Prediction Meaning of Words with Word2Vec on Whatsapp Data for Disaster Topic

Edy Subowo[1*], *Tri Retnaningsih Soeprobowati*[1], and *Aris Puji Widodo*[1]

[1]Doctor Program of Information System, School of Postgraduate Studies, Diponegoro University, Semarang - Indonesia

Abstract. The community's need for conditions that are safe from disasters is the basis for the problems of the system to be created. The system that will be created is able to detect the meaning of words from Whatsapp data obtained from the public. The data used is Whatsapp social media data, systematics of data processing starts from pre-processing and ends with an extended factorization matrix Word2Vec analysis, which is called Continuous Bag-of-Word (CBOW) to get the meaning of sentences as early detection of disaster locations. The system can extract 43% required Whatsapp data in the total of data processed by the system, and total accuracy on Word2Vec is 79%.

1 Background

In the current era of globalization, people often travel from one place to another, so the need for safe conditions from disasters is a problem that must be resolved. BNPB (National Board of Disaster Management) data show that from 2010 to 2020 it was recorded in the Indonesian Disaster Data and Information Management Database (DIBI) that there were 24,969 incidents with a total of 5,060,778 casualties and 4,400,809 houses affected and 19,169 damaged public facilities throughout Indonesia [1]. The condition of a disaster-related area often changes dynamically. A systematic, contextual and real-time reporting system is needed to obtain a fast and precise area condition. Social media messaging such as Whatsapp can solve the reporting system in question [2].

Whatsapp provides a fast approach by informing many members of the population, besides providing user location information based on the country code of the telephone number [3]. Social media can provide location metadata such as latitude and longitude coordinates where users submit social media content [4]. Inappropriate social media content that is often sent is events that occur around users, such as air pollution, floods, and landslides. The population of social media content at one location provides accuracy [5, 6].

Social media can be used for early detection of planning, warning and response to natural disasters such as tsunamis, floods and landslides by reviewing ongoing challenges such as considering technical, social and policy issues and challenges to science and practical challenges in implementing systems [7].

* Corresponding author: edy.subowo@gmail.com

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (http://creativecommons.org/licenses/by/4.0/).

Based on previous research, there are several methods used for the extraction of social media features, such as the use of a decision tree to support a place recommendation decision based on the number of words per content on a particular label so that the C4.5 algorithm is 92% accurate [8]. Other studies regarding data accuracy using the Support Vector Machine (SVM) method to determine the level of congestion at a location using social media data resulted in an accuracy rate of 97% [9]. The results of this accuracy were obtained in manual labelling with two classes (+) and (-) so that when applied to unigram data with multilabel, the accuracy decreased to 74% in C4.5 and 83% in SVM [10]. The latest multilabel text data analysis method with skip-gram relationships between words introduced by Google is Word2Vec [11].

The Word2Vec model can process unstructured text data by taking a corpus of words as input and generating a word vector. One of the main advantages of the Word2Vec model is that it represents features as dense vectors rather than conventional tenuous representations, which are generally able to solve the synonym and homonym problems that are often encountered in NLP tasks so that this method produces an accuracy of 89% [12].

The use of word2vec in the classification model through CNN based on news articles and tweets by comparing the performance of the two word2vec learning algorithms, namely CBOW and Skip-gram, found CBOW performed better when used in news articles. The Skip-gram algorithm showed better performance when used on tweets because news articles usually show a more uniform format when compared to tweets [13]. Because the Whatsapp data, in this case, is uniform and contextual to the disaster, the CBOW method was chosen in this study.

Sentences with multiple meanings are often found on social media; besides that, SPOK is also incomplete. When stemming is done, it is not able to find the proper feature extraction. How to produce a system that can extract social media data based on syllables with multilabel classification and Contextual Bias Matrix Factorization to be applied to the Word2Vec Collection Bag of Words (CBOW) the method with the object of Disaster topic will be the main problem in this topic.

The system created can detect the meaning of words from Whatsapp data obtained from the public as a reporter of air pollution events, areas of impact of floods and landslides on each word view. The data used is Whatsapp social media data, where the GeoTag feature is available to get the precise location where users send social media content. WhatsApp data is collected every hour with keywords are air pollution, areas of impact of floods, earthquake, and landslide events. The input types are id phone number, text systematics of data processing starting from pre-processing, feature extraction with TF-IDF, and ends with an extended factorization matrix Word2Vec analysis, which is called Continuous Bag-of-Word (CBOW). The accuracy of the system and meaning of sentences as early detection of disaster locations will be the output of the study.

2 Literature review

The vector representation of a word with a relatively fast time and with a large enough dataset is the word's syntax similarity and semantic similarity. The accuracy of the results with neural networks techniques has better results[14]. The Continuous Skip Gram Model has higher quality vector representation and increases the speed of the training dataset. Additional methods can be applied to a large enough dataset fairly quickly using hierarchical softmax and negative sampling approaches. When creating the word2vec model using Skip-gram and Negative Sampling, you must also create a library called Word2Vec to implement the method[15]. Recommended methods with libraries such as content-based and collaborative filtering make Word2vec quite promising when used as a recommendation system [16]. Word2Vec can be used in all languages, such as incorrect Arabic spelling using the

Levenshtein Distance algorithm and the Bi-gram model. The result is that the wrong word can be corrected effectively[17]. Another system for correcting writing errors in Russian automatically uses the Edit Distance method to select the right word candidates for correctness. The correct word candidates are re-ranked using Logistic Regression with F1-Measure. The results of this study are also exceptionally high at 75%. [18].

2.1 Text Preprocessing

Text Preprocessing is a stage of the TF-IDF initial process of text to prepare the text into data for further processing. A text cannot be processed directly by the search algorithm. Therefore text preprocessing is needed to convert text into numeric data. The preprocessing steps starting from stemming to labelling can be seen in Figure 1.

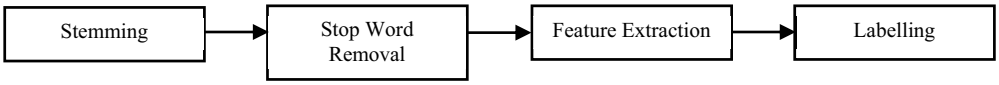


Fig. 1. Text Pre-processing.

The stemming process is carried out on Twitter data obtained by breaking down each word and removing the @, #, and numbers symbols. The stemming result data is then carried out by a stop word removal process to remove the auxiliary words, conjunctions so that feature extraction can be carried out. Furthermore, the feature extraction process is carried out to obtain the data results required for classification. Data extraction is being used to make labelling data manually.

The Term Frequency Inverse Document Frequency (TFIDF) method gives weight to the relationship of a word (term) to a document. TFIDF is a statistical measure used to evaluate how important a word is in a document or a group of words. For a single document, each sentence is considered a document. The frequency with which the word appears in a given document shows how important it is in the document. The number of times a document contains this word indicates how common it is. The weight of the word is more significant if it appears in a document frequently, and it is smaller if it appears in multiple documents. In the TF-IDF algorithm, a formula is used to calculate the weight (W) of each document against keywords with the formula, namely:

$$Wdt = tfdt * Idft \quad (1)$$

Where:
Wdt= the weight of the ked document against the t word
tfdt = the number of words that are searched for in a document
Idft = Inverse Document Frequency (log (N/df))
N = total of document
df = many documents contain the word being searched.

2.2 Word2Vec

Word2vec understands and vectorizes the meaning of words in a document based on the hypothesis that words with similar meanings in a given context exhibit close distances. Fig 2 shows the model architectures of CBOW and Skip-gram, learning algorithms of word2vec proposed by Mikolov. The learning algorithms exhibit Input, Projection, and Output layers, although their output derivation processes are different. The input layer receives

$$Wn = \{W(t-2),W(t-1),\ldots,W(t+1),W(t+2)\} \quad (2)$$

As arguments, where Wn denotes words, the projection layer corresponds to an array of multidimensional vectors and stores the sum of several vectors. The output layer corresponds to the layer that outputs the results of the vectors from the projection layer. Specifically, CBOW is similar to the feedforward Neural Network Language Model (NNLM) and predicts the output word from other near word vectors. The basic principle of CBOW involves predicting when a particular word appears via analyzing neighbouring words. The projection layer of CBOW projects all words at the same position, and thus, the vectors of all words maintain an average and share the positions of all words. The structure of CBOW exhibits the advantage of uniformly organizing the information distributed in the data set.

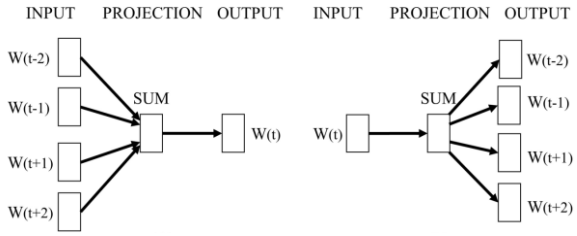


Fig. 2. Model architecture of (A) CBOW and (B) Skip-gram.

Conversely, the Skip-gram exhibits a structure for predicting vectors of other words from one word. The basic principle of Skip-gram involves predicting other words that appear around a particular word. The projection layer of the Skip-gram predicts neighbouring words around the word inserted into the input layer. The structure of the Skip-gram exhibits the advantage of vectorizing when new words appear. Based on the study by Mikolov, CBOW is faster and better suited when compared to Skip-gram when the data size is large, and Skipgram exhibits better performance when compared to CBOW while learning new words. However, other studies that compare the performance of CBOW and Skip-gram state that the performance of Skip-gram exceeds that of CBOW.

2.3 Cross Fold Validation

Evaluation of the results using the cross fold method so that the confusion matrix is obtained. A *confusion matrix* is a tool used to evaluate classification models to estimate objects that are right or wrong. The process flow begins by reading the * .csv data file then the data is divided between training data and test data to find predictive values and accuracy. In this study, there are two folds, so that the 2x2 confusion matrix was obtained to obtain different accuracy values for each distribution of training data and test data.

3 Result and Discussion

3.1 Data Collection

In the Natural Language Preprocessing schema, two data schemes have the same context but have different roles. The two data schemes are training data and testing data. The machine learning method uses training data, where the data is used to get a system scheme. In this case, the system workflow in creating knowledge to be used in the testing process with testing data to obtain the overall system accuracy.

The Whatsapp data collection method is sent to the Whatsapp group or sent directly to the author's number with the format "@ incident (space) text message". The events in

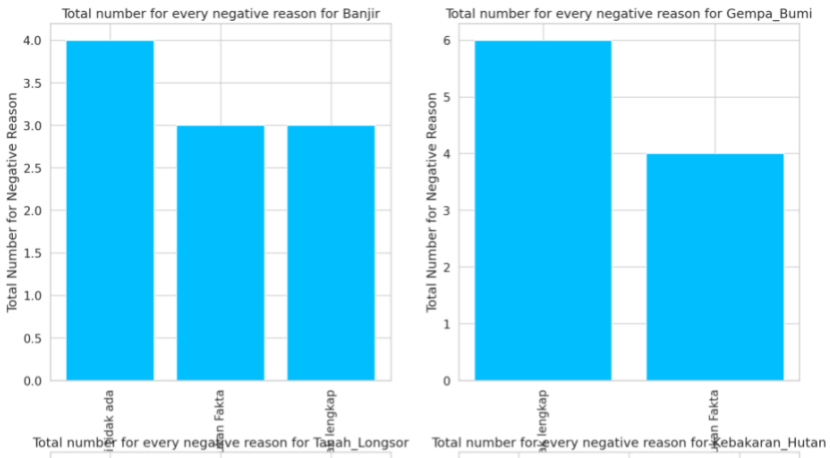
question are natural disasters consisting of floods, landslides, forest fires, and earthquakes. Below is the pseudocode to collect WhatsApp.

```
Intent sendMsg = new Intent(Intent.ACTION_VIEW);
        String url = "https://api.whatsapp.com/send?phone=" + "+6285892574457" + "&text=" +
URLEncoder.encode("@Tanah_Longsor/@Banjir/@Kebakaran/@Gempa  TextInformasi", "UTF-8");
        sendMsg.setPackage("com.whatsapp");
        sendMsg.setData(Uri.parse(url));
        if (sendMsg.resolveActivity(getPackageManager()) != null) {
           startActivity(sendMsg); }
```

Figure 3 shows an example of the captured data. Training Data were taken from Jan 5, 2021 - Apr 29, 2021, with Twitter API for the dummy because it looks like WhatsApp data 500data. The data was inserted with @label and labelled as positive and negative. The data is positive because it has elements of fact formation, namely precise location and exact time. Meanwhile, harmful data is incomplete, the location does not exist, and opinions are not facts. The figure shows the reasons for harmful data along with the amount of data based on the reasons.

	Data_id	sentimen_bencana	fraksi_sentimen_bencana	negatif	negativereason_confidence	Bencana	nama	negativereason_gold	retWA_count	text
0	5.700000e+17	positif	1.0000	NaN	1.0000	Tanah_Longsor	cairdin	NaN	0	@Tanah_Longsor Jalur Cipanas Warungbanten putu...
1	5.700000e+17	positif	0.3486	NaN	0.0000	Tanah_Longsor	jnardino	NaN	0	@Tanah_Longsor Tanah longsor yang menutup dua ...
2	5.700000e+17	negatif	0.6837	Data tidak lengkap	1.0000	Tanah_Longsor	yvonnalynn	NaN	0	@Tanah_Longsor hujan deras Balikpapan dan seki...
3	5.700000e+17	positif	1.0000	NaN	0.7033	Tanah_Longsor	jnardino	NaN	0	@Tanah_Longsor Bencana Alam di Pedalaman Sinta...
4	5.700000e+17	positif	1.0000	NaN	0.6842	Tanah_Longsor	jnardino	NaN	0	@Tanah_Longsor Polres Majene Amankan Lokasi ke...

Fig. 3. Labeling Structure on Data.



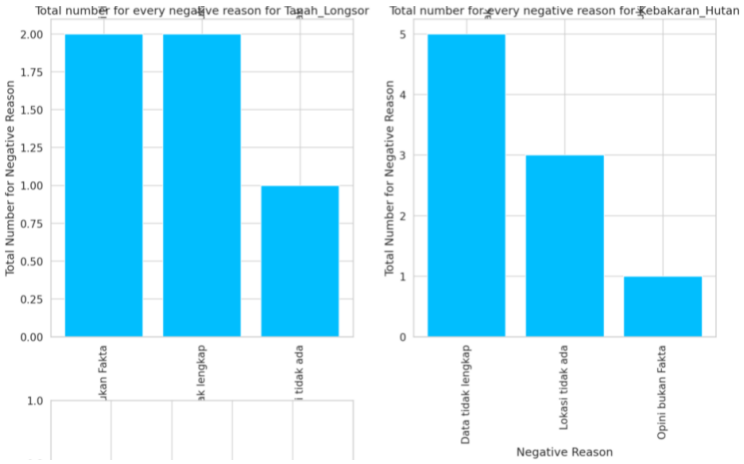


Fig. 4. The reason for the negative data

The majority of the reasons for harmful data are that the data is incomplete, such as the location of the disaster or the date of the disaster, while the subsequent negative data is because the data does not have a precise location of the incident, and at least is opinion data, meaning that data does not have elements of fact such as the form of the incident, location, and precise time. This shows the need for education regarding content that must be entered into the Whatsapp format to provide good and precise information. Whereas in the comparison of the quantity of positive and negative data, in Figure 5, it is explained that the number of harmful data has 7% more than positive data. Hence, the percentage of the system that can extract Whatsapp data is 43% of the total data processed by the system. The harmful data obtained is diffuse in nature, and few words indicate the exact location of natural disasters; this is different from positive data where the word that often appears is flood because flooding is a natural disaster that often hits Indonesia, as shown in Figure 6.

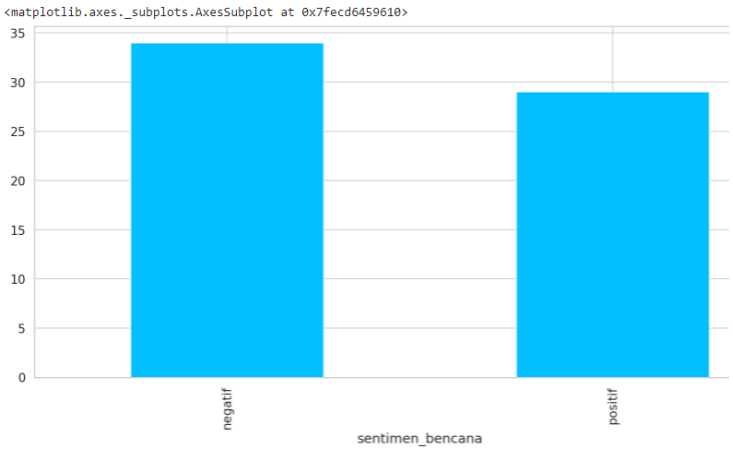


Fig. 5. Comparison of the number of positive and negative data

Fig. 6. Words that often appear in positive and negative data

3.2 Sentiment classification models

The model used is a library for sentiment analysis, TextBlob, because it is widely used in the industry. In the function definition for prediction, the threshold parameters are selected using web recommendations from the library users. Since the sentiment prediction is made in the sentence, the mean of the total sentiment score is taken for review. Figure 7 shows the confusion matrix of the TextBlob model, which shows the number of errors in the prediction of neutral sentiment on harmful sentiment data.

	precision	recall	f1-score	support
negatif	0.33	0.01	0.01	400
netral	0.34	0.93	0.49	214
positif	0.00	0.00	0.00	8
accuracy			0.32	622
macro avg	0.22	0.31	0.17	622
weighted avg	0.33	0.32	0.18	622

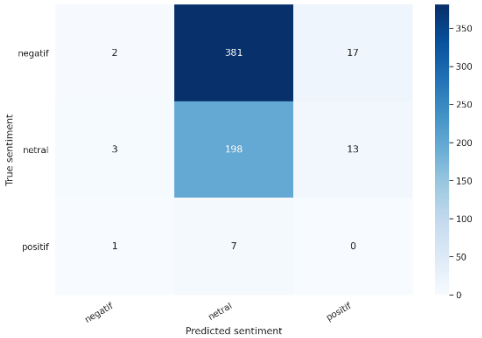


Fig. 7. *Confusion matrik* model TextBlob.

In this study, the default dictionary SentiStrength lexicon is translated and adjusted according to Indonesian rules. However, the terms in the default dictionary are not entirely by the needs of Indonesian vocabulary, so that the text blob accuracy value tends to be below.

In this section, we will have a big focus on representing our review with feature vectors, using word2vec. The purpose of this next step is to visualize all of our test documents in the document features space, highlighted by his sentiments. Since the doc2vec model only works in training data, one can test how well it generalizes. As seen further below, we can see that the three sentiment classes are very separate in the test document feature space, proving that

our doc2vec model can generalize to invisible data, as shown in Figure 8. From the data presented, the accuracy for the embedding sentiment is 89% for the negative sentiment, knowing that it is the main class in our dataset with the most occurrences. It seems very difficult to classify positive reviews with 12% accuracy due to the lack of positive review data. The total accuracy on Word2Vec is 79%.

	precision	recall	f1-score	support
negatif	0.81	0.89	0.85	400
netral	0.75	0.63	0.69	214
positif	1.00	0.12	0.22	8
accuracy			0.79	622
macro avg	0.85	0.55	0.58	622
weighted avg	0.79	0.79	0.78	622

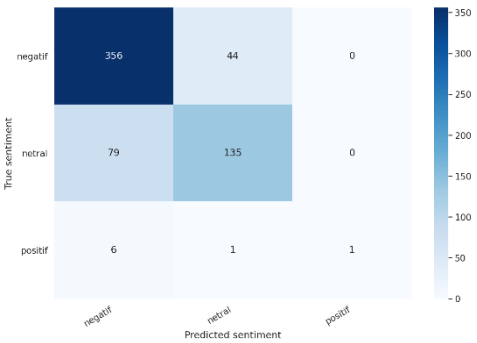


Fig.8. Hasil *confusion matric* pada Model Word2Vec

4 Conclusion

The community's need for conditions that are safe from disasters is the basis for the problems of the system to be created. The system created can detect the meaning of words from Whatsapp data obtained from the public. The data used is Whatsapp social media data; systematics of data processing starts from pre-processing and ends with an extended factorization matrix Word2Vec analysis, which is called Continuous Bag-of-Word (CBOW) to get the meaning of sentences as early detection of disaster locations. The system can extract 43% of required Whatsapp data in the total of data processed by the system, and total accuracy on Word2Vec is 79%.

5 Reference

[1] BNPB, "No Title," *DesInventar - Profile*, 2020. https://dibi.bnpb.go.id/DesInventar/profilet%0Aab.jsp?countrycode=id&continue=y.

[2] P. Lei, G. Marfia, G. Pau, and R. Tse, "Can we monitor the natural environment analyzing online social network posts? A literature review," *Online Soc. Networks Media*, vol. 5, pp. 51–60, 2018, doi: 10.1016/j.osnem.2017.12.001.

[3] K. Garimella and G. Tyson, "Whatsapp, doc? A first look at Whatsapp public group data," *12th Int. AAAI Conf. Web Soc. Media, ICWSM 2018*, no. Icwsm, pp. 511–517, 2018.

[4] K. Stock, "Mining location from social media: A systematic review," *Comput.*

Environ. Urban Syst., vol. 71, no. May, pp. 209–240, 2018, doi: 10.1016/j.compenvurbsys.2018.05.007.

[5] C. Fan, M. Esparza, J. Dargin, F. Wu, B. Oztekin, and A. Mostafavi, "Spatial biases in crowdsourced data: Social media content attention concentrates on populous areas in disasters," *Comput. Environ. Urban Syst.*, vol. 83, no. May, p. 101514, 2020, doi: 10.1016/j.compenvurbsys.2020.101514.

[6] A. Ghermandi and M. Sinclair, "Passive crowdsourcing of social media in environmental research: A systematic map," *Glob. Environ. Chang.*, vol. 55, no. January, pp. 36–47, 2019, doi: 10.1016/j.gloenvcha.2019.02.003.

[7] P. M. Landwehr, W. Wei, M. Kowalchuck, and K. M. Carley, "Using tweets to support disaster planning, warning and response," *Saf. Sci.*, vol. 90, pp. 33–47, 2016, doi: 10.1016/j.ssci.2016.04.012.

[8] E. Subowo, I. Rosyadi, and H. H. Kusumawardhani, "Twitter Data as Decision Tree Parameter for Analysis of Tourism Potential Policies," vol. 436, pp. 474–478, 2020, doi: 10.2991/assehr.k.200529.099.

[9] E. Subowo, E. Sediyono, and Farikhin, "Ant Colony Algorithm for Determining Dynamic Travel Routes Based on Traffic Information from Twitter," *E3S Web Conf.*, vol. 125, no. 201 9, 2019, doi: 10.1051/e3sconf/201912523012.

[10] A. Krouska, C. Troussas, and M. Virvou, "The effect of preprocessing techniques on Twitter sentiment analysis," *IISA 2016 - 7th Int. Conf. Information, Intell. Syst. Appl.*, no. July, 2016, doi: 10.1109/IISA.2016.7785373.

[11] K. W. Church, "Emerging Trends: Word2Vec," *Nat. Lang. Eng.*, vol. 23, no. 1, pp. 155–162, 2017, doi: 10.1017/S1351324916000334.

[12] R. P. Nawangsari, R. Kusumaningrum, and A. Wibowo, "Word2vec for Indonesian sentiment analysis towards hotel reviews: An evaluation study," *Procedia Comput. Sci.*, vol. 157, pp. 360–366, 2019, doi: 10.1016/j.procs.2019.08.178.

[13] B. Jang, I. Kim, and J. W. Kim, "Word2vec convolutional neural networks for classification of news articles and tweets," *PLoS One*, vol. 14, no. 8, pp. 1–20, 2019, doi: 10.1371/journal.pone.0220976.

[14] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *1st Int. Conf. Learn. Represent. ICLR 2013 - Work. Track Proc.*, pp. 1–12, 2013.

[15] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations ofwords and phrases and their compositionality," *Adv. Neural Inf. Process. Syst.*, pp. 1–9, 2013.

[16] M. G. Ozsoy, "From Word Embeddings to Item Recommendation," 2016, [Online]. Available: http://arxiv.org/abs/1601.01356.

[17] A. S. Lhoussain, G. Hicham, and Y. Abdellah, "Adaptating the levenshtein distance to contextual spelling correction," *Int. J. Comput. Sci. Appl.*, vol. 12, no. 1, pp. 127–133, 2015.

[18] A. A. Sorokin and T. O. Shavrina, "Automatic spelling correction for Russian social media texts," *Komp'juternaja Lingvistika i Intellektual'nye Tehnol.*, pp. 688–701, 2016.