# Combating Fake News/Content with NLP-Based Detection for a Safer Online Environment

Lifu (Mario) Ma

Xianzhang (Jenny) Deng

Zhengjia (Vicky) Chen

Yifan Zhu

# Table of Contents

## 1. Business Objective

By March 2023, smartphone penetration in the United States reached approximately 85%[1]. Social media apps like Twitter and TikTok, along with email apps, have become the primary channels for disseminating information. Posting fake news or content on social media can lead to public panic, instill misguided values in impressionable teenagers, and deceive isolated elderly individuals who struggle to determine the authenticity of the content. As a result, content safety has increasingly become a critical concern for social media platforms. To help organizations maintain the safety and integrity of their app content and protect users' rights, we employ Natural Language Processing (NLP) related algorithms to develop models that automatically identify fake content. This enables organizations to intercept and prevent the spread of such content during the content review and publishing stages.

Our business model is focused on B2B, which aims to provide our cutting-edge "fake news/content detection" algorithm to organizations that develop social media apps, email apps, or engage in public relations. This innovative solution will assist these organizations in preserving the safety and integrity of their platforms by proactively identifying and curbing the dissemination of fake news and misleading content, ultimately safeguarding users' rights and interests.

## 2. Key Actionable Business Initiative

### 2.1 Metrics of Success

Before training models, we establish a function named "metrics" that comprises various performance measures, including confusion matrices. A confusion matrix is a table that describes the performance of a classification model by outlining the number of correct and incorrect predictions for each class. This matrix provides a comprehensive overview of the model's performance.

In our specific context, where we aim to distinguish between fake news (label = 0) and real news (label = 1), the true negative rate is of paramount importance. We primarily focus on specificity and accuracy as key performance metrics, given our objective is to accurately identify when news items are truly fake. Therefore, our goal is to optimize for both specificity and accuracy.

Specificity, also known as the True Negative Rate, measures the proportion of actual negatives (in our case, fake news) that are correctly identified. It is an especially valuable metric when the cost of a False Positive is high. The formula for specificity is TN / (TN + FP).

Accuracy, on the other hand, is a widely used metric that calculates the proportion of correct predictions made by the model, irrespective of the class. The formula for accuracy is (TP + TN) / (TP + TN + FP + FN). This overall measure of correct classification is also vital in our context, as we want the model to accurately distinguish both fake and real news.

### 2.2 Role of Analytics

Analytics plays a crucial role in assisting organizations that develop social media apps, email apps, or engage in public relations in combating fake news and misleading content. It provides value through early detection and prevention, real-time monitoring and alerting, risk assessment and mitigation, and continuous improvement.

Firstly, analytics enables early detection and prevention by analyzing large volumes of data using NLP techniques and machine learning algorithms. This allows organizations to proactively identify instances of fake news, taking swift action to prevent its dissemination and safeguard users' rights and interests.

Secondly, analytics provides real-time monitoring and alerting capabilities. By continuously scanning content and user interactions on the platform, analytics can quickly flag potentially fake news or misleading content. This empowers organizations to investigate, verify, and mitigate the impact of such content, ensuring that their platforms remain safe and trusted environments for users.

Thirdly, analytics offers risk assessment and mitigation strategies. By analyzing historical data, patterns, and user behavior, organizations can identify high-risk content sources, users, or topics. This enables them to prioritize efforts in monitoring, reviewing, and moderating content, effectively reducing the likelihood of fake news dissemination, and mitigating potential risks to users.

Additionally, analytics facilitates continuous improvement and adaptation of the fake news/content detection system. Through analyzing the performance of the algorithms, evaluating metrics, and incorporating user feedback, organizations can refine and enhance the accuracy and effectiveness of the system over time. This ensures that the system remains agile, staying ahead of evolving fake news techniques and maintaining a safe and reliable platform for users.

## 3. Analytics Methodology

### 3.1 Data Description

Data Source:

The dataset is sourced from Kaggle and can be accessed through this link. The project relies on existing observational data rather than designing and conducting experiments. The dataset was obtained from an external source and is not generated as part of the project served as a valuable resource to train and evaluate the NLP-based detection algorithm.

Data Structure:

The dataset comprises three columns: (1) the title of the news, (2) the text content of the news, and (3) the label indicating whether the news is real or fake. The dataset consists of a total of 6335 rows, each representing a news article.

Data Validation:

The dataset has undergone validation, ensuring that the news articles included are not fabricated and are available from credible sources on the internet. This validation process confirms the reliability and authenticity of the dataset for use in the project.

Target and Explanatory Variables/Features:

The target or outcome variable in the dataset is the label indicating whether a news article is real or fake. The explanatory variables or features are the title and text content of the news, which can be used to analyze patterns, extract meaningful insights, and build models for fake news detection.

Data Preprocessing and Data Exploration:

By doing steps listed below, the dataset is prepared for further analysis and model development, ensuring missing values are handled appropriately and the text data is tokenized and filtered to remove irrelevant elements.

Step 1 - *Handling Missing Values and EDA*: Remove any rows that have missing values in the "text" column. This ensures that the dataset is clean and does not contain incomplete data.

Step 2 - *Tokenizing and Stemming*: Utilize the NLTK library's English stop words to exclude common words that do not carry significant meaning. Then, tokenize the document, converting it into a list of strings representing individual words. Filter out any tokens that do not contain letters, such as numeric tokens or raw punctuation. This step helps in creating a more focused and meaningful representation of the text data.

## 3.2 Type of Analytics & Methodology

### 3.2.1 Analytical Methodology Overview

The goal of this project is to accurately predict whether a given news is real or fake, and we took two approaches to solve this problem. Our first approach relies on the frequency of words in the news. For instance, people commonly assume that fake news tends to be more emotionally exaggerated, as it strives to convince readers of its authenticity and validity. Therefore, we anticipate that words with strong targeting attributes will be frequently used in such news, and we want to detect these words using NLP techniques, such as Term Frequency – Inverse Document Frequency (TF-IDF) analysis. To understand how the frequency of words can be used to detect fake news, we also applied Recurrent Neural Network (RNN) using TF-IDF score as model inputs.

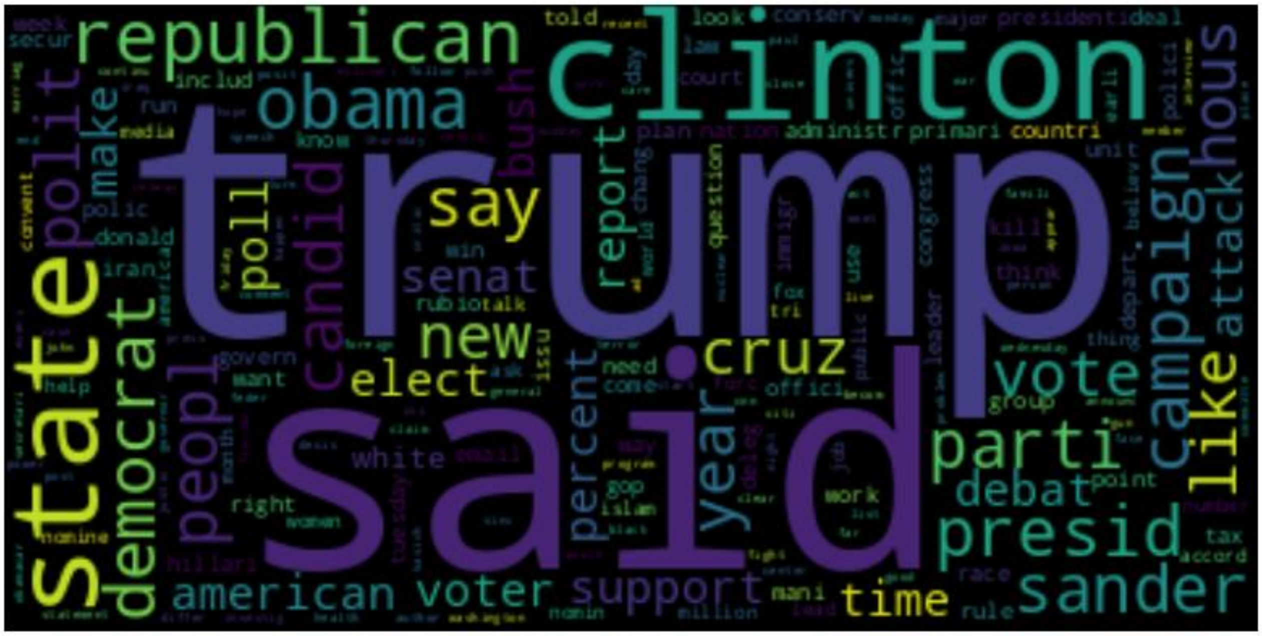
The second approach we took mainly focused on Content-Dependent features. We believe that fake news, either written by humans or machines, tends to be logically inconsistent, and such inconsistency can be captured after vectorizing words using embedding techniques like Word2Vec. This rationale underpinned our preprocessing steps, including tokenization and stemming. To understand how the semantic meaning of words would contribute to detect fake news, we subsequently pass the result of Word2Vec to a RNN model. By deploying these two distinct approaches and comparing model metrics, we aim to gain valuable insights that will lead to a comprehensive strategy to combat the proliferation of fake news in today’s digital landscape.

### 3.2.2 Descriptive Analytics – NLP

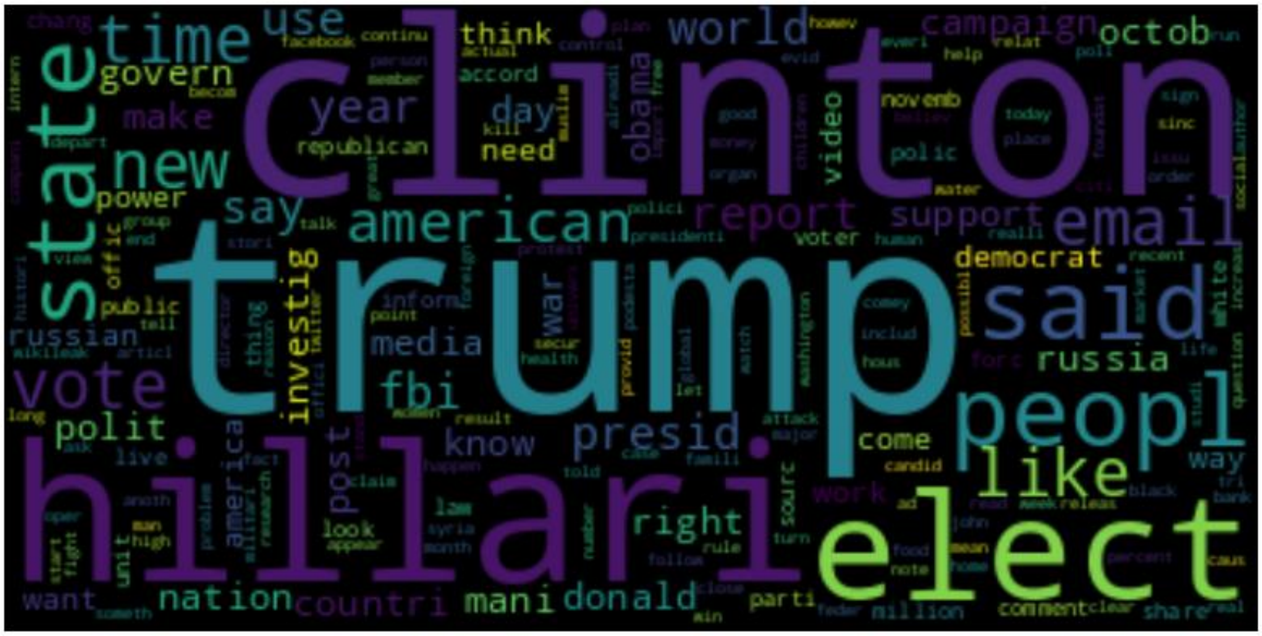
The first NLP related analytical method we used is Frequency-Inverse Document Frequency (TF-IDF)**.** Once the text data has been tokenized and filtered, we create a TF-IDF representation. TF-IDF is a numerical statistic that reflects the importance of a term in a document within a collection or corpus of documents. We use TfidfVectorizer offered by the Scikit-learn library to generate the TF-IDF representation. Finally, we calculate the TF-IDF scores for each token in the documents and create a matrix representation where each row corresponds to a document and each column corresponds to a token, with the cell value representing the TF-IDF score. And we also select the top 1000 words with the highest frequency.

In an effort to identify potential keywords predominantly used in the dissemination of fake news, we employed the use of a word cloud based on the accumulated TF-IDF score for the top 1000 words, independently analyzed across all documents for both authentic and fraudulent news. However, the results revealed a significant overlap between the two word clouds, implying a high degree of similarity in their lexical compositions. Consequently, our findings suggest that there is no single word or common phrase that can be broadly associated with the propagation of fake news. This prompts us to conclude that the assumption of universally applicable 'fake news' keywords is largely unfounded.

Word Cloud (Real News)



Word Cloud (Fake News)



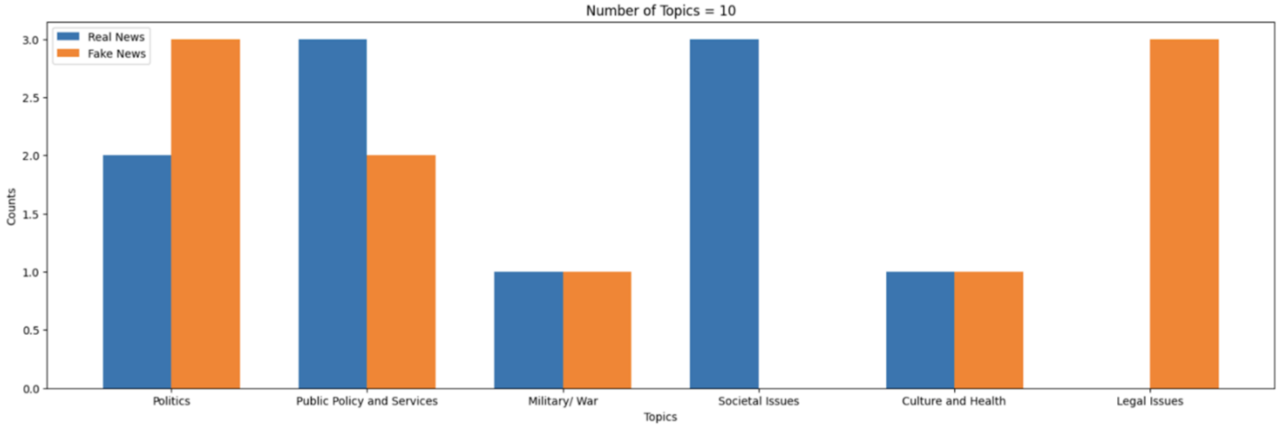
The second NLP related analytical method we used is Word2Vec**.** As mentioned in the Analytical Methodology Overview, our second approach is to predict fake news by understanding the semantic relationship between words. Thus, we decided to use Word2Vec for embedding. Word2Vec is a popular algorithm in the field of NLP that uses a shallow neural network to learn word embeddings from a text corpus. These embeddings represent words in a high-dimensional vector space where the semantic relationship between words can be captured. In our case, we set the embedding size to be equal to 100, which means that each word in the news will be represented as a 100-dimension vector, and each dimension captures some aspect of the word's meaning. Considering our objective to classify news as real or fake, we compute the average of all word vectors within a news article. This resultant single vector then serves as a comprehensive representation of the entire piece. Vectors for each news would then be used to train our model and make predictions.

### 3.2.3 Descriptive Analytics – K-means Algorithm & LDA

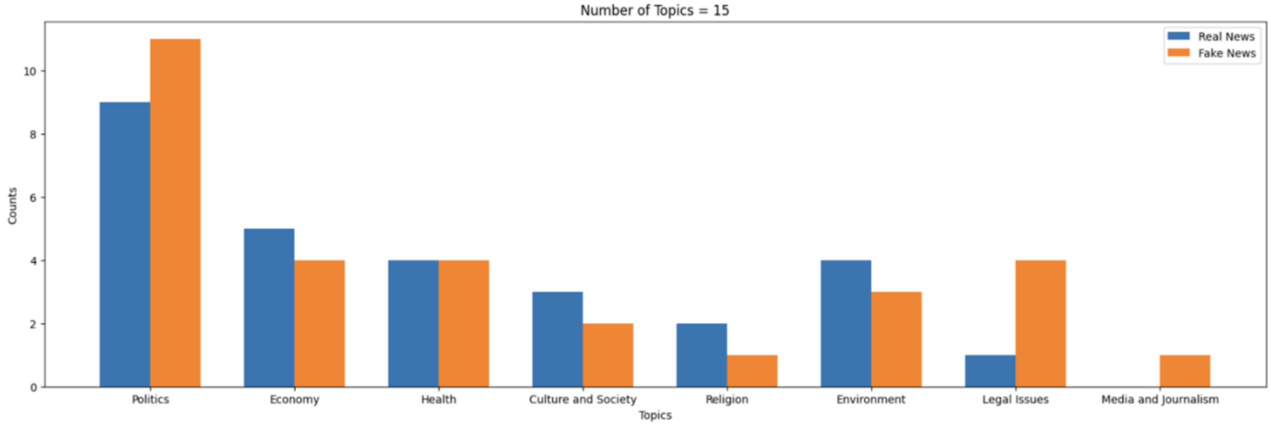
Before building a machine learning model that can accurately predict whether a given news is real or fake, there are two important premises that we want to check. Firstly, we want to ensure that a broad range of content is being utilized to train our model. So, the model can handle any types of text content and give accurate prediction based on that. Secondly, we want to ensure that the model's predictability isn't influenced by content differences between real and fake datasets, that is we want to control for new topics and reduce the potential bias caused by it.

Based on these two important premises, we introduced a powerful technique called LDA, which stands for Latent Dirichlet Allocation. LDA is a generative probabilistic model widely used for topic modeling. It assumes that every document is a mixture of certain number of topics, and every topic is a mixture of words. The logic behind LDA is kind similar to K-means algorithm, where you started by setting the number of topics, then you go through a random assignment until meet convergence. To ensure each topic has more concentrated information, we have also implemented a text processing technique known as part-of-speech (POS) tagging from the Natural Language Toolkit (NLTK) library. The POS tagging function takes a list of words as input and returns a list of tuples, where the first element of the tuple is the word, and the second element is the POS tag for that word. In our case, we specifically filtered for nouns as nouns typically possess the highest explanatory power. The four types of nouns we filtered are singular noun (NN), proper noun (NNP), plural noun (NNS), and plural proper noun (NNPS). To analyze the distribution of the number of topics in both real and fake news datasets, we constructed three comparative graphs. These were plotted side by side, each representing the situation where the number of topics was set to 10, 15, and 20, respectively.

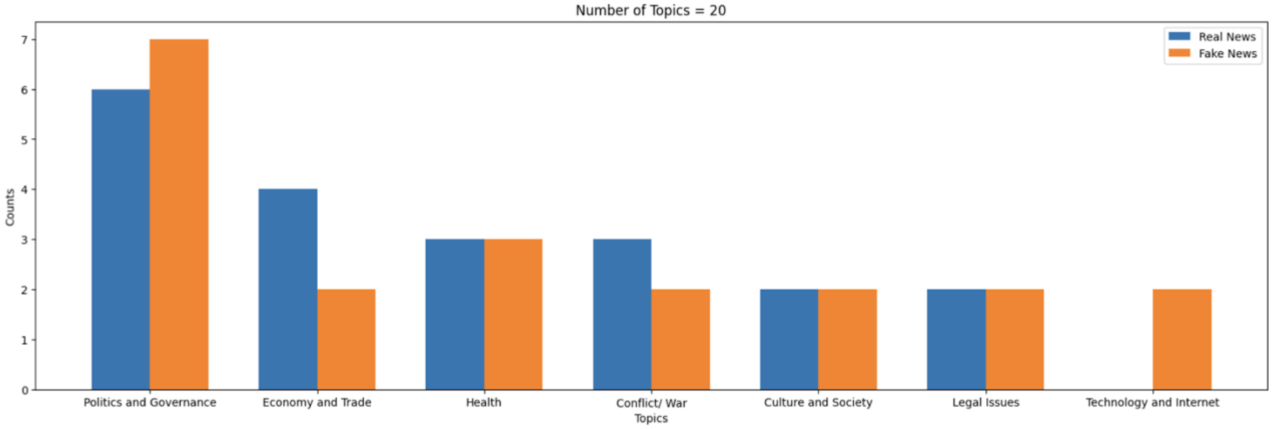
Number of Topics = 10



Number of Topics = 15



Number of Topics = 20



Based on three graphs showed above, we see that the distribution of the number of topics in both real and fake news datasets get more and more close. This indicate that the top distribution is similar for both real and fake news datasets. Therefore, we are confident to conclude that a diverse range of text content is used for training our model, and model's predictability isn't influenced by content differences between real and fake news datasets.

### 3.2.4 Predictive Analytics – RNN

*Deep Learning (RNN)-TF-IDF:* In this project, we leveraged the power of a recurrent neural network (RNN) with a Long Short-Term Memory (LSTM) architecture to address the task of predicting the authenticity of news articles. To accomplish this, we developed a user-defined function (UDF) with five essential hyperparameters.

The *input_size* hyperparameter determines the size of the input features for each time step, enabling the model to effectively process sequential information. Meanwhile, the *hidden_size* hyperparameter controls the number of units in the LSTM's hidden state, influencing the model's ability to capture and retain relevant information over time.

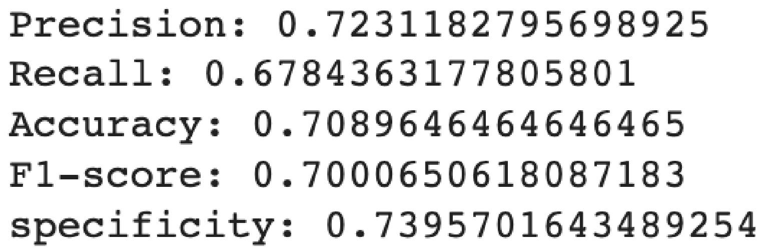
We also considered the *num_layers* hyperparameter, which determines the number of LSTM layers within the network. This choice impacts the model's depth and its capacity to learn intricate patterns and dependencies in the data.

To define the output of the model, we specified the *output_size* hyperparameter, which determines the size of the output features. This dimensionality defines the number of classes or the nature of the prediction task.

In order to prevent overfitting and enhance generalization, we incorporated a regularization technique called *dropout*. The dropout hyperparameter represents the dropout rate, which determines the proportion of neural network units that are randomly deactivated during training.

To identify the optimal combination of hyperparameter values, we employed the Optuna package. Optuna facilitated the hyperparameter optimization process, assisting us in finding the set of hyperparameters that yielded the best performance for the task at hand.

After following the aforementioned steps, we obtained the final model result using TF-IDF inputs. By leveraging the TF-IDF representation as input, our model was able to capture the significance and context of words in the news articles. This representation allowed us to effectively train and evaluate the model's performance in distinguishing between real and fake news. The result as following:



```
Precision: 0.7231182795698925
Recall: 0.6784363177805801
Accuracy: 0.7089646464646465
F1-score: 0.7000650618087183
specificity: 0.7395701643489254
```

As mentioned before, in this project, we focused on evaluating the model's performance using two important metrics: accuracy and specificity.

The accuracy metric, with a value of 0.70, represents the overall correctness of the model's predictions. It indicates the proportion of correctly classified articles out of the total number of articles in the dataset. An accuracy of 0.70 demonstrates a high level of accuracy in determining the authenticity of news articles.

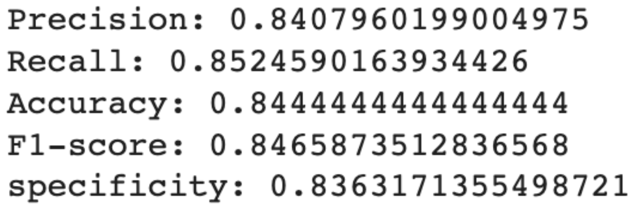
Additionally, we measured the specificity metric, which quantifies the model's ability to correctly identify fake news articles. With a specificity of 0.73, the model demonstrates a strong capability to correctly classify fake news articles as genuine.

*Deep Learning (RNN)-Word2Vec:* In addition to fitting the RNN model with Word2Vec inputs, we also explored the utilization of Word2Vec embeddings as inputs. By incorporating Word2Vec representations, our model captured the semantic meaning, contextual information, and relationships between words in the news articles. This enabled the model to grasp the nuances of semantic similarities and differences between words, enhancing its understanding of the textual data.

In contrast, TF-IDF places emphasis on the importance of terms within individual documents relative to the entire corpus. It quantifies the significance of terms based on their frequency and rarity across the corpus, allowing for effective feature representation in text-based tasks.

By comparing the predictive results between these two different NLP techniques, we aimed to assess their respective contributions and impacts on the model's performance. This analysis enabled us to evaluate how the semantic understanding and contextual representation provided by Word2Vec embeddings differ from the document-centric importance highlighted by TF-IDF.

we find that the accuracy metric, which yielded a value of 0.84, and the specificity metric obtained a value of 0.83, detailed results shown as following:

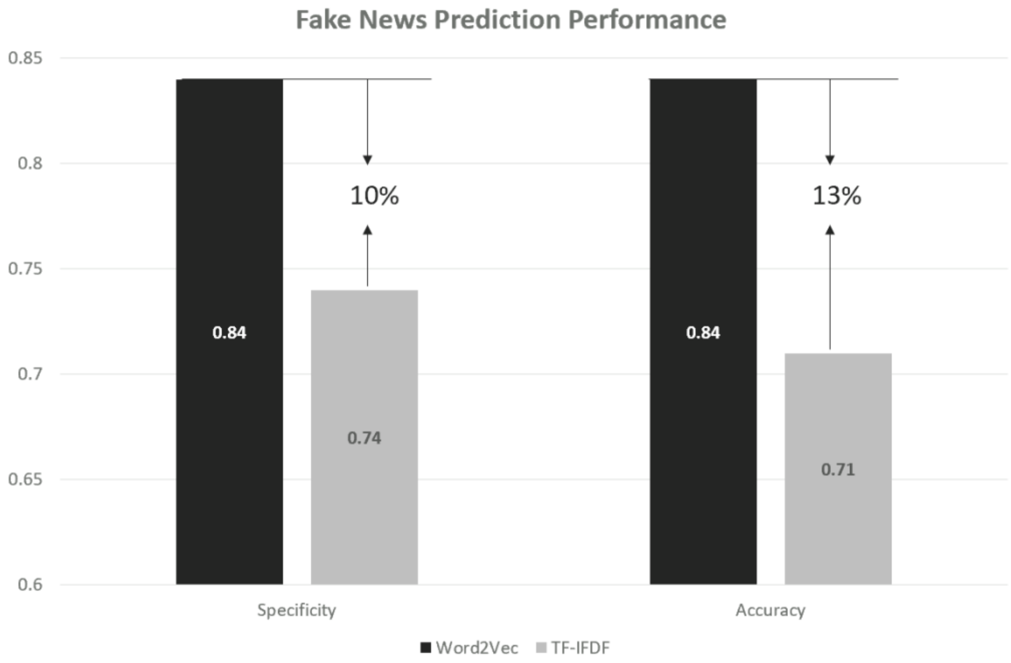


```
Precision: 0.8407960199004975
Recall: 0.8524590163934426
Accuracy: 0.8444444444444444
F1-score: 0.8465873512836568
specificity: 0.8363171355498721
```

## 4.Main Analytics Results

### 4.1 Finding from two predictive models (TF-IDF vs Word2Vec)

From our predictive analytics, we discovered that the specificity metric derived from 'Word2Vec' substantially exceeds that of 'TF-IDF', by approximately 10%. Additionally, the accuracy metric is enhanced by 13% (refer to Figure 1).



Upon comparing the performance between Word2Vec and TF-IDF, we observed that the Word2Vec embeddings achieved higher scores compared to TF-IDF. This outcome suggests that the semantic understanding and contextual information captured by Word2Vec embeddings have positively influenced the model's predictive capabilities, which could gain better performance than just focus on frequency-based representations.

### 4.2 Model Usage

Our business operates on a B2B model. We specialize in the development of fake news detection models which can be integrated into social media and email applications. This system scrutinizes the authenticity of the content prior to its publication. If the content is flagged as potentially false, users are alerted that their content requires review before it can be posted, at which point it undergoes manual verification. For new media companies, this model serves as an invaluable tool.

It significantly reduces operational costs related to content safety, mitigates legal risks, and helps maintain their corporate reputation.

### 4.3 Future Scope

At present, we have developed robust deep learning models that capably discern between real and fake news. Nevertheless, there's a wealth of future analytics to be explored. For instance, we could enrich our dataset by introducing additional columns. Currently, our data is limited to the title, content, and label of the articles. A beneficial addition could be a column representing the author of the article. Given that these articles are readily available online, we have the opportunity to employ web-scraping tools such as Selenium and BeautifulSoup. These tools can automatically extract author information by searching for the article titles online. Following this, we could employ innovative methods like LIWC (Linguistic Inquiry and Word Count), a technique used to gauge the frequency of diverse lexical or language features within a text. These features may include personal pronouns, emotion words, social words, cognitive process words, time and space words, among others. By applying this method, we can gain a deeper understanding of the distinct differences between fake and real news authors, and probably also helping enhance the reliability of our models.

# Reference

[1] Zippia. "25+ Incredible US Smartphone Industry Statistics [2023]: How Many Americans Have Smartphoness" Zippia.com. Mar. 2, 2023, https://www.zippia.com/advice/us-smartphone-industry-statistics/