News and Market Sentiment Analysis Exam

Gustav Emil Lange
Gulan24@student.sdu.dk

12 - 19 Dec, 2025

Contents

1 Introduction

In order to conduct a risk and media analysis for the large organisation, the initial step is to examine the data set that is being used as a basis for the analysis[2]. The assignment involves the analysis of two distinct data sets: one comprising fake news and the other true news. The data set under consideration was collected between 2015 and 2018 and consists of articles from American news outlets that were determined to be either true or fake.

The objective of this study is to demonstrate the distinguishing characteristics of true news and fake news in terms of style, tone and topics. This analysis will explore how such misinformation can potentially distort sentiment-based trading models and reputation risk models.

In order to accomplish this objective, it is first necessary to conduct a thorough exploration of the available data. This will provide an overview of the number of documents, the average length of our texts, and the number of words, among other metrics. In addition, an analysis will be conducted to ascertain which words are present most frequently in the texts, and which words are most indicative of fake news and real news, respectively. Consequently, this approach can also be employed in the context of sentiment analysis. Finally, in our data exploration, we will perform word embedding using the TF-IDF method to see how similar the two classes are.

Following the exploratory analysis of the data, a sentiment analysis will be conducted using a robust pre-trained model, DistilBERT. The following methodology will be employed to evaluate the extent to which the two data sets are negative or positive. Furthermore, an analysis of the text corpus will be conducted to identify the linguistic characteristics that distinguish between the two types of news. In combination with the DistilBERT sentiment analysis, this will facilitate the identification of the tone and style that is characteristic of the two distinct categories of news.

In the subsequent stage of the research, an LDA model will be employed to analyse which topics are typical of real and true news, respectively. The objective of this analysis is to ascertain whether there is a significant difference between the topics in the two different classes. Furthermore, an analysis will be conducted of the terms that are most prevalent in the two classes, and, based on this, a manual form of topic modelling will be performed.

In the concluding section of the analysis, a Naive Bayes classifier and a DistilBERT classifier are developed. Here we investigate potential differences between the performance of a model trained exclusively on the available dataset and that of a model pretrained on a large-scale corpus, such as DistilBERT. This comparison aims to put light on the potential influence of fake news on sentiment-based trading strategies and why we also should be careful when concluding on how good pre-trained models are.

ChatGPT has been utilised in certain contexts to support the coding of models and the refinement of code structure. All analyses, methodological choices, and interpretations presented in this report have been developed by me, who bears full responsibility for the results and conclusions. In addition, ChatGPT has been used for language polishing and grammatical corrections. The use of ChatGPT is reflected where relevant in notebook sections and the accompanying text.

2 Preprocessing

Prior to the exploration of the data, an explanation of the types of pre-processing that have been carried out is necessary. This approach has been formulated on the basis of an in-depth analysis of the content of the individual texts, in conjunction with Spacy preprocessing, a methodology with which the class has been introduced to [8].

In the initial phase of the process, it is imperative to eliminate unnecessary columns. These include the subject, date, and title columns. This is done to simplify the dataset. Furthermore, an analysis of the texts reveals the presence of a substantial amount of 'metadata'. This encompasses links to Twitter posts, YouTube content, and other relevant materials. Furthermore, a number of filler words such as screenshot, image, getty, flickr, etc. were identified. It is considered that these words constitute 'noise' and could have an effect on the models. It is determined that their removal is necessary, as they are considered to provide minimal contextual information of the texts.

Upon examination of the dataset concerning true news, it appears that the data has been sourced from Reuters, as it is preceded by an American city name and the Reuters brand. This is also considered to be a form of noise, as it would likely constitute a relatively straightforward method for our models to differentiate between true and fake news. Consequently, these are eliminated. Furthermore, both datasets were examined for the presence of duplicates. These were identified and subsequently removed.

spaCy is used as the central tool for preprocessing, as the library enables tokenization and lemmatization based on pre-trained linguistic models. By processing the text with spaCy, each word is analyzed in its linguistic context, ensuring a more accurate division into tokens and consistent reduction to base forms. This helps to reduce variation in word forms and thus reduce the dimensionality of the subsequent text basis. In addition, it is also used because it is designed for efficiency and ease of use in large-scale text processing tasks.The inspiration to use spaCy comes from the LDA code that our teacher shares in his GitHub repository and from slides in the teaching[8].

The preprocessing pipeline takes a collection of raw text documents as input. First, the text is cleaned up, with everything converted to lowercase and technical and structural noise removed using regular expressions. This includes URLs, media references, and other web artifacts that do not contribute linguistic or semantic content (as mentioned earlier). The purpose of this step is to ensure that the subsequent linguistic processing is based on text that is as uniform and relevant as possible.

Next, spaCy is used to tokenize and lemmatize the cleaned text. spaCy divides the text into individual tokens and reduces each token to its base form, which reduces variation in word forms and contributes to a more consistent vocabulary. During this process, stop words, non-alphabetic tokens, and very short words are filtered out to reduce noise and preserve words with higher information value.

The output of the preprocessing pipeline is the variable *clean_text*, which contains the final, cleaned text representation. *clean_text* thus consists of semantically meaningful, lemmatized words and forms a compact and structured text basis, where the focus has shifted from raw text to content-based linguistic information, suitable for further analysis [6].

3 Data Exploration

After preprocessing, data exploration is performed. The purpose of this is to get to know our two data sets better, but also to see if there are any differences to be found based on this basic analysis. The concatenated dataset consists of 21,192 true news articles and 17,445 fake news articles. This means that we have a majority of true news articles. The average length of the articles in the two datasets is examined before and after preprocessing:

	char_count_text	word_count_text	char_count_clean	word_count_clean
fake	2549.81	425.26	1421.15	200.85
true	2357.07	381.71	1503.66	205.90

Table 1: Average text length before and after preprocessing

Preprocessing reduces the text length in both datasets, indicating that part of the raw text consists of words and elements with low semantic value. At the same time, it can be seen that fake texts are longer than true texts in their raw form, but that this difference almost disappears after preprocessing. In the cleaned text, true texts are even slightly longer. This could indicate that fake texts contain relatively more redundant language, while true texts retain a larger proportion of meaningful content words.

In addition, we explore which words occur most frequently in the two datasets. The lists show the 10 (top 50 in script) most frequently occurring words in true and fake news, respectively, after preprocessing, where each word is counted according to its total occurrence across all documents in the dataset. They thus provide an overview of the dominant vocabulary in the two text collections and serve as an initial exploratory insight into both content and linguistic style:

Word	Count
say	106816
trump	54109
president	28387
state	25163
year	22325
government	19675
republican	17704
house	16891
new	15800
tell	15317

Table 2: Top 10 most frequent words in the true news dataset

Word	Count
trump	65025
say	32256
people	20770
president	20473
donald	14882
like	14862
go	14475
know	13303
clinton	13264
year	13221

Table 3: Top 10 most frequent words in the fake news dataset

The results show that fake and true news share a large proportion of their most frequently used words, especially names of political actors, institutions, and general action verbs. This reflects that both datasets deal with the same overall context. Since the data is from 2015–2018, which covers the US presidential election campaign and Donald Trump's election and early presidency, it's expected that names like Trump, Clinton, and Obama show up a lot in both datasets.

At the same time, linguistic differences can already be identified at this level. true news contains relatively more formal and institutional terms, such as government, official, court, and administration, while fake news uses more informal and subjective words such as like, think, know, and want. This indicates that the differences between fake and true news are less related to the choice of topic and more to linguistic style and framing [6].

To examine these linguistic differences more systematically, a text corpus difference analysis is then performed. After preprocessing, fake and true news are treated as two separate text corpora, where word frequencies are compared across classes. To ensure that the analysis is based on stable and representative patterns, only words that appear at least 20 times in both fake and true news are included. The analysis thus focuses on relative differences in word usage rather than on whether certain words appear or not [6][7].

An analysis of the words most characteristic of fake news reveals a clear pattern in linguistic tone and style. A large proportion of the highest-ranked words are clearly judgmental, emotional, or confrontational in nature. Examples include words such as disgusting, stupid, awful, liar, hypocrisy, hateful, racist, and sexist, all of which express explicit negative judgments. These words contribute to a linguistic framing in which people or events are not merely described, but actively evaluated and condemned.

In addition, an informal and conversational style is evident in fake news through words such as literally, yeah, hey, okay, funny, and epic. This points to a more colloquial register that differs from traditional journalistic writing style. This linguistic style may contribute to a more emotionally engaging and polarizing presentation.

In contrast, the list of words most characteristic of true News shows a significantly different linguistic pattern. Here, geographical names (Bangladesh, Beijing, Spain, Lebanon), international political actors (Macron, Theresa, Jinping, Erdogan, Duterte), and institutional and formal terms such as parliament, ministry, referendum, bilateral, and tariff dominate. This reflects a more global, institutional, and reporting focus. The language in true news appears more factual and descriptive, with an emphasis on international relations, political processes, and official actors. Overall, the interpretation of the characteristic words shows that fake and true news differ significantly in linguistic tone and style: fake news is characterized by evaluative, emotional, and informal language, while true news uses a more formal, institutional, and informative register. Next we will analyse how similar the two classes are, by using the cosine similarity measure.

4 Word Embeddings and Cosine Similarity

To examine linguistic similarity between fake and true news at the document level, each article is represented as a TF-IDF embedding, where the text is converted into a numerical vector reflecting the relative importance of words in the article and in the full corpus. For each class, a centroid is computed as the average of all document vectors, representing the typical linguistic profile of fake and true news respectively. Cosine similarity is then used to measure how close articles are to each centroid and how similar the two class centroids are. The difference between an article's similarity to its own class centroid and to the opposite class, referred to as the centroid margin, is used as a simple measure of linguistic alignment. The method is unsupervised and focuses on describing overall linguistic structure rather than prediction of a class, which we will cover later in the analysis[1] [8][7].

The results show that articles in both classes are, on average, linguistically closest to their own class centroid. Fake news articles have a mean cosine similarity of 0.1377 to the fake centroid and 0.1026 to the true centroid, resulting in a centroid margin of -0.0351. This indicates that fake news articles tend to resemble the general fake-news language profile more than the true-news profile. Likewise, true news articles show a mean cosine similarity of 0.1348 to the true centroid and 0.1004 to the fake centroid, yielding a positive centroid margin of 0.0344. The near-identical size of the margins across the two classes suggests a stable and systematic linguistic distinction.

Although the absolute cosine similarity values are relatively low and lie in a narrow range between approximately 0.10 and 0.14, this also indicates substantial linguistic overlap between fake and true news. This overlap reflects that both types of articles address similar topics and share many stylistic conventions. The central result is therefore not the absolute similarity, but the consistent difference of around 0.03–0.04 between similarity to the own class and to the opposing class. This shows that fake and true news largely inhabit the same linguistic space, yet still differ in how topics are framed and organized linguistically rather than in what topics are covered. In the next part we will take a look on the topics and see if they look pretty much the same, with the use of LDA topic modelling.

Label	sim_fake_centroid	sim_true_centroid	centroid_margin
Fake (0)	0.137682	0.102566	-0.035116
True (1)	0.100427	0.134811	0.034384

Table 4: Mean cosine similarity between articles and class centroids for Fake and True news

5 Topic Modelling - LDA

Latent Dirichlet Allocation (LDA) is an unsupervised, generative probabilistic topic model that aims to identify latent topics in a text corpus. The model assumes that each document can be described as a mixture of several topics, and that each topic consists of a probability distribution over words. LDA operates on a bag-of-words representation, where word order and syntactic information are ignored, and topics are estimated based on statistical patterns in the co-occurrence of words across documents [7][8][4]. In this analysis, LDA is applied separately to fake and true news, where the pre-processed text is first converted into a document-term matrix using a CountVectorizer. For each dataset, an LDA model with eight topics is trained, which is a predetermined and exploratively selected hyperparameter. The model is used exclusively for interpretation, where the most probable words for each topic are used to identify dominant themes and compare the thematic organization in fake and true news.

Topic	Top words
1	people, woman, say, muslim, black, man, right, group, video, like, life, white
2	trump, president, obama, people, say, vote, white, republican, right, go, image, donald
3	clinton, hillary, state, email, say, fbi, election, investigation, report, department, information, campaign
4	say, cruz, candidate, campaign, republican, like, school, sander, hillary, clinton, woman, go
5	police, say, gun, officer, year, report, kill, shoot, tell, attack, shooting, man
6	trump, donald, say, president, image, news, go, know, like, tweet, think, time
7	world, president, obama, war, united, government, american, new, say, state, country, military
8	state, people, bill, law, say, million, year, pay, health, federal, plan, government

Table 5: LDA topics for fake news (8 topics)

Topic	Top words
1	say, trump, president, russia, russian, house, official, campaign, white, washington, committee, intelligence
2	say, north, trump, united, korea, states, nuclear, china, president, trade, deal, sanction
3	say, court, eu, government, law, trump, order, britain, judge, case, justice, ban
4	say, year, government, china, myanmar, right, country, state, official, reuters, people, security
5	say, tax, house, republican, senate, bill, congress, state, trump, plan, republicans, year
6	say, state, government, force, people, islamic, attack, group, reuters, year, kill, syria
7	say, trump, company, year, million, president, new, bank, business, billion, percent, mexico
8	trump, say, election, party, republican, clinton, campaign, presidential, vote, state, candidate, new

Table 6: LDA topics for true news (8 topics)

The LDA analysis shows that the overall topics in fake and true news overlap to a large extent. Both datasets contain topics centered around the same political actors and events, including Donald Trump, Hillary Clinton, election campaigns, government institutions, and international conflicts. This confirms the earlier observations from the data exploration that fake and true news primarily deal with the same thematic content, reflecting the common political context in the period 2015–2018. Seen in this light, the LDA results indicate that the differences between the two classes cannot be explained by significantly different topic choices.

Although the topics are thematically similar, there are differences in the linguistic organization and framing of these topics. Topics in fake news are more characterized by a focus on individuals, identity categories, and more informal or evaluative language, while topics in true news contain more institutional, legal, and geopolitical terms. This is consistent with the results of both the corpus difference analysis we have made earlier and supports the conclusion that the difference between fake and true news is primarily manifested in linguistic style and framing rather than in thematic content. The LDA analysis thus serves as further confirmation that fake and true news share topics but differ in the way these topics are linguistically structured and communicated.

6 Sentiment

Following the exploratory data analysis and the topic modelling analysis, a sentiment analysis based on the pre-trained model distilbert-base-uncased-finetuned-sst-2-english is used. The model receives raw text as input and performs the necessary internal preprocessing itself, including tokenization and conversion to numerical representations. The analysis is thus based on the original text of the articles and not on the preprocessed text used in the other analyses. The model is pre-trained on large English text corpora and subsequently fine-tuned on the Stanford Sentiment Treebank (SST-2), which consists of short film reviews annotated with binary sentiment (positive or negative)[3][7].

When applying the model to the present dataset, it is important to note that the model uses a maximum of the first 512 tokens from each article, as longer texts are truncated. This means that the sentiment assessment does not necessarily see all tokens in the individual news articles. At the same time, the model has not been trained on news articles, but on a different text domain, which may affect the model's interpretation of sentiment in journalistic language. These circumstances mean that the results should not be understood as precise measurements of the emotional tone of news texts, but rather as indicative patterns. We use the model to highlight potential differences in linguistic framing and, despite the above, we also use it to as an indicater of the emotional charge between fake and true news, but with the caveat that it is not domain-specific.

	Negative sentiment	Positive sentiment
fake news	86.0%	14.0%
true news	81.8%	18.2%

Table 7: Distribution of sentiment labels for fake and true news based on DistilBERT sentiment analysis

The results of the sentiment analysis indicate that both fake and true news articles are predominantly classified as negative in sentiment. Specifically, approximately 86% of fake news articles are classified as negative, while about 14% are classified as positive. For true news, a similar but slightly less skewed distribution is observed, with around 82% of articles classified as negative and approximately 18% as positive. Overall, these results suggest that fake news exhibits a more negative sentiment profile on average than true news. However, the difference between the two classes is relatively moderate, indicating that negative sentiment is a common characteristic across both types of news content.

However, it is important to note that the predominantly negative sentiment is not necessarily a characteristic unique to fake news. News articles, especially in political journalism, often deal with conflicts, controversies, and issues, which can naturally result in a more negative linguistic tone. As our articles primarily deal with American politics, as mentioned earlier, this is particularly relevant in our case. The high proportion of negative sentiment in both data sets may thus reflect the genre and subject area rather than a clear distinction between fake and true news.

The difference in sentiment distribution between fake and true news can therefore primarily be interpreted as a difference in the degree of negative framing rather than a binary difference in emotional tone. fake news seems to use more polarized and negatively charged language to a slightly greater extent, which is consistent with the other findings from the data exploration, including the analysis of which words characterize the two classes. Here, words such as sexist, racist, etc. were very prominent for the fake news. Overall, the sentiment analysis thus supports the conclusion that the differences between fake and true news relate more to linguistic tone and framing than to thematic content, while the results should be interpreted with caution in light of the model's domain limitations.

7 Classifiers

Until now, the analysis has focused on identifying the linguistic and thematic characteristics of fake and true news. Through data exploration, corpus analysis, and topic modeling, it has been shown that the two classes deal with largely the same topics but differ in tone, style, and linguistic framing. Based on these findings, it is relevant to investigate whether such differences can also be exploited operationally for automatic classification. Therefore, a Naive Bayes classifier is first used, which is well suited for text data and functions as a transparent baseline model that explicitly exploits differences in word occurrences between classes. Next, a more complex, pretrained model in the form of DistilBERT is tested, which represents a contextual approach to text classification. The two models thus enable a comparison between a classic, word-based method and a modern transformer-based approach.

7.1 Naive Bayes

A multinomial Naive Bayes classifier based on a bag-of-words representation of the preprocessed text is used as the baseline classification model. The model is trained on 80%

of the dataset and evaluated on the remaining 20% using a stratified split. Naive Bayes assumes conditional independence between words given the class, which makes it a simple model; however, in practice it is often effective for text classification tasks where differences in word occurrences carry significant informative value. [8]

The model achieves relatively solid classification results with an overall accuracy of 0.923, indicating that correct classifications are made in the vast majority of cases. For fake news, a precision of 0.91 and a recall of 0.93 are achieved, while true news attains a precision of 0.94 and a recall of 0.92. Recall measures the proportion of actual articles in a given class that the model correctly identifies. A recall of 0.93 for fake news therefore implies that 93% of genuinely fake articles are correctly classified as fake, while only 7% are misclassified as true. Similarly, a recall of 0.92 for true news indicates that the vast majority of true news articles are correctly identified. The high recall values for both classes suggest that the model rarely overlooks articles from either class, which is particularly important in contexts where failing to identify misinformation may have significant consequences for risk assessment and decision support.

Class	Precision	Recall	F1-score
Fake (0)	0.91	0.93	0.92
True (1)	0.94	0.92	0.93

Table 8: Precision, recall and F1-score for the Naive Bayes classifier

The confusion matrix shows that 3,231 fake news articles and 3,906 true news articles are classified correctly. At the same time, 260 fake articles are misclassified as true and 333 true articles as fake, indicating relatively few errors compared to the total number of observations.

	Predicted Fake	Predicted True
Fake (0)	3231	260
True (1)	333	3906

Table 9: Confusion matrix for the Naive Bayes classifier

All together, these results show that a simple, word-based classifier can exploit systematic differences in lexical choice between fake and true news with high accuracy. However, the strong performance should be interpreted with caution. Because Naive Bayes relies on a bag-of-words representation and assumes conditional independence between words, it captures correlations in word usage rather than deeper semantic or contextual meaning. As a result, the model may overemphasize surface-level lexical cues that are specific to the dataset or time period, potentially limiting its robustness and generalizability. To examine whether a model with contextual language understanding can provide a more nuanced and potentially more robust distinction between fake and true news, the analysis therefore proceeds with a pretrained transformer-based model, DistilBERT, which leverages contextualized representations learned from large-scale text corpora.

7.2 DistilBERT classifier

The DistilBERT classifier is used to assess how effectively transformer-based language models can perform text classification compared to simpler word-based methods such as Naive Bayes. DistilBERT is a compressed version of BERT, pretrained on large English text corpora using masked language modeling, and is chosen for its favorable balance between computational efficiency and representational power. The model takes raw text as input and performs its own preprocessing through subword tokenization, producing contextual embeddings in which word representations depend on their surrounding context. Due to a maximum input length of 512 tokens, longer news articles are truncated, meaning that not all textual content necessarily contributes to the classification [5] [7].

Despite the strong baseline performance, DistilBERT is included to examine whether contextual language understanding provides additional value beyond lexical cues. While earlier analyses suggest that differences between fake and true news largely manifest at the lexico-stylistic level, applying DistilBERT allows for a robustness check of these findings and enables a principled comparison between interpretable word-based models and more complex contextual approaches in misinformation detection.

The DistilBERT classifier achieves near-perfect performance, with an accuracy of approximately 0.998 on the test set and almost identical results on the training set. Precision, recall, and F1-scores are effectively equal to 1.00 for both classes, indicating that the model makes very few classification errors. While this demonstrates the strong capacity of transformer-based language models for text classification, such results also warrant a critical interpretation. One plausible explanation for the exceptionally high performance is that DistilBERT is pretrained on extremely large and diverse text corpora, enabling it to recognize linguistic patterns, stylistic cues, and semantic structures that generalize well to news data.

Class	Precision	Recall	F1-score
Fake (0)	1.00	1.00	1.00
True (1)	1.00	1.00	1.00

Table 10: Classification performance of the DistilBERT classifier

The confusion matrix shows that DistilBERT classifies almost all articles correctly. Out of 3,491 fake news articles, only 9 are misclassified as true news, while 7 out of 4,239 true news articles are misclassified as fake.

	Pred Fake	Pred True
Fake (0)	3482	9
True (1)	7	4232

Table 11: Confusion matrix for the DistilBERT classifier

As a result, the model may already encode substantial prior knowledge relevant to political discourse and journalistic language. At the same time, the near-perfect scores raise concerns about potential overfitting to dataset-specific regularities or annotation artifacts, particularly given that fake and true news in this dataset differ strongly in explicit lexical and stylistic features. Moreover, the truncation of long articles to 512 tokens means that the model's decisions may rely disproportionately on introductory sections, which often contain strong framing signals. Compared to Naive Bayes, DistilBERT clearly achieves higher predictive performance, but this comes at the cost of reduced interpretability and increased computational complexity. Importantly, the strong performance of the simpler Naive Bayes model suggests that much of the discriminative signal is already present at the lexical level, implying that DistilBERT's gains primarily reflect its ability to exploit and generalize these patterns rather than uncover fundamentally different linguistic distinctions.

8 Conclusion

In relation to the task description's focus on misinformation and its potential impact on sentiment-based trading, the data exploration shows that fake and true news largely share the same thematic starting point. Both datasets are strongly dominated by political actors, institutions, and events from the same historical period, which means that misinformation cannot be identified based on topic selection alone. This insight is crucial, as it indicates that systems based primarily on subject- or topic-based analysis will have difficulty distinguishing between fake and true news in political contexts.

In contrast, lexical and stylistic analyses show that linguistic choices are a much stronger signal. The high performance of the Naive Bayes classifier demonstrates that differences in word choice, tone, and style are sufficiently systematic to enable high-precision classification. This is particularly relevant in relation to the task description, as it indicates that sentiment-based models and risk assessment systems may be influenced by linguistic framing rather than factual content, which can lead to biased assessments if this dimension is not explicitly addressed.

The application of DistilBERT shows that even advanced transformer-based language models largely confirm the patterns already identified using simple, lexically based methods. The near-perfect classification performance indicates that misinformation in this dataset can be largely recognized through linguistic and stylistic signals. At the same time, the results raise important questions about the generalizability of the models and the risk that they primarily exploit dataset-specific regularities rather than more general characteristics of misinformation. In relation to the task, this emphasizes that high classification performance does not necessarily reflect deeper linguistic understanding, and that the use of such models in practice should be accompanied by critical considerations regarding robustness, interpretability, and the potential consequences for decision support in financial and media contexts.

The overall conclusion is that when sentiment-based trading and reputation risk models use news articles as input, it is crucial to take into account that some of this content may potentially be misinformation. The analysis shows that the differences between fake and true news are largely reflected in tone, linguistic framing, and the degree of evaluative language rather than in the choice of topics. Models used in this context should therefore be particularly sensitive to such linguistic and stylistic differences, as these are key indicators of misinformation and can have a significant impact on risk assessments and decision support in financial and media contexts. In addition, it is important to take into account that analyses and models such as those used in this assignment are highly dependent on the temporal context of the data on which they are trained and evaluated. The dataset used covers the period 2015–2018, which is clearly reflected in the dominant political topics and actors appearing in the articles, including frequent references to Hillary Clinton, Donald Trump, and contemporary international conflicts. Using more recent text data would likely reveal a different thematic focus, for example on Ukraine, Russia, and Vladimir Putin. This emphasizes that the results should be interpreted in light of the historical context of the dataset and that models based on news text must necessarily be used with an awareness that linguistic patterns and political references change over time.

References

[1] Adem Akdogan. Word embedding techniques: Word2vec and tf-idf explained. `https://medium.com/data-science/word-embedding-techniques-word2vec-and-tf-idf-explained-c5d02e34d08`, 2021.

[2] Clément Bisaillon. Fake and real news dataset. Kaggle dataset, 2018.

[3] DT12the. Dt12the/distilbert-sentiment-analysis. `https://huggingface.co/DT12the/distilbert-sentiment-analysis`, 2024.

[4] Thushan Ganegedara. Light on math: Machine learning intuitive guide to latent dirichlet allocation. Towards Data Science (Medium), 2021.

[5] Hugging Face. Distilbert. `https://huggingface.co/docs/transformers/main/en/model_doc/distilbert`, 2025.

[6] Gustav Emil Lange. exam-news-gel.ipynb. Source code, 2025.

[7] OpenAI. Chatgpt. Large language model, 2025. Accessed via chat.openai.com.

[8] Christian Vedels. News and market sentiment analytics. GitHub repository, 2024.