**Detection of Originality in Turkish News Websites Using Bert Model**

**Istanbul University**

**ISTEC**

Github Repository: <https://github.com/istec-iuc/TrNewsSim>

**Authors:**

Yigit Ahmed Emik

Mustafa Abduruf

Osman Cagri Simsek

**Summary**

Due to the high number of news portals producing news on a daily basis in Turkey, similar news are being published in different news portals. Therefore, the problem of re-publishing news arises for news portals with high news similarity. In a study conducted on this issue, the extent to which original content is produced in Turkish news and the dependence on news agencies were investigated using content analysis. In our study, the similarity rates of news were determined using the natural language processing model Bert. The titles of 30 randomly selected news sites were scraped and pre-processing was performed to remove html tags and unnecessary characters from the collected data. A pre-trained Turkish Bert model was used to detect the similarity of the news. The Bert model converted the sentences into vectors and assigned a numerical value. The similarity matrix was processed based on these vectors to obtain the similarity ratio between the news. By analyzing the results, three different outcomes were found: the number of similar news, the similarity ratio of news, and the arithmetic average of the news site.

1. **Introduction**

The increase in the number of internet users with the development of technology has played an important role in the evolution of traditional journalism, and with the spread of internet journalism, the number of news portals has increased day by day. Unlike traditional journalism, news writers in internet journalism sometimes publish news independently, making it difficult to distinguish original news portals from others.

There are some manual methods to analyze the similarity of news. Content analysis method means the in-depth examination and organization of independent qualitative and quantitative studies conducted on a certain subject or field (Ültay, E., Akyurt, H. & Ültay, N. 2021). Content analysis is a comprehensive method, but it is a costly manual method in terms of time and labor.

With the development of technology today, technologies that facilitate our work have been produced in many areas of our lives. Instead of digital communication in human-computer interaction, a concept called natural language processing has come into play.

Natural language processing is a subcategory of artificial intelligence and linguistics known as NLP (Natural Language Processing). It is a field of research that studies the processing and use of natural languages ​​such as Turkish, English, German, and French (Wikipedia). With natural language processing, many tasks such as text classification, content labeling, sentiment analysis, and text similarity can be performed. There are models of natural language processing such as Transformers, Natural Language Toolkit (NLTK), Spacy, and BERT for analyzing texts. The BERT model has been quite successful in text analysis compared to other models because it has a bidirectional evaluation feature.

In our study, the headlines of prominent news on the homepages of 30 news portals we selected were pulled using the Scrapy library of the Python programming language. The collected data is daily and has been cleansed of html tags and meaningless symbols that occurred during the pulling process by going through a preprocessing process. The prominent feature of this study is that the Bert model, which is a natural language processing tool, is used to calculate the similarity rates of the obtained data set. The aim is to convert preprocessed texts into vectors using the Bert model and assign them a numerical value. After each text has taken its unique numerical value, a comparison was made using the cosine similarity formula. The comparison was made with a matrix formula, and the similarity score of each news site was matched to each other. The obtained data was analyzed, and the similarity rate, ratio of news, and arithmetic mean of news sites were found.

The aim of our research is to determine the similarity rates of Turkish news portals using the Bert model.

1. **Similar Studies**

There are various studies conducted on Turkish news portals or news portals with related natural language processing and content analysis methods. In our literature review, we found studies conducted on Turkish news portals or news portals using natural language processing and machine learning with content analysis method.

**2.1 Studies conducted with natural language processing method**

TORUN and Burak INNER [1] have identified similar news by summarizing Turkish news. To calculate the similarity, the summarization method was used first to summarize the news texts on the news pages. In this study, 12,000 news texts collected from different sources within a 4-month period were summarized by selecting sentences with the summarization method, scoring the sentences based on their word frequencies, and summarizing the sentences by sorting them according to this scoring. Similarity detection was made using the word frequency in the summarized news. As a result of this summarization, they have identified that there are 1,500 similar news out of 12,000 news.

Anıl KARADAĞ and Hidayet TAKÇI [2] focused on similar news detection with text mining. The method of this study was to assign a list of keywords (or tags) to each news, and then the similarity between the news was evaluated by comparing the tag lists of the news, and a system that classifies similar news for a dynamic content news site was developed.

Mehmet BOZUYLA [3] compared traditional methods such as machine learning algorithms Naive Bayes and Random Forest with the BertTurk model proposed by the author to detect fake news in Turkish. As a result of the study, the BertTurk model achieved 99.90% accuracy in defining fake news.

[4] They presented a system (BERTEmotion) for propaganda detection in news. They used natural language processing techniques such as BERT, BilSTM, and Capsule to classify sentences. As a result of the study, the BertEomtion model developed by the authors achieved a higher success rate than other models in propaganda detection with a F1 score of 0.6338.

ABDULKADİR KARACA [5] has developed an automatic title generation application for Turkish news texts. Suitable news texts were extracted from the SuDer news corpus as the dataset and were pre-processed. The keras library was used for numerical representation of verbal data. The transformer architecture commonly used for model training and the abstract summarization method were used. After the 20th and 25th epochs of model training, they succeeded in producing highly accurate titles for news texts by reaching approximately 75% and 85% accuracy respectively.

[6] The authors applied sentiment analysis to Turkish political news. Four supervised machine learning algorithms, Naive Bayes, Maximum Entropy, SVM, and character-based N-Gram Language Model, were compared. As a result of the study, Maximum Entropy and N-gram achieved higher accuracy than other models with 65% to 77% accuracy.

Yavuz Kömeçoğlu, Başak Kömeçoğlu, and Burcu Yılmaz [7] prepared a system that clusters real-time news with similar content in their study of detecting news with similar content in the Turkish news flow. They used Apache Solr database and the MoreLikeThis search component to capture the texts of the news from their sources. They achieved an F-score of over 90% by using the word embedding method for similarity clustering.

**2.2 Content Analysis and Studies Using Different Methods**

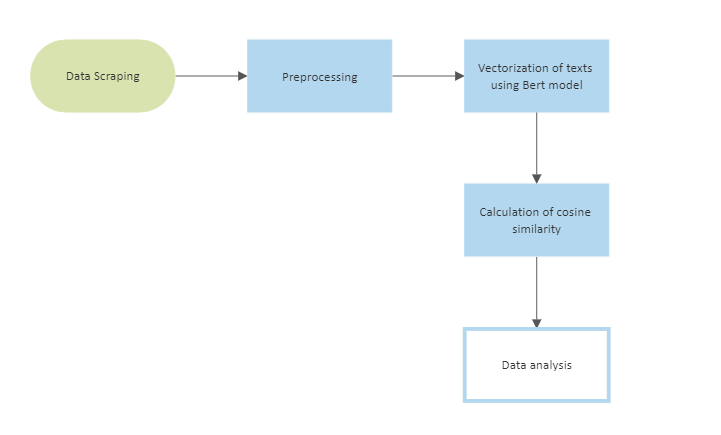
In his study, Halil İbrahim Gürcan [8] analyzed the similarity rates of content in Turkish news published on Turkish news websites. He tried to find similarities based on the prominent news stories of 20 Turkish news portals randomly selected. The content analysis method was used to analyze the news in the study. As a result of the study, it was found that 13 out of 20 sites (%65) used the same or similar headlines. It was also determined that 7 of the 20 sites (%35) used original expressions in their headlines even though they covered the same topic.

Duygu Temel [9] determined the extent to which Turkish news websites are dependent on news agencies in their news production using the content analysis method and discussed the reasons.

In the study by Hüseyin Pola and Ali Ünişen [10], 885 news stories related to teachers were examined from four Turkish news sites in December 2014 using the content analysis method. It was found that 57.06% of the news stories were negative, 28.59% were positive, and 14.35% were neutral.

1. **Method**

This section includes the methods and techniques used in our study. The data collection and preprocessing processes are explained under the heading 3.1. The process of detecting the similarity of texts using the natural language processing technique, BERT, is found in 3.2. Finally, in 3.3, the output data was saved in the database and analyzed.



**Picture-1.** Flowchart of Methods.

**3.1 Data scraping**

Large amounts of data are stored on websites, and it can be difficult to collect this data without any tools. Therefore, the data scraping method comes into play to obtain data faster and more easily. Data scraping is also known as web data scraping, and it involves collecting general data on the web automatically.

There are many libraries that we can use for data scraping, some of which are Python libraries such as "LXML", "Selenium", "Beautiful soup4", and "Scrapy". For our work, we chose the Scrapy library because it has a simpler structure and can quickly extract large amounts of data.

Scrapy is a web crawling library written in Python. This library can collect all structured information on a website. Using this library, we extracted all the news headlines on the homepage of a news website in Python (the codes are given in the reference section), and saved this data in a JSON file in the format of the news website name and news headline.

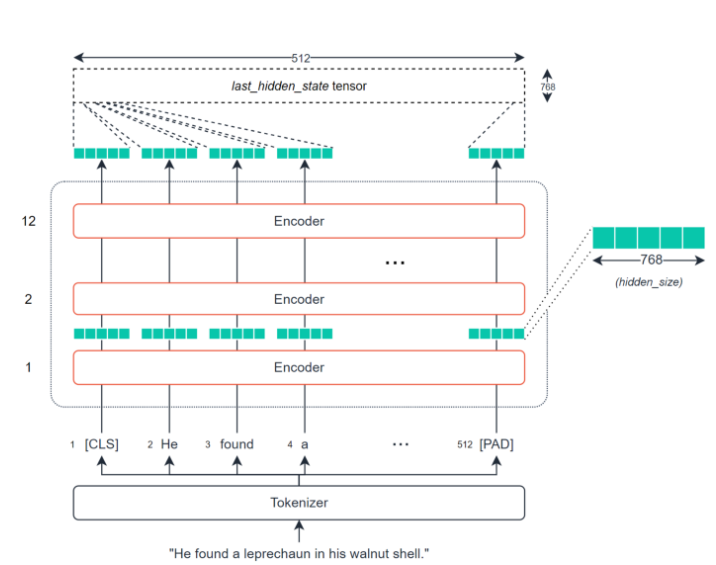
Sometimes, HTML tags can also be extracted with the data. Therefore, all the saved data was cleaned by going through a preprocessing stage to remove characters consisting of symbols, HTML tags, and so on. Then, this data was collected in a file (the files are given in the document section).

**3.2 Bert Model**

Natural Language Processing (NLP) is a subcategory of artificial intelligence and linguistics that studies the processing and use of natural languages such as Turkish, English, German, and French (Wikipedia). NLP aims to analyze and understand the structured nature of natural languages or to reproduce them. The advantages of this analysis for humans can be summarized by automatic translation of written documents, question-answering machines, automatic speech and command understanding, speech synthesis, speech generation, automatic text summarization, and information retrieval. Among the techniques of natural language processing, we chose to use Bert in our work because it has a higher accuracy rate compared to other models.

BERT is an open-source natural language processing model proposed by researchers at Google Research in 2018. BERT also uses many previous NLP algorithms and architectures such as semi-supervised training, OpenAI transformers, ELMo embeddings, ULMFit, and Transformers. Unlike other models, it evaluates the sentence both from left to right and from right to left. Thus, it aims to extract the meaning and relationships between words better and receives corresponding results.

The logic of the Bert model we used for text similarity is as follows. First, a text is converted into a vector containing numerical values. The text is tokenized, and it changes to a vector containing 768 values per word. These 768 words represent numerical values and can also be used as word embeddings.



**Figure-1 Bert Model**

Since each word takes a separate vector value (the output of each encoder), we can call these numerical values 768-dimensional tensors. We can take these tensors and convert them into an input sequence to create semantic representations. Then, we can use a similarity matrix to calculate the similarity score between sequences.

The last\_hidden\_state tensor has a value of 512\*768 and is used to measure the similarity of a vector. A pooling process is used to create a single "text vector" of 768 values by taking the average of each token's 768 values and compressing them.

This single tensor is suitable for calculating the similarity between texts. However, natural language processing comes into play to perform the similarity calculation using a similarity matrix.

To convert human language into a format that computers can read, dense vectors are generally used. Artificial neural networks typically produce dense vectors, allowing us to transform words and sentences into high-dimensional vectors where each vector's geometric position can carry meaning.

The similarity score is then determined using the cosine formula with these vectors carrying meaning.

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

**Figure-2** Cosine similarity formula

Cosine similarity considers vector direction regardless of vector magnitude.

The first thing we need to know in this formula is that the numerator is actually a dot product that takes into account both magnitude and direction.

In the denominator, we have two vertical bars that represent 'length'. So, we have the product of the length of u and the length of v. Length takes magnitude into account.

When we take a function that considers both magnitude and direction and divide it by a function that only considers magnitude, these two magnitudes cancel out, leaving us with a function that considers direction independent of magnitude.

An example of this is given below:

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

**Figure-3** Cosine similarity of a and b is close to 1

Using Python's "Sklearn" library, cosine similarity has been implemented in our text (text previously converted to vector).

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

**Figure-4** Cosine similarity operation with the Sklearn library

In Figure 4, the cosine similarity formula was used to calculate the similarity score between two texts converted into vectors. This formula was used to determine the similarity score of the 30 news sites we collected. To determine the similarity score, we needed to match each news site with each other. We had to match all the news of site A with all the news of site B to find the number of similar news between A and B, and choose the one with the highest similarity score.

To do this, we used a matrix formula to match each news article in site A with all the news articles in site B. For example, the first news article of site A was matched with all the news articles of site B, and the one with the highest similarity score was selected. This process continued until all the news articles of site A were matched with all the news articles of site B. An example of this is shown in the table below.

|  |  |  |
| --- | --- | --- |
| **sendika** | **sondakika** | **Score** |
| İşverenin azmettirmesiyle dövülen 60 yaşındaki Altın Erdoğan’ın davası 22 Kasım’a ertelendi: Sağlık raporu istendi | Takım değiştiren Ümit Karan'ın yeni adresi belli oldu | 0.1519 |
| İşverenin azmettirmesiyle dövülen 60 yaşındaki Altın Erdoğan’ın davası 22 Kasım’a ertelendi: Sağlık raporu istendi | Jahrein'in Mahmut Ustaosmanoğlu'nun vefatına ilişkin attığı tweet gündem oldu: Azrail'e iki tokat atıp yollayamamış mı? | 0.3212 |
| İşverenin azmettirmesiyle dövülen 60 yaşındaki Altın Erdoğan’ın davası 22 Kasım’a ertelendi: Sağlık raporu istendi | Konut kredilerine ilişkin yeni kararlar açıklandı! Evin fiyatına göre bankadan alınacak tutar değişecek | 0.0335 |

**Table-1** Matching samples between “sendika” and “sondakika” websites.

The first news of the union news site has been matched with the news of the breaking news site in Table-1. The news with the highest similarity score will be selected as a result of the matching.

As a result, the number of similar news between site A and site B has been obtained. Using the same logic, the similarity rate between site A and all other sites, then site B, C, D, and so on, has been determined by first converting the headlines of all 30 news sites into vectors and then using the cosine similarity formula to determine the similarity rate.

**Results**

After passing through all the stages in the method section, the data was analyzed, and three different results were obtained: 5.1 Number of similar news, 5.2 Similarity rate of news, 5.3 Arithmetic mean of similar news.

**5.1 Number of similar news**

As mentioned in the similarity matrix section, we recorded the highest matching score of each news. To distinguish similar news from each other, we set a criterion and defined the news pair with a score of over 0.5 as similar news. Thus, we filtered the news pairs that matched between site A and site B with a score of over 0.5 using SQL commands and recorded them in an Excel table.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **A/B**  **Similarity** | **Altinok** | **t24** | **sendika** | **OdaTV** | **AnadoluAjansi** | **Dikgazete** | **F5** | **Gerçekgündem** |
| **bianet** | 2 | 2 | 3 | 1 | 6 | 1 | 0 | 8 |
| **dokuzsutun** | 2 | 1 | 2 | 1 | 4 | 1 | 3 | 2 |
| **haksozhaber** | 1 | 1 | 2 | 1 | 2 | 6 | 1 | 3 |
| **IHA** | 1 | 2 | 0 | 0 | 5 | 1 | 0 | 0 |
| **Ilerigazetesi** | 0 | 0 | 1 | 0 | 0 | 3 | 1 | 0 |
| **Timeturk** | 2 | 4 | 0 | 5 | 10 | 7 | 2 | 10 |
| **turkuazgazetesi** | 0 | 3 | 1 | 2 | 5 | 1 | 1 | 8 |
| **ahaber** | 1 | 3 | 1 | 6 | 6 | 1 | 4 | 4 |
| **ensonhaber** | 2 | 4 | 0 | 3 | 2 | 3 | 1 | 4 |
| **F5** | 0 | 1 | 1 | 1 | 0 | 1 |  | 3 |
| **TRThaber** | 2 | 7 | 1 | 4 | 8 | 7 | 1 | 7 |
| **Mynet** | 0 | 1 | 1 | 5 | 2 | 4 | 3 | 4 |
| **Gerçekgündem** | 1 | 10 | 1 | 6 | 7 | 0 | 3 | nan |

**Table-2** Number of similar news

Table-2 is a matrix with a size of 30x30, where the rows represent the A news site and the columns represent the B news site that it is compared to. Each A news site is matched with the other 29 news sites one by one, and the number of similar news stories with a similarity score of above 0.5 is recorded in the table.

**Table-2** Full version: <https://github.com/istec-iuc/TrNewsSim/blob/main/Analyzed%20data/Table_news%20number%20of%20similarities.csv>

**5.2 Similarity rate between news sites**

The difference between the similarity rate and the number of similar stories is that it is calculated based on the number of news stories in each site. The number of news stories in the two matched news sites is not the same, so the similarity rate is determined using the number of shared news stories and the total number of news stories in each site.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **A/B**  **Similarity** | **Altinok** | **t24** | **sendika** | **OdaTV** | **AnadoluAjansi** | **Dikgazete** | **F5** | **Gerçekgündem** |
| **bianet** | 3.57% | 3.57% | 5.36% | 1.79% | 10.71% | 1.79% | 0.00% | 14.29% |
| **dokuzsutun** | 4.76% | 2.38% | 4.76% | 2.38% | 9.52% | 2.38% | 7.14% | 4.76% |
| **haksozhaber** | 1.35% | 1.35% | 2.70% | 1.35% | 2.70% | 8.11% | 1.35% | 4.05% |
| **IHA** | 3.45% | 6.90% | 0.00% | 0.00% | 17.24% | 3.45% | 0.00% | 0.00% |
| **Ilerigazetesi** | 0.00% | 0.00% | 2.70% | 0.00% | 0.00% | 8.11% | 2.70% | 0.00% |
| **Timeturk** | 2.27% | 4.55% | 0.00% | 5.68% | 11.36% | 7.95% | 2.27% | 11.36% |
| **turkuazgazetesi** | 0.00% | 3.95% | 1.32% | 2.63% | 6.58% | 1.32% | 1.32% | 10.53% |
| **ahaber** | 1.59% | 4.76% | 1.59% | 9.52% | 9.52% | 1.59% | 6.35% | 6.35% |
| **ensonhaber** | 6.06% | 12.12% | 0.00% | 9.09% | 6.06% | 9.09% | 3.03% | 12.12% |
| **F5** | 0.00% | 11.11% | 11.11% | 11.11% | 0.00% | 11.11% | nan | 33.33% |
| **TRThaber** | 4.65% | 16.28% | 2.33% | 9.30% | 18.60% | 16.28% | 2.33% | 16.28% |
| **Mynet** | 0.00% | 1.82% | 1.82% | 9.09% | 3.64% | 7.27% | 5.45% | 7.27% |
| **Gerçekgündem** | 1.47% | 14.71% | 1.47% | 8.82% | 10.29% | 0.00% | 4.41% | nan |

**Table-3** Similarity Ratio of News

A similarity ratio with B = common number of news/total number of news of A\*100

Table-3 shows the similarity ratio of news websites in a 30\*30 table. The calculation method is based on dividing the common news count of two websites by the total news count of the matching website. For example, if website A has 60 news and website B has 30 news, and they have 5 common news, the similarity ratio of website A is calculated by finding what percentage of 60 is 5. The results are recorded in the table.

The full version of **Table-3** can be found at:

<https://github.com/istec-iuc/TrNewsSim/blob/main/Analyzed%20data/Table_ratio%20of%20news.csv>

To obtain a general similarity ratio, all news websites are matched with each other, and an arithmetic calculation is performed by adding the similarity ratio of the website with all other websites. Then, the top 10 most similar and top 10 least similar news websites are determined.

**5.3 Arithmetic average of similar news**

There is a similarity ratio between news sites and all other news sites, but this is the result of matching each news site with the other news sites one by one. To determine a general similarity ratio, the similarity ratio of the news site with all other news is calculated by an arithmetic calculation. Then the top 10 news sites with the highest similarity and the bottom 10 news sites with the lowest similarity are obtained.

|  |  |  |
| --- | --- | --- |
| **Website** | **Similarity** | **Number of News** |
| **F5** | 15,55% | 9 |
| **HaberAktuel** | 14,10% | 35 |
| **Altinok** | 9,82% | 18 |
| **TRThaber** | 9,69% | 43 |
| **ensonhaber** | 8,48% | 33 |
| **AnadoluAjansi** | 8,44% | 47 |
| **sondakika** | 8,12% | 78 |
| **Hakkariobjektifhaber** | 7,66% | 57 |
| **Ahaber** | 7,25% | 63 |

**Table-4:** Top 10 Most Similar Websites

In Table-4, an arithmetic average was calculated for each news site by adding the similarity ratios of each news site in Table-3. To find the arithmetic average of Site F5, the ratios resulting from matching with the other 29 news sites were added and divided by the total number of news sites (30). The same method was applied to all other news sites. As a result, the top 10 most similar news sites were filtered and recorded in the table.

|  |  |  |
| --- | --- | --- |
| **Website** | **Similarity** | **Number of News** |
| **Denizgazete** | 0,93% | 25 |
| **buyuklimanpostasi** | 2,18% | 26 |
| **haksozhaber** | 2,25% | 74 |
| **sendika** | 2,31% | 39 |
| **Ilerigazetesi** | 2,52% | 37 |
| **Dikgazete** | 2,79% | 92 |
| **pressturk** | 3,33% | 28 |
| **bianet** | 3,33% | 56 |
| **NTV** | 3,55% | 107 |
| **dokuzsutun** | 4,68% | 42 |

**Table-5**: Top 10 Less Similar Websites

**6. Discussion**

Table 5 lists the 10 least similar news sites, determined by calculating their arithmetic means using the same method as Table 4. In this study, we aimed to determine the similarity between Turkish news sites by focusing solely on news headlines. All evaluations were made based on the headlines. Our research focused on the extent to which news sites produced unique news headlines. The data used for analysis was a single-day data pulled on July 7th, 2022.

According to our study, F5 news site showed the highest similarity with other news sites at 15.55%. Denizgazete, on the other hand, had the highest rate of unique headline production at 0.93%. However, it should be noted that the number of news sites included in this study directly affected the similarity rate. During data collection, it was observed that some news sites had significantly different numbers of news articles. For instance, F5 had a total of 9 news articles, while NTV had 107. The presence of a shared news article on F5's news site would result in a similarity rate of 11.11%.

Among the 30 sites used for comparison, the news agencies iha haber ajansi and Anadolu ajansi were included. Despite being a news agency, Anadolu ajansi showed a low similarity rate of 8.44% compared to other news sites.

**References:**

[1] Torun, H., & Inner, A. B. (2018, May 1). Detecting similar news by summarizing Turkish news. <https://doi.org/10.1109/SIU.2018.8404826>

[2] Anıl,KARADAĞ.,& Hidayet, TAKÇI.Metin Madenciliği ile Benzer Haber Tespiti.

<https://www.researchgate.net/publication/228415390_Metin_Madenciligi_ile_Benzer_Haber_Tespiti>

[3] BOZUYLA, M. (2022). Çift Yönlü Transformatör Kodlayıcı Temsilleriyle Gelişmiş Türkçe Sahte Haber Tahmini. Konya Journal of Engineering Sciences, 750–761. <https://doi.org/10.36306/konjes.995060>

[4] George-Alexandru Vlad, Mircea-Adrian Tanase, Cristian Onose, and Dumitru-Clementin Cercel. 2019. Sentence-Level Propaganda Detection in News Articles with Transfer Learning and BERT-BiLSTM-Capsule Model. In Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda, pages 148–154, Hong Kong, China. Association for Computational Linguistics.

[5] Karaca, A., & Aydın, Ö. (2021). Derin öğrenme ile Türkçe haber metinlerine başlık üretme. *Dspace.trakya.edu.tr*. https://dspace.trakya.edu.tr/xmlui/handle/trakya/5304‌

[6] Kaya, M., Fidan, G., & Toroslu, I. H. (2012). Sentiment Analysis of Turkish Political News. Içinde 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology. 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT). IEEE. <https://doi.org/10.1109/wi-iat.2012.115>

[7]Kömeçoğlu, Y., Kömeçoğlu, B. B., & Yılmaz, B. (2021). Real-Time News Grouping: Detecting the Same-Content News on Turkish News Stream. Advances in Intelligent Systems and Computing, 14–25. <https://doi.org/10.1007/978-3-030-66501-2_2>

[8] [*https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.600.2959&rep=rep1&type=pdf*](https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.600.2959&rep=rep1&type=pdf)

[9] TEMEL, D. (2020). Çevrimiçi Haberin Kaynağı: İnternet Gazetelerinde Kopyala-Yapıştır Habercilik. Galatasaray Üniversitesi İleti-ş-Im Dergisi. <https://doi.org/10.16878/gsuilet.661144>

[10] Polat, H., & Ünişen, A. (2016). An analysis of teacher news in Turkish printed media within the context of teachers’ image. Cogent Education, 3(1). <https://doi.org/10.1080/2331186x.2016.1179614>

Briggs, J. (2021, September 2). BERT For Measuring Text Similarity. Medium. <https://towardsdatascience.com/bert-for-measuring-text-similarity-eec91c6bf9e1>

Turkish  BERT model : <https://huggingface.co/emrecan/bert-base-turkish-cased-mean-nli-stsb-tr>