Detection of fake text messages in real time with machine learning – literature review

Aneta Poniszewska-Maranda
Institute of Information Technology/Lodz University of Technology
Lodz, Poland *aneta.poniszewska-maranda@p.lodz.pl*

Adam Czyzewski
Institute of Information Technology/Lodz University of Technology
Lodz, Poland *adam.czyzewski@dokt.p.lodz.pl*

Lumbardha Hasimi
Institute of Information Technology/Lodz University of Technology
Lodz, Poland *lumbardha.hasimi@dokt.p.lodz.pl*

Abstract

In the era of growing digital threats, real-time detection of fake text messages has become a key challenge to ensure systems and services integrity. The literature review examines the latest developments in the field, focusing on the use of advanced machine learning techniques and deep learning. To have a full understanding of the most recent developments, common techniques and challenges, the review covers both—methodological and practical aspects of implementing real-time systems, and discusses other aspects related to efficiency, scalability and data management. The study underscores the critical role of advanced analytical techniques in combating the rapid dissemination of misinformation in contemporary digital environments.

Keywords: Fake text messages, Recognition of fake news, Real-time detection, Machine learning.

1. Introduction

In the digital age, the spread of fake text messages, commonly referred to as fake news, as well as SMS spam or phishing attempts, has become a serious problem for individuals and organizations, with societal implications in general. The false or only partially manipulated messages are often intended to spread false information on Internet, and often to extract confidential information or spread malware, posing serious security and privacy risks. To combat such a growing threat, researchers have increasingly turned to machine learning techniques that can detect and filter fake text messages.

However, detection alone has shown little to no impact, considering the rapid pace and time-sensitive nature of fake news dissemination. Numerous studies emphasize that the speed and patterns of spreading are critically important. Research has found that fake news or deceptive content spreads most rapidly within the first hours, often peaking in the initial minutes of sharing with around 70% faster spreading that real content [9], [23]. Many studies have recommended the development of advanced real-time detection systems that utilize machine learning (ML) and deep learning (DL) techniques that allow the analysis of vast amounts of data quickly and accurately [18], [22]. Given these findings, it is suggested that real-time detection is the most crucial and sensitive aspect of addressing the spread of fake news, considering the existing gaps and potential improvements in the field.

Machine learning, with its ability to analyze massive amounts of data and identify patterns, offers a promising solution to the problem of fake text messages [5]. Traditional methods such as rule-based filtering have proven insufficient to keep up with the evolving tactics of malicious actors. Machine learning models, especially those using natural language processing (NLP) techniques, can dynamically adapt to new types of fake news, spam, as well as phishing attempts [5], [19], [21]. By training on large datasets of labeled messages, these models can learn to distinguish legitimate from fake communications with high accuracy.

The research community has made significant progress in developing and refining machine learning approaches. To enhance the detection capabilities of these models, techniques such as machine learning, deep learning, and reinforcement learning are used [18, 19]. Moreover, the integration of real-time detection mechanisms ensures that users are protected immediately, without any noticeable delays. This article describes the methods, approaches, datasets, and evaluation metrics used in current research, providing a comprehensive overview of the state-of-the-art in real-time fake text message detection. This paper discusses the current state of research taking into consideration the critical aspect of real-time detection, while highlighting recent developments and challenges.

2. Research methodology

This section outlines the methodology employed in the study, aimed at exploring machine learning solutions for real-time detection of fake news. Given the rapidly evolving nature of misinformation tactics, it is crucial to employ advanced analytical techniques. Our methodology focuses on identifying the most effective ML methods for text classification, enhancing computational efficiency, addressing challenges in real-time detection, and optimizing ML systems.

Following established frameworks in software engineering literature reviews, such as those outlined in [18] this study employs a structured approach to investigate the efficacy of ML methods in real-time fake news detection. In PRISMA protocol the definition of research questions is based on PICO that can be defined as follows: (1) P (Population): text messages (SMS, messaging texts); (2) I (Intervention): ML algorithms for fake text message detection; (3) C (Comparison): (Implicit) human-based detection, rule-based systems or alternative ML methods; (4) O (Outcome): accuracy, precision, recall, F1-score, detection time (or other measures of real-time performance). By defining the following research questions, we aim to systematically address critical aspects of the field, ensuring comprehensive understanding and robust analysis:

1. RQ1: What are the most common machine learning techniques used for text classification?
2. RQ2: What strategies are employed to enhance the computational efficiency of machine learning models?
3. RQ3: What approaches excessively optimize (leading to overkilling) or minimally impact the performance of machine learning models?
4. RQ4: What are the primary challenges associated with detecting fake news in real-time systems?
5. RQ5: What methods are utilized to optimize real-time systems?

The focus was on finding solutions related to text classification, but there are also publications related to multi-modal classification. Information was sought about systems processing and classifying in real time.

The main purposes of the research questions of this study are: through (RQ1) and (RQ2), we cover both the theoretical and practical aspects of model development. By investigating methods as stated in (RQ3) we will shed light on identifying best practices and potential pitfalls, which is crucial for refining detection models. Moreover, focusing on the challenges specific to real-time systems (RQ4) and optimization techniques (RQ5) we ensure that our findings are directly applicable to the development of effective, timely solutions for fake news detection.

To analyze the research works in the field of artificial intelligence used to detect text fake news and process the data in real time, the following databases were selected: (1) ACM Digital Library; (2) IEEE Digital Library; (3) Elsevier Science Direct; (4) Springer Library. The used search string with keywords: *("fake text message" OR "SMS fraud" OR "spam detection" OR "phishing detection") AND ("machine learning" OR "artificial intelligence" OR "deep learning") AND ("real-time" OR "online detection" OR "streaming data").* The pre-selection criteria defined in this way allowed us to find 770 scientific publications: 423 papers from Science@Direct, 163 from ACM Digital Library, 94 from IEEE Digital Library, 90 from ISI Web of Science.

The initial selection of papers was guided by following defined criteria – inclusion/exclusion criteria: (1) The works concern issues related to machine learning (ML) or deep learning (DL) in the field of classification of false information – a necessary criterion. (2) The papers not published before 2018 – it was decided to include works from a maximum of 7 years ago. (3) The works contain a AI / ML / DL model along with the achieved result – the final goal of the research work is to create a model, hence the state of knowledge should include achievements not only in the theoretical sphere, but also provide the information about the created solutions in the field of artificial intelligence / machine learning. Initially, models achieving less than 90% accuracy should be rejected, but this criterion has changed. In some cases, it was observed that although the results are slightly worse, the created method may influence the development of a new solution. (4) English as the target language – this criterion was adopted due to the fact that our research work is based on the classification of fake news in English. Solutions that included multilingual models were accepted. (5) Text classification – the lack of text classification in the work did not assess it as useful for defining the state of knowledge within the considered problem. Some works contain both text and image classification, therefore only the part relating to the textual aspect was taken into account.

The lack of fulfilling these criteria resulted in the rejection of the given study. It is worth mentioning that works focusing on the classification of fake news in real time were also searched for. However, due to very small number of such studies, it was not decided to include it as a criteria. Of the total 770 papers collected, 165 were duplicates. After excluding conference papers and review papers, 51 research papers were selected for the final analysis.

Following the selection based on the above-mentioned criteria, the selected papers were further analyzed. Each article was rated on a scale of 0-4. This constitutes the scale of usefulness of scientific work in searching for the state of knowledge about the given problem [18]. The evaluation criteria were formulated in the form of following questions:

- Does the study contribute to detecting text fake news?
- Does the study concern real-time system solutions?
- Are the research objectives clearly defined by the authors?
- Is the solution clearly presented and justified?

Each question of evaluation criteria could be answered with "no" giving 0 points, "partially" giving 0.5 points and "yes" giving 1 point. Each scientific work that received at least 1.5 points according to the grades was taken into account as a work relating to a given problem. As a result of this assessment, the papers were obtained presenting the current state of knowledge.

3. Analysis of review results

This section presents the results of the state-of-the-art analysis based on determined research questions. The answers to predefined questions constitute the current state of knowledge about detecting the text fake news. Both methods and challenges important in the context of exploring the topic of NLP (in particular false information) are discussed.

3.1. RQ1: What are the predominant machine-learning techniques used for text classification?

Model building and data analysis. The researchers often use different methods for data processing and classification. Most of the studies follow the AI/ML standard methods described in this section. However, it is worth attention to some procedures that can optimize the sample classification process. The purpose of summarizing the methods is to select the best suited approaches to solve a given problem. Analysis of the achievements of other works allows us to determine the level of performance of individual approach.

Preprocessing. There are various methods for obtaining the feature vectors in analyzed works. Currently, the standard is to clean the texts of punctuation marks and stop words. When choosing this through NLTK module, it is necessary to transform the text into feature vectors. The most popular algorithms are Word2VEC, TF-IDF and BERT model. Sometimes the LIWC algorithm is also used. The characteristics and variations of algorithms are more important than the principle of operation itself – especially in the context of real-time processing. The results achieved by most algorithms in combination with the classifiers are usually similar, but the dominant popularity of some does not come out of nowhere.

Word2Vec is a prevalent approach used for producing the word embeddings, which are vectorized representations of words in a continuous vector space. This method seeks to acquire distributed word representations by analyzing their contextual usage in a given text corpus. The algorithm presents two primary frameworks: Continuous Bag of Words (CBOW) and Skip-gram. The CBOW architecture predicts the target word by taking into account its context words, using a context window of words as input. The goal is to optimize the likelihood of accurately guessing the target word based on its surrounding context. Conversely, the Skip-gram architecture aims to forecast context words based on target word, with objective of maximizing the probability of context words given the target word.

The training procedure encompasses the preparation of an extensive corpus of text, the preprocessing of this text, and the creation of a vocabulary wherein each word is assigned a distinct vector representation. The model's architecture comprises an input layer, embedding layer, and a layer that can be either CBOW or Skip-gram. The training objective aims to optimize the probability of target word given its context (CBOW) or the probability of context words given the target word (Skip-gram), usually by utilizing negative log likelihood as the loss function. Optimization techniques, such as Stochastic Gradient Descent (SGD), are utilized to modify the weights (word vectors) and improve the predicted accuracy of the model.

After undergoing training, Word2Vec generates vector representations for every word in the vocabulary, effectively capturing the semantic associations between words. The embeddings in question capture contextual similarity, meaning that words that appear in comparable situations would have similar vector representations. The spatial arrangement of word vectors in the vector space indicates semantic associations. The influence of Word2Vec extends to diverse natural language processing (NLP) problems, such as text categorization, sentiment analysis, and machine translation, owing to its capability to offer significant and efficient word representations.

The duration of training Word2Vec models is contingent upon various parameters. The size of the corpus is a significant factor, as larger datasets necessitate more processing time. Furthermore, the selected model architecture, such as CBOW or Skip-gram, has an impact on computing requirements, with Skip-gram models typically being more resource-intensive. The dimensionality of word vectors is an additional factor to consider, as it impacts the level of complexity in the computations required.

The overall training duration is influenced by hyperparameters, including the learning rate, context window size, and the number of negative samples. Adjusting these parameters may need extra computational resources. Hardware accelerators, such as GPUs or TPUs, can greatly speed up the training process by efficiently handling the matrix operations involved in neural network training. In addition, Word2Vec implementations utilizing widely-used deep learning frameworks such as TensorFlow or PyTorch can facilitate the parallelization across several CPUs or GPUs, resulting in additional reduction of training time.

Optimization methods such as hierarchical softmax or negative sampling are commonly used to accelerate training by approximating the complete softmax function. Early termination can also be incorporated, enabling the training process to end once a desirable level of performance is reached, so preserving computational resources.

Term Frequency-Inverse text Frequency (TF-IDF) algorithm is a technique that assigns weights to words based on their importance within a text compared to their frequency throughout a whole collection of documents. The computation of the TF-IDF score for a term in a document involves the multiplication of its TF (Term Frequency) and IDF (Inverse Document Frequency) values. This rating accounts for both the relative value of the term within the document and its overall importance throughout the full collection of texts. Words with higher TF-IDF scores are seen as more significant inside a document, and ensuing sparse vector form is typical in TF-IDF results.

This approach is extensively employed in information retrieval, text categorization, and keyword extraction because to its effective ability to capture the significance of phrases in a collection of documents. TF-IDF is typically efficient in terms of computing time, making it well-suited for a range of text analysis applications, especially when compared to more computationally intensive embedding models [20].

Bidirectional Encoder Representations from Transformers (BERT) unlike other extraction algorithms, it is architecturally based on a deep learning solution. Moreover, its advantage is the fact that it is pre-trained on a huge amount of data and this network has been tested in many fields other than just detecting the fake news.

In the context of fake news, it is possible to use the fine-tuning method on the collected data, without modifying the architecture (possibly adding a feature vector at the end to match the rest of the solution is not a problem). Training reduces generalization and allows for the extraction of text features characteristic of the separation of fake news and true news. An important advantage of BERT is also the ability to capture the context of the word used. This may be especially important when using everyday language, as some words may have two meanings [8].

The Linguistic Inquiry and Word Count (LIWC) algorithm is an infrequently used feature extraction method in the context of fake news detection solutions. The primary objective of this tool is to examine the written or spoken language and extract different linguistic and psychological characteristics from the text in order to gain the understanding into the psychological and emotional aspects of language [24].

The primary objective of the extraction process is to generate the features based on word count, categorization, and percentage computation:

1. Word counting – an algorithm that tallies the frequency of particular words or word categories in a provided text. The terms are predetermined and classified into categories such as emotions, cognitive processes, social processes.

2. Categorization involves the classification of words into predetermined linguistic and psychological groups. Lexical items denoting positive or negative affect, definite and indefinite articles, personal and possessive pronouns, and other linguistic element that are systematically classified into distinct categories.

3. Percentage calculation – computes the proportion of words in a given text that fall into specific established categories, offering a standardized measure for comparing texts of

varying lengths.

The computational duration of LIWC is contingent upon the magnitude of text and the intricacy of categories. It is typically effective for small to medium-sized texts but can be computationally costly for really large datasets. The time complexity is determined by the quantity of words and categories being examined. The underlying idea of LIWC is that certain language patterns and word categories might serve as indicators of psychological and emotional states. The program categorizes words into predetermined categories and examines their distribution to gain the insights on the psychological attributes of the text.

AI models architectures. In the category of artificial intelligence models considered in the application of text processing, it is worth dividing this category into two parts: machine learning and deep learning. Machine learning category uses basic algorithms such as SVM and it is currently no longer the primary method for processing the text samples. Solutions based on recurrent networks have become particularly popular. Thanks to them, it is possible to achieve the fake news classification results of 95.99%.

Very often, well-known machine learning algorithms are used after feature extraction. They usually provide a kind of reference point for more proposed deep networks. The results achieved on well-categorized LIAR and PHEME sets range from 85% to 95%, depending on the method of feature extraction and data preparation used. In most cases, a benchmark in the form of a comparison of SVM, RF, DT, MLP algorithms may not be enough to fully legitimize the artificial intelligence model. They may constitute the basis for testing new/modified feature extraction methods, but combining them with deep learning algorithms will not bring anything new to the current state of art [5].

Due to their lower computational complexity, their use seems to be more reasonable and optimal from the point of view of processing speed. Three different network recently gained the most popularity and constitute a reference point in the research are the following:

1. Long Short-Term Memory (LSTM) is designed to solve the problem of vanishing gradient and to capture the long-range connections in sequential data. The fundamental components are memory cells, responsible for preserving a state of long-term memory. The movement of information into and out of the memory cell is regulated by three gates: Forget, Input, and Output. The Forget gate is responsible for selecting the elements to be removed from the cell state, while the Input gate is responsible for identifying and storing the new information. The Output gate, on the other hand, regulates the information that will be produced as output. The cell state retains information throughout time, while the hidden state acts as the output of LSTM cell, gathering pertinent information from the input [15].

2. Bidirectional LSTM (bi-LSTM) differs from standard LSTMs in processing input sequences bidirectionally, considering both past and future information concurrently [10]. This processing of the network enables it to assimilate information from both preceding and subsequent contexts, hence enhancing its ability to comprehend and capture interdependencies within the input sequence. A Bi-LSTM architecture comprises two distinct LSTM layers: one that processes the input sequence in the forward direction and another that processes it in the backward manner. Every LSTM layer retains its own distinct hidden state and memory cell. The outputs from both directions are commonly concatenated at each time step or mixed in a different way to create the ultimate output sequence [10], [12].

3. Gated Recurrent Unit (GRU), like LSTM, employs the same mechanisms, but with a more straightforward structure consisting of two primary gates: Reset Gate and Update Gate. Reset Gate is responsible for determining the extent to which previous information should

be disregarded or forgotten. The model utilizes the preceding hidden state and the current input to generate a value ranging from 0 to 1 for each element in the hidden state. A value of 1 indicates total retention, while a value of 0 signifies complete disregard. Update gate controls the extent to which the current concealed state should integrate the new information. It determines the extent to which previous information should be transmitted to subsequent instances. Like the reset gate, the update gate vector is generated by processing the previous concealed state and the current input. The GRU model does not have a distinct cell state, unlike the LSTM model [2], [11]. Alternatively, there exists a concealed candidate state that merges information from the preceding concealed state with the present input. The candidate hidden state is calculated using the reset gate and it is subsequently combined with the previous hidden state using the update gate through linear interpolation [14].

In addition to the recurrent network architectures defined in this way, convolutional networks are also used (despite their purpose for image analysis). Overall, these 3 or 4 architectures are often compared to see which one achieves the best results with a particular preprocessing method. The Word2Vec algorithm is most often used, where in combination with recurrent networks results achieves around 95%-99% on ISOT [17], LIAR, FakeNewsNet sets [16], [15], [10], [2], [23]. On the other hand, some studies indicate good results achieved by combining BERT and convolutional networks. They achieved similar results on the same harvests – also around 95% on all quality measures [7].

Results comparison. When examining results, standard quality measures are most often used, i.e. accuracy, precision, sensitivity and F1 measure. The simplest method used for validation and evidence in support of the thesis being tested is to train a standard classifier (and several similar ones, e.g. GRU and LSTM), and refer the tested variation of the method to classical approach. Previously, a method was also used to train a deep network and compare it to machine learning algorithms. Currently, a large part of the solutions differs slightly, and the achieved results reach approximately 99% of each quality measure when selecting appropriate architectures and hyperparameters.

When comparing the results, it is also worth recalling testing the architecture on standard datasets (e.g. ISOT or LIAR). This facilitates subsequent comparison to other research works and legitimizes the proposed approach much more. The model's behavior can be tested on own data (this can be the first step to analyzing the model in a production environment). Due to the satisfactory results achieved by recurrent networks in combination with classical extraction methods, researchers are trying to look for less computationally expensive methods that maintain the quality of classification [23]. Hence, considerations for real-time sample processing may gain the popularity in near future.

3.2. RQ2: What strategies are employed to enhance the computational efficiency of machine learning models?

The subject of research was not always the use of methods based on the search for the best possible solution. When looking for appropriate methods, researchers often use machine learning algorithms. The summary of research discussed in this subsection constitutes a basis for improving and optimizing the learning process.

Preprocessing. The use of standard feature extraction methods allows for the classic extraction method related to natural language processing. However, each given problem may require a separate extraction method. For example, the study [11], [13] proves an improvement in data separability through the use of CV (one-hot encoding) method. CV converts the text document in a histogram vector, where each element represents the number of appearances of a word in the document. The vector length depends on the number of unique words in the corpus. This simplified extraction method allowed the SVM classifier to achieve 96% accuracy and 98% sensitivity on the LIAR and FakeNewsNet datasets [11], [14]. The achieved results can be compared with deep learning models. Moreover, the found solution is optimal, which may be a potential solution in the context of e.g. real processing systems.

Training the model based on several different sets. Training on several datasets and evaluation allows to create a model that is not only generalized in a sense, but also provides solid legitimization in the context of comparison with existing solutions. There is a possibility to enrich existing datasets based on collected data [22]. This approach was used by the authors of the WELfake system [20], where results of around 96% were obtained on the BERT transformer. According to the authors of this system, training on several data sets improves the model's generalization and adaptation to diverse data.

3.3. RQ3: What approaches excessively optimize or minimally impact the performance of machine learning models?

The review of the existing works allowed for the identification of several factors that make it difficult to create an optimal and effective solution/model. The aspects considered in this subsection are described from the point of view of the considered problem of textual classification of fake news in real time.

Lack of proper feature extraction. Feature extraction seems to be crucial in determining the appropriate representativeness of the data. Currently used classic methods such as Word2Vec seem to solve this problem quite well. On the other hand, they do not reflect emotional content, but are based on stylometry. Some studies tried to avoid the aspect related to stylometric analysis and decided to partially or completely abandon such feature extraction.

An example of algorithm used alone that works well in sentiment and emotional analysis is the LIWC algorithm. The studies show, that relying solely on features related to the LIWC algorithm does not fully represent the data, and classifiers have problems with the learning. In the context of tweets [4], [1] proposed methods based solely on tweet-related statistics – the number of retweets, comments, number of words, etc. It turns out that the lack of features related to text analysis creates a problem in the classification of fake news. Hence the summarisation that it is also worth using standard extraction methods. In the sense, we then reach the curse of dimensionality, but based on the research conducted, it can be concluded that it is better to extract more features than too few.

Creation of solutions heavy computationally. From the point of view of creating a system that processes data in real time, one of the factors is creating a model that is too large compared to the demand. This fact was observed when analyzing the works whose aim was to achieve the best possible results, disregarding the efficiency aspect.

A potential overkill for a real-time processing system may be the creation of a classifier based on a voting method, i.e. training many models that solve the same problem and make the final classification based on voting. This method was used and brought the good results. The authors of WELfake system [20] trained all available configurations of extraction methods and deep networks and then combined the selected ones into a voting system. The results were achieved on the set they created (it is worth noting that the solution is legitimized by making the set publicly available) and achieved 96.73% accuracy and 96.56% F1 measure.

Own dataset usage not related with other works. When conducting the research related to the study of a specific model or solution, not all research works use widely used data sets. Despite the use of data sets that were created to research a given problem and achieve satisfactory results, it is difficult to legitimize a solution based on artificial intelligence [9]. A given model could, for example, be overfitted and not fully solve a given problem. Taking into account the previously mentioned arguments, it is worth using an approach that uses existing popular collections and own data.

3.4. RQ4: What are the primary challenges associated with detecting fake news in real-time systems?

Some studies prioritize research not solely aimed at achieving the optimal model, but rather focus on elements that are practical and applicable in the intended use. This is justified, the vast majority of solutions currently achieve results above 90% on most standard sets (e.g. LIAR or ISOT). Hence, solutions based on semi-supervised or unsupervised learning (testing proportion of marked to unmarked samples) or optimal combination of various architectures are sought.

Semi-supervised learning. This type of learning is an approach used in situations where the available training data is partially labelled and some of data is unlabelled. This approach is useful in cases where collecting large amounts of labeled data is expensive or time-consuming. The use of semi-supervised learning can help to improve the performance of machine learning models by leveraging available unlabeled data in the training process. Currently, the subject of research is to understand the impact of the percentage of marked to unmarked samples on the quality of classification. For this purpose, the markings of a selected percentage of samples were removed from the training data set and the results were examined on the test set. The research carried out on models based on deep networks on the LIAR set are presented in table 1.

Table 1. Semi-supervised learning – results for deep networks on LIAR set.

Percentage of labeled samples	Accuracy	Precision	Recall	F1 score
100%	85.01%	83.57%	99.94%	91.02%
1%	82.04%	83.60%	97.64%	90.08%
10%	82.29%	83.49%	98.13%	90.21%
30%	83.52%	83.58%	99.90%	91.01%

While the given problem does not address the issue of semi-supervised learning, the results suggest that it is possible to use the self-learning in the future. The proposed system can be enriched with this concept in future so that the model adapts itself to changing the trends.

Optimal usage of AI models. Currently, the fake news classification reaches over 95% in almost all deep learning models. The considered problem concerns the issues related to real-time processing. The search for an optimal solution assumes not only a qualitative classification, which research has already been achieved, but also an optimal one in terms of resource consumption. Not much research has been carried out on this issue, but there are methods that allow for parallelizing certain classification processes (e.g. the use of graphics processors) [3].

However, it was noted that works based on real-time detection of fake news tried to reduce the number of calls to the AI model through prior filtering based on similarity testing. The authors of HOAX system [19] proposed the following method: (1) When collecting the samples, each of them is examined based on the measure of similarity with existing fake news in the database. (2) If the similarity measure reached a certain level, the sample is sent to the classification model.

This approach made it possible to circumvent the use of time-intensive methods for sample processing. On the other hand, this approach could prove more time-consuming if most of the samples were fake news. Certainly, the challenge in relation to the problem of real-time processing is to develop the similar methods that reduce the system load.

Multimodal fake news (similarity measurement). In works related to the detection of fake news, some studies partially address the issue of text samples. This is due to the fact that most forms of knowledge transfer allow the attachment of photos, videos, etc. The main problem is the search for appropriate features in order to be able to use some type of similarity measure (e.g. cosine similarity) [21], [13]. On the other hand, the problems related to multimodal classification seems to be unsolved.

3.5. RQ5: What methods are utilized to optimize real-time systems?

The topic of presenting in this paper research work is related to real-time systems. In the context of searching for solutions, this term can be understood in two ways: (1) From the moment the sample enters the system, the response delivery process should not exceed 3 seconds. (2) The system constantly monitors the platforms in search of an article, entry or post that is fake news (detection is carried out in real time). In this case, it is difficult to determine the time frame for such supervision. A key issue in the operation of such a system is the ability to scan the specific pages of article publishers or platforms.

Hence, it was decided to investigate both of these possibilities and find the current state of affairs for both interpretations.

Real time as real-time processing of samples. Currently, there is no research on the sample classification time. There are a huge number of variables that influence this process. However, during the test, it is worth measuring the sample classification time to determine the system's efficiency. Additionally, when creating an artificial intelligence system, it is worth embedding the model on a graphics processor. This allows the model to parallelize impulse processing. The use of Pytorch library may be particularly useful here (it has easy-to-use mechanisms that do not generate the errors).

Real time as data processing from external sources. Currently, the real-time data processing focuses more on monitoring various types of online platforms. The purpose of such a system is to process data that appears on a given platform (e.g. Google News) and classify it (usually as true or false) or only false based on existing data.

In recent years, several solutions have been developed to detect fake news in real time. Such a prototype is, for example, the FADE system [6]. In the optimization approach, it was decided to use a representation of 19 (relatively few) features. The SVM classification method was also used, which is relatively simple without the use of deep networks. The results achieved an accuracy of up to 90%. A similar approach was used to create the HOAX inspector system, where only machine learning algorithms were used in the research. The system normally collects data from selected websites and then classifies them.

4. Conclusions

In summary, real-time detection of fake text messages using machine learning is a dynamically developing field that is gaining importance in the context of growing digital threats. The development of advanced algorithms and natural language processing techniques allows for more effective and precise differentiation of authentic messages from fraudulent ones. Research in this area has shown that machine learning models can achieve high accuracy and efficiency, which are crucial for the battle against such threats.

With the effective application of machine learning models and their optimal performance, several challenges must be addressed. Key issues include data management, such as the requirement for extensive and representative datasets for training, and concerns surrounding data privacy and security, which present substantial barriers. Furthermore, deploying real-time detection systems necessitates robust infrastructure and efficient monitoring mechanisms, entailing additional overheads and technological prerequisites.

In our study we delve into the methodologies and findings in the detection of fake news using machine learning techniques. By addressing critical research questions, such as identifying prevalent machine learning methods for text classification, strategies for enhancing computational efficiency, and challenges specific to real-time detection systems, this paper provides a comprehensive overview of the current state of the field. The findings highlight the importance of optimized feature extraction methods and model architectures in achieving high accuracy rates. Strategies to improve computational efficiency, such as parallelization and pre-filtering based on similarity measures, are examined to address the operational demands of real-time systems.

Apart from laying the groundwork for future advancements in AI-driven fake news detection systems, this study highlights the need for collaboration between academia, industry and government which also impact the developing effective solutions. Integrating modern technologies such as artificial intelligence and big data and developing standards and regulations for the security of electronic communications will contribute to creating a more secure digital environment. As technology advances, more and more advanced and effective tools can protect against fake text messages.

References

[1] Aguilera, A., Quinteros, P., Dongo, I., Cardinale, Y.: Credibot: Applying bot detection for credibility analysis on twitter. IEEE Access 11, pp. 108365–108385 (2023)

[2] Chen, M.Y., Lai, Y.W., Lian, J.W.: Using deep learning models to detect fake news about covid-19. ACM Trans. Internet Technol. 23(2) (2023), `https://doi.org/10.1145/3533431`

[3] Dadkhah, S., Shoeleh, F., Yadollahi, M.M., Zhang, X., Ghorbani, A.A.: A real-time hostile activities analyses and detection system. Applied Soft Computing 104, pp. 107175 (2021)

[4] Demirci, S., Sagiroglu, S.: Twitterbulletin: An intelligent and real-time automated news categorization tool for twitter. Journal of Universal Computer Science 28(4), pp. 345–377 (2022)

[5] Dutta, A.K.: Detecting phishing websites using machine learning technique. PLoS ONE 16(10), pp. e0258361 (2021)

[6] Jabiyev, B., Pehlivanoglu, S., Onarlioglu, K., Kirda, E.: Fade: Detecting fake news articles on the web. Association for Computing Machinery, New York, NY, USA (2021), `https://doi.org/10.1145/3465481.3465751`

[7] Kalra, S., Kumar, C.H.S., Sharma, Y., Chauhan, G.S.: Multimodal fake news detection on fakeddit dataset using transformer-based architectures. In: Khare, N., Tomar, D.S., Ahirwal, M.K., Semwal, V.B., Soni, V. (eds.) Machine Learning, Image Processing, Network Security and Data Sciences. pp. 281–292. Springer Nature Switzerland (2022)

[8] Kasnesis, P., Heartfield, R., Liang, X., Toumanidis, L., Sakellari, G., Patrikakis, C., Loukas, G.: Transformer-based identification of stochastic information cascades in social networks using text and image similarity. Applied Soft Computing 108, pp. 107413 (2021)

[9] Khanday, A.M.U.D., Rabani, S.T., Khan, Q.R., Malik, S.H.: Detecting twitter hate speech in covid-19 era using machine learning and ensemble learning techniques. International Journal of Information Management Data Insights 2(2), pp. 100120 (2022)

[10] Kishore, V., Kumar, M.: Enhanced multimodal fake news detection with optimal feature fusion and modified bi-lstm architecture. Cybernetics and Systems 0(0), pp. 1–31 (2023)

[11] Luo, Y., Ma, J., Yeo, C.K.: Bcmm: A novel post-based augmentation representation for early rumour detection on social media. Pattern Recognition 113, pp. 107818 (2021)

[12] Majumdar, B., RafiuzzamanBhuiyan, M., Hasan, M.A., Islam, M.S., Noori, S.R.H.: Multi class fake news detection using lstm approach. In: 2021 10th International Conference on System Modeling Advancement in Research Trends (SMART). pp. 75–79 (2021)

[13] Nadeem, M.I., Ahmed, K., Zheng, Z., Li, D., Assam, M., Ghadi, Y.Y., Alghamedy, F.H., Eldin, E.T.: Ssm: Stylometric and semantic similarity oriented multimodal fake news detection. Journal of King Saud University - Computer and Information Sciences 35(5), pp. 101559 (2023)

[14] Nasir, J.A., Khan, O.S., Varlamis, I.: Fake news detection: A hybrid cnn-rnn based deep learning approach. International Journal of Information Management Data Insights 1(1), pp. 100007 (2021)

[15] Okunoye, O.B., Ibor, A.E.: Hybrid fake news detection technique with genetic search and deep learning. Computers and Electrical Engineering 103, pp. 108344 (2022)

[16] Saxena, N., Sinha, A., Bansal, T., Wadhwa, A.: A statistical approach for reducing misinformation propagation on twitter social media. Information Processing & Management 60(4), pp. 103360 (2023)

[17] Sheikhi, S.: An effective fake news detection method using woa-xgbtree algorithm and content-based features. Applied Soft Computing 109, pp. 107559 (2021)

[18] Shen, Y., Liu, Q., Guo, N., Yuan, J., Yang, Y.: Fake news detection on social networks: A survey. Applied Sciences 13(21), pp. 11877 (2023), https://www.mdpi.com/2076-3417/13/21/11877

[19] Varshney, D., Vishwakarma, D.K.: Hoax news-inspector: a real-time prediction of fake news using content resemblance over web search results for authenticating the credibility of news articles. Journal of Ambient Intelligence and Humanized Computing 12(9), pp. 8961–8974 (2021)

[20] Verma, P.K., Agrawal, P., Amorim, I., Prodan, R.: Welfake: Word embedding over linguistic features for fake news detection. IEEE Transactions on Computational Social Systems 8(4), pp. 881–893 (2021)

[21] Xue, J., Wang, Y., Tian, Y., Li, Y., Shi, L., Wei, L.: Detecting fake news by exploring the consistency of multimodal data. Information Processing & Management 58(5), pp. 102610 (2021)

[22] Zhang, C., Gupta, A., Kauten, C., Deokar, A.V., Qin, X.: Detecting fake news for reducing misinformation risks using analytics approaches. European Journal of Operational Research 279(3), pp. 1036–1052 (2019)

[23] Zhang, Q., Guo, Z., Zhu, Y., Vijayakumar, P., Castiglione, A., Gupta, B.B.: A deep learning-based fast fake news detection model for cyber-physical social services. Pattern Recognition Letters 168, pp. 31–38 (2023)

[24] Zhao, Y., Da, J., Yan, J.: Detecting health misinformation in online health communities: Incorporating behavioral features into machine learning based approaches. Information Processing & Management 58(1), pp. 102390 (2021)