Journal of Soft Computing and Artificial Intelligence
Journal homepage: https://dergipark.org.tr/en/pub/jscai

International Open Access
Volume 04
Issue 01
June, 2023

Research ***Article***

Classification of News Texts from Different Languages with Machine Learning Algorithms

Sidar Ağduk [1], ***Emrah Aydemir*** [2], ***Ayfer Polat*** [3]

[1] *Management Information Systems, Faculty of Economics and Administrative Sciences. Tarsus University, 33000, Tarsus, Turkey*
[2] *Management Information Systems, Institute of Business. Sakarya University, 54000, Sakarya, Turkey*
[3] *SAP Applications Business Analyst, BIZE Project Development Inc., 16000, Bursa, Turkey*

ARTICLE INFO

Article history:
Received **June 8, 2023**
Revised **June 12, 2023**
Accepted **June 19, 2023**

Keywords:
Text Mining
Text Classification
Machine Learning
Weka
Naive Bayes

ABSTRACT

As a result of the developments in technology, the internet is accepted as one of the most important sources of information today. Although it is possible to access a large number of data in a short time thanks to the Internet, it is critical to analyze this data correctly. The need for text mining is increasing day by day by processing and analyzing the increasingly irregular text type data in the digital environment and classifying them in a meaningful way. In this study, news texts obtained from online German, Spanish, English and Turkish news sites were separated according to predetermined world, sports, economy and politics categories. The data set consisting of 4000 news texts was classified using 41 different machine learning algorithms in the Weka program. The highest successful classification was obtained with Naive Bayes Multinominal and Naive Bayes Multinominal Updateable algorithms, and 93.5% for German news texts, 93.3% for English news texts, 82.8% for Spanish news texts and 88.8% for Turkish news texts.

1. Introduction

Thanks to the advancements in internet and information technologies, access to information has become significantly easier [1, 2]. In particular, the increasing use of the internet has led to the vast expansion of accessible data [3]. Therefore, in the age of information and technology we find ourselves in, it is crucial to be able to quickly access the desired accurate data [4]. Data mining techniques, which vary depending on the type of data stack, are used to extract meaningful information, process, and analyze the complex array of data found online [5]. Text mining, one of the types of data mining, is used to extract meaningful information from textually stored data such as emails, web pages, reports, articles, and official documents [6]. Text mining, resulting from the combined use of natural language processing and data mining techniques, uncovers the hidden meanings in textual data stacks with the help of computer systems [7]. In text mining applications, textual data is classified and categorized using natural language processing or data mining methods, and a model is created. Text mining performs prediction when encountering a new text that is not included in the dataset, based on the established model [8]. Text classification, used for the purpose of classification in text mining, enables the assignment of categories to newly encountered texts from existing categories [9].

The rest of the study is organized as follows: the second section provides information about text

[1] Corresponding author
e-mail: sidaragduk@tarsus.edu.tr
DOI: 10.55195/jscai.1311380

classification and the third section presents the studies on text classification found in the literature. The fourth section discusses the purpose of the study, the dataset and preprocessing stages, feature extraction, performance evaluation criteria, and the classification methods used, under the heading of methodology. The fifth section presents the findings of the study, while the final section includes evaluations of the results.

2. Text Classification

Due to the advancements in information and communication technologies, the number of documents created in the online environment has been increasing every day [10]. While the increase in accessible information brings many benefits, it also presents some challenges [2]. The classification of texts found in the online environment is among these challenges. Simply put, text classification is the process of determining to which previously defined category or categories a given text data belongs. In other words, text classification involves determining whether the textual data in set $B=\{b_1,b_2,...,b_n\}$ belongs to the classes in set $S=\{s_1,s_2,...,s_m\}$ that have been predetermined. Therefore, it is necessary to generate a value, true or false, for $(b_j, s_i) \in B \times S$. A function g can be represented as $g : D \times C \rightarrow \{true, false\}$, where g produces the actual results, i.e., true if the j_{th} document belongs to the i_{th} class, and false otherwise. Accordingly, a similar function f that operates in a similar manner can be created using machine learning methods, represented as $f : D \times C \rightarrow \{true, false\}$. The aim is for the results produced by the generated *f* function to be as similar as possible to the results of the *g* function. A model is created using machine learning methods, and an *f* function (classifier) that operates similarly to the g function is implemented. Finally, the similarity between the *f* function and the actual results, represented by *g*, is compared [11].

With the advancements in technology and the increasing use of the internet, there is a growing need for data analysis and categorization [1]. News agencies, one of the most important sources of information today, have incorporated the online environment into their publishing activities as a result of technological developments. Proper classification, labeling, and presentation of the content offered to readers are of critical importance in enabling access to accurate news texts [12, 13]. Text classification, which involves automatically separating documents into specific semantic categories, effectively utilizes machine learning techniques. Documents consisting of textual data can be uncategorized or composed of content belonging to one or more categories. In order to classify texts automatically using machine learning, textual data needs to be transformed into numerical form using various approaches. TF-IDF, Word2Vec, and FastText methods are among the approaches used to extract vector models of texts [10]. Upon reviewing the conducted studies, it can be observed that various classification algorithms such as Random Forest [1, 2, 4, 10, 14-16], K-Nearest Neighbor [12, 16-18], Naive Bayes [2, 4, 10, 12, 14-16], Support Vector Machines [4, 10, 12, 15, 16, 19], [15], C4.5 [12, 14], Artificial Neural Networks [10, 20-25], and Logistic Regression [10] are utilized in the classification of textual data.

3. Relevant Literature

Aydemir et al. classified 2248 news texts from a Turkish-language news website into eight different categories using Multinomial Naive Bayes Algorithm (MNBA) and Random Forest (RF) algorithms based on predefined news categories. The study achieved a classification accuracy rate of 95.24% with MNBA and 99.86% with RF algorithm [2]. Başkaya and Aydın classified a total of 80 news texts, consisting of four different categories and 20 news texts for each category, from different news websites and newspapers using Naive Bayes (NB), J48 Decision Trees, Support Vector Machine (SVM), and RF. The highest successful result was achieved with the Random Forest algorithm, with a success rate of 100% in all four classification types [1]. Uslu and Akyol performed text classification using 4900 Turkish news texts. The news texts consisted of seven different categories, with 700 news texts in each category. SVM, RF, and NB algorithms were used for the classification of Turkish news text contents in the study. The analysis results showed a successful classification rate of 89% with SVM, 87% with RF, and 91% with NB [4]. Acı and Çırak used the widely used Turkish Text Classification 3600 dataset for the classification of Turkish news contents. The dataset consists of 3600 news data, with 600 news texts in each of the six different categories. Convolutional Neural Networks and Word2Vec method were used for the text classification process, resulting in a success rate of 93.3% [3]. Çelik and Koç performed text classification on a dataset of 12,000 data samples

from different Turkish news sources belonging to six different categories. The news texts, vectorized using Tfidfvectorizer, Word2Vec, and FastText methods, were then classified using DVM, NB, LR, RF, and ANN methods. The study achieved a highest success rate of 95.75%, obtained by classifying FastText vectorized news texts with DVM [10]. Şimşek and Aydemir classified 1017 emails obtained from 20 different Gmail and Hotmail accounts as spam or legitimate emails using 45 different classification algorithms. The study yielded the highest accurate classification rate of 94.78% with Naive Bayes Multinomial and Naive Bayes Multinomial Updateable algorithms [26]. Table 1 provides information about the studies included in the literature related to text classification.

Table 1 Related Literature

Study & Author	Data Size	Classification Method	Success Rate (%)
[1] Başkaya & Aydın (2017)	80	Naive Bayes Support Vector Machine C4.5 Algorithm Random Forest	90 95 65 100
[2] Aydemir et al. (2021)	2248	Multinomial Naive Bayes Random Forest	95.24 99.6
[3] Acı & Çırak (2019)	3600	Convolutional Neural Networks	93.3
[4] Uslu & Akyol (2019)	4900	Support Vector Machine Random Forest Naive Bayes	89 87 91
[15] Cusmuliuc et al. (2018)	10000	Naive Bayes Support Vector Machine Random Forest	92.43 95 95.93
[17] Aşlıyan & Günel (2010)	250	k-Nearest Neighbors	76.8
[27] Dilrukshi et al. (2013)	3569	Support Vector Machine	75
[28] Deniz et al. (2019)	799	Logistic Regression Naive Bayes Decision Tree Random Forest Support Vector Machine k-Nearest Neighbors	73.12 78.12 59.37 60.62 78.75 78.12
[29] Sel et al. (2019)	18878	MaxEnt Classification	94.54
[30] Jehad & Yousif (2020)	20800	C4.5 Algorithm Random Forest	89.11 84.97
[31]	4964	Support Vector Machine Artificial Neural Network	74.62 72.99
Shahi & Pant (2018)		Naive Bayes	68.31

4. Method

4.1. Research Objective

Despite the significant advancements in technology that bring great convenience to our lives, they also come with certain drawbacks. Particularly, the widespread use of the internet as the primary tool for accessing information leads to the generation of a large volume of data in real-time in the online environment. News agencies, being one of the most important sources of information, have also incorporated the online platform into their publishing activities due to these technological developments. In the face of increasing data on the internet, proper classification, labeling, and presentation of content to readers have become critically important for ensuring access to accurate news articles [12, 13]. In line with this, this study aims to successfully classify German, Turkish, Spanish, and English news texts into predefined categories such as world, economy, politics, and sports using various classification algorithms. The flowchart of the study is presented in Figure 1.

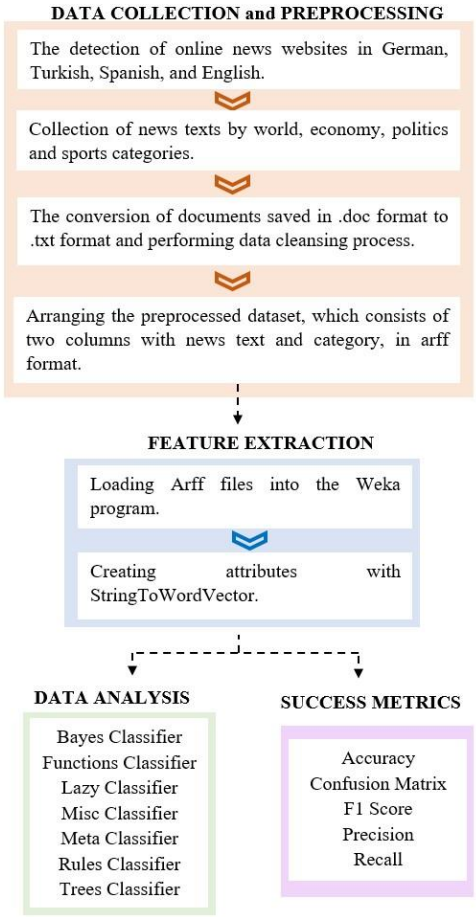

Figure 1 Study Flowchart

4.2. Data Set

In the study, a total of 4,000 news articles were obtained from online news websites publishing in German, English, Spanish, and Turkish languages, covering the categories of world, economy, politics, and sports. Each category consists of 250 news articles. The dataset used in the study has been publicly published on Kaggle [33]. Detailed information about the dataset used in the study is presented in Table 2.

Table 2 Data Set

Language of News	Category Type			
	World	***Economy***	***Politics***	***Sports***
German	250	250	250	250
Spanish	250	250	250	250
English	250	250	250	250
Turkish	250	250	250	250
Total	1000	1000	1000	1000

4.3. Feature Extraction

In order for machine learning classification algorithms to understand the dataset consisting of news texts, the texts need to be converted into numerical format. For this purpose, the StringToWordVector filter in the Weka program is used, which employs techniques such as TF-IDF and n-grams to transform the texts into numerical vectors [26]. Firstly, in the study, the "RegExpFromFile" command available in the Weka program is selected to determine whether a word is a stopword or not. In this step, the "WordTokenizer" command is also chosen to tokenize the words for the vectorization process. The preprocessed .arff format news data, which have gone through various preprocessing stages, are loaded into the Weka program, and then the attributes are extracted using the "StringToWordVector" filter. The "StringToWordVector" filter, which covers all words, generates attributes in numerical values indicating the frequencies of the words [32]. The vectorized form of the word frequencies is used in the classification phase. In the study, 2257 word vectors are extracted as features for the English news dataset, 2088 for the Spanish news dataset, 2257 for the German news dataset, and a total of 2572 for the Turkish news dataset. The parameter provided for the "StringToWordVector" function is as follows:

- weka.filters.unsupervised.attribute.StringToWordVector -R first -W 1000 -prune-rate -1.0 -N 0 -stemmer weka.core.stemmers.NullStemmer -stopwords-handler "weka.core.stopwords.RegExpFromFile -stopwords \"C:\\\\Program Files\\\\Weka-3-8-6\"" -M 1 -tokenizer "weka.core.tokenizers.WordTokenizer -delimiters \" \\r\\n\\t.,;:\\\'\\\"()?!\""

4.4. Classification Method

The data consisting of news texts were analyzed using the Weka program. The Weka program, which takes its name from the initials of "Waikato Environment for Knowledge Analysis," was developed at Waikato University in New Zealand. This program, which is free, open-source, and Java-based, enables various operations such as Classification, Clustering, Association, Data Preprocessing, and Visualization. The program includes commonly used machine learning algorithms [32]. In this study, 41 different classification methods belonging to the Bayes Classifiers, Tree Algorithms, Rule-Based Classifiers, Function Classifiers, Lazy Algorithms, Various Classifiers, and Meta-Learning Algorithms were used in the Classify tab of the Weka program for the analysis of the news data. Before the classification process, the dataset needs to be split into training and test sets. The main goal of machine learning algorithms is to generate models that make accurate predictions on the separated training dataset and evaluate the accuracy of the model on new data. The data used to test the accuracy of the model constitute the test dataset. The simplest approach used to split the dataset for training and testing is to randomly assign a percentage, for example, 80% for training and 20% for testing. Splitting the data percentage-wise may introduce some errors in determining the training and test data based on data distribution. To overcome this issue, the cross-validation method was used to split all data into training and test sets within themselves. With this method, the data is initially divided into 10 separate groups, and one group is used for testing while the remaining nine groups are used for training, repeated 10 times. Then, the average of classification performances in each iteration is calculated to obtain the final success rate. This process is visually explained in Figure 2 below.

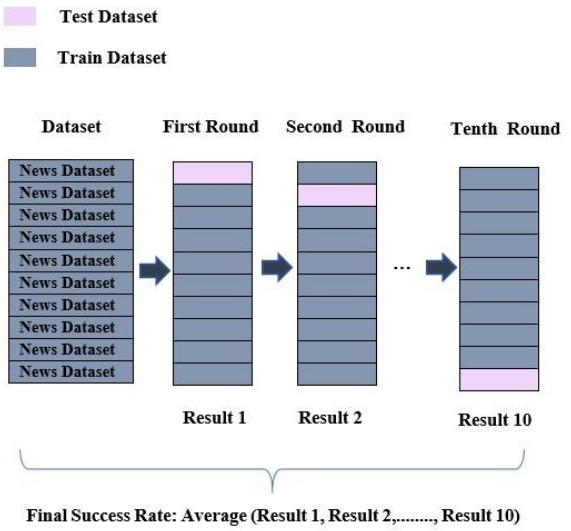

Figure 2 K-Fold Cross Validation

4.5. Performance Measures

In machine learning, the "Confusion Matrix," also known as the "Error Matrix," is used to compare the predicted and true values and interpret the performance of classification models. This matrix provides information about the correct or incorrect placement of test data into classes [32]. Along with the confusion matrix presented in Table 3 below, the following performance measures are obtained:

- Accuracy
- Recall
- Precision
- F1 score

Table 3 Confusion Matrix

		True Value	
		True	**False**
Prediction Value	**True**	True Positive (TP)	False Positive (FP)
	False	False Negative (FN)	True Negative (TN)

The definitions related to the confusion matrix in Table 3 are provided below:

- True Positive (TP): The instances that are correctly predicted as positive when the true value is positive.
- False Negative (FN): The instances that are incorrectly predicted as negative when the true value is positive.
- False Positive (FP): The instances that are incorrectly predicted as positive when the true value is negative.
- True Negative (TN): The instances that are correctly predicted as negative when the true value is negative.

In addition to these categorical values, precision (1), recall (2), F-score (3), and accuracy rate (4) are used when predicting categorical values. These values are calculated using the formulas provided below.

$$\frac{TP}{TP + FP} \quad (1)$$

$$\frac{TP}{TP + FN} \quad (2)$$

$$2 * \frac{Precision * Recall}{Precision + Recall} \quad (3)$$

$$\frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

5. Results

The features of the dataset consisting of news texts in different languages were determined using the "StringToWordVector" function in the Preprocess tab of the Weka program. The news data with extracted features were then tested with 41 different machine learning algorithms using the widely accepted 10-fold cross-validation method in the Classify tab, and the findings are presented in the following tables.

Table 4 Confusion Matrix

Algorithm		**Success Rate by Languages (%)**			
		German	English	Spanish	Turkish
BAYES	Bayes Net	85.2	85.1	73.3	84.2
	Naive Bayes	86.7	86.7	72.7	84
	Naive Bayes Multinominal	93.5	93.3	82.8	88.8
	Naive Bayes Multinominal Text	25	25	25	25
	Naive Bayes Multinominal Updateable	93.5	93.3	82.8	88.8
	Naive Bayes Updateable	86.7	86.7	72.7	84
TREE	Decision Stump	38.6	38.6	39.5	33.9
	Hoeffding Tree	25	25	25	25
	J48	80.6	80.7	66.8	77.2
	LMT	90.8	90.9	78.6	86.8
	Random Forest	88.3	88.8	77.9	87.3
	Random Tree	64.6	64.6	50.3	59.2
	REP Tree	76.2	76.7	64.5	73.3
RULES	Decision Table	71.8	71.8	65.6	63.8
	JRip	78.1	79.5	65.9	72.5
	OneR	40.3	40.3	41.8	36.4
	PART	82.5	82.5	65.6	77.8
	ZeroR	25	25	25	25
	Simple Logistic	90.8	90.9	78.9	86.6

	SMO	93.3	93.2	80.5	87.4
LAZY	IBk	68	68.1	49.1	64.7
	KStar	69.4	69.3	51.3	66.5
	LWL	41.7	41.7	41.9	40.3
MISC	Input Mapped Classifier	25	25	25	25
META	AdaBoostM1	38.6	38.6	39.5	33.9
	Attribute Selected Classifier	81.1	80.8	68.3	80.5
	Bagging	83.3	83.9	73.4	76.9
	Classication Via Regression	80.9	80.1	67.8	66.5
	CV Parameter Selection	25	25	25	25
	Filtered Classifier	82.7	82.5	68.2	82.2
	Iterative Classifier Optimizer	86.6	86.2	73.2	82.2
	Logit Boost	86.6	86.2	73.1	82.2
	Multi Class Classifier	87.1	87.4	50.7	77.7
	Multi Class Classifier Updateable	92.4	92.4	76.5	85.7
	Random Committee	82	83.6	68.8	81.9
	Randomizable Filtered Classifier	42.3	42.3	38.4	35.3
	Random Sub Space	84.6	84.7	74.5	82.7
	Stacking	25	25	25	25
	Vote	25	25	25	25
	Weighted Instances Handler Wrapper	25	25	25	25
	Multi Scheme	25	25	25	25

When examining Table 4, it is observed that the highest classification results for German, English, Spanish, and Turkish news data belong to the Naive Bayes Multinomial and Naive Bayes Multinomial Updateable classifiers. The analysis results for these algorithms regarding German, English, Spanish, and Turkish languages, including values such as True Positive Rate (TP), False Positive Rate (FP), and F-Score, are presented in Table 5. The confusion matrices for each language are shown in Figures 3, 4, 5, and 6, respectively.

Table 5 Other performance metrics for the top classification algorithms

News Dataset	Algorithm	Accuracy Rate (%)	Precision	Recall	F-Score	TP	FP
German	Naive Bayes Multinominal	93.5	0.937	0.935	0.935	0.935	0.022
	Naive Bayes Multinominal Updateable	93.5	0.937	0.935	0.935	0.935	0.022
English	Naive Bayes Multinominal	93.3	0.935	0.933	0.933	0.933	0.022
	Naive Bayes Multinominal Updateable	93.3	0.935	0.933	0.933	0.933	0.022
Spanish	Naive Bayes Multinominal	82.8	0.830	0.828	0.829	0.828	0.057
	Naive Bayes Multinominal Updateable	82.8	0.830	0.828	0.829	0.828	0.057
Turkish	Naive Bayes Multinominal	88.8	0.890	0.888	0.889	0.888	0.037
	Naive Bayes Multinominal Updateable	88.8	0.890	0.888	0.889	0.888	0.037

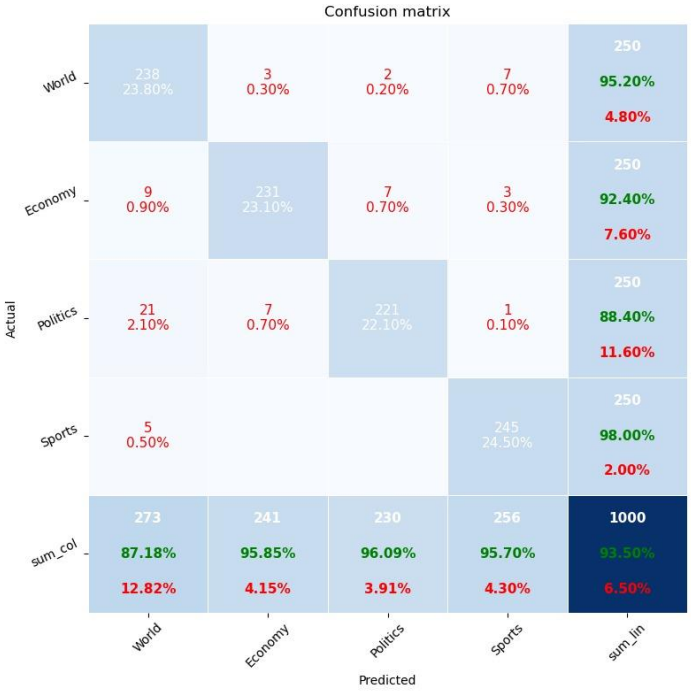

Figure 3 Confusion Matrix for German News Dataset

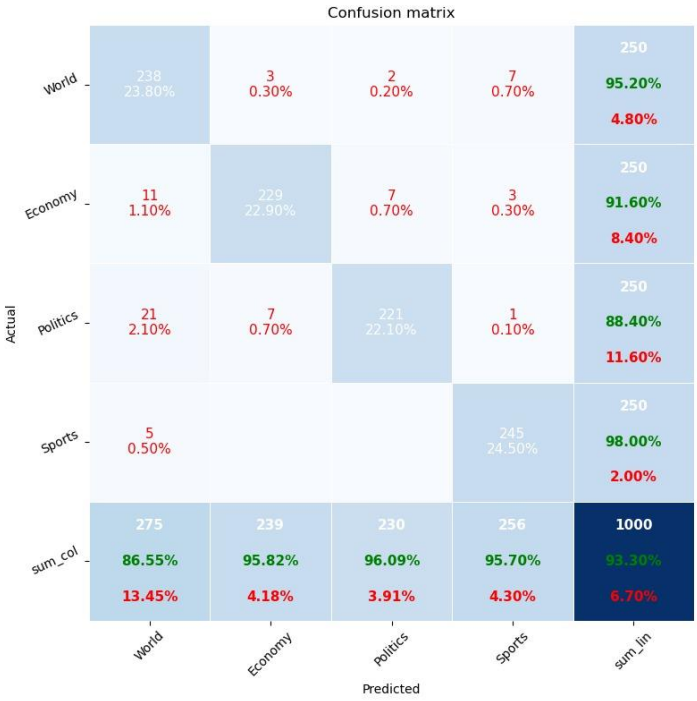

Figure 4 Confusion Matrix for English News Dataset

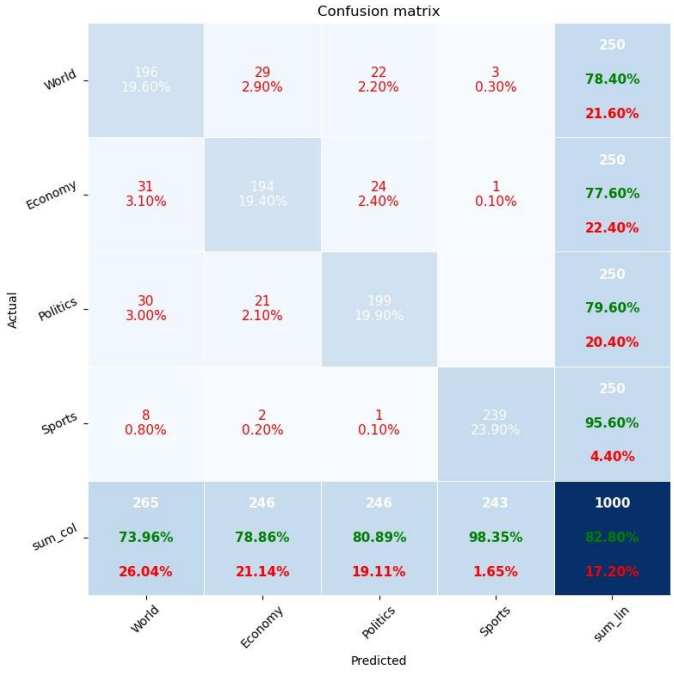

Figure 5 Confusion Matrix for Spanish News Dataset

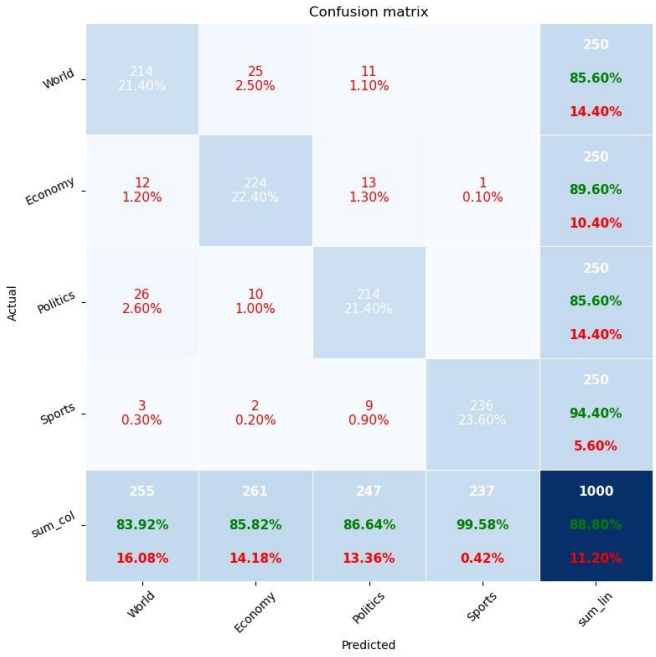

Figure 6 Confusion Matrix for Turkish News Dataset

6. Discussion and Conclusion

With the rapid increase of a large amount of textual data, particularly news articles, in online platforms and other sources, it has become increasingly important to effectively analyze and comprehend this data. Text classification of news articles serves as a fundamental step in categorizing and extracting meaningful information from these data. It assists many individuals in understanding news articles within large datasets, identifying trends, and making informed decisions. Therefore, accurate classification of news articles facilitates easy access to information, saves time, and plays an effective role in information management.

Aydemir et al. classified 2248 news texts from a

Turkish-language news website into eight different categories using Multinomial Naive Bayes Algorithm (MNBA) and Random Forest (RF) algorithms based on predefined news categories. The study achieved a classification accuracy rate of 95.24% with MNBA and 99.86% with RF algorithm [2]. Uslu and Akyol performed text classification using 4900 Turkish news texts. The news texts consisted of seven different categories, with 700 news texts in each category. SVM, RF, and NB algorithms were used for the classification of Turkish news text contents in the study. The analysis results showed a successful classification rate of 89% with SVM, 87% with RF, and 91% with NB [4].

In this study, German, English, Spanish, and Turkish news texts were classified according to the categories of world, economy, politics, and sports. A dataset consisting of 4000 news texts was tested using 41 classification algorithms in the Weka program. As a result, the highest classification performance was achieved with the Naive Bayes Multinomial and Naive Bayes Multinomial Updateable algorithms, belonging to the Bayes classifier, for all news texts. The success rates were determined as 93.5% for German news texts, 93.3% for English news texts, 82.8% for Spanish news texts, and 88.8% for Turkish news texts. Additionally, among other successful classification algorithms, it was observed that the SMO algorithm of the Functions classifier and the Multi Class Classifier Updateable algorithm of the Meta classifier were prominent. For the SMO algorithm, success rates of 93.3%, 93.2%, 80.5%, and 87.4% were obtained for German, Spanish, English, and Turkish news texts, respectively. For the Multi Class Classifier Updateable algorithm, success rates of 92.4%, 92.4%, 76.5%, and 85.7% were obtained for the same languages. Finally, it was determined that the Naive Bayes Multinomial Text, Hoeffding Tree, ZeroR, Input Mapped Classifier, CV Parameter Selection, Stacking, Vote, Weighted Instances Handler Wrapper, and Multi Scheme algorithms had the lowest success rates, indicating that the news texts were classified only into one category. In conclusion, this study demonstrates that Naive Bayes Multinomial and Naive Bayes Multinomial Updateable algorithms achieve high success rates when compared to similar studies in the literature. Furthermore, considering the success rates of other classification algorithms, it can be said that this study makes a significant contribution in terms of classification performance.

References

[1]. Başkaya, F., & Aydın, İ. Haber Metinlerinin Farklı Metin Madenciliği Yöntemleriyle Sınıflandırılması, In 2017 International Artificial Intelligence and Data Processing Symposium (IDAP), 2017, pp. 1-5. IEEE.

[2]. Aydemir, E. , Işık, M. & Tuncer, T. Türkçe Haber Metinlerinin Çok Terimli Naive Bayes Algoritması Kullanılarak Sınıflandırılması, Fırat Üniversitesi Mühendislik Bilimleri Dergisi, 2021, 33(2), pp. 519-526. doi: 10.35234/fumbd.871986

[3]. Acı, Ç. & Çırak, A. Türkçe Haber Metinlerinin Konvolüsyonel Sinir Ağları ve Word2Vec Kullanılarak Sınıflandırılması, Bilişim Teknolojileri Dergisi, 2019, 12(3), pp. 219-228. doi: 10.17671/gazibtd.457917.

[4]. Uslu, O., & Akyol, S. Türkçe Haber Metinlerinin Makine Öğrenmesi Yöntemleri Kullanılarak Sınıflandırılması, ESTUDAM Bilişim Dergisi, 2019, 2(1), pp. 15-20.

[5]. Doğan, K., & Arslantekin, S. Büyük Veri: Önemi, Yapısı Ve Günümüzdeki Durum, Ankara Üniversitesi Dil ve Tarih-Coğrafya Fakültesi Dergisi, 2016, 56(1), pp.15-36.

[6]. Bach, M. P., Krstić, Ž., Seljan, S., & Turulja, L. Text mining for big data analysis in financial sector: A literature review, Sustainability, 2019, 11(5), pp. 1-27.

[7]. Tan, A. H. Text mining: The state of the art and the challenges, In Proceedings of the pakdd 1999 workshop on knowledge disocovery from advanced databases, 1999, pp. 65-70.

[8]. Coşkun, C., & Baykal, A. Veri Madenciliğinde Sınıflandırma Algoritmalarının Bir Örnek Üzerinde Karşılaştırılması. Akademik Bilişim, 2011, 11, pp. 51-58.

[9]. Dalal, M. K., & Zaveri, M. A. Automatic Text Classification: A Technical Review, International Journal of Computer Applications, 2011, 28(2), pp. 37-40.

[10]. Çelik, Ö., & Koç, B. C. TF-IDF, Word2vec ve Fasttext Vektör Model Yöntemleri ile Türkçe Haber Metinlerinin Sınıflandırılması, Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi, 2021, 23(67), pp. 121-127.

[11]. Tantuğ, A. C. Metin Sınıflandırma, Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi, 2016, 5(2).

[12]. Toraman, C., Can, F., & Koçberber, S. Developing A Text Categorization Template For Turkish News

Portals, In 2011 International Symposium on Innovations in Intelligent Systems and Applications, 2011, pp. 379-383. IEEE.

[13]. Yıldırım, S., & Yıldız, T. Türkçe İçin Karşılaştırmalı Metin Sınıflandırma Analizi, Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi, 2018, 24(5), pp. 879-886.

[14]. Amasyalı, M. F., Diri, B., & Türkoğlu, F. Farklı Özellik Vektörleri İle Türkçe Dokümanların Yazarlarının Belirlenmesi, In The Fifteenth Turkish Symposium on Artificial Intelligence and Neural Networks (TAINN'2006), 2006, pp. 4.

[15]. Cusmuliuc, C. G., Coca, L. G. and Iftene, A. Identifying Fake News on Twitter using Naive Bayes, SVM and Random Forest Distributed Algorithms, In Proceedings of The 13th Edition of the International Conference on Linguistic Resources and Tools for Processing Romanian Language, 2018, pp.177-188.

[16]. Doğan, S., & Diri, B. Türkçe Dokümanlar için N-Gram Tabanlı Yeni Bir Sınıflandırma (Ng-İnd): Yazar, Tür ve Cinsiyet, Türkiye Bilişim Vakfi Bilgisayar Bilimleri ve Mühendisliği Dergisi, 2010, 3(1), pp. 11-19.

[17]. Aşlıyan, R., & Günel, K. Metin İçerikli Türkçe Dokümanların Sınıflandırılması, Akademik Bilişim Konferansı, 2010, pp. 659-665.

[18]. Soucy, P., & Mineau, G. W. A Simple KNN Algorithm For Text Categorization, In Proceedings 2001 IEEE international conference on data mining, 2001, pp. 647-648. IEEE.

[19]. Joachims, T. Text Categorization With Support Vector Machines: Learning With Many Relevant Features, In European conference on machine learning, 1998, pp. 137-142.

[20]. Ma, L., Shepherd, J., & Zhang, Y. Enhancing Text Classification Using Synopses Extraction, In Proceedings of the Fourth International Conference on Web Information Systems Engineering, 2003, pp. 115-124. IEEE.

[21]. Lam, S. L., & Lee, D. L. Feature Reduction For Neural Network Based Text Categorization, In Proceedings. 6th international conference on advanced systems for advanced applications, 1999, pp. 195-202. IEEE.

[22]. Ng, H. T., Goh, W. B., & Low, K. L. Feature Selection, Perceptron Learning, And A Usability Case Study For Text Categorization, In Proceedings Of The 20th Annual International ACM SIGIR Conference On Research And Development In Information Retrieval, 1997, pp. 67-73.

[23]. Nakayama, M., & Shimizu, Y. Subject Categorization for Web Educational Resources using MLP, In ESANN, 2003, pp. 9-14.

[24]. Srinivasan, P., & Ruiz, M. E. Automatic Text Categorization Using Neural Network, In Proceedings of the 8th ASIS SIG/CR Workshop on Classification Research, 1998, pp. 59-72.

[25]. Ma, S., & Ji, C. A Unified Approach on Fast Training of Feedforward and Recurrent Networks Using EM Algorithm, IEEE transactions on signal processing, 1998, 46(8), pp. 2270-2274. IEEE.

[26]. Şimşek, H. & Aydemir, E. Classification of Unwanted E-Mails (Spam) with Turkish Text by Different Algorithms in Weka Program, Journal of Soft Computing and Artificial Intelligence, 2022, 3(1) , pp. 1-10. doi: 10.55195/jscai.1104694

[27]. Dilrukshi, I., De Zoysa, K., & Caldera, A. Twitter News Classification Using SVM, In 2013 8th International Conference on Computer Science & Education, 2013, pp. 287-291. IEEE.

[28]. Deniz, E., Erbay, H., & Coşar, M. Classification Of Turkish E-Mails With Doc2Vec, In 2019 1st International Informatics and Software Engineering Conference (UBMYK), 2019, pp. 1-4. IEEE.

[29]. Sel, İ., Karci, A., & Hanbay, D. Feature Selection for Text Classification Using Mutual Information, In 2019 International Artificial Intelligence and Data Processing Symposium (IDAP), 2019, pp. 1-4. IEEE.

[30]. Jehad, R., & Yousif, S. A. Fake News Classification Using Random Forest and Decision Tree (J48), Al-Nahrain Journal of Science, 2020, 23(4), pp. 49-55.

[31]. Shahi, T. B., & Pant, A. K. Nepali News Classification Using Naïve Bayes, Support Vector Machines and Neural Networks, In 2018 International Conference on Communication Information and Computing Technology (ICCICT), 2018, pp. 1-5. IEEE.

[32]. Aydemir, E. Weka İle Yapay Zeka. Seçkin Yayınevi, 2018, Ankara.

[33]. Ağduk, S., Aydemir, E. & Polat, A. (2022). News Texts by Category in Different Languages [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/3572093