Topic Modeling Using LDA and BERT Techniques: Teknofest Example

Conference Paper · September 2021

DOI: 10.1109/UBMK52708.2021.9558988

3 authors, including:

Ercan Atagün
Duzce University
15 PUBLICATIONS **22** CITATIONS
SEE PROFILE

Ahmet Albayrak
Karadeniz Technical University
40 PUBLICATIONS **96** CITATIONS
SEE PROFILE

Topic Modeling Using LDA and BERT Techniques: Teknofest Example

Ercan Atagün
Duzce University
Computer Engineering
Duzce, Turkey
ercan-atagun@hotmail.com

Bengisu Hartoka
Duzce University
Faculty of Technology
Duzce, Turkey
bengisu.hartoka@gmail.com

Ahmet Albayrak
Duzce University
Computer Engineering
Duzce, Turkey
ahmetalbayrak@duzce.edu.tr

2021 6th International Conference on Computer Science and Engineering (UBMK) | 978-1-6654-2908-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/UBMK52708.2021.9558988

***Abstract*—This paper is a natural language processing study and includes models used in natural language processing. In this paper, topic modeling, which is one of the sub-fields of natural language processing, has been studied. In order to make topic modeling, the data set was obtained by using the data scraping method, which has been very popular in recent years, over social media. The dataset is related to Teknofest competitions. The dataset was created by utilizing the Selenium library, one of the popular libraries used for the data scraping method. In order to be able to analyze on the prepared data set and to ensure the consistency of the clustering process, the text to be used before the analysis was preprocessed. After text preprocessing, clustering was performed on the data set with natural language processing techniques such as BERT and LDA.**

Keywords—Data Scraping, LDA, Bert, Topic Modeling, Teknofest.

I. Introduction

Teknofest (Aviation, Space and Technology Festival) is a technology, aviation and space technology festival held in Turkey. It is the first and only festival organized by the Turkish Technology Team Foundation (also known as T3) and the Ministry of Industry and Technology, aiming to develop Turkey's national technologies. In general, it aims to promote and develop national technologies and to raise awareness of people in this field. While seminars, award-winning technology competitions, domestic technology exhibitions and an international entrepreneurship summit are taking place in the festival; Parachuting, SOLO TURK, Turkish Stars and similar local-international organizations air shows are also held. The festival is held one year in Istanbul and one year in one of the Anatolian provinces. In the festival, where many ministries and presidency are stakeholders, the stakeholder institutions vary according to the years, but these stakeholders are generally ASELSAN, ROKETSAN, Baykar, BMC, HAVELSAN, TÜBİTAK, STM, Turkish Aerospace(TUSAŞ), Turkish Airlines, Turkish Aeronautical Association, Türksat. and as supporters and participants such as Turkcell; is in establishment. In addition, the leading universities of our country provide support by providing academic consultancy services[1].

BERT is a complex trained model released by Google in 2018 that represents sentences in accordance with their word meanings. Since it can represent the sentence very well, it has performed very well in almost all sub-fields of NLP. Context-free models such as Word2vec and GloVe create a single word embedding for each word in the vocabulary [2]. Instead, contextual models such as BERT produce a representation of each word based on other words in the sentence. BERT captures these relationships bidirectionally as a contextual model. BERT models have succeeded in gaining superiority over other models in many areas in the field of natural language processing in recent years [3]. These models are very complex in terms of their inner workings and require a significant amount of data and processing power for training. However, there are many ready-made BERT models. It is possible to achieve good results by using them and even by feeding them with additional training sets depending on the task to be done. The main goal of the BERT model is to create a fixed-length vector representing sentences, taking into account the meaning and order of words. BERT models differ depending on the language spoken. For Turkish, a German researcher "Stefan Schweter" published a Turkish-specific model, which can be accessed in Python with the transformers package [4], [5].

LDA is a probability-based topic modeling method. The model generates topics based on word weight from a set of documents. LDA is an example of topic modeling where each document is assumed to be a collection of topics and each word in the document corresponds to one of these topics. LDA is an unsupervised learning algorithm, it does not need predefined words. After the number of topics is determined, labels are assigned to the topics according to the classes[6]. Since the Bert model was used in this study, the use of GPU was needed. For this reason, Google Colaboratory was preferred as the coding environment for topic modeling. The most important reason for choosing Google Colaboratory is that it provides free GPU, unlike other cloud services. For topic modeling, vectors obtained from BERT and LDA were combined and clustering was applied [7], [8]. This study consists of data set cleaning and preparation, model development and experimental studies, and conclusions.

II. Material and Method

A. Data Set

While preparing the data set used in this study, Visual Studio Code was used as the coding environment. Visual Studio Code is a fast, lightweight and software development tool that can be used on different operating systems. The tool, which has a simple text editor view, supports many programming languages such as Python, C/C++, C#, Javascript, thanks to plugins. In this study, the Python programming language was used [9].

The data set used in this study was prepared by using the Selenium library with the Web Scraping method, which has been very popular in recent years. Web Scraping is a technique used to retrieve data from web pages. It is an automated process in which the HTML codes of the web page are processed to extract data for manipulation, by executing on an

978-1-6654-2908-5/21/$31.00 ©2021 IEEE

application's web page, copying it to a local database or a generated file for later analysis [9]. In this study, a bot was written to pull tweets on Twitter with Python programming language and HTML codes using Selenium library [10]. Using this bot, a total of 737 tweets (May 15, 2016 - November 20, 2020) were pulled from the Twitter pages of Teknofest and T3 foundation chairman of the board of trustees, Selçuk Bayraktar. Then, the captured tweets were saved in the csv file created in a sequential and orderly manner. In Figure 1, the block diagram showing the process steps is given.

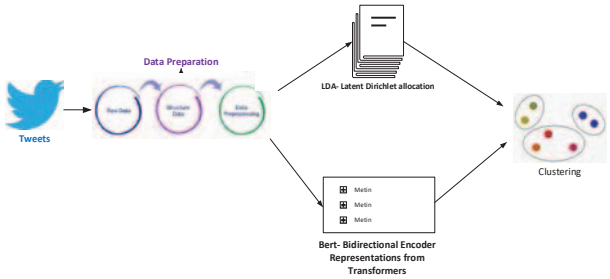

Fig. 1. Block diagram of the study

B. Text Preprocessing

In order to be able to analyze the text and to ensure the consistency of the clustering process, the text to be used before the analysis needs to be pre-processed. These processes are tokenization, removal of frequently used and meaningless stop words, conversion of all characters to lowercase, removal of information such as ip address or e-mail, numbers in tweets, but extracting them from the text, all kinds of punctuation and can be given as deletion of special characters. In Figure 1, these processes are given as data preparation. All these operations were carried out by defining separate functions. In this study, NLTK (Natural Language Toolkit) library was used for these operations. NLTK (Natural Language Toolkit) stands for natural language toolkit [11]. It is an open source natural language processing library [12]–[14] developed with the Python programming language to work with human language data and built with over 50 corpus and lexical resources under development. In Table 1, the raw text data prepared for this study is given before the text preprocessing process is applied.

TABLE I. ALGORITHM AND SCORE

Id	*Raw Data (tweet)*	*Text Preprocessing (tweet)*	*Tokenize (tweet)*
445	İHA dediğin doğa dostu olmalı... #MilliTeknolojiHamlesi	iha dediğin doğa dostu olmalı milli teknoloji hamlesi	‘İha’,’dediğin’,doğa’,’dostu’ , ‘milli’ ‘teknoloji’ ‘hamlesi’
306	Milli Eğitim Bakanlığı'nın en eğlenceli okulu #TEKNOFEST'te @tcmeb alanında! #MilliTeknolojiHamlesi 20 Eylül Atatürk Havalimanı	milli eğitim bakanlığının eğlenceli okulu teknofestte alanında milli teknoloji hamlesi 20 eylül atatürk havalimanı	'milli', 'eğitim', 'bakanlığının', 'okulu','teknofestte','alanında', 'milli','teknoloji', 'hamlesi', 'eylül','atatürk', 'havalimanı'
38	Çevre ve Enerji Teknolojileri Yarışması #TEKNOFEST yarışmacıları temiz bir çevre, verimli	çevre ve enerji teknolojileri yarışması teknofest yarışmacıları	'enerji','teknolojileri', 'yarışması','yarışmacıları', 'çevre','verimli','gelecek',
	bir gelecek için yarışıyorlar. @SankoholdingAs 24-27 Eylül Gaziantep Ortadoğu Fuar Merkez	temiz bir çevre verimli bir gelecek için yarışıyorlar 2427 eylül gaziantep ortadoğu fuar merkezi	'yarışıyorlar','gaziantep', 'fuar','merkezi'

Text preprocessing was applied in two stages. First, simple preprocessing was applied, all letters were converted to lowercase, non-alphanumeric characters were cleared by writing a function. Table 1 shows tweets with simple text preprocessing. As the last step of the text preprocessing applied in this study, the stop words of each tweet and the numerical values they contain were cleared and separated word by word, that is, tokenized. The result of this process is given in Table 1. After all these processes, the data set is ready to be given to the prepared model.

C. Model Development and Experimental Studies

This paper is a natural language processing study and has been studied on topic modeling, which is one of the sub-fields of natural language processing. There are generally two ways to define a topic. Similar structures in vector space can be clustered by resorting to hierarchical Bayesian models such as LDA (Latent Dirichlet Allocation) or by embedding the document in vector spaces. LDA is known to not do very well at processing short texts when there is not much text to model. Therefore, this study requires a technique that places the entire content of the sentence and can then cluster similar topics [15]–[17]. Therefore, creating vector representations and clustering may be the solution. Methods such as TD-IF, BERT have been tried to obtain vector representations from texts. TF-IDF is obtained by multiplying the term frequency (TF) of terms with the inverse document frequency (IDF). Class information of terms is not used in this weighting. Therefore, it is an unsupervised weighting method [8], [18], [19].

To cluster embeddings from BERT and LDA models and to find topics in clusters, it is necessary to ensure that documents with similar topics are clustered together. WordCloud was used for the visualization process. WordCloud is a data visualization technique used to represent text data where the size of each word indicates its frequency or importance. Important textual data points can be highlighted using a word cloud. WordCloud is widely used to analyze data from social networking websites. The libraries required to create WordCloud in Python are matplotlib, pandas and wordcloud [20], [21].

In this study, the result sets created by using the TF-IDF model were not balanced and sufficient for this study. Because TF-IDF is also word bag-based (ignoring grammar and word order), it loses contextual information, so a good enough result could not be obtained with the TF-IDF method. As given in Figure 2, the vectors obtained with TF-IDF do not distribute well enough and the processing results are not good in clustering.

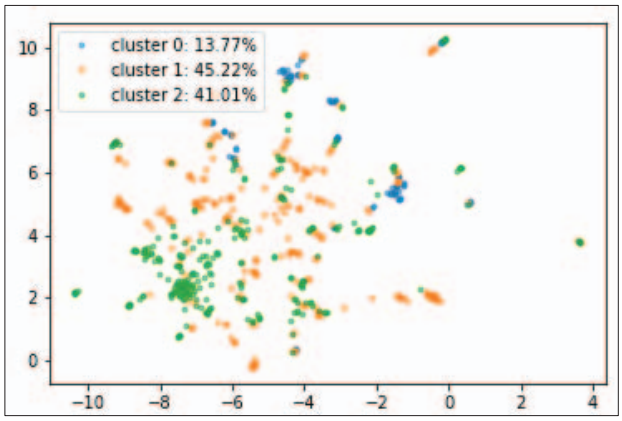

Fig. 2. Clustering result of vectors obtained with TF-IDF

The word distributions formed by the TF-IDF process are given in Figure 3 with word clouds.

Fig. 3. Word clouds obtained as a result of TF-IDF process

Vector representations of the tweets in the dataset were created using the BERT model. The Distilbert model was used for this study. This is because it provides a nice balance between speed and performance, and this package also includes a language model for Turkish [22]. When the vector representations from the BERT model were clustered, good enough results were not obtained. The clustering results made with the vectors obtained with the BERT model are given in Figure 4.

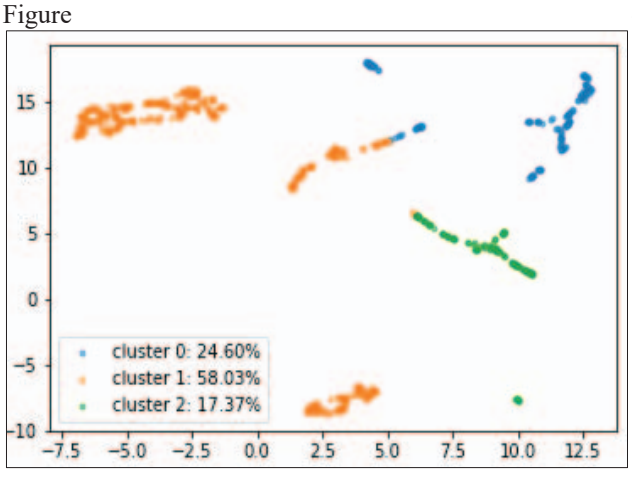

Fig. 4. Clustering result of vectors obtained with BERT

An important feature of the BERT model is that it places special start and end marks at the beginning and end of each text. The model outputs corresponding to the starting mark can be considered as the features that the model gives to the whole text [15]. As seen in Figure 4, the distribution obtained with the BERT model has been much more successful than the TF-IDF model, but it is still not a good enough result. In Figure 5, the word cloud obtained as a result of using the BERT model is given.

Fig. 5. Word clouds obtained as a result of the BERT model

In this study, the Clustering process was applied by combining the vector representations from BERT and LDA to make topic modeling, and the best result was obtained in this case. By combining LDA, BERT and clustering, semantic information can be preserved and contextual issue identification can be achieved. The model was designed by combining the probabilistic topic assignment vector of LDA obtained from the LDA model and the sentence vectors obtained from the BERT model [23]–[25]. Figure 6 shows the distribution resulting from Bert, LDA and clustering.

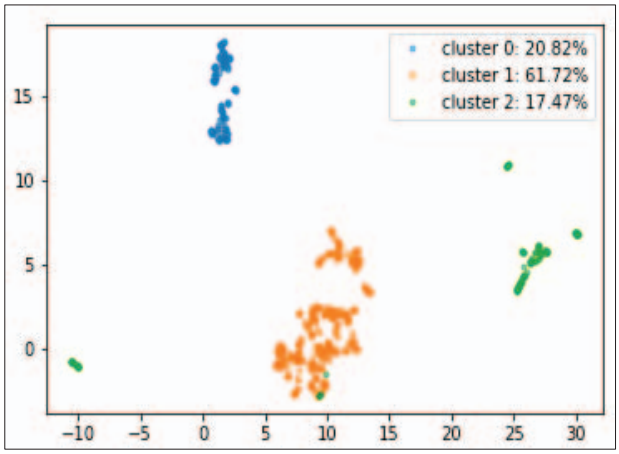

Fig. 6. Clustering result of vectors obtained with BERT and LDA

The word clouds of the topic modeling as a result of BERT, LDA and clustering are given in Figure 7.

Fig. 7. Word clouds formed as a result of BERT + LDA model

III. Conclusion

This study is a natural language processing study and natural language processing models are used. The data set was prepared by using the data scraping method, which has been very popular in recent years, on social media for topic modeling. The prepared data set was given to the model as a result of the preprocessing. Vector representations obtained with TF-IDF and BERT models were clustered and it was observed that these methods did not give good enough results. It has been determined that the most appropriate method for this study is topic modeling by clustering vector representations obtained from BERT and LDA models. The model was designed by combining the probabilistic subject assignment vector of LDA obtained from the LDA model and the sentence vectors obtained from the BERT model. By combining LDA, BERT and clustering, semantic information was preserved and contextual topic modeling was created. The best performance with balanced and separate clusters was obtained with this model. In experimental studies, while TF-IDF + clustering was 30%, LDA+ clustering was 43%, Bert+ clustering was 54%, Bert+LDA+ clustering resulted in 83% good results. Silhouette coefficient was taken into account as a success criterion in topic modeling. The success rates were found to be 0.21, 0.33, 0.67 and 0.83, respectively.

Teknofest events, which started in the recent past in our country and are very popular, are wanted to be included in the literature with this study. Because, following this study, it is desired to carry out academic studies on Teknofest competitions. Among the studies to be done in the future, the analysis of the feelings shared by Teknofest participants during the events, and the comparison of other international technology festivals to ensure the participation of wider segments of the event can be given.

References

[1] Wikipedia "Teknofest." https://tr.wikipedia.org/wiki/Teknofest (accessed Jun. 30, 2021)

[2] R. E. Lopez-Martinez and G. Sierra, "Research trends in the international literature on natural language processing, 2000-2019 - A bibliometric study," J. Scientometr. Res., vol. 9, no. 3, pp. 310–318, 2020, doi: 10.5530/JSCIRES.9.3.38.

[3] S. Arts, J. Hou, and J. C. Gomez, "Natural language processing to identify the creation and impact of new technologies in patent text: Code, data, and new measures," Res. Policy, vol. 50, no. 2, p. 104144, 2021, doi: 10.1016/j.respol.2020.104144.

[4] Y. Piris and A. C. Gay, "Customer satisfaction and natural language processing," J. Bus. Res., vol. 124, no. January 2020, pp. 264–271, 2021, doi: 10.1016/j.jbusres.2020.11.065

[5] "Doğal Dil İşleme Teknikleri Kullanılarak Disiplinler Arası Lisansüstü Ders İçeriği Hazırlanması Preparing Interdisciplinary Graduate Course Contents Using Natural Language Processing Techniques," pp. 373–383, 2020, doi: 10.17671/gazibtd.714447.

[6] T. M. Georgescu, "Natural language processing model for automatic analysis of cybersecurity-related documents," Symmetry (Basel)., vol. 12, no. 3, 2020, doi: 10.3390/sym12030354.

[7] Z. Zhu, J. Liang, D. Li, H. Yu, and G. Liu, "Hot Topic Detection Based on a Refined TF-IDF Algorithm," IEEE Access, vol. 7, pp. 26996–27007, 2019, doi: 10.1109/ACCESS.2019.2893980.

[8] Z. Hu, J. Luo, C. Zhang, and W. Li, "A Natural Language Process-Based Framework for Automatic Association Word Extraction," IEEE Access, vol. 8, pp. 1986–1997, 2020, doi: 10.1109/ACCESS.2019.2962154.

[9] H. Aldabbas, A. Bajahzar, M. Alruily, A. A. Qureshi, R. M. Amir Latif, and M. Farhan, "Google Play Content Scraping and Knowledge Engineering using Natural Language Processing Techniques with the Analysis of User Reviews," J. Intell. Syst., vol. 30, no. 1, pp. 192–208, 2020, doi: 10.1515/jisys-2019-0197.

[10] T. K. Mackey et al., "Big data, natural language processing, and deep learning to detect and characterize illicit COVID-19 product sales: Infoveillance study on Twitter and Instagram," JMIR Public Heal. Surveill., vol. 6, no. 3, pp. 1–17, 2020, doi: 10.2196/20794.

[11] S. O. Park and N. Hassairi, What predicts legislative success of early care and education policies?: Applications of machine learning and natural language processing in a cross-state early childhood policy analysis, vol. 16, no. 2 Febuary. 2021.

[12] A. Ortigosa, J. M. Martín, and R. M. Carro, "Sentiment analysis in Facebook and its application to e-learning," Comput. Human Behav., vol. 31, no. 1, pp. 527–541, 2014, doi: 10.1016/j.chb.2013.05.024.

[13] M. Heilman and N. Madnani, "ETS: Domain adaptation and stacking for short answer scoring," *SEM 2013 - 2nd Jt. Conf. Lex. Comput. Semant., vol. 2, no. SemEval, pp. 275–279, 2013i.

[14] H. Jang, Y. Jeong, and B. Yoon, "TechWord: Development of a technology lexical database for structuring textual technology information based on natural language processing," Expert Syst. Appl., vol. 164, no. September 2020, p. 114042, 2021, doi: 10.1016/j.eswa.2020.114042.

[15] A. Mahadevan and M. Arock, "Integrated topic modeling and sentiment analysis: A review rating prediction approach for recommender systems," Turkish J. Electr. Eng. Comput. Sci., vol. 28, no. 1, pp. 107–123, 2020, doi: 10.3906/elk-1905-114.

[16] L. Yao, Y. Zhang, Q. Chen, H. Qian, B. Wei, and Z. Hu, "Mining coherent topics in documents using word embeddings and large-scale text data," Eng. Appl. Artif. Intell., vol. 64, no. January, pp. 432–439, 2017, doi: 10.1016/j.engappai.2017.06.024.

[17] J. C. Campbell, A. Hindle, and E. Stroulia, "Latent Dirichlet Allocation: Extracting Topics from Software Engineering Data," Art Sci. Anal. Softw. Data, vol. 3, pp. 139–159, 2015, doi: 10.1016/B978-0-12-411519-4.00006-9.

[18] F. B. Enstitüsü, B. M. A. B. D, T. Fakültesi, and B. Mühendisliği, "Ver i b i l i m i derg i s i," vol. 4, no. 1, pp. 53–60, 2021

[19] J. Beel, B. Gipp, S. Langer, and C. Breitinger, "Research-paper recommender systems: a literature survey," Int. J. Digit. Libr., vol. 17, no. 4, pp. 305–338, 2016, doi: 10.1007/s00799-015-0156-0.

[20] E. Gümüşsu, "'Tamam' ve 'Devam' Etiketleriyle Atılan Tweetler ile Cumhur ve Millet İttifakı Grupları Arasındaki İlişkinin İncelenmesi," Bilişim Teknol. Derg., pp. 287–298, 2019, doi: 10.17671/gazibtd.453272.

[21] I. Chaturvedi, E. Cambria, R. E. Welsch, and F. Herrera, "Distinguishing between facts and opinions for sentiment analysis : Survey and challenges," Inf. Fusion, vol. 44, no. June 2017, pp. 65–77, 2018, doi: 10.1016/j.inffus.2017.12.006.

[22] C. Zuheros, E. Martínez-Cámara, E. Herrera-Viedma, and F. Herrera, "Sentiment Analysis based Multi-Person Multi-criteria Decision Making methodology using natural language processing and deep learning for smarter decision aid. Case study of restaurant choice using TripAdvisor reviews," Inf. Fusion, vol. 68, no. October 2020, pp. 22–36, 2021, doi: 10.1016/j.inffus.2020.10.019.

[23] D. Dessì, F. Osborne, D. Reforgiato Recupero, D. Buscaldi, and E. Motta, "Generating knowledge graphs by employing Natural Language Processing and Machine Learning techniques within the scholarly domain," Futur. Gener. Comput. Syst., vol. 116, pp. 253–264, 2021, doi: 10.1016/j.future.2020.10.026.

[24] D. Mazzei, F. Chiarello, and G. Fantoni, “Analyzing Social Robotics Research with Natural Language Processing Techniques,” Cognit. Comput., vol. 13, no. 2, pp. 308–321, 2021, doi: 10.1007/s12559-020-09799-1.

[25] L. Lucy, D. Demszky, P. Bromley, and D. Jurafsky, “Content Analysis of Textbooks via Natural Language Processing: Findings on Gender, Race, and Ethnicity in Texas U.S. History Textbooks,” AERA Open, vol. 6, no. 3, p. 233285842094031, 2020, doi: 10.1177/2332858420940312.