European Journal of
Educational & Social Sciences
Volume 6 Issue 1
May 2021
ISSN: 2564-6621

Article

Analyzing User Comments on Covid-19 Pandemic with Word2Vec Technique

Nimet AKSOY[1] **Özlem TÜLEK**[3]

Özlem AYDIN[2] **Erol ÖZÇEKİÇ**[4]

Manuscript information:

Received: February 1, 2021
Revised: May 30, 2021
Accepted: May 31, 2021

Author 1
Electric - Electronics Engineering, Institute of Science, Balıkesir University,, TURKEY
E-mail:
nimet.aksoy@balikesir.edu.tr

Author 2
Industrial Engineering, Institute of Science, Balıkesir University, TURKEY
E-mail:
ozlem.aydin@balikesir.edu.tr

Author 3
Computer and Information Engineering/ Institute of Science/ Sakarya University, TURKEY. E-mail:
ozlemtulek@balikesir.edu.tr

Author 4
Lecturer, Bigadiç Vocational School, Balıkesir University, TURKEY. E-mail:
erolozcekic@balikesir.edu.tr

Abstract

In Covid-19 pandemic, people spend more time at home than before the pandemic. Due to this reason, more time is spent on the internet than before. People expressed their views and assessments about Covid-19 pandemic on social media. Within the scope of this study, we collected people's comments on different topics about Covid-19 pandemic on the internet and we evaluated them using Word2Vec technique. With this technique, vectors of words in a document are calculated and the semantic relationship between words is captured. The collected data include March and April data, so we compared the results of the two months. As a result of this study, many different results were found about people's views and opinions about the pandemic. The results of this study can be used in the future as automatic psychological evaluation studies with natural language processing techniques. And the trained model will be shared on internet platforms.

Keywords: Word Embedding, Word2Vec, NLP, Covid-19

Cite as:
Aksoy, N.; Tülek, Ö.; Aydın, Ö. & Özçekiç, E. (2021). Analyzing user comments on Covid-19 pandemic with Word2Vec technique. *European Journal of Educational and Social Sciences, 6 (1),* 119 – 129.

INTRODUCTION

With the introduction of the internet into our lives, the amount of data is constantly increasing. Especially when it comes to the present day, people share their opinions, ideas and evaluations about daily news on social media platforms. Although there are many data on the internet, valuable knowledge is very difficult to find. Therefore, text mining researches are very valuable. Text mining is used to analyze the resulting textual data and it has experienced advances due to web 2.0 and social networking applications (Aggarwal & Zhai, 2013).

Text mining can be used by business to examine customer and competitor data for improving competitiveness, by pharmaceutical industry for improving drug discovery within academic research, and in the field of telecommunications for information filtering and routing etc. (Dasri, Barde, Shivajirao, & Bainwad, 2017). In this article, we focus on examining how people's opinions about Covid-19 changed during March and April, based on their internet posts.

The pandemic process we are experiencing has greatly affected the lives of people both in our country and in the world. In order to prevent the spread of Covid-19 disease some measures have been taken in our country, such as closing some workplaces, imposing a travel ban, restricting public events and suspending education in schools. It has been recommended that people stay at home unless necessary, work from home if possible, and stay at least two meters away from each other when they go out. In the following process, curfews have started to be applied.

The fact that people spend more time at home has caused them to use social media more. People use social media to share information, raise concerns about the pandemic or just spend time. Covid-19 and its effects have become one of the most talked about topics on social media platforms (Wiederhold, 2020). Among the topics discussed are the fear and stress caused by Covid-19, economic concerns, the necessity of wearing a mask and calls to stay at home (Abd-Alrazaq, Alhuwail, Househ, Hamdi, & Shah, 2020).

Within the scope of this study, in order to analyze people’s feelings about this great change in their lives and Covid-19 pandemic, user comments on different topics related to the pandemic were collected from the "Ekşi Sözlük" platform and evaluated using Word2Vec technique.

MATERIALS and METHODS

Texts are not suitable data sources for machine learning algorithms. Therefore, this data need to be digitized with feature extraction methods. Main aims of these methods are to reduce data size and find valuable data for machine learning algorithms (Khalid, Khalil, & Nasreen, 2014).

There are different methods used in feature extraction methods. Some of these methods are bag of words approach and others are the semantic approach. In this title, we used countvectorizer and word2vec technique as feature extraction methods.

According to the countvectorize, the words in the text and the number of usages of the words are kept in a matrix (Tripathy, Agrawal, & Rath, 2016). In this way, the most frequent words in the document are found. In this study, we calculated numbers of words in our documents for March and April. The sample table of this is as follows;

Table 1. Number of Words in March

Word	Num. of Word in document
test / test	1544
vaka / case	1390
sokağa / to the street	1051
gün / day	1029
çıkma / going out	951
sayısı / number	872
yasağı / curfew	813

Word embedding is a basic Natural Language processing procedure that semantic and syntactic features are found from unlabeled text data (Li & Yang, 2018). Due to these methods, the extracted features can be organized in low dimensional space. There are different word embedding techniques, some of these are; LSA (Latent Semantic Analysis), Word2Vec and Glove (Naili, Habacha, & Ghezala, 2017). In this study we used Word2Vec technique and we calculated similarity of words thus semantic relationship between words were captured. The similarity table of the word "vaka" as shown follows:

Table 2. Similarity Rate of Word of "Vaka"

Word	Similarity Rate
artarak / increasingly	0,906
ölü / dead	0,900
bekliyordum / was waiting for	0,897
önceki / previous	0,895
yükseldi / increased	0,888
sayısında / in the number of	0,887
artış / increase	0,885
yükselmiş / risen	0,882
yeni / new	0,881
3000	0,879

One of the word embedding techniques is Word2Vec. This technique was created and published in 2013 by a team of researchers led by Tomas Mikolov at Google. In this technique, vectors of unique words within a corpus are calculated and the vectors are kept in a matrix. Similar words in the corpus are positioned closely in the vector space (Mikolov, Kai, Corrado, & Dean, 2013). After

calculating the word vectors, we can find similarities and differences between words and documents. Thus, we can capture the meaning of the texts without using the labeled data.

While calculating vectors of words, word2vec uses two different techniques that is known as skip-gram and CBOW.

In CBOW approach, target word in a sentence is predicted according to the predicted neighboring words (Mikolov, Le, & Sutskever, Exploiting similarities among languages for machine translation, 2013). Numbers of neighborhoods are assigned according to the window size of the target word. Unlike CBOW, the skip-gram technique uses target word to learn other words. The figure showing these techniques is as follows:

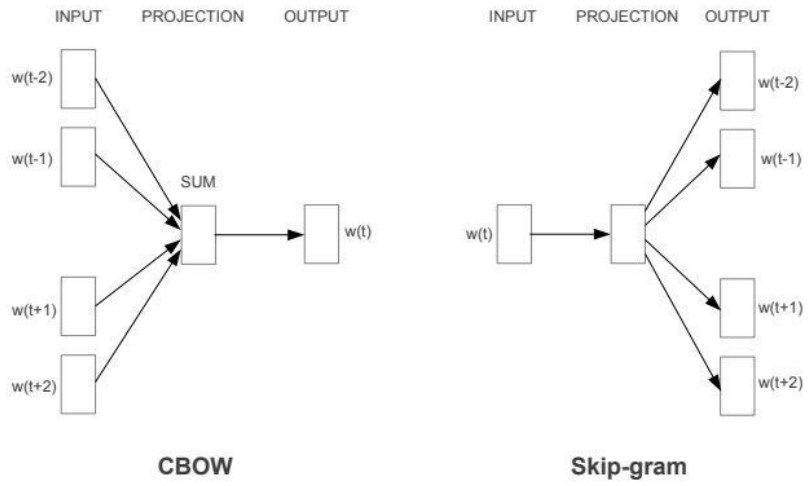

Figure 1. Skip-gram / CBOW (Mikolov, Le, & Sutskever, Exploiting similarities among languages for machine translation, 2013)

As a result, the working principles of these two approaches are different from each other. The skip-gram technique gives better results in smaller corpus and captures similarities between rare words better. On the other hand The CBOW (continues bag of words) technique, captures similarities between common words better and works faster than the skip-gram approach (Mikolov, Grave, Bojanowski, Puhrsch, & Joulin, 2018).

While training a model in big dataset, it is advantageous to use pre-trained word vectors. Because this pre-trained word vectors representation is used in new training model (Mikolov, Grave, Bojanowski, Puhrsch, & Joulin, 2018). When using this method, word similarities will be stronger and more understandable thanks to the pre-trained model.

RESULTS and DISCUSSION

In the study, we collected 12500 comments from people on "ekşi sözlük" platform about Covid-19 pandemic on several topics. Some of these topics are

- Koronavirüs
- corona-virusu-sayesinde-fark-edilen-gercekler

- sokaga-cikma-yasagi
- sars-cov-2
- covid-19

These comments belong to March and April. Thus, the data of the two months were compared with each other and different topics related to the comments of the users were captured. First of all, the pre-processing step is executed on the collected data, and in this step, stop words and punctuation marks in the data are removed. After the pre-processing step, corpus is made from each user comment. In our study, the most frequently mentioned words in March and April data were found using the CountVectorizer method.

The lists that belong to March and April are as follows:

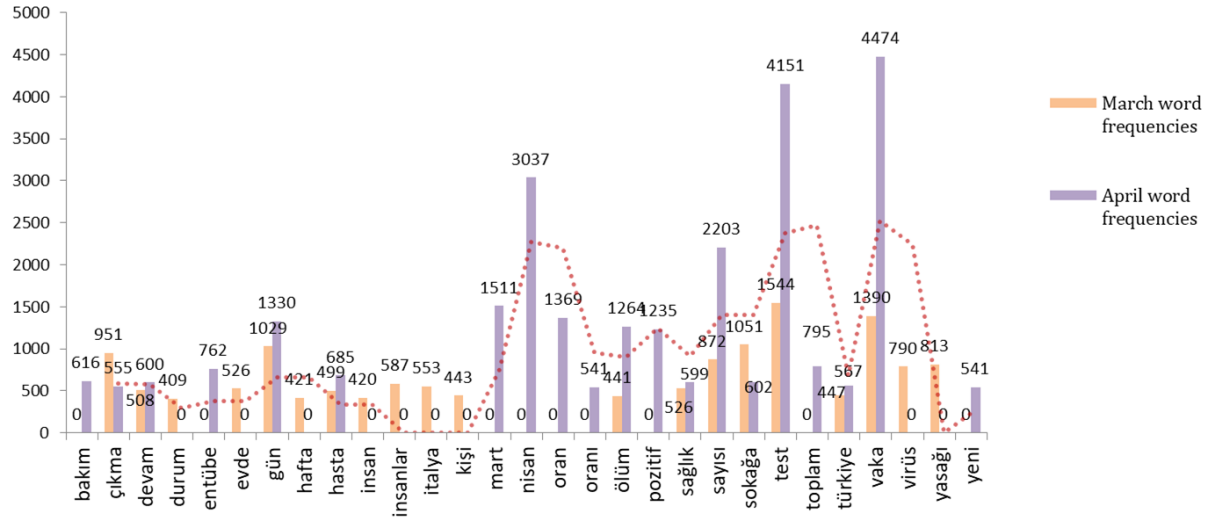

Figure 2. The 20 most frequent words in March and April datasets and their frequencies

When the results are examined, there are some similar words in two months and the number of occurrences and orders of these words are different. Although some words were found in March, these words were not found in April. The opposite was found. Some words are similar for two months. As an example, the word "vaka" is passed in two months, but the number of this word are found is different, in March this word passed 1390 times, but in April it passed 4474 times and the order of it is different.

We understood from these results that within two months, people thought different things in Covid-19 pandemic and as a result they expressed different opinions.

We used word2vec technique to understand similarity and difference between the words of each month. As a result of this, we understood what similar words or different words mean. In this study, we selected common words for two months according to the countvectorizer results. These words are “test, vaka, sokağa, çıkma, devam, ölüm".

While applying the word2vec technique, we trained models for each month and then tried the models on selected words. As an example, the word "vaka" similarity words chart is as follows:

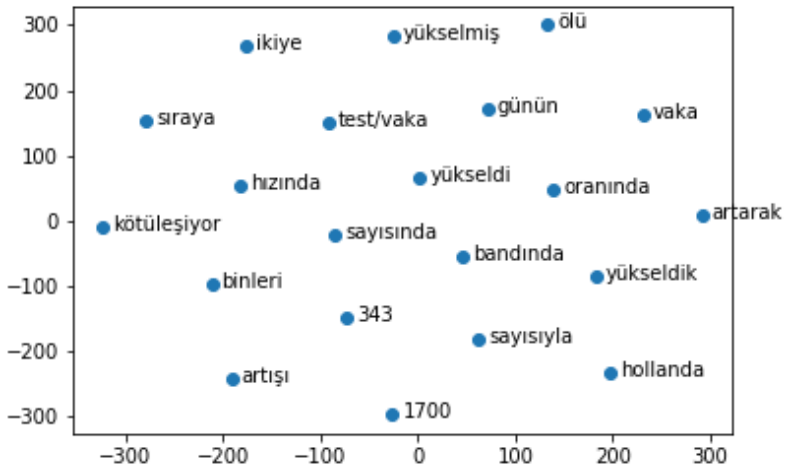

Figure 3. Words similar to the word "vaka" for March

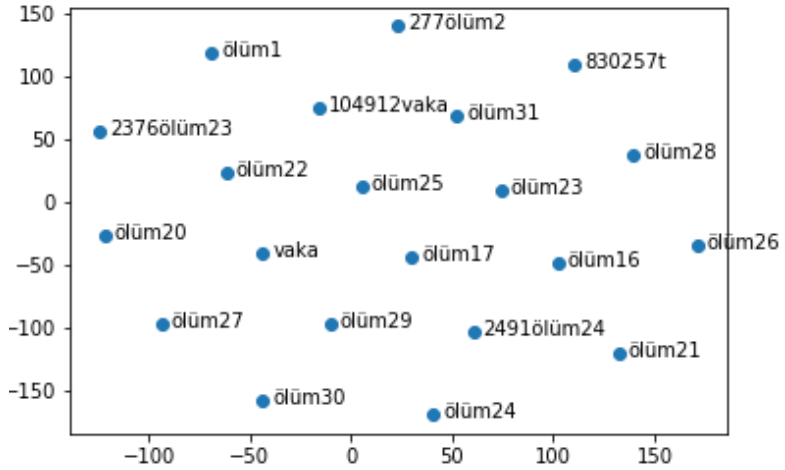

Figure 4. Words similar to the word "vaka" for April

As shown in the graphics, the similar words found are rather different. In March, this word was associated with words "binlere, bandında, yükseldik, hızında" but in April it was associated with word to "ölüm" and number of "ölüm". In the March graph, words similar to the case word appear to be close to words that emphasize the trend of the cases. In April, it is seen that the word case is directly associated with the number of deaths. When we look at the Covidien-19's in Turkey in April, it is observed that the number of cases and deaths peaked therefore it seems to be semantically associated with each of the people that said the two concepts fit to the progress in our country.

Another example is the word to “devam”. People seem to have different meanings for this word. The charts for March and April are as follows:

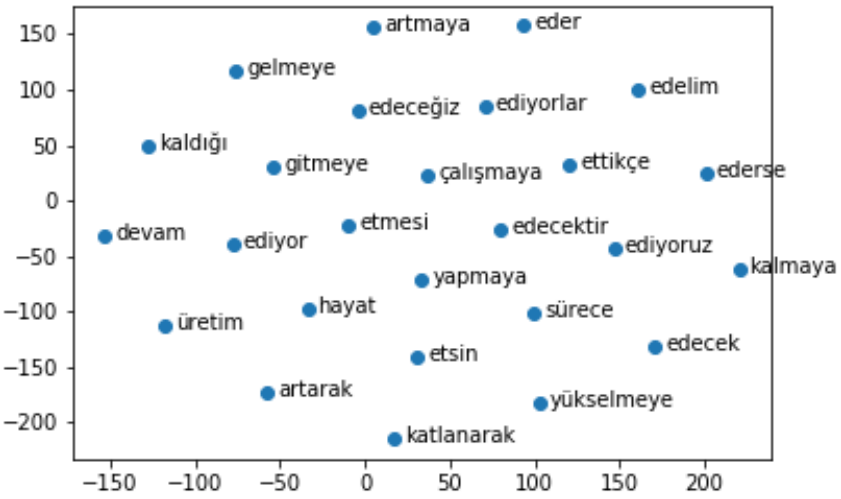

Figure 5. Words similar to the word “Devam” for March

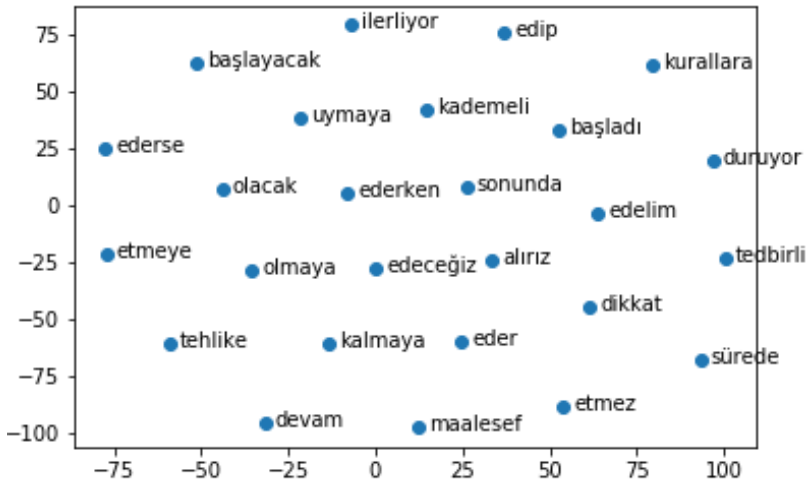

Figure 6. Words similar to the word “Devam” for April

As shown in Figure 5 and Figure 6, the word "devam" has different meanings to each month. In the graph of March "üretim, çalışmaya, hayat, devam, etsin " words are notable. In March, people used the word “devam” instead of how COVİD-19 affects their physical health or the measures they need to take to protect themselves from the disease, instead of words that correspond to daily concerns such as “üretim, çalışmaya , hayat, devam, etsin” and “artmaya, katlanarak, yükselmeye”. They are associated with words for the course of the disease.

By April, words that have semantic affinity with the word “devam” are "dikkat, edelim, tehlike, kurallara, uymaya, tedbirli". Here, a semantic affinity is observed with the words that emphasize the threat “tehlike, tedbirli, etc.” created by harm and the ways of protection from it such as “dikkat, edelim, kurallara, uymaya”.

There are similar words except for these words in two months. Some of these have the same meaning in two months, like the word to "sokağa". Similar words of this word are associated with curfew. Sample of these is shown as follows:

Similar Word in March	Similar Word in April
Çıkma / Going out	İlan / Announcement
Yasağı / Curfew	Edilmeli / Must be
İlan / Announcement	Çıkma / Going out
Getirilmeli / Should be imposed	Yasağı / Curfew
Uygulansın / Apply	Gelmezse / If not come
Ohal / State of emergency	Süreli / Periodic

Figure 7. Words similar to the word "Sokağa" in March and April

As shown in Figure 7. people asked for curfew to be protected from Covid-19 pandemic both in the month March and April. Based on the word similarity calculations, it can be said that the general opinion is that the curfew is necessary in both March and April.

Although there are some words included in the most frequently mentioned word list in March these words are not included in the most frequently mentioned word list in April. One of these words is "italya". People compared the course of Covid-19 in Turkey with İtalya and other countries in March. And also, in April this comparison continued but the names of some countries changed. The changing country names follow the spread trend of the disease as in the example of "England". The graphics of March and April are as follows:

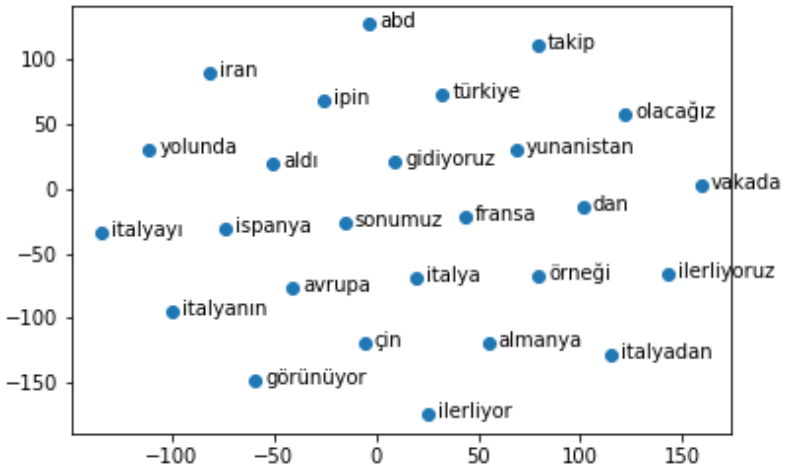

Figure 8. Words similar to the word "İtalya" for March

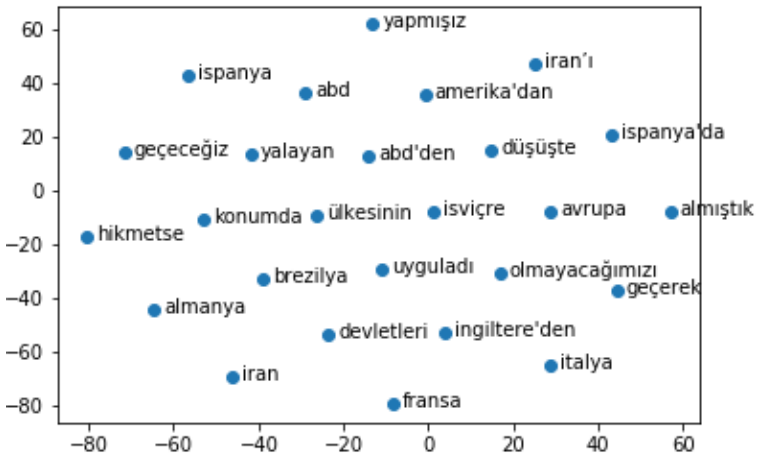

Figure 9. Words similar to the word "İtalya" for April

The words clustered around the word Italy have often been other countries of the world struggling with Covid-19. According to the course of the process, it is seen that countries such as England and Argentina are included in the chart in April. In the light of the graphs, it can be said that the authors followed the processes of other countries in this process.

Besides of this, some words that were included in the most frequently mentioned word list in April were not included in the most frequently mentioned word list in March. One of these words is "Entübe". The graphics of March and are as follows:

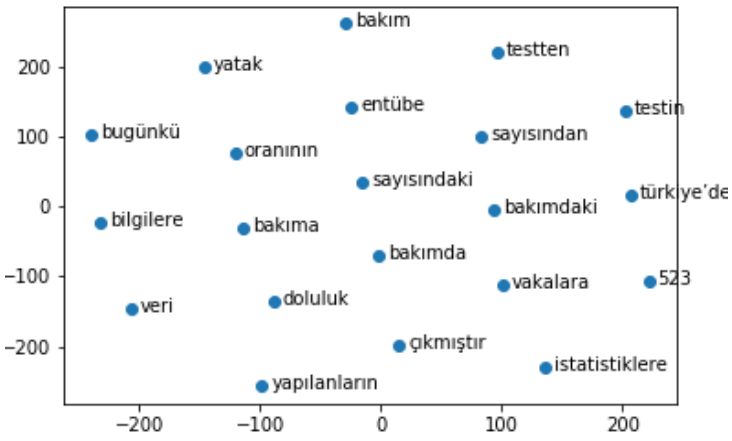

Figure 10. Words similar to the word "Entübe" in March

When Figure 10 and Figure 11 are examined, people associated words "doluluk, oran, bakım" with this word both in March and in April. The graphics of April are as follows:

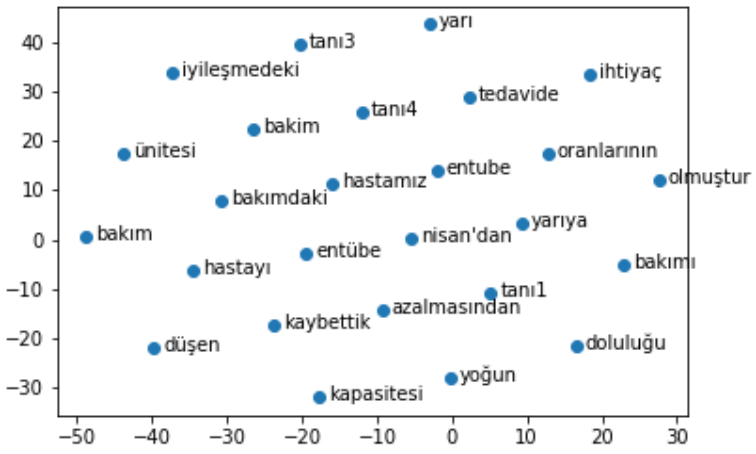

Figure 11. Words similar to the word "Entübe" in April

In addition, in this study we used pre-trained models with word2vec technique. In this technique, firstly we trained a model for words in March and secondly, a new model was trained consisting of words in April with this pre-trained model. Thus, a word gained a meaning consisting of March and April words. This new model has more meaning about Covid-19 pandemic. For example, when reviewing the word "vaka" in April, people associated it with death, but in March they gave it a more general meaning. If we use pre-trained model to train the words in April the results will be different. We used pre-trained model to train the words in April and we have achieved the following result:

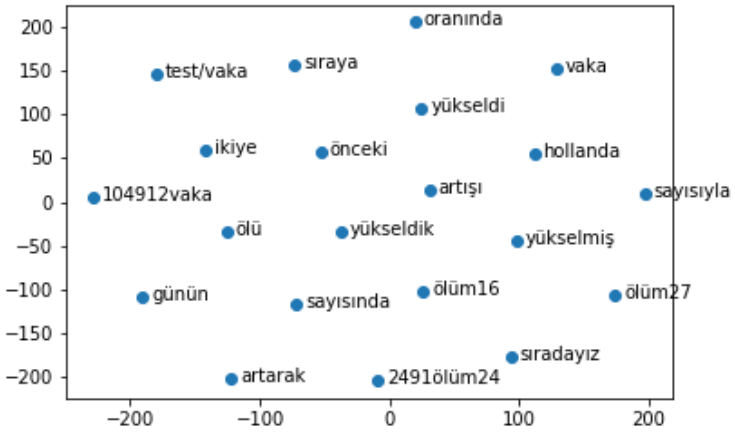

Figure 12. With Pre-Trained Model Result for "Vaka"

When reviewing the Figure 12, we have seen that the word to "vaka" included meaning of March and April words. Thus, the results are more understandable.

REFERENCES

Abd-Alrazaq, A., Alhuwail, D., Househ, M., Hamdi, M., & Shah, Z. (2020). Top concerns of tweeters during the COVID-19 pandemic: infoveillance study. *Journal of Medical Internet Research, 22*(4).

Aggarwal, C., & Zhai, C. (2013). *Mining text data*.

Dasri, Y. B., Barde, B. V., Shivajirao, N. P., & Bainwad, M. A. (2017). Text mining framework, methods and techniques. *IOSR Journal of Computer Engineering, 19*(4), 19-22.

Khalid, S., Khalil, T., & Nasreen, S. (2014). A survey of feature selection and feature extraction techniques in machine learning. *Proceedings of 2014 Science and Information Conference, SAI 2014*.

Li, Y., & Yang, T. (2018). Word embedding for understanding natural language: a survey. *Guide to big data applications* (s. 83-104). içinde Springer International Publishing.

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., & Joulin, A. (2018). Advances in pre-training distributed word representations. *LREC 2018 - 11th International Conference on Language Resources and Evaluation.* Miyazaki, Japan.

Mikolov, T., Kai, C., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings.*

Mikolov, T., Le, Q. V., & Sutskever, I. (2013). Exploiting similarities among languages for machine translation. *arXive:1309.4168*.

Naili, M., Habacha, A., & Ghezala, H. B. (2017). Comparative study of word embedding methods in topic segmentation. *Procedia Computer Science, 112,* 340-349.

Tripathy, A., Agrawal, A., & Rath, S. (2016). Classification of sentiment reviews using n-gram machine learning approach. *Expert Systems with Applications, 57*(October 2017), 117-126.

Wiederhold, B. K. (2020). Using social media to our advantage: alleviating anxiety during a pandemic. *Cyberpsychology, Behavior and Social Networking, 23*(4).