Available online at www.alphanumericjournal.com

alphanumeric journal

The Journal of Operations Research, Statistics, Econometrics and
Management Information Systems

Volume 10, Issue 2, 2022

Received: July 27, 2022
Accepted: October 20, 2022
Published Online: December 31, 2022

AJ ID: 2022.10.01.MIS.02
DOI: 10.17093/alphanumeric.1149753
Research Article

Classification of News Texts by Categories Using Machine Learning Methods

Mehmet Kayakuş, Ph.D. *

Assist. Prof., Department of Business Information Systems, Manavgat Faculty of Social and Human Sciences, Akdeniz University, Antalya, Turkiye, mehmetkayakus@akdeniz.edu.tr

Fatma Yiğit Açıkgöz

Lect., Social Sciences Vocational School, Akdeniz University, Antalya, Turkiye, fatmayigit@akdeniz.edu.tr

* Manavgat Sosyal Ve Beşeri Bilimler Fakültesi 07600 Manavgat/Antalya, Türkiye

ABSTRACT

In parallel with the advances in technology, digital journalism is preferred more than printed journalism day by day. Due to the fast and up-to-date sense of journalism provided by digital journalism and its ubiquitous accessibility features, it is read more by users. In addition to these advantages provided by digital journalism, it also has some difficulties compared to printed journalism. The stage of preparation and delivery of the news to the user requires more technological knowledge and equipment compared to printed journalism. The processes of title selection, text creation, photo selection and determination of the appropriate news category in the preparation phase of the news are designed to be both faster and user-friendly compared to printed publishing. The news created to be presented to the target audience may belong to one or more of different categories such as economy, politics, sports, technology, and health. The inclusion of the news in the appropriate category provides convenience in terms of reaching the right audience and archiving the news correctly. In this study, news texts were classified according to their categories based on the machine learning methods. In the study, news of five newspapers in three different categories were used. Bayesian classifier and decision tree methods were used to classify the news in the dataset including a total of 10.500 news. In the results of the study, it was observed that the Bayesian classifier classified the news more successfully according to their categories.

Keywords: News, Category, Classification, Machine Learning

1. Introduction

The changes in communication technologies along with the digital age have radically changed journalism practices, both for readers and publishers, as in every field. Online technologies that allow direct communication to have provided alternative communication opportunities to traditional media (Bardoel, 1996). People can access news whenever and wherever they want through web applications, e-mail systems, smart phones and social media channels. On the other hand, contrary to popular belief, the origins of digital journalism are not based on the spread of the Internet or the computer, but on Teletext broadcasting, a videotext standard used to display text and basic graphics on appropriately equipped televisions invented in the 1970s. In the popular application patented by the BBC, news texts were broadcast daily. In the 1990s, digital publishing opportunities expanded rapidly along with the development of Web 2.0 software. The Internet was used only for military purposes by the US Department of Defense in the 1960s. However, as of 1993, the internet was made accessible to individual users and became more widespread with the introduction of computers into homes. Online journalism also quickly became popular in parallel with the spread of the internet, as in many areas. Internet is the most important factor in the transformation of journalism. Nevertheless, internet journalism has gone through different processes until today. Internet journalism, that started with the process of copying the contents of newspapers published in print in the first years of its emergence, continued with the sharing of original contents for the web (Aydoğan, 2013).

When the first examples of internet journalism in the world and in Turkey are examined, the transfer of the printed publications of New York Times, The Washington Times and the International Herald Tribune and the Daily Mirror in Europe to the Internet in 1995 is considered as the beginning of internet journalism. In 1995, the transfer of the content of Aktüel Magazine to the Internet in Turkey started the Internet journalism in our country. In 1996, important newspapers of the Turkish press such as Milliyet, Türkiye, Hürriyet and Sabah started to broadcast on the internet (Çakır, 2007). As of the 2000s, independent internet news sites were established under the leadership of important journalists with the effect of both the rapid change in the field of communication and the crisis in the media sector. Nowadays, Internet journalism has reached very important points both in Turkey and in the world. The advantages of internet journalism compared to printed newspapers play a significant role in the prevalence of internet journalism. The factors such as integrated processes into web pages, news sites, blogs and increased interest in social media have made digital journalism more preferable than printed journalism.

Internet news, which is made available to the public regardless of time and place, has brought along the prominence of different factors for publishing, and moreover, the changes in journalism practices. In traditional journalism practices, the behaviours of the target audience with one-way communication cannot be evaluated. Considering the Internet journalism practices, instantaneity, interactive, convenience, globality, rapid sharing, and participation come to the forefront. The differences between the two broadcasting have also radically changed the news production and distribution practices.

When digital journalism is evaluated in general, it offers many advantages to the reader and imposes new responsibilities on the reporters. The issues such as the delivery of the news to the reader, continuous updating of news, the production of interesting contents, and harmonious presentation of audio, video and content, and also the need for technical equipment are some of these responsibilities.

In the field of digital journalism where publishers compete with each other to present the best to the target audience, another issue that journalists should focus on is that the news is divided into categories and presented to the reader. Presenting the news under the correct category in accordance with its content provides convenience in issues such as reaching the right audience and archiving the news correctly. From this point of view, in this study, news texts were classified according to their categories based on the machine learning methods.

Although there are many studies on text classification in the literature, most of the studies focused on English texts, and there are few studies in Turkish in the literature.

Amasyalı and Yıldırım used the Naive Bayes, Vector Quantization and Multilayer Classifier to classify the news texts in five categories collected from the web pages of the newspapers and achieved a success by 76% (Amasyalı and Yıldırım, 2004).

Amasyalı et al. determined that which of 18 different authors with predetermined authorship characteristics an anonymous document belonged to by using the Naive Bayes, Support Vector Machine, C4.5 and Random Forest algorithms (Amasyalı et al. 2006).

Amasyalı and Beken used a hypothesis that "the semantic similarity of two words to each other is directly proportional to the number of documents in which the words are mentioned together" in order to classify Turkish news texts, and they achieved more successful results compared to traditional methods (Amasyalı and Beken, 2009).

Aşlıyan and Günel created two different collections consisting of Turkish documents in five categories to compare the performances of the Nearest Neighbor and k-Nearest Neighbor methods and determined that the highest correct classification rate was 88.4% with the Nearest Neighbor method (Aşlıyan and Günel, 2010).

In their study, Doğan and Diri determined the type of the document and the gender of the author of the document as well as identifying the author of the document. To this end, the feature vectors of different sizes were created by extracting the 2, 3 and 4 grams of the Turkish language (Doğan and Diri, 2010).

In their study, Toraman et al. aimed to provide a high-accuracy classifier with automatic text categorization to be used in Turkish news portals. The C4.5, KNN, Naive Bayes and SVM methods were applied to two Turkish test datasets with different features created using Bilkent News Portal, and the results were discussed. In the study in which the results of four different methods were compared, it was also recommended to evaluate other root detection algorithms in the classification of news texts (Toraman, 2011).

In the study of Tüfekçi et al. in which web-based news texts were classified using Turkish grammar features, the relationship between the size of the feature vector used in the classifier and the success of the classifier was examined, and a method in which the success value did not decrease despite the size reduction was proposed. In the study, the Naive Bayes, SVM, C4.5 and Random Forest classification methods were analyzed and it was stated that the highest success rate in the use of reduced feature vectors was achieved with the Naive Bayes algorithm (Tüfekçi et al., 2012).

Levent and Diri solved the problem of identifying the authors of Turkish texts with Artificial Neural Networks and obtained results close to the previous algorithms used for author identification (Levent and Diri, 2014).

In the study of Başkaya and Aydın, a total of 80 news texts were collected by collecting 20 news from different news sites belonging to 4 different categories. While 60 of the collected news were used for training, 20 of them were used for testing. They successfully classified all the news they collected (Başkaya and Aydın, 2017).

Acı and Çırak classified the Turkish news texts using Convolutional Neural Networks and Word2Vec. The text classification was made on the Turkish Text Classification 3600 data set used in the study, and it was compared with the previous study results of the authors. Accordingly, it was indicated that the Word2Vec method provided higher performance compared to classical statistical and machine learning-based classification algorithms (Acı and Çırak, 2019).

In a study conducted by Usmani and Shamsi, a news headline classification algorithm was developed. In this algorithm, they achieved a success of 88% with the simple natural language processing techniques (Usmani and Shamsi, 2020).

In their study, Uslu and Akyol performed the classification of Turkish news texts using machine learning methods. A dataset containing many news texts and news categories was used as news content. In the study, the results of the analysis performed according to the support vector classifier, random forest and Naive Bayes Classifier were compared, and it was concluded that the method with the most successful performance was the Naive Bayes Classifier with an accuracy of 91% (Uslu and Akyol, 2021).

2. Material and Method

The study consists of five main stages. The first stage is the creation of the dataset. A dataset created from 10.500 news data collected from news sites was used in the study. The second stage is the removal of noisy data in the dataset and the conversion of the data into the desired format. In the feature extraction stage, each word root and word level in the text is determined. Term weighting is the frequency with which a feature appears in the text. A high weight of a term in a document means that this term has a distinctive feature for that text. In the classification stage, it is determined to which category the news content belongs. Three news categories were used in the study. The main stages of the study are presented in Figure 1.

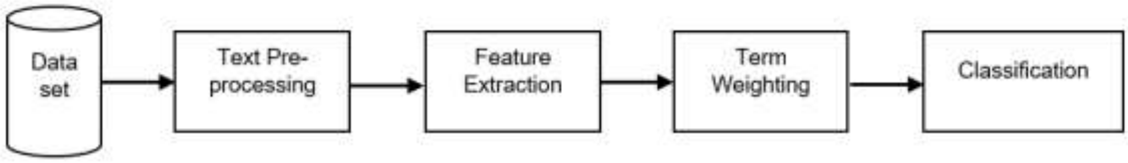

Figure 1. Stages of the study

2.1. Dataset

For the creation of dataset of the study, the data were collected regularly for one week using the RSS information of five different news sites. RSS is a monitoring system used to keep users up to date with new streams and to follow up-to-date information on websites with new information that are regularly updated. The data were taken from there on a daily basis and transferred to the Excel table using the Knime application, and the dataset consisting of a total of 10.500 was created. The RSS information retrieval scheme in the Knime program is presented in Figure 2.

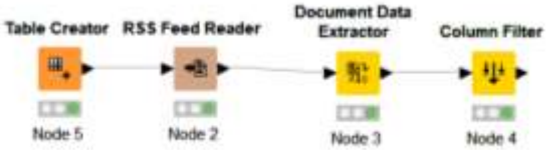

Figure 2. Information retrieval scheme from Knime news sites

Five news sites with high number of visitors and interaction in digital broadcasting in Turkey were included in the study. These news sites are as follows:

- Haber Türk
- A Haber
- CNN Türk
- Mynet
- Hürriyet

100 news were received daily from each category of each news site, and thus, 300 news were received from a news site at the end of the day. At the end of a week, a dataset consisting of 10.500 news in three categories from five news sites was created. The dataset, in which the analysis of machine learning methods was applied in the study, included Turkish news texts in text form and consisting of two columns. In the dataset, while the first column will represent the news text, the second column will contain the category headline of the news. The news texts are divided into three categories: economy, sports and world. News sites share RSS values on their websites according to their categories.

2.2. Text Pre-processing

The pre-processing stage may differ by each dataset and study. Furthermore, the importance of the pre-processing stage is an undeniable fact since removing the noise in the text and making the text structural significantly affect the classification success rate (Başkaya and Aydın, 2017).

The preparation stage was carried out for the removal and classification of the data in the dataset created in the pre-processing stage. In this stage, the texts, punctuation marks, and commands that are included in the news but evaluated in content analysis are first extracted. Before pre-processing, the dataset includes news texts in XML format taken from the RSS feed. The information under the <text> heading of the XML data obtained for classification studies was considered. In this stage, Java Script codes under this heading, all html tags and contents in the textual expression, operators, punctuation, non-printable characters, conjunctions, words that are meaningless on their own such as interrogative particles, and unnecessary data such as advertisements are also removed (Acı and Çırak, 2019).

2.3. Feature Extraction

Two types of feature methods, including each root word and word level n-gram in the text, were used in this study. The word level n-gram, which consists of word combinations of different lengths extracted from the text, is used. The words can also be meaningful in groups of 2,3. The words are grouped as unigram, bigram, trigram etc. and analysed to examine these word groups. The sentence "Turkish is an agglutinative language" is expressed as below with word level 1-gram (unigram), 2-gram (bigram) and 3-gram (trigram) (Başkaya and Aydın, 2017): Turkish is an agglutinative language

Unigram: "Türkçe", "sondan", "eklemeli", "bir","dildir"

Bigram: "Türkçe_sondan", "sondan_eklemeli", "eklemeli_bir", "bir_dildir"

Trigram: "Türkçe_sondan_eklemeli","sondan_eklemeli_bir", "eklemeli_bir_dildir"

2.4. Term Weighting

It is the frequency with which a feature appears in the text. A high weight of a term in a document means that this term has a distinctive feature for that text. Therefore, the weighting of terms is one of the factors that affect the classification performance (Dayıbaşı, 2022). 3.4.1. Binary Scoring, the vector is weighted as (1,0) according to the presence or absence of the relevant word in the document. Raw Frequency refers to the number of times the term occurs in the document / the number of words in the document. Inverse Document Freq (IDF) tries to understand whether this word is a term or not as a conjunction etc. (Stop Words) by finding the number of occurrences of the word in multiple documents. For this purpose, the absolute value of the logarithm of the Number of Documents in which the Term is Used/the Number of Documents is taken. Zipf Law performs sequencing from the most frequent word to the least frequent word in a text that exceeds 100 words. The weight/score of the word is calculated by dividing the index in this sequencing by the number of words.

2.5. Classification

Most basically, the text classification process, in which an existing text is determined to be included in which predetermined classes, is carried out by determining whether each text or document in the set T={t_1,t_2,...,t_n} belongs to the classes in the predefined set C={c_1,c_2,...,c_m} (Tantuğ, 2012). For a t_i document or text that is evaluated accordingly, the value is produced according to whether it is included in any class from the C set or not (Uslu and Akyol, 2021). Decision trees and Naive Bayes classification algorithm were used for the classification processes in the study.

2.5.1. Decision Trees

Decision trees are used to classify an object into a predefined set of classes based on its features. It can be used in complex data sets. A decision tree is a structure used to divide a dataset containing many records into smaller sets by applying a set of decision rules.

It is based on creating a tree by breaking apart the variables. It is a very useful technique because of its tree structure and easy rule extraction. Decision trees are composed of decision nodes and leaf nodes according to feature and target. While dividing a dataset into smaller and smaller subsets, an associated decision tree is progressively developed at the same time. Each feature of the dataset becomes a root node, and the leaf nodes represent the results.

Many algorithms such as ID3, C4.5, CART, Random Forest can be used in decision trees. The most important step in the creation of decision trees is to determine the criteria for branching in the tree or according to which feature values the tree structure will be created. In the literature, there are various approaches developed to solve this problem. The most used algorithms in decision trees are Gini and Entropy (Gain) algorithms.

The left and right Gini values are calculated before calculating the Gini value of a feature. The calculation of the left Gini and the calculation of the right are presented in equation 1 and equation 2, respectively (Adak and Yurtay, 2013).

$$Gini_{left} = 1 - \sum_{i=1}^{k} \left[\frac{L_i}{|T_{left}|} \right]^2 \tag{1}$$

$$Gini_{right} = 1 - \sum_{i=1}^{k} \left[\frac{R_i}{|T_{right}|} \right]^2 \tag{2}$$

Here, k represents the number of classes, T represents the number of samples in a node, T_{left} represents the number of samples in the left arm, T_{right} represents the number of samples in the right arm, L_i represents the number of samples in category i in the left arm, and R_i represents the number of samples in category i.

The calculated left and right values are used in calculating the Gini value of the feature. The Gini value of a feature is calculated according to Equation 3.

$$Gini_j = \frac{1}{n} (|T_{left}| Gini_{left} + |T_{right}| Gini_{right}) \tag{3}$$

Among the Gini values calculated for each feature, the smallest one is selected, and the division takes place over this feature.

2.5.2. Naive Bayes Classifier

The Bayes' theorem is an important topic studied within probability theory. This theorem shows the relationship between conditional probabilities and marginal probabilities within the probability distribution for a random variable. Naive Bayes classifier is based on Bayes' theorem. The working logic of the algorithm is to calculate the probability of each situation for an element and classify it according to the highest probability value. Success can be achieved with a little training data.

C= {C_1, C_2, ..., C_m} represents a collection of different classification sets and represents the number m, and X represents an example of a case with an unknown classification. P(C_i) is the previous probability of C_i. P (X I C_i) is the probability of X case samples, provided that the CI assumption is acceptable.

While classifying texts according to Bayes' theorem, the probability that the d_j document belongs to a class c is calculated as follows (Uslu and Akyol, 2021).

$$p(c|d_j) = \frac{p(d_j|c)p(c)}{p(d_j)} = \frac{p(d_j|c)p(c)}{p(d_j|c)p(c) + p(d_j|\overline{c})p(\overline{c})} \tag{4}$$

$$p(c|d_j) = \frac{\frac{p(d_j|c)}{p(d_j|\overline{c})} \cdot p(c)}{\frac{p(d_j|c)}{p(d_j|\overline{c})} \cdot p(c) + p(c)} \tag{5}$$

3. Results and Discussion

In the study, 1.500 news were collected daily from five different digital newspapers, and at the end of 15 days, a dataset consisting of a total of 10.500 news was created. Text pre-processing, feature extraction and term weighting operations, which are text processing operations, were performed on these data. Sample news in the dataset used in the study are presented in Table 1.

Sample news	Categories
Ücretli çalışan sayısı arttı Sanayi, inşaat ve ticaret-hizmet sektörleri toplamında ücretli çalışan sayısı 2022 Mayıs ayında bir önceki yılın aynı ayına göre yüzde 5,7 arttı.	Economy
"THY'den çifte rekor Türk Hava Yolları THY önceki gün 543 sefer 260 bini aşan yolcu sayısıyla tarihinin yoğun gününü yaşadı Yolcu uçuş rakamlarında önceki gün THY tarihinin yoğun günü yaşandı"	
"Konut satışları yüzde 107,5 arttı Türkiye İstatistik Verileri ' ne göre mayıs ayında konut satışı geçtiğimiz yılın aynı dönemine göre yüzde 107,5 oranında arttı"	
"Antalyaspor Haji Wright 3 1 yıllık sözleşmeye imza atacak Spor Toto Süper Lig ekiplerinden Fraport TAV Antalyaspor geçen sezon kiralık kadrosunda bulunan transfer döneminde ismi Türk takımıyla anılan ABD'li forvet Haji Wright 3 1 yıllık sözleşme imzalayacak"	Sport
"Burak Yılmaz Fenerbahçe avantajlı görüyorum Lille sözleşmesinin sona ermesinin ardından Fortuna Sittard transfer Burak Yılmaz sezon Türkiye Ligi nde şampiyonluk favorisini açıkladı"	
"Milliler Akdeniz Oyunları'nı 108 madalya tamamladı 19 Akdeniz Oyunları' son gününde ay-yıldızlılar 3 altın 1 gümüş 1 bronz madalya kazanırken oyunları 45 altın 26 gümüş 37 bronz madalya tamamladı"	
"İngiltere kırmızı alarm verildi Ülke tarihinde ilk yaşanacak İngiltere kırmızı alarm verildi Ülkede pazartesi salı günleri hava sıcaklığının ilk 40 dereceye ulaşması beklenirken öngörülen sıcaklıklar olağan dışı nitelendirildi Aşırı sıcaklıkların altyapıyı etkileyebileceği olumsuz sağlık koşullarına yol açabileceği konusunda uyarıda bulunuldu İngiltere Meteoroloji Ofisi Sözcüsü Grahame Madge sıcaklıkların 40 dereceye ulaşmasının tarihi an olacağını söyledi"	World
"Sri Lanka olağanüstü hâl uzatıldı Sri Lanka geçici devlet başkanı Ranil Wickremesinghe ülkede protestoların ardından ilan edilen OHAL uzattığını açıkladı"	
"Çin sel felaketi 12 ölü 12 kayıp Çin şiddetli yağışın sel felaketinde 12 kişi hayatını kaybetti 12 kişi sel sularında kayboldu"	

Table 1. Sample dataset

Precision, sensitivity, and accuracy metrics will be used to compare the results of the study with similar studies in the literature. The model's statements showing TP (true positive) and TN (true negative) correct classifications and FP (false positive), and FN (false negative) misclassifications are first analysed in determining these values.

Precision (π_i) indicates the probability that this classification is correct if any document d is included in class ci.

$$\pi_i = \frac{TP_i}{TP_i + FP_i} \tag{6}$$

Sensitivity (p_i) is defined as how many of the documents that should be under class ci are included in this class.

$$p_i = \frac{TP_i}{TP_i + FN_i} \tag{7}$$

Accuracy (A_i) indicates the ability of the classifier to obtain accurate results and is calculated as in equation 8. The error rate is found as 1's complement of this value.

$$A_i = \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i} \tag{8}$$

F1-score represents the harmonic mean of the Precision and Sensitivity values to avoid ignoring extreme cases.

$$F_1 = 2 \cdot \frac{\pi_i \cdot p_i}{\pi_i + p_i} \tag{9}$$

The dataset used in the study was divided into two as training and test data. The training data is used in determining the parameters of the model. The test dataset is used to test and analyse the performance of the created model. Although there is no definite rule about the separation of data as training and testing, this rate is 80%-20%, 70%-30% in the literature. In this study, 70% of the data, in other words 7.350 news items were used for training purposes, while 3.150 news items were used to test the generated model. Take from top, linear sampling and draw randomly are some of the data selection methods that can be used. In the study, linear sampling method was preferred in data selection to compare the results of the two models.

In the study, the Knime program was used for text mining and machine learning method. KNIME program consists of the abbreviation of "Konstanz Information Miner". KNIME is an open-source program and a platform used for data analysis, reporting and integration processes. The model created in the Knime program is presented in Figure 3.

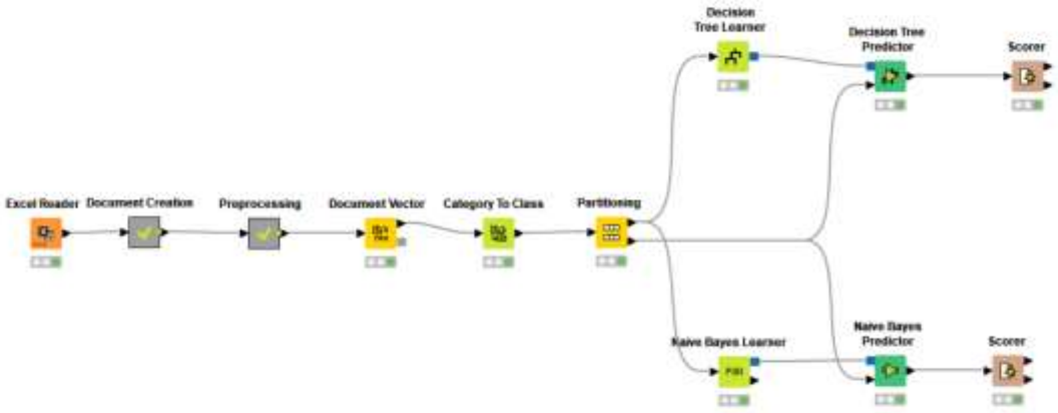

Figure 3. Study model

When the results of the study were evaluated according to the F1 score, it was observed that the success of the Naive Bayes classifier was 88.66% and the success of the decision trees was 82.96%. According to these results, both algorithms can be considered sufficient and successful. In the study, it was observed that the most successful algorithm between the two models was the Naive Bayes classifier.

Amasyalı and Yıldırım used Naive Bayes, Vector Quantization and Multilayer Classifiers with 76% success (Amasyalı and Yıldırım, 2004). Aşlıyan and Günel achieved 88.4% success in their studies using the nearest neighbour algorithm (Aşlıyan and Günel, 2010). Usmani and Shamsi achieved 88% success in their studies using natural language processing techniques (Usmani and Shamsi, 2020). In the Uslu and Akyol study, the most successful method was the Naive Bayes Classifier with 91% of the studies they carried out according to the support vector classifier, random forest and Naive Bayes Classifier (Uslu and Akyol, 2021). The study is similar to the studies in the literature. According to the F1 score, Naive Bayes was 88.66%; the decision tree method was 82.96% successful.

4. Conclusions

Internet, which is one of the most important opportunities provided by the developing technology, has deeply affected all areas of life and the field of press media. Newspapers, that met the readers as printed publications as of the 17th century although their history dates to ancient times, started to reach its readers through digital channels along with the widespread use of the internet. Digital journalism provides many advantages to both readers and publishers, and thus, the most remarkable element among these advantages is the fast news flow, which enables the news to reach millions of people within minutes. Rapid news production and transfer processes have required that the issues such as the accurate transfer of news, the quality and reliability of the news should also be addressed carefully. While publishers are trying to offer the best in the field of digital journalism, the utilization of technological opportunities provides many conveniences to both the publisher and the reader. In this context, the news was classified according to categories using machine learning methods in this study.

In the study, an original dataset consisting of 10.500 news in three different categories taken from five different digital newspapers for one week was used. The results obtained according to the F1 score were 88.66% for Naive Bayes and 82.96% according to the decision tree method. According to these results, it was concluded that both models showed acceptable success, however, Naive Bayes was more successful. When the study was compared with other studies in the literature, it was observed that it was successful compared to similar studies in terms of F1 score.

Nevertheless, increasing the data acquisition time and receiving news by including different news categories will increase the number and diversity of news in the dataset, which will have a positive effect on the success rate of the study. Another important element that will increase the success of the study is the text pre-processing stage. In this stage, the removal of noisy data by testing different algorithms and non-inclusion of the news below a certain number of words in the dataset will positively affect the success of the study. Studies to be conducted using support vector machines and deep learning algorithms, which are other machine learning methods, will also affect the success rate.

The categorization of news by using machine learning methods in digital journalism offers advantages such as reaching the right audience, categorizing the news quickly and accurately, archiving and ease of workforce.

References

Acı, Ç.İ., Çırak, A. 2019. "Türkçe Haber Metinlerinin Konvolüsyonel Sinir Ağları ve Word2Vec Kullanılarak Sınıflandırılması", Bilişim Teknolojileri Dergisi, 12(3), 219–228.

Adak, M.F., Yurtay, N. 2013. "Gini Algoritmasını Kullanarak Karar Ağacı Oluşturmayı Sağlayan Bir Yazılımın Geliştirilmesi," Internatıonal Journal of Informatics Technologies, 6(3), 1-6.

Amasyalı, M.F., Yıldırım, T. 2004. "Otomatik haber metinleri sınıflandırma", 13. Sinyal İşleme ve Uygulama Kurultayı, 224–226, Kuşadası, Türkiye.

Amasyalı, M.F., Beken, A. 2009. "Türkçe Kelimelerin Anlamsal Benzerliklerinin Ölçülmesi ve Metin Sınıflandırmada Kullanılması", IEEE 17. Sinyal İşleme ve İletişim Uygulamaları Kurultayı, Antalya, Türkiye.

Amasyalı, M.F., Diri, B., Türkoğlu, F. 2006. "Farklı özellik vektörleri ile Türkçe dokümanların yazarlarının belirlenmesi", 15th Turkish Symposium on Artificial Intelligence and Neural Network, Muğla, Türkiye.

Aşlıyan, R., Günel, K. 2010. "Metin İçerikli Türkçe Dokümanların Sınıflandırılması", Akademik Bilişim Konferansı, 659–665, Muğla, Türkiye.

Aydoğan, D. 2013. Türkiye'de dijital gazetecilik: Habertürk ve Hürriyet gazeteleri örneği. Turkish Online Journal of Design Art and Communication, 3(3), 26-40.

Bardoel, J. (1996). Beyond journalism: A profession between ınformation society and civil society. European Journal of Communication, 11(3), 283-302.

Başkaya, F., Aydin, İ. 2017. "Haber metinlerinin farklı metin madenciliği yöntemleriyle sınıflandırılması", International Artificial Intelligence and Data Processing Symposium (IDAP), Malatya, Turkey.

Çakır, H., 2007. "Geleneksel Gazetecilik Karşısında İnternet Gazeteciliği". Erciyes Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 22(1), 123-149

Dayıbaşı, O. 2022. "Metin Madenciliği'nde Kavramlar 1", medium.com, https://medium.com/algorithms-data-structures/metin-madencili%C4%9Finde-text-mining-kavramlar-1-e11b87b28847, Son erişim tarihi: 29 Nisan 2022

Doğan, S., Diri, B., 2010. "Türkçe dokümanlar için N-gram tabanlı yeni bir sınıflandırma (Ng-ind): yazar, tür ve cinsiyet", Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi, 3(1), 11-19.

Levent, V.E., Diri, B. 2014. "Türkçe dokümanlarda yapay sinir ağları ile yazar tanıma", 15. Akademik Bilişim Konferansı, 735–741, Mersin, Türkiye.

Toraman, C., Can, F., Koçberber, S. 2011. "Developing a Text Categorization Template for Turkish News Portals", International Symposium on Inovations in Intelligent Systems and Applications, İstanbul, Turkey.

Tüfekci, P., Uzun, E., Sevinç, B. 2012. "Türkçe Dilbilgisi Özelliklerini Kullanarak Web Tabanlı Haber Metinlerinin Sınıflandırılması", 21. IEEE Sinyal İşleme ve İletişim Uygulamaları Kurultayı, Girne, KKTC.

Uslu, Osman, Akyol, S. 2021. "Türkçe Haber Metinlerinin Makine Öğrenmesi Yöntemleri Kullanılarak Sınıflandırılması", Eskişehir Türk Dünyası Uygulama ve Araştırma Merkezi Bilişim Dergisi, 2(1), 15-20.

Usmani S, Shamsi J.A. 2020. "News Headlines Categorization Scheme for Unlabelled Data", International Conference on Emerging Trends in Smart Technologies (ICETST), Karachi, Pakistan.