## Original Paper

# Fuzzy Rank-based Ensemble Approach of Language Models for Detecting Misinformation of Long-Term Effects of COVID-19

## Abstract

**Background:** During the COVID-19 pandemic, the continuous spread of misinformation on the internet poses an ongoing threat to public trust and understanding of epidemic prevention policies. Even with the pandemic under control, information regarding the risks of long-term COVID-19 and reinfection still needs to be integrated into COVID-19 policies.

**Objective:** The study introduces a deep learning approach combing language models with fuzzy rank-based ensemble method for detecting misinformation concerning the long-term impacts of COVID-19.

**Methods:** The data, comprising 566 genuine and 2361 fake samples, was collected and refined from reliable open sources using data processing techniques. Afterward, deep learning models such as HAN, BERT, and XLNet were trained based on the collected data to detect misinformation about the long-term impacts of COVID-19. This study employed the fuzzy rank-based ensemble technique, combining different deep models to improve the performance further.

**Results:** After training on the dataset, various classification methods were evaluated on the test set, including the fuzzy rank-based method and state-of-the-art large language models. The fuzzy rank-based ensemble method, which combines multiple language models, achieved an F1-score of 96.03%.

**Conclusions:** The fusion of ensemble learning with PLMs and the Gompertz function, employing fuzzy rank-based methodology, introduces a novel prediction approach with prospects for enhancing accuracy and reliability. Additionally, experimental results imply that training solely on textual content can yield high prediction accuracy.

**Keywords:**

Misinformation; COVID-19; Ensemble models; Fuzzy ranks

## Introduction

### Background

From 2019 to 2022, the global community faced the challenges posed by the coronavirus disease 2019 (COVID-19) pandemic. In response, governments worldwide and the World Health Organization (WHO) collaborated extensively to reduce the virus's spread. An increased demand for trustworthy information sources and accurate health guidance arose during this global health crisis. However, the surge in these information needs overlapped with the rapid spread of misinformation and false news through social media platforms, leading to widespread public confusion.

The WHO used the term "infodemic" to describe the spread of misinformation during the pandemic 1. They emphasized the potential threat that such misinformation posed to national epidemic prevention policies. Trust in incorrect or misleading information could result in adverse health behaviors and non-compliance with health policies, worsening the pandemic's challenges.

While the distribution of COVID-19 vaccines contributed to the gradual control of the pandemic, the virus persisted, giving rise to post-infection symptoms known as long COVID, confirmed in at least 10% of those who contracted the virus2. Additionally, instances of reinfection after initial recovery were observed, with research from the U.S. Department of Veterans Affairs indicating increased risks of mortality, hospitalization, and post-symptomatic conditions for reinfected patients3.

Despite the diminishing immediate threat of COVID-19, the ongoing risks associated with long COVID and reinfection make public attention necessary to COVID-19-related policies and information. The challenges of fake news and misinformation persist as the world transitions into a post-pandemic era coexisting with the virus. Specifically, issues related to long COVID and reinfection continue to be crucial points for misinformation. Therefore, the timely and accurate identification and classification of such misinformation becomes critical.

### Prior Work

Throughout the COVID-19 pandemic, some research studies have used machine learning and deep learning techniques to address the challenge of detecting fake news and misinformation.

Patwa et al.4 collected COVID-19-related texts from publicly available fact-checking websites and social media platforms. Their approach involved TF-IDF for feature extraction and applying various machine learning algorithms, such as logistic regression, support vector machines, decision trees, and gradient boosting, for binary classification of fake news. Das et al.5 employed pre-trained language models (PLMs), including RoBERTa and XLNet, for preprocessing and training on the same dataset. By combining predictions from multiple models through voting, they achieved admirable results in the CONSTRAINT2021 COVID-19 Fake News Detection competition.

Paka et al.6 claimed that more than relying on textual features might be required for accurate fake news classification. To address this, they gathered COVID-19-related tweets from Twitter, incorporating additional data such as the number of likes for a tweet, URL links, and the follower count of the tweeters. Introducing a multi-feature classification approach for fake news using a cross-stitch unit combined with an LSTM architecture enhanced the classification accuracy.

Furthermore, research teams focusing on the Chinese language employed deep learning frameworks like RNN, CNN, and Transformer to classify COVID-19 fake news in Chinese text7. These endeavors highlight the global commitment to addressing the infodemic and giving accurate and reliable information.

Additionally, the emergence of LLM based on transformers, such as ChatGPT8, has become prominent in recent years. These models, capable of understanding natural language and interacting with users, hold the potential to contribute to the development of more robust tools for combating the spread of false news across various domains.

### Goal of This Study

Considering advancements in deep learning technologies and natural language processing (NLP), this study investigates the performance of various deep learning models in detecting fake news. The objective is to provide a scientific and efficient method for fake news detection in the post-pandemic era. Texts about long COVID and reinfection were collected from open-source databases and through web crawling, followed by a preprocessing phase to clean and refine the data. Afterward, various machine and deep learning models were trained and evaluated based on their performance after preprocessing. Finally, a fuzzy rank-based ensemble approach combined multiple models. The performance of this ensemble method was then compared with the state-of-the-art large language models (LLMs) methods.

The proposed method achieved an F1-score of 96.03%, which can significantly help people to classify misinformation in real-time. The results also demonstrate the effectiveness of language models in distinguishing misinformation.

## Methods

The method comprises four key stages: data collection, data preprocessing, data analysis, and modeling. Initially, information was collected from diverse publicly available online sources. Given the inherent inconsistency of online sources, the gathered data underwent a preprocessing phase to enhance its cleanliness. After the preprocessing, a foundational analysis was undertaken to understand the dataset's characteristics better. In the end, various deep learning models were trained and compared with other text classification methods. We introduce a fuzzy ranking method with the Gompertz function, which adjusts weights based on the confidence scores of each classifier to generate final predictions for each sample. This fusion of ensemble learning and the Gompertz function offers a fresh perspective on prediction methodologies.

### Data Collection

Articles and claims related to COVID-19 were collected from diverse online sources. The gathered materials underwent a filter phase using keywords associated with long COVID and reinfection, such as chronic, long-term, persistent, after-effects, sequelae, complications, recovery, post covid, post-covid, omicron, subvariant, reinfection, immune, and variant. The resulting dataset, categorized as either "genuine" or "fake," originated from three primary sources:

#### Open Source Dataset

**Fighting an Infodemic**4: This dataset includes COVID-19-related topics from platforms like Twitter, Facebook, and fact-checking websites. Utilized for the Constraint@AAAI2021 - COVID19 Fake News Detection in English competition, only labeled data from this dataset were used and are available on GitHub9.

**CTF (COVID-19 Twitter Fake News)**6: Focused on tweets from Twitter, this dataset includes labeled and unlabeled data concerning genuine and fake COVID-19 news. For this study, only the labeled text content data was utilized.

**CoAID (Covid-19 heAlthcare mIsinformation Dataset)**10: A diverse COVID-19 fake news dataset containing news from the internet and social media platforms, user engagements, tweets, and labels appearing on Twitter.

**FibVID (Fake news information-broadcasting dataset of COVID-19)**11: Collects claims from fact-checking websites like Snopes and Politifact, along with related discourse from Twitter. This dataset includes COVID and non-COVID topics, divided into four labels. For this study, only data related to COVID-19 from categories 0 and 1 were used.

**FaCOV**12: Collected from 13 English fact-checking websites related to COVID-19, this dataset includes article titles, URLs, claims, and abstracts. Data with two category labels were used by merging titles and article contents.

#### FactCheck Website

While open-source databases offer significant support, they often have limitations regarding data timeframe. To conquer the restrictions, web scraping and data cleaning techniques were employed to gather more recent data from verified fact-checking websites Snopes13 and PolitiFact14, which are certified by the International Fact-Checking Network (IFCN). Web crawling method was employed to systematically extract articles classified under "CORONAVIRUS" and "COVID-19" from Snopes (data up to August 31, 2023) and PolitiFact (data up to July 31, 2023).

Alongside the article contents, labels were collected for model training. 1500 and 806 texts were extracted from Snopes and PolitiFact, respectively. Subsequently, the collected data underwent keyword filtering to align more closely with the topic.

In PolitiFact, articles are categorized into six labels: pants-on-fire, false, barely-true, half-true, mostly-true, and true. In contrast, Snopes classifies articles into fourteen labels: true, mostly-true, mixture, mostly-false, false, unproven, outdated, miscaptioned, correct-attribution, misattributed, scam, legend, labeled-satire, and lost-legend. Based on the research by Khan et al.15, the labels from diverse sources were reclassified into two categories: genuine and fake.

#### Governmental Bodies

Government and public institution websites, such as the World Health Organization (WHO) and the Centers for Disease Control and Prevention (CDC), were also regarded as primary sources. These institutions have consistently distributed up-to-date information and guidelines throughout the pandemic, establishing them as widely acknowledged reliable and accurate data sources. Articles related to "long COVID" and "Reinfection" were collected from the COVID-19 sections of these websites. Because the original content might be lengthy, ChatGPT 8 was used to refine and reorganize the content as short claims. The structured claims ensured appropriate length and clarity.

Consequently, each claim was labeled as "genuine" owing to its reputable source. The dataset used for model training encapsulated the latest information attained through these procedural steps. [Table 1](#Table1) presents the filtered sample counts from various data sources.

Table 1. Sample size from different sources.

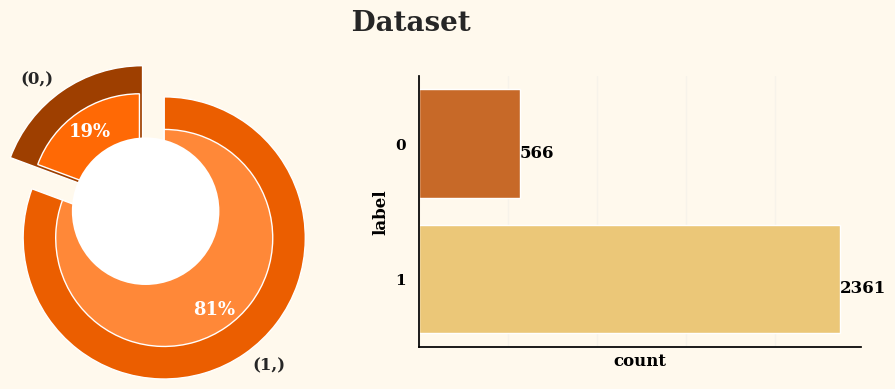
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Source | Time until | Sample size | Fake label | Genuine label |
| CTF | ~2021 | 1292 | 1130 | 162 |
| Fighting an Infodemic | ~2021 | 218 | 62 | 156 |
| CoAID | ~2020 | 70 | 0 | 70 |
| FibVID | ~2020 | 615 | 318 | 297 |
| FaCOV | ~2021 | 811 | 811 | 0 |
| PolitiFact | ~2023 | 87 | 42 | 45 |
| Snopes | ~2023 | 15 | 9 | 6 |
| CDC+WHO | ~2023 | 58 | 0 | 58 |
| Total |  | 3166 | 2372 | 794 |

### Data preprocessing

Ensuring the absence of duplicate entries in the merged open-source datasets is crucial. To achieve this, cross-referencing the data with existing open-source datasets was undertaken. Any identified duplicate samples were eliminated to reduce redundancy, thereby preventing potential impacts on the following model training and analysis performance.

During the preprocessing phase, social media posts and articles, known to include emojis and external links (URLs) frequently, underwent processing to enhance their suitability for distinguishing between genuine and fake news. The Tweet-preprocessor package was employed to eliminate emojis and URLs from the texts because of their low distribution in classification.

All the labels were encoded as "0" (stands for genuine) and "1" (stands for fake), respectively. The distribution of our dataset is illustrated in [Figure 1](#Figure1). As the public data sources exhibited a bias towards the "fake" category, an imbalanced label distribution was observed in the dataset. Therefore, we employed stratified sampling, allocating 10% of the data for testing and 90% for training. This sampling approach ensures consistent label proportions in both the test and training sets, thereby preventing the potential insufficient of "genuine" samples that may arise from random sampling.

Figure 1. Data distribution after preprocessing. 0 stands for genuine, and 1 stands for fake.

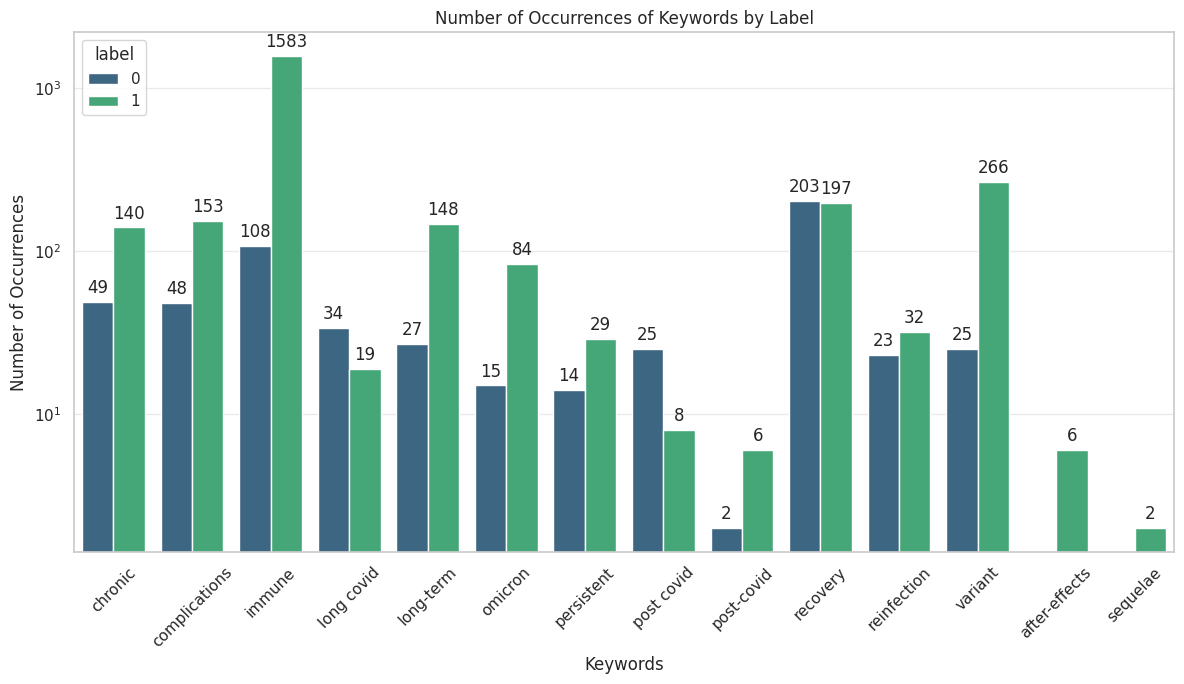
### Data analysis

The data analysis included keyword occurrences, sentiment analysis, and subjectivity values. The analysis aimed to get deeper insights from the collected dataset and identified different distributions between genuine and fake articles.

#### Keywords occurrences

Through the analysis of the frequency of certain keywords in the data, we can gain a better understanding of the public's interests. In [Figure 2](#Figure2), we can see that over 50% of the fake samples contain the term "immune," while only a few genuine samples feature it. The term "recovery" is primarily used in genuine samples but is also present in a similar proportion of fake samples. Other keywords such as "variant," "complication," and "chronic" also appear more frequently in fake samples. This suggests that such terms are often used to spread fake news.

Figure 2. Number of keywords occurrences by label.



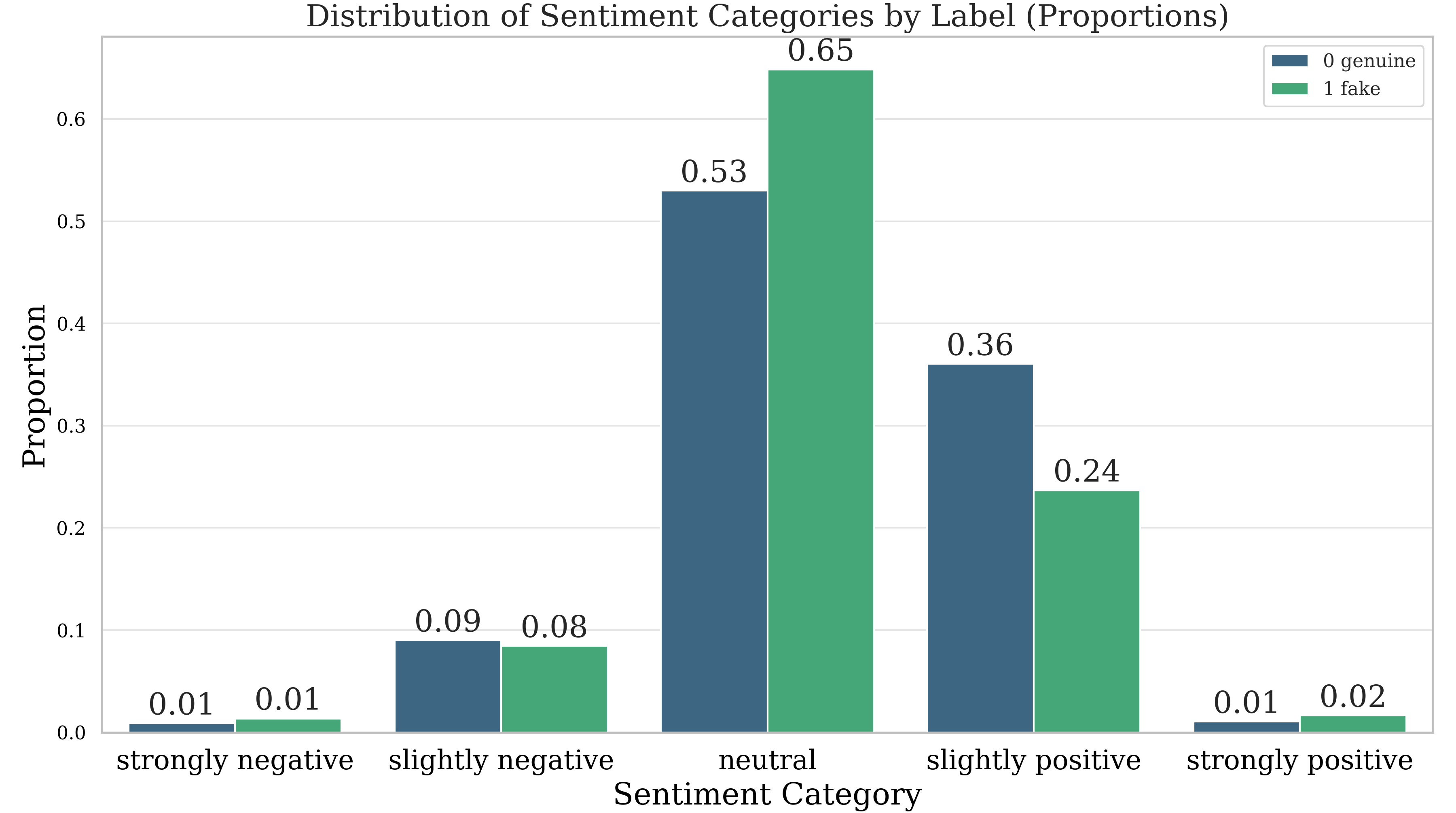
#### Sentiment analysis

Sentiment analysis can explain an article's tone, whether positive, negative, or neutral. Additionally, it can estimate the subjectivity of the text, distinguishing between genuine information and the author's opinions. To gain deeper insights into the textual data sourced from open-source databases and fact-checking websites, the TextBlob package16 was used for sentiment analysis. TextBlob's sentiment analysis assigns a polarity value ranging from -1 to 1, indicating the sentiment from entirely negative to entirely positive. To classify sentiments based on polarity, we categorized them into the following five groups:

* strongly negative: polarity values between -1 and -0.5.
* slightly negative: polarity values between -0.5 and -0.1.
* neutral: polarity values between -0.1 and 0.1.
* slightly positive: polarity values between 0.1 and 0.5.
* strongly positive: polarity values between 0.5 and 1.

The sentiment analysis reveals the distribution of sentiments within the collected textual data. In [Figure 3](#Figure3), most of the content falls into the "neutral" category. Approximately 65% of the content in fake texts and 53% in genuine texts are classified as "neutral." Genuine texts show a higher proportion (36%) in the "slightly positive" category compared to fake texts (24%), indicating a tendency for genuine content to include more positive language. Conversely, there are no significant disparities between the distributions of the "strongly negative," "slightly negative," and "strongly positive" sentiment categories across the two labels.

Figure 3. Data distribution (percentage) of different sentiment polarity groups.

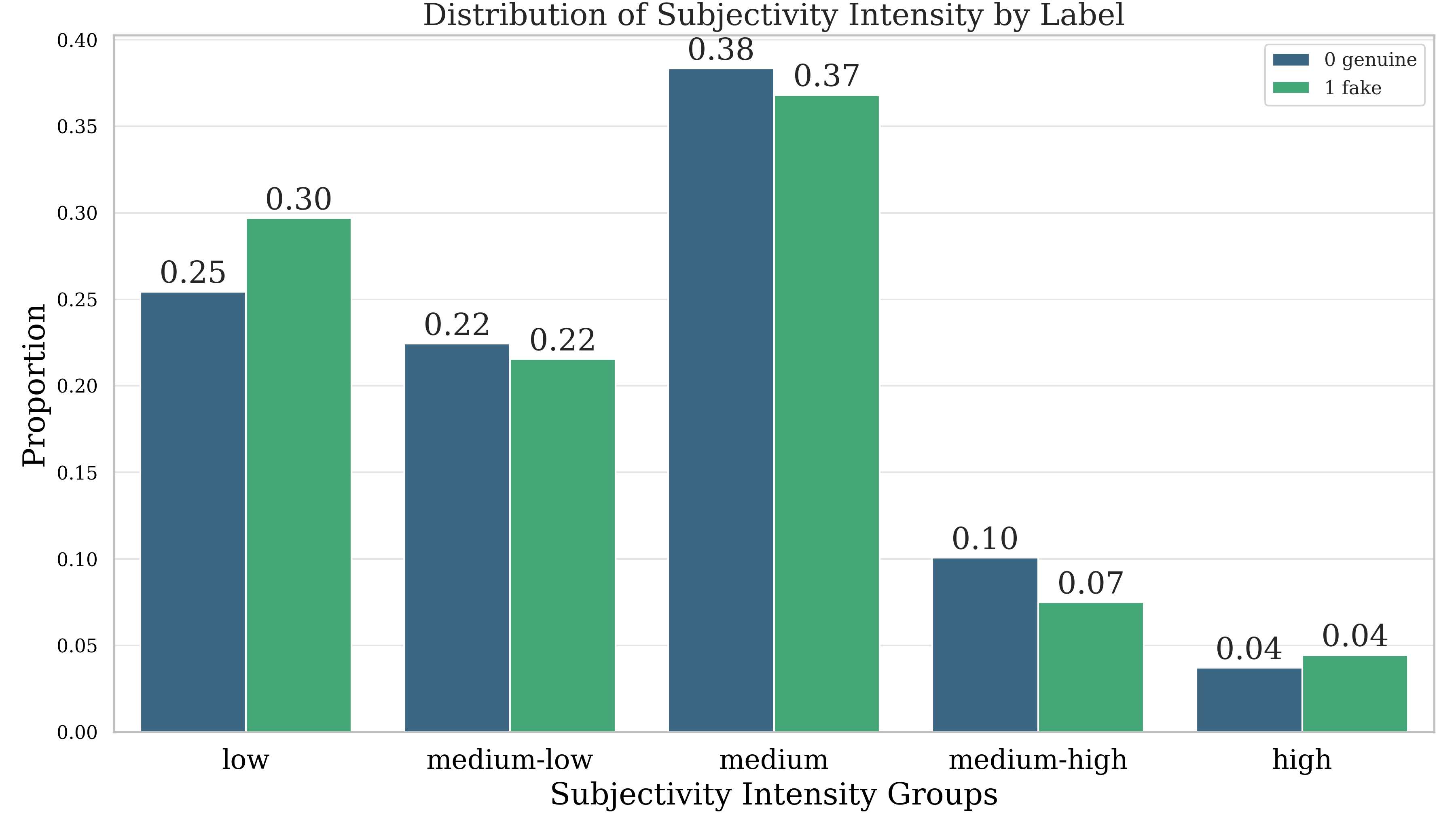


In addition to sentiment analysis, TextBlob also provides a subjectivity value ranging from 0 to 1, indicating the degree of Subjectivity within the text, ranging from entirely objective to entirely subjective. To simplify comprehension, we categorized the degree of Subjectivity into the following five groups:

* low subjectivity: values between 0 and 0.2.
* medium-low subjectivity: values between 0.2 and 0.4.
* medium subjectivity: values between 0.4 and 0.6.
* medium-high subjectivity: values between 0.6 and 0.8.
* high subjectivity: values between 0.8 and 1.

Based on the analysis illustrated in [Figure 4](#Figure4), the distribution of subjectivity levels between genuine and fake texts appears quite similar, mainly concentrated in the "medium subjectivity" category. Genuine texts exhibit 38% of their content in this category, while fake texts contain 37%. There is a 5% higher prevalence of fake texts in the "low subjectivity" category than genuine texts. Conversely, in the "medium-high subjectivity" category, genuine texts surpass fake texts by a margin of 3%.

Figure 4. Data distribution (percentage) of different subjectivity groups.



While minor differences are observed in the distribution of sentiment and subjectivity between genuine and fake texts, these differences may not serve as definitive classification criteria. Moreover, sentiment analysis encounters challenges in natural language comprehension, such as accurately identifying sarcasm. Therefore, more precise approaches, such as machine learning algorithms and deep learning models, are required to differentiate genuine from fake news concerning long COVID and reinfections.

### Models

In this study, a thorough comparison was executed using various classification methods. Traditional machine learning algorithms that used text content features were employed to establish a baseline. Deep learning models, from the HAN to BERT series, were also used to take advantage of the advanced abilities to handle complex textual data information. Also, embedding models based on LLMs were used to compare the performance between fee-required models and open-source deep models. The experiment approach helps evaluate different methods in distinguishing articles with genuine and fake information, especially regarding long-term COVID-19 and reinfections.

#### SVM

To establish a baseline model, this study initially selected SVM (Support Vector Machine)17. In text classification, linear classifiers are commonly considered strong baseline models. By comparing the performance of the linear classifier with that of deep learning models, we can verify their effectiveness when fine-tuned and employed18. To apply SVM for classification tasks, uni-gram TF-IDF features were generated from the training set data, and the SVM model was trained using these features for binary classification.

#### HAN

HAN (Hierarchical Attention Networks)19 integrates attention mechanisms at multiple levels, focusing on word and sentence levels, to capture diverse hierarchical structures within documents. The authors used recurrent neural networks combined with word-attention and sentence-attention layers for text classification, resulting in state-of-the-art performance across six datasets. Following the training of this model on the training set, it will be compared with the method we proposed.

#### PLMs

In addition to deep learning models like HAN, this study also fine-tuned PLMs to use their text-understanding capabilities for fake news detection. PLMs, including BERT20, RoBERTa21, DeBERTa22, and XLNet23, are state-of-the-art models widely used in natural language processing tasks. BERT, introduced by Google, uses masked language modeling and next-sentence prediction to generate contextualized word representations. RoBERTa, an enhancement of BERT by Facebook, removes next-sentence prediction and incorporates optimization strategies for improved performance. DeBERTa introduces disentangled attention and enhanced mask decoder mechanisms to refine self-attention further, achieving superior results across NLP tasks. XLNet, developed by Google, employs permutation language modeling and Transformer-XL architecture to effectively understand bidirectional contextual comprehension, especially in processing long texts, surpassing BERT and RoBERTa in various benchmarks.

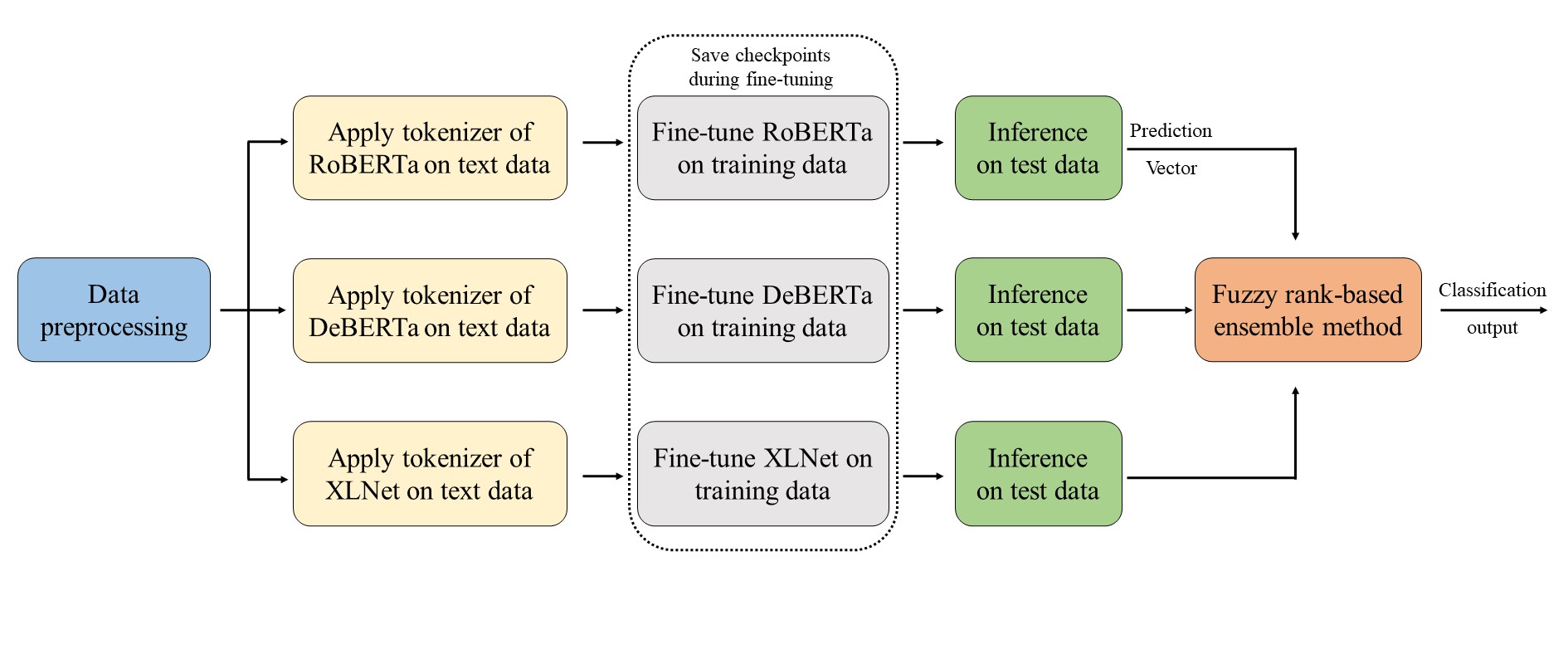
#### LLMs

The success of LLMs in recent years has made significant contributions to applications in the field of NLP. For the classification task in this study, we use OpenAI's GPT embedding model "text-embedding-ada-002"24,25 and Google's Gemini embedding model26 to transform the training set data, combining it with machine learning methods such as SVM for training. Integrating knowledge from large language models and the training dataset helps predictions of text classification. Furthermore, we directly apply OpenAI's GPT-427, the state-of-the-art LLM, to infer the texts in the test set. This allows us to compare the performance of LLMs with our proposed method.

### *Fuzzy rank-based ensemble technique*

Ensemble learning combines individual models' strengths to yield predictions that outperform any contributing model. This research proposes an approach to enhance prediction performance by incorporating the Gompertz function into ensemble learning techniques.

We employed a fuzzy rank-based ensemble technique28, where the confidence of each classifier in its predictions was given priority for each test case. This differs from traditional ensemble methods like the average or weighted average rules, which assign pre-defined fixed weights to classifiers. Moreover, the re-parameterized Gompertz function was used to compute the fuzzy ranks of each pre-trained model for detection. Incorporating state-of-the-art PLMs like RoBERTa, DeBERTa, and XLNet further improved our approach. These PLMs bring advanced language understanding abilities to the ensemble, contributing to its robust performance. Afterward, the predictions of three models were fused. [Figure 5](#Figure5) provides a process of the proposed method.

Figure 5. Diagram of the fuzzy ensemble process using multiple language models.

### Implementation

The data was randomly split using the Scikit-learn29 package. This study employed 5-fold cross-validation to train the SVM and determine the best-performing model for final evaluation. Fine-tuning of PLMs was conducted using checkpoints provided by Hugging Face, using the AdamW optimizer30 and cross-entropy loss. The learning rate was set to 2e-5, and training was carried out for 20 epochs. HAN was trained for the same number of epochs with its default settings. The training procedure was repeated five times with different random initial weights, and checkpoints of models were selected for final evaluation based on the highest validation F1-score31. HAN was trained on the Tesla T4, while all other deep models were trained on the RTX A5000.

## Results

Model performances were evaluated using well-known metrics: accuracy, precision, recall, F1-score, and AUC (Area Under the ROC Curve). The results presented in [Table 2](#Table2) compare the proposed fuzzy method with other approaches, including traditional machine learning algorithms, deep learning networks, pre-trained networks, and state-of-the-art LLMs.

TF-IDF with SVM achieved an accuracy of 89.08%, precision of 91.46%, recall of 95.34%, F1-score of 93.36%, and AUC of 92.02%. While these metrics show acceptable performance, they are still beaten by attention-based deep models like the HAN and BERT series in the experiments.

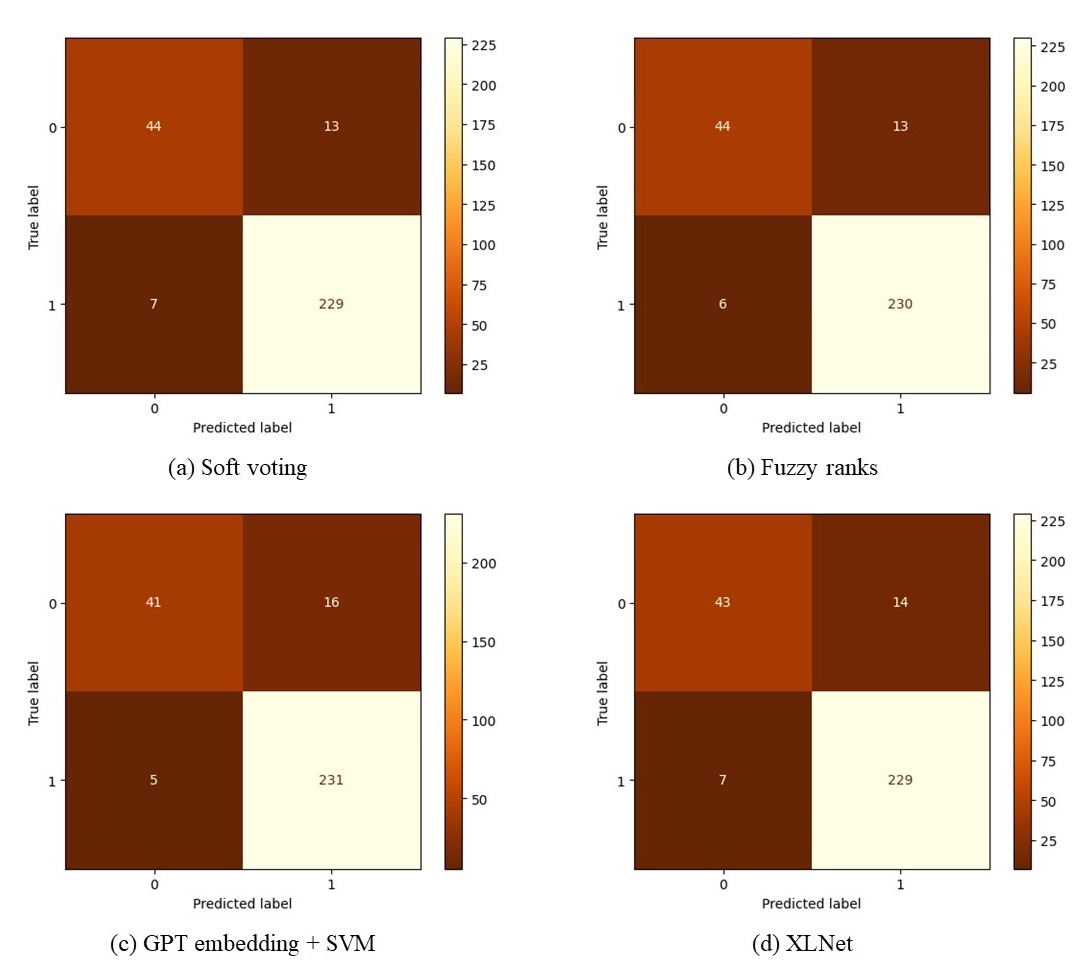
Comparing HAN with TF-IDF and BERT, we can observe that HAN outperforms TF-IDF in terms of accuracy (90.78% vs. 89.08%), F1-score(94.29% vs. 93.36%), and AUC (93.09% vs. 92.02%), indicating its superior performance in the fake news detection task. However, BERT achieved slightly higher accuracy and AUC compared to HAN.

Table 2. Comparison of the performance of different models for test data using various evaluation metrics includes accuracy, precision, recall, F1-score, and AUC.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1-score | AUC |
| TF-IDF + SVM | 89.08% | 91.46% | 95.34% | 93.36% | 92.02% |
| HAN | 90.78% | 94.09% | 94.49% | 94.29% | 93.09% |
| BERT | 91.13% | 93.75% | 95.34% | 94.54% | 96.31% |
| RoBERTa | 91.81% | 93.44% | 96.61% | 95.00% | 96.59% |
| DeBERTa | 91.81% | 94.17% | 95.76% | 94.96% | 95.75% |
| XLNet | 92.83% | 94.24% | 97.03% | 95.62% | 96.12% |
| GPT embedding + SVM | 92.83% | 93.52% | 97.88% | 95.65% | 94.88% |
| Gemini embedding + SVM | 91.47% | 92.71% | 97.03% | 94.82% | 93.27% |
| GPT-4 | 82.25% | 91.82% | 85.59% | 88.60% | N/A |
| Soft voting | 93.17% | 94.63% | 97.03% | 95.82% | 97.14% |
| Fuzzy rank-based method | 93.52% | 94.65% | 97.46% | 96.03% | 97.15% |

aBoth fuzzy and soft voting methods combine the RoBERTa, DeBERTa, and XLNet models.

The fuzzy method achieved impressive results on the test set, with an accuracy of 93.52%, precision of 94.65%, F1-score of 96.03%, and AUC of 97.15%, presenting the highest performance among all methods estimated. The conventional soft voting ensemble method on three PLMs also yielded high AUC of 97.14% as well. While combining the GPT embedding model with SVM achieved the highest recall of 97.88%. The experiment demonstrates the effectiveness of the proposed approach across multiple evaluation metrics, showcasing its potential for robust classification tasks. [Figure 6](#Figure6) shows the confusion matrix with the top 4 F1-score in the experiment.

Figure 6. Confusion matrix of (a) soft voting, (b) fuzzy rank-based method, (c) GPT embedding method and (d)XLNet on the held-out test dataset.

[Table 3](#Table3) is the actual case of using the fuzzy method to detect fake and genuine news that was not included in the test set. We used four isolated samples, two genuine and two fake, varying in short and long length. The result shows that our method accurately detected the genuineness of the content regardless of its length, revealing its robustness across different text lengths.

Table 3. Real case inference using fuzzy rank ensemble method.

|  |  |  |  |
| --- | --- | --- | --- |
| Content / Mainly claim | Length | Prediction | Ground truth |
| A German study has revealed long COVID is linked to the vaccine. | 15 | Fake | Fake |
| Covid vaccination before infection strongly linked to reduced risk of developing long covid. | 15 | Genuine | Genuine |
| Long COVID's causes and risk factors remain a subject of ongoing research, with potential factors including reactivation of SARS-CoV-2 particles, overactive immune responses, and the development of autoantibodies attacking organs. Certain groups, such as those with severe COVID-19 history, underlying health conditions, or lacking vaccination, are at higher risk, alongside other factors like sex, age, initial immune response, and viral variants. Health inequities may also contribute, especially affecting racial or ethnic minority groups and individuals with disabilities. | 367 | Genuine | Genuine |
| While Omicron's subvariants find new ways to evade vaccines and destabilize immune systems, another pandemic has overwhelmed officials who are supposed to be in charge of public health. In any case, COVID, a novel virus that can wreak havoc with vital organs in the body, continues to evolve at a furious pace. In response officials have largely abandoned any coherent response, including masking, testing, tracing and even basic data collection. Yes, the people have been abandoned. So don't expect "normal" to return to your hospital, your airport, your nation, your community or your life anytime soon. | 469 | Fake | Fake |

## Discussion

### Principal Findings

The experimental results demonstrate that using language models, particularly XLNet, for fake news detection outperforms traditional TF-IDF features combined with SVM or deep models like HAN. Some insights can be observed in [Table 4](#Table4), showing a trend where models with more parameters generally deliver better classification accuracy. Models like BERT, RoBERTa, DeBERTa, and XLNet, which possess significantly larger parameters than HAN, demonstrated superior performance across various evaluation metrics, including accuracy, precision, recall, F1-score, and AUC.Despite XLNet having fewer parameters than other models in the BERT series, it still outperformed them in the experiment. This result suggests that classification effectiveness depends not only on the number of parameters but also on the model architecture, training methods, and optimization techniques. XLNet's success can be attributed to its permutation language modeling approach, which enhances bidirectional context understanding while reducing some limitations in BERT.

Table 4. Parameter count of each deep model used in the experiment.

|  |  |
| --- | --- |
| Model | Count of Parameters |
| HAN | 2,343,202 |
| BERT | 109,483,778 |
| RoBERTa | 124,647,170 |
| DeBERTa | 139,193,858 |
| XLNet | 117,310,466 |
| GPT embedding | unknown |
| Gemini-embedding | unknown |
| GPT-4 | unknown |

Additionally, using LLMs' embedding models exhibits superior performance compared to traditional TF-IDF features. In this study, the GPT embedding model performs slightly better than Gemini, possibly due to differences in the length of the embedding vectors. GPT defaults to 1536, while Gemini defaults to 768.  Despite yielding acceptable outcomes, directly using GPT-4 falls short compared to training SVM on vector-transformed training data using embedding models. This result implies that LLMs still benefit from training data in fake news detection tasks. Although the LLMs’ embedding models showed remarkable performance in the experiment, accessing this kind of embedding model via API needs charges. However, the fuzzy method can be combined with open-source PLMs to achieve even better results. Moreover, compared to using a single language model or soft voting method with pre-defined weights, the fuzzy fusion-based technique allows for determining ensemble model weights for each test case, resulting in superior performance.

### Limitations

One limitation of the study is the presence of data imbalance. The imbalance in the data suggests a potential bias towards the prevalence of fake information on the internet. Addressing this issue would require gathering a more extensive and up-to-date dataset of genuine information to achieve better balance and representativeness in the training data.

### Conclusions

This study used open-source databases and reputable websites to collect textual data concerning long-term COVID and reinfections. Through information engineering techniques, we gained insights into the characteristics of articles and claims about these topics. AI models, including transformer-based models and linear classifiers, have shown strength in detecting misleading or inaccurate information, thereby serving as valuable tools for the public to discern between genuine and fake information. Ensemble models have demonstrated robust performance by minimizing prediction errors' variance and reducing dispersion. The fuzzy rank-based fusion of ensemble learning with PLMs and the Gompertz function presents a novel approach to prediction methodologies, offering potential improvements in accuracy and reliability. Moreover, experimental results indicate that training solely on textual content can achieve high prediction accuracy.

### Conflicts of Interest

None declared.

### Abbreviations

COVID-19: coronavirus disease 2019

WHO: World Health Organization

PLMs: pre-trained language models

NLP: natural language processing

LLMs: large language models

CTF: COVID-19 Twitter Fake News

CoAID: Covid-19 heAlthcare mIsinformation Dataset

FibVID: Fake news information-broadcasting dataset of COVID-19

IFCN: International Fact-Checking Network

CDC: Centers for Disease Control and Prevention

SVM: Support Vector Machine

HAN: Hierarchical Attention Networks

BERT: Bidirectional Encoder Representations from Transformers

RoBERTa: Robustly Optimized BERT Pretraining Approach

DeBERTa: Decoding-enhanced BERT with disentangled attention

AUC: Area Under the ROC Curve

## References

1. Infodemic. Accessed September 28, 2023. https://www.who.int/health-topics/infodemic

2. Davis HE, McCorkell L, Vogel JM, Topol EJ. Long COVID: major findings, mechanisms and recommendations. *Nat Rev Microbiol*. 2023;21(3):133-146. doi:10.1038/s41579-022-00846-2

3. Bowe B, Xie Y, Al-Aly Z. Acute and postacute sequelae associated with SARS-CoV-2 reinfection. *Nat Med*. 2022;28(11):2398-2405. doi:10.1038/s41591-022-02051-3

4. Patwa P, Sharma S, Pykl S, et al. Fighting an Infodemic: COVID-19 Fake News Dataset. In: Vol 1402. ; 2021:21-29. doi:10.1007/978-3-030-73696-5\_3

5. Das SD, Basak A, Dutta S. A Heuristic-driven Ensemble Framework for COVID-19 Fake News Detection. Published online January 10, 2021. doi:10.48550/arXiv.2101.03545

6. Paka WS, Bansal R, Kaushik A, Sengupta S, Chakraborty T. Cross-SEAN: A cross-stitch semi-supervised neural attention model for COVID-19 fake news detection. *Appl Soft Comput*. 2021;107:107393. doi:10.1016/j.asoc.2021.107393

7. Yang C, Zhou X, Zafarani R. CHECKED: Chinese COVID-19 fake news dataset. *Soc Netw Anal Min*. 2021;11(1):58. doi:10.1007/s13278-021-00766-8

8. ChatGPT. Accessed October 19, 2023. https://chat.openai.com

9. covid\_fake\_news/data at main · diptamath/covid\_fake\_news. GitHub. Accessed October 18, 2023. https://github.com/diptamath/covid\_fake\_news/tree/main/data

10. Cui L, Lee D. CoAID: COVID-19 Healthcare Misinformation Dataset. Published online November 3, 2020. doi:10.48550/arXiv.2006.00885

11. Kim J, Aum J, Lee S, Jang Y, Park E, Choi D. FibVID: Comprehensive fake news diffusion dataset during the COVID-19 period. *Telemat Inform*. 2021;64:101688. doi:10.1016/j.tele.2021.101688

12. Sharma S, Agrawal E, Sharma R, Datta A. FaCov: COVID-19 Viral News and Rumors Fact-Check Articles Dataset. *Proc Int AAAI Conf Web Soc Media*. 2022;16:1312-1321. doi:10.1609/icwsm.v16i1.19383

13. COVID Archives | Snopes.com. Accessed October 19, 2023. https://www.snopes.com/tag/covid-19/

14. Fact-checks | PolitiFact. Accessed October 19, 2023. https://www.politifact.com/factchecks/list/?category=coronavirus

15. Khan JY, Khondaker MdTI, Afroz S, Uddin G, Iqbal A. A benchmark study of machine learning models for online fake news detection. *Mach Learn Appl*. 2021;4:100032. doi:10.1016/j.mlwa.2021.100032

16. TextBlob: Simplified Text Processing — TextBlob 0.16.0 documentation. Accessed October 12, 2023. https://textblob.readthedocs.io/en/dev/

17. Cortes C, Vapnik V. Support-vector networks. *Mach Learn*. 1995;20(3):273-297. doi:10.1007/BF00994018

18. Lin YC, Chen SA, Liu JJ, Lin CJ. Linear Classifier: An Often-Forgotten Baseline for Text Classification. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics; 2023:1876-1888. doi:10.18653/v1/2023.acl-short.160

19. Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E. Hierarchical Attention Networks for Document Classification. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics; 2016:1480-1489. doi:10.18653/v1/N16-1174

20. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Published online May 24, 2019. doi:10.48550/arXiv.1810.04805

21. Liu Y, Ott M, Goyal N, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. Published online July 26, 2019. doi:10.48550/arXiv.1907.11692

22. He P, Liu X, Gao J, Chen W. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. Published online October 6, 2021. doi:10.48550/arXiv.2006.03654

23. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov RR, Le QV. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In: *Advances in Neural Information Processing Systems*. Vol 32. Curran Associates, Inc.; 2019. Accessed July 12, 2023. https://papers.nips.cc/paper\_files/paper/2019/hash/dc6a7e655d7e5840e66733e9ee67cc69-Abstract.html

24. OpenAI Platform. Accessed December 4, 2023. https://platform.openai.com

25. New and improved embedding model. Accessed March 25, 2024. https://openai.com/blog/new-and-improved-embedding-model

26. Gemini Team, Anil R, Borgeaud S, et al. Gemini: A Family of Highly Capable Multimodal Models. Published online December 18, 2023. doi:10.48550/arXiv.2312.11805

27. OpenAI. GPT-4 Technical Report. Published online March 27, 2023. Accessed December 4, 2023. http://arxiv.org/abs/2303.08774

28. Kundu R, Basak H, Singh PK, Ahmadian A, Ferrara M, Sarkar R. Fuzzy rank-based fusion of CNN models using Gompertz function for screening COVID-19 CT-scans. *Sci Rep*. 2021;11(1):14133. doi:10.1038/s41598-021-93658-y

29. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res*. 2011;12(85):2825-2830.

30. Loshchilov I, Hutter F. Decoupled Weight Decay Regularization. Published online January 4, 2019. doi:10.48550/arXiv.1711.05101

31. Chalkidis I, Jana A, Hartung D, et al. LexGLUE: A Benchmark Dataset for Legal Language Understanding in English. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics; 2022:4310-4330. doi:10.18653/v1/2022.acl-long.297