SMU Data Science Review

Volume 7 | Number 1 Article 3

Content-Based Unsupervised Fake News Detection on Ukraine-Russia War

Yucheol Shin
Southern Methodist University, yucheol.shin92@gmail.com

Yvan Sojdehei
Southern Methodist University, sojdehei007@gmail.com

Limin Zheng
Southern Methodist University, minmin2005216@gmail.com

Brad Blanchard
Southern Methodist University, bablanchard@mail.smu.edu

Content-Based Unsupervised Fake News Detection on Ukraine-Russia War

Yucheol Shin[1], Yvan Sojdehei[1], Limin Zheng[1], Bradley Blanchard,
[1] Master of Science in Data Science, Southern Methodist University,
Dallas, TX 75275 USA
shinc@smu.edu, ysojdehei@smu.edu, lzheng@smu.edu

Abstract. The Ukrainian-Russian war has garnered significant attention worldwide, with fake news obstructing the formation of public opinion and disseminating false information. This scholarly paper explores the use of unsupervised learning methods and the Bidirectional Encoder Representations from Transformers (BERT) to detect fake news in news articles from various sources. BERT topic modeling is applied to cluster news articles by their respective topics, followed by summarization to measure the similarity scores. The hypothesis posits that topics with larger variances are more likely to contain fake news. The proposed method was evaluated using a dataset of approximately 1000 labeled news articles related to the Syrian war. The study found that while unsupervised content clustering with topic similarity was insufficient to detect fake news, it demonstrated the prevalence of fake news content and its potential for clustering by topic.

1 Introduction

On February 24, 2022, Russia launched its biggest invasion in Europe since World War II, attacking Ukraine [1]. This military action has drawn worldwide attention and concern, with suspicions being raised that Russian media sources may be flooding the international media with false information for political propaganda purposes [2]. In times of war, the dissemination of fake news can be particularly divisive and used to justify acts of aggression. Therefore, it is crucial to decipher false information to prevent further geopolitical uncertainty.

During wartime, the media and social outlets have a significant impact on how people perceive conflicts. The continuous flow of news creates anxiety and confusion among the public [3], as people become increasingly worried about nuclear weapons, chemical weapons, refugees, NATO escalation [4], soaring oil prices [5], and food shortages. Allowing fake news to spread further can exacerbate people's fears and spread inaccurate and illegitimate sources, which can lead to greater anxiety among the public.

In the current era of information overload, the proliferation of fake news has become an increasingly prevalent and problematic issue. Research conducted by NewsGuard [2] highlights the fact that certain TikTok users have created fake live streams of Ukraine-Russian conflicts by utilizing old clips of unrelated incidents, which can appear realistic and deceive viewers. This false information is then disseminated and used to collect donations and charities, which ultimately end up in the hands of the creators. With the continued development of social networking services like Twitter, Instagram, TikTok, and Facebook, unverified information spreads rapidly, often accompanied by misleading images and videos. For many people, differentiating between true and false information based on their limited knowledge and experience has become increasingly difficult. Fake news has permeated various platforms, intermingling misleading information with facts. In most cases, by the time news platforms have taken the necessary steps to debunk questionable articles, the damage has already been done, with millions of viewers having received false information, leading to negative consequences for public perception of specific issues or topics.

With the increasing accessibility of online news to the public, differentiating

between legitimate and false sources becomes more challenging for authorities and organizations. In today's world, digital technologies have enabled individuals to access vast amounts of information. Unfortunately, some malicious actors use this technology to disseminate misinformation that is often incomplete or untrue. The irresponsible spread of fake news can be particularly damaging, especially during wartime, as it can undermine efforts to stabilize public perception and opinion. Given that leaders often take public opinion into account when making decisions and formulating strategies, it is strategically critical to limit the media's influence on public opinion during wartime. Fake news can be employed as a weapon to plant seeds of misinformation in the public psyche and sway opinion [6].

In the context of wartime, news coverage can sometimes be viewed as political propaganda, and the term "fake news" is often used to describe the intentional dissemination of misleading or false information [7]. Such fake news can harm the reputations and integrity of individuals and organizations, and is often employed as a means of swaying public opinion. Moreover, linguistic differences may result in misinterpretations and inaccurate translations, leading to further misunderstandings, particularly when news sources are reporting on events in Ukraine or Russia. Notably, there is a significant informational discrepancy between Western and Russian media regarding the conflict in the region [8]. To mitigate this bias, the current study draws on news sources from 70 different countries, thereby avoiding the influence of national origin on the collection of information.

Obtaining labeled data presents the greatest challenge in identifying fake news. The process of cross-checking information, verifying sources, and conducting investigations require significant time and manual labor, making it difficult to obtain labeled data and identify fake news accurately. Moreover, creating machine learning models that can classify fake news for specific topics, such as wartime situations, poses significant challenges. The content of fake news varies across different conflicts due to various factors, such as the region, weapons used, and political, racial, and religious influences. For instance, news coverage of the Syrian war that began on March 15, 2011 [9] primarily focused on chemical weapons and neighboring countries like Turkey and Saudi Arabia, while reports of missile bombing and civilian casualties dominate the news about the Ukrainian-Russian conflict [10].

The identification of fake news is a crucial step in combating the spread of misinformation on a global scale. This study seeks to achieve this objective by utilizing a content-only approach to determine the degree of misinformation in news articles. Given the challenges of obtaining labeled datasets, this study will employ unsupervised training to overcome this hurdle. Specifically, this research will use a comprehensive database of news relating to the Ukrainian-Russian conflict from a diverse range of news sources, languages, and regions, spanning from the outbreak of the conflict until the present day.

2 Literature Review

With the widespread availability of digital electronics, disseminating information through computer and communication technologies has become effortless. Online news media and social networks have become the most common ways of spreading information. As a result, the digital age has come to be known as the "information age," with misinformation emerging as a significant issue alongside technological advancement. In this age, it has become essential for people to distinguish between true and false news. Various private sector entities, such as Facebook, Google, TikTok, and Twitter, have invested significant time and resources in fighting fake news [11].

2.1 Fake News in War Time

The outbreak of the Ukraine and Russian war has given rise to the proliferation of fake news and disinformation. In March 2022, Russian President Putin enacted a law that imposes severe penalties, including up to 15 years in prison, on those who publish fake news [12]. This law empowers the authorities to block websites that spread false information or impose fines or imprisonment for disseminating unapproved information. A survey conducted in France revealed that approximately 60% of respondents believed that fake news and disinformation likely affected news in the Russia-Ukraine war [13]. The subjective and often edited nature of videos and pictures made it difficult for the public to discern the truth. As a result, it has become widely accepted that information warfare is a part of modern warfare. Lovelace's study of newspapers and social media during modern warfare, including Twitter and TikTok, demonstrated the media's potential to influence public opinions and exert pressure on decision-makers [14]. In Ciuriak's study, social media networks were identified as a new battlefield in modern warfare that could be used as weapons with detrimental effects since authorities find it challenging to maintain control and censorship over these platforms [6].

Eric et al. studied the prevalence of fake digital images on social media platforms, such as Twitter, during war times and found that distorted digital images easily spread fake news over social networks [15]. Montaged digital images are more persuasive than narratives, and people tend to believe what they see more than what they hear. Inundated with a flood of fake digital images spread on social networks, people face difficulties in distinguishing between authentic and edited images. Additionally, once the images or articles are verified for authenticity, they have already circulated widely through the media worldwide.

During the Ukraine war, false videos on social media platforms like TikTok garnered millions of views, with one account featuring disinformation on the Ukraine-Russia war amassing almost 30 million views [16]. However, these videos were clips from old military training back in 2017. Such videos are misleading and can have adverse effects on public perceptions. TikTok's policy encourages creators to gain attention from the public by increasing their influence, likes, reposts, and comments, which can directly convert to income. Therefore, creators may intentionally post fake news without taking responsibility to attract public attention and follow the latest trends.

2.2 Natural Language Processing (NLP) in Fake News Detection

NLP has shown potential in detecting fake news by analyzing patterns in news content. Researchers have explored various methods to improve the accuracy and reliability of fake news detection through machine learning and AI advancements. For instance, Nasir et al. [17] proposed a hybrid CNN-RNN method for fake news detection and compared it with several traditional classification methods, including logistic regression and random forest. The proposed Modified Bi-directional LSTM method also showed promising results, achieving an accuracy of 91% in PolitiFact's dataset [18].

However, conventional supervised machine learning faces challenges such as limited amounts of labeled data and concept drift, which can reduce the effectiveness of models and impair fake news classification. The labeling of datasets is often time-consuming and costly, and past experiences may not determine future occurrences. Different categories of fake news may also have different patterns and keywords, requiring constant updates on the model.

While supervised methods are straightforward, labeled datasets are rare and difficult to obtain in real-world situations when fake news is circulating. In contrast, semi-supervised and unsupervised fake news detection can help improve efficiency with a small number of labeled instances without sacrificing accuracy [19, 20, 21]. Unsupervised algorithms can be trained without labeled data and can use clustering methods to achieve

topic extraction at the cluster level. Documentation can then be analyzed for similarity, and a method for authenticity can be determined through classification.

2.3 Unsupervised Fake News Detection

The limitations of supervised and semi-supervised learning methods for fake news detection rely on having a correctly annotated dataset. However, in real-time situations, where there is a sudden influx of news on a critical incident, such as the Russia-Ukraine War, labeled news data is often unavailable. To overcome these limitations, new research has proposed unsupervised machine learning methods based on trustworthy sources, user information, dissemination patterns, and inter-user behavior [22].

Most social networking platforms utilize fake news detection algorithms that analyze viewer content, user behavior, and user posts to determine the authenticity of accounts. Gaglani et al. [23] implemented semantic similarity technology for social network fake news detection. This unsupervised method is content-based, but it neglects the social context of social media. Gangireddy et al. [24] proposed an unsupervised method to detect fake news using user behavior, but this method does not include contents in the modeling. Li [25] used an autoencoder as the unsupervised machine learning method; four features, including content, propagation, user, and image, are extracted from social media. Yang et al. [26] proposed an unsupervised approach using a user's engagement and credibility without the contents of news to detect fake news. Hosseinimotlagh et al. [27] proposed a content-based unsupervised learning method to detect fake news using the tensor decomposition method and ensemble method in model building.

Unfortunately, there is limited fake news detection research on news articles, possibly because there are no useful features except for contents and media sources in model building. This study focuses on content-based fake news detection and utilizes BERT, bidirectional encoder representations from transformers, to extract contextual information from news articles for cluster and summary analysis. Training the BERT model on corpora will allow extraction of the core summary of the out clusters from the topic model output and potentially analyze these results at the summary level, allowing for the analysis of the prevalence of fake news at the topic level.

2.4 BERT

LSTM and Gated Recurrent Units (GRU), recurrent neural network-based networks, have traditionally been used in NLP tasks, but they are being overtaken by transformers. As such, various pre-trained language models, such as BERT and Generative Pre-trained Transformer (GPT), continue to develop. BERT, released by Google researchers Devlin et al. [28] in 2018, is a particularly noteworthy model in the field of NLP. Built on Transformers, BERT is a pre-trained language model that utilizes unlabeled text data from sources such as Wikipedia (2.5 billion words) and BooksCorpus (800 million words) [28].

BERT has found widespread application in the detection of fake news, thanks in part to its bidirectional Transformer that enables it to interpret sentence phrases. Its ability to understand contextual information in more detail by learning words bidirectionally, using a pre-trained language model and additional fine-tuning, makes it particularly suitable for NLP tasks [28]. For instance, Jwa et al. [29] utilized BERT to detect fake news automatically by analyzing the headline and contents of news articles. By using pre-training to improve performance, Jwa achieved a 0.746 F1 score, outperforming similar models such as TalosCNN and TalosTree (with F1 scores of 0.308 and 0.570, respectively) [29]. In another study, Farokhian et al. [30] applied MWPBert, which is constructed with two parallel BERT layers, to detect whether news articles were fake or

not. One layer is used to encode the news title, and the other is used to encode the actual news contents. Using this model, fake news was detected with 0.854 accuracy [30].

2.4.2 BERT Input Embedding

BERT, like other deep learning models such as ELMo and GPT-1, uses contextual embedding. The input to BERT is in the form of embedding vectors that have been processed by the embedding layer. BERT's input embedding is made up of three embedding layers: Token Embedding, Segment Embedding, and Positional Embedding [28]. These three embeddings are combined using layer normalization and dropout before being used as input to BERT.

Token Embedding utilizes the Word Piece embedding method and a sub-word tokenizer that breaks the input text into smaller units than words. This sub-word tokenizer adds frequently appearing words to the word set but divides infrequently appearing words into smaller sub-words that are then added to the word set. Once the word set is created, tokenization is performed based on this set. Segment Embedding is used for sentence separation. In BERT, two sentences are separated by inserting a delimiter ([SEP]) and then designated as one segment. BERT limits this one segment to 512 sub-words in length.

To represent the positional information of words, Transformer uses Positional Encoding, which utilizes sine and cosine functions to create a matrix with different values depending on the position of the word, and then adds word vectors to it. In contrast, BERT uses a method called Position Embedding, which is similar to Positional Encoding but is obtained through learning rather than through creating sine and cosine functions. Position Embedding assigns a positional indicator to each token, such that if the length of a sentence is 4, it trains 4 position embedding vectors and designates the positions of each word. Then, the first word is assigned to the first positional embedding, the second word to the second embedding, and so on.

2.4.3 BERT Fine-Tuning

Fine-tuning is the crucial step in applying the pre-trained BERT language model to solve specific NLP tasks. This process involves transferring the learned knowledge from the pre-training stage to a new task by fine-tuning the model with labeled data for the target task. Unlike semi-supervised learning, where the model learns from unlabeled data, fine-tuning is a supervised learning process that requires labeled data for the target task. By fine-tuning BERT's language model, it is possible to achieve state-of-the-art performance on various NLP tasks such as Named Entity Recognition (NER) and Question Answering (QA). This approach eliminates the need to create a separate algorithm or language model for each task, making the process more efficient and effective.

In addition to its ability to learn rich language representations, BERT also includes a binary next-sentence prediction task that enables it to detect biased or fake content in news articles. This feature allows the model to classify unlabeled news articles more accurately by performing a deep analysis of the sentence prediction task. Overall, fine-tuning BERT's language model has revolutionized the field of NLP by allowing for efficient transfer learning and achieving state-of-the-art results on various tasks.

2.5 Topic Modeling

Topic modeling is a crucial natural language processing technique used to identify themes or topics within a corpus of documents. It involves analyzing both the structural

and contextual information of the documents to discover common word and phrase patterns. This approach is used in various applications, including search engines, where identifying the topic of a document is paramount, and customer complaint systems, where extracting key topics and sentiments can be useful.

Although Document-Term Matrix (DTM) and Term Frequency-Inverse Document Frequency (TF-IDF) based on Bag of Words are commonly used, they have a significant drawback in that they cannot account for the meaning of words. This is because they rely on the frequency of words, which is a purely numerical method. Latent Semantic Analysis (LSA) is a method that addresses this issue by deriving the latent meaning of DTM. Although LSA is not optimized for topic modeling, it is a useful algorithm that provides insights into the field. LSA has the advantage of being quick and easy to implement, and it can extract the potential meaning of words. LSA also performs well in calculating the similarity of documents. However, its singular value decomposition (SVD) method has limitations when it comes to adding new data.

Recent research has shown that artificial neural network-based methods, such as Word2Vec, are gaining more attention than LSA. Latent Dirichlet Allocation (LDA) is an algorithm that improves upon the shortcomings of LSA and is better suited for topic modeling. LDA assumes that words are represented as topics, and documents are represented as a collection of topics. This algorithm sorts documents into topics based on probability distribution. It is assumed that the input to LDA has undergone preprocessing to remove the subject and unnecessary propositions. In other words, preprocessed DTM becomes the input to LDA, which provides the topic and word distributions in a document within each topic.

Before selecting the final method for topic modeling, various techniques such as TD-IDF, LDA, and BERT were considered. Ultimately, BERT was selected for this research because it is more suitable for training against larger datasets like news articles. While alternative algorithms such as LDA and LSTM outperform BERT on a smaller scale, they lag behind when dealing with more complex datasets. Despite its greater computational demand, BERT's wide variety of corpora and adjustable parameters make it the better choice for this study.

2.6 Similarity

The determination of document similarity is a fundamental topic in natural language processing. Human perception of document similarity primarily depends on the shared usage of the same or similar words between documents. This also applies to machine-based approaches, where the performance of document similarity relies on the numerical representation of words in each document, such as Document-Term Matrix, TF-IDF Vectors, or Word2Vec, as well as the method used to calculate differences between words, such as Euclidean distance and cosine similarity.

In the field of fake news detection, researchers have explored similarity-based methods. For instance, Gaglani et al. [23] used semantic similarity technology to identify fake news on social networks. Their method involved collecting news articles from credible sources, extracting keywords, performing text summarization, comparing network claims with real news article summarization, and using similarity scores to evaluate claims on social networks. Similarly, Zhou et al. [31] proposed a similarity detection method by comparing text and image information in an article. Mhatre and Masurkar [32] also proposed fake news detection methods based on semantic similarity.

This study aims to employ unsupervised techniques to classify news articles as fake or factual. Contextual information will be extracted through summarization techniques and grouped through topic modeling. Within each topic cluster, the similarity will be measured to identify the core article, which is assumed to be factual. Deviation

from the core summarization within a specific topic will be utilized to measure the "fakeness" of news articles.

3 Methods

3.1 Data Structure

A logic was created in this research to retrieve a diverse data set from various sources, which involved fetching data from Newscatcher and NewsAPI news APIs. These APIs contained articles from numerous international sources, including but not limited to major news outlets like CNN, NBC, and The Guardian. News data was collected daily, resulting in a large data frame comprising over 50,000 rows, each representing a distinct news article. The data frame consisted of several columns, including news sources (rights), authors, Country of Origin (country), and Summary, among others. The API included a feature to query articles based on language, which was utilized to obtain Russian language articles from Ukrainian and Russian news sources in addition to English language sources. These articles were then translated and added to the data frame. The data retrieval process was initiated by executing a custom webscrape3.py script [33] with the necessary parameters and desired queries.

3.2 Data Preprocessing

News articles are textual data and require preprocessing before making a machine learning model. There are various data processing techniques for NLP tasks, and it is essential to make a predictive model focusing more on important words. In this research, the following steps were taken: unnecessary web scrape text removal, punctuation removal, stop words removal, and lemmatization.

The data used in this research was collected through web scraping. Sometimes there are irrelevant texts included during the collection of the news articles. For example, phrases like *"This is a carousel. Use Next and Previous buttons to navigate"* were included as part of the news articles, and these kinds of web scrape phrases are removed as part of the data processing. Additionally, there are multiple new lines (*/n*) and tabs (*/t*) due to the web scraping, so using regex, these unnecessary texts are cleaned.

Numbers were not removed from the scraped data and were omitted from the cleanup process. Typically in NLP data processing, numbers are not considered very meaningful, so they are usually removed. However, due to the characteristics of news articles, the numbers contain important meanings. For example, news articles related to war may use numbers as an indicator of the date of incidents, the number of victims, or the financial amount of damages. Specific clock time, however, was removed from the sentences for simplicity. Likewise, they have important meaning to understand the contextual information of the news articles. These kinds of numbers were kept in the data processing.

After the removal of unnecessary words, punctuation and stop words were removed as well. Stop words are words that are commonly used but do not provide much context in a language. They are words like conjunctions, such as *"and"* and *"but"*, or prepositions, such as *"in"* and *"to"*, which connect the phrases or words of the sentences. Since they do not have much importance to the meaning of the texts, they are usually removed as part of the data processing, which helps to emphasize the other essential words.

Lastly, there are stemming and lemmatization as methods to simplify the words.

Both are techniques that remove prefixes or suffixes of the words to simplify them into common words. The difference between the two techniques is that stemming reduces words into stems by removing prefixes or suffixes of the words. Lemmatization, on the other hand, reduces words properly by morphologically analyzing the use of vocabulary and part of speech (POS) tagging, and then it returns the words into the dictionary form of a word. Because it uses POS tagging, lemmatization is also able to distinguish between verbs and nouns. In this research, for a better interpretation of the contextual meaning of the sentences, lemmatization is used rather than stemming.

As shown in *Figure 1* below, after data processing, the news articles are processed and simplified for implementing a better predictive model.

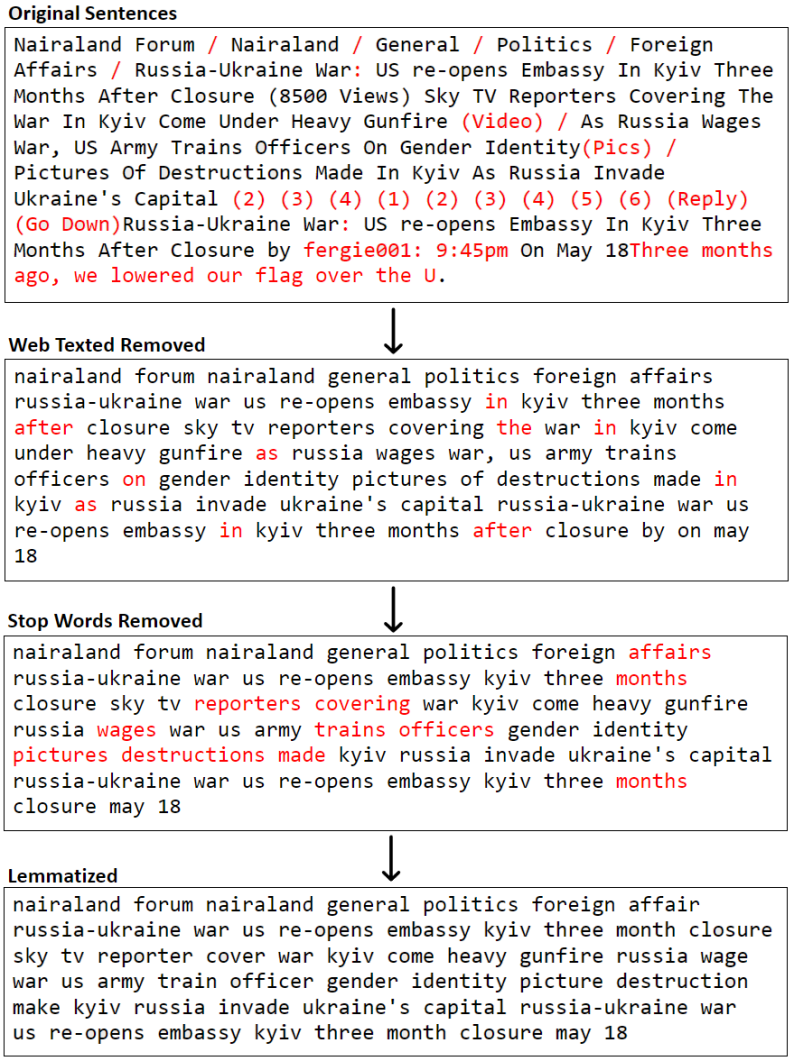


Figure 1. Example of Data Processing

In this research, data processing was conducted in two stages, with the initial step entailing the removal of redundant text prior to employing the BERT models. It has been established that the BERT models yield better results when provided with the unprocessed text to better comprehend the contextual meaning of the given text. Therefore, during the topic modeling and summarization phases, the text data underwent preprocessing solely for the removal of extraneous text. Subsequently, following the BERT model processing, the text data was further preprocessed to eliminate stop words and to perform

lemmatization for similarity measurement.

3.3 Proposed method

The proposed method for detecting fake news utilizes both similarity and variance ranking methods in conjunction with BERT to identify the most relevant articles among content with similar topics. The coherence score is used as an indicator of how frequently words appear together in documents and alone, with a higher score indicating greater coherence. Silhouette analysis is employed to evaluate the effectiveness and quality of clustering techniques.

The proposed method consists of two major steps. The first step involves clustering documents using contextual topic modeling to extract an overall summary of each topic. The second step involves measuring the similarity between news articles and the core summary within the same topic. Through summarization and contextual topic modeling, relevant articles with similar contextual information are grouped. Within articles that share similar information, their similarity and variance are measured to identify potential topics with a higher likelihood of containing fake news.

Based on the hypothesis that news articles on each topic should share similar contents and that the extracted summary should contain core contextual information, high similarity variance can be an indication of contextual differences within that topic. Finally, the similarity variance is used to rank topics that have a higher chance of containing fake news. *Figure 2* illustrates the overall architecture of the proposed method.

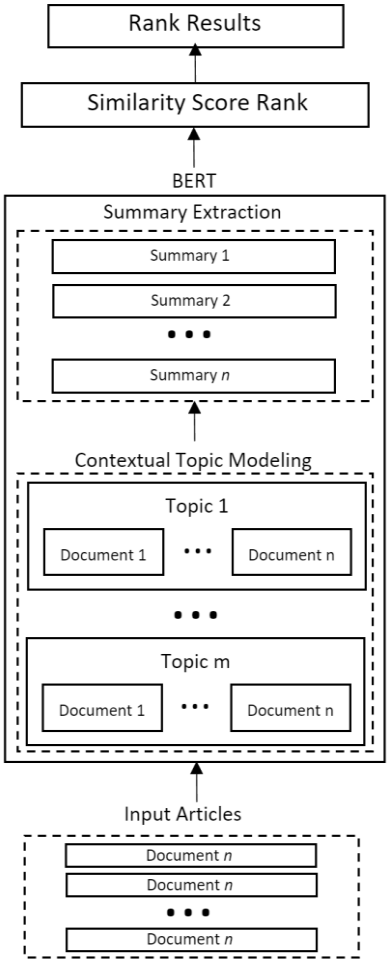


Figure 2. The overview architecture of the proposed BERT Model

3.4 Topic Modeling

Contextualized Topic Modeling is an algorithmic approach that utilizes BERT's document embedding capabilities to capture context and the unsupervised learning ability of topic models to extract relevant topics from documents. By summarizing sentences and clustering them into groups of similar articles, Contextual Topic Model identifies relevant articles for each topic. This study uses BERT sentence embedding to cluster sentences according to their contextual topics, and to achieve this goal, it employs BERTopic, which combines BERT's embedding capabilities with a sentence transformer model.

BERTopic is an improvement over traditional topic modeling techniques such as LDA, providing continuous topic modeling instead of a discrete approach, as well as stochastic modeling, resulting in greater diversity in the topics identified. BERTopic also leverages state-of-the-art language transformer models, including advanced tools such as HBSCAN and UMAP, for model adjustment, resulting in more accurate and meaningful topic identification. Researchers can easily extract the most important topics for further analysis and interpretation after the modeling process is complete. Overall, BERTopic is a superior approach to topic modeling, delivering more precise and useful insights than traditional techniques like LDA.

In this particular use case, fine-tuning the model required adjusting the model parameters, such as UMAP, to ensure proper clustering of news articles by topic and to minimize the number of detected outliers through the use of HBSCAN. Reducing the number of outliers was crucial, which led to the implementation of a multi-layered topic modeling process.

The modeling process began with dimensionality reduction, employing UMAP as a modeling parameter in the BERTopic model. UMAP can capture both local and global high-dimensional spaces in lower dimensions, allowing for various dimensionality reduction methods. The 'n_neighbors' parameter was used to balance the data structure, while a cluster size of 70 allowed for more data points in the clustering process, providing better coverage of the broader data structure. The minimum distance parameter controlled the distance at which data points were from one another, and a value of '0.0' was used for this clustering process. The 'n_components' parameter determined the dimensions of the reduced dimension space, and a value of 10 was necessary to emphasize the overall density of the clusters. Finally, the 'metric' parameter, which controlled how distance was computed within the input data, used a 'cosine' correlation metric for the model.

To adjust the model parameters and achieve proper clustering by news article topic while minimizing the number of detected outliers, a multi-layered topic modeling process was used. Starting with dimensionality reduction, UMAP was used as a modeling parameter. A cluster size of 70, a minimum distance parameter of 0.0, a value of 10 for the 'n_components' parameter, and a 'cosine' correlation metric were used. After reducing the dimensionality of the embeddings, HBSCAN was used to cluster them with similar embeddings to extract the topics. A 'min_cluster_size' value of 70, a 'min_samples' value of 10, and 'eom' as the cluster selection method were used. Initially, 11 topics were obtained with one outlier topic containing 14537 outliers, with the rest of the data model into 11 topics. To account for this, a multi-layered topic modeling process was used, resulting in better clustering and fewer outliers as shown below in *Table 1*.

	Topic	Count	Name
0	-1	14537	-1_to_and_ukraine_russian
1	0	16322	0_and_ukraine_russian_war
2	1	2213	1_nuclear_zaporizhzhia_nuclear power_power plant
3	2	553	2_erdogan_turkish_tayyip erdogan_recep tayyip
4	3	454	3_donbas_region_the donbas_donbas region
5	4	435	4_investors_stocks_markets_ukraine
6	5	431	5_indian_students_medical_indian students
7	6	403	6_pipeline_nord stream_gazprom_the nord stream
8	7	377	7_refugees_refugee_million_fleeing
9	8	341	8_uefa_champions_world cup_champions league
10	9	299	9_kherson_kherson region_dnipro_ukrainian

	Topic	Count	Name
0	-1	3669	-1_and_ukraine_russian_war
1	0	9182	0_to_ukraine_war_russia
2	1	211	1_missiles_hypersonic_hypersonic missiles_russian
3	2	165	2_trial_guilty_pleaded_on trial
4	3	103	3_hospital_pokrovsk_veronika_babies
5	4	96	4_soldiers_british_troops_carney
6	5	96	5_crimea_explosions_russianannexed crimea_ukraine
7	6	86	6_russian_strengthen_ukraine_decree
8	7	84	7_biden_quad_summit_leaders
9	8	82	8_kitten_firefighters_emergency services_ukraines
10	9	81	9_space station_international space_internatio...
11	10	81	10_sean_sean penn_film festival_sean penn is
12	11	79	11_state department_evidence_conflict observat...
13	12	79	12_australian national university_australian n...
14	13	76	13_embassy_embassy in_embassy in kyiv_us embassy
15	14	76	14_carlson_tucker carlson_tucker_invasion of iraq
16	15	75	15_filtration_deported_forcibly deported_detai...
17	16	74	16_philippines_the philippines_philippine_rodr...
18	17	71	17_higgins_letter_irish_michael higgins
19	18	71	18_prices_swiss_central bank_axpo

Table 1. Second-layer Topic Modeling done on the First layer's Outliers

BERTopic is a superior approach to traditional techniques like LDA because it delivers more precise and useful insights, leverages state-of-the-art language transformer models, and generates different results each time it is run, leading to greater diversity in the topics identified. Finally, the most important topics can be easily outputted for further analysis and interpretation after the modeling process is complete as shown below in *Figure 3*.

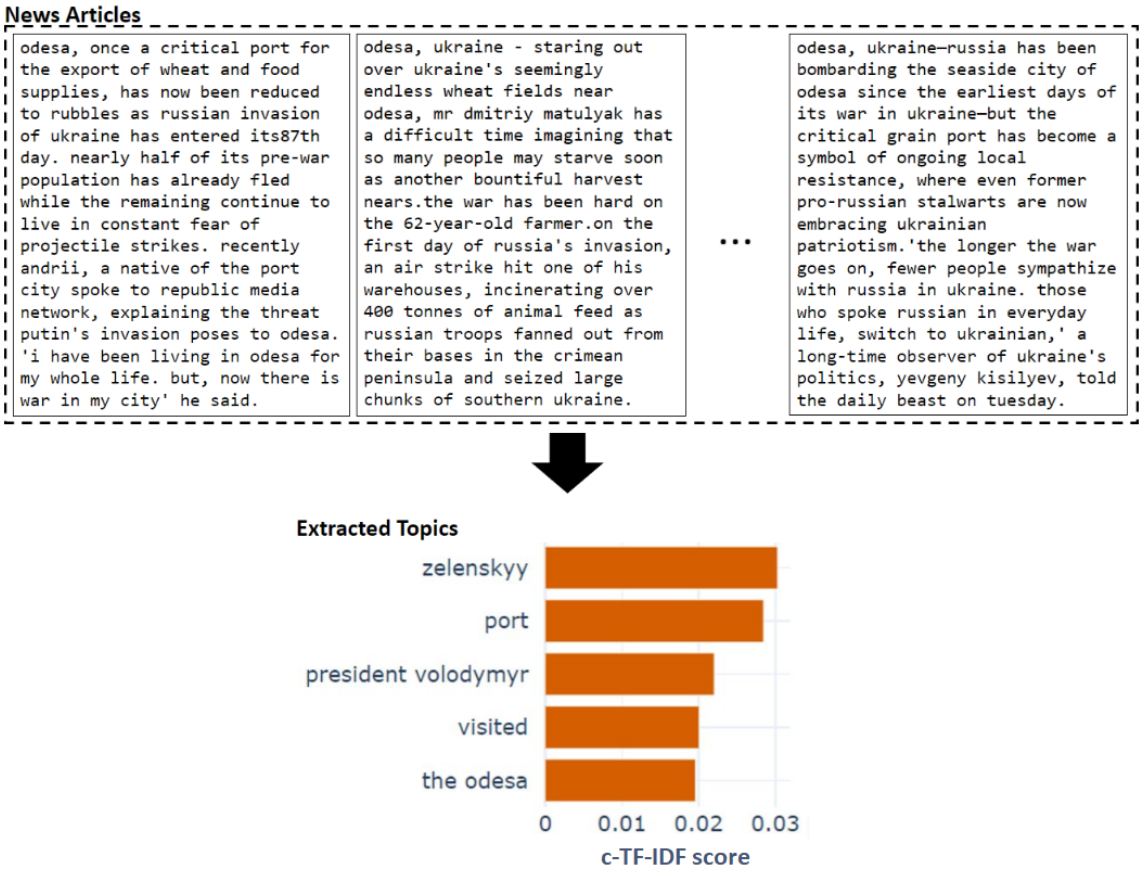


Figure 3. Topic Extraction example for Topic 175: Topics on Zelensky Visited Odesa

3.5 Summarization

To group articles by relevant contexts, key information is extracted from them. The summarization layer utilizes BERTSUM, a novel method developed by Liu [34], which is pre-trained with language encoders and uses Inter-Sentence Transformer and Recurrent Neural Network to extract the summary of the texts. The input format of BERTSUM differs from the base BERT model, where a token is added at the start of each sentence to separate multiple sentences and collect features of the preceding sentence [39]. Each sentence is assigned an embedding and then sent into further layers, where BERTSUM assigns scores to each sentence, representing the value that the sentence adds to the overall document [39].

To generate summaries of the news articles in this research, the Bert Extractive Summarizer (BES) algorithm is used, which utilizes the HuggingFace Pytorch transformers library to extract summaries [41]. This is achieved by embedding sentences and running the clustering algorithm process to find the sentences that makeup what is known as the cluster centroids. The ‘distilbert-base-uncased’ transformation model, a smaller, more flexible version of base BERT, which is pre-trained with raw text only and can use a vast array of public data, with automatic processes put into place to generate inputs and labels from text, is utilized [40]. In this research, a concise summary of a given topic is critical for contextual analysis. BERTSUM takes into account the articles on a topic that comprise the core news articles, or the articles that are most similar to one another. This process results in an extracted summary, which is what BERTSUM has determined to be the summarization of all the input core articles for the given topic (see *Figure 4*).

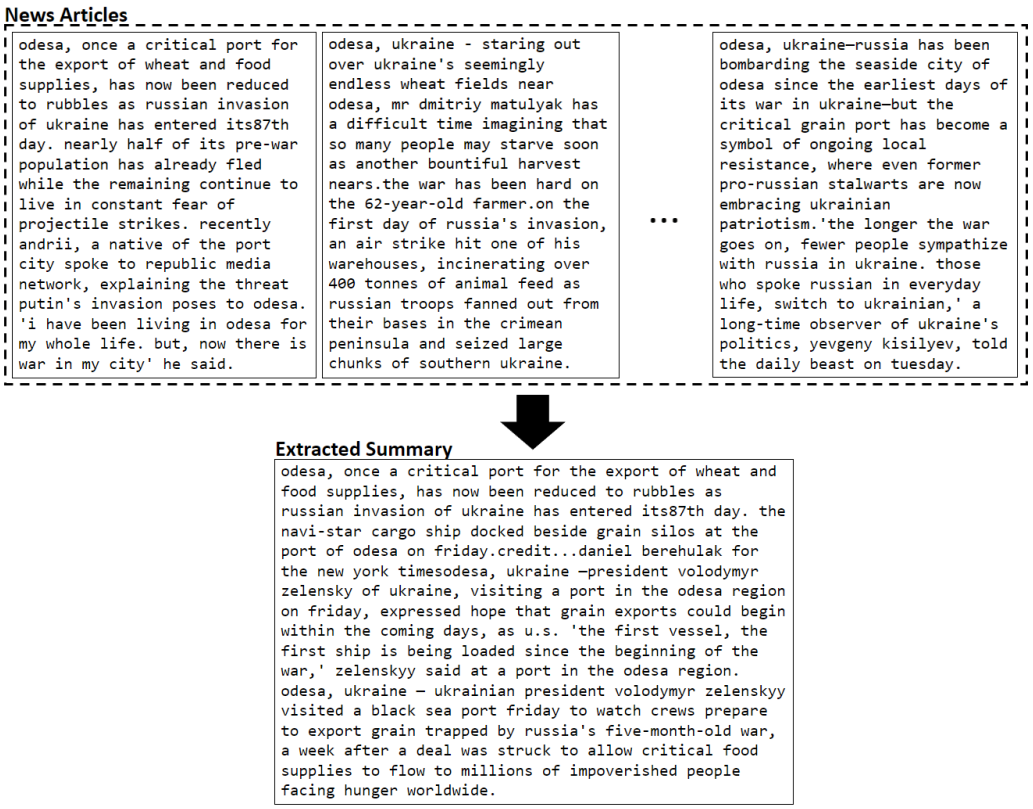


Figure 4. Example of Summary of Topic 175: Summary about a port city in Ukraine

The accuracy of the extraction process is verified through TF-IDF comparison of the most common words/phrases within the extracted core article summaries and the resulting extracted summary, as seen in *Figure 5*. The majority of the most common words from the articles match those in the output extracted summary.

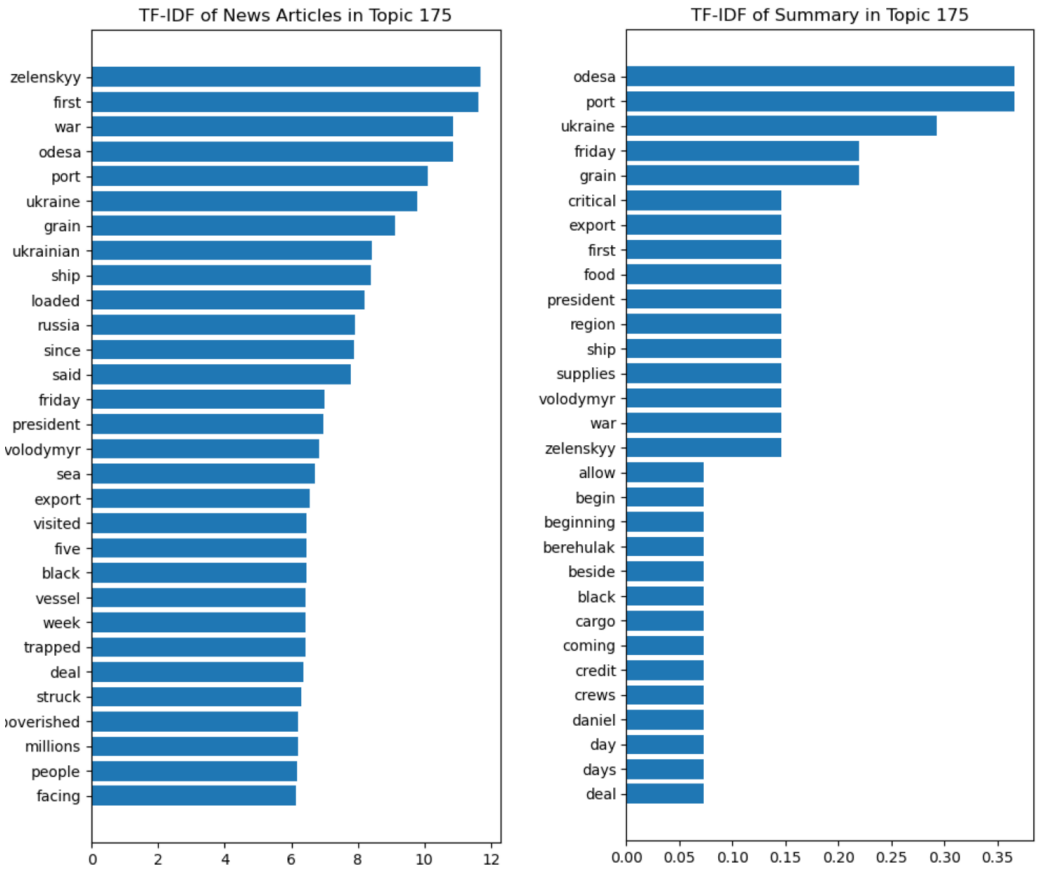


Figure 5. Comparison of TF-IDF of overall news articles and TF-IDF of summary

The article similarity score is also examined, and of the tested transformers, layer 4 distilbert-base-uncased yielded the most optimal summary extraction results. This can be attributed to its higher average similarity, as well as the better quality that the transformer has in the summary of the topic. The summary integrity is a metric that takes into account the conciseness of the extracted summary in comparison to the word count of the article summaries in the input topic to ensure that no summary was too far off from its input topic summary lengths, with layer 4 distilbert-base-uncased having both the higher mean similarity score and summary (see *Table 2*).

BERTSUM Transformation Models	Average Similarity Scores	Length of Articles / Length of Summary
distilbert-base-uncased (layer 3)	0.5254	0.2823
distilbert-base-uncased (layer 4)	0.6457	0.4925
bert-large-uncased (layer 3)	0.6206	0.4204
bert-large-uncased (layer 4)	0.5971	0.3866
paraphrase-MiniLM-L6-v2	0.5693	0.4641
all-MiniLM-L6-v2	0.5767	0.4403

Table 2. Comparison of BERT Sum model

3.6 Similarity Scores

After the completion of word embeddings and identification of the core summary sentence containing the most contextual information, the similarity between the summary sentence in each topic was measured to rank the scores. Cosine similarity, a fundamental method of measuring document similarity, was employed in this research. Cosine similarity computes the similarity between two vectors by taking the cosine angle between them. As cosine similarity focuses on the direction (pattern) of vectors when measuring similarity, it is suitable for making relatively fair comparisons between documents of varying lengths.

In this study, doc2vec was chosen as the preferred approach to measure the similarity scores of news articles. Doc2vec was deemed a better approach since this research focuses on comparing news articles by their contextual meaning. Using doc2vec, the extracted summary was compared with each document in the topic clusters, and the similarity was calculated for each. For each topic cluster, news articles were tokenized to determine their dimensionality and create a compact corpus. Since the corpus was separately created, the doc2vec model was trained specifically for each topic.

The variances of similarity for each topic are presented in the charts below. *Figure 6* displays the top 10 topics with the least similarity variance and their extracted topic examples. As similarity scores were closely measured, it was assumed that articles on these topics contain similar information. *Figure 7* shows the top 10 topics with the highest similarity variances. News articles in topic 155 contained news about soldiers and killings, but the context information may have been irregular due to its high variance. Based on these similarity variances, the following hypotheses were formulated: a topic with high variance indicates that the documents within that topic are highly diverse, while a topic with low variance suggests that news articles within that topic have a higher chance of having similar news.

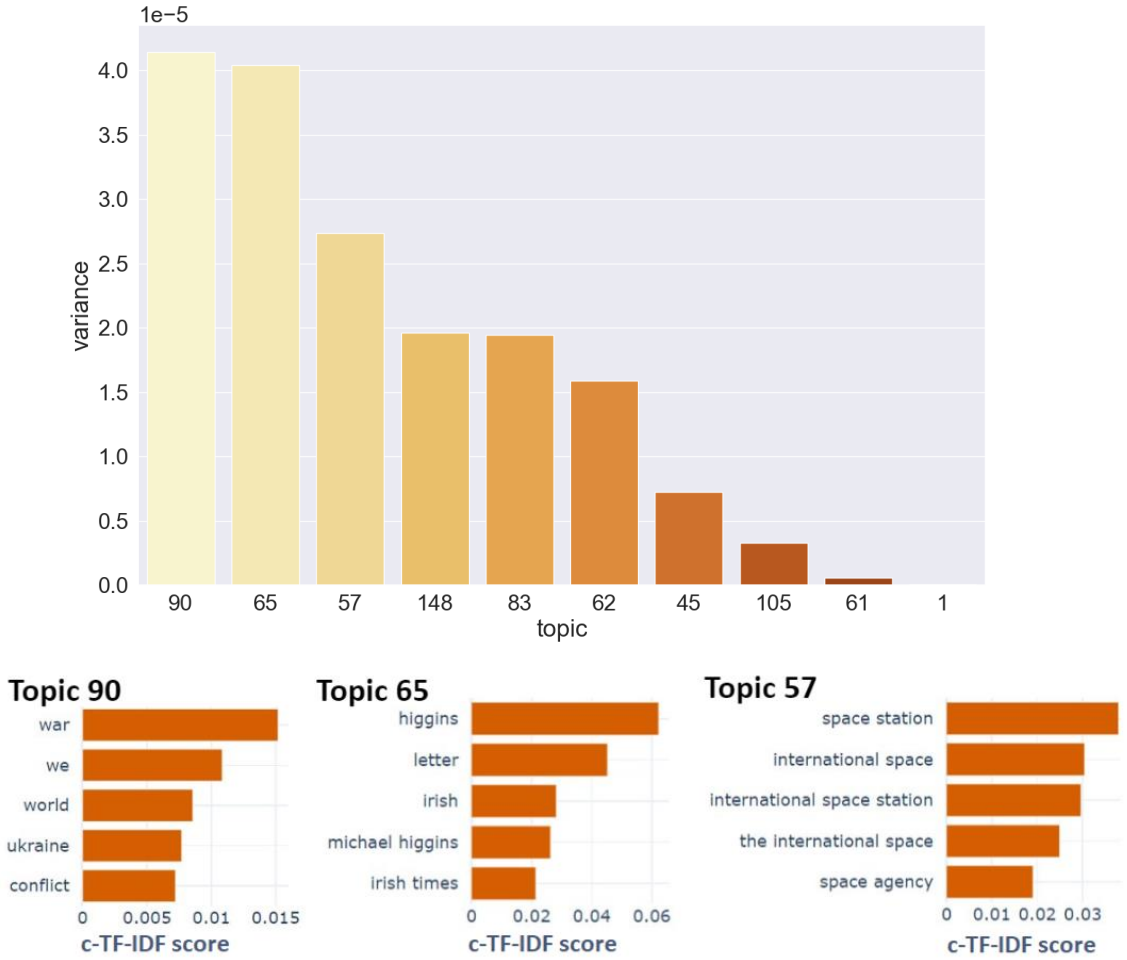


Figure 6. Topics with the least similarity variance and their topics

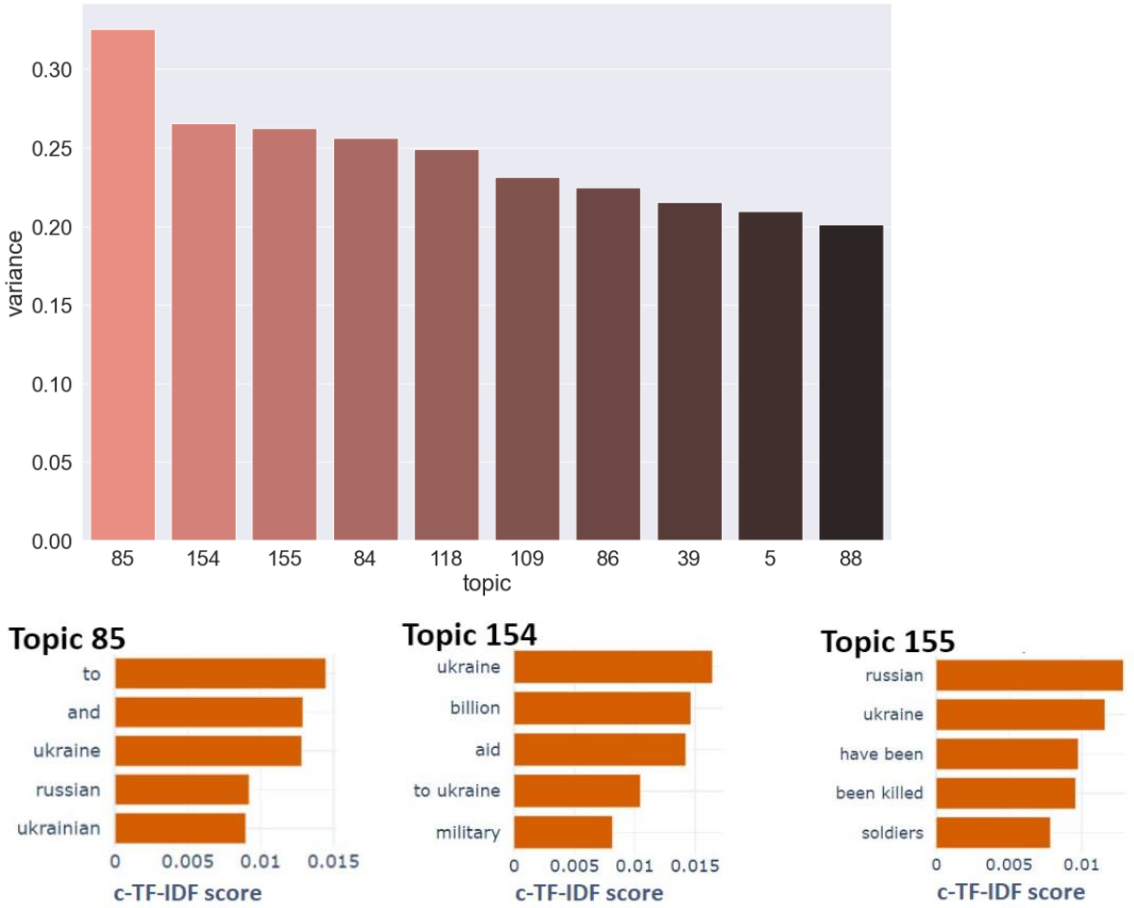


Figure 7. Topics with the most similarity variance and their topics

3.7 Ranking Topics

To determine the prevalence of misinformation within each topic, the similarity between each article per topic was computed. This similarity score was based on how closely the articles were related to the core summary of the topic. The resulting similarity variances were used to identify topics with a higher density of dissimilar articles, which in turn served as a metric for measuring the level of misinformation per topic.

The topics were ranked based on the distribution of similarity variances, with peaks indicating a dense cluster of articles within the topic as shown in *Figure 8 and 9*. The ranking scale ranged from 1 to 11, with rank 1 indicating the lowest variance and rank 11 indicating the highest variance. As per the hypothesis, articles on higher-ranked topics were assumed to have a higher potential for containing fake news. For instance, rank 11 topics, such as those related to the Ukraine-Russian war and India's prime minister (as shown in *Table 4*), suggested a greater likelihood of inconsistent and inaccurate information. In contrast, rank 1 topics, such as war crimes and hypersonic missiles (as shown in *Table 3*), were more likely to contain consistent and precise contextual information.

By employing this ranking, topics that might have a higher risk of containing inconsistent information could be identified, thereby providing a useful metric for further research. The outcome of this research presents a new approach for detecting potential fake news articles by identifying suspicious topics.

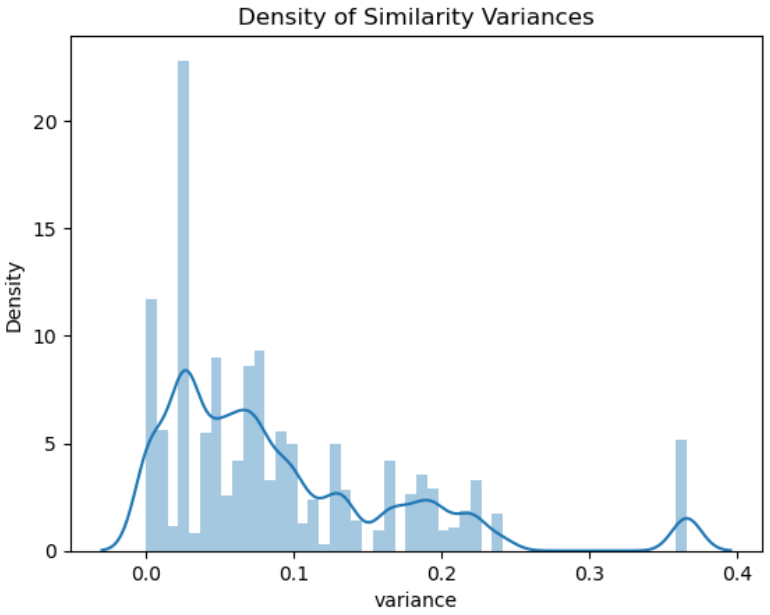


Figure 8. Distribution Plot of Similarity Variance

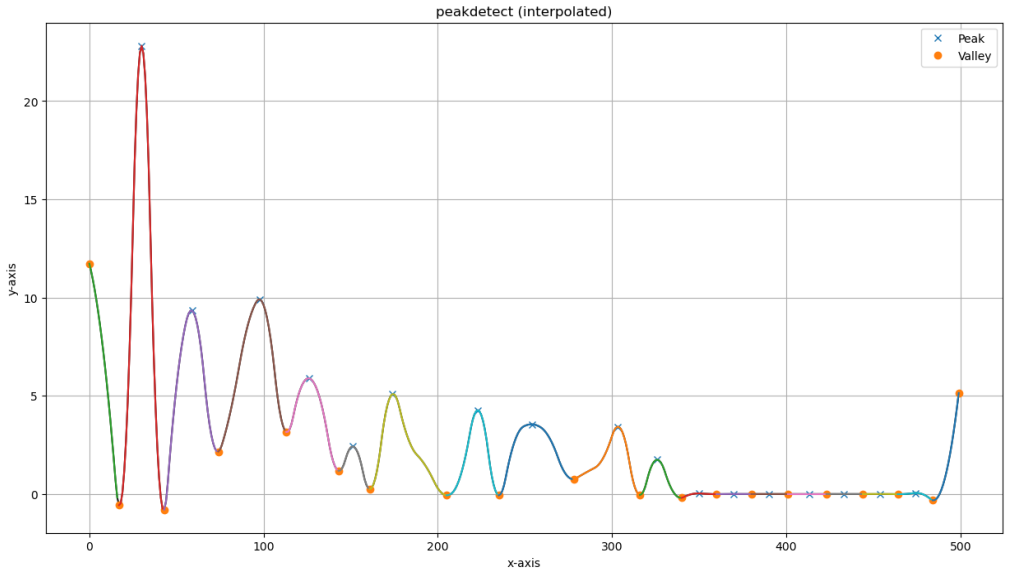


Figure 9. Peak Detection from *Figure 8,* distribution of similarity variance

Topic ID	Topic Lables
94	artist_mural_ukraine_paint
138	lavrov_russian foreign minister_lavrov say_sergei lavrov
127	ukraine rule_ukraine rule ceasefire_concession moscow_concession moscow russia
42	sanction_maximum sanction_maximum sanction russia_call maximum sanction
66	russian_war crime_russian soldier_ukrainian
101	laser_hypersonic_hypersonic missile_missile

Table 3. List of topics in Rank 1 and its topic labels

Topic ID	Topic Lables
80	ukraine_war_russia_russias
6	india_minister_prime minister_liz

Table 4. List of topics in Rank 11 and its topic labels

3.8 Evaluation

The evaluation of the proposed method presents a challenge due to the lack of labeled data, which makes it difficult to identify whether the news is fake or not. Given the near impossibility of manual fact-checking, an alternative labeled dataset, the Syrian War dataset, was utilized to validate and test the proposed hypothesis. This dataset contains 1000 labeled news articles about the Syrian War.

The proposed model was trained using the Ukraine War dataset, and the validation of the model was conducted using the Syrian War dataset. To determine whether the hypothesis can be rejected or not, the labeled Syrian War dataset was utilized for evaluation. Specifically, the performance of the proposed method was evaluated based on its ability to accurately identify fake news articles within the labeled dataset.

4 Results

The evaluation of an unsupervised model can be challenging, and as such, the proposed model was applied to the labeled Syrian War dataset to test the hypothesis. The hypothesis predicted that the ranking of news articles would have a higher percentage of fake news in the higher-ranking categories and a lower percentage in the lower-ranking categories. The results, as shown in *Figure 10*, supported the hypothesis, indicating that topics in rank 3 had the highest percentage of fake news on their topics, with 19.42% of news articles containing fake news. This finding aligned with the hypothesis that rank 3 had the highest likelihood of containing fake news. However, rank 1, which was expected to have the lowest percentage of fake news, had the second-highest percentage of fake news at 17.99%, indicating the presence of some errors.

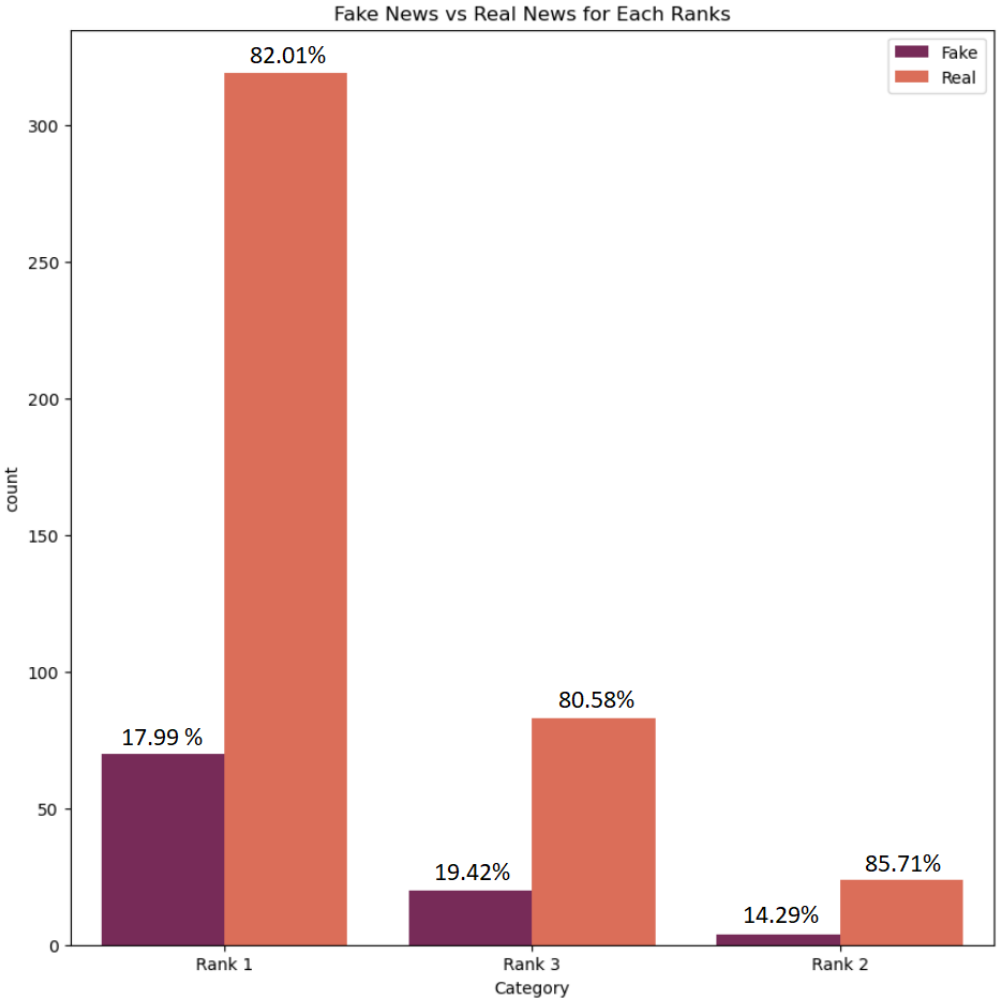
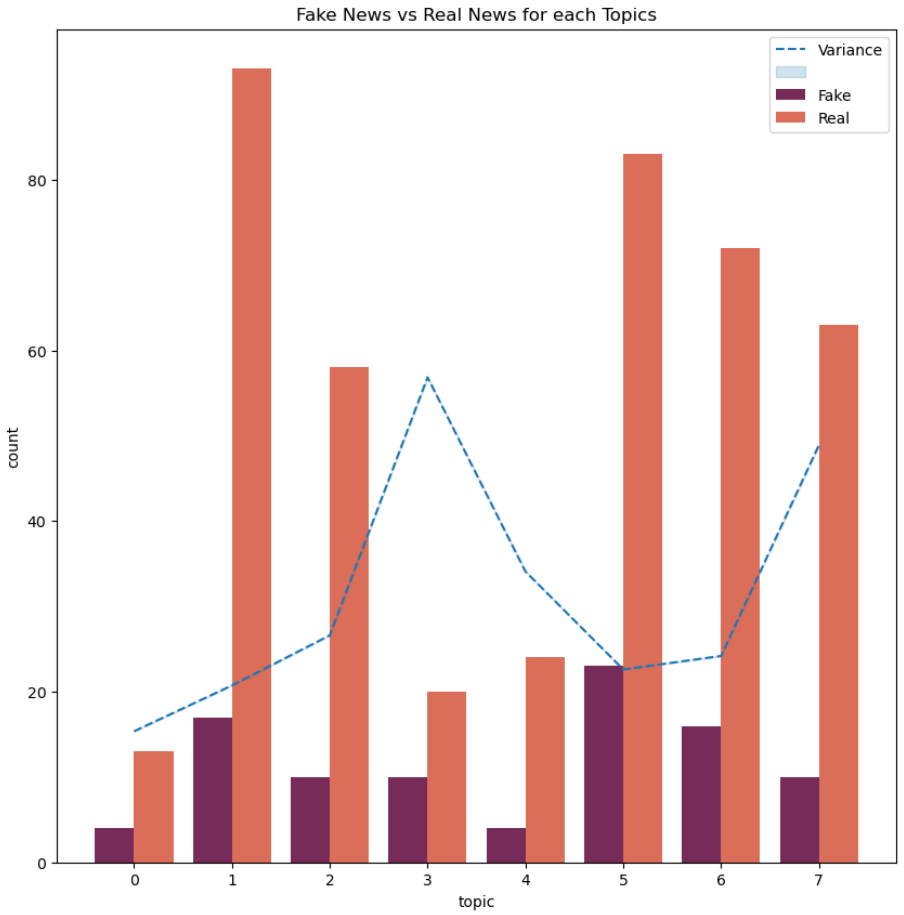


Figure 10. Evaluation: Rank Categorization of news articles for Syrian News data

To further investigate the errors, the topical level was analyzed to explore the relationship between similarity variance and the number of fake news articles. *Figure 11* presents the results of the topical-level analysis. Based on the hypothesis, topic 3, which had the highest similarity variance, was expected to have the highest percentage of fake news. Indeed, topic 3 had the highest percentage of fake news, with 33.33%. On the other hand, topic 0, which had the lowest similarity variance, was expected to have the lowest percentage of fake news. However, topic 0 had 23.53% of fake news, which was not the lowest percentage. This inconsistency suggests that the hypothesis did not entirely hold. Although the model was able to identify the topic and rank with the highest percentage of fake news by similarity variance, it was not able to detect the topics with a low percentage of fake news. As such, the hypothesis that topics with larger variances have a higher likelihood of containing fake news was rejected.



Topic	Fake News Ratio (%)	Real News Ratio (%)
0	23.53	76.47
1	15.4	84.6
2	14.71	85.29
3	33.33	66.67
4	14.29	85.71
5	21.7	78.3
6	18.18	81.82
7	13.7	86.3

Figure 11. Evaluation: Topical level categorization of news articles for Syrian News data

5 Discussion

In this study, the detection of fake news was conducted using unsupervised methods and unlabeled data. One critical aspect was the development of concise topic modeling to generate news topics. Given the lack of labels and the goal of comparing documents based on contextual information, reliable results depended on topic modeling. Although the hypothesis was rejected, this study explored various new approaches. The BERTopic tool was employed, and prior research has not applied it to fake news detection. Additionally, the multi-layer BERTopic was used to subdivide topic modeling into deeper levels of clustering, leading to enhanced model performance. As a result, this study improved the BERT model by applying multi-layer topic modeling, allowing it to extract topics at a denser level and reduce the number of outliers.

5.1 Limitations

The content-based unsupervised BERT model, while effective in classifying news articles based on their content, was limited in identifying fake news due to the lack of additional parameters such as the reliability of news sources, time of publication, languages, and geographic locations. Future model training could benefit from incorporating such parameters.

Moreover, the news article dataset used in this study did not include viewers' reactions, such as comments, shares, and likes. The number of views and the demographics of the viewers are also unknown, limiting the ability to analyze dissemination patterns. As a result, the news article dataset may have generic limitations compared to social network datasets.

Furthermore, this study was constrained by the unavailability of labeled news articles related to the Ukraine-Russia war. As a substitute, a dataset from a similar conflict in Syria was utilized for model evaluation. However, the evaluation dataset was limited in size, which may have affected the overall performance of the model.

5.2 Implications

After testing the hypothesis that the summary of each article cluster is factual, it was determined that the deviation from core summarization in a specific topic is not sufficient to measure the "fakeness" of news articles. The results suggest that relying solely on content summary in each cluster is inadequate in determining the reliability of the news. The implications of this study demonstrate that the proposed method of topic clustering with consecutive content summarization is not biased toward real or fake news. While fake news may prevail in certain topics, factual news may prevail in others, and all news, whether fake or true, can be clustered and have centralized patterns. This feature of disguisable fake news highlights the importance of developing new indicators for unsupervised fake news detection, as content cluster summarization and deviations are not effective indicators.

5.3 Ethical Concerns

In this study, news articles from various news sources worldwide were included to avoid any potential geopolitical bias in the results. However, the subjective nature of fake news detection should always be considered, as there is no clear definition of "fake" news. Furthermore, certain media sources may publish incomplete or misleading information in order to be the first to break a story, which can contribute to the spread of misinformation.

Future research on the prevalence of fake news in the Ukraine-Russia war could explore the potential impact of country of origin or geographic location on levels of accuracy in the model. For instance, how would the model perform when analyzing only Russian news sources versus American news sources? Additionally, it would be valuable to investigate the role of language in identifying fake news articles. However, the lack of available features for analysis was a limitation of this study, and further work is needed to better understand the ethical implications of fake news detection.

5.4 Future Research

The findings of this study suggest that the use of clustering and content summary alone may not be sufficient to identify patterns of fake news. To address this limitation, future research could be conducted to compare news bias across different social media platforms, countries, and languages. For instance, researchers could obtain two datasets of news articles from English and Russian media sources through web scraping. Using the same method as in this study, two models could be developed from each dataset, and news articles could be fed into both models. By comparing the results, it may be possible to identify instances of wartime news propaganda within certain clusters.

Furthermore, this multilayered method could be applied to improve the detection of fake news in social media. Additional layers of information, such as sources, user profiles, and dissemination patterns, could be incorporated into the model. This approach could serve as a potential avenue for future research.

6 Conclusion

In conclusion, this study proposed an unsupervised method for detecting fake news that involved clustering unlabeled news data by topic using BERT summarization and BERTopic modeling. The similarity scores for each topic were calculated from the extracted core summaries, and the deviation from the core summary in each topic was used to measure the "fakeness" of the articles in the given topic cluster.

The results of this research indicated that fake news cannot be detected by merely using the deviation from content clustering. This means that the core summarization from a specific topic can be either true or false, and fake news contents can prevail and be centralized in certain topics. Therefore, although this method did not successfully classify fake news, it provides valuable insights for future research on avoiding content clustering with a similarity deviation measurement approach in unsupervised fake news detection.

It is worth noting that while the proposed method did not achieve the desired level of accuracy in detecting fake news, it has the potential to be further improved through the use of more advanced NLP techniques and a larger and more diverse dataset. Moreover, the approach used in this study can be used as a starting point for developing more effective methods for detecting fake news, especially in the absence of labeled data or ground truth. Overall, this study contributes to the ongoing efforts to combat the spread of fake news and disinformation in today's digital age.

References

1. Nikolskaya, P. & Osborn, A. (2022). Russia's Putin Authorises "Special Military Operation" against Ukraine. In *Reuters, Moscow. Archived from the Original On, vol. 24*.
2. Stănescu, G. (2022). Ukraine Conflict: The Challenge of Informational War. In *Social Sciences and Education Research Review, vol. 9, issue 1, 146–148.*
3. Guadagno, R. E. & Guttieri, K. (2019). Fake News and Information Warfare: An Examination of the Political and Psychological Processes From the Digital Sphere to the Real World. In *Research Anthology on Fake News, Political Warfare, and Combating the Spread of Misinformation, IGI Global, Hershey, Pennsylvania (701 E. Chocolate Avenue, Hershey, Pennsylvania, 17033, USA), 2020, 167–168.*
4. Lukiv, J. (2022). Ukraine War: Nato Pledges to Provide More Weapons and Fix Power Grid. In *BBC News, November 29, 2022, www.bbc.com/news/world-europe-63798506.*
5. Liang, A. (2022). Ukraine War: Oil Prices Rise as Cap on Russian Crude Kicks In. In *BBC News, December 5, 2022, www.bbc.com/news/business-63855030.*
6. Ciuriak, D. (2022). The Role of Social Media in Russia's War on Ukraine. Available at *SSRN: http://dx.doi.org/10.2139/ssrn.4078863.*

7. Gelfer, A. (2018). Fake news: A definition. In *Informal Logic 38, 1 (2018), 84–117.*
8. Khaldarova, I, & Mervi P. (2016). Fake News: The narrative battle over Ukrainian conflict. In *Journalism Practice, vol. 10, no. 7, 2016, 891–901*.
9. Slackman, M. (2011). Syrian Troops Open Fire on Protestors in Several Cities. In *The New York Times, vol. 21, 2011*.
10. Lukiv, J. & Bachega, H. (2022). Ukraine War: Russian Missile Strikes Force Emergency Power Shutdowns. In *BBC News, December 6, 2022*.
11. Grant, N., & Hsu, T. (2022). Google finds 'inoculating' people against misinformation helps blunt its power. In *The New York Times. Retrieved October 29, 2022, from https://www.nytimes.com/2022/08/24/technology/google-search-misinformation.html*
12. Russia Criminalizes Sanctions Calls, 'Fake News' on Military (2022). In *Bloomberg News March 4, 2022.*
13. IFOP (2022). The French and Information on the Conflict between Russia and Ukraine.
14. Lovelace, A. G. (2022). Tomorrow's Wars and the Media. In *The US Army War College Quarterly: Parameters, 52(2), 117-134*.
15. YarAdua, S. M. (2022). Influence of Digital Images on the Propagation of Fake News on Twitter in Russia and Ukraine Crisis. Available at *SSRN: http://dx.doi.org/10.2139/ssrn.4062502*
16. Sardarizadeh, S. (2022). Ukraine war: False TikTok videos draw millions of views. In *BBC News April 25.*
17. Nasir, J.A., Khan, O.S. & Varlamis, I. (2021). Fake News Detection: A hybrid CNN-RNN based deep learning approach. In *International Journal of Information Management Data Insights. Volume 1, Issue 1, 100007*
18. Agrawal, C., Pandey, A., & Goyal, S. (2022). Fake news detection system based on modified bi-directional long short term memory. In *Multimedia Tools and Applications, 1-25.*
19. Saha, S.K. & Hasan, K.T. (2022). Improving Classification Efficiency of Fake News using Semi-Supervised Meth*od.* PREPRINT (Version 1) available at *Research Square [https://doi.org/10.21203/rs.3.rs-1201074/v1]*
20. Li, X., Lu, P., Hu, L., Wang, X., & Lu, L. (2022). A novel self-learning semi-supervised deep learning network to detect fake news on social media. In *Multimedia Tools and Applications, 81(14), 19341-19349*.
21. Zhang, C., & Abdul-Mageed, M. (2019). No army, no navy: Bert semi-supervised learning of arabic dialects. *In Proceedings of the Fourth Arabic Natural Language Processing Workshop (pp. 279-284)*.
22. P, D. (2021). On Unsupervised Methods for Fake News Detection. In *Data Science for Fake News. The Information Retrieval Series, vol 42. Springer, Cham. https://doi.org/10.1007/978-3-030-62696-9_2*
23. Gaglani, J., Gandhi, Y., Gogate, S., & Halbe, A. (2020). Unsupervised whatsapp fake news detection using semantic search. In *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS) (pp. 285-289). IEEE*.
24. Gangireddy, S. C. R., Long, C., & Chakraborty, T. (2020). Unsupervised fake news detection: A graph-based approach. In *Proceedings of the 31st ACM conference on hypertext and social media (pp. 75-83).*
25. Li, D., Guo, H., Wang, Z., & Zheng, Z. (2021). Unsupervised fake news detection based on autoencoder. In *IEEE Access, 9, 29356-29365*.
26. Yang, S., Shu, K., Wang, S., Gu, R., Wu, F., & Liu, H. (2019). Unsupervised fake news detection on social media: A generative approach. In *Proceedings of the AAAI conference on artificial intelligence (Vol. 33, No. 01, pp. 5644-5651).*
27. Hosseinimotlagh, S., & Papalexakis, E. E. (2018). Unsupervised content-based identification of fake news articles with tensor decomposition ensembles. In *Proceedings of the Workshop on Misinformation and Misbehavior Mining on the Web (MIS2)*.
28. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. In *arXiv preprint, arXiv:1810.04805*.
29. Jwa, H., Oh, D., Park, K., Kang, J. M., & Lim, H. (2019). exBAKE: Automatic fake news detection model based on bidirectional encoder representations from transformers (bert).

In *Applied Sciences, 9(19), 4062*.

30. Farokhian, M., Rafe, V., & Veisi, H. (2022). Fake news detection using parallel BERT deep neural networks. In *arXiv preprint, arXiv:2204.04793.*
31. Zhou, X., Wu, J., & Zafarani, R. (2020). Safe: similarity-aware multi-modal fake news detection (2020). In *arXiv preprint, arXiv: 2003.04981*.
32. Davies, M. (2016). Corpus of News on the Web (NOW). Dataset. Available online at: *https://www.english-corpora.org/now/*
33. Yvan S. *(2022). Webscrape3.py (Appendix)*
34. Liu, Y. (2019). Fine-tune BERT for extractive summarization. In *arXiv preprint, arXiv:1903.10318.*
35. Yoo, W. J. 19-06 Bert 기반 복합 토픽 모델(Combined Topic Models, CTM). In *Introduction to Deep Learning for Natural Language Processing, Wikidocs, Won Joon Yoo, https://wikidocs.net/161310*.
36. McInnes, L. & Healy, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. In *arXiv preprint, arXiv: 1802.03426.*
37. Grootendorst, W. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. In *arXiv preprint, arXiv: 2203.05794.*
38. McInnes, L. & Healy, J. (2017). Accelerated Hierarchical Density Based Clustering. In: *IEEE International Conference on Data Mining Workshops (ICDMW), IEEE, 33-42.*
39. Madhukar, B. (2020, August 21). Hands-on Guide To Extractive Text Summarization With BERTSum. *https://analyticsindiamag.com/hands-on-guide-to-extractive-text-summarization-with-bertsum/.*
40. Sanh,V., Debut, L., Chaumond, J. & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *arXiv preprint, arXiv:1910.01108.*
41. Miller, D. (2019). Leveraging BERT for Extractive Text Summarization on Lectures. In *arXiv preprint, arXiv:1906.04165.*

Appendix

Webscrape3.py

```
# Import packages
# Default packages
import time
import csv
import os
import json

# Preinstalled packages
import requests
import pandas as pd

# Define desired work folder, where you want to save your .csv files
# Windows Example
os.chdir('/Users/ysojdehei/Documents/GradSchool/Capstone/Fake News Classifier')

# URL of our News API
base_url =
'https://na01.safelinks.protection.outlook.com/?url=https%3A%2F%2Fapi.newscatcherapi.com%2Fv2%2Fsearch&amp;data=05%
7C01%7C%7Ccb3ce7b65d844203e74608dad421e8e9%7C84df9e7fe9f640afb435aaaaaaaaaaa%7C1%7C0%7C638055539348857
426%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3
000%7C%7C%7C&amp;sdata=2WRkEMnUvUtrPETU2rVESG9nxpc4cuxMIU7pGCq7WRw%3D&amp;reserved=0'

# Your API key
X_API_KEY = 'wyxhlFuAuTwcd_gtZrVny-LrV-axlwUkQn4o8Ks2ZRA'

# Put your API key to headers in order to be authorized to perform a call
headers = {'x-api-key': X_API_KEY}

# Define your desired parameters
params = {
    'q': 'War AND Ukraine AND Russia',
    'lang': 'en',
```

```
    'to_rank': 10000,
    'page_size': 100,
    'page': 1
    }

# Make a simple call with both headers and params
response = requests.get(base_url, headers=headers, params=params)

# Encode received results
results = json.loads(response.text.encode())
if response.status_code == 200:
    print('Done')
else:
    print(results)
    print('ERROR: API call failed.')

# Variable to store all found news articles
all_news_articles = []

# Ensure that we start from page 1
params['page'] = 1

# Infinite loop which ends when all articles are extracted
while True:

    # Wait for 1 second between each call
    time.sleep(1)

    # GET Call from previous section enriched with some logs
    response = requests.get(base_url, headers=headers, params=params)
    results = json.loads(response.text.encode())
    if response.status_code == 200:
        print(f'Done for page number => {params["page"]}')


        # Adding your parameters to each result to be able to explore afterwards
        for i in results['articles']:
            i['used_params'] = str(params)


        # Storing all found articles
        all_news_articles.extend(results['articles'])

        # Ensuring to cover all pages by incrementing "page" value at each iteration
        params['page'] += 1
        if params['page'] > results['total_pages']:
            print("All articles have been extracted")
            break
        else:
            print(f'Proceed extracting page number => {params["page"]}')
    else:
        print(results)
        print(f'ERROR: API call failed for page number => {params["page"]}')
        break

print(f'Number of extracted articles => {str(len(all_news_articles))}')



# Generate CSV from Pandas table
# Create Pandas table
pandas_table = pd.DataFrame(all_news_articles)

# Generate CSV
pandas_table.to_csv('extracted_news_articles4.csv', encoding='utf-8', sep=';')
```