DOI: 10.31643/2023/6445.36
Engineering and Technology

Classification of texts on emergency situations in Almaty

[1]Andirov M.Y., [1*]Assan Zh.Zh., [2]Nopembri S., [3]Seilkhan A.M., [1]Myrzakhmetov D.E.

[1]Al-Farabi Kazakh National University, Almaty, Kazakhstan
[2]Universitas Negeri Yogyakarta, Yogyakarta, Indonesia
[3]Aktobe RSU named after K.K. Zhubanov, Aktobe, Kazakhstan

** Corresponding author email: zh.assanova98@gmail.com*

Received: *December 24, 2022*
Peer-reviewed: *January 18, 2023*
Accepted: *January 31, 2023*

ABSTRACT

Text classification is a process that includes stages and approaches for the effective classification of texts that are diverse in their structure. In this article, machine learning algorithms are implemented, such as the support vector method, logistic regression, and the k nearest neighborhood method for classifying texts collected from emergency news sites in Almaty. During the experiment, a special role was played by the data collection stage, as well as their subsequent processing. Prior to the classification of the data set, preliminary data processing was performed, which includes such steps as the removal of stop words, tokenization, stemming, lemmatization, feature extraction, and the construction of feature vectors. The data was obtained by automated collection of information from open sources using a script. Experimental results show that the classifier based on logistic regression provides the best performance results compared to other types of algorithms. The performance indicators of each algorithm were obtained, which allows us to perform a comparative analysis between them.

Keywords: machine learning, text classification, support vector machine, logistic regression, KNN, NLP, preprocessing, emergencies.

Andirov Mussa Yerezhepbayuly	*2nd year Master's student, Computer Science, Faculty of Information Technology, Al-Farabi Kazakh National University, Almaty, Kazakhstan. Email: andirov2610@gmail.com*
Assan Zhanelya Zheniskyzy	*2nd year Master's student, Computer Science, Faculty of Information Technology, Al-Farabi Kazakh National University, Almaty, Kazakhstan. Email: zh.assanova98@gmail.com*
Nopembri Soni	*Professor, Universitas Negeri Yogyakarta, Yogyakarta, Indonesia. Email: soni_nopembri@uny.ac.id*
Seilkhan Abilmansur Meiramgaliuly	*2nd year Master's student, Computer science and information technology, Faculty of Physics and Mathematics, K. Zhubanov Aktobe Regional University, Aktobe, Kazakhstan. Email: seilkhan.mansur@gmail.com*
Myrzakhmetov Dias Erlanuly	*2nd year Master's student, Computer Science, Faculty of Information Technology, Al-Farabi Kazakh National University, Almaty, Kazakhstan. Email: diko.17.04@gmail.com*

Introduction

On the Internet, the amount of data in the form of text is increasing every day, that leads to the necessity for studying and processing of this information. Basically, the texts are unstructured, but there is data that can be classified as partially structured, they include email messages, blogs and chats on social networks, articles on news sites, electronic libraries, etc. This alignment requires timely collection, monitoring and correct classification of the incoming information flow [1]. While classifying texts, the correct and fast work is required. It achieved by training on a pre-classified data-set. After training the classifier, the selected algorithm will properly categorize the provided data [2].

Science covers a large number of articles and researches based on theoretical evidences. Quite a few studies are devoted to the practical evidences of text classification. In the article [3], classification algorithms were applied on the database collected from social articles, consisting of unstructured and raw texts.

The authors used ten algorithms, five of which are based on machine learning methods, the rest are vocabulary-based. The authors came to the conclusion that almost all classifiers work correctly when classifying English texts, but when classifying foreign texts, more time and effort will be required.

The article [4] implements a classifier of a foreign language text. On a database consisting of texts in the Persian language, classification methods based on machine learning were applied. In this article, support vector, logistic regression, and k nearest neighborhood methods were used and compared. The k nearest neighborhood (KNN) method is a metric classification method based on machine learning. The selected sample is classified by calculating the distance of this object from other samples. If the distance is the minimum threshold value, the object is assigned this class. A large number of positive characteristics stand out, such as a simple theoretical basis, the ability to select metrics to increase efficiency.

Support Vector Machine is a supervised machine learning method. In this method, with the help of calculations, the optimal hyperplane is found, which divides the data set into several classes. The definition and application of this method is described in the article [5].

The logistic regression method is a supervised machine learning method. This algorithm finds the optimal hyperplane. This hyperplane will divide the test set into classes [6]. If the value of the target or target variable is of a categorical character, it is advisable to use the logistic regression algorithm. In the article [7], the authors consider the method of logistic regression largely theoretical, exactly the basic formulas for calculating the position of the hyperplane, advantages and disadvantages.

The research work [8] provides a model for classifying textual data in English. This work indicates the stages of data collection, text preprocessing, vector representation of text, as well as classification algorithms based on machine learning. Through a comparative analysis by metrics, we choose the most effective machine learning algorithm. In the article [9], the classification was carried out between unlabeled data and data with a positive class.

As a result, the authors obtained a classifier that separates the new data set into a positive and unlabeled class. The main feature of this article is the moment of selecting parameters for each algorithm separately, which gives excellent results. The work [10] represents a number of articles on the topic of text classification. The best and most significant results in this topic are indicated. When categorizing text by topic, it is necessary to highlight a number of words that are suitable for describing each class. The article [11] demonstrated the methods by which this goal is achieved. More and more information is being accumulated in social networks, and email plays an important role among them. In the article [12], such methods as SVM, classifier based on neural networks, J48, classifier based on naive Bayes are used to categorize texts received by e-mail. The data set consisted of unnormalized spam and non-spam messages. A feature of this article is that a simple classifier based on J48 showed the best results, although it is based on building a binary tree.

In the article [13], the authors noted that the main factor of correct classification is the presentation of the text. During the classification process, four methods of text representation were compared, such as phrases, RDR, key words, and N-grams. To increase the efficiency indicators, it is worth sorting out these methods and choosing the best one. Many of us notice that information on the Internet and social networks is becoming more and more personalized. The research paper [14] deals with news collected from newspapers. Thus, each person receives the information in which he is interested.

2. Materials and methods

2.1 Problem statement and data

Here considered the emergency situations of technogenic and natural character of Almaty city. Data on emergency situations were collected from the news site https://tengrinews.kz, then a data table was compiled for them.

The used dataset consists of a classification of 4 types of news about emergency situations. They are:

- Road traffic accidents are events involving vehicles, as a result of which people or environmental objects died and suffered, namely: cargo, structures, property, etc.
- Flooding is an event that carry the nature of submersion the environment as a result of a rise in the water level in a local river, sea, lake due to such causes as rain, congestion, snow.
- A fire is an event in which an uncontrolled fire occurs, which entails material damage to a person and his property.
- An earthquake is an event in which tremors and vibrations occur during various interventions in the earth's crust.

The dataset has 1712 lines and 3 columns. Each line represents specific events, and each column has different indications of those events. Each line of the data set contains the following fields:

- Data-date of news publication;
- Content-content and description of news;
- Category-category of the event.

2.2. Text classification

In the course of work the categorization of the text, several stages were performed. Figure 1 shows the general structure for categorizing emergency news articles. Depending on the goal and preferences of the performer, as well as the expected result, the steps may change, but the overall structure remains the same. As texts, documents can be selected the data from open sources or collected manually. The preprocessing step can be omitted if necessary or have a different structure.

In this study, special attention is paid to this particular stage, since the data has irrelevant elements and noise in the form of html language characters. The stages of indexing and feature selection cannot be skipped when classifying text using machine learning algorithms. The fact is that machine learning algorithms cannot accept natural language data, which leads to the need to bring the data into some numerical form to train the classifier. These stages are interconnected and perform the transformation of the text into a numerical form. Text classification algorithms can be probabilistic for example Naive Bayes, metric: k nearest neighborhood algorithm, logical: decision tree, linear: support vector machine, logistic regression, neural network-based methods: RNN, CNN.

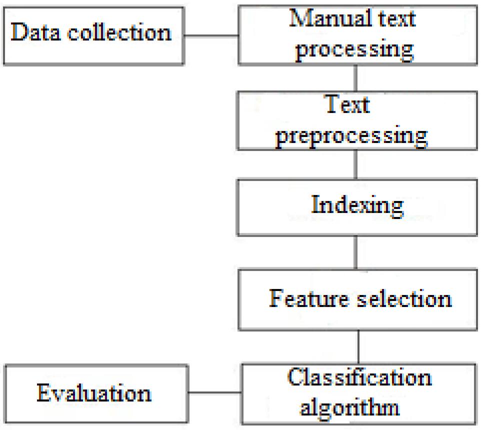


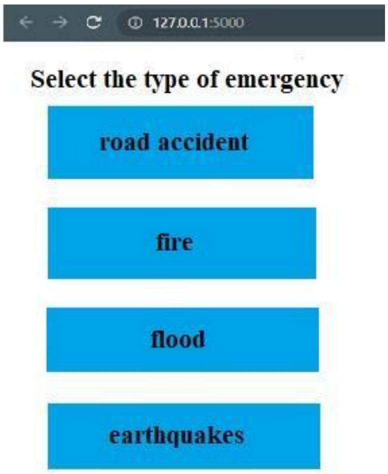
Figure 1 - Stages of text classification

2.2.1 Data collection

The website tengrinews.kz was chosen as the data source. The web scraping program is written in Python. And there are used libraries such as BeautifulSoup, requests, fake-useragent, as well as regular expressions. tengrinews.kz is a news website of Kazakhstan that publishes information about events on various topics [15]. For the classification of texts on emergency situations, the topics of traffic accidents, fire, earthquake and flood were considered. In the process, a request of the type - get was executed.

For a visual demonstration and ease of perception, a site in the Python language was designed. From the numerous sets of frameworks provided by this language, was chosen the Flask framework.

Figure 2 demonstrates a significant site for data collection. To collect identification to identify identified cases of dangerous situations. As a result, the database contains an xls file.



Select the type of emergency

Figure 2 - Site for data collection

Table 1 provides general information about the collected database

Table 1 - Database information

Volume	6,07 MB
Number of columns	3
Number of lines (news)	1712

2.2.2 Manual processing of the text

When classifying text without manual processing, the performance indicators of the selected machine learning algorithms were low, and during the derivation of a set of features related to each category, these features did not accurately describe the category data, which prompted us to take the process of collecting data manually. Manual processing was carried out in an environment for working with spreadsheets excel. We sorted the news and filtered out irrelevant news into categories. It was necessary to collect data on other settings. As a result, the data was ready for software processing.

2.2.3 Preprocessing the dataset

Text pre-processing is a technique implemented during the initial stages of text classification systems. This stage is obligatory and may have a different structure depending on the tasks set.

Text pre-processing is the actions that must be performed when working with text in order to bring the text into a suitable form for further work. This process may change depending on the task and preferences.

- Convert text to lowercase;

- Partial or complete removal of numbers;

- Removing punctuation and punctuation marks in the text.

- Tokenization;

- Remove stop words. Such as connecting words, prepositions, conjunctions, interjections;

- Stemming;

- Lemmatization;

-Vector representation of words using CountVectorizer and TF-IDF.

When converting letters to lowercase, the ready-made language function lower () was used. Regular expressions were used to remove punctuation and punctuation marks, as well as to remove unwanted symbols and numbers. All of these manipulations on texts were placed in a function that we used repeatedly throughout the experiment. Tokenization was performed by a ready-made function, which makes it possible to quickly complete this stage. For a large amount of data, it is also necessary that the words have the initial form, which leads to the fact that the dimension decreases and the speed of the program execution increases.

Morphological analyzer pymorphy2 written in python converts words to normal form and returns the grammatical basis of words. This library is quite fast and uses the OpenCorpora dictionary. To remove stop words, we used a dictionary from the nltk library designed to perform all natural language treatment processes.

2.2.4 Indexing

Indexing is the process of converting text into numbers. Achieved using different models. For this work, the “bag of words” models was tested. As a result, the model calculated the weight of each word in the overall text. H ereinafter, indexing is necessary for the selection of features.

2.2. Feature Selection

One of the main stages in the text classification system is the stage of feature selection. In the process of performing this stage, factors such as the number of keywords, the correspondence of features to each category were taken into account. With a large number of unigrams, the system is considered ineffective. Table 2 presents the results of the work on the selection of features.

Table 2 - Unigrams and Bigrams

	fire	flood	Earthquake	road accident
Uni-gram	burn, ignition, fire, firefighter, flame	thousand, level, river, water, flood	underground, epicenter, push, magnitude, earthquake	Car crash, police, car, driver, road accident
Bigram	tremor, earthquake, magnitude			

2.3 Algorithm of classification

When performing this stage, the data is divided into test and training samples, the training data is used during the training of the algorithm, the test data is used to evaluate the efficiency of the classifier. In this work, machine learning algorithms such as k nearest neighborhood algorithm, logistic regression, and support vector machines were used to classify news about emergency situations.

2.3.1 K nearest neighbor (KNN) method

The knn algorithm consists of two stages: training and classification. During training, the algorithm remembers the vectors of each observation feature, in our case, the text, as well as the class labels of each object. It is necessary to set the parameter k, which is responsible for the number of neighborhoods required for object classification. During the stage of classifying an object for which a class label is not specified, neighborhood is determined and classification takes place based on these calculations [16].

2.3.2 Support Vector Machine (SVM*)*

Support Vector Machine is a supervised learning method that is known to be successful in a wide variety of applications. The high generalizing ability of the method makes it suitable especially for large data such as text. The principle of operation of this algorithm is to find the most correct hyperplane (line) that will divide the data into two or more classes. The algorithm receives at the input a certain set of classified data for training, after which, when submitting unclassified texts, it outputs a class based on the separating plane [[17], [18]].

The advantages of the algorithm:

- trainability of the algorithm even with a small data set;
- the quality of the algorithm execution.

Disadvantages:

- problems in the presence of an outlier in the data;
- Difficulty in selecting parameters.

2.3.3 Logistic Regression Method (LR)

During the research work, it was decided to use multinomial logistic regression, due to the fact that the number of classes is more than two. Multinomial logistic regression is a variation of regular logistic regression, but in which the number of categories is greater than two. In our case, the number of categories in the dependent variable is four. In this algorithm, for each class in the number of the dependent variable, it is necessary to construct an equation as in binary (binary) logistic regression. One of the categories becomes the main pillar, the other categories of the dependent variable are compared with it [[19], [20]].

Advantages:

- has incremental learning.

The disadvantages of this method are similar to those of the SVM algorithm. Based on this information, these methods were selected for research and classification of texts on emergency situations.

2.5 Research data analysis

Before proceeding with the preliminary processing of a data set, it is advisable to conduct a research analysis of the data obtained. In text classification, the main problem faced by a large number of studies is the imbalance of data. This means that the data has the same amount across classes. Let's say if there are two classes and 95 ways tof processing unbalanced data. The first method assumes undersampling the majority class and resampling the minority class.

The second method implies the use of other metrics to evaluate the error, such as f1-score, recall, precision. When we analyze the data, we have indicators in the form of percentage of observations corresponding to each class. Figure 3 shows that the classes are balanced, so we did not use methods to compensate or remove the sample.

The next variable is the length of news by category, the diagram shows the distribution of length by category. From Figure 4, we can see that all four categories of emergency situations are almost the same length from 1000 to 2000 symbols. But news about a fire has a little more symbols, and news about an earthquake, on the contrary, has less symbols compared to other categories.

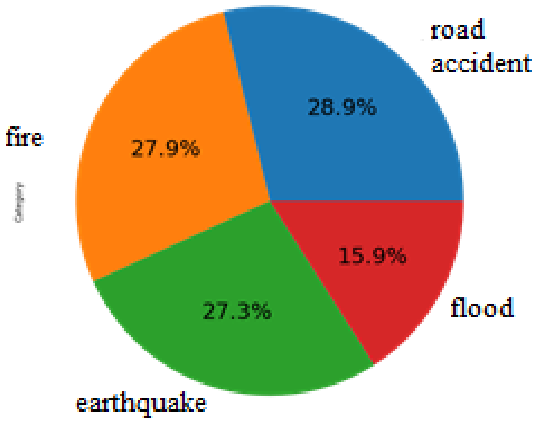


Figure 3 - Percentage of the number of news by category

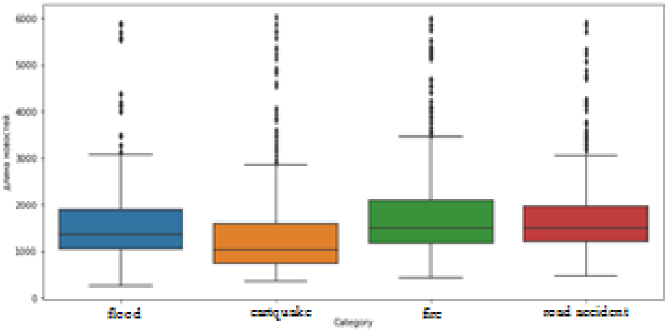


Figure 4 - Length of news by category

Results

For correct classification, it is necessary to take into account the aspects of dividing data into train and test. As practice shows, it is best to divide the data in proportions of 20% to 80% or 30% to 70%.

Data consisting of texts about emergency situations are divided in the proportions of 20-test and 80-train. Further, on this data, machine learning algorithms were used and the results of the effectiveness of each of them were provided. The k nearest neighborhood algorithm is calculated using the Euclidean distance. The number of neighborhoods affects the efficiency, so during the experiment, their values were chosen according to high rates.

As shown in Figure 5, the classifier determined the class of the 97 texts correctly. The values 1, 1, 3 in the first line indicate that the classifier made a mistake 5 times in the course of work.

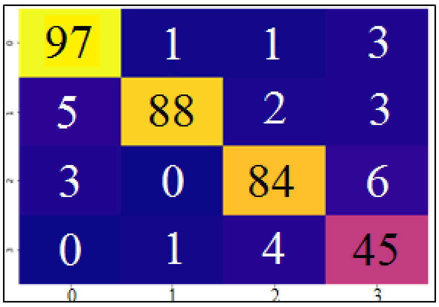


Figure 5 - KNN- confusion matrix

The metrics reflecting the efficiency of the classifier are shown in table 3. The highest indicators were obtained with the number of neighborhood equal to 8.

Table 3 - knn algorithm indicators

KNN	Accuracy	Recall	Precision	F1
6	0.92128	0.92062	0.91270	0.91517
3	0.91253	0.91021	0.90863	0.90879
8	0.93294	0.92842	0.92488	0.92597

The support vector machine, as well as the k nearest neighborhood algorithm, has reached high values. In this method, the parameter, called the core, was chosen as linear. In addition to the core, the value of the "C" parameter was adjusted, which is equal to 0.1. The effectiveness of this method was evaluated using the same quality metrics, the indicators of which are shown in Table 4.

Table 4 - Indicators of the SVM algorithm

	Accuracy	Recall	Precision	F1
SVM	0.96209	0.96179	0.95253	0.95658

In the Figure 6, you can see that the support vector classifier correctly classified 98 out of 102 texts in the first category.

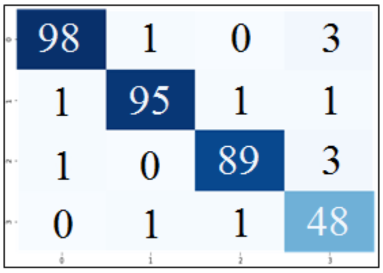


Figure 6 - SVM - confusion matrix

The results of the logistic regression method showed the highest values, which can be seen in Table 5 and the figure 7.

Table 5 - Indicators of the LR algorithm

	Accuracy	Recall	Precision	F1
Logistic Regression	0.97667	0.97452	0.97452	0.9734

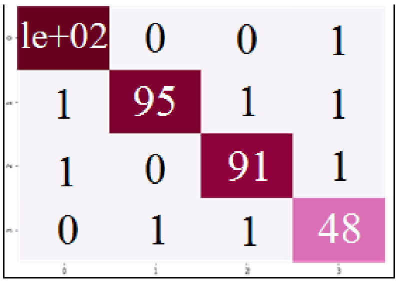


Figure 7 - LogisticRegression - confusion matrix

Conclusion

The aim of this article is to collect information about emergency situations and the subsequent classification of this data. Several methods of machine learning were chosen as algorithms.

- Data collection was carried out from the site tengrinews.kz.
- A text pre-processing process was performed, which includes the steps of clearing and converting the text to a number format.
- Before program processing, manual text processing was performed in the excel environment.
- We checked the set of features describing each category using the "bag of words" method.
- Comparing the methods of support vector machines, logistic regression and k nearest neighborhood , we came to the conclusion that logistic regression is in many ways superior to other machine learning algorithms. For comparison of methods, were used the metrics such as precision, f1, score, accuracy, recall.

- We compared methods such as logistic regression, k nearest neighborhood , support vector machine. According to the result of the study, all three methods: support vector machine method, nearest neighbor method and logistic regression gave good results, but logistic regression is superior to other machine learning classification algorithms for the collected dataset. Method comparisons were achieved using metrics such as accuracy, precision, recall, f1 measure.

Conflict of interests

On behalf of all authors, the correspondent author declares that there is no conflict of interest.

Cite this article as: Kabdrakhova SS, Assan ZhZh, Nopembri S, Seilhan AM, Myrzakhmetov DE. Classification of texts on emergency situations in Almaty. Kompleksnoe Ispolzovanie Mineralnogo Syra = Complex Use of Mineral Resources. 2023; 327(4):23-31. https://doi.org/10.31643/2023/6445.36

Алматы қаласындағы төтенше жағдайлары бойынша мәтіндерді жіктеу

[1]Андиров М.Е., [1*]Асан Ж.Ж., [2]Nopembri S., [3]Сейлхан Ә.М., [1]Мырзахметов Д.Е.

[1]Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан
[2] Йогьякарта Мемлекеттік Университеті, Йогьякарта, Индонезия
[3]Қ.Қ.Жұбанов атындағы Ақтөбе ЕҰУ, Ақтөбе, Қазақстан

Мақала келді: *24 желтоқсан 2022*
Сараптамадан өтті: *18 қаңтар 2023*
Қабылданды: *31 қаңтар 2023*

ТҮЙІНДЕМЕ

Мәтіндерді жіктеу-бұл құрылымы бойынша әртүрлі мәтіндерді тиімді жіктеудің кезеңдері мен тәсілдерін қамтитын процесс. Бұл мақалада Алматы қаласының төтенше жағдайлар жөніндегі жаңалықтар сайттарынан жиналған мәтіндерді жіктеу үшін тірек векторлар әдісі, логистикалық регрессия, жақын көршілердің k әдісі сияқты машиналық оқыту алгоритмдері жүзеге асырылады. Тәжірибе барысында деректерді жинау кезеңі, сондай-ақ кейіннен оларды өңдеу ерекше рөл атқарды. Деректердің жиынтығын жіктеуден бұрын деректерге алдын-ала өңдеу жүргізілді, оған стоп сөздерін алып тастау, токенизация, стемминг, лемматизация, белгілерді алу, белгілердің векторларын құру сияқты қадамдар кіреді. Деректер арнайы скрипт көмегімен автоматты түрде жаңалықтар сайтынан алынды. Эксперименттік нәтижелер логистикалық регрессияға негізделген жіктеуіш алгоритмдердің басқа түрлерімен салыстырғанда ең жақсы өнімділік нәтижелерін беретінін көрсетеді. Әр алгоритмнің тиімділік көрсеткіштері алынды, бұл олардың арасында салыстырмалы талдау жасауға мүмкіндік береді.

Түйін сөздер: машиналық оқыту, мәтіндерді жіктеу, тірек векторлар әдісі, логистикалық регрессия, KNN, NLP, алдын-ала өңдеу, төтенше жағдайлар.

	Авторлар туралы ақпарат:
Андиров Муса Ережепбайұлы	*Магистрант 2 курс, компьютерлік ғылымдар, ақпараттық технологиялар факультеті, әл-Фараби атындағы ҚазҰУ, Алматы қ., Қазақстан. Email: andirov2610@gmail.com*
Асан Жанеля Жеңісқызы	*Магистрант 2 курс, компьютерлік ғылымдар, ақпараттық технологиялар факультеті, әл-Фараби атындағы ҚазҰУ, Алматы қ., Қазақстан. Email: zh.assanova98@gmail.com*
Nopembri Soni	*Профессор, Йогьякарта Мемлекеттік Университеті, Йогьякарта, Индонезия. Email: soni_nopembri@uny.ac.id*
Сейлхан Әбілмансұр Мейрамғалиұлы	*Магистрант 2 курс, информатика және ақпараттық технологиялар, физика-математика факультеті, Қ. Жұбанов атындағы АӨУ , Ақтөбе, Қазақстан. Email: seilkhan.mansur@gmail.com*
Мырзахметов Диас Ерланұлы	*Магистрант 2 курс, компьютерлік ғылымдар, ақпараттық технологиялар факультеті, әл-Фараби атындағы ҚазҰУ, Алматы қ., Қазақстан. Email: diko.17.04@gmail.com*

Классификация текстов по чрезвычайным ситуациям г. Алматы

[1]Андиров М.Е., [1*]Асан Ж.Ж., [2]Nopembri S., [3]Сейлхан А.М., [1]Мырзахметов Д.Е.

[1]Казахский национальный университет им. аль-Фараби, Алматы, Казахстан
[2]Джокьякартский государственный университет, Джокьякарта, Индонезия
[3]Актюбинский государственный университет имени К.К. Жубанова, Актобе, Казахстан

Поступила: *24 декабрь 2022*
Рецензирование: *18 января 2023*
Принята в печать: *31 января 2023*

АННОТАЦИЯ

Классификация текстов — это процесс, включающий в себя этапы и подходы для эффективной классификации разновидных по своей структуре текстов. В данной статье реализуются алгоритмы машинного обучения, такие как метод опорных векторов, логистическая регрессия, метод k ближайших соседей для классификации текстов собранных с новостных сайтов по чрезвычайным ситуациям г. Алматы. В ходе эксперимента особую роль играл этап сбора данных, а также их последующая обработка. Перед классификацией набора данных производилась предварительная обработка данных, которая включает в себя такие этапы как удаление стоп-слов, токенизация, стемминг, лемматизация, извлечение признаков, построение векторов признаков. Данные были получены с помощью автоматизированного сбора информации из открытых источников с помощью скрипта. Экспериментальные результаты показывают, что классификатор на основе логистической регрессии обеспечивает наилучшие результаты производительности по сравнению с другими видами алгоритмов. Были получены показатели эффективности каждого алгоритма, что дает нам выполнить сравнительный анализ между ними.

Ключевые слова: машинное обучение, классификация текстов, метод опорных векторов, логистическая регрессия, KNN, NLP, предобработка, чрезвычайные ситуации.

	Информация об авторах:
Андиров Муса Ережепбайулы	*Магистрант 2 курс, компьютерные науки, факультет информационных технологий, КазНУ имени аль-Фараби, г. Алматы, Казахстан. Email: andirov2610@gmail.com*
Асан Жанеля Женискызы	*Магистрант 2 курс, компьютерные науки, факультет информационных технологий, КазНУ имени аль-Фараби, г. Алматы, Казахстан. Email: zh.assanova98@gmail.com*
Nopembri Soni	*Профессор Джокьякартского государственного университета, Джокьякарта, Индонезия. Email: soni_nopembri@uny.ac.id*
Сейлхан Абильмансур Мейрамгалиулы	*Магистрант 2 курс, информатика и информационные технологии, физико-математический факультет, Актюбинский РУ им. К.Жубанова, г. Актобе, Казахстан. Email: seilkhan.mansur@gmail.com*
Мырзахметов Диас Ерланулы	*Магистрант 2 курс, компьютерные науки, факультет информационных технологий, КазНУ имени аль-Фараби, г. Алматы, Казахстан. Email: diko.17.04@gmail.com*

References

[1] A Review of Machine Learning Algorithms for Text-Documents Classification. Aurangzeb Khan and Baharum Baharudin and Lam Hong Lee and Khairullah khan. JOURNAL OF ADVANCES IN INFORMATION TECHNOLOGY. 2010; 1:4-20.

[2] KNN based Machine Learning Approach for Text and Document Mining. Vishwanath Bijalwan and Vinay Kumar and Pinki Kumari and Jordan Pascual. International Journal of Database Theory and Application. 2014; 7:61-70.

[3] Krasnyansky MN, Obukhov AD, Solomatina EM, Voyakina AA. Sravnitel'nyj analiz metodov mashinnogo obucheniya dlya resheniya zadachi klassifikacii dokumentov nauchno-obrazovatel'nogo uchrezhdeniya [Comparative analysis of machine learning methods for solving the problem of classifying documents of the scientific and educational institution ques]. Vestnik VGU. 2018; 3:173-182. (in Russ.).

[4] Applying machine learning algorithms for automatic Persian text classification. Mojgan Farhoodi and Alireza Yari. International Conference on Advanced Information Management and Service. 2010; 6:318-323.

[5] Text Classification with Machine Learning Algorithms. Nasim VasfiSisi and Mohammad Reza and Feizi Derakhshi. Journal of Basic and Applied Scientific Research. 2013; 1:31-35.

[6] A Novel Active Learning Method Using SVM for Text Classification. Mohamed Goudjil and Mouloud Koudil and M. Bedda. International Journal of Automation and Computing. 2018; 15:290-298.

[7] Performance Analysis of Supervised Machine Learning Algorithms for Text Classification. Sadia Zaman Mishu and Rafiuddin SM. International Conference on Computer and Information Technology. 2016; 19:409-413.

[8] Study on SVM Compared with the other Text Classification Methods. Xiaoyu Luo. Alexandria Engineering Journal. 2021; 60:3401-3409.

[9] Text Classification Using Machine Learning Techniques. Ikonomakis, Emmanouil and Kotsiantis and Sotiris and Tampakas, V. WSEAS transactions on computers. 2005; 4:966-974.

[10] A survey of text classification algorithms. Aggarwal Charu C. and ChengXiang Zhai. Mining text data. Springer. 2012; 4:163-222.

[11] Text classification by labeling words. Liu and Bing. AAAI. 2004, 4.

[12] A Comparative Study for Email Classification, Advances and Innovations in Systems. Seongwook Youn and Dennis McLeod. Computing Sciences and Software Engineering. 2007, 387-391.

[13] Keikha, Mostafa and Razavian, Narjes and Oroumchian, Farhad and Razi, Hassan, Document Representation and Quality of Text: An Analysis, Survey of Text Mining II: Clustering, Classification, and Retrieval. 2008, 219-232.

[14] Ontolo gy-Based Classification Of News In An Electronic Newspaper. Lena Tenenboim and Bracha Shapira and Peretz Shoval. International Conference Intelligent Information and Engineering Systems. 2008.

[15] The news site is tengrinews.kz [Electronic resource]. Access mode: https://tengrinews.kz/kazakhstan_news/devushka-sportkare-sovershila-dtp-vyiezjaya-kluba-almatyi-466091/

[16] Li Qian, Hao Peng, Jianxin Li, Congying Xia, Renyu Yang, Lichao Sun, Philip S. Yu, and Lifang He. A Survey on Text Classification: From Traditional to Deep Learning. ACM Transactions on Intelligent Systems and Technology (TIST) 13. 2022; 2:1-41.

[17] Study on SVM Compared with the other Text Classification Methods. Zhijie Liu and Xueqiang Lv and Kun Liu and Shuicai Shi. 2010 Second International Workshop on Education Technology and Computer Science. 2010; 1:219-222.

[18] An Optimal SVM-Based Text Classification Algorithm. Zi-qiang Wang and Xia Sun and De-xian Zhang and Xin Li. International Conference on Machine Learning and Cybernetics. 2006; 60:1378-1381.

[19] Li Qian, Hao Peng, Jianxin Li, Congying Xia, Renyu Yang, Lichao Sun, Philip S. Yu, and Lifang He. A Survey on Text Classification: From Traditional to Deep Learning. ACM Transactions on Intelligent Systems and Technology (TIST) 13. 2022; 2:1-41.

[20] Sabri T, El Beggar O, and Kissi M. Comparative study of Arabic text classification using feature vectorization methods. Procedia Computer Science. 2022; 198:269-275.

[21] Wadud MAH, Kabir MM, Mridha MF, Ali MA, Hamid MA, and Monowar MM. How can we manage offensive text in social media-a text classification approach using LSTM-BOOST. International Journal of Information Management Data Insights. 2022; 2:100095.