publication, the name of the author, the text of the tweet, the class to which the text belongs (positive, negative, neutral), the number of times the post was favourite, the number of retweets, the number of the users friends, the number of the users followers, and the number of lists the user is included in.

3.3. Uniformity of the corpus

To build a vocabulary of emotional words, it is necessary to have a collection of posts containing a fairly large number of lemmas. The Russian language is rich and various, but not all of its words are used for communication on social networking sites. One of the objectives of this study is to collect a sufficiently representative corpus for building a vocabulary of emotional words. A "sufficiently representative corpus" means that adding new tweets to the collection would lead to a very small number of new lemmas being added. To check whether the corpus is sufficiently representative, the three collections were combined into one. Then the number of unique lemmas was calculated for collections of different size. Figure 1 shows that when

the number of tweets is small, adding new posts to the collection causes an increase in the number of unique lemmas. But when the number of unique lemmas is about 340,000, adding new tweets to the collection does not result in a significant increase in the number of unique lemmas.

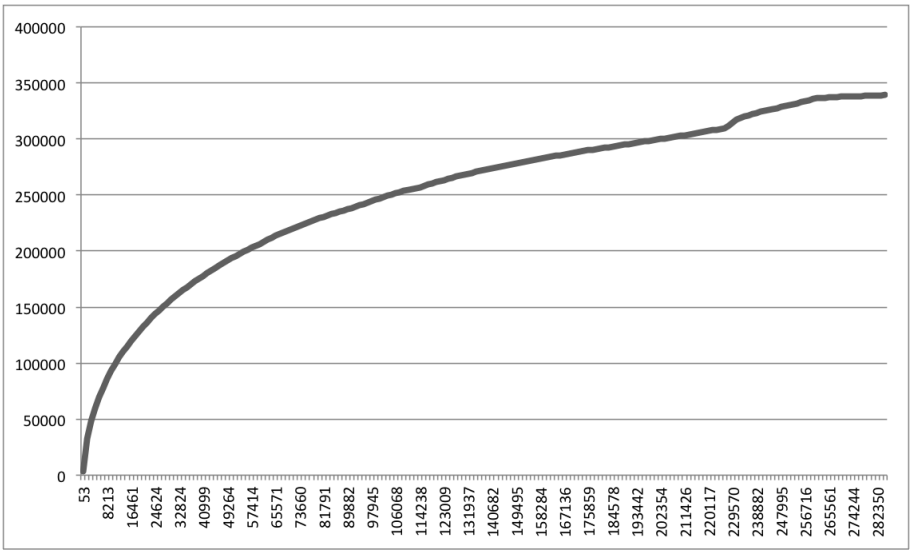

Figure 1. Distribution of the number of unique terms, depending on the number of tweets

4. Morphological tagging

There exist several morphological analysers of the Russian language. For example, mystem is a morphological analyser developed by the search engine Yandex. Mystem is able to guess the normal forms of words not present in the system dictionary [12], but it does not resolve morphological ambiguity and outputs all possible paradigms as a result. Another morphological analyser, Myaso[2], is a language-independent tool for the morphological tagging of texts based on Hidden Markov Models. It uses remote interpolation and the Viterbi algorithm. Unfortunately, the current implementation of Myaso has two significant drawbacks:

1. the Myaso tagger crashes when it discovers a word in the text that is not present in the vocabulary;
2. though a text is tagged rather quickly, the initial launch of the Myaso analyser needs much time (about a minute).

[2] `http://nlpub.ru/Myaso`

TreeTagger for the Russian language [11, 13] was selected to tag the training corpus. TreeTagger is a probabilistic tool for text tagging that allows us to resolve the morphosyntactic ambiguities of the Russian language. The task of studying the corpus consists of two subtasks:

1. Identifying the distribution patterns of the parts of speech between collections known to be either including or not including emotive expressions.
2. Identifying the distribution patterns of the parts of speech between the "positive" and "negative" collections.

The analysis showed that the authors are inclined to use different parts of speech and forms depending on whether they are expressing a positive/negative sentiment or not. For example, the authors of tweets containing positive emotions often use verbs in the perfective aspect of the present tense passive, third-person singular. In turn, the authors of tweets with negative sentiments most commonly use verbs in the perfective aspect of the past tense active, feminine singular in their posts Figure 2. The parts of speech in bigrams were also analysed. In the terms of word pairs, the authors of positive tweets often use the combination of a perfect passive masculine singular verb in the past tense and a cardinal numeral. The authors of negative tweets often use a combination of an inanimate feminine plural accusative noun and a third-person dative plural pronoun.

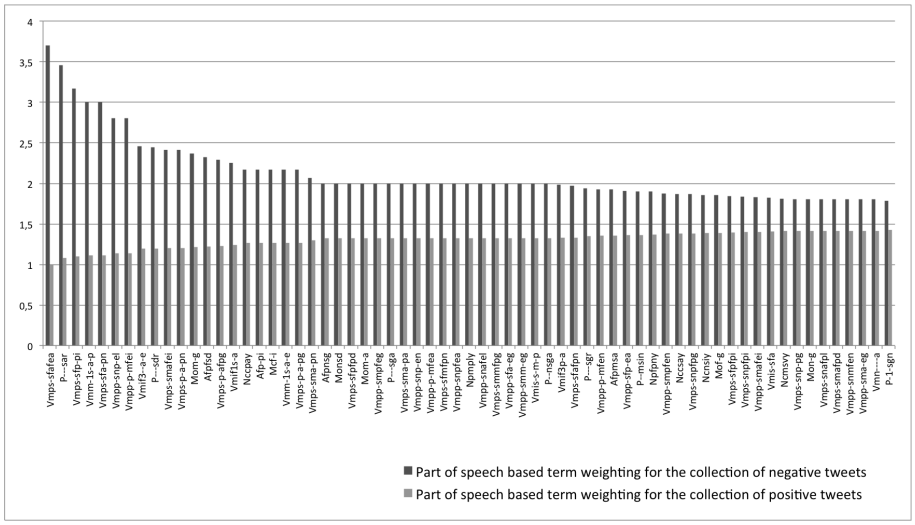

Figure 2. A graph comparing the usage frequency of parts of speech in the collections of positive and negative texts ordered relative to their usage frequency in the negative texts.

5. Building an emotional vocabulary

5.1. A vocabulary of emotional words

In the terms of corpus linguistics, the extraction methods based on the measure of significance of a term for the collection are widely used, for example, the term weighting scheme called TF-IDF [9]. It is shown in [7] that the RF scheme demonstrates better results when calculating the weight of a term with allowances made for its belonging to different classes. The essence of the method based on RF is that the weight of a word (term) is calculated based on the information about the distribution of the term in the texts of a collection taking into account belonging of each of these texts to a certain class (positive, negative, or neutral).

Therefore, in this study the weight of each term in each collection was calculated using the term weighting scheme RF.

Let a be the number of tweets containing a term T and belonging to a certain class C, and c be the number of tweets that contain the same term T and do not belong to the class C. In this case, the significance (rf) of the term T for the class C is expressed as follows:

$rf = log(2 + \frac{a}{max(1,c)})$.

The table shows five most significant terms for the collections of positive and negative tweets based on the weighting schema RF.

Table. Five most significant terms for the collections of positive and negative tweets based on the weighting schema RF

Positive word	**Weight of the word in the collection of positive tweets**	**Negative word**	**Weight of the word in the collection of negative tweets**
ofigenny (awesome <short form plural>)	7.3487	pogibshikh (dead <genitive plural>)	6.2668
pozitiva (positive <gen sing>)	6.6439	terakt (terrorist attack)	6.2479
pozdravlyaem (we congratulate)	5.9542	nekrolog (obituary)	6.1898
razblokiruyut (will unblock)	5.9542	Ohoo	5.858
oglyadyvayutsya (they look around)	5.9307	skorbim (we mourn)	5.8074

The developed vocabulary consists of 19,039 word forms that do not be-

long to any predetermined domain. The comparison of the results obtained in this study with the results of a previous one [10] shows that a natural language is constantly changing and, to solve more accurately the problem of classification of short texts into three classes (positive, negative, and neutral), it is necessary to keep up-to-date the text corpora and the weights of terms in the vocabulary of emotional words.

Most of the existing automatic and semi-automatic term weighting schemes are based on the assumption that all the data are known in advance, accessible and static. For example, to use the term weighting schemes, such as TF-IDF [9] and TF-RF [7], it is necessary to know the frequency of occurrence in the document for each term; therefore the data set should not be changed during calculation. This is a serious complication in case data should be calculated in real time. For example, when adding a new text to a collection, it is necessary to recalculate the weights for all terms in the collection. The computational complexity of recalculating all weights in the collection is $O(N^2)$. The use of the term weighting schemes TF-ICF [8] to calculate the weights of terms in a collection in real time is one of the prospects for the future study.

5.2. Vocabulary of Emotional Bigrams

In addition to the vocabulary of unigrams used in the collections, a bigram vocabulary was constructed. Bigram weights were also calculated according to the RF term weighting scheme. The most significant bigrams for the collections of positive tweets are: “begal kak (ran like), “den pozitiva (a day of positive), “idiot tsely (total idiot); for the collection of negative tweets: “kto nenavidit (who hates), “nenavidit probki (hates traffic jams), “probki retvit (traffic jams retweet).

6. Conclusion

The tools for corpus constructing and updating have been developed. Using these tools, a corpus of short Russian texts has been built on the basis of emotional posts from Twitter. The corpus was automatically divided into three classes: “positive”, “negative” and “neutral”, and now it contains 111,923 positive, 114,911 negative, and 107,990 neutral tweets.

Each text in the corpus is supplied with a set of attributes that allow us to draw the conclusions about the relevance of the statement, the strength of its impact on the reader and the importance of the post. The corpus is morphologically tagged. Following the morphological tagging, the patterns of dependence of the post sentiment on the parts of speech used in it were revealed. Based on the corpus, a general topic vocabulary of emotional words including both unigrams and bigrams was also built.

The corpus is suitable to train a sentiment classifier operating at the sentence level.

The program module developed during this study allows us to monitor and take into account lexical changes in a spoken language. The corpus in the form of a database is available for public access at http://study.mokoron.com.

References

[1] Jonathon R. Using emoticons to reduce dependency in machine learning techniques for sentiment classification // Proc. of the ACL Student Research Workshop – Association for Computational Linguistics, 2005. – Ann Arbor, Michigan, 2005. – P. 43–48.

[2] Chetviorkin I. I., Loukachevitch N. V. Crossdomain opinion word extraction model // Proc. of 6-th Russian Yong Schientists Conf. in Information Retrieval. – Yaroslavl, 2012. – P. 5–15.

[3] Pang B, Lillian L, Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques // Proc. of the 2002 Conf. on Empirical Methods in Natural Language Processing. – Univ. of Pennsylvania, 2002. – P. 79–86.

[4] Pang B., Lee L. Seeing stars: exploiting class relationships for sentiment categorization with respect of rating scales // Proc. of ACL, 43rd Meeting of the Association for Computational Linguistics 2005. – Ann Arbor, Michigan, 2005. – P. 115–124.

[5] ROMIP: Russian Information Retrieval Evaluation Seminar. – http://romip.ru/ru/collections/imhonet-films.html.

[6] Wilson T., Wiebe J., Hoffmann P. Recognizing contextual polarity in phreselevel sentiment analysis // Proc. of Human Languages Technologies Conf. / Conf. on Emperical Methods in Natural Language Processing – HLT/EMNLP 2005. – Vancouver, CA, 2005.

[7] Lan M., Tan C. L., Su J., Lu Y. Supervised and traditional term weighting methods for automatic text categorization // IEEE Trans. on Pattern Analysis and Machine Intelligence. – 2009. – Vol. 31, No. 4. – P. 721–735.

[8] Reed J. W., Jiao Yu, Potok T. E., Klump B. A., Elmore M. T., Hurson A. R. TF-ICF: A new term weighting scheme for clustering dynamic data streams // Proc. of Machine Learning and Applications – ICMLA '06. – 2006. – P. 258–263.

[9] Salton G.,Buckley C. Term-weighting approaches in automatic text retrieval // J. of Information Processing and Management. – 1988. – Vol. 24(5). – P. 513–523.

[10] Rubtsova Y. V. A method for development and analysis of short text corpus for the review classification task // Proc. of the XVth All-Russian Scientific Conf. RCDL–2013. – Russia, 2013. – P. 269–275.

[11] Schmid H. Probabilistic part-of-speech tagging using decision trees // Proc. of the Internat. Conf. on New Methods in Language Processing, 2004. – P. 44–49.

[12] Segalovich I. A fast morphological algorithm with unknown word guessing induced by a dictionary for a Web Search Engine // MLMTA. – 2003. – P. 273–280.

[13] Sharoff S., Nivre J. The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge // Proc. of Dialogue – Russian Conf. on Computational Linguistics, 2011.