Sentiment Classification into Three Classes Applying Multinomial Bayes Algorithm, N-grams, and Thesaurus

Ksenia Lagutina, Vladislav Larionov, Vladislav Petryakov, Nadezhda Lagutina, Ilya Paramonov, Ivan Shchitov
P.G. Demidov Yaroslavl State University
Yaroslavl, Russia
ksenia.lagutina@fruct.org, vladlarionov998@gmail.com, petryakov.v@inbox.ru, lagutinans@gmail.com,
Ilya.Paramonov@fruct.org, ivan.shchitov@fruct.org

***Abstract*—The paper is devoted to development of the method that classifies texts in English and Russian by sentiments into positive, negative, and neutral. The proposed method is based on the Multinomial Naive Bayes classifier with additional n-grams application. The classifier is trained either on three classes, or on two contrasting classes with a threshold to separate neutral texts. Experiments with texts on various topics showed significant improvement of classification quality for reviews from a particular domain. Besides, the analysis of thesaurus relationships application to sentiment classification into three classes was done, however it did not show significant improvement of the classification results.**

I. Introduction

One of the main sentiment classification tasks that is deeply investigated by researchers, is classification of texts by overall sentiment. This level of classification implies that the text expresses only one sentiment or opinion about a topic. Such texts are primarily short texts like tweets, microblogs, or reviews [1].

The most common classifiers of texts distinguish two or three classes. In the first case the classes are contrasting: positive and negative [2]. In the second case these classes are supplemented by a neutral sentiment [3].

Usually investigators use two types of methods for development of sentiment analysis algorithms: lexical, which are based primarily on dictionaries or databases with sentiment vocabulary [4], and machine learning approaches [5].

Researchers note [6] that sentiment analysis for English texts generally involves machine learning methods, which show higher quality than the others. However, for other languages, particularly, for Russian, usage of lexical methods prevails [7]. One of the reasons is a lack of labeled data collections that can be used for training of machine learning models. Sometimes, during creation of classifiers for national languages with use of sentiment dictionaries, developers even have to translate the data of English-language dictionaries into the desired language [8]. Thus, an area of sentiment analysis tools for national texts needs more deep investigations.

The goal of our research is to develop a method of sentiment text classification into three classes that can be suitable for two languages: English and Russian. We choose machine learning algorithms as the basis of the method. Additionally, we set a subtask to study the possibility of using an automatically generated thesaurus for the sentiment classification of texts.

The paper is structured as follows. In Section II we describe state-of-the-art in sentiment text classification into three classes for English and Russian languages. Section III introduces the method of classification based on machine learning techniques, n-grams, and thesauri. In Section IV we provide results of our experiments with our method and its modification that uses a thesaurus. Section V discusses advantages and limitations of our method, including thesaurus application, and future investigations. Conclusion summarizes the paper.

II. State-of-the-art

Most researchers in the field of sentiment classification into three texts use corpora of tweets as a dataset for experiments, less frequently they classify reviews. Other types of texts appears in similar studies very rarely and usually are split into several parts to be classified by sentiments.

Sentiment classification of tweets into three classes was the topic of the SemEval-2017 contest, as a part of the "Sentiment Analysis in Twitter" task [9]. Participants proposed many methods for English-language tweets classification based on different mathematical, statistical, and linguistic algorithms. The best results had average recall, F-measure, and accuracy around 0.63–0.68, and most of them were achieved by neural networks.

Generally in recent years neural networks combined with linguistic approaches show one of the highest sentiment classification quality. For example, Vo and Zhang [10] apply word embeddings and context features of tweets as an input in neural pooling functions. This method's accuracy is 0.71 and F-measure is 0.70. Cates et al. [11] construct feature vectors for Youtube comments basing on emoticons' presence or absence and classify them using the Naive Bayes model or recurrent neural network. Both classifiers show high accuracy: 0.86 by Naive Bayes and 0.81 by the neural network.

The idea of combining different linguistic and mathematical approaches is state-of-the-art. Kolovou et al. [12] proposed the fusion of several classification systems. They calculate vectors with different statistical and semantic types of features, process them by various classifiers: convolutional

neural network (CNN), Word2vec, Naive Bayes, Webis [13], and take the average result. Such a system was one of the best in SemEval-2017.

Machine learning techniques also give high classification quality not only for tweets [14], [15], but also for longer texts and reviews. The system developed by Tripathy et al. [16] achieves around 0.80–0.88 accuracy for the IMDb Dataset calculating statistical features for n-grams and applying the most popular machine learning algorithms: Naive Bayes, maximum entropy, SVM, stochastic gradient descent. Kaur et al. [17] use a similar approach combining Gini Index with Random Forest classifier and SVM and get accuracy and F-measure about 0.75–0.80.

Unfortunately, classification of English-language reviews into three classes remain understudied: articles devoted to this problem are less common than investigations of tweets classification and binary sentiment classification of short texts or reviews. Linguistic approaches are rarely applied to three-class sentiment classification, although they show their usefulness in combinations with other algorithms for tweets [10], [12] and achieve 0.65–0.70 average recall and F-measure.

Most articles state classification problems and propose solutions for English-language texts. If we investigate the field of Russian text classification by sentiments, we can see that the task of three-class classification is not very popular: most of researchers talk about binary classification. One of the biggest investigations for sentiment classification into three classes was SentiRuEval [18], where participants classified reviews about banks and telecom companies. Results were quite low: 0.54 for the telecom domain and 0.37 for the bank domain, in comparison with English language, for which the best results of standard metrics achieve 0.8 and higher.

Volkova et al. [19] create sentiment lexicons for English, Spanish, and Russian languages. They use bootstrapping methods and translate terms from the English lexicon to other languages. The application of this lexicon combined with the emoticons and hashtags use, allows to achieve 0.67–0.70 F-measure for Spanish and Russian. Therefore, lexical methods can improve classification quality for national languages. Thus, sentiment classification of reviews into three classes requires additional research for both English and Russian language.

III. Sentiment classification method

A. Method overview

The method for sentiment classification into three classes (positive, negative, and neutral) continues our previous investigations in the field of text classification [20], where we performed sentiment classification of Russian-language tweets, blogs, and reviews into two classes: positive and negative. The method proposed in this paper, also applies the Multinomial Naive Bayes machine learning classifier [21], but uses different features for text vectors and two schemes for the training phase to be suitable for better search of neutral texts.

The main stages of the method are the following:

1) Preprocessing texts for classification: lemmatization and calculation of numbers of word occurrences.
2) Training the classifier on two or three classes.
3) Classifying texts into three classes.

On the preprocessing stage we split texts into not only unigrams, but also bigrams. It allows to find more thesaurus terms in texts and apply more relationships between them. For lemmatization we use the Snowball stemmer.

Last two stages represent the supervised classification method based on the Multinomial Naive Bayes model [21] that we train on two or three classes. Training and testing on three classes is the baseline method in our experiments. Training on two classes requires an additional procedure on the test step that extracts neutral texts, basing on their probabilities of belonging to positive or negative classes.

Besides, we can combine this algorithm with a procedure that computes sentiments using not only occurrences of words in texts, but also thesaurus relationships between them.

Let us discuss main steps of the method and thesaurus application in more details.

B. Training and classification

We train the classifier using two classes only: positive and negative, or using all three classes. We compare a priori probabilities that a particular word belongs to a particular class. These probabilities are based on numbers of occurrences in texts from different classes. The term gets the sentiment of the class, where it appears more frequently.

During the testing step we calculate a posteriori probabilities of each word or bigram that they belong to particular classes, as sums of logarithmic a priori probabilities of classes and their words and bigrams. The highest probability defines the sentiment of the word or bigram.

All calculated probabilities become features for classification vectors. Each text has the vector, which elements correspond all words and bigrams found in the corpus. If a text contains a particular phrase, the corresponding feature equals the phrase's a posteriori probability, else it equals zero. Vectors are classified by the Naive Bayes algorithm.

If we train the algorithm on all classes, on the last stage we compute probabilities of each text that it belongs to the particular class and choose the highest score. It is standard classification into three separate classes, when we assume that each class has its own set of features that are revealed on the training phase and applied for classification.

The algorithm with the training on positive and negative classes only is based on the idea that neutral class is intermediate between positive and negative. Neutral texts either contain both positive and negative words or do not have emotional phrases.

If we train the algorithm on positive and negative classes only, on the last stage we classify all texts from the test set into two classes, i.e., compute only two probabilities. If one probability is significantly greater than another one, that means the text has a particular emotional sentiment: positive or negative. In the case when probabilities are equal or slightly different, we interpret this text as neutral. The threshold for difference between probabilities is the method's parameter and can be varied.

If our method discovers both sentiments in the same text and computes similar probabilities that the text is positive or negative, it considers it properly as neutral. If our method does not find phrases and emotional sentiments and calculate low probabilities to belong to the particular class, it also makes this text neutral.

C. Thesaurus application

The additional subtask of our research is thesaurus application to sentiment classification into three texts and analysis of influence of its terms and relationships on the result.

1) Thesaurus generation: Our method generates a specialized thesaurus fully automatically, use its relationships for word sentiment calculation, and apply sentiments as features for the Multinomial Naive Bayes machine learning algorithm. In this modification the method starts from creation of the domain-specific thesaurus. It processes all texts used for training and classification, find their keyphrases and also semantic relationships between them. They become thesaurus terms and relationships respectively.

Algorithms of thesauri construction for both English and Russian languages were taken from our research about classifying English newspaper articles [22] and Russian short texts [20]. For each language we use the method that already showed its efficiency. Both methods have similar structure and differ only by several inner algorithms for search of semantic relationships. They extract terms by TextRank, associations by latent semantic analysis, synonyms by the Levenshtein distance, hypernyms and hyponyms by morpho-syntactic rules.

Besides, both thesauri were filled with relationships from existing linguistic resources constructed manually: WordNet for English, RuThes and Synmaster dictionary of synonyms for Russian language.

Several methods of relationships search became suitable for one language only: measurement of term information quantity [23] and lexico-syntactic patterns for English, word2vec for Russian.

The procedure of thesaurus creation works fully automatically and does not require expert's participation, so its main advantage consists in very fast and quite qualified thesaurus creation for raw texts from specific domain.

2) Classification with the automatically generated thesaurus: For text classification with the thesaurus we calculate sentiments of thesaurus terms basing on their occurrences in texts with known sentiments or on their thesaurus neighbors following the thesaurus relationships.

The calculation of sentiments basing on thesaurus relationships is applied for each of two or three classes. Sentiment of the word or bigram from the training set equals the highest fraction of word occurrences in texts of the particular class among all its occurrences. Sentiment of the word or bigram from the test set of texts is the sum of its thesaurus neighbors' sentiments. In such a way the term gets the sentiment that is common between terms closest to it. Such algorithm allows to take into account structure of the chosen domain and spread known sentiments to new words.

On the last stage of the method with the embedded thesaurus procedure we use thesaurus term sentiments as features for classifier's vectors.

D. Evaluation

To estimate our classificator's quality we chose the most popular standard metrics: precision and recall computed for each class separately, accuracy, and macro F-measure computed for all classes.

Precision and recall are fractions of texts actually belonging to the given class among all found texts and all texts actually belonging to the class respectively. The metrics allow to see quality of classification for the particular class.

Accuracy is the fraction of texts for which the classifier made a correct decision. The macro F-measure is the average of F-measures for all classes. Both metrics allow to evaluate classificator's quality on the whole.

IV. Experiments

A. Description of text corpora

We use four text corpora for experiments. The corpora are divided into two groups by the subject: texts from a specific domain and texts that are not united by one topic.

The first group includes Russian and English reviews of hotels from the website trivago.ru. The corpus of Russian reviews contains 392 texts, 13 242 words totally. The average length of a review is 34 words. The corpus of English reviews contains 702 texts, 31 907 words totally. The average length of a review is 45 words.

The second group includes two corpora of English texts: tweets and reviews. Tweets were taken from the website http://help.sentiment140.com/for-students. The corpus contains 494 texts, 6 357 words totally. The average length of a tweet is 13 words. Reviews were taken from the site https://www.yelp.com/dataset/. The corpus contains 318 686 texts, 29 307 823 words totally. The average length of the text is 92 words.

B. Experiments stages

We conducted experiments for each corpus separately with a thesaurus and without it, for each case we varied the number of classes for training and the use of bigrams.

When training the algorithm on two classes, neutral messages are excluded from the training set. We consider a text as neutral, if difference between ratings of positive and negative classes is less than a certain number H. H is in the range from 0 to 10 with the step 1, for which the maximum classification accuracy is achieved. We experiment with all possible combinations of these modifications. The results are presented in the tables below.

C. Experiments on texts from a specific domain

Table I represents the results of experiments on a corpus of Russian reviews about hotels. When training on three classes the training set contains 100 positive, 55 negative, and 55 neutral texts. The test set contains 100 positive, 20 negative,

TABLE I. SENTIMENT CLASSIFICATION OF RUSSIAN-LANGUAGE OPINIONS ABOUT HOTELS

# of classes	Bigrams	H	Accuracy	P_{pos}	R_{pos}	P_{neg}	R_{neg}	P_{neu}	R_{neu}	F_{macro}
Three	-	-	0.690	0.694	0.930	0.700	0.350	0.675	0.403	0.588
Three	+	-	0.674	0.683	0.930	0.600	0.300	0.657	0.370	0.554
Two	-	7	0.762	0.886	0.860	0.571	0.200	0.623	0.774	0.620
Two	+	7	0.767	0.815	0.930	0.666	0.300	0.689	0.645	0.649

TABLE II. SENTIMENT CLASSIFICATION OF ENGLISH-LANGUAGE OPINIONS ABOUT HOTELS

# of classes	Bigrams	H	Accuracy	P_{pos}	R_{pos}	P_{neg}	R_{neg}	P_{neu}	R_{neu}	F_{macro}
Three	-	-	0.468	0.551	0.615	0.662	0.513	0.235	0.256	0.468
Three	+	-	0.500	0.557	0.746	0.701	0.486	0.252	0.230	0.484
Two	-	5	0.669	0.733	0.700	0.773	0.738	0.516	0.566	0.670
Two	+	5	0.646	0.626	0.838	0.767	0.684	0.543	0.389	0.631

TABLE III. SENTIMENT CLASSIFICATION OF ENGLISH-LANGUAGE REVIEWS ABOUT SOME GOODS AND SERVICES

# of classes	Bigrams	H	Accuracy	P_{pos}	R_{pos}	P_{neg}	R_{neg}	P_{neu}	R_{neu}	F_{macro}
Three	-	-	0.586	0.705	0.458	0.655	0.640	0.480	0.659	0.585
Three	+	-	0.625	0.733	0.545	0.665	0.710	0.521	0.619	0.626
Two	-	7	0.570	0.615	0.645	0.651	0.695	0.420	0.372	0.565
Two	+	7	0.565	0.611	0.639	0.652	0.683	0.413	0.374	0.561

TABLE IV. SENTIMENT CLASSIFICATION OF ENGLISH-LANGUAGE TWEETS

# of classes	Bigrams	H	Accuracy	P_{pos}	R_{pos}	P_{neg}	R_{neg}	P_{neu}	R_{neu}	F_{macro}
Three	-	-	0.638	0.601	0.681	0.678	0.703	0.645	0.462	0.623
Three	+	-	0.628	0.611	0.681	0.654	0.685	0.612	0.447	0.610
Two	-	1	0.565	0.614	0.663	0.759	0.555	0.321	0.417	0.547
Two	+	1	0.590	0.636	0.681	0.709	0.611	0.364	0.402	0.565

TABLE V. SENTIMENT CLASSIFICATION OF RUSSIAN-LANGUAGE OPINIONS ABOUT HOTELS WITH THESAURUS

Thesaurus	H	Accuracy	P_{pos}	R_{pos}	P_{neg}	R_{neg}	P_{neu}	R_{neu}	F_{macro}
without relationships	1	0.580	0.636	0.750	1.000	0.050	0.468	0.468	0.417
synonyms	7	0.768	0.816	0.930	0.667	0.300	0.690	0.645	0.650
associations	7	0.773	0.817	0.940	0.667	0.300	0.702	0.645	0.653
hypernyms	7	0.768	0.816	0.930	0.667	0.300	0.690	0.645	0.650
hyponyms	7	0.768	0.816	0.930	0.667	0.300	0.690	0.645	0.650
sy+as+hr+hp	7	0.762	0.809	0.930	0.667	0.300	0.684	0.629	0.645

TABLE VI. SENTIMENT CLASSIFICATION OF ENGLISH-LANGUAGE OPINIONS ABOUT HOTELS WITH THESAURUS

Thesaurus	H	Accuracy	P_{pos}	R_{pos}	P_{neg}	R_{neg}	P_{neu}	R_{neu}	F_{macro}
without relationships	2	0.554	0.572	0.700	0.768	0.568	0.372	0.372	0.551
synonyms	5	0.658	0.723	0.662	0.779	0.730	0.504	0.584	0.662
associations	5	0.661	0.715	0.677	0.788	0.739	0.504	0.566	0.664
hypernyms	5	0.647	0.733	0.654	0.772	0.703	0.482	0.584	0.652
hyponyms	5	0.669	0.734	0.700	0.783	0.748	0.508	0.558	0.671
sy+as+hr+hp	5	0.638	0.629	0.846	0.765	0.676	0.506	0.363	0.621

TABLE VII. SENTIMENT CLASSIFICATION OF ENGLISH-LANGUAGE TWEETS WITH THESAURUS

Thesaurus	Accuracy	P_{pos}	R_{pos}	P_{neg}	R_{neg}	P_{neu}	R_{neu}	F_{macro}
without relationships	0.576	0.607	0.582	0.635	0.645	0.489	0.502	0.576
synonyms	0.649	0.614	0.690	0.709	0.722	0.608	0.463	0.630
associations	0.659	0.624	0.690	0.699	0.731	0.640	0.478	0.639
hypernyms	0.663	0.639	0.690	0.702	0.741	0.635	0.493	0.646
hyponyms	0.653	0.619	0.690	0.716	0.722	0.604	0.478	0.635
sy+as+hr+hp	0.649	0.611	0.681	0.718	0.731	0.596	0.463	0.630

and 62 neutral texts. In experiments on a set of Russian reviews of hotels the best result is achieved by the classification algorithm trained on two classes, taking into account bigrams. This algorithm shows 7 % improvement in accuracy compared to the best classifier trained on three classes.

Table II represents the results of experiments on a corpus of English reviews about hotels. When training on three classes the training set contains 127 positive, 107 negative, and 114 neutral text. The test set contains 130 positive, 110 negative, and 113 neutral texts. In experiments on a set of English reviews of hotels the best result is achieved by the classification algorithm trained on two classes, without taking into account bigrams. This algorithm shows 16 % improvement in accuracy compared to the best classifier trained on three classes.

From these results we can see that the algorithm for text classification into three classes should differ from the similar algorithm for binary classification. This is because neutral texts have more complex structure than positive and negative ones. They mainly consist of positive and negative words and almost do not contain neutral ones. Besides, the classifier with bigrams does not always improve three-class classification accuracy. In several cases the classifier trained on two classes, is better.

D. Experiments on texts not united by one topic

Table III represents the results of experiments on a corpus of English reviews of different products and services. When training on three classes the training set contains 88 173 positive, 107 381 negative, and 93 134 neutral text. The test set contains 10 000 positive, 10 000 negative, and 10 000 neutral texts.

In the experiments on the set of English reviews, the best result is achieved by the classification algorithm trained on three classes, taking bigrams into account. This algorithm shows 5 % improvement in accuracy compared to the best classifier trained on two classes.

Table IV represents the results of experiments on a corpus of English tweets. When training on three classes the training set contains 69 positive, 69 negative, and 68 neutral texts. The test set contains 113 positive, 108 negative, and 67 neutral texts.

In experiments on a set of English tweets, the best result is achieved by the classification algorithm trained on three classes, without taking into account bigrams. This algorithm shows 4 % improvement in accuracy compared to the best classifier trained on two classes.

From these results we can see that the best accuracy of classification is observed for algorithms trained on three classes. This may be due to the fact that in such texts neutral words appear more frequently than in texts from a specific domain. The classifier with bigrams, as in the case of specific domain texts, does not always improve the accuracy of the classification. In several cases the classifier trained on three classes, is better.

E. Experiments with thesauri

We conducted experiments to study how using thesaurus relationships affect quality of text classification.

We use bigrams in all experiments with a thesaurus. In the experiments with Russian reviews (Table V) we use the algorithm trained on two classes. The algorithm with associations shows the best accuracy. In the experiments with English reviews (Table VI) we use the algorithm trained on two classes. The algorithm with hyponyms shows the best accuracy. In the experiments with English tweets (Table VII) we use the algorithm trained on three classes. The algorithm with hyperonyms shows the best accuracy.

As a result, we can see that the algorithm using no thesaurus relationships always shows worse accuracy than algorithms with relationships. The algorithm using all relationships always shows worse accuracy than the algorithms using one relationship of every type. Finally, these experiments show that involving a thesaurus does not greatly improve accuracy of classification.

V. Discussion

The research results allow us to make conclusions about the advantages and disadvantages of the proposed method, as well as about prospects of its development.

The proposed method works better for corpora from particular domains than for corpora with texts that are not united by one topic. For example, for the corpus of heterogeneous tweets classification results were the lowest. The main reason of this is connected with peculiarities of use of words and terms within a specific domain, in particular, with the problem of lexical ambiguity. In the vocabulary of a particular domain term ambiguity, especially the possibility of different sentiment polarities for one term, occurs significantly less frequently. For example, the adjective “fresh” in the field of biology has a neutral color: fresh water, but in the field of cooking — positive: fresh food.

Experiments with the thesaurus look especially illustrative. The automatically generated thesaurus reflects well relationships between terms of the particular domain and allows to effectively take into account sentiments of words within this domain as we established in our earlier research [20]. It is caused by the presence of different types of semantic relationships. In general purpose thesauri synonyms, hyponyms, and hypernyms of a term can have different sentiment polarities. But specialized thesauri do not have such a feature. That is why spreading of sentiments from terms to their thesaurus neighbors improves classification quality in specialized domains.

Another possible factor that affects classification quality, is ambiguity of the markup of texts in existing corpora created manually. Each text is considered as belonging to only one particular class. However, expert’s evaluation of even a small number of texts from it shows presence of elements with an ambiguous markup. For example, two reviews, classified as neutral by the corpus’ markup, contain both positive and negative characteristics: “Very clean and quiet place, excellent location. The absence of room temperature control, too few towels and very limited breakfast options would not allow me to call it a 4 star, but it is a solid 3-star facility” and “Good breakfast, amazing restaurant staff and nice room design. But bedding is poor, I didn’t sleep well. Mattress and pillows wasn’t comfortable”. Our method classified the first one correctly as neutral, but the latter one as positive.

Since in any text the author's opinion about a topic can change from positive to negative and vice versa, computational linguistics deals not only with sentiment classification of texts as a whole, but also with aspect-level classification of short phrases and expressions [24]. The proposed method can be easily applied to identify fine-grained sentiment information in particular parts of the text.

VI. Conclusion

In the paper we proposed the method of sentiment classification based on n-grams and Multinomial Naive Bayes classifier. The method allows training on two or three classes and classification into positive, negative, and neutral texts for both Russian and English languages. Experiments on several corpora proved efficiency of our method for reviews classification on two languages.

Thesaurus embedding into the method showed that thesaurus relationships might provide more qualified classification than application of the thesaurus terms set only, especially for texts from a specific domain.

The promising area for future research is study of methods that combine using of neural networks and thesauri. Such an approach can be also applied to national languages. Neural networks show great efficiency of solving problems of sentiment classification. The thesaurus allows to take into account characteristics of use of domain vocabulary in different national languages. These are the signs that combining them both would allow to achieve further results improvement.

References

[1] H. H. Lek and D. C. Poo, "Aspect-based twitter sentiment classification," in *Proceedings of the 25th International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, 2013, pp. 366–373.

[2] Y. Liu, X. Huang, A. An, and X. Yu, "ARSA: a sentiment-aware model for predicting sales performance using blogs," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2007, pp. 607–614.

[3] J. Gratch, G. Lucas, N. Malandrakis, E. Szablowski, E. Fessler, and J. Nichols, "GOAALLL!: Using sentiment in the world cup to explore theories of emotion," in *Proceedings of the 2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2015, pp. 898–903.

[4] A. Yousefpour, R. Ibrahim, and H. N. A. Hamed, "Ordinal-based and frequency-based integration of feature selection methods for sentiment analysis," *Expert Systems with Applications*, vol. 75, pp. 80–93, 2017.

[5] F. H. Khan, U. Qamar, and S. Bashir, "eSAP: a decision support framework for enhanced sentiment analysis and polarity classification," *Information Sciences*, vol. 367, pp. 862–873, 2016.

[6] X. Zhu, S. Kiritchenko, and S. Mohammad, "NRC-Canada-2014: Recent improvements in the sentiment analysis of tweets," in *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, 2014, pp. 443–447.

[7] N. Loukachevitch and Y. Rubtsova, "Entity-oriented sentiment analysis of tweets: results and problems," in *International Conference on Text, Speech, and Dialogue*. Springer, 2015, pp. 551–559.

[8] H. K. Aldayel and A. M. Azmi, "Arabic tweets sentiment analysis–a hybrid scheme," *Journal of Information Science*, vol. 42, no. 6, pp. 782–797, 2016.

[9] S. Rosenthal, N. Farra, and P. Nakov, "SemEval-2017 task 4: Sentiment analysis in Twitter," in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 2017, pp. 502–518.

[10] D.-T. Vo and Y. Zhang, "Target-dependent twitter sentiment classification with rich automatic features," in *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)*, 2015, pp. 1347–1353.

[11] K. Cates, P. Xiao, Z. Zhang, and C. Dailey, "Can emoticons be used to predict sentiment?" *Journal of Data Science*, vol. 16, no. 2, pp. 355–375, 2018.

[12] A. Kolovou, F. Kokkinos, A. Fergadis, P. Papalampidi, E. Iosif, N. Malandrakis, E. Palogiannidi, H. Papageorgiou, S. Narayanan, and A. Potamianos, "Tweester at SemEval-2017 Task 4: Fusion of Semantic-Affective and pairwise classification models for sentiment analysis in Twitter," in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 2017, pp. 675–682.

[13] M. Hagen, M. Potthast, M. Büchner, and B. Stein, "Webis: An ensemble for twitter sentiment detection," in *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, 2015, pp. 582–589.

[14] M. Jabreel and A. Moreno, "SiTAKA at SemEval-2017 Task 4: Sentiment Analysis in Twitter Based on a Rich Set of Features," in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 2017, pp. 694–699.

[15] H. K. Kumar and B. Harish, "Classification of short text using various preprocessing techniques: An empirical evaluation," in *Recent Findings in Intelligent Computing Techniques*. Springer, 2018, pp. 19–30.

[16] A. Tripathy, A. Agrawal, and S. K. Rath, "Classification of sentiment reviews using n-gram machine learning approach," *Expert Systems with Applications*, vol. 57, pp. 117–126, 2016.

[17] R. Kaur and P. Verma, "Sentiment analysis of movie reviews: A study of machine learning algorithms with various feature selection methods," *International Journal of Computer Sciences and Engineering*, vol. 5, no. 9, pp. 113–121, 2017.

[18] N. Loukachevitch, P. Blinov, E. Kotelnikov, Y. Rubtsova, V. Ivanov, and E. Tutubalina, "SentiRuEval: testing object-oriented sentiment analysis systems in Russian," in *Proceedings of International Conference Dialog*, vol. 2, 2015, pp. 3–13.

[19] S. Volkova, T. Wilson, and D. Yarowsky, "Exploring demographic language variations to improve multilingual sentiment analysis in social media," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013, pp. 1815–1827.

[20] K. Lagutina, V. Larionov, V. Petryakov, N. Lagutina, and I. Paramonov, "Sentiment classification of russian texts using automatically generated thesaurus," in *Proceedings of the 23rd Conference of Open Innovations Association FRUCT*. IEEE, 2018, pp. 217–222.

[21] C. D. Manning, P. Raghavan, H. Schütze *et al.*, *Introduction to information retrieval*. Cambridge university press, 2008.

[22] I. Shchitov, K. Lagutina, N. Lagutina, and I. Paramonov, "Sentiment classification of long newspaper articles based on automatically generated thesaurus with various semantic relationships," in *Proceedings of 21st Conference of Open Innovations Association FRUCT*. IEEE, 2017, pp. 290–295.

[23] E. Mozzherina, "Avtomaticheskoe postroenie ontologii po kollektsii tekstovyh dokumentov [Automatic creation of ontology from collection of text documents]," in *Digital Libraries: Advanced Methods and Technologies (RCDL-2011)*, 2011, pp. 293–298, (in Russian).

[24] K. Schouten and F. Frasincar, "Survey on aspect-level sentiment analysis," *IEEE Transactions on Knowledge & Data Engineering*, vol. 75, no. 3, pp. 813–830, 2016.