ИНФОРМАТИКА

COMPUTER SCIENCE

IRSTI 28.23.02

10.51889/2959-5894.2024.85.1.010

A.A. Abibullayeva[1*], G.N. Kazbekova[1], N.M. Zhunissov [1]

[1] Khoja Akhmet Yassawi International Kazakh-Turkish University, Turkistan, Kazakhstan

[*]e-mail: aiman.abibullayeva@ayu.edu.kz

KEYWORD EXTRACTION FROM KAZAKH TEXT WITH MACHINE LEARNING ALGORITHMS

Abstract

Browsing information on the internet in daily life has become a common activity for computer users. Since thousands of Internet news are published on the Internet every day, it is difficult to effectively retrieve and summarize the relevant documents. Therefore, the keyword or keyphrase extraction technique is used to provide the main content of a particular web page. Due to such needs, the use of keywords allows the reader to access the sought-after information easily and quickly. In this article, Random Forest and XgBoost (Extreme Gradient Boosting) algorithms, which are machine learning algorithms, were tested The results were obtained on the 500N-KPCrowd dataset, which consists of English-language news content widely used in the literature, and compared with the results obtained from the Kazakh language datasets. For the Kazakh data set, the highest result in the literature was achieved with the best F_1 score of 0.97. For the 500N-KPCrowd data set, the best F_1 score of 0.70 was obtained.

Keywords: keyword extraction, machine learning, Random Forest, XgBoost, statistical features, graphical features.

А.А.Абибуллаева[1], Г.Н.Казбекова.[1], Н.М.Жунисов[1]

[1] Қожа Ахмет Ясауи атындағы Халықаралық қазақ-түрік университеті, Түркістан қ., Казахстан

МАШИНАЛЫҚ ОҚЫТУ АЛГОРИТМДЕРІНІҢ КӨМЕГІМЕН ҚАЗАҚ ТІЛІНДЕГІ МӘТІННЕН ТҮЙІН СӨЗДЕРДІ АЛЫП АЛУ

Аңдатпа

Күнделікті өмірде интернеттегі ақпаратты шолу компьютер пайдаланушылары үшін әдеттегі әрекетке айналды. Интернетте күн сайын мыңдаған интернет жаңалықтары жарияланатындықтан, тиісті құжаттарды тиімді түрде алу және қорытындылау қиын. Сондықтан белгілі бір веб-беттің негізгі мазмұнын қамтамасыз ету үшін кілт сөзді немесе түйінді фразаны алу әдісі қолданылады. Осындай қажеттіліктерге байланысты түйінді сөздерді қолдану оқырманға қажетті ақпаратқа оңай және жылдам қол жеткізуге мүмкіндік береді. Бұл мақалада машиналық оқыту алгоритмдері болып табылатын Кездейсоқ орман және Градиентті күшейту алгоритмдері тексерілді. Нәтижелер әдебиетте кеңінен қолданылатын ағылшын тіліндегі жаңалықтар мазмұнынан тұратын 500N-KPCrowd деректер жинағында алынды және қазақ тіліндегі деректер жинақтарынан алынған нәтижелермен салыстырылды. Қазақ деректер жинағы үшін әдебиеттегі ең жоғары нәтиже 0,97 ең жақсы F_1 ұпайымен қол жеткізілді. 500N-KPCrowd деректер жинағы үшін 0,70 ең жақсы F_1 ұпайы алынды.

Түйін сөздер: кілт сөзді шығару, машиналық оқыту, Кездейсоқ орман, XgBoost, статистикалық ерекшеліктер, графикалық ерекшеліктер.

А.А. Абибуллаева[1], Г.Н. Казбекова[1], Н.М. Жунисов[1]
[1] Международный казахско-турецкий университет имени Ходжи Ахмеда Ясави,
г.Туркестан, Казахстан

ИЗВЛЕЧЕНИЕ КЛЮЧЕВЫХ СЛОВ ИЗ КАЗАХСКОГО ТЕКСТА С ПОМОЩЬЮ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ

Аннотация

Просмотр информации в Интернете в повседневной жизни стал обычным занятием для пользователей компьютеров. Поскольку каждый день в Интернете публикуются тысячи интернет-новостей, эффективно найти и обобщить соответствующие документы сложно. Таким образом, метод извлечения ключевых слов или ключевых фраз используется для предоставления основного содержимого конкретной веб-страницы. В связи с такими потребностями использование ключевых слов позволяет читателю легко и быстро получить доступ к необходимой информации. В этой статье были протестированы алгоритмы Случайног леса и Экстремального повышения градиента, являющиеся алгоритмами машинного обучения. Результаты были получены на наборе данных 500N-KPCrowd, который состоит из новостного контента на английском языке, широко используемом в литературе, и сравнивались с результатами, полученными на наборах данных на казахском языке. Для казахстанского набора данных самый высокий результат в литературе был достигнут с лучшим показателем F_1 равным 0,97. Для набора данных 500N-KPCrowd был получен лучший показатель F_1 равный 0,70.

Ключевые слова: извлечение ключевых слов, машинное обучение, Случайный лес, XgBoost, статистические особенности, графические особенности.

Introduction

The amount of digital data produced, consumed and stored all over the world is rapidly increasing. While the digital data produced in 2015 was approximately 15 zettabytes, it is estimated that this rate will be approximately 180 zettabytes in 2025. In Figure 1, digital data production has followed an increasing trend over the years [1].

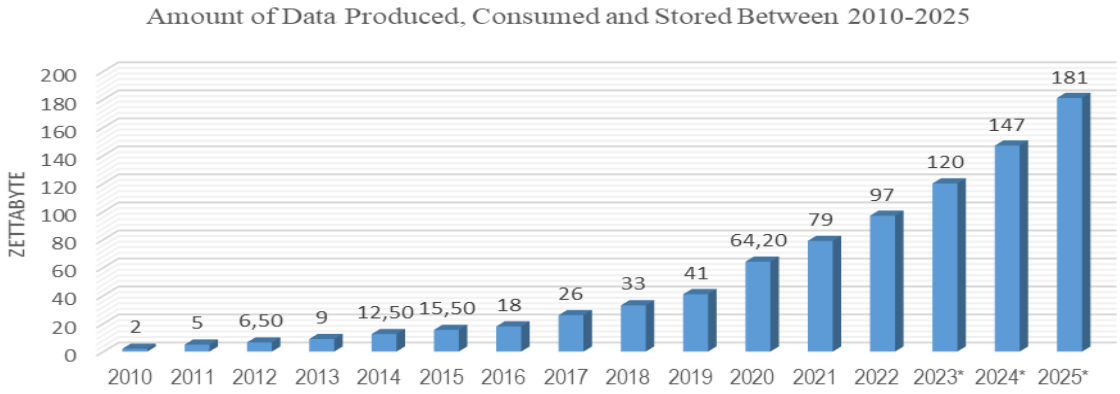


Figure 1. Estimated data growth over the years

With the rapid increase in digital content, finding the information sought among textual data masses has become a problem that needs to be solved. In order to access the desired information quickly and easily, keywords must be assigned to textual content. However, the keyword extraction problem is the main problem that needs to be solved in order to develop systems such as summarization, document linking and clustering. Keyword extraction can be done manually or automatically. Manual keyword extraction takes a long time and is not cost-effective for a mass of digital text. Therefore, researchers in the field of Natural Language Processing (NLP) have continuously focused on developing methods to automate this process.

When we look at the literature, the keyword is:

- index terms containing the most important information [2],

- a set of terms that provide summary information about the content for the reader [3],
- words that capture the main headings of a document [4],
- small pieces that capture the main idea or title of the text [5],
- words or groups of words that show the text in the clearest way [6],
- expressions that capture the main topic discussed in a document [7],
- words expressing all important aspects of the content of the document [8],
- words that give information about the content of the text [9].

Many models have been proposed for keyword extraction in the literature. However, when looking at the performance results of the proposed models, their problem solving performance is still far below expectations. These models are basically grouped under two headings: supervised and unsupervised. While supervised models require a pre-labeled training set, unsupervised methods do not require a pre-compiled dataset. Most unsupervised algorithms perform the task of keyword extraction using a single input document rather than a corpus. Previously, keyword extraction was solved with statistical methods or Natural Language Processing (NLP) techniques. With the emergence of machine learning technology in recent years, it has started to be solved with deep learning algorithms and artificial neural networks and better results have been obtained. Unsupervised models were first developed using statistical features such as Term Frequency (TF) and Inverse Document Frequency (TDF) [10]. In the same years, the problem of keyword extraction was addressed using linguistic features such as Part of Speech (PoS) and n-gram [11], [12]. Statistical and linguistic models provide powerful linguistic and statistical information about input terms. However, these methods cannot describe the semantic relationship between words and sentences. Graph-based [13], [14], [15]) and embedding-based [16], [17], [18], [19] models have been proposed to describe the semantic relationship between words and sentences. Graph-based Text Rank [13], Page Rank [14] and Graph-Based Technique for Extracting Keyphrases in a Single-Document (GTEK) [15] models extract attributes by drawing the word co-occurrence graph of the input text. These graphs count the number of times words co-occur as edges in a sliding window and aim to capture semantic information by calculating centralities. When examined in terms of the Kazakh language, although a limited number of natural language processing studies have been carried out with machine learning and deep learning methods, the issue of keyword extraction in the Kazakh language has not yet been addressed except for a study by Abibullayeva and Çetin [20].

Models were created using algorithms for keyword extraction in the literature, and the prediction performances of the obtained models were compared and it was examined which algorithm created more successful models in the data source used. Although there have been many studies on keyword extraction in English and other languages to date, the situation is different for the Kazakh language. The issue of keyword extraction in the Kazakh language has not yet been addressed. Machine learning and deep learning methods and natural language processing studies are limited to the Kazakh language. There is no model trained with deep learning yet for keyword extraction from Kazakh texts. For these reasons, it is important to conduct studies and make suggestions in the field of extracting keywords from news texts. It is known that the Kazakh people have been using the alphabet system based on Arabic graphics for centuries. From 1929 to 1940, the alphabet based on the Latin alphabet was included in the writing system, and since 1940 the Cyrillic alphabet has been used. In 2017, the new Latin alphabet of the Kazakh language was approved by the decree of the President of the Republic of Kazakhstan on October 26. It is planned to switch to a new alphabet between 2017-2025.

Currently, the issue of switching from Cyrillic to Latin is widely discussed in society. The transition to the Latin alphabet, which has become the language of all advanced technologies, is important for Kazakh art and culture. Kazakhstan's transition to the Latin alphabet is important both socio-economically and politically, as well as raising the Kazakh language to its deserved position in world civilization. The most important problem of the alphabet change is forgetting the old heritage. When the alphabet changes, access to the texts written in that alphabet becomes difficult and the connection with the past begins to decrease over time. Since there has been no previous study on keyword extraction in the Kazakh language, it is thought that this research will not only contribute to

the academic literature but also facilitate access to works written in the Cyrillic alphabet in the Kazakh language, thus supporting the preservation of cultural heritage. After the transition from the Cyrillic alphabet to the Latin alphabet, the derivation of keywords will enable connections between documents. In this way, accessing documents and searching for content will be easier.

Research methodology

With the rapid increase in the amount of data produced, consumed and stored, the problem of filtering information from big data has emerged. To solve this problem, automatic keyword assignment is made. The difficulty of the keyword extraction problem arises from the variable characteristics of different data sets. In Figure 2, data was first collected from the news sites zhasalash.kz, aikyn.kz, bilgimdinews.kz using the "data scraping" method using BeautifulSoup and Request libraries.

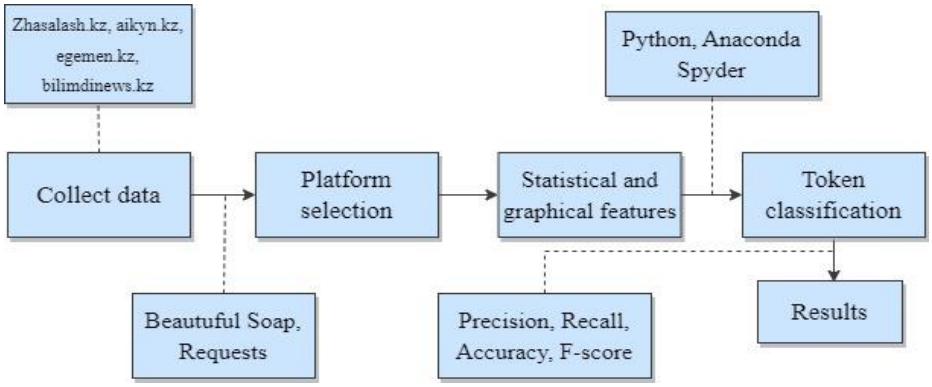


Figure 2. Process followed in model development

The KazakhNews dataset was compiled from web pages originally published in Kazakh language. Later, the Anaconda Spyder environment, which was widely used by researchers, was chosen as the platform because it had strong scientific features. Statistical and graphical attributes were calculated for each content text of the newly compiled data sets. These calculated features were passed through the Community classification module one by one and the sequence labeling task was completed. In the Token Classification module, Random Forest and Extreme Gradient Boosting (XgBoost) classification algorithms were trained and tested separately for each data set. In addition, the model was trained for the 500N-KPCrowd dataset, which consists of English-language news content frequently used in the literature, and the tested results are presented in a table. Summary statistics of the data sets are shown in Table 1, Token Classification module Random Forest and XgBoost classification algorithms were trained and tested separately for each dataset.

Table 1. Dataset summary statistics

Dataset	*Language*	*Field*	*Title*	*#document*	*#keyword*
KazakhNews	*Kazakh*	*News website*	*Politics, literature, etc.*	*1000*	*5*
500N-KPCrowd	*English*	*News website*	*art and culture, crime, fashion, business, health, world politics, politics, sports, science, and technology*	*400*	*49*

We made the KazakhNews dataset publicly available at https://github.com/Aiman128792/Kazakh_News.

Random forest algorithm is a managerial machine learning method that consists of multiple decision trees. Each decision tree reaches the bottom leaf of the tree by visiting all random nodes according to the input given in the training set and depending on the conditions in these nodes. Within the scope of this study, each tree was trained with words and word groups in the nodes of the decision trees, and with the class of the Internet page in the leaf node.

During this training, criterion variables such as tree depth, number of trees, gini-entropy, which are the parameters of the random forest algorithm, were optimized with grid search. It was used to solve the multi-class classification problem. The attributes used as input for the Random Forest algorithm are grouped under statistical attributes and word graph attributes.

Results of the study

In the Token Classification module, Random Forest and XgBoost ensemble classification algorithms were trained and tested separately for each dataset.

Table 2 shows the performance results of the proposed model for the KazakhNews dataset. When the table is examined, with the combination of statistical features and graphical features, Extreme Gradient Boosting gives a F1-score of 0.88 for the Kazakh data set, while the Random Forest model has the best results with an F1-score of 0.97.

Table 2. Performance results for the KazakhNews dataset

Token Classification	*Metrics*	*Statistical Features*	*GraphicalFeatures*	*StatisticalFeatures + Graphical Features*
Random Forest	*Accuracy*	*0,994*	*0,987*	*0,995*
	Precision	*0,978*	*0,904*	*0,988*
	Recall	*0,956*	*0,956*	*0,958*
	F_1-score	***0,967***	***0,929***	***0,973***
XgBoost	*Accuracy*	*0,978*	*0,958*	*0,981*
	Precision	*0,919*	*0,838*	*0,936*
	Recall	*0,821*	*0,641*	*0,847*
	F_1-score	***0,867***	***0,726***	***0,889***

Evaluation Metrics

In order to accurately evaluate the performance of the model, in addition to the accuracy value, the so-called F1 score was also monitored. The success of the algorithms used is evaluated using criteria such as accuracy, precision, sensitivity and f-criterion (F-Score), which determine the degree of performance of the created models. True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN) values are used in calculating the F1 score. In this case, TP and TN are considered correct results, and FP and FN are considered incorrect results. The accuracy value is calculated by the ratio of the TP and TN values correctly predicted by the model to all predicted TP, TN, FP, FN values.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

The precision value is the ratio of the number of TP values predicted by the model to the number of TP and FP values, which are all positive results produced by the model.

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Recall - Calculates the proportion of positive values that are correctly predicted.

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

To evaluate the performance of the model, the F1-score, which is the harmonic mean between accuracy and precision, is measured.

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

F_1-score criterion is used in evaluating keyword extraction algorithms. In calculating this score, the confusion matrix created by looking at the actual value/predicted value numbers of the predicted values is used. Figure 3, shows the complexity matrix.

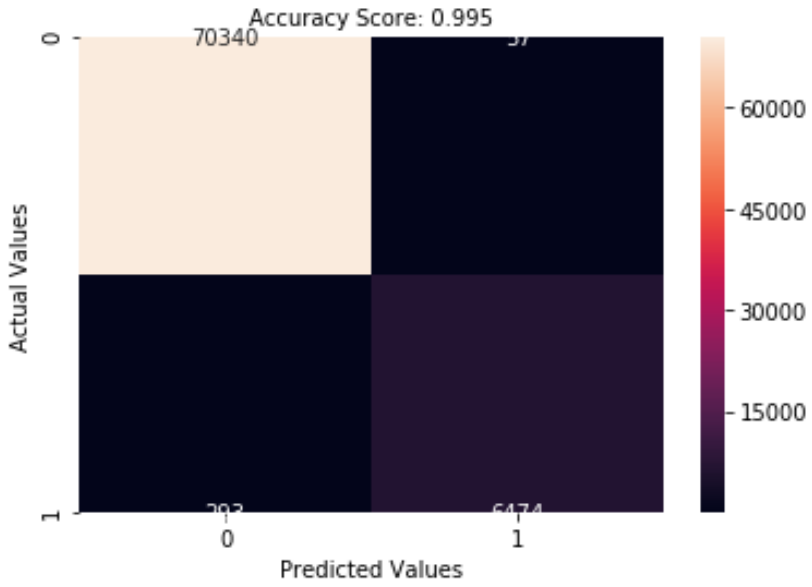


Figure 3. Confusion matrix of Random Forest model for Kazakh dataset

There is the Confusion matrix in Table 2 The matrix contains TP values, which are actually keywords and are predicted as keywords, FP, which are not actually keywords but are predicted as keywords, FN, which are not actually keywords but are marked as keywords, and finally TN, which are not actually keywords but are marked as not keywords.

Table 2. Confusion matrix for Kazakh data set

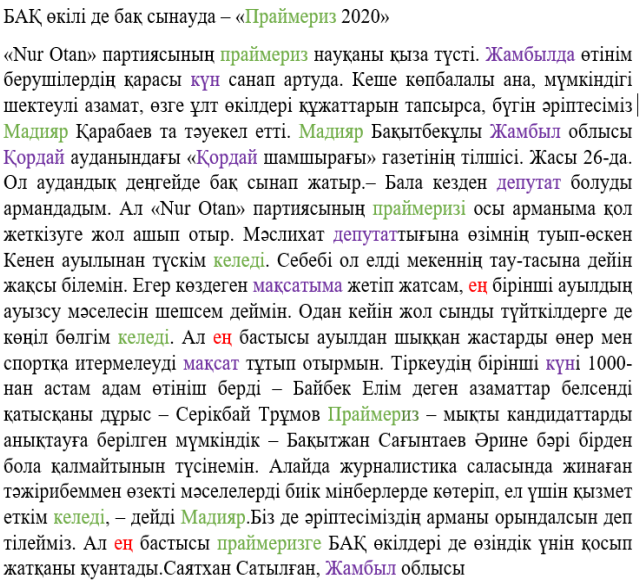
	Positive	*Negative*
Positive	*GN= 70340*	*YP=57*
Negative	*YN= 293*	*GP=6474*

Table 3 shows the performance results of my model for the 500N-KPCrowd dataset. For this dataset, the highest F1-score of XgBoost of 0.57 and Random Forest of 0.70 were obtained. For the 500N-KPCrowd dataset, the Random Forest algorithm used both feature groups together to increase the performance [21].

Table 3. Performance results for the 500N-KPCrowd dataset

Token Classification	*Metrics*	*Statistical Features*	*Graphical Features*	*Statistical Features + Graphical Features*
Random Forest	*Accuracy*	*0,791*	*0,738*	*0,803*
	Precision	*0,725*	*0,643*	*0,779*
	Recall	*0,688*	*0,638*	*0,643*
	F_1-score	***0,706***	***0,641***	***0,705***
XgBoost	*Accuracy*	*0,746*	*0,706*	*0,750*
	Precision	*0,754*	*0,714*	*0,763*
	Recall	*0,451*	*0,322*	*0,456*
	F_1-score	***0,565***	***0,444***	***0,571***

Figure 4 shows the labeling result of the model for an example from the KazakhNews dataset. By simulating tagged keywords for a summary from the KazakhNews dataset, true positives are in green and false negatives are in red. Purple represents keys that are false positives.



БАҚ өкілі де бақ сынауда – «Праймериз 2020»

«Nur Otan» партиясының праймериз науқаны қыза түсті. Жамбылда өтінім берушілердің қарасы күн санап артуда. Кеше көпбалалы ана, мүмкіндігі шектеулі азамат, өзге ұлт өкілдері құжаттарын тапсырса, бүгін әріптесіміз Мадияр Қарабаев та тәуекел етті. Мадияр Бақытбекұлы Жамбыл облысы Қордай ауданындағы «Қордай шамшырағы» газетінің тілшісі. Жасы 26-да. Ол аудандық деңгейде бақ сынап жатыр.– Бала кезден депутат болуды армандадым. Ал «Nur Otan» партиясының праймеризі осы арманыма қол жеткізуге жол ашып отыр. Мәслихат депутаттығына өзімнің туып-өскен Кенен ауылынан түскім келеді. Себебі ол елді мекеннің тау-тасына дейін жақсы білемін. Егер көздеген мақсатыма жетіп жатсам, ең бірінші ауылдың ауызсу мәселесін шешсем деймін. Одан кейін жол сынды түйткілдерге де көңіл бөлгім келеді. Ал ең бастысы ауылдан шыққан жастарды өнер мен спортқа итермелеуді мақсат тұтып отырмын. Тіркеудің бірінші күні 1000-нан астам адам өтініш берді – Байбек Елім деген азаматтар белсенді қатысқаны дұрыс – Серікбай Трұмов Праймериз – мықты кандидаттарды анықтауға берілген мүмкіндік – Бақытжан Сағынтаев Әрине бәрі бірден бола қалмайтынын түсінемін. Алайда журналистика саласында жинаған тәжірибеммен өзекті мәселелерді биік мінберлерде көтеріп, ел үшін қызмет еткім келеді, – дейді Мадияр.Біз де әріптесіміздің арманы орындалсын деп тілейміз. Ал ең бастысы праймеризге БАҚ өкілдері де өзіндік үнін қосып жатқаны қуантады.Саятхан Сатылған, Жамбыл облысы

Figure 4. Labeling result of the model for an example for KazakhNews

Discussion

Experimental results show that the proposed model can label independently of the domain and other characteristic features of the dataset.

Conclusion

In this article, Random Forest and XgBoost algorithms were tested for keyword extraction from Kazakh news texts. In the study, the statistical and graphical features of the text were tested both separately and in combination with each other. Two new data sets, KazakNews, using the Cyrillic alphabet, were created to use in training and testing the model and to compare the performance of the model in different languages. In addition to the data set, the performance results of the model were obtained for the Latin 500N-KPCrowd data set, which is widely used in the literature and contains news texts. The model was trained separately for each language with different ML algorithms and different data sets. In the study, it was seen that the Random Forest algorithms had very similar performance for the two data sets. The highest result (0.97 f-score) for the KazakhNews dataset was obtained by using Random Forest together with statistical and graphical features.

References

[1] Reinsel, D., Gantz, J., & Rydning, J. The digitization of the World, IDC, 2018. 1-28.

[2] Kaur, J., Gupta, V. Effective approaches for extraction of keywords. International Journal of Computer Science Issues, 2010, (IJCSI), 7(6), 144.

[3] Liu, Z., Huang, W., Zheng, Y., and Sun, M. Automatic keyphrase extraction via topic decomposition. In Proceedings of the 2010 conference on empirical methods in natural language processing, 2010, October. - pp. 366-376. Association for Computational Linguistics.

[4] Kim, S. N., Medelyan, O., Kan, M. Y., and Baldwin, T. Automatic keyphrase extraction from scientific articles. Language Resources and Evaluation, 2013, 47(3), 723-742.

[5] Awajan, A. A. Unsupervised Approach for Automatic Keyword Extraction from Arabic Documents. In Proceedings of the 26th Conference on Computational Linguistics and Speech Processing (ROCLING 2014), 2014, October. - pp. 175-184.

[6] Birdevrim, A. S., Boyacı, A., and S Al Thani, D. A. İyileştirilmiş otomatik anahtar kelime çıkarımı BHO-AKÇ. İstanbul Ticaret Üniversitesi Teknoloji ve Uygulamalı Bilimler Dergisi, 2018. 1(1), 11-19.

[7] Basaldella, M., Antolli, E., Serra, G., and Tasso, C. Bidirectional lstm recurrent neural network for keyphrase extraction. In Italian Research Conference on Digital Libraries, 2018, January. pp. 180-187. Springer, Cham.

[8] Papagiannopoulou, E., Tsoumakas, G. A review of keyphrase extraction. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2020. 10(2), e1339. 1-59.

[9] Kılıç Ünlü, H., & Çetin, A. Keyword extraction as sequence labeling with classification algorithms. Neural Computing and Applications, 2023. 35(4), 3413-3422.

[10] Ramos, J. Using tf-idf to determine word relevance in document queries. In Proceedings of the first instructional conference on machine learning. 2003, December. Vol. 242, No. 1, pp. 29-48.

[11] Tomokiyo, T., & Hurst, M. A language model approach to keyphrase extraction. In Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment, 2003, July. - pp. 33-40.

[12] Haddoud, M., Mokhtari, A., Lecroq, T., & Abdeddaïm, S. Accurate Keyphrase Extraction from Scientific Papers by Mining Linguistic Information. In CLBib@ ISSI, 2015, June. - pp. 12-17.

[13] Mihalcea, R., Tarau, P. TextRank: Brining order into texts. In Proceedings of EMNLP 2004, Association for Computational Linguistics, Barcelona, Spain, 2004. – p 404-411.

[14] Zhao, W. X., Jiang, J., He, J., Song, Y., Achananuparp, P., Lim, E. P., & Li, X. Topical keyphrase extraction from twitter. In Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1, June (pp. 379-388). Association for Computational Linguistics. 2011, p 1-10.

[15] Alfarra, M. R., & Alfarra, A. Graph-Based Technique for Extracting Keyphrases in a Single-Document (GTEK). In 2018 International Conference on Promising Electronic Technologies (ICPET), 2018, October. (pp. 92-97). IEEE.

[16] Benani-Smires, K., Musat, C., Hossmann, A., Baeriswyl, M., & Jaggi, M. Simple unsupervised keyphrase extraction using sentence embeddings. arXiv preprint arXiv:1801.04470, 2018. p 1-9.

[17] Sun, Y., Qiu, H., Zheng, Y., Wang, Z., & Zhang, C. SIFRank: A New Baseline for Unsupervised Keyphrase Extraction Based on Pre-Trained Language Model. IEEE Access, 8, 2020, 10896-10906.

[18] Liang, X., Wu, S., Li, M., & Li, Z. Unsupervised keyphrase extraction by jointly modeling local and global context. arXiv preprint arXiv:2109.07293, 2021.p 1-10.

[19] Ajallouda, L., Fagroud, F. Z., Zellou, A., & Lahmar, E. B. KP-USE: An Unsupervised Approach for Key-Phrases Extraction from Documents. International Journal of Advanced Computer Science and Applications, 2022. 13(4). 283-289.

[20] Abibullayeva, A., & Çetin, A. Keyword Extraction from Kazakh News Dataset with BERT. El-Cezeri, 9(4), 2022, pp 1193-1200.

[21] Abibullayeva, A. A novel ensemble keyword extraction model in the kazakh language with machine learning // Ph. D. Thesis, Gazi university, Turkey. – 2023. 1-106.