Gümüşhane University Journal of Science

GUFBD / *GUJS* (2024) 14(3): 764-776 doi: 10.17714/gumusfenbil.1396652 Research Article

A study on effective feature extraction and genetic algorithm based feature selection method in fake news detection classification using machine learning approaches

Makine öğrenimi yaklaşımları kullanılarak sahte haber tespiti sınıflandırmasında etkili özellik çıkarma ve genetik algoritma tabanlı özellik seçimi yöntemi üzerine bir çalışma

Ramazan İNCİR*[1], Mete YAĞANOĞLU[2], Ferhat BOZKURT[2]

[1]*Gümüşhane Üniversitesi, Kelkit Aydın Doğan Meslek Yüksekokulu, Bilgisayar Teknolojileri Bölümü, 29100, Gümüşhane*
[2]*Atatürk Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, 25240, Erzurum*

• Received: 27.11.2024 • Accepted: 06.05.2024

Abstract

In today's technology, information spreads quickly through online social networks, making our lives easier. However, when false news is shared without critical evaluation, it can harm society and affect social, political and economic aspects as it reaches a wide audience. At this point, it is important to develop content verification and confirmation systems. In this study, the aim is to conduct monolingual and cross-lingual classification on a multi-class dataset containing English and German news content. We applied data preprocessing, including CountVectorizer and stylometric feature extraction, before classification. Feature selection was made using the genetic algorithm, which is an algorithm based on the idea of evolution in nature. Selected features were classified by Random Forest, Logistic Regression, Multinomial Naive Bayes, Decision Tree and KNearest Neighbors machine learning algorithms. In the classification process, Multinomial Naive Bayes achieved 58.49% Accuracy and 42.97% macro-F1 for monolingual English news texts, while Logistic Regression achieved 45.39% Accuracy and 37.70% macro-F1 in Cross-lingual classification using English and German news texts. Significantly successful results were obtained compared to studies conducted with the same dataset. In addition, the same methodology was applied to the ISOT dataset. 99.48% and 99.62% macro-F1 were obtained by Logistic Regression and Decision Tree algorithms, respectively.

Keywords: Cross-lingual classification, Fake news detection, Genetic algorithm, Machine learning, Monolingual classification

Öz

Günümüz teknolojisinde bilgi çevrimiçi sosyal ağlar aracılığıyla hızla yayılarak hayatımızı kolaylaştırmaktadır. Ancak sahte haberler eleştirel bir değerlendirme yapılmadan paylaşıldığında geniş kitlelere kolaylıkla ulaştığı için topluma zarar verebilmekte ve sosyal, politik ve ekonomik yönleri etkileyebilmektedir. Bu noktada içerik doğrulama ve teyit sistemlerinin geliştirilmesi önem arz etmektedir. Bu çalışmada İngilizce ve Almanca haber içeriklerinin yer aldığı çok sınıflı bir veri seti üzerinde tek dilli ve diller arası bir sınıflandırma yapılması amaçlanmıştır. Sınıflandırmadan önce CountVectorizer ve stilometrik özellik çıkarımı da dâhil olmak üzere veri ön işleme uygulanmıştır. Özellik seçimi, doğadaki evrim fikrine dayanan bir algoritma olan genetik algoritma kullanılarak yapılmıştır. Seçilen özellikler Rastgele Orman, Lojistik Regresyon, Multinomial Naive Bayes, Karar Ağacı ve K-En Yakın komşu makine öğrenme algoritmaları ile sınıflandırılmıştır. Sınıflandırma sonucunda tek dilli İngilizce haber metinleri için Multinomial Naive Bayes algoritması ile %58.49 Doğruluk ve %42.97 makro-F1 elde edilirken, İngilizce ve Almanca haber metinleri kullanılarak diller arası sınıflandırmada Lojistik Regresyon algoritması ile %45.39 Doğruluk ve %37.70 makro-F1 elde edilmiştir. Aynı veri seti ile yapılan çalışmalara göre oldukça başarılı sonuçlar elde edildiği gözlemlenmiştir. Ayrıca ISOT veri setine de aynı metodoloji uygulanmıştır. Lojistik Regresyon ve Karar Ağacı algoritmaları ile sırasıyla %99.48 ve %99.62 makro-F1 elde edilmiştir.

Anahtar kelimeler*: Diller arası sınıflandırma, Sahte haber tespiti, Genetik algoritma, Makine öğrenimi, Tek dilli sınıflandırma*

* Ramazan İNCİR; ramazan.incir@gumushane.edu.tr

1. Introduction

With the development of internet technology, modern life has become more convenient, leading to increased communication and information sharing that significantly simplifies our lives. Almost every piece of information has become easily accessible. However, the content available can be created and disseminated by ordinary people, whether it's based on false or real information. Fake news refers to news that contains false or misleading content, giving the impression of being accurate but failing to reflect the truth.

In recent years, there has been an increase in the amount of false information and rumors in print newspapers, television, radio, the internet, and social media (Lima et al., 2022). With the prevalence of social media, suspicious and untrue news content has reached large masses unchecked. Users have contributed to the credibility of such news by sharing fake news on various platforms with the prejudice of believing, without questioning the accuracy of the information shared by their friends (Jain & Kasbe., 2018). Fake news spreads faster, deeper, and wider than real news, resulting in significant consequences (Zhang et al., 2022). The fact that major rumors on social media platforms such as Facebook and Twitter influenced the outcome of the 2016 US presidential election confirms this situation (DiFranzo & Gloria-Garcia., 2017).

The mass media plays a crucial role in communicating and disseminating messages in various fields, including politics and health. However, due to the huge amount of information produced, mass media may have difficulty in communicating the correct information effectively. Given the large volume of data, manual validation is impractical (Ghayoomi & Mousavian., 2022). Therefore, it is of great importance to develop detection systems to prevent the spread of fake news that can have serious consequences.

In this study, both monolingual and cross-lingual classifications were conducted on news articles, which encompassed classes of true, false, partially false, and others. These classifications were presented comparatively using five different machine learning algorithms for classification. Feature extraction from news articles was performed using CountVectorizer and Stylometric feature extraction methods. Feature selection was performed using the genetic algorithm applied to the extracted features. Our approach was also compared with studies on the same dataset. To further evaluate the performance of the methodology, classification was performed using the ISOT dataset and the resulting performance measures were presented.

The novelty of our work lies in feature extraction using CountVectorizer, combining these features with stylometric features, and enhancing performance through feature selection using the genetic algorithm. The scope and contributions of this study can be summarized as follows:

- Effective feature extraction, feature selection and machine learning algorithms are proposed for fake news detection.
- A feature set was created by extracting features from the dataset via Stylometry and CountVectorizer.
- Feature selection was performed using a genetic algorithm on the created feature set.
- To assess the robustness of our methodology, classification was conducted using the ISOT dataset.
- In this study, experimental studies were carried out using five different machine learning techniques on a fake news article dataset with four distinct classes, and a comparative performance evaluation was conducted.

In the literature, there are studies that classify fake news using various methods. Ahmed et al. (2017) proposed a solution based on n-gram analysis and machine learning algorithms for fake news detection. In addition to six different machine learning, two different feature extraction equipment were used: Term Frequency (TF) and Term Frequency Inverted Document Frequency (TF-IDF). In the experimental evaluation, Linear Support Vector Machine (LSVM), one of the TF-IDF-based machine learning programs, reached 92% accuracy. Aborisade et al. (2018) aimed to detect fake news on 46,895 tweets they collected from Twitter. After the preprocessing steps such as short text cleaning, stop words cleaning, punctuation mark cleaning and tokenization, the classification process was conducted. The results showed an accuracy of 91.1% with logistic regression and 89.8% with a Naive Bayes classifier. Vogel et al. (2019) created the GermanFakeNC dataset consisting of German news texts for fake news detection. They transformed the train and test datasets into tf-idf feature vectors. As a result of the evaluation, it was stated that 0.72 and 0.89 Cohen's Kappa, 0.74 and 0.90 F1 Score were obtained with SVM and CNN classifiers, respectively.

Ahmad et al. (2020), have created a series of combinations of various machine learning algorithms. As a result of the experimental evaluations, they stated that the average accuracy values of 88.16% were obtained, in which ensemble learners performed better than individual learners on average. The lowest performance was obtained with Wang-Bi-LSTM. Nasir et al. (2021), presented a hybrid deep learning model combining Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) for fake news detection. Classification was done on ISOT and FA-KES datasets. They stated that the proposed hybrid model was more successful than other models. They stated that as a result of the classification made on the FA-KES and ISOT datasets, an average accuracy value of 60% and 99% was obtained, respectively.

Fayaz et al. (2022) extracted various features on the ISOT dataset to detect fake news. They used feature importance, information gain, univariate and chi2 for feature selection. The best 14 features were selected from 23 features. From the experimental results, an accuracy of 97.33% was obtained with chi2 in feature selection and random forest algorithm, one of the machine learning algorithms for classification. Taboubi et al. (2022), used pre-trained BERT base uncased and RoBERTa models. Preprocessing steps such as stop words and lemmatization were applied. The BERT model used consists of two sub-models. One of the models processed the title while the other processed the text. As a result, the macro F1 measure of 0.339 was reached with the BERT model. Martinez-Rico et al. (2022) LIWC explicit features and transformer models are used to extract hidden features in the text. At the end of the evaluations, the highest F1 measure value of 0.3324 was obtained as a result of using the third approach. Tran et al. (2022) aimed to classify fake news cross-linguals and monolinguals on a dataset consisting of English and German texts. As a result of the classification process, they obtained a macro-F1 score of 39.54% with T5-3B classifier for monolingual English news texts. They obtained XLM-RLarge 30.06% macro-F1 score for cross-lingual English and German news texts. LekshmiAmmal et al. (2022) utilized a Transformer-based model for the classification process after Text Preprocessing and Tokenization. As a result, they reported F1 scores of 0.2980 for monolingual and 0.2245 for cross-lingual texts.

Truică et al. (2022) used BiLSTM with BART sentence converters on English articles for monolingual fake news detection. For cross-lingual fake news detection, they used English and German articles and BiLSM with XLM sentence converters. As a result, they obtained an F1 Score of 0.32 and an accuracy value of 0.53 in monolingual classification, and an F1 Score of 0.19 and an accuracy value of 0.28 in cross-lingual classification. Pritzkau et al. (2022) addressed a multi-class classification problem. They applied pre-processing steps such as removing duplicate data on the dataset consisting of English news articles. They made classification using RoBERTa and Longformer models. In their study, they obtained Macro-F1 at a rate of 0.3076. Althabiti et al. (2022) used a statistical measure called TF-IDF to convert all words to vectors in the text preprocessing stage on the dataset. Machine algorithms and transformer-based BERT, XLNet, RoBERTa and DistilBERT algorithms were used for the classification process. After classification, they obtained an F1 measure of 0.305 with the bert-large-cased model. Arif et al. (2022) used passive-aggressive classification, Bi-LSTM and RoBERTa for news classification. For the monolingual task consisting of English news articles, RoBERTa achieved the best performance with an F1 score of 28.60%. They achieved an F1 score of 17.21% with Bi-LSTM in the cross-lingual classification of news articles in English and German.

La Barbera et al. (2022) aimed to reveal passages matching Wikipedia data in articles using the word bag approach, validate the accuracy of passage claims using the T5 transformer, and classify with the BERT model. They obtained an F1 score of 0.275 in the classification results. Porto-Capetillo et al. (2022) proposed traditional machine learning algorithms and BERT embeddings from deep learning-based pre-trained architectures for the classification process. In order to increase the classification success of the models, they used additional stylometric features. At the end of the study, they obtained a F1-macro score of 0.2951 as a result. Blanc et al. (2022) proposed a two-component BERT-based system. In the first component, they determined whether the text content sample belongs to the classes. The second component is used to assign a truth value to the first component. They have fine-tuned BERT to improve classification performance and obtained an F1 score of 0.2549. Ludwig et al. (2022) proposed a metaheuristic feature selection algorithm based on human behavior on texts. They used SVM to show the effect of selected features on classification. They obtained an F1-score value of 0.251 as a result of the classification made for the purpose of obtaining higher performance by selecting some of the features. Schütz et al. (2022) used the monolingual and cross-lingual transformer model XLM-RoBERTa in their work. They achieved an F1 score of 15.48% in the monolingual study after classification. In the cross-lingual study, they achieved an F1 score of 19.46% in the classification of German texts.

When examining the existing studies, it is evident that there is no suitable evolutionary classification technique proposed for selecting the fundamental characteristics of fake news. This deficiency underscores the necessity for developing more effective solutions to cope with the increasing prevalence of fake news. Particularly, there is a need for new models that take into account the fundamental distribution of data. Therefore, this study aimed to achieve promising performance in detecting fake news by extracting features from fake news using tools such as Stylometry and CountVectorizer. Genetic Algorithm was employed for selecting these features. Subsequently, the selected features were classified using five different machine learning algorithms. Observations indicate that better performance was achieved compared to studies conducted with the same dataset in the literature.

2. Material and method

The aim of our study is to conduct an optimal classification analysis by applying effective feature extraction and feature selection techniques to a multi-class dataset containing fake news articles. The system design flowchart is presented in Figure 1, and the subsequent sections will discuss the various stages.

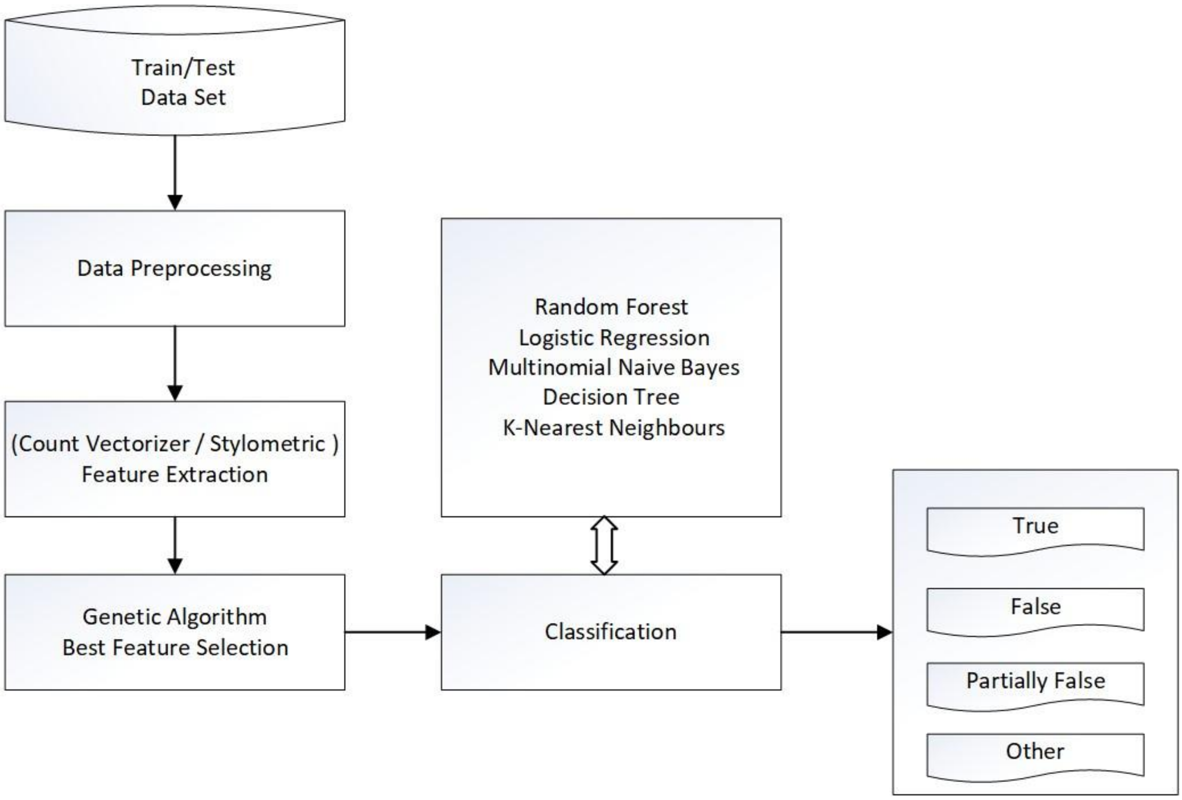

Figure 1. General flowchart of system design.

2.1. Dataset

In this study, A Multilingual dataset for Fake News Detection was used to detect fake news (Shahi et al., 2022). In the dataset, the statistical information of which is given in Table 1, there are a total of 1264 English news articles, including a training set of 900 lines and a development set of 394 lines. Each news article belongs to four classes: true, false, partially false, and other.

News articles have ID, text, title and our rating information. ID is a unique value that identifies news articles. Text contains the content information of news articles. Title is the title of the news article. Our rating indicates the trust rating and set of the news article, true, false, partially false, and others. The True and False classes are provable classes. Partially false class is specified as the class that cannot evaluate to 100% and contains true and false information. The other class is specified as the class that cannot be categorized as true, false, and partially false. When examining the percentages of data in the training set classes, there are a total of 1264 news articles, with 16.69% for the True class, 45.73% for the False class, 28.32% for the Partially False class, and 9.26% for the Other class.

The study aims to test the training set, which contains 1264 news articles, in two different ways: monolingual and cross-lingual. Statistical information from 612 English and 586 German news articles, as provided in Table 1, was utilized. The same training dataset was used for both monolingual and cross-lingual classification.

Table 1. Training and Test dataset statistics

Class	Train Data Count	Test Data Count English	Test Data Count German
True	211	210	243
False	578	315	191
Partially False	358	56	97
Other	117	31	55
Total	**1264**	**612**	**586**

2.2. Data preprocessing

In machine learning studies, one of the crucial steps that require emphasis is data preparation. In everyday life, during data collection, there can be instances of missing, incorrect, and contradictory data, which significantly impact the performance of the generated data. Data preprocessing plays a substantial role, especially in overcoming many challenges that may arise in the field of natural language processing. This contributes to obtaining more accurate and reliable results.

Various preprocessing steps have been applied to the text and headlines of the news articles used in our study. These preprocessing steps include converting uppercase letters to lowercase, removing numerical content, eliminating duplicate records, removing rows with empty values, cleaning URLs, removing HTML content, cleaning unprintable characters, cleaning repeated characters, and removing stop words. Additionally, punctuation cleansing and lemmatization have been performed, which we believe positively influence the performance of the classification process.

2.3. Feature extraction

When performing classification using machine learning algorithms, it is essential to transform the data into a format that the machine can comprehend. Textual data needs to be represented using numerical values for classification. This process is known as feature extraction. CountVectorizer is a method used to convert text into numerical data. It vectorizes the text based on the frequency of each word in the entire text. This means that every word in each text in the dataset is transformed into a vector. CountVectorizer represents each unique word in a column of the matrix, and the value in each cell denotes the frequency of the respective word in the corresponding text.

In the study, CountVectorizer was used for feature extraction from fake news text. When feature extraction is performed using CountVectorizer, terms are arranged based on their frequency. The top 5000 features, sorted by term frequency, were selected from the extracted features. Additionally, various stylometric features were extracted from the text and headline contents. Stylometry is a linguistic subfield that involves the statistical analysis of linguistically extracted features from text (Zheng et al., 2006). These features are commonly used in natural language processing, especially in text classification (Lagutina et al., 2019). Stylometry focuses on examining the linguistic features of content and is often employed to authenticate the originality or source of content based on linguistic processing style. Linguistic features help distinguish deceptive essence used to camouflage the author's unique composition style. With the assistance of stylometry, models can be trained to distinguish whether news articles are fake or real based on features extracted from the content of the written articles.

After a comprehensive examination of the news articles in the dataset, notable stylometric features were identified. A total of 26 stylometric features are listed in Table 2, with 14 related to text information, 11 related to title information, and 1 related to both title and text information in fake news articles.

Table 2. Stylometric features extracted on fake news texts

Feature	Text	Title
Length	✓	✓
Punctuation Count	✓	✓
Numeric Count	✓	✓
Http Count	✓	
Non Printable	✓	✓
Sent Count	✓	✓
Word Count	✓	✓
Char Count	✓	✓
Mentions Count	✓	
Caps Words Count	✓	
Avg Word Length	✓	✓
Propn Count	✓	✓
Noun Count	✓	✓
Punctuation Per	✓	✓
Title Text Ratio	✓	

2.4. Classification algorithms used in the study

The article uses various machine learning methods to accomplish the task of classifying and predicting fake news. The algorithms used are the result of a careful process that forms the methodology of the study. When determining the algorithms, the characteristics of the dataset were analyzed, literature review was conducted, and the problem context was carefully evaluated. In the initial stage, the size, distribution, feature type, and relationships of the dataset were carefully examined. Then, machine learning algorithms used in the literature were studied to understand the type of results these algorithms provide. Factors such as the performance, complexity, and interpretability of the examined algorithms in the literature were taken into account. The best algorithms serving the purposes of the study have been identified. Each of these algorithms contributes to the comprehensive analysis and identification of fake news articles within the dataset, bringing unique strengths and characteristics to the classification task. Machine learning methods used in the article for fake news classification and prediction are:

2.4.1. Random forest classifier (RFC)

The RFC algorithm, one of the machine learning algorithms, is a supervised algorithm that often yields excellent performance results and is easy to use. RFC is commonly used for both classification and regression tasks (Breiman., 2001). It aims to achieve more accurate and stable predictions by using multiple decision trees. Since it consists of a collection of decision trees, it generally performs better than single-tree classifications. The RFC algorithm considers the majority votes by looking at the predictions of multiple decision trees rather than relying on a single decision tree, as it contains a large number of decision trees in its subsets. It predicts the final output by examining the final votes, thereby achieving high accuracy and mitigating overfitting issues.

2.4.2. Logistic regression (LR)

LR, one of the supervised machine learning algorithms, is popularly used today. Its general purpose is to model the relationship between two or more dependent variables and independent variables in order to predict the outcome in the classification process. In LR, the dependent variable is categorical and measures its relationship with the independent variable (Choudhary & Jain., 2017).

2.4.3. Multinomial naive bayes (MNB)

MNB is a widely used and computationally efficient machine learning algorithm, primarily employed in text classification problems, especially in Natural Language Processing (NLP). It is based on Bayes' theorem and is designed to determine the frequency of a term, i.e., how often a term appears in a text (Mc- Callum et al., 1998). The specificity of a term in a document is what makes that term significant in the classification phase of the MNB model. In MNB, term frequency is crucial, so attention should be paid to the relationship between high-frequency terms and the text. Terms with high frequency but no semantic value in the text need preprocessing.

2.4.4. Decision tree (DT)

Tree-based algorithms are one of the most frequently used algorithms. It can be used to solve many problems such as classification and regression. DT is an algorithm used to determine the class of data whose class is unknown, based on the class of data. In DT, the node attribute represents the branch decision rules and the leaf represents the result. The top node is called the root node. The classification process takes place in the leaves. The results are stored in the branches of the tree. When classifying a sample, it is classified by starting from the root node and progressing down the tree branch by testing it according to the attributes in the nodes.

2.4.5. K-nearest neighbours (KNN)

The KNN algorithm is a machine learning algorithm that is frequently used in classification and regression problems. It focuses on keeping similar structures closest to each other. The KNN algorithm tries to predict the correct class by finding the similarity according to the distance between the training points and the data to be tested. K is the number of neighbors. It works by taking the closest K data as a reference while determining the class of new data.

2.5. Genetic algorithm based feature selection

To train a model created using machine learning algorithms and improve its learning, a large amount of data is collected. The degree to which the collected data is used in model training is a critical aspect. Some data may be useful, while others may be unnecessary. Unnecessary data can slow down model training, negatively impact model performance, and lead to incorrect results. The purpose of feature selection for machine learning algorithms is to retain useful data while removing unnecessary data. The primary goal is to create a subset of features from the existing features in the data without making any changes to the data itself. In this context, we aim to select the best feature subset using GA from the feature set extracted from English and German news texts.

GA is inspired by the idea of evolutionary mechanisms in nature, where the fittest survive. First introduced by John Holland in 1975, this theory draws inspiration from Darwin's theory of natural selection (Hayes-Roth., 1975). An intuitive algorithm, GA brings evolutionary processes into the realm of computer-based problem-solving. Following the logic of biological steps, it adheres to the principle of optimization. The general-purpose of GA is to find the most useful and suitable solution from within a problem-solving set. GAs mimic biological processes, such as selection, crossover, and mutation, to produce high-quality solutions. When examining the fundamental terminology of GAs, terms like Chromosome, Population, Fitness Function, Selection, Crossover, and Mutation are encountered.

- **Chromosome:** A collection of genes that are passed down from parents to offspring and to subsequent generations. A chromosome can be exemplified as a sequence of bits, with each bit representing a gene. It represents a candidate solution in GA.
- **Population:** A group of chromosomes. In GA, the population contains a series of possible solutions.
- **Fitness Function:** The stage where fitness is evaluated in each iteration. It determines the most fit individuals.
- **Selection:** The process of determining which individuals from the population will be selected and passed on to future generations based on their fitness values.
- **Crossover:** The stage where genetic exchange occurs between two selected individuals to create new offspring.
- **Mutation:** A random change process within a chromosome to produce better-optimized solutions.

The purpose of the GA used in our study is to determine the best subset of features from the features we generate for machine learning models and to better predict the target variable. Setting a user-defined parameter is important before predicting the target variable using GA. The optimal parameters were updated based on the success of experimental results. Among these parameters, the mutation rate was set to 0.01, and the population size and the number of iterations were both set to 100. The flowchart of the GA used in our study is shown in Figure 2.

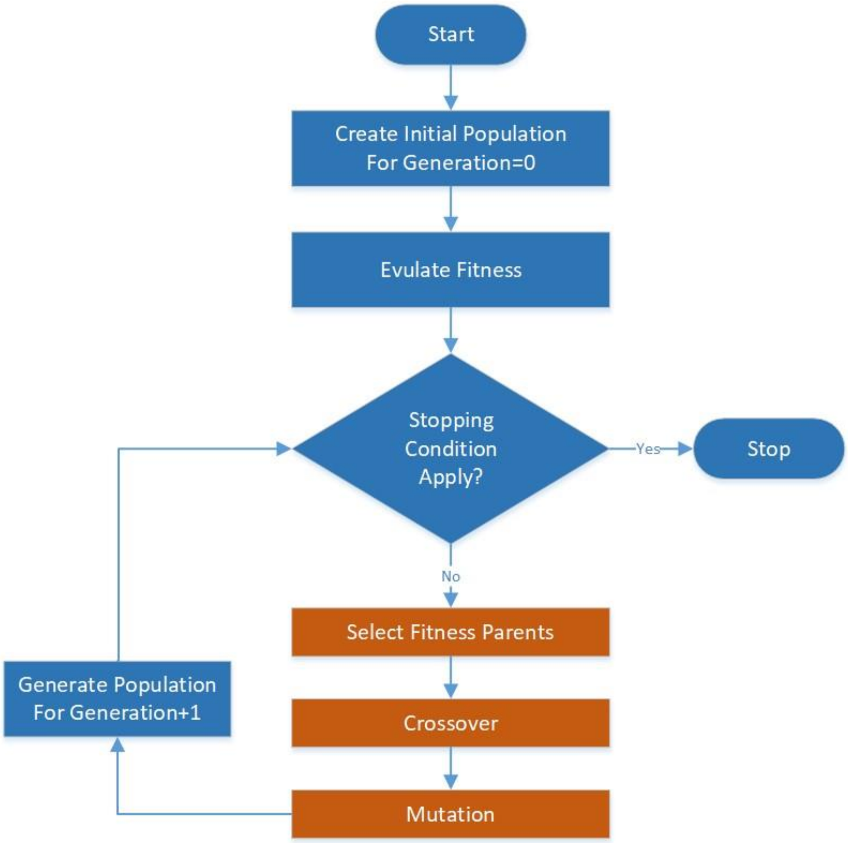

Figure 2. Basic flow chart of GA.

2.6. Performance metrics

The metrics used to determine the performance of machine learning models are given in Table 3.

Table 3. Performance Metrics

Metric	Mathematical Notation
Accuracy (ACC)	$ACC = \frac{TP + TN}{TP + TN + FN + FP}$ TP = True Positive, TN = True Negative, FN = False Negative, FP = False Positive.
Precision (P)	$P = \frac{TP}{TP + FP}$
Recall (R)	$R = \frac{TP}{TP + FN}$
F1-Score (F1)	$F1 = 2 * \frac{P * R}{P + R}$

3. Results

To detect and classify fake news, RFC, LR, MNB, DT, and KNN machine learning algorithms were used. In the classification process, 1264 English training data were used for training purposes. During testing, 612 English and 586 German news articles were used. Python's scikit-learn library was used for the classification process and performance evaluation. The experiments were conducted on a computer with an i7-11800H @2.30GHz processor, an RTX 3070 8GB graphics card, and 16GB of primary memory hardware. The development was carried out in Python 3.9, using the Jupyter Notebook development environment. The German data was translated into English for cross-lingual fake news detection using the Google API.

The goal was to find the best classification result using different machine learning methods. Stylometric features extracted from the text and title in the dataset were combined with CountVectorizer features with different maximum feature values and passed through a genetic algorithm (GA). In this way, the GA algorithm, inspired by the idea of evolution in nature, was used to select the best features. Feature sets of different sizes were classified five times by each machine learning algorithm. The experimental results showed that the best result was achieved with a maximum feature limit of 5000.

In the comparison of performance values obtained after the classification process, accuracy, Macro-F1 score, and class-specific F1 score were used as performance criteria. The results of the performance criteria for the classification process are presented in Table 4 for English news articles and in Table 5 for German news articles.

Table 4. F1 score, Accuracy and Macro-F1 score evaluation results by class for monolingual English news articles

Method	True Class F1Score	False Class F1Score	Partially False Class F1Score	Other Class F1Score	Accuracy	Macro-F1
RFC	0.11	0.70	0.18	0.00	0.5228	0.2495
KNN	0.20	0.60	0.13	0.21	0.4019	0.2847
DT	0.38	0.69	0.27	0.18	0.5228	0.3802
LR	0.40	0.72	0.25	0.23	0.5637	0.4016
MNB	0.41	0.72	0.30	0.29	0.5849	0.4297

Table 5. F1 score, Accuracy and Macro-F1 score evaluation results by class for cross-lingual English and German news articles

Method	True Class F1Score	False Class F1Score	Partially False Class F1Score	Other Class F1Score	Accuracy	Macro-F1
RFC	0.09	0.53	0.16	0.00	0.3515	0.1933
KNN	0.16	0.45	0.25	0.19	0.3122	0.2650
MNB	0.17	0.50	0.24	0.25	0.3412	0.2907
DT	0.42	0.50	0.24	0.30	0.4112	0.3646
LR	0.46	0.54	0.33	0.18	0.4539	0.3770

When Table 4 is examined, the most successful result for monolingual English news articles according to the Macro-F1 score is 42.97% with the MNB algorithm. When Table 5 is examined, according to the Macro-F1 score, the most successful result for cross-lingual English and German news articles was obtained with the LR algorithm, with a value of 37.70%. In light of these results, the confusion matrix obtained with the MNB algorithm for a monolingual as a result of the classification is indicated in Figure 3a the confusion matrix obtained with the LR algorithm as a result of the cross-lingual classification is shown in Figure 3b.

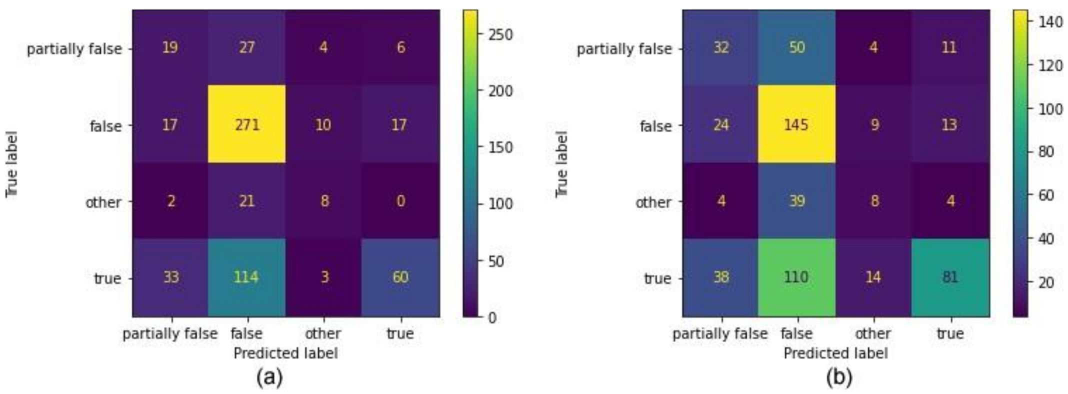

Figure 3. (a) Confusion matrix after classification of English news articles with MNB algorithm. (b) Confusion matrix of German and English news articles after classification with LR algorithm.

The robustness and performance of the system design indicated in Figure 1 have also been demonstrated by experiments on the "ISOT" dataset, created by Victoria University, in addition to the "A Multilingual dataset for Fake News Detection". The same steps of the system design indicated in Figure 1 were applied to the ISOT dataset, which consists of 23,481 fake news and 21,417 real news (ISOT, 2022). Based on the results we obtained, the classification process was performed using LR, MNB, and DT algorithms, which provided the best performance. In contrast to the previous dataset, the ISOT dataset includes both text and class attributes, so only 7 Stylometric features were extracted from the text. These Stylometric features include Length, Number of Https, Word Count, Character Count, Average Word Length, Number of Proper Nouns, and Number of Nouns. The number of CountVectorizer features is limited to 1000. As a result of the classification with 1007 features, the results specified in Table 6 were obtained. Within the framework of the methodology we used in our study, F1 Scores of 99.48% and 99.62% were achieved with LR and DT algorithms, respectively.

Table 6. ISOT dataset classification results

Method	Precision	Accuracy	Macro-F1
MNB	0.9601	0.9600	0.9600
LR	0.9948	0.9948	0.9948
DT	0.9962	0.9962	0.9962

4. Discussion

In recent years, there has been a growing body of research on fake news detection. When examining these studies, it becomes evident that people are increasingly exposed to fake news, and they often struggle to identify them. The rising number of such news articles is seen to significantly impact economic and social life. In this context, the presence of automatic detection systems becomes crucial.

This study addresses the classification problem of single-language and cross-language fake news detection. Machine learning algorithms, including RFC, LR, MNB, DT, and KNN, were employed for the detection and classification of fake news. The classification process utilized 1264 English training data, 684 English test data, and 586 German test data. The German data was translated into English using Google API. Feature extraction was performed on the text and headlines of the dataset using CountVectorizer and Stylometric features. The obtained features were processed through GA (Genetic Algorithm) to select the best features for classification. As a result, for the single-language case, MNB achieved an accuracy of 58.49% and a macro F1 score of 42.97%, while for cross-language detection, LR achieved an accuracy of 45.39% and a macro F1 score of 37.70%. The ISOT dataset was also used to validate the methodology outlined in the system design, and successful results were obtained.

Comparative studies using the same dataset are presented in Tables 7 and 8.

Table 7. English: Performance comparison of similar studies in the literature according to the macro-F1 score

Author	Method	True	False	Partially False	Other	Accuracy	Macro-F1
(Tran et al., 2022)	T5-3B	-	-	-	-	-	0.3954
(Taboubi et al., 2022)	BERT	0.383	0.721	0.173	0.080	0.5470	0.3390
(Martinez-Rico et al., 2022)	Longformer	-	-	-	-	0.5410	0.3324
(Truică et al., 2022)	BiLSTM	0.328	0.744	0.185	0.035	0.5310	0.3230
(Pritzkau et al., 2022)	RoBERTa & Longformer	0.339	0.707	0.184	0.000	0.5131	0.3076
(Althabiti et al., 2022)	Bert-Large-Cased	-	-	-	-	0.5260	0.3050
(LekshmiAmmal et al., 2022)	Transformer based	-	-	-	-	-	0.2980
(Kumar et al., 2022)	RoBERT	0.276	0.619	0.137	0.155	0.4420	0.2960
(Porto-Capetillo et al., 2022)	BERT & Stylometric	-	-	-	-	0.5458	0.2951
(Arif et al., 2022)	RoBERTa	-	-	-	-	0.4700	0.2860
(La Barbera et al., 2022)	BERT	0.207	0.694	0.140	0.062	0.4720	0.2750
(Blanc et al., 2022)	BERT	0.126	0.661	0.202	0.028	0.4444	0.2549
(Ludwig et al., 2022)	SVM	0.141	0.670	0.169	0.022	0.4620	0.2510
(Schütz et al., 2022)	XLMRoBERTa	0.280	0.146	0.153	0.392	0.1993	0.1548
Our study	**MNB**	**0.410**	**0.720**	**0.300**	**0.290**	**0.5849**	**0.4297**

When examining studies conducted on English news using the same dataset, Tran et al. (2022) achieved the best result with a 39.54% F1 score. Comparing our results, it is observed that our study obtained a 3.43% higher macro F1 score compared to this study. Similarly, when studies conducted on German news using the same dataset were examined, Tran et al. (2022) achieved the best result with a 30.06% F1 score. Comparing our results, it is observed that our study obtained a 7.64% higher macro F1 score compared to this study.

Table 8. German: Performance comparison of similar studies in the literature according to the macro-F1 score

Author	Method	True	False	Partially False	Other	Accuracy	Macro-F1
(Tran et al., 2022)	XLM-RLarge	-	-	-	-	-	0.3006
(LekshmiAmmal et al., 2022)	Transformer based	-	-	-	-	-	0.2245
(Schütz et al., 2022)	XLMRoBERTa	0.378	0.168	0.151	0.808	0.2542	0.1946
(Truică et al., 2022)	BiLSTM+XML	0.098	0.452	0.193	0.000	0.2832	0.1859
(Arif et al., 2022)	BiLSTM	-	-	-	-	0.2800	0.1721
Our study	**LR**	**0.460**	**0.540**	**0.330**	**0.180**	**0.4539**	**0.3770**

5. Conclusions

Fake news detection is a challenging task in the field of text classification. Various researchers have made numerous attempts in this regard. This study presents a comparative analysis of news classification into four different classes: true, false, partially false, and other, using five different machine learning algorithms. CountVectorizer was combined with feature extraction and stylometric feature extraction for effective feature extraction. Genetic algorithms were used to select effective features. The classification results of the employed machine learning methods are presented comparatively. The study was compared with other studies using the same dataset, and successful results were achieved. The same methodology was also tested on the ISOT dataset, confirming the robustness of our approach.

In the future, data augmentation on the dataset and the use of different meta-heuristic algorithms and deep learning algorithms are planned to achieve better results.

Acknowledgement

We sincerely thank the referees and journal editors for their careful review and valuable contributions.

Author contribution

All authors had the same contribution. All authors read and approved the final manuscript.

Declaration of ethical code

The authors of this article declare that the materials and methods used in this study do not require ethical committee approval and/or legal-specific permission.

Conflicts of interest

The authors declare that they have no conflict of interest.

References

Aborisade, O., & Anwar, M. (2018). Classification for authorship of tweets by comparing logistic regression and naive bayes classifiers. *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, 269–276. https://doi.org/10.1109/IRI.2018.00049.

Ahmad, I., Yousaf, M., Yousaf, S., & Ahmad, M. O. (2020). Fake news detection using machine learning ensemble methods. *Complexity*, *2020*, 1–11. https://doi.org/10.1155/2020/8885861.

Ahmed, H., Traore, I., & Saad, S. (2017). Detection of online fake news using n-gram analysis and machine learning techniques. *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments: First International Conference, ISDDC 2017, Vancouver, BC, Canada, October 26-28, 2017, Proceedings 1*, 127–138. https://doi.org/10.1007/978-3-319-69155-8_9.

Althabiti, S., Alsalka, M. A., & Atwell, E. (2022). SCUoL at CheckThat! 2022: fake news detection using transformer-based models. *CEUR Workshop Proceedings*, *3180*, 428–433.

Arif, M., Tonja, A. L., Ameer, I., Kolesnikova, O., Gelbukh, A., Sidorov, G., & Meque, A. G. M. (2022). CIC at CheckThat! 2022: multi-class and cross-lingual fake news detection. *Working Notes of CLEF*.

Blanc, O., Pritzkau, A., Schade, U., & Geierhos, M. (2022). CODE at CheckThat! 2022: multi-class fake news detection of news articles with BERT. *Working Notes of CLEF*.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32. https://doi.org/10.1023/a:1010933404324.

Choudhary, N., & Jain, A. K. (2017). Towards filtering of SMS spam messages using machine learning based technique. *Advanced Informatics for Computing Research: First International Conference, ICAICR 2017, Jalandhar, India, March 17--18, 2017, Revised Selected Papers*, 18–30. https://doi.org/10.1007/978-981-10-5780-9_2.

DiFranzo, D., & Gloria-Garcia, K. (2017). Filter bubbles and fake news. *XRDS: Crossroads, The ACM Magazine for Students*, *23*(3), 32–35. https://doi.org/10.1145/3055153.

Fayaz, M., Khan, A., Bilal, M., & Khan, S. U. (2022). Machine learning for fake news classification with optimal feature selection. *Soft Computing*, *26*(16), 7763–7771. https://doi.org/10.1007/s00500-022-06773-x.

Ghayoomi, M., & Mousavian, M. (2022). Deep transfer learning for COVID-19 fake news detection in Persian. *Expert Systems*, *39*(8), e13008. https://doi.org/10.1111/exsy.13008.

Hayes-Roth, F. (1975). Review of" Adaptation in Natural and Artificial Systems by John H. Holland", The U. of Michigan Press, 1975. *ACM SIGART Bulletin*, *53*, 15–15.

ISOT Fake News Dataset. (2023, February 10). https://onlineacademiccommunity.uvic.ca/isot/2022/11/27/fake-news-detection-datasets/

Jain, A., & Kasbe, A. (2018). Fake News Detection. *2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*, 1–5. https://doi.org/10.1109/SCEECS.2018.8546944.

Jose, X., Kumar, S. D. M., & Chandran, P. (2021). Characterization, Classification and Detection of Fake News in Online Social Media Networks. *2021 IEEE Mysore Sub Section International Conference (MysuruCon)*, 759–765. https://doi.org/10.1109/MysuruCon52639.2021.9641517.

Kumar, S., Kumar, G., & Singh, S. R. (2022). TextMinor at CheckThat! 2022: fake news article detection using RoBERT. *Working Notes of CLEF*.

La Barbera, D., Roitero, K., Mackenzie, J., Damiano, S., Demartini, G., & Mizzaro, S. (2022). BUM at CheckThat! 2022: a composite deep learning approach to fake news detection using evidence retrieval. *Working Notes of CLEF*.

Lagutina, K., Lagutina, N., Boychuk, E., Vorontsova, I., Shliakhtina, E., Belyaeva, O., Paramonov, I., & Demidov, P. G. (2019). A survey on stylometric text features. *2019 25th Conference of Open Innovations Association (FRUCT)*, 184–195. https://doi.org/10.23919/FRUCT48121.2019.8981504.

LekshmiAmmal, H. R., & Madasamy, A. K. (2022). *NITK-IT NLP at CheckThat! 2022: Window based approach for Fake News Detection using transformers*.

Lima, G. B., Chaves, T. de M., Freitas, W. W. L., & de Souza, R. M. (2022). Statistical learning from Brazilian fake news. *Expert Systems*, e13171. https://doi.org/10.1111/exsy.13171.

Ludwig, A., Felser, J., Xi, J., Labudde, D., & Spranger, M. (2022). FoSIL at CheckThat! 2022: using human behaviour-based optimization for text classification. *Working Notes of CLEF*.

Martinez-Rico, J. R., Martinez-Romo, J., & Araujo, L. (2022). NLP &IRUNED at CheckThat! 2022: ensemble of classifiers for fake news detection. *Working Notes of CLEF*.

McCallum, A., Nigam, K., & Others. (1998). A comparison of event models for naive bayes text classification. *AAAI-98 Workshop on Learning for Text Categorization*, *752*, 41–48.

Nasir, J. A., Khan, O. S., & Varlamis, I. (2021). Fake news detection: A hybrid CNN-RNN based deep learning approach. *International Journal of Information Management Data Insights*, *1*(1), 100007. https://doi.org/10.1016/j.jjimei.2020.100007.

Porto-Capetillo, C., Lecuona-Gómez, D., Gómez-Adorno, H., Arroyo-Fernández, I., & Neri-Chávez, J. (2022). *HBDCI at CheckThat! 2022: Fake News Detection Using a Combination of stylometric Features and Deep Learning*.

Pritzkau, A., Blanc, O., Geierhos, M., & Schade, U. (2022). NLytics at CheckThat! 2022: hierarchical multi-class fake news detection of news articles exploiting the topic structure. *Working Notes of CLEF*.

Schütz, M., Böck, J., Andresel, M., Kirchknopf, A., Liakhovets, D., Slijepčević, D., & Schindler, A. (2022). AIT FHSTP at CheckThat! 2022: cross-lingual fake news detection with a large pre-trained transformer. *Working Notes of CLEF*.

Shahi, G. K., Struß, J. M., Mandl, T., Köhler, J., Wiegand, M., & Siegel, M. (2022, May 16). *CT-FAN: A Multilingual dataset for Fake News Detection*. Zenodo. https://zenodo.org/records/6555293

Taboubi, B., Nessir, M. A. B., & Haddad, H. (2022). iCompass at CheckThat! 2022: combining deep language models for fake news detection. *Working Notes of CLEF*.

Tran, H. N., & Kruschwitz, U. (2022). ur-iw-hnt at CheckThat! 2022: cross-lingual text summarization for fake news detection. *Working Notes of CLEF*.

Truică, C.-O., Apostol, E.-S., & Paschke, A. (2022). Awakened at CheckThat! 2022: Fake news detection using BiLSTM and sentence transformer. *Working Notes of CLEF*.

Vogel, I., & Jiang, P. (2019). Fake news detection with the new German dataset "GermanFakeNC." *Digital Libraries for Open Knowledge: 23rd International Conference on Theory and Practice of Digital Libraries, TPDL 2019, Oslo, Norway, September 9-12, 2019, Proceedings 23*, 288–295. https://doi.org/10.1007/978-3-030-30760-8_25.

Zhang, D., Xu, J., Zadorozhny, V., & Grant, J. (2022). Fake news detection based on statement conflict. *Journal of Intelligent Information Systems*, *59*(1), 173–192. https://doi.org/10.1007/s10844-021-00678-1.

Zheng, R., Li, J., Chen, H., & Huang, Z. (2006). A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American Society for Information Science and Technology*, *57*(3), 378–393. https://doi.org/10.1002/asi.20316.