computation

MDPI

Article

Unveiling AI-Generated Financial Text: A Computational Approach Using Natural Language Processing and Generative Artificial Intelligence

Muhammad Asad Arshed [1,*], Ștefan Cristian Gherghina [2,*], Christine Dewi [3,4], Asma Iqbal [1] and Shahzad Mumtaz [5,6]

[1] Department of Software Engineering, University of Management and Technology, Lahore 54770, Pakistan; asma_iqbal0900@outlook.com
[2] Department of Finance, Bucharest University of Economic Studies, 6 Piata Romana, 010374 Bucharest, Romania
[3] Department of Information Technology, Satya Wacana Christian University, Salatiga 50715, Indonesia; christine.dewi@uksw.edu
[4] School of Information Technology, Deakin University, Campus 221 Burwood Hwy, Burwood, VIC 3125, Australia
[5] Department of Data Science, The Islamia University of Bahawalpur, Bahawalpur 63100, Pakistan; shahzadmumtaz22@gmail.com
[6] School of Natural and Computing Sciences, University of Aberdeen, Aberdeen AB24 3FX, Scotland, UK
* Correspondence: asad.arshed@umt.edu.pk (M.A.A.); stefan.gherghina@fin.ase.ro (Ș.C.G.)

Abstract: This study is an in-depth exploration of the nascent field of Natural Language Processing (NLP) and generative Artificial Intelligence (AI), and it concentrates on the vital task of distinguishing between human-generated text and content that has been produced by AI models. Particularly, this research pioneers the identification of financial text derived from AI models such as ChatGPT and paraphrasing tools like QuillBot. While our primary focus is on financial content, we have also pinpointed texts generated by paragraph rewriting tools and utilized ChatGPT for various contexts this multiclass identification was missing in previous studies. In this paper, we use a comprehensive feature extraction methodology that combines TF–IDF with Word2Vec, along with individual feature extraction methods. Importantly, combining a Random Forest model with Word2Vec results in impressive outcomes. Moreover, this study investigates the significance of the window size parameters in the Word2Vec approach, revealing that a window size of one produces outstanding scores across various metrics, including accuracy, precision, recall and the F1 measure, all reaching a notable value of 0.74. In addition to this, our developed model performs well in classification, attaining AUC values of 0.94 for the 'GPT' class; 0.77 for the 'Quil' class; and 0.89 for the 'Real' class. We also achieved an accuracy of 0.72, precision of 0.71, recall of 0.72, and F1 of 0.71 for our extended prepared dataset. This study contributes significantly to the evolving landscape of AI text identification, providing valuable insights and promising directions for future research.

Keywords: ChatGPT; QuillBot; generative artificial intelligence; natural language processing; text identification; machine learning

Citation: Arshed, M.A.; Gherghina, Ș.C.; Dewi, C.; Iqbal, A.; Mumtaz, S. Unveiling AI-Generated Financial Text: A Computational Approach Using Natural Language Processing and Generative Artificial Intelligence. *Computation* **2024**, *12*, 101. https://doi.org/10.3390/computation12050101

Academic Editor: Demos T. Tsahalis

Received: 21 March 2024
Revised: 6 May 2024
Accepted: 13 May 2024
Published: 15 May 2024

1. Introduction

Nowadays, Natural Language Processing (NLP) has arisen as a central subfield of AI with keen execution of human–computer interactions and information processing [1,2]. The continuous action to create a machine and then to figure out and correspond in mortal vocabulary is an extraordinary improvement. It is transforming how we commit to technology [3].

In previous years, the domain of NLP has made many innovations in cultured models that have seen exceptional improvements [2,4]. These models use detailed procedures of the

Machine Learning (ML) algorithms that can learn from the huge amount of data, following which they can be translated into the human language without any problem. Moreover, they are truly extraordinary at memorizing from a massive quantity of text [5]. Such models use a massive quantity of data to learn how people address and figure out specialties like grammar and significance. These language models are frequently instructed on enormous datasets, including text from the internet, books, papers, and additional resources to grasp the nominal facts of the language too. They use the obtained knowledge to develop readable and dependent accurate text in response to different infusions and assignments. The improvement in this domain is due to the availability of big data, strong computing resources, cultured Machine Learning (ML), and generative Artificial Intelligence (AI) methods. This mixture of language models has led to the expansion of language models that can accomplish various tasks like text generation, translation, summarization, and even answering questions. The model's capacity to develop human-like language is very efficacious in understanding and propagating the complexity of human communication. If the NLP domain keeps rising and advancing, these language models will evolve more significantly in how we utilize the technology and converse with computers [6].

Nonetheless, the widespread integration of groundbreaking AI-driven chatbots, such as ChatGPT, underscores the significance of being able to distinguish between text composed by humans and text generated by AI. These potential consequences could significantly impact various domains, particularly those related to digital forensics and information security. In information security, the ability to distinguish AI-generated text is critical for detecting and guarding against malicious applications of AI, such as social engineering attacks or the dissemination of misinformation and disinformation. Developing techniques to identify AI-generated texts is imperative for ensuring the precision and reliability of data. This necessity becomes especially apparent in sensitive sectors like finance and banking, political campaigns, and legal documents, including customer reviews for movies, restaurants, or products.

The remaining sections of this paper are structured as follows: Section 2 outlines details concerning the background and related works, while Section 3 expands on the dataset and proposed methodology. Following that, Section 4 puts forth the analytical results and discussion, leading to the conclusion in Section 5.

2. Background

2.1. Machine Learning

In ML, a set of algorithms is taught to a machine (computer) using data where there are no set rules. The machine learns by finding the patterns and similarities in the data. ML algorithms use patterns to be utilized. The machine decides what to carry out using the data it has determined. Like humans, computers also learn more efficiently with more data [2]. Data are essential for ML. Nowadays, due to the internet, we can discover these data efficiently. Due to the adopted strategies, ML algorithms are organized into supervised and unsupervised. As referred to in the introduction, various algorithms were utilized in this work, like the K-Nearest Neighbor (K-NN), logistic regression (LR), Support Vector Machine (SVM), Random Forest (RF), Naive Bayes (NB), and Long Short-Term Memory (LTSM) algorithms. Moreover, LTSM is operated as an unsupervised DL algorithm.

2.1.1. Supervised Learning

Supervised learning is a type of ML that uses labeled data to predict outcomes. The ML model uses a sample dataset. The preliminary purpose is to acquire the descriptions of the data class. Supervised learning algorithms aim to explore the connections and correlations between inputs and outputs [7]. So basically, in supervised learning, the data we use are associated with the class they belong to. These data are termed labeled data. In straightforward terms, classification is a strategy of supervised learning where prior data labels are utilized to predict the classification of new instances. Different classifiers, like rule-based classifiers, DT classifiers, neural network classifiers, Neuro fuzzy classifiers,

SVMs, and many more, can be used for this. In the mentioned paper, different ML models, such as SVM, Naive Bayes, logistic regression, and random forest, were manipulated with labeled textual data.

2.1.2. Unsupervised Learning

In unsupervised learning, computers identify patterns in data without any precise output or labels. No unique outcomes or associated data are available in the dataset. They look for the similarities and patterns in the input to create models. Unsupervised learning discovers patterns and relationships between the datasets that may be positive or negative. In additional terms, unsupervised learning uncovers patterns of similarities or differences across datasets. As no sample organizer or dataset is available, this strategy belongs to the unsupervised learning group. Unsupervised ML strategies have been employed to comprise the Hidden Markov Model, k-means Clustering, Long Short-Term Memory (LSTM) method, and Singular-Value Decomposition. This study also operates the k-means algorithm.

2.2. Twitter

Some microblogging platforms have raised social media networking services, which include Twitter, Facebook, and Instagram [8]. Among all of these, Twitter stands at the top because of the widely embraced SNS in which users are allowed to change the brief 140-character messages [9], which are normally known as tweets. Twitter has impressively increased its user portion to 330 million active users [10]. Because of its user-friendly domain and the comfort of sharing contexts, Twitter has been expanded into a central source of user-generated data. In the next couple of sections, we list and examine the main important characteristics that have made Twitter an essential outlet for data exchange and information.

2.3. ChatGPT-3.5

Due to its unique methodology and impressive performance on tasks involving language, OpenAI's Generative Pretrained Transformer (GPT) series has attracted a lot of interest in the field of natural language processing (NLP). A GPT is a deep learning model that initially absorbs knowledge from a sizable amount of text input, more precisely from a transformer architecture. The model gains an understanding of a language's grammar, structure, and suggestions during this early learning period.

The transformer design in the GPT is renowned for effectively managing intricate linguistic tasks. By utilizing a technique known as "self-attention" to evaluate the significance of words within a phrase, the GPT generates coherent and contextually relevant text. Its advantages extend across a variety of NLP applications, and post-initial training, the GPT showcases remarkable adaptability to diverse tasks. For tasks such as translation, summarization, question answering, and idea analysis, the GPT can be fine-tuned with specific data. The versatility and accessibility of GPTs have made them a central focus of NLP research. Moreover, beyond the realm of research, GPTs are influencing how we interact with and process text. In communication, education, and healthcare, large language models like GPTs are used. They enable chatbots, virtual assistants, and customer care applications, improving the naturalness of human–machine interactions. They tailor learning opportunities for students in education. They aid in the analysis of medical data, enhance patient care, and automate administrative processes in healthcare, resulting in more effective and user-friendly solutions.

In summary, GPTs and similar models are improving communication, education, and healthcare services across numerous industries, which is advantageous to both individuals and society as a whole.

2.4. QuillBot

QuillBot has arisen as an important landmark in the geography of NLP and text generation [11]. QuillBot was developed by AI engineers and researchers as a language model, made to face the challenges of automated paraphrasing text and growth. With the expansion of online material, the need for gadgets will enhance the grade and correctness of the text written. QuillBot answers these needs by using the significant techniques of NLP and ML for better quality text extraction, academic script, content creation, or skilled transmission [12]. However, the across-the-board acceptance of extreme AI-driven chatbots, such as Chat Generative Pretrained Transformer (ChatGPT), and text improvement devices, like QuillBot, emphasizes the significance of the capacity to distinguish between text written by humans and that developed by AI. This dissimilarity holds considerable significance across varied domains, especially within digital forensics and information security. The ability to differentiate between AI-generated and human-authored text recreates a key role in different environments, especially in the domain of information security, where it is involved in catching sight of and weakening hazards such as social engineering attacks and the spreading of misinformation. Assuring data exactness and reliability is most important in sensitive domains like accounting, banking, politics, e-commerce, and law, where AI-generated text could be utilized for scheming actions, controlling public opinion, and lawful reality. In the economic sector, punctual identification of AI-driven content protection against fraudulent economic advice and frauds. Additionally, in lawful contexts, it assures the reality of lawful papers, agreements, and academic property ownership. Therefore, the growth of strong methods for recognizing AI-generated text carries far-reaching importance for data integrity, client safety, and the protection of democratic regulations in our increasingly automated and AI-based world [13].

A moral Artificial Intelligence Generated Content (AIGC) strategy was brought up within the healthcare domain by Liao et al. The immediate goal of AIGC was to examine discrepancies between medical texts induced by ChatGPT and those written by humans. Additionally, to distinguish and specify medical texts developed from ChatGPT, they designed machine learning workflows. Originally, the researchers' datasets included medical texts generated by ChatGPT and those written by humans. Later, to specify the concept of the generated medical content, they executed and composed machine learning techniques [14].

In the moral area of students using AI tools, with a certain priority on the Large Language Model (LLM), ChatGPT, in proper educational assessments, broad research was performed by Perkins et al. [15]. Their research contained an assessment of the growth of these AI-driven tools and pointed towards the potential routes through which LLMs contribute to students' education in the digital writing domain. These contributions contained various factors, including composition and writing instructions, the collective potential between AI systems and human writers, a boost to Automated Writing Evaluation (AWE), and increased support for English as a Foreign Language (EFL) learners.

Zellers et al. [16] recommended improving the abilities of their powerful GROVER language model by combining a linear variety layer. They declared that GROVER's ability in text generation also gives it the possibility to serve as a challenging text detector. In their work, they formulated the GROVER-Mega detector to determine content developed by GROVER-Mega.

Alamleh et al. [17] evaluated the usefulness of machine learning (ML) methods in the discrimination of AI-based text from human-authored content. To fulfil this objective, these investigators collected reactions from computer science students, containing both essay and programming assignments. Thereafter, imposing this dataset, the crew carefully assessed and instructed a various array of ML methods, containing Support Vector Machines (SVMs), logistic regression (LR), neural networks (NNs), Random Forests (RFs), and Decision Trees (DTs).

While the publications deliver a restricted number of studies involving the detection of ChatGPT-generated text as well, there is a significant lack of studies that, together,

deal with the detection of text generated by paragraph rewriting tools like QuillBot and ChatGPT. Consequently, there is a specific need to improve the effectiveness of detection procedures for both ChatGPT-generated text and paragraph-rewritten content. This study aims to bridge this gap and contribute to additional complete and precise text identification procedures. In response to the detailed prerequisites of the economic sector, the immediate priority of this investigation is directed towards the designation of AI-generated content within the domain of finance. The significant contributions of this study are listed below:

- We meticulously compiled a substantial dataset of finance-related tweets sourced from Twitter, forming the cornerstone of our research.
- Utilizing this financial tweet dataset, we harnessed advanced language models, including ChatGPT and QuillBot, to craft pertinent content, augmenting the depth and breadth of our collection for in-depth studying.
- Our exploration commences with a strategic approach to address the complexities of this scenario as a multiclass categorization problem, thus enabling us to achieve high-precision classification of economic content.
- To validate the efficacy of our methodology, we conducted a thorough evaluation of a spectrum of cutting-edge machine learning models, systematically gauging their applicability to this specific task.
- We used a comprehensive feature extraction methodology that combines TF–IDF with Word2Vec, enhancing the effectiveness of our model.
- By combining a Random Forest model with Word2Vec, we achieved impressive outcomes in accuracy, precision, recall, and F1 measure.
- We investigated the significance of window size parameters in the Word2Vec approach, highlighting the effectiveness of a window size of one.

Overall, our study contributes valuable insights and promising directions for future research on AI text identification.

3. Materials and Methods

In this section, we dig into the vital factors of our proposed study, enclosing dataset details, preprocessing steps, and a comprehensive explanation of the machine learning (ML) models and techniques utilized. These elements are fundamental to understanding the framework and procedures behind our study. This study is based on dataset collection from Twitter and preparation with ChatGPT and QuillBot. The features were extracted using a hybrid approach based on TF–IDF [18] and Word2Vec [19] for embedding. Figure 1 illustrates the basic flow of our proposed study on an abstract level.

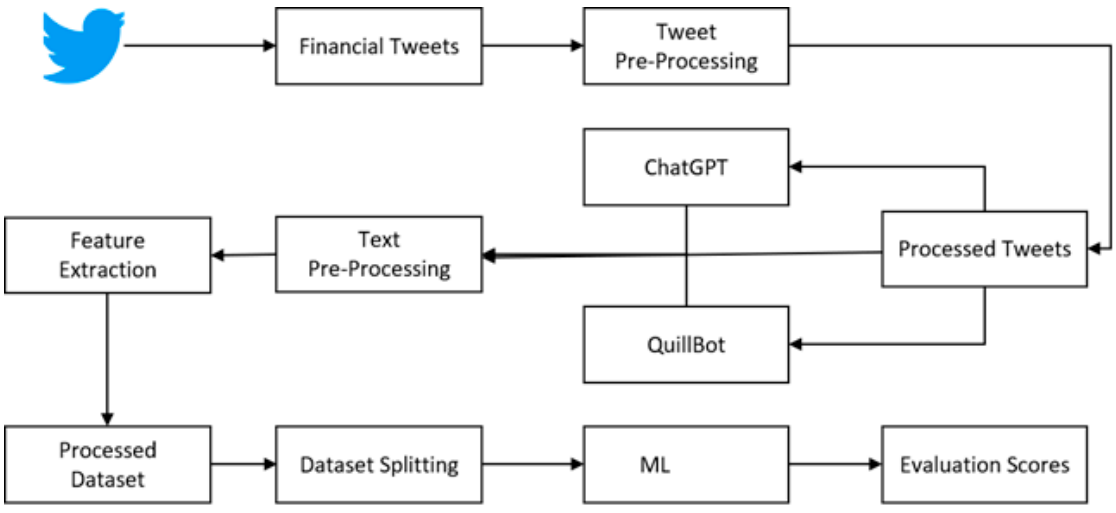


Figure 1. Abstract diagram for the proposed study.

3.1. Dataset

In this study, we compiled a dataset of tweets from Twitter using the search API. To curate appropriate tweets for our study, we utilized prejudged financial tenures, especially 'debit' and 'credit' (see Table 1 for the complete search keyword list), as our inquiry measures. In this study, initially, we only retrieved tweets from Pakistan. Likewise, we moved to renew the gathered tweets utilizing both ChatGPT-3.5 and QuillBot. To convey the objections of possible overfitting originating from category inequality, we carefully assembled a flat dataset, including 500 samples for each category. This comprehensive dataset encloses a total of 1500 text data samples, comprising 500 actual tweet texts, 500 regenerated by ChatGPT-3.5, and 500 regenerated by QuillBot (see Figure 2).

Table 1. Keywords to retrieve finance-related tweets (https://business.gov.au/finance/financial-tools-and-templates/key-financial-terms (accessed 18 April 2024) and https://www.fluentu.com/blog/english/english-for-accounting/ (accessed 18 April 2024)).

Financial Terms			
Accounts payable	Debt	Liquidity	Variable cost
Accounts receivable	Debt consolidation	Loan	Venture capital
Accounts receivable finance	Debt finance	Loan to value ratio (LVR)	Working capital
Accrual accounting	Debtor	Margin	Average cost
Amortisation	Debtors finance	Margin call	Bank draft
Assets	Default	Mark down	Bank rate
Audit	Depreciation	Mark up	Bond
Bad debts	Disbursements	Maturity date	Borrowing
Balance sheet	Discount	Net assets	Capital Good
Balloon payment	Double-entry bookkeeping	Net income	Capital inflow
Bank reconciliation	Drawings	Net profit	Capital infusion
Bankrupt	Drip pricing	Net worth	Capital loss
Bankruptcy	Employee share schemes	Overdraft facility	Capital market
Benchmark	Encumbered	Overdrawn account	Capital movement
Benchmarking	Equity	Overheads	Capital stock
Bill of sale	Equity finance	Owner's equity	Constant dollars
Bookkeeping	Excise duty	Personal property	Consumer price index
Bootstrapping	Facility	Personal Property Security Register (PPSR)	Conversion price
Bottom line	Factoring	Petty cash	Currency
Break-even point	Transactions	Plant and equipment	income tax
Budget	Financial year	Principal	Cost of capital
Capital	Financial statement	Profit	Allowance
Capital cost	Fixed asset	Profit and loss statement	Price
Capital gain	Fixed cost	Profit margin	Deposit
Capital growth	Fixed interest rate	Projection	Money
Cash	Float	R&D	Dollar
Cash accounting	Forecast	Receipts	Income
Cash book	Fringe benefits	Record keeping	Economy
Cash flow	Fully drawn advance	Refinance	Exchange Rate

Table 1. *Cont.*

Financial Terms			
Cash incoming	Goodwill	Rent to buy	Import
Cash outgoing	Gross income	Repossess	Export
Chart of accounts	Gross profit	Retention of title	Foreign Aid
Chattel mortgage	Guarantor	Return on investment (ROI)	Funds
Collateral	Hire-purchase	Return on investment (ROI) formula example	Payment
Commercial bill	Initial public offering (IPO)	Revenue	Wealth
Contingent liability	Insolvent	Single-entry bookkeeping	World Bank
Cost of goods sold	Intangible assets	Scam	Recession
Credit	Interest rate	Security	Mortgage
Creditor	Inventory	Shareholder's equity	Overdraft
Credit limit	Investment	SMSF	Shares
Credit rating	Invoice	Stock	Stocks
Credit history	Invoice finance	Stocktaking	Rally
Crowdfunding	Liability	Superannuation	Bull market
Current asset	Line of credit	Tax invoice	Bear market
Current liability	Liquidate	Turnover	Ruppee
Debit	Liquidation	Variable interest rate	
Billion	Million	Trillion	

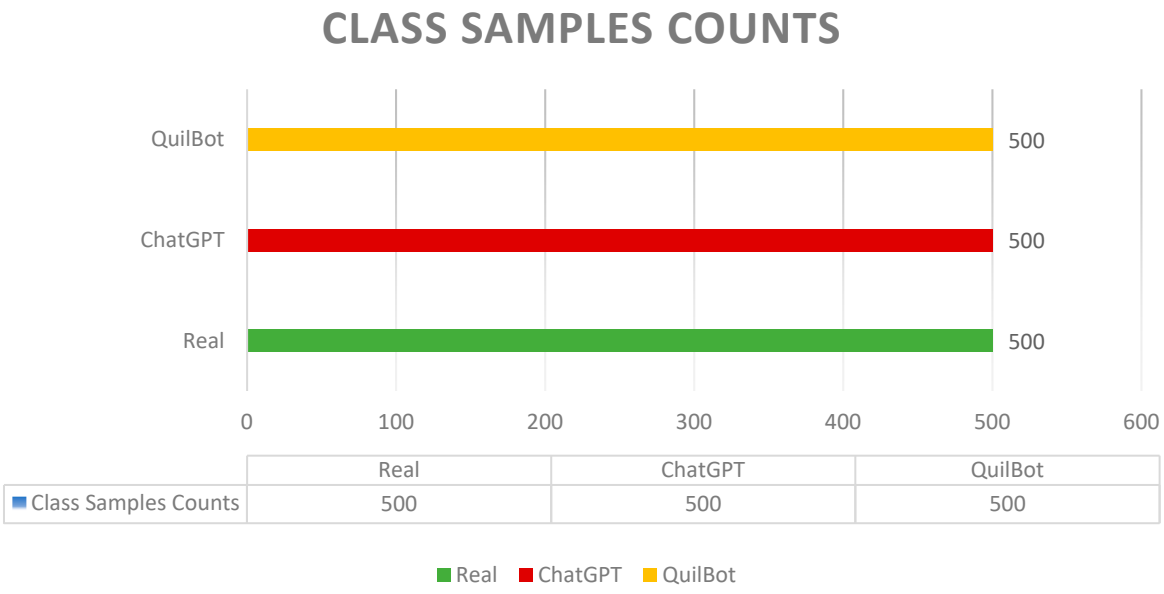


Figure 2. Dataset class distribution in terms of original financial tweets and regenerated tweets (ChatGPT and QuillBot).

Tweet Preprocessing

Tweet preprocessing is necessary for NLP tasks like sentiment and text classification [20]. We had to preprocess the tweets with the NLP toolkit i.e., the Tweet Preprocessor library [21], which involved eliminating special characters, emoticons, hashtags, punctuation, URLs, numbers, mentions, and unnecessary symbols from the text. The dataset used

in this study is formless and includes unnecessary data. Different techniques were used to remove senseless information from tweets.

- Irregular term removal: A significant ratio of users consists of elements such as URL links, hashtags, numerals, emojis, and emoticons in their tweet statements. In the context of our study, these elements encompass no materiality and may cover up the factual judgment of the tweet's sentiment. Hence, it is critical to implement a term reduction process to wipe out these odd elements and improve the accuracy of tweets.
- Accept only English tweets: In this study, our introductory focus revolves around English-language text. It is worth pointing out that both ChatGPT and QuillBot are too firmly embedded in the English language, further reinforcing our conclusion to eliminate tweets not documented in complete English from our dataset. This strategy ensures that our study remains consistent with the grammatical context of these language models.

The financial tweet text was transformed using QuillBot and ChatGPT-3.5 after cleaning the tweets, which included the removal of hashtags and emojis. The prepared dataset samples for better understanding can be seen in Table 2. The word cloud after the tweet pre-processing can be seen in Figure 3.

Table 2. Prepared dataset samples and the transformation of the original tweet to ChatGPT and QuillBot.

Original Tweet	ChatGPT Text	QuillBot Text
"trillion even issue trillion roughly budget year period anyone bit knowledge government able figure along medium curb perhaps also worth talk medium capability."	"A trillion is a matter that cannot be dismissed lightly. With approximately a trillion allocated within the annual budget, anyone possessing even a modicum of knowledge about government finance should be capable of discerning the significance. The idea of a medium-term approach to rein in such expenditures warrants consideration, and it might be worth discussing our medium-term capabilities as well."	"infinity even problem trillion around budget year period anyone with some understanding of how the government would calculate a medium curb, possibly also worth discussing a medium capacity."
"one trillion rupees one thousand billion entire cost billion only exaggeration ma humble suggestion in future ppp dispensation suggest consider ministry finance."	"A sum of one trillion rupees... equating to one thousand billion, the entire cost is in the billions, not merely an exaggeration. In my humble suggestion, for future PPP dispensations, I propose that the Ministry of Finance takes into account this consideration."	"One trillion rupees, or $1000 billion in total cost, is merely an exaggeration. My modest recommendation is that in future PPP administrations, ministry finance be taken into consideration."
"the exchange rate lock."	"The stability of the exchange rate remains unchanged."	"a lock on the exchange rate."

Figure 3. Word cloud after tweet pre-processing.

3.2. Text Preprocessing

Text preprocessing transforms unstructured data into a useful format, allowing models to identify patterns and extract valuable information. It involves several steps to convert unstructured text into structured data, making training easier and faster. In this study, we carried out a series of preprocessing steps to productively wipe out this noise from the tweet texts. For effective feature extraction, we focused on removing stop words and performing lemmatization.

- **Stop word removal:** This involves the removal of common words that lack significant meaning or contribute minimally to the overall understanding of the text. Stop words such as "the", "is", and "and", are examples of these. The elimination of these words can enhance the efficiency and accuracy of text identification tasks [22].
- **Lemmatization:** Lemmatization, an essential preprocessing phase in Natural Language Processing (NLP), involves specifying the morphological root of terms [23]. The word cloud after the post-processing can be seen in Figure 4.

Figure 4. Word cloud after post-processing.

3.3. Feature Extraction

The next step in this study is feature extraction, a pivotal aspect in tackling classification problems. To sift through words and separate the wheat from the chaff, we employed two different approaches: vectorization techniques and feature embedding. Specifically, we employed TF–IDF for vectorization, and to complement that, we harnessed pre-trained Word2Vec as a hybrid approach.

3.3.1. Term Frequency–Inverse Document Frequency (TF–IDF)

TF–IDF serves as a foundational preprocessing step for converting tweet text data into numerical representations before the application of any classification model [24]. It involves two statistical methods: TF (term frequency), which quantifies the total number of times a word appears in a document, and IDF (inverse document frequency), which measures the overall occurrence of terms across documents [7]. The weight assigned to a term is calculated as the product of TF and IDF, thereby gauging its relevance and significance within a specific document. Equations (1) and (2) depict the formulas for computing *TF* and *IDF*, respectively. In these equations, 't' represents the term with a frequency of 'f'; 'd' denotes the document; and 'N' signifies the total number of documents containing the term 't' [18]. TF–IDF is a product of Equations (1) and (2):

$$TF\ (t,d) = \frac{f_t}{f} \tag{1}$$

$$IDF\ (d) = \frac{N_d}{N} \tag{2}$$

3.3.2. Word2Vec

Word2Vec [19] is a widely employed technique for acquiring word embeddings through the utilization of neural networks. This trained model leverages mathematical operations on the text corpus to position words with similar meanings within the same vector space. There exist two primary approaches for Word2Vec: one is the skip-gram method, which is focused on predicting context words based on a given word, and the other is the continuous bag of words (CBOW) method, where the predicted word relies on its context. In the context of our research, we implemented the CBOW algorithm, which was trained on the corpus using specific parameters, including a window size (W) of 5, a minimum word frequency of 5, and a dimensionality (D) of 100.

3.4. Classification Algorithms

For our classification, we employed a variety of machine learning classifiers, i.e., random forests, decision trees, logistic regression, KNN, and SVMs. We opted for these algorithms due to their extensive use in text classification, known for their exceptional accuracy. Our objective was to assess and compare the performance of each classifier to determine the optimal model.

3.4.1. Logistic Regression

Logistic regression, a pioneering method in machine learning, was originally developed by David Cox in 1958 [25]. Over time, it has emerged as one of the most extensively applied methodologies in the field. This approach utilizes probabilities to describe and anticipate results, rendering it especially suitable for tasks centered around categorical classification. In our specific context, we implemented multinomial logistic regression for multiclass classification, capitalizing on the multinomial probability distribution for predictive modeling. The core concept underpinning logistic regression is to predict the class with the highest posterior probability. This decision-making principle is elucidated in Equation (3). Essentially, logistic regression empowers us to make well-informed predic-

tions by assessing the probabilities associated with various outcomes, establishing it as a crucial tool within the domain of machine learning.

$$\check{L} = argmax_i P(L = i|T) \quad (3)$$

In this context, P represents the posterior probabilities; $\check{L}$ stands for the predicted label; i signifies the total number of labels; and T corresponds to the input text.

3.4.2. KNN

K-Nearest Neighbors (KNN) is a classification algorithm that was developed by Evelyn Fix and Joseph Hodges in 1951 [26], and it was further refined and popularized by Thomas Cover in 1967 [27]. It classifies data points by determining the majority class among their k-nearest neighbors, employing a distance metric to gauge proximity in a multidimensional feature space, as shown in Equation (4). The selection of the distance metric and the value of k hold significance as they are pivotal parameters affecting the algorithm's effectiveness. Unlike certain machine learning approaches with explicit equations, KNN does not have a straightforward mathematical formula. Instead, it operates on the principle of identifying the most similar data points in the training set to inform classification decisions. Despite lacking a formal equation, KNN remains a straightforward and extensively utilized tool for diverse classification tasks.

$$D(X, Y) = \sqrt{\sum_{i=1}^{n} (X_i - Y_i)^2} \quad (4)$$

3.4.3. SVM

A SVM is a sophisticated supervised learning algorithm that is usually used for classification tasks [28]. It works by expressing input information as vectors and then projecting them into a higher-dimensional space to achieve unambiguous segregation between various classes. SVMs constitute a powerful technique that may be used with a variety of kernel functions, including Gaussian, radial, linear, and polynomial kernels. This adaptability allows for the efficient handling of a wide range of datasets. A specific kernel function was chosen and used in the scope of this research [29]. Selecting the kernel function plays a crucial role in determining how the SVM models the relationships between data points, thereby impacting the success of the classification task. The versatility and effectiveness of SVMs position them as a valuable tool applicable across a wide range of domains, including, but not limited to, image recognition and natural language processing.

3.4.4. Random Forest

The random forest classifier belongs to the family of ensemble approaches [30]. It uses the combined strength of numerous decision trees, acting as base learners, rather than a single decision tree. Each of these individual trees is trained independently, and their combined predicted accuracy is improved by averaging the dataset's results. This ensemble strategy, which employs a diverse set of decision trees, is critical in improving forecast accuracy and reducing the risk of overfitting. As a result, it contributes to the development of more durable and accurate models.

3.4.5. Decision Tree

The theoretical foundation of RFs is rooted in decision trees. Decision trees employ a recursive process to partition the feature space into rectangular regions, utilizing modes or means as forecasts for observations within these areas. This approach is often referred to as the decision tree method, as it represents the division criteria of the feature space in the form of a tree structure. In regression tasks, data with similar response values are grouped, and each resulting region is associated with a fixed value, typically the mean.

4. Results and Discussion

In this section, we present the findings derived from our experimental exploration of various feature extraction approaches coupled with machine learning models for identifying financial machine-generated content. We conducted a series of three feature extraction experiments to analyze this approach.

4.1. Evaluation Matrics

Evaluation metrics are pivotal for appraising the performance of machine learning. They bear considerable significance within the domains of machine learning and statistical research. In this study, our emphasis lies on a selection of essential evaluation metrics to measure the efficiency of ML models:

- Accuracy: Assesses the global accuracy of the model's predictions by determining the ratio of correctly classified samples to the total number of samples. However, accuracy alone may prove inadequate for evaluation, particularly when handling imbalanced datasets or situations where different types of errors carry differing consequences.

$$Accuracy = (TP + TN)/((TP + FP + FN + TN) \tag{5}$$

- Precision: Quantifies the model's capacity to accurately detect positive samples among the predicted positives. It computes the ratio of true positives to the sum of true positives and false positives. Precision primarily evaluates the trustworthiness of positive predictions.

$$Precision = \frac{(TP)}{(TP + FP)} \tag{6}$$

- Recall: Often referred to as sensitivity or the true positive rate, recall assesses the model's capability to correctly identify positive samples among all the genuine positives. It computes the ratio of true positives to the sum of true positives and false negatives. Recall primarily assesses the comprehensiveness of positive predictions.

$$Recall = \frac{(TP)}{(TP + FN)} \tag{7}$$

- F1 score: The harmonic mean of precision and recall. It provides a single metric that balances both precision and recall, making it useful for when there is an uneven class distribution or an equal emphasis on both types of errors. The *F1* score ranges from 0 to 1, with 1 denoting the best performance.

$$F - Measure = \frac{2 \times Precision \times Recall}{(Precision + Recall\ )} \tag{8}$$

In the context of multiclass classification, accuracy is determined by dividing the number of correct predictions (true positives and true negatives) by the total number of predictions, irrespective of the class. Conversely, for multiclass classification, precision, recall, and *F1* scores are computed in the form of weighted averages. Weighted averaging assigns a weight to each class based on its representation in the dataset. To derive weighted metrics, precision, recall, and *F1* scores for each class are multiplied by their respective weights and subsequently summed. The total is then divided by the overall weight, a strategy that effectively addresses class imbalance within the dataset.

4.2. Experimental Setup

In our investigation, we chose a cloud-based strategy, with Google Colab [31] serving as our initial computing platform. This cloud environment gives you access to powerful GPUs and TPUs, which speed up the experimentation and research procedures for machine

learning models. Using Google Colab enables computational scalability, collaboration, and streamlined research, allowing us to focus on our study's key objectives without the requirement for specific local hardware support.

4.3. Experimental Results and Discussion

We tested three different feature extraction algorithms in our experiments: TF–IDF, Word2Vec, and Ensemble. The goal was to find the best feature extraction method for detecting machine-generated financial material. We highlight the results and discoveries, emphasizing the approach that produced superior results and digging into the significance of our findings. This investigation establishes the framework for future research in this area. Table 2 shows the outcomes of the machine learning models with TF–IDF vectorization and a max_length of 1000.

Our testing experiments were conducted on a dataset comprising 150 samples, as depicted in Figure 5, encompassing various financial content categories. These results highlight the RF model's effectiveness in accurately classifying financial content into multiple classes, providing valuable insights for real-world applications. Future research avenues may explore feature engineering, model interpretability, and the evaluation of different ensemble methods, as well as address considerations of scalability and efficiency for practical deployment.

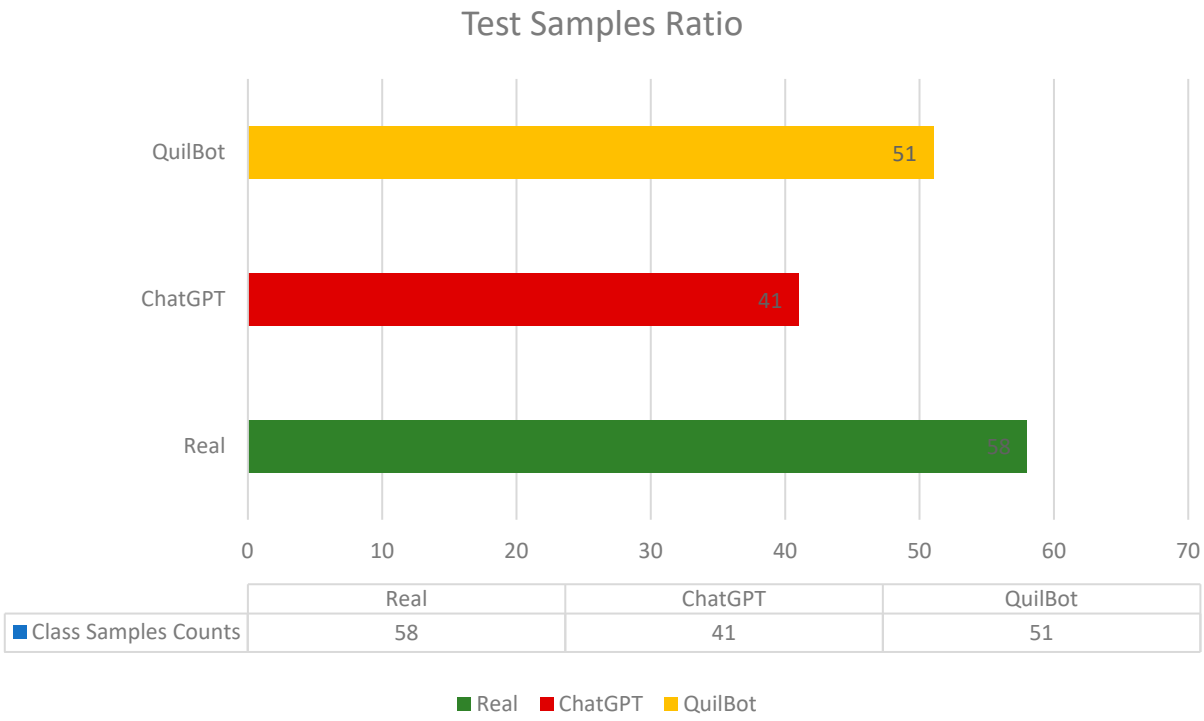


Figure 5. Test set sample class ratio.

We conducted a detailed statistical analysis to validate the appropriateness of the classes for model training and evaluation. This analysis included examining the distribution of classes within the dataset to ensure that each class was adequately represented. We also assessed the balance and bias of the classes to ensure that the model would not be skewed towards any particular class during training. Additionally, we performed statistical tests to compare the characteristics of each class and identify any significant differences that could impact the model's performance.

In our study of multiclass financial machine-generated content identification, Random Forest (RF) emerged as the most accurate model, as evidenced in Table 3. Employing 'GridSearchCV', we fine-tuned the crucial hyperparameter 'n_estimators' to optimize the RF's performance [32].

Table 3. Evaluation scores of ML models for machine-generated financial content identification using TF–IDF.

ML Models	Accuracy	Weighted Precision	Weighted Recall	Weighted F1
Random Forest	0.39	0.40	0.39	0.39
LR	0.37	0.39	0.37	0.37
KNN	0.24	0.11	0.24	0.15
Decision tree	0.33	0.35	0.33	0.33
SVM	0.23	0.25	0.23	0.23

In our research, Table 4 assumes pivotal importance as it visually represents the confusion matrices of ML models [33] with TF–IDF-based features. These matrices serve as a comprehensive source of evaluation metrics extracted from a four-term matrix. By showcasing the true positives, true negatives, false positives, and false negatives, the confusion matrices provide a deeper understanding of the model's performance and their capacity to make accurate predictions across multiple classes. These metrics are essential for assessing the model's precision, recall, F1 score, and overall effectiveness in the context of multiclass financial machine-generated content identification. They enable us to delve into the intricate dynamics of model performance, thereby contributing to more informed decisions and further research in this domain.

Table 4. Confusion matrix for machine-generated financial content identification using TF–IDF.

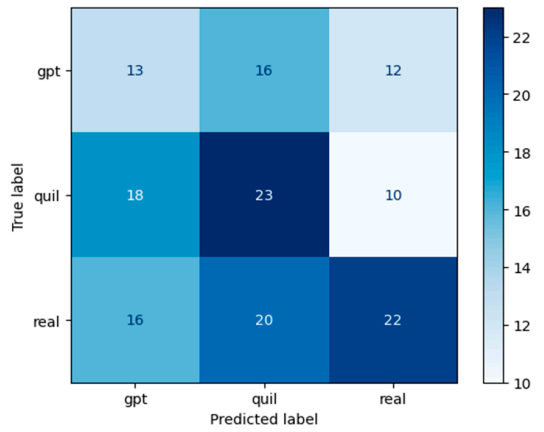
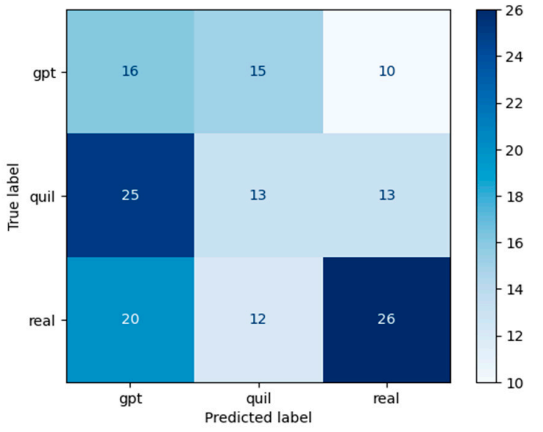
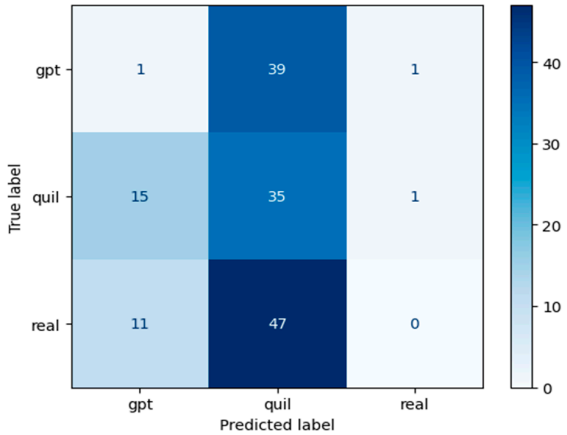
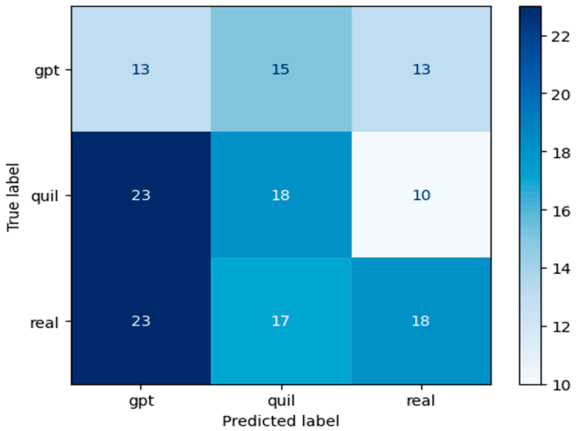
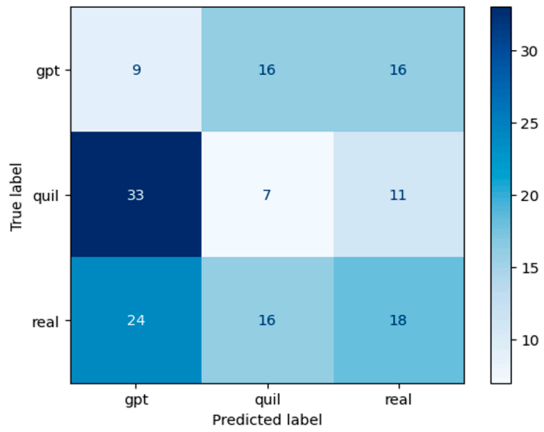
Confusion Matrix (TF-IDF)	
 (RF)	 (LR)
 (KNN)	 (Decision Tree)

Table 4. *Cont.*

Confusion Matrix (TF-IDF)

(SVM)

In this study, the TF–IDF feature extraction method, although pairing effectively with the Random Forest machine learning model, yielded an accuracy score of 0.39, which falls short of the acceptable threshold for model effectiveness. To address this challenge and enhance our evaluation scores, we adopted the Word2Vec approach with a window size of 5 and a vector size of 100 (see Table 5), keeping the parameters consistent with those used in the TF–IDF scenario.

Table 5. Machine-generated financial content identification using Word2Vec.

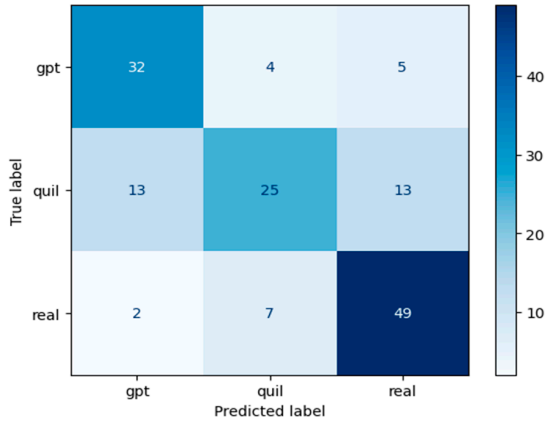
ML Models	Accuracy	Weighted Precision	Weighted Recall	Weighted F1
Random Forest	0.71	0.70	0.71	0.70
LR	0.58	0.56	0.58	0.56
KNN	0.61	0.63	0.61	0.61
Decision tree	0.56	0.57	0.56	0.56
SVM	0.59	0.64	0.59	0.61

In this study, we achieved a noteworthy accuracy of 0.71 by implementing the Random Forest model in combination with the Word2Vec feature extraction approach. This accuracy represents a substantial improvement compared to our prior use of the TF–IDF approach. The significance of this achievement lies in its real-world applicability, especially in the realm of financial machine-generated content identification. The higher accuracy underscores the potential for more precise content classification, carrying implications for applications in finance, information retrieval, and content filtering.

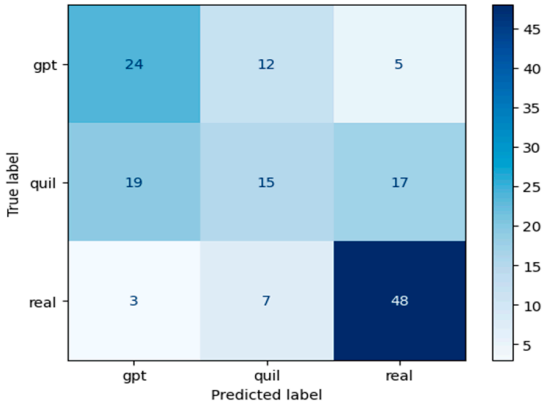
The effectiveness of the Random Forest and Word2Vec methods highlight the critical role of feature extraction methods in machine learning. Word2Vec, with its ability to capture semantic relationships within financial text data, contributed significantly to the improved accuracy [34–38]. This outcome, coupled with the robustness of the Random Forest model, suggests a more reliable and accurate means of classifying financial machine-generated content. In essence, our research not only advances our understanding of model performance but also establishes a strong foundation for future developments in automated content identification within the financial sector and beyond (see Table 6 for Word2Vec-based confusion matrices of ML models).

Table 6. Confusion matrix for machine-generated financial content identification using Word2Vec.

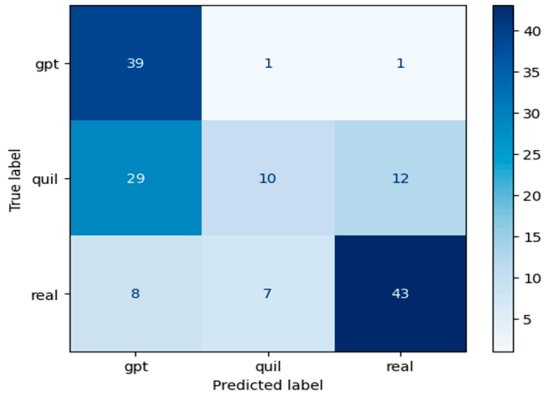
Confusion Matrix (Word2Vec)



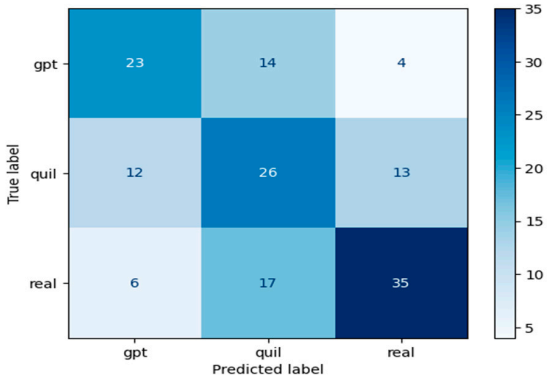
(RF)



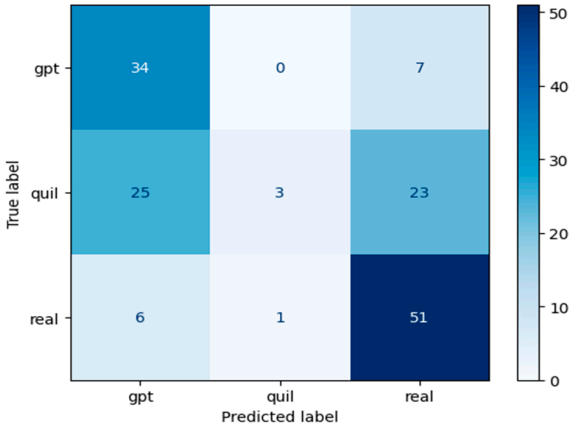
(LR)



(KNN)



(Decision Tree)



(SVM)

In the extension of our study, we ventured into an intriguing exploration of ensemble methods by combining two distinctive feature extraction techniques—TF–IDF and Word2Vec. This strategy is noteworthy because it investigates the combination of two complementary approaches, each with their own special advantages in the area of machine-generated content recognition. This addition is valuable since it aims to provide better performance and synergy. A traditional but effective feature extraction technique called TF–IDF is excellent at collecting document specificity and term frequency, giving important information about the importance of individual words in a document. However, Word2Vec, which is based on distributed

word representations, is particularly good at capturing contextual information and semantic linkages, which are essential for understanding meaning and context in textual data.

By combining TF–IDF and Word2Vec (See Table 7), we aim to harness the strengths of both techniques. This study has implications for sentiment analysis, automated categorization, and information retrieval, among other areas where textual content analysis is crucial. These applications extend beyond the finance industry. As data-driven decision-making and information processing become more and more demanding, this represents a significant step towards maximizing machine-generated content identification across many industries.

Table 7. Machine-generated financial content identification using an ensemble approach (TF–IDF + Word2Vec).

ML Models	Accuracy	Weighted Precision	Weighted Recall	Weighted F1
Random Forest	0.67	0.67	0.67	0.67
LR	0.65	0.63	0.65	0.63
KNN	0.32	0.38	0.32	0.33
Decision tree	0.53	0.54	0.53	0.53
SVM	0.63	0.61	0.63	0.61

In this study, we introduced and evaluated an ensemble strategy for recognizing machine-generated financial information that combines the strengths of the TF–IDF and Word2Vec feature extraction techniques. The findings, shown in Table 6, showed an accuracy of 0.67, indicating a significant improvement over the TF–IDF approach alone. Although it falls somewhat short of the accuracy reached with the Word2Vec alone strategy, this result demonstrates the potential of ensemble methods in improving the model's capacity to categorize financial content effectively. The relevance of this discovery resides in the adaptability and versatility of ensemble techniques, which allow us to capitalize on the distinct advantages of various feature extraction methods. This method creates a balance between word frequency-based understanding (TF–IDF) and Word2Vec's semantic context.

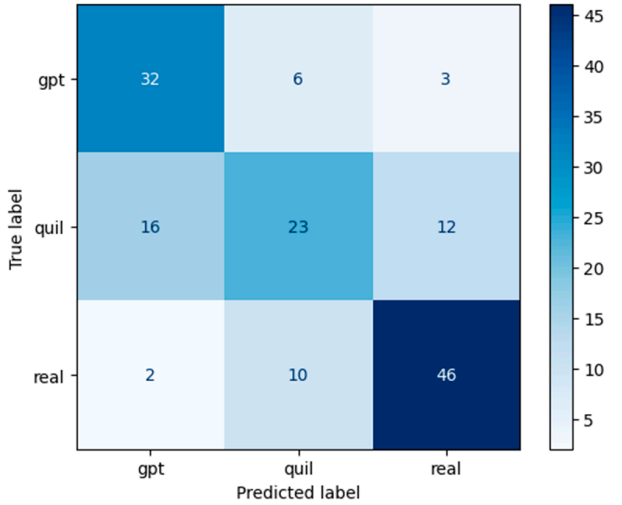
The confusion matrix, shown in Table 8, provides more insight into the model's performance, acting as a significant resource for measuring precision, recall, and overall efficacy. This study opens the door to additional investigations into ensemble approaches and their function in refining content identification in domains where precise classification is critical. It establishes the groundwork for future advances in machine-generated content identification with real-world applications in finance and beyond.

These results demonstrate that Word2Vec is quite effective in the setting of our investigation. Furthermore, we employed a systematic approach to evaluate how different window sizes affect Word2Vec when paired with the Random Forest model, a frequently used word embedding technique. Our goal was to understand the impact of different window sizes on the quality of word embeddings and how they affect machine learning models through a series of experiments; see Figure 6. What we discovered underscores the vital relevance of selecting an appropriate window size based on the unique objectives of natural language processing activities. It emphasizes the trade-off between catching local syntactic links and comprehending the broader semantic context.

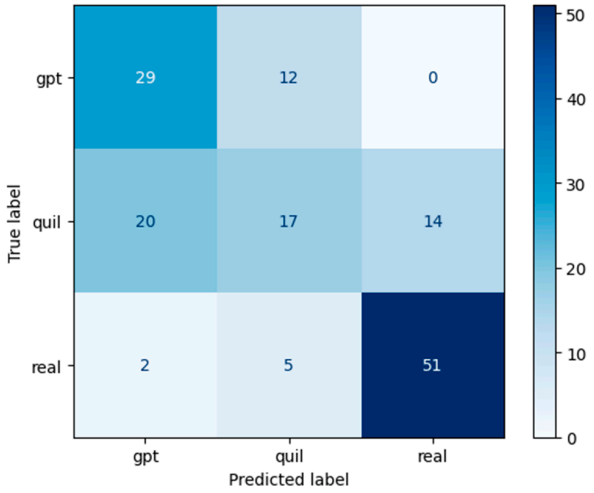
Figure 6 reveals the critical significance of the window size parameter in Word2Vec, exhibiting a significant performance gain when using a smaller window size of 1, resulting in an amazing accuracy, precision, recall, and F1 score all of 0.74; see Table 9. This underscores the importance of adjusting hyperparameters, particularly window size, in natural language processing tasks. The choice of window size has a substantial impact on the model's capacity to grasp complex relationships between words, both syntactic and semantic. Being able to pick the right window size is essential for retrieving the most out of word embeddings, making it an important aspect for researchers and practitioners looking to improve the performance of different natural language processing applications.

Table 8. Confusion matrix for machine-generated financial content identification using TF–IDF + Word2Vec.

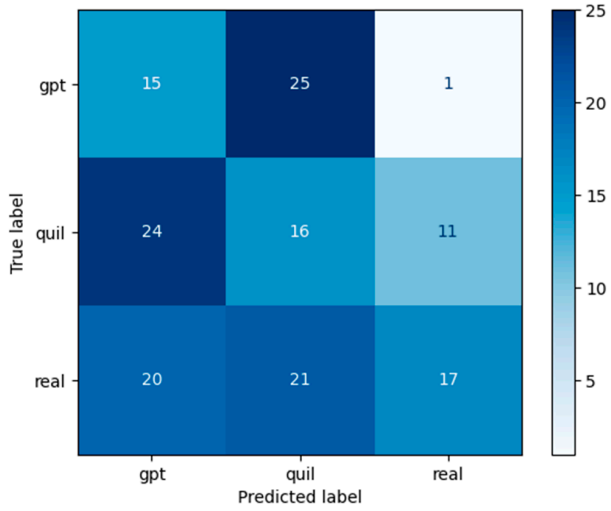
Confusion Matrix (TF–IDF + Word2Vec)



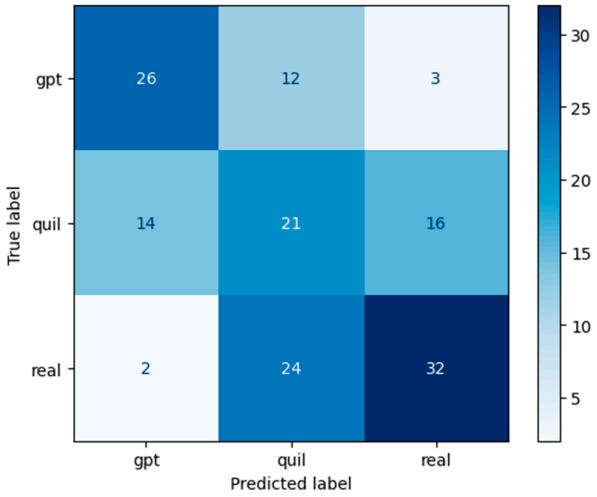
(RF)



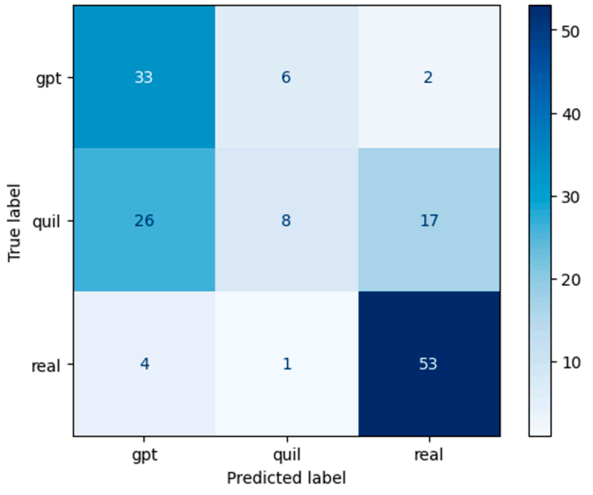
(LR)



(KNN)



(Decision Tree)



(SVM)

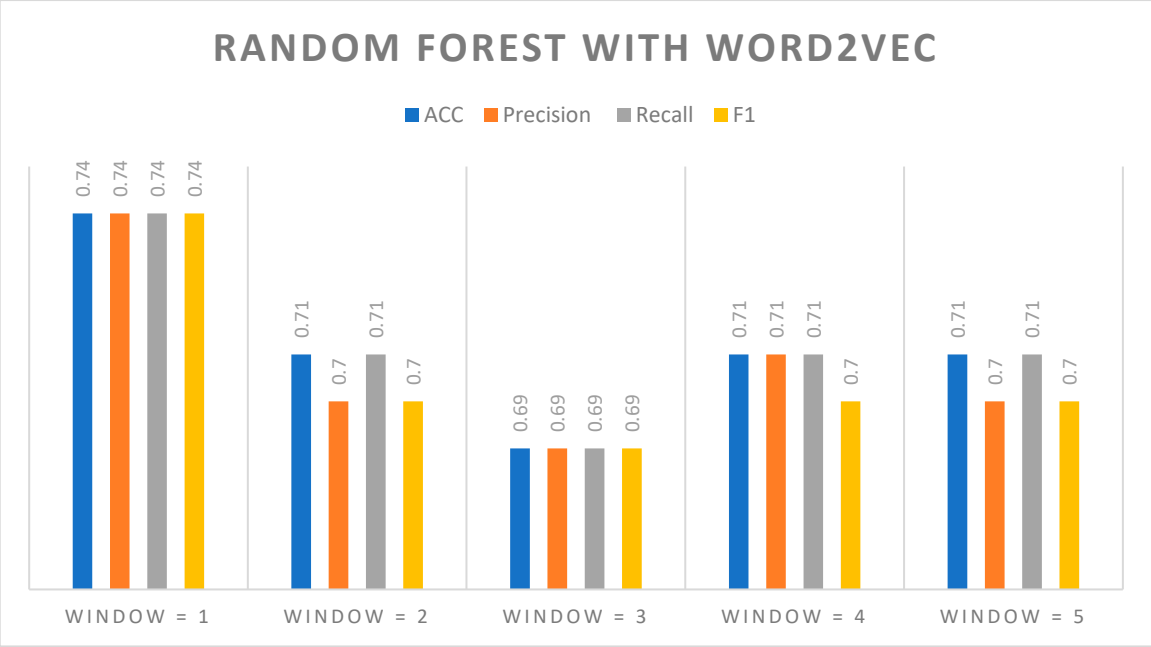


Figure 6. Window size importance for machine-generated financial content identification using Word2Vec and the RF model.

Table 9. Classification report of the RF model for machine-generated financial content identification using Word2Vec (window = 1). The macro and weighted averages were calculated based on the evaluation scores of three individual classes, i.e., gpt, quil, and real.

	Precision	**Recall**	**F1**	**Support**
Gpt	0.80	0.80	0.80	41
Quil	0.71	0.59	0.65	51
Real	0.72	0.83	0.77	58
Macro avg	0.75	0.74	0.74	150
Weighted avg	0.74	0.74	0.74	150
Overall Accuracy = 0.74				

Figure 7 illustrates the confusion matrix for the RF model using the Word2Vec approach with a window size of one, which proves to be more effective than other window sizes.

In Figure 8, we present the ROC curve for the RF model with Word2Vec, specifically using a window size of one. This curve is like a report card for this model, showing how well it can tell apart different classes. It is worth noting that the model performed well, scoring 0.94 for the 'GPT' class, 0.77 for the 'Quil' class, and 0.89 for the 'Real' class. These scores highlight how effective this model is at accurately categorizing data, especially in distinguishing between these specific classes.

In addition to our extensive experiments, we have created an advanced dataset to further bolster our findings. This new dataset comprises 1000 samples each of real, ChatGPT-regenerated text, and QuillBot-regenerated content. Including these additional data enhances the robustness of our study and provides further support for our results. From this extensive dataset, we have considered 10% of the dataset as a test set with the proper balance distribution to avoid overfitting and to observe the model nature; see Figure 9 for the train and test dataset split by class-wise distribution. We applied the same RF model with Word2Vec (Window Size = 1) for a fair comparison.

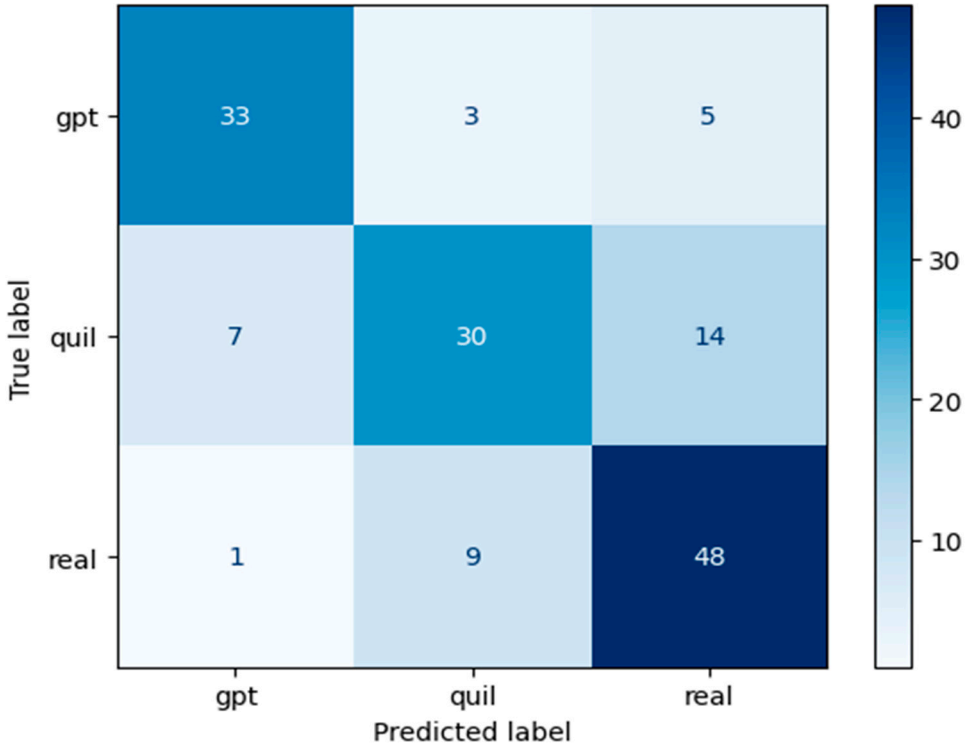


Figure 7. Confusion matrix of the RF model using Word2Vec (window = 1).

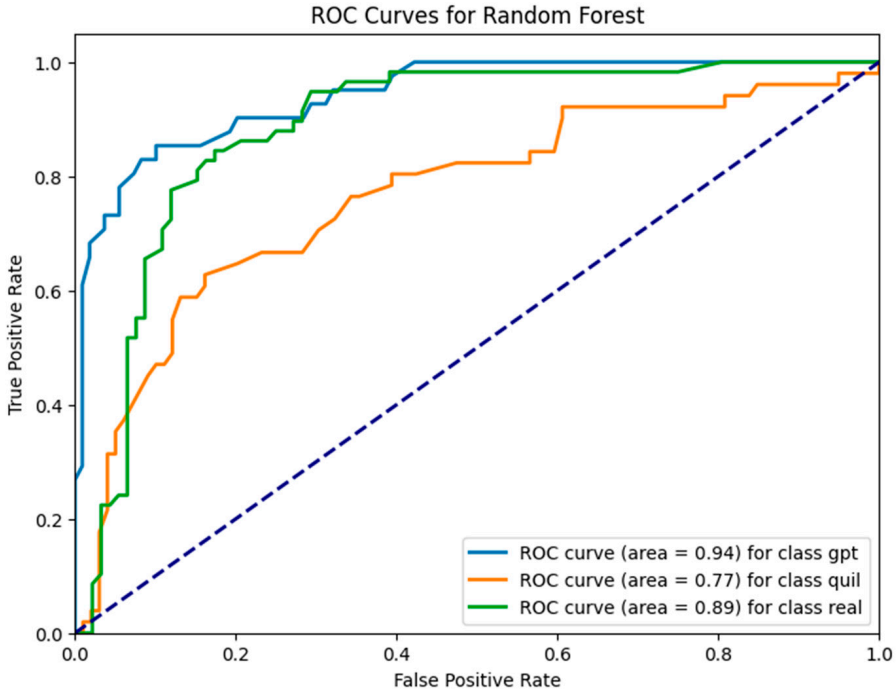


Figure 8. ROC curve of the RF model using Word2Vec (window = 1).

The word cloud for this advanced dataset can be seen in Figure 10. The Pakistan word is prominent as the tweets were from the Pakistan region.

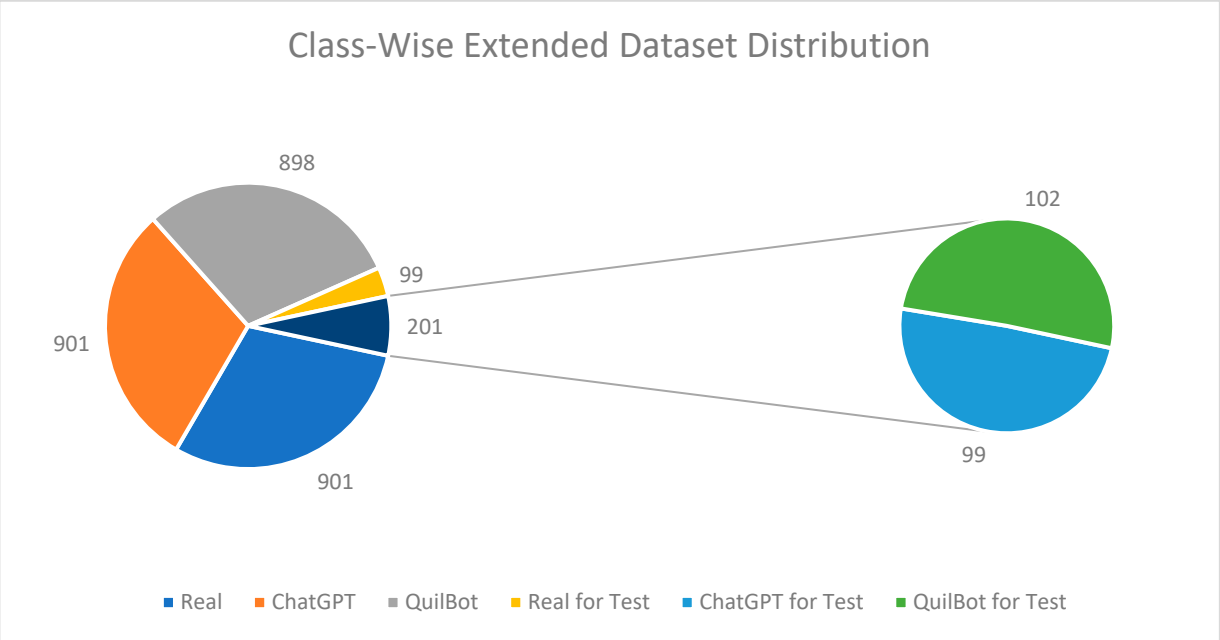


Figure 9. Advanced dataset class distribution in terms of original financial tweets and regenerated tweets (ChatGPT and QuillBot).

Figure 10. Advanced dataset word cloud.

Our study initially achieved an accuracy of 0.74 with the initial dataset size. However, after increasing the dataset size, we achieved a slightly lower but still impressive accuracy of 0.72. This demonstrates the positive impact of a larger dataset on the performance of our model, highlighting the importance of having a sufficient amount of data for training machine learning models in our domain; see Table 10 for the classification report and Figure 11 for the confusion matrix.

Table 10. Advanced dataset classification report of the RF model for machine-generated financial content identification using Word2Vec (window = 1). The macro and weighted averages were calculated based on the evaluation scores of three individual classes, i.e., gpt, quil, and real.

	Precision	Recall	F1	Support
Gpt	0.69	0.74	0.71	99
Quil	0.64	0.47	0.54	102
Real	0.86	0.96	0.90	99
Macro avg	0.71	0.72	0.71	300
Weighted avg	0.71	0.72	0.71	300
Overall Accuracy = 0.72				

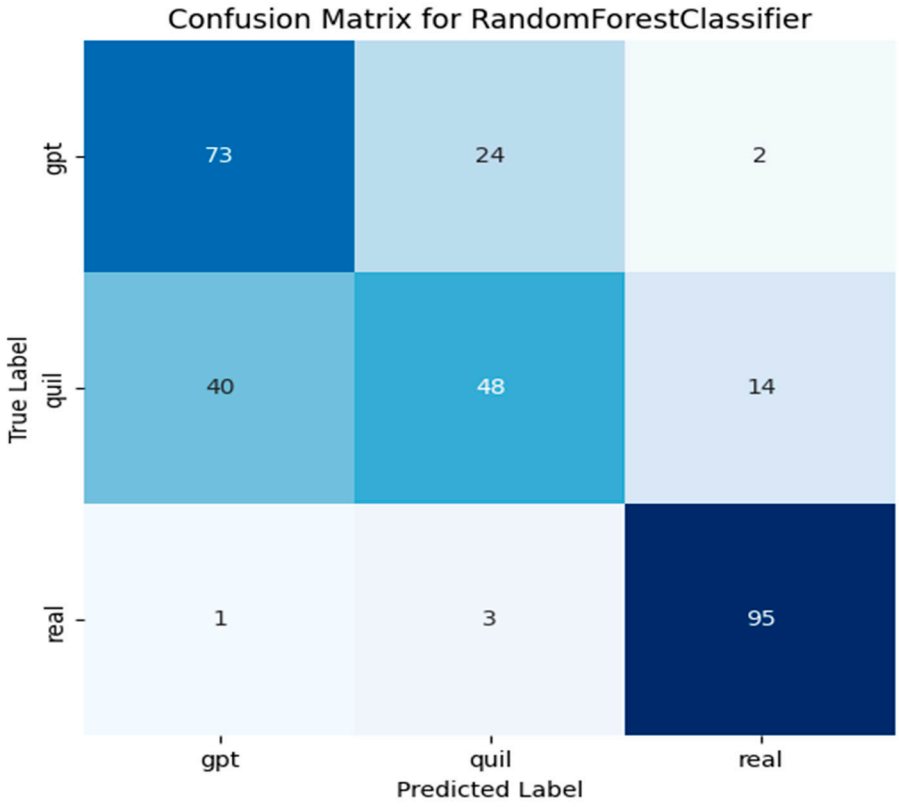


Figure 11. Confusion matrix of the RF model using Word2Vec (window = 1) for the extended dataset.

5. Conclusions

In this research, we put together a thoughtfully selected dataset with the precise goal of spotting machine-generated financial content. We utilized various machine learning models and techniques for extracting features to understand how these models make decisions and reveal hidden patterns. Our emphasis was on the design and implementation of TF–IDF, Word2Vec, and ensemble approaches for feature extraction, with the Random Forest machine learning model at the forefront of our classification efforts. Notably, the Random Forest model, when paired with the Word2Vec approach, proved to be the most effective combination, yielding remarkable results. Furthermore, we explored the impact of window size in the Word2Vec approach, finding that a window size of one produced the highest scores across metrics such as accuracy, precision, recall, and F1 score, all achieving an impressive value of 0.74. We also achieved an accuracy of 0.72, precision of 0.71, recall of 0.72, and F1 score of 0.71 for our extended prepared dataset. While our study undeniably demonstrates significant promise and yields commendable results, we maintain transparency by acknowledging the limitations posed by our relatively small dataset. This

acknowledgement not only underscores the conscientiousness of our research but also invites future endeavors to build upon our findings and explore their generalizability to larger datasets. In essence, our work lays a solid foundation, providing valuable insights that pave the way for further exploration and application in the realm of feature extraction and machine learning classification. Future enhancements are envisioned through the expansion of the dataset, offering the potential for further refinement and improved model performance.

Author Contributions: Conceptualization, M.A.A., Ș.C.G., C.D., A.I. and S.M.; methodology, M.A.A., Ș.C.G., C.D., A.I. and S.M.; validation, M.A.A., Ș.C.G., C.D., A.I. and S.M.; investigation, M.A.A., Ș.C.G., C.D., A.I. and S.M.; data curation, M.A.A., Ș.C.G., C.D., A.I. and S.M.; writing—original draft preparation, M.A.A., Ș.C.G., C.D., A.I. and S.M.; writing— review and editing, M.A.A., Ș.C.G., C.D., A.I. and S.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data will be made available on request.

Acknowledgments: We acknowledge the use of ChatGPT (https://chat.openai.com/, 1 January 2024) and QuillBot (https://quillbot.com/, 1 January 2024) for fake machine-generated dataset preparation.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Muneer, A.; Alwadain, A.; Ragab, M.G.; Alqushaibi, A. Cyberbullying Detection on Social Media Using Stacking Ensemble Learning and Enhanced BERT. *Information* **2023**, *14*, 467. [CrossRef]
2. Hadi, M.U.; Al Tashi, Q.; Qureshi, R.; Shah, A.; Muneer, A.; Irfan, M.; Zafar, A.; Shaikh, M.B.; Akhtar, N.; Wu, J.; et al. Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations, and Future Prospects. *Authorea Prepr.* **2023**. [CrossRef]
3. Tyagi, N.; Bhushan, B. Demystifying the Role of Natural Language Processing (NLP) in Smart City Applications: Background, Motivation, Recent Advances, and Future Research Directions. *Wirel. Pers. Commun.* **2023**, *130*, 857–908. [CrossRef]
4. Khurana, D.; Koli, A.; Khatter, K.; Singh, S. Natural language processing: State of the art, current trends and challenges. *Multimed. Tools Appl.* **2023**, *82*, 3713–3744. [CrossRef]
5. Pavlik, J.V. Collaborating with ChatGPT: Considering the Implications of Generative Artificial Intelligence for Journalism and Media Education. *J. Mass Commun. Educ.* **2023**, *78*, 84–93. [CrossRef]
6. Yew, A.N.J.; Schraagen, M.; Otte, W.M.; van Diessen, E. Transforming epilepsy research: A systematic review on natural language processing applications. *Epilepsia* **2022**, *64*, 292–305. [CrossRef]
7. Muneer, A.; Fati, S.M. A Comparative Analysis of Machine Learning Techniques for Cyberbullying Detection on Twitter. *Futur. Internet* **2020**, *12*, 187. [CrossRef]
8. Fati, S.M.; Muneer, A.; Alwadain, A.; Balogun, A.O. Cyberbullying Detection on Twitter Using Deep Learning-Based Attention Mechanisms and Continuous Bag of Words Feature Extraction. *Mathematics* **2023**, *11*, 3567. [CrossRef]
9. Gligorić, K.; Anderson, A.; West, R. Adoption of Twitter's New Length Limit: Is 280 the New 140? *arXiv* **2020**, arXiv:2009.07661.
10. How Many Users Does Twitter Have? Available online: https://www.bankmycell.com/blog/how-many-users-does-twitter-have (accessed on 4 September 2023).
11. Fitria, T.N. QuillBot as an online tool: Students' alternative in paraphrasing and rewriting of English writing. *Englisia J.* **2021**, *9*, 183–196. [CrossRef]
12. Nurmayanti, N.; Suryadi, S. The Effectiveness of Using Quillbot In Improving Writing for Students of English Education Study Program. *J. Teknol. Pendidik.* **2023**, *8*, 32–40. [CrossRef]
13. Alawida, M.; Mejri, S.; Mehmood, A.; Chikhaoui, B.; Abiodun, O.I. A Comprehensive Study of ChatGPT: Advancements, Limitations, and Ethical Considerations in Natural Language Processing and Cybersecurity. *Information* **2023**, *14*, 462. [CrossRef]
14. Liao, W.; Liu, Z.; Dai, H.; Xu, S.; Wu, Z.; Zhang, Y.; Liu, T. Differentiate ChatGPT-Generated and Human-Written Medical Texts. *arXiv* **2023**, arXiv:2304.11567.
15. Perkins, M. Academic integrity considerations of AI Large Language Models in the post-pandemic era: ChatGPT and beyond. *J. Univ. Teach. Learn. Pract.* **2023**, *20*, 7. [CrossRef]
16. Zellers, R.; Holtzman, A.; Rashkin, H.; Bisk, Y.; Farhadi, A.; Roesner, F.; Choi, Y. Defending Against Neural Fake News. *Adv. Neural Inf. Process. Syst.* **2019**, *32*. Available online: https://arxiv.org/abs/1905.12616v3 (accessed on 5 September 2023).

17. Alamleh, H.; AlQahtani, A.A.S.; ElSaid, A. Distinguishing Human-Written and ChatGPT-Generated Text Using Machine Learning. In Proceedings of the 2023 Systems and Information Engineering Design Symposium (SIEDS), Charlottesville, VA, USA, 27–28 April 2023.
18. Das, M.; Kamalanathan, S.; Alphonse, P. A Comparative Study on TF-IDF Feature Weighting Method and Its Analysis Using Unstructured Dataset. *arXiv* **2023**, arXiv:2308.04037.
19. Jang, B.; Kim, I.; Kim, J.W. Word2vec convolutional neural networks for classification of news articles and tweets. *PLoS ONE* **2019**, *14*, e0220976. [CrossRef]
20. Haddi, E.; Liu, X.; Shi, Y. The Role of Text Pre-processing in Sentiment Analysis. *Procedia Comput. Sci.* **2013**, *17*, 26–32. [CrossRef]
21. Tweet-Preprocessor · PyPI. Available online: https://pypi.org/project/tweet-preprocessor/ (accessed on 20 March 2022).
22. Makrehchi, M.; Kamel, M.S. Extracting domain-specific stopwords for text classifiers. *Intell. Data Anal.* **2017**, *21*, 39–62. [CrossRef]
23. Kanerva, J.; Ginter, F.; Salakoski, T. Universal Lemmatizer: A sequence-to-sequence model for lemmatizing Universal Dependencies treebanks. *Nat. Lang. Eng.* **2020**, *27*, 545–574. [CrossRef]
24. Zhou, H. Research of Text Classification Based on TF-IDF and CNN-LSTM. *J. Physics* **2022**, *2171*, 012021. [CrossRef]
25. Cox, D.R. The Regression Analysis of Binary Sequences. *J. R. Stat. Soc. Ser. B* **1958**, *20*, 215–232. [CrossRef]
26. Fix, E.; Hodges, J.L., Jr. Discriminatory analysis. Nonparametric discrimination: Consistency properties. *Int. Stat. Rev./Rev. Int. Stat.* **1989**, *57*, 238–247. [CrossRef]
27. Altman, N.S. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *Am. Stat.* **1992**, *46*, 175–185. [CrossRef]
28. Han, T. Research on Chinese Patent Text Classification Based on SVM. In Proceedings of the 2nd International Conference on Mathematical Statistics and Economic Analysis, MSEA 2023, Nanjing, China, 26–28 May 2023.
29. Altin, F.G.; Budak, I.; Özcan, F. Predicting the amount of medical waste using kernel-based SVM and deep learning methods for a private hospital in Turkey. *Sustain. Chem. Pharm.* **2023**, *33*, 101060. [CrossRef]
30. Ho, T.K. Random decision forests. In Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, Canada, 14–16 August 1995.
31. Colab.Google. Available online: https://colab.google/ (accessed on 9 September 2023).
32. Zeini, H.A.; Al-Jeznawi, D.; Imran, H.; Bernardo, L.F.A.; Al-Khafaji, Z.; Ostrowski, K.A. Random Forest Algorithm for the Strength Prediction of Geopolymer Stabilized Clayey Soil. *Sustainability* **2023**, *15*, 1408. [CrossRef]
33. Valero-Carreras, D.; Alcaraz, J.; Landete, M. Comparing two SVM models through different metrics based on the confusion matrix. *Comput. Oper. Res.* **2023**, *152*, 106131. [CrossRef]
34. Aoumeur, N.E.; Li, Z.; Alshari, E.M.M. Improving the Polarity of Text through word2vec Embedding for Primary Classical Arabic Sentiment Analysis. *Neural Process. Lett.* **2023**, *55*, 2249–2264. [CrossRef]
35. Kale, A.S.; Pandya, V.; Di Troia, F.; Stamp, M. Malware classification with Word2Vec, HMM2Vec, BERT, and ELMo. *J. Comput. Virol. Hacking Tech.* **2023**, *19*, 1–16. [CrossRef]
36. Wei, L.; Wang, L.; Liu, F.; Qian, Z. Clustering Analysis of Wind Turbine Alarm Sequences Based on Domain Knowledge-Fused Word2vec. *Appl. Sci.* **2023**, *13*, 10114. [CrossRef]
37. Zhu, J.-J.; Ren, Z.J. The evolution of research in resources, conservation & recycling revealed by Word2vec-enhanced data mining. *Resour. Conserv. Recycl.* **2023**, *190*, 106876. [CrossRef]
38. Sharma, A.; Kumar, S. Ontology-based semantic retrieval of documents using Word2vec model. *Data Knowl. Eng.* **2023**, *144*, 102110. [CrossRef]