Open Peer Review on Qeios

Automatic Content Analysis Systems: Detecting Disinformation in Social Networks

Roman Romanchuk[1], Victoria Vysotska[1]

1 Lviv Polytechnic National University

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.

Abstract

In the 21st century, the rapid rise of disinformation and propaganda has become a significant global issue, undermining democratic processes and socio-political institutions. Disinformation, defined as intentionally false or misleading information, aims to manipulate public opinion and cause economic harm. This paper explores the use of computational linguistics and machine learning methods to detect disinformation. Techniques such as text preprocessing, feature extraction, and classification algorithms (e.g., SVM, naive Bayes) are adapted for identifying fake news. Recent studies demonstrate the effectiveness of these methods in social media and news platforms, highlighting the importance of advanced models like GPT-4 in improving detection accuracy and combating the spread of disinformation.

Roman Romanchuk[a], and **Victoria Vysotska**[b]

Lviv Polytechnic National University, Lviv, Ukraine

[a] ORCID iD: 0009-0004-4352-1073

[b] ORCID iD: 0000-0001-6417-3689

Keywords: propaganda, computational linguistics, machine learning, disinformation, fake, disinformation detection, SVM, naive Bayes, identifying fake news, GPT-4.

Introduction

The modern development of computer technologies and the global Internet network increases their role in people's lives every day. The large-scale implementation of simple, fast, and primarily free online services forms the basis of communication for significant parts of society, its access to information, the actualization of economic interests, and the political positions of citizens [1].

Modern Internet technologies have significantly changed not only the methods of news distribution but also the methods and means of news creation. Various social networks and messaging services (messengers) increasingly serve as the main source of information for people. The lack of availability of information hygiene skills, along with the ease of creating and distributing news, has dramatically increased the spread and use of disinformation.

According to the Cambridge Dictionary, disinformation is the dissemination of false information with the aim of misleading people. Representatives of international organizations such as the UN, OSCE, and others emphasized in their Joint Declaration that the purpose of disinformation is to distort or present false, provocative information as true to mislead the public, political opponents, and competitors. In his report dated April 13, 2021, the UN Special Rapporteur on the promotion and protection of freedom of opinion and expression emphasized that disinformation is false information intentionally disseminated to cause serious social harm. According to the Code of Practice of the European Union on countering disinformation, this type of information is aimed at obtaining economic benefits for the distributor [2].

The results of sociological research conducted in Germany[3] show that 59% of Internet users have come across fake news. More than 80% of survey participants agreed that fake news is a threat, particularly to democracy.

Due to the simplification of the processes of creating and distributing news via the Internet, as well as the physical impossibility of checking the large volumes of information circulating on the network, the volume of disinformation and fake news distribution has increased significantly.

In this regard, the detection of fake news becomes a critical task. This not only ensures the provision of verified and reliable information to users but also helps prevent the manipulation of public consciousness. Strengthening control over the credibility of news is important for maintaining a reliable information ecosystem.

Various methods are used to combat the spread of disinformation, including checking sources of information and facts, evaluating the authority and reliability of the sources, carefully studying the context of the information and other techniques. Additionally, the use of information technologies, including artificial intelligence, to combat the spread of disinformation is becoming increasingly popular.

Artificial intelligence plays an important role in the detection and analysis of misinformation in the educational environment, thanks to its machine learning algorithms that allow for the quick and efficient analysis of large volumes of information. Automated systems can check the facts, plausibility of sources, and context of information, which helps identify potential cases of misinformation. This increases information literacy among pupils and students and helps teachers and educational institutions filter out false or manipulative information from educational materials.

Machine learning and big data analysis algorithms can be useful for pattern recognition and the detection of false information. Pattern recognition is the process of identifying recurring or typical structures characteristic of a particular set of data or context. In the context of disinformation analysis and the use of artificial intelligence technologies, this term refers to the identification of specific patterns or deviations that may indicate a disinformation campaign. For example, pattern recognition may include analyzing the structure of textual information, identifying keywords or phrases, establishing connections between different sources, and analyzing rhetoric. This helps identify specific signs that may indicate inaccurate information or attempts to manipulate the audience [4].

The use of machine learning algorithms is becoming an increasingly effective method for detecting and classifying materials containing misinformation in educational and scientific resources. Such algorithms analyze large and ultra-large volumes of data and can identify patterns indicative of disinformation. As a result, it is possible to build automated systems for detecting suspicious materials. This creates a safer and more favourable environment for educational and scientific processes.

Related works

In the 21st century, information has become a key tool for national governance, influencing non-state actors and changing the dynamics of influence. Recently, there has been a rapid development of disinformation and propaganda worldwide. Adversaries use propaganda and disinformation to manipulate public opinion, destroy socio-political institutions, and weaken democracy. A deep understanding of the information environment, its impact on geopolitical events, and people's behaviour is becoming critical for maintaining security and making important political decisions. The main challenge for modern society is the protection of democratic processes and individuals from the spread of misinformation and propaganda.

Disinformation is the dissemination of false or misleading information with the intent to mislead[5]. The main purposes of disinformation are to cause economic damage, manipulate public opinion, or make a profit. In modern times, misinformation is often accompanied by falsified, out-of-context, and manipulated images or videos. The main channels for spreading misinformation include Internet forums, news sites, and social networks [6].

It is important to note that the European Council first recognized the threat of disinformation in 2015 when Russia increased its efforts in this direction within the European Union. Disinformation campaigns, especially those conducted by third countries, often become components of hybrid warfare, including cyberattacks and hacking networks.

In the "EU Action Plan on Disinformation" from 2018, it is stated that disinformation is an integral part of hybrid campaigns aimed at dividing European society and undermining citizens' trust in democratic processes and EU institutions. The authors of the study "Information Disorder: Towards an Interdisciplinary Framework for Research and Policy-Making," published in 2017 with the support of the Council of Europe, are Claire Wardle and Hossein Derakhshan. According to this study, information disorder encompasses three main concepts: false information, misinformation, and harmful information.

Misinformation is information that is inaccurate but created without the intent to cause harm. This type of information often results from inadvertent errors, incorrect photo or video captions, incorrect dates, or incorrect statistics.

Disinformation is intentionally created false information to harm a person, social group, organization, or country. This may include fabricated or purposefully altered audiovisual content, as well as the creation of various conspiracy theories or rumours. Harmful information (malinformation) is information that is based on real facts but is used to harm a person, organization, or country. This includes intentionally disclosing personal information for personal gain and purposefully changing the context of posts.

In the European Union, there is a problem with agreeing on common terminology such as "disinformation," "propaganda," and "fake news," due to the lack of uniform definitions used in the EU. The term "fake news" can have different interpretations, but the Cambridge Dictionary provides a clear one: "false stories that, masquerading as news reports, are distributed on the Internet or used in other media, usually to influence political beliefs or as humour."

The High-Level Group on Fake News and Online Disinformation provided the first formal attempt to define the term "disinformation" in its report, "A Multidimensional Approach to Disinformation," which was released in 2018 [7]. It is noted that the problems of disinformation are interconnected with the development of digital media and are managed by various actors, such as state and non-state organizations, commercial structures, mass media, and citizens. These problems also have manipulative manifestations in communication mechanisms.

Methods and tools

To solve the problem of identifying and combating the spread of fake information, it is important to understand the principles by which it is spread. In this context, computational linguistics is used to uncover rules and patterns by which disinformation elements can be identified in a stream of textual data. Studies [8] indicate that a certain set of rules are used to create fake information that makes the news believable.

False content has specific characteristics, such as the use of abbreviations, the transmission of a limited amount of information and the presence of a negative character. The methods of computer linguistics allow analyzing these features, contributing to the detection and identification of potentially fake information in text sources.

The problem of detecting fake sources of information can be similar to the tasks of detecting spam, especially when using statistical methods of machine learning. Machine learning techniques used to classify text such as tweets or emails to determine whether it is spam are already successfully used to detect spam.

These machine learning techniques can include pre-processing the text, identifying features (for example, using a "bag of words"), and cutting out unnecessary features to improve accuracy on the test data set. The principle of operation consists of using a training data set on which the model is "learned" to recognize signs of spam, and then using this model to classify new text as spam or not spam.

This approach can be adapted to detect fake sources of information, using a suitable training data set and developing models to identify characteristic features of fakes.

Once the features are defined, they can be used for classification using various machine-learning techniques that involve training with a teacher. In particular, methods such as the naive Bayesian classifier, the support vector method, TF-IDF, or the K-nearest neighbour method can be used.

The task of disinformation detection is similar to the task of spam detection, as both aim to separate genuine texts from fake or false ones. However, it is important to consider these tasks separately due to several differences.

To detect elements of disinformation, similar techniques to those used in spam detection are applied. Part-of-speech (POS) tagging is the process of tagging words in text according to their parts of speech (such as nouns, verbs, adjectives, etc.). This process is an important step in many natural language processing tasks, such as parsing, machine translation, and semantic analysis. The use of POS tagging helps machines more accurately understand the structure and meaning of sentences, thereby improving the quality and accuracy of text processing.

A Word N-gram is a model based on sequences of N words in a text. This technique is used in natural language processing to analyze the context and dependencies between words. N-grams help improve text comprehension by facilitating more accurate translations, creating speech patterns, and recognizing speech patterns. They are often used in next-word prediction and automatic text generation tasks, providing the basis for many modern word processing systems.

A Char N-gram is a model that uses sequences of N characters in text for analysis. It allows the detection of patterns and dependencies between symbols in speech, which can be useful in natural language processing tasks such as text classification, speech recognition, language modelling, and spam filtering. Thanks to Char N-grams, you can effectively analyze text regardless of its language and vocabulary, making it a universal tool for processing text information.

Word2Vec is an algorithm for vectorizing words in text, which converts words into vectors of numbers in space. It uses neural networks to train word representations such that semantically related words are close to each other in the vector space. This method allows you to apply mathematical operations to work with words, such as subtracting or adding vectors, to reproduce the semantic relations between them. Word2Vec is widely used in various natural language processing tasks such as text clustering, machine translation, sentiment analysis, and many others.

BERT (Bidirectional Encoder Representations from Transformers) is a model for word representations based on transformers, which is capable of achieving impressive results in various natural language processing tasks. It uses contextual vectors of words, which allows taking into account dependencies between words in a sentence in both directions. BERT is trained on a large text corpus by training on two tasks: predicting the next word in a sentence and predicting whether a sentence from a set of text is consecutive. The model can be refined to perform various natural language processing tasks by fine-tuning specific data for a specific task. BERT is one of the most effective and popular models in the field of natural language processing, due to its ability to understand words in context and high accuracy in various tasks.

ELMo (Embeddings from Language Models) is a method that uses contextual word vectors trained on a large text corpus to represent words in sentences. It is based on LSTM (Long Short-Term Memory) technology and uses a context transfer mechanism to ensure that each word has a vector representation that takes into account its meaning in a specific context. This approach makes it possible to better take into account the semantic connections between words in a sentence. ELMo can be used for various natural language processing tasks, such as text classification, machine translation, named entity recognition, and others, thanks to its ability to contextually understand words in the text.

TF-IDF (Term Frequency-Inverse Document Frequency) is a statistical method used to determine the importance of terms in documents relative to the corpus of texts. TF-IDF is calculated by multiplying two values: TF, which represents the frequency of a term in a document, and IDF, which represents the inverse frequency of the document containing the term in the entire corpus. A high TF-IDF significance indicates that a term occurs frequently in a particular document but rarely occurs in other documents in the corpus, making it important for that particular document. TF-IDF is often used for the weighted extraction of important words in documents for various natural language processing tasks such as text classification, clustering, indexing, etc.

Bag-of-Words (BoW) is a simple but powerful text vectorization method used in natural language processing. In this method, each sentence or document is represented as a vector, where each value of this vector corresponds to the number of occurrences of each word from a predefined dictionary in the document. BoW is used to create a "bag of words" by ignoring the order of words in a document and focusing only on their frequency. This method is often used to build models for text classification, clustering, and sentiment analysis.

Machine learning techniques such as Naive Bayes, Support Vector Machines, TF-IDF or K-Nearest Neighbors can be applied to analyze text and classify it as potentially misleading. Feature analysis, "bag of words" detection and clipping of unnecessary features can also be used to improve detection accuracy.

It is important to consider that detecting elements of disinformation in the text is a more difficult task compared to detecting spam. This can include subtle nuances, such as replacing a person's name, which can change the meaning of the text. Therefore, the model needs to be trained based on a wide range of other news and articles on a specific topic to achieve high accuracy in identifying misinformation.

In addition, tracking the spread of fake news can be an effective tool for vetting stories and determining their credibility through cross-checking and fact-checking.

Looking at words in isolation may not be enough to effectively predict fake news. Therefore, some scholars use other linguistic strategies, including analyzing the syntax and grammar of the language.

One such approach involves the use of probabilistic context-free grammars (PCFGs). PCFG is used to transform sentences into parse trees that describe the structure of the sentences and transform them into a recursive syntactic structure. The parsing that PCFG can use is widely used to analyze the mood of a text (semantic parsing). This approach can help detect syntactic anomalies that may indicate fake or manipulative information.

Text parsing leads to its transformation into a data structure, which is usually presented in the form of a tree. This tree displays the syntactic structure of the input sequence, providing a clear visual representation of the syntactic relationships between words and phrases. This tree can be used for further text processing. The process of parsing is usually considered as two-stage:

- Lexical analysis (Tokenization): At this stage, the text is divided into tokens, which are the basic units of syntactic analysis. Lexical analysis determines which words or symbols make up tokens and how they are related to each other.

- Creation of a parsing tree (Parsing): At this stage, the rules of language grammar are used to create a data structure in the form of a tree that reflects the syntactic structure of the text. A parse tree allows you to determine the relationships between words and phrases in a sentence.

In the context of detecting fake news or disinformation, parsing can help detect anomalies in text structure that may be characteristic of manipulative information.

Also, analyzing comments and comparing them with similar articles can be a useful method to assess the veracity of news or text. If the majority of such articles do not corroborate the data in the news, this may indicate that the news may be biased or fake.

Analysis of comments can also serve as an indicator. If comments indicate doubt, lack of confirmation, or question the authenticity of the article, this may indicate the possibility of false information.

However, it is important to be careful when using this method, as some fake news can be widely distributed through social media or other channels, and they can have a significant volume of comments from people who support them. Also, consider the possibility of creating artificial comments to support fake news. However, comments can serve as an important source of additional information when assessing the credibility of content.

Manual and automatic verification methods are also used to find disinformation.

Yes, manual fact-checking can be divided into two main categories: expert fact-checking and crowd-sourced fact-checking.

Expert fact-checking:

- Carried out by experts who have specialized knowledge in a specific area or subject.
- Usually used to test a limited amount of data or specific facts.
- Has high accuracy, but can be expensive and not scalable when processing a large amount of information.

Crowdsourced fact-checking:

- Involves a wide range of users or enthusiasts to fact-check or verify information.
- Is more scalable and able to process large amounts of data.
- Usually less accurate compared to peer review because it involves different levels of expertise.

Each of these methods has its advantages and limitations. The choice of a particular method may depend on the volume of information to be verified and the availability of resources to involve experts or a wide range of users.

Often, given the volume and speed of news on social media, manual fact-checking becomes impractical to effectively detect and respond to misinformation. That is why developers use automatic fact-checking methods based on information retrieval (IR) and natural language processing (NLP) technologies.

Information retrieval (IR):

- Uses search methods to find information in textual sources.
- Search algorithms allow you to quickly highlight and extract relevant facts from various sources.

Natural Language Processing (NLP):

- Used to understand and process human speech.
- NLP algorithms can analyze text, recognize semantics and discover key facts and aspects.

The combination of IR and NLP allows systems to automatically analyze and track information to detect potential misinformation or fake news. It is also important to consider context, source of information, and other factors to accurately determine credibility. Such automatic methods can help in the real-time detection and resolution of problems related to the spread of misinformation in social networks.

Due to the widespread dissemination of disinformation campaigns, an increasing number of researchers are endeavouring to address the challenges associated with combating them. In particular, in [9], the authors provide a general overview of "false news" (fake news). Additionally, emerging studies focus on the rapidly growing imbalance between true and false news in social networks [10][11], fact-checking during political debates[12][13], fact-checking in community Q&A forums[14], and other related areas.

In their publication [15], the authors propose categorizing fake news into clickbait, influential, and satirical types. Techniques such as spam detection, positioning, and benchmarking have been employed to combat the spread of fake news. Furthermore, the authors delve into sentiment analysis, which is a component of natural language processing methods. One notable example of fake news discussed is the story of China's Airport Security Robot "Electroshok," which surfaced in 2016 and led to over 12,000 fake news stories in China, disseminated across 244 different websites as sources.

Recent research exploring issues of bias and misinformation has increasingly focused on social networks and other specific online platforms. In publications [16][17], the authors investigate the credibility and bias of messages on Twitter (now X) and methods to counter toxic comments on various forums [18]. Concurrently, studies examine the veracity of articles in digital and print media [19], as well as the credibility of news sources for the media[20].

In terms of feature extraction, most previous studies rely on stylistic and linguistic features that may be independent of the topic or genre. In other words, regardless of whether the event is mentioned in a specific news article or is biased, the features should contain information that helps the ML model make the right decisions. This is a basic design principle used in user profiling to check whether questionable content is written by the same user who publishes a wide range of similar content [21].

Different features have varying effects on the final result. For example, the length of the content and the specifics of the topic play a minor role in solving such problems. In contrast, character N-grams (char N-grams) have a significant impact on the final result [22]. This type of trait has been observed to be more influential and significant in determining advocacy and bias levels. Previous research on bias detection and advocacy has also investigated this type of trait. To test the credibility of news claims, the authors [23] used stylistic features that include the occurrence of assertive, factual verbs, as well as indirect speech verbs, mitigating and implicative words, and discourse markers that were manually extracted using manually created lexicons.

In the study [24], the authors drew attention to the significant increase in the influence of fake news on everyday life. They analyzed three methods for identifying fake news: naive Bayes, neural networks, and the support vector method. According to their study, the naive Bayesian method had an accuracy of 96.08% in detecting fake news, while the support vector method showed an accuracy of 99.90%.

To categorically identify biased opinions that give answers as "exactly yes" or "exactly no," it is useful to pay attention to stylometric characteristics. For example, in the study [25], the authors investigated the differences between true news, satirical news, and fake news. They compiled a dataset consisting of news articles collected from nine different web sources, manually checked by journalists. For identification, they utilized a stylometric technique developed [26] for verifying authorship to predict factuality and the degree of bias. They hypothesized that biased content exhibits a typical writing style. Another case study developed a model for detecting fake news [27]. Their conclusion demonstrated that the use of proper names and the structure of headlines are crucial characteristics for distinguishing fake news from real news. Another method was presented to distinguish between real, fake, and satirical news using indicators of writing style and complexity. Among the features were various part-of-speech tags, swear words, slang words, stop words, punctuation, and negation. For complexity measures, they employed different readability indices and concluded that "fake news" had low readability scores, shorter length, plain language usage, and fewer technical words. Additionally, they found that fake news is more closely aligned with satire. Initially, more than 130 characteristics were selected, but most of them were not deemed useful.

In [28], the author team presented a model that utilizes a support vector machine (SVM) algorithm to collect articles and determine their authenticity. Using the SVM algorithm for binary classification, the model organizes the articles and categorizes them as genuine or fake. The developed models consist of three main components: an aggregator for collecting articles, an authenticator for verifying their authenticity, and a recommendation system. To further validate the veracity of the articles, the team employed a naive Bayesian algorithm, which, combined with SVM and NLP, achieved an accuracy of 93.50% in news classification.

To classify news articles into four categories (including genuine, satirical, hoax, and propaganda), researchers [29] constructed a corpus of news articles sourced from the English Gigaword and seven different news sources. They concluded that the use of verbal N-grams impaired performance when tested on unknown sources. In 2017, a binary propaganda classification model was developed [30] using a publicly available dataset containing information from 15,000 bots. News articles were tagged based on published content, which could have been created by bots or humans. Their model achieved a 95% AUC using the Random Forest method. Later, researchers [31] developed an identification model using 130 features based on content collected from literary sources.

In the study [32], the authors analyzed methods of detecting fake news based on Twitter posts. They focused on determining whether posts were real or fake, specifically examining incidents like the 2010 earthquake in Chile and the US presidential election. The primary fake news detection method they proposed was based on natural language processing, involving classifying the news before applying various machine learning models to produce the results. The authors paid attention to improving the effectiveness of detecting fake news by incorporating word length into their methods. The study utilized five different machine learning algorithms: Naive Bayesian, logistic regression, support vector method (SVM), recurrent neural networks, and long short-term memory. The authors introduced four types of feature vectors: count vectors, word-level vectors, N-gram vectors, and character-level vectors, among which the support vector method proved to be the most effective for identifying fake news.

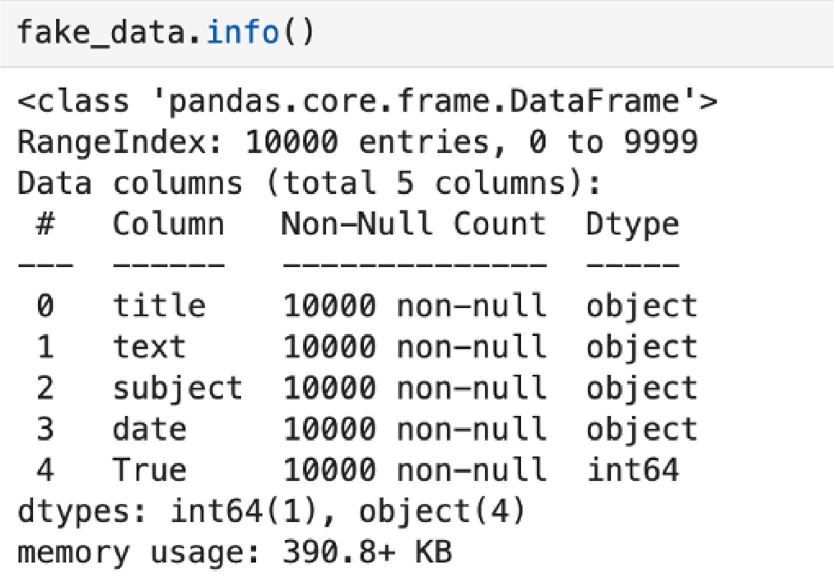
In 2019, a data analysis involving 293,570,101 articles[33] was conducted using neural networks. The research results included sentence-level classification and fragment-level classification. The obtained data showed that classification at the sentence level yielded an F1-score of 60.82%, while classification at the fragment level resulted in an F1-score of 22.58%. Later, the authors [34] addressed the problem of identifying suspicious terms related to radical content using machine learning methods. Their results indicated that Random Forest was the best classifier, achieving an accuracy of 94%. Subsequently, a binary model for propaganda identification was developed [35]. The authors compiled the Qprop dataset, which consists of real news articles collected from 104 different news publications. Their approach was based on comparisons between various textual and linguistic models. Their evaluation demonstrated that symbolic N-grams, combined with Nela, yielded the best results. Similar approaches are presented in [36] and [37], which utilized Tf-IDF, POS, lexicon-based, and Word2vec vectors for propaganda identification.

Additionally, the study [38] developed a method for identifying fake news based on its distribution using a graph neural network. The authors evaluated their model on both known and unknown datasets. [39] proposed a two-stage system for identifying propaganda in news articles. Experiments with a dataset consisting of 550 news articles showed that their model outperformed state-of-the-art methods. Furthermore, the authors in [40] proposed a method that combines sentiment analysis with the use of Word2vec, claiming that this combination of semantic and emotional analysis leads to an improvement in the task of identifying propaganda.

Also, large language models (LLMs) have recently been gaining popularity. In [41], the authors investigate the effectiveness of modern large language models, such as GPT-3 and GPT-4, for detecting propaganda. Experiments were conducted using the SemEval-2020 Task 11 dataset, which contains news articles tagged with 14 propaganda techniques, framed as a multi-label classification problem. Five variants of GPT-3 and GPT-4 were used, each employing different strategies to develop hints and bring the models to high accuracy. The results show that GPT-4 achieves comparative results with the best state-of-the-art approaches. Additionally, this study analyzes the potential and challenges of LLMs in demanding tasks such as propaganda detection.

Experiments, results and discussion

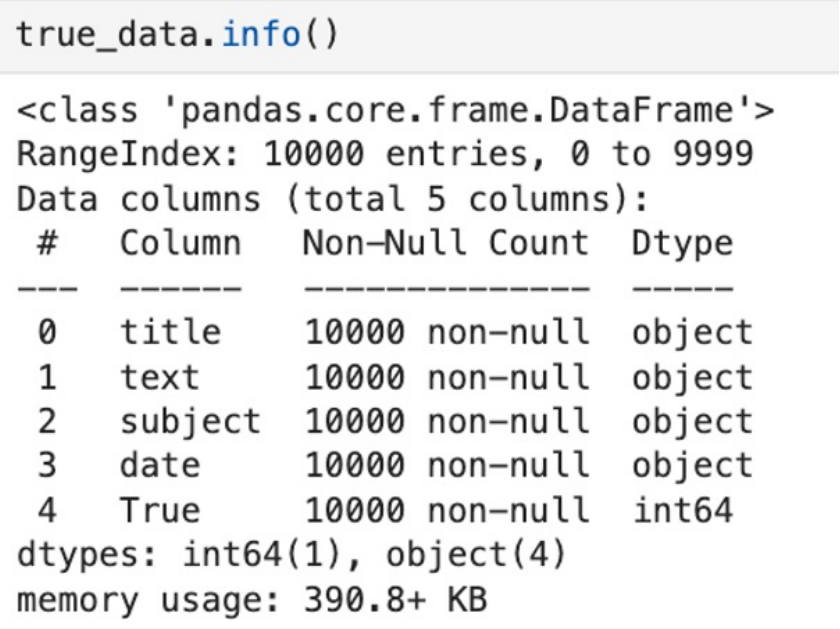
For our experiment, we use a dataset with a total of 20,000 articles: 10,000 entries for fake news (Fig. 1) and 10,000 for non-fake news (Fig. 2). Most of the articles are related to politics.



```
fake_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 5 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   title    10000 non-null  object
 1   text     10000 non-null  object
 2   subject  10000 non-null  object
 3   date     10000 non-null  object
 4   True     10000 non-null  int64
dtypes: int64(1), object(4)
memory usage: 390.8+ KB
```

Figure 1. *DataFrame information for the fake sub-dataset*



```
true_data.info()
```

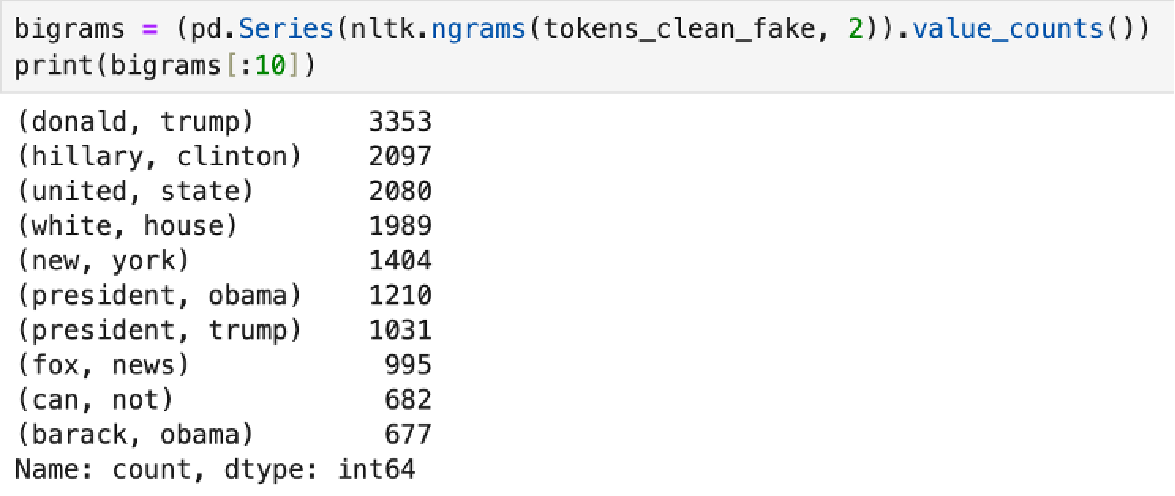
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 5 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   title    10000 non-null  object
 1   text     10000 non-null  object
 2   subject  10000 non-null  object
 3   date     10000 non-null  object
 4   True     10000 non-null  int64
dtypes: int64(1), object(4)
memory usage: 390.8+ KB
```

Figure 2. DataFrame information *for the non-fake sub-dataset*

For both sub-datasets, basic text clean-up procedures were performed, such as changing text to lowercase, removing punctuation, cleaning up location and author tags, and removing stop words, etc.

After the cleanup, tokenizing and lemmatizing were performed. For better lemmatization results, each token was POS tagged. The use of POS tagging helps perform lemmatization more accurately.

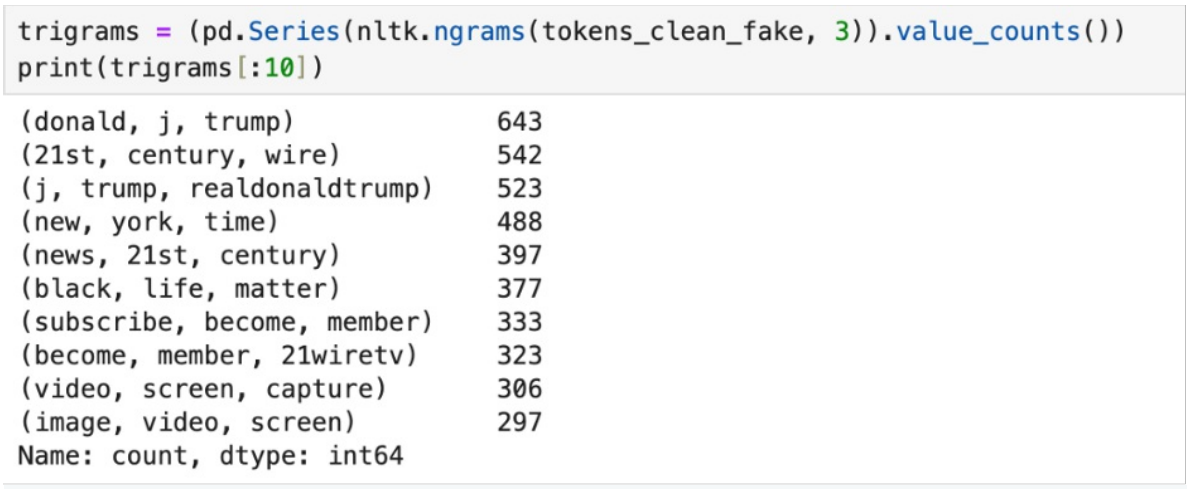
For both sub-datasets, bigrams and trigrams were built to better understand the context of the articles in the dataset.



```
bigrams = (pd.Series(nltk.ngrams(tokens_clean_fake, 2)).value_counts())
print(bigrams[:10])
```

```
(donald, trump)       3353
(hillary, clinton)    2097
(united, state)       2080
(white, house)        1989
(new, york)           1404
(president, obama)    1210
(president, trump)    1031
(fox, news)            995
(can, not)             682
(barack, obama)        677
Name: count, dtype: int64
```

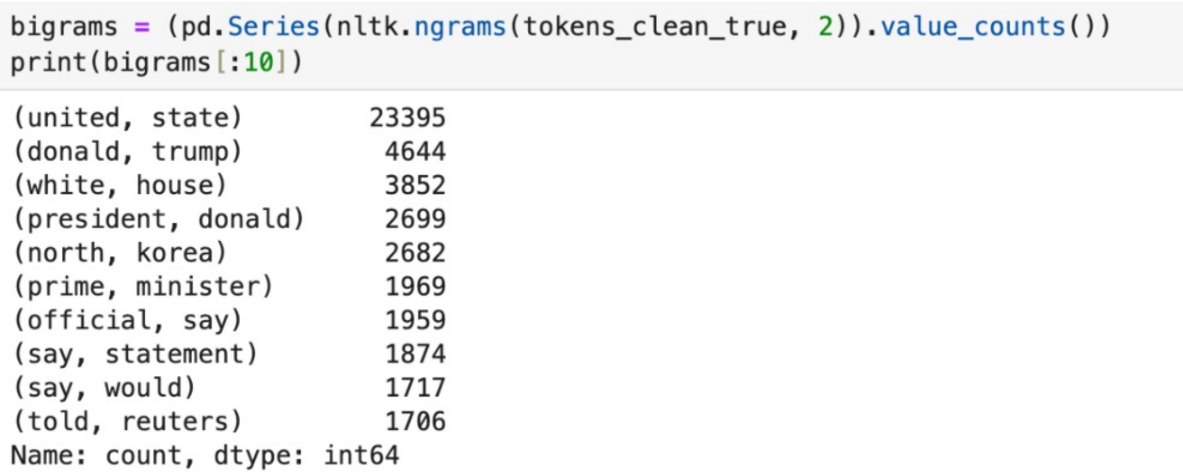
Figure 3. *TOP-10 bigrams from fake news sub-dataset*



```
trigrams = (pd.Series(nltk.ngrams(tokens_clean_fake, 3)).value_counts())
print(trigrams[:10])
```

```
(donald, j, trump)              643
(21st, century, wire)           542
(j, trump, realdonaldtrump)     523
(new, york, time)               488
(news, 21st, century)           397
(black, life, matter)           377
(subscribe, become, member)     333
(become, member, 21wiretv)      323
(video, screen, capture)        306
(image, video, screen)          297
Name: count, dtype: int64
```

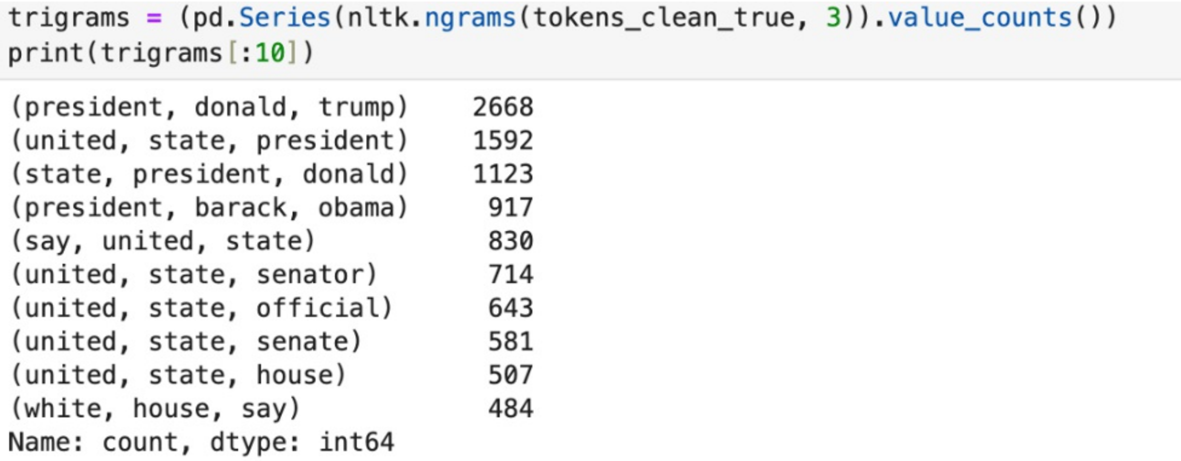
Figure 4. *TOP-10 trigrams from fake news sub-dataset*



```
bigrams = (pd.Series(nltk.ngrams(tokens_clean_true, 2)).value_counts())
print(bigrams[:10])
```

```
(united, state)        23395
(donald, trump)         4644
(white, house)          3852
(president, donald)     2699
(north, korea)          2682
(prime, minister)       1969
(official, say)         1959
(say, statement)        1874
(say, would)            1717
(told, reuters)         1706
Name: count, dtype: int64
```

Figure 5. *TOP-10 bigrams from non-fake news sub-dataset*



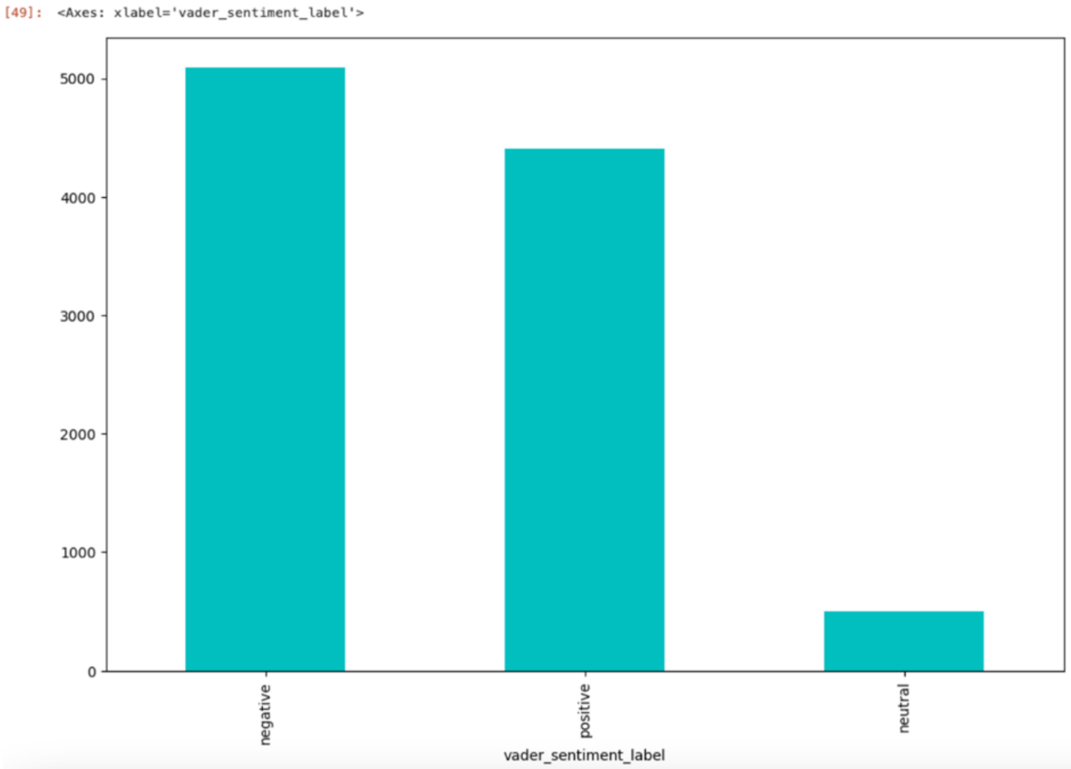
```
trigrams = (pd.Series(nltk.ngrams(tokens_clean_true, 3)).value_counts())
print(trigrams[:10])
```

```
(president, donald, trump)    2668
(united, state, president)    1592
(state, president, donald)    1123
(president, barack, obama)     917
(say, united, state)           830
(united, state, senator)       714
(united, state, official)      643
(united, state, senate)        581
(united, state, house)         507
(white, house, say)            484
Name: count, dtype: int64
```

Figure 6. *TOP-10 bigrams from non-fake news sub-dataset*

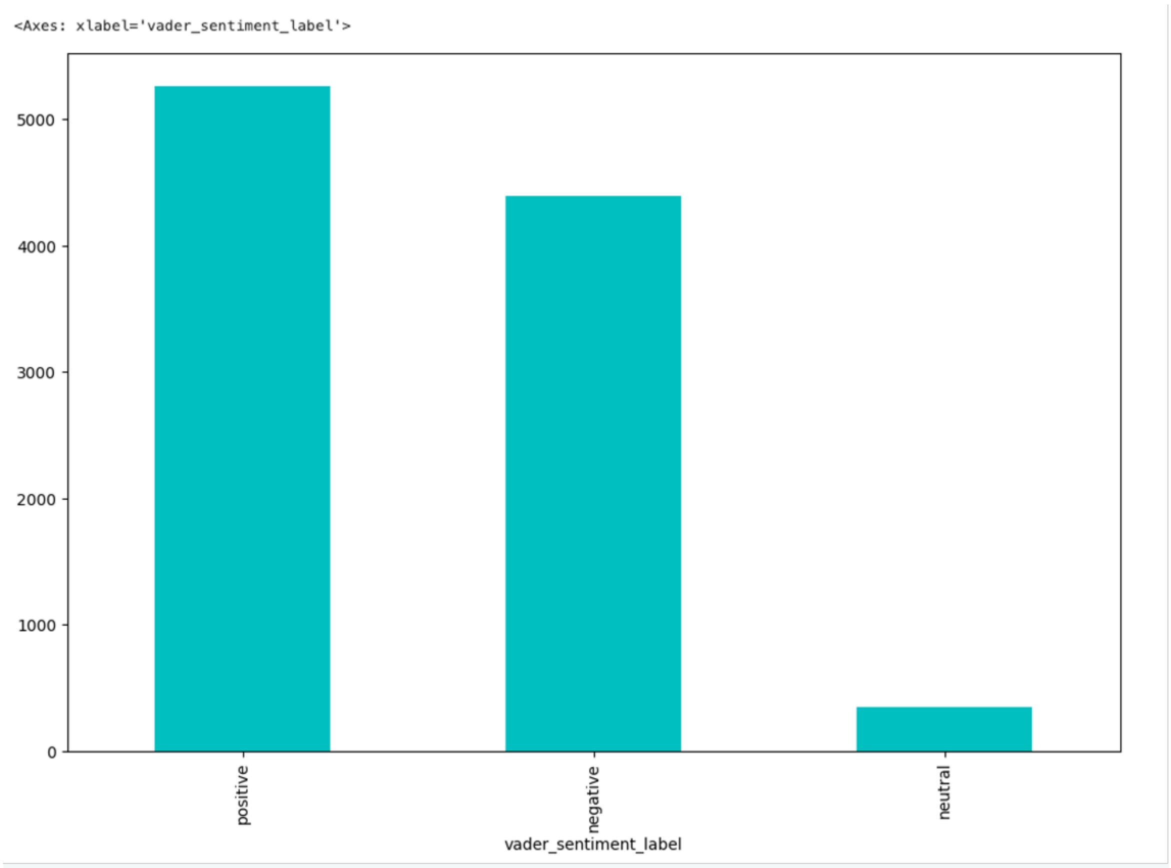
As we can see from Figs. 3-6, a more official language style is used in the non-fake news. For example, the top trigram from the fake sub-dataset is '(donald, j, trump)', whereas for the non-fake sub-dataset, it is '(president, donald, trump)'.

Next, we performed sentiment analysis on both sub-datasets (Figs. 7-8). The results show that the fake sub-dataset contains more negative scores, while the non-fake sub-dataset mostly has positive scores.



```
[49]: <Axes: xlabel='vader_sentiment_label'>
```

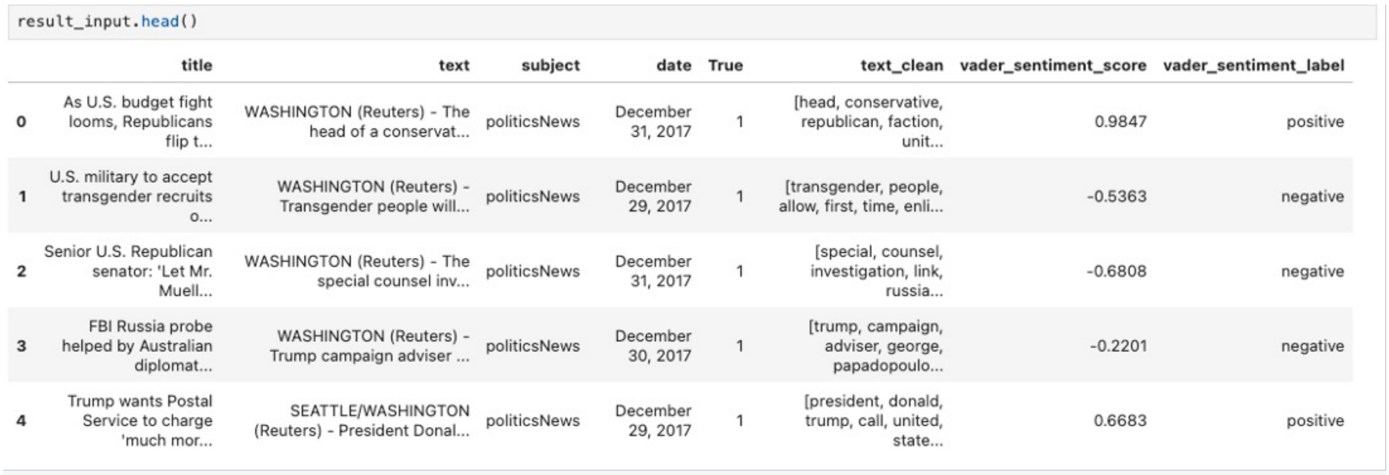
Figure 7. *Sentiment labels plot for fake news sub-dataset*



```
<Axes: xlabel='vader_sentiment_label'>
```

Figure 8. *Sentiment labels plot for non-fake news sub-dataset*

The sub-datasets were merged before building the prediction model. In Fig. 9, the head of the resulting dataset is shown.

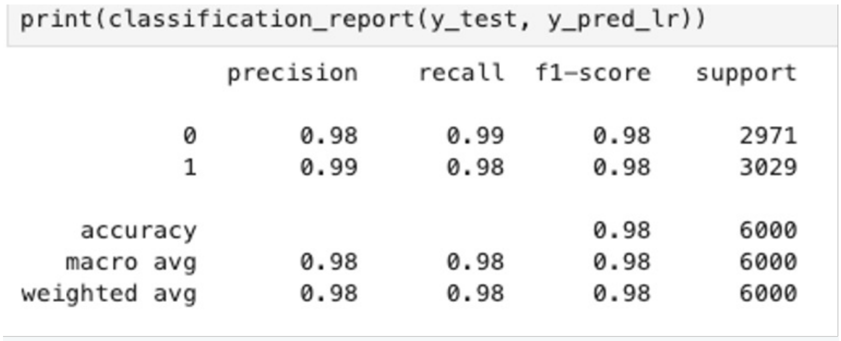


```
result_input.head()
```

	title	text	subject	date	True	text_clean	vader_sentiment_score	vader_sentiment_label
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017	1	[head, conservative, republican, faction, unit...	0.9847	positive
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017	1	[transgender, people, allow, first, time, enli...	-0.5363	negative
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017	1	[special, counsel, investigation, link, russia...	-0.6808	negative
3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	December 30, 2017	1	[trump, campaign, adviser, george, papadopoulo...	-0.2201	negative
4	Trump wants Postal Service to charge 'much mor...	SEATTLE/WASHINGTON (Reuters) - President Donal...	politicsNews	December 29, 2017	1	[president, donald, trump, call, united, state...	0.6683	positive

Figure 9. *Dataset's head after cleanup and sentiment analysis*

For the prediction model, the BOW and Logistic Regression functions were used. The results of the model are shown in Fig. 10.



```
print(classification_report(y_test, y_pred_lr))

              precision    recall  f1-score   support

           0       0.98      0.99      0.98      2971
           1       0.99      0.98      0.98      3029

    accuracy                           0.98      6000
   macro avg       0.98      0.98      0.98      6000
weighted avg       0.98      0.98      0.98      6000
```

Figure 10. *Results of model*

The F1-score is 0.98 for both classes (0-fake, 1-non-fake). Such good results can be explained by the 'laboratory' quality of the dataset. In further experiments, we want to focus on verifying the model on real-time news.

Conclusions

At the current stage, various methods exist for detecting fake news. A promising direction is the use of machine learning models that can be integrated into news monitoring systems and information portals to automatically detect fake news in texts. Such an approach can significantly reduce the spread of misinformation, improve the quality of information available to users, and increase trust in mass media.

In the world of the information society, where a large amount of information constantly becomes available through various communication channels, the problem of fake news becomes extremely urgent. It can influence public opinion, political decisions, and international relations. Thus, the creation of machine learning models to detect fake news in texts is a key task for scientific research and practical application.

In today's world, where social networks have become an unlimited source of information and influence, the problem of disinformation has become extremely relevant. Thanks to the rapid spread of news and personal recommendation systems, social networks have not only become platforms for information exchange but also powerful tools for manipulating public opinion. This is where the creation of disinformation detection systems becomes crucial.

First, disinformation detection systems are a necessary tool for maintaining user trust in social networks. Thanks to the development of deep learning technologies and big data analysis, it has become possible to automatically detect patterns and characteristics of disinformation. This allows for the prompt recognition and removal of unreliable information, ensuring higher-quality content and protecting public opinion.

Second, disinformation detection systems help preserve democratic processes. This is especially relevant during elections when the spread of fake news can significantly affect electoral decisions and the stability of society. Ensuring the reliability of information on social networks is becoming an integral part of maintaining honesty and transparency in political processes.

Third, disinformation detection systems enable us to maintain security and screen out threats to public safety. Given the role of social networks in mobilizing society and spreading extremist ideas, it is important to recognize and suppress dangerous content on time. This helps maintain the stability and security of the online community.

References

1. ^*Tyshchenko V., Muzhanova T. Disinformation and fake news: features and methods of detection on the internet. Cybersecurity: education, science, technique. 2022. T. 2, № 18. C. 175–186. URL: https://doi.org/10.28925/2663-4023.2022.18.175186 (дата звернення: 16.05.2024).*
2. ^*Myronyuk, O. Disinformation: How to Recognize and Combat It. Faculty of Law, Yuriy Fedkovych Chernivtsi National University. URL: https://law.chnu.edu.ua/dezinformatsiia-yak-rozpiznaty-ta-borotysia/ (accessed on: 16.05.2024).*
3. ^*Reuter, C., Hartwig, K., Kirchner, J., & Schlegel, N. (2019). Fake news perception in Germany: A representative study of people's attitudes and approaches to counteract disinformation.*
4. ^*Luchko, Yu. I. (2023). The Role of Artificial Intelligence Technologies in the Spread and Fight Against Disinformation. Countering Disinformation in the Context of Russian Aggression Against Ukraine: Challenges and Perspectives. Scientific Research Institute of Public Policy and Social Sciences. https://doi.org/10.32782/ppss.2023.1.26*
5. ^*Komisariv, M. The Problem of the Spread of False Information in the Media and Social Media. Abstract for the presentation at the 20th Central Asian Media Conference "The Future of Journalism". URL: https://www.osce.org/representative-on-freedom-of-media.*
6. ^*Marchi R. With facebook, blogs, and fake news, teens reject journalistic "objectivity". Journal of communication inquiry. 2012. Vol. 36, no. 3. P. 246–262. URL: https://doi.org/10.1177/0196859912458700 (date of access: 10.12.2023).*
7. ^*Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake News Detection on Social Media: A Data Mining Perspective. SIGKDD Explor. Newsl. 19, 1 (June 2017), 22–36. https://doi.org/10.1145/3137597.3137600*
8. ^*Zhou, Z.; Guan, H.; Bhat, M. and Hsu, J. (2019). Fake News Detection via NLP is Vulnerable to Adversarial Attacks. In Proceedings of the 11th International Conference on Agents and Artificial Intelligence - Volume 2: ICAART; ISBN 978-989-758-350-6; ISSN 2184-433X, SciTePress, pages 794-800. DOI: 10.5220/0007566307940800*
9. ^*Lazer DM, Baum MA, Benkler Y, Berinsky AJ, Greenhill KM, Menczer F, Metzger MJ, Nyhan B, Pennycook G, and Rothschild D. 2018. The science of fake news. Science 359:1094-1096*
10. ^*Vosoughi S, Roy D, and Aral S. 2018. The spread of true and false news online. Science 359:1146- 1151.*
11. ^*Zuo C, Karakas A, and Banerjee R. 2018. A hybrid recognition system for check-worthy claims using heuristics and supervised learning. CEUR workshop proceedings.*
12. ^*Hansen C, Hansen C, Simonsen JG, and Lioma C. 2018. The Copenhagen Team Participation in the Check-Worthiness Task of the Competition of Automatic Identification and Verification of Claims in Political Debates of the CLEF-2018 CheckThat! Lab. CLEF (Working Notes).*
13. ^*Thorne J, and Vlachos A. 2018. Automated fact checking: Task formulations, methods and future directions. arXiv preprint arXiv:180607687.*

14. ^*Mihaylova T, Karadjov G, Atanasova P, Baly R, Mohtarami M, and Nakov P. 2019. SemEval- 2019 task 8: Fact checking in community question answering forums. arXiv preprint arXiv:190601727.*
15. ^*O'Brien, N. (2018).Machine learning fordetection of fake news. Diss. Massachusetts Institute of Technology. https://dspace.mit.edu/handle/1721.1/1197279.*
16. ^*Canini KR, Suh B, and Pirolli PL. 2011. Finding credible information sources in social networks based on content and social structure. 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing: IEEE. p 1-8.*
17. ^*Derczynski L, Bontcheva K, Liakata M, Procter R, Hoi GWS, and Zubiaga A. 2017. SemEval- 2017 Task 8: RumourEval: Determining rumour veracity and support for rumours. arXiv preprint arXiv:170405972.*
18. ^*Baly R, Karadzhov G, Alexandrov D, Glass J, and Nakov P. 2018. Predicting factuality of reporting and bias of news media sources. arXiv preprint arXiv:181001765.*
19. ^*Hardalov M, Koychev I, and Nakov P. 2016. In search of credible news. International Conference on Artificial Intelligence: Methodology, Systems, and Applications: Springer. p 172-180.*
20. ^*Baly R, Karadzhov G, Alexandrov D, Glass J, and Nakov P. 2018. Predicting factuality of reporting and bias of news media sources. arXiv preprint arXiv:181001765.*
21. ^*Juola P. 2012. An Overview of the Traditional Authorship Attribution Subtask. CLEF (Online Working Notes/Labs/Workshop): Citeseer.*
22. ^*Stamatatos E. 2009. A survey of modern authorship attribution methods. Journal of the American Society for information Science and Technology 60:538-556.*
23. ^*Popat K, Mukherjee S, Strˆtgen J, and Weikum G. 2017. Where the truth lies: Explaining the credibility of emerging claims on the web and social media. Proceedings of the 26th International Conference on World Wide Web Companion. p 1003-1012.*
24. ^*Aphiwongsophon, S., &Chongstitvatana, P. (2018). Detecting fake newswith machine learning methods.2018 15thInternational Conference on Electrical Engineering/Electronics, Computer, Telecommunications & Information Technology (ECTI-CON). https://ieeexplore.ieee.org/document/862005110.*
25. ^*Potthast M, Kiesel J, Reinartz K, Bevendorff J, and Stein B. 2017. A stylometric inquiry into hyperpartisan and fake news. arXiv preprint arXiv:170205638.*
26. ^*Koppel M, Schler J, and Bonchek-Dokow E. 2007. Measuring Differentiability: Unmasking Pseudonymous Authors. Journal of Machine Learning Research 8.*
27. ^*Horne B, and Adali S. 2017. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. Proceedings of the International AAAI Conference on Web and Social Media.*
28. ^*A. Jain, A. Shakya, H. Khatter and A. K. Gupta, "A smart System for Fake News Detection Using Machine Learning," 2019 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), Ghaziabad, India, 2019, pp. 1-4, doi: 10.1109/ICICT46931.2019.8977659.*
29. ^*Rashkin H, Choi E, Jang JY, Volkova S, and Choi Y. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. Proceedings of the 2017 conference on empirical methods in natural language processing. p 2931-2937.*

30. ^Shao C, Ciampaglia GL, Varol O, Flammini A, and Menczer F. 2017. The spread of fake news by social bots. arXiv preprint arXiv:170707592 96:104.
31. ^Horne BD, Khedr S, and Adali S. 2018. Sampling the news producers: A large news and feature data set for the study of the complex media landscape. Twelfth International AAAI Conference on Web and Social Media.
32. ^Abdullah-All-Tanvir, et al. (2019). Detecting Fake News using Machine Learning and Deep Learning Algorithms.2019 7thInternational Conference on Smart Computing and Communications (ICSCC). https://ieeexplore.ieee.org/document/884361211.
33. ^Da San Martino G, Yu S, BarrÛn-Cedeno A, Petrov R, and Nakov P. 2019. Fine-grained analysis of propaganda in news article. Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP). p 5636-5646.
34. ^Nouh M, Nurse JR, and Goldsmith M. 2019. Understanding the radical mind: Identifying signals to detect extremist content on twitter. 2019 IEEE International Conference on Intelligence and Security Informatics (ISI): IEEE. p 98-103.
35. ^Barrón-Cedeño A., Jaradat I, Da San Martino G, and Nakov P. 2019. Proppy:Organizingthenewsbasedontheirpropagandisticcontent. https://wwwsciencedirectcom/science/article/abs/pii/S0306457318306058:16.
36. ^Oliinyk V-A, Vysotska V, Burov Y, Mykich K, and Fernandes VB. 2020. Propaganda Detection in Text Data Based on NLP and Machine Learning. MoMLeT+ DS. p 132-144.
37. ^Altiti O, Abdullah M, and Obiedat R. 2020. JUST at SemEval-2020 task 11: Detecting propaganda techniques using BERT pre-trained model. Proceedings of the Fourteenth Workshop on Semantic Evaluation. p 1749-1755.
38. ^Han Y, Karunasekera S, and Leckie C. 2020. Graph neural networks with continual learning for fake news detection from social media. arXiv preprint arXiv:200703316.
39. ^Li W, Li S, Liu C, Lu L, Shi Z, and Wen S. 2021. Span identification and technique classification of propaganda in news articles. Complex & Intelligent Systems:1-10.
40. ^Polonijo B, äuman S, and äimac I. 2021. Propaganda Detection Using Sentiment Aware Ensemble Deep Learning. 2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO): IEEE. p 199-204.
41. ^Sprenkamp, K., Jones, D. G., & Zavolokina, L. (2023). Large Language Models for Propaganda Detection. arXiv preprint arXiv:2310.06422.