National College of Ireland

National College of Ireland

Project Submission Sheet

Student Name:	Olamide Abioro
Student ID:	23428902
Programme:	MSCDAD_C **Year:** 2025
Module:	Analytics Programming and Data Visualisation
Lecturer:	Furqan Rustam
Submission Due Date:	25th April 2025
Project Title:	Fake News Detection Through Textual Feature Analysis
Word Count:	

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
<u>ALL</u> internet material must be referenced in the references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	April 24, 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. Projects should be submitted to your Programme Coordinator.
3. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**
5. All projects must be submitted and passed in order to successfully complete the year. **Any project/assignment not submitted will be marked as a fail.**

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Fake News Detection Through Textual Feature Analysis

Liton Nath
School of Computing (MSCDAD_C)
National College of Ireland
Dublin, Ireland
x23402661@student.ncirl.ie

Olamide Abioro
School of Computing (MSCDAD_C)
National College of Ireland
Dublin, Ireland
x23428902@student.ncirl.ie

Muhammad Osama Hassan Khan
School of Computing (MSCDAD_C)
National College of Ireland
Dublin, Ireland
x24137782@student.ncirl.ie

***Abstract*—In this age of rapid information technology development, distinguishing fact from fiction has become increasingly difficult. This project examines fake news detection using multiple datasets, including the CIC TruthSeeker 2023 and Politifact, as well as one obtained from an API and another from Kaggle. An ETL pipeline was established to extract, transform, and load data efficiently for analysis and visualization. The approach focuses on text processing of structured and unstructured data using TF-IDF vectorization and Random Forest classification to achieve high predictive accuracy. Additional analysis includes sentiment profiling, keyword frequency, subjectivity comparison, and graphical depiction of statement characteristics. Results, with a high accuracy of 99.84%, affirm the effectiveness of machine learning and NLP in identifying misinformation and offer insights into the emotional and linguistic patterns distinguishing real from fake news. Technical challenges and data transformation processes are also discussed.**

***Index Terms*—Fake News Detection, Text Analysis, ETL Pipeline, Natural Language Processing (NLP), Data Preprocessing, API Dataset, Kaggle Dataset, CIC TruthSeeker 2023 dataset, Politifact dataset, Feature Extraction, Data Visualization, Misinformation Detection.**

I. Introduction

The spread of misinformation on online platforms has become a major obstacle to journalism, policy-making, and the general working of democracy. Surprisingly, these fake stories spread faster than true news, raising alarms about the well-being of the society. Recently, in a society where information is a driving force, the fast spread of disinformation has become a concern, affecting social, political, and economic processes. This project was inspired by the need to understand and combat this issue through a data-driven approach.

Our goal was to select datasets that are practical and can benefit the solution of real-world problems. We employed three datasets: one retrieved from The Guardian News public API, which provided access to real news stories from a proven source; another from the open-source platform Kaggle and politifact, with labeled real and fake news entries; and the CIC TruthSeeker 2023 dataset, containing over 134,000 news items marked as either true or false. By combining these datasets, we carried out a robust comparative study and developed an awareness of critical textual features that distinguish fake news from credible journalism.

To achieve this, we analyzed sentiment, keyword choice, and linguistic structure using Natural Language Processing (NLP) techniques. The study is supported by a customized ETL (Extract, Transform, Load) pipeline for automated data ingestion, cleaning, transformation, and visualization. This project not only investigates the structural and emotional patterns of misinformation but also lays the groundwork for scalable solutions that can help platforms and users flag false information in real time.

II. Related Work

Fake news detection has been given significant attention, with research aiming at linguistic, psycholinguistic, and readability attributes to distinguish between false and factual information.

A. Linguistic & Psycholinguistic Features:

Prior work (Rubin et al., 2016; Ott et al., 2011) suggests that fraudulent text deviates in word usage and tone. We use TF-IDF encoded n-grams, patterns in punctuation, and LIWC-based psycholinguistic features (e.g., emotional tone, cognitive and social processes) to detect these nuances [1].

B. Readability & Syntax:

Deceptive news varies significantly in readability. Tests like Flesch-Kincaid and Gunning Fog allow us to test for clarity of content. Syntactic features based on CFG trees (Klein & Manning, 2003) also highlight structural irregularities present in misleading articles (Feng et al., 2012) [1].

C. ML and NLP Approaches to Fake News Detection

Several researchers have explored how machine learning and natural language processing can be used to detect fake news. Ahmed et al. [2] presented a comprehensive survey of machine learning techniques for fake news detection. They outlined the advantages and disadvantages of different algorithms, including Naive Bayes, Support Vector Machines, and deep learning models. Their paper illustrated how effective these tools are, but they also said that a model's performance relies a lot on the quality and balance of the dataset.

D. Multimedia-Based Fake News Detection

Sharma et al. [3] focused on detecting fake news using not just text but also multimedia content. This is useful in cases where images or videos are manipulated to spread misinformation. However, their work is limited in scope when it comes to text-only detection, which is the main focus of this project. Their research did help shape our understanding of how different types of content can contribute to misinformation, even though our study is limited to textual data.

E. Social Context in Fake News Detection

Shu et al. [4] looked at the detection of fake news using data mining. They did not only look at what is being reported by the news but also at the social context, i.e., how users interact with it on platforms like Twitter and Facebook. This is a strong method since it collects more signals, especially those relating to the spread of misinformation. However, our project focuses only on the textual content, due to dataset constraints.

F. Theoretical Foundations of Fake News Detection

Zhou and Zafarani [5] conducted an extensive review of the entire fake news detection pipeline, covering datasets, algorithms, and psychological theories behind misinformation. Their work helped us understand the strengths and weaknesses of various detection techniques. While their paper is more theoretical and broad, it offered a solid foundation on which to build our practical implementation.

In summary, the previous works provided useful frameworks and highlighted potential challenges. But many of them either focused too much on the algorithm or ignored the value of analysis and visualization. Our work tries to address this gap by combining reliable classification with meaningful analysis and clear visual presentation.

Evaluation of Our Work

This work adds to existing fake news detection research by:

- Training and testing models using live and historical news data from public APIs like *The Guardian* and Kaggle datasets.
- Extracting multi-level linguistic, psycholinguistic, and syntactic features like TF-IDF n-grams, LIWC-based categories, readability scores, and CFG-based syntactic rules.
- Employing scalable data processing and storage utilities (e.g., MongoDB, PostgreSQL) and Python-based pipelines to enable efficient analysis and reproducibility [1].

Compared to previous research that has been inclined to focus on fixed data or a single feature set, our method offers a complete linguistic feature set and incorporates live news streams, providing actionable insights regarding misinformation patterns across sources and time horizons. While previous works provided useful frameworks and highlighted potential challenges, many of them either focused too heavily on algorithmic performance or overlooked the importance of analysis and visualization. Our work addresses this gap by combining reliable classification with meaningful analysis and clear visual presentation.

III. METHODOLOGY

This study follows a contemporary ETL (Extract, Transform, Load) approach, using two separate databases to distinguish between unprocessed and processed data. The goal was to build a structured pipeline that supports automation and enables efficient analysis. Critical visualizations were also performed to extract meaningful insights from the datasets, which support decision-making and serve as preparation for applying machine learning models

A. Dataset Overview

The datasets used in this study were gathered from both public APIs and open-source repositories. Specifically:

- *The Guardian News API*: It was utilized to collect live and historical news articles for real-time analysis of misinformation patterns.

	category	title	weburl	pub_date	body	label
364	science	Gene-editing crops to be colourful could aid w...	https://www.theguardian.com/science/2024/apr/1...	2024-04-17 20:30:05	Genetically engineering crops to be colourful ...	1
1224	NaN	Super Tuesday : The Politics Daily - The Atlantic	https://www.theatlantic.com/politics/archive/2...	NaT	white socialism might have once viscerally con...	0
1696	NaN	Corona ... Report sparks controversy over "con...	https://www.albiddu.net/world/36029883.html	NaT	the controversy over the origin of the new cor...	1
373	science	Private moon lander lifts off aiming for first-...	https://www.theguardian.com/science/2024/feb/1...	2024-02-15 13:49:22	A solar-powered lunar lander designed by a for...	1
1408	NaN	Japan slides towards recession after typhoon	https://www.thetimes.co.uk/edition/business/ja...	NaT	japan slides towards recession after typhoon j...	0
142	world	India summons envoy after US criticises Delhi ...	https://www.theguardian.com/world/2024/mar/20/...	2024-03-29 02:27:02	The chief minister of Delhi has been remanded ...	1
851	Education	University of Buckingham's Vice-Chancellor Cle...	https://www.fakenews.com/university-buckingham...	2021-10-15 17:30:00	In a shocking turn of events, the University o...	0
761	NaN	Coronavirus : Shoppers bulk - buy hand sanitis...	https://www.thetimes.co.uk/article/online-shop...	NaT	coronavirus shoppers bulk buy hand sanitiser b...	0
922	NaN	Australian composer Brett Dean hospitalized wi...	https://apnews.com/23603c42a404a9dd8a04783cc72...	NaT	australian composer brett dean hospitalized wi...	0
1467	NaN	Notable & Quotable : Wisconsin	https://www.wsj.com/articles/notable-quotable-...	NaT	opinion commentary notable quotable notable-q...	0

Fig. 1. Dataset - Guardian Live News

This data set is made up of news articles collected from various sources, including real-time data from The Guardian API. Each record in the data set contains metadata such as the article category, title, publication URL, publication date, article body, and a label for whether the news is real (0) or not (1).

The data has a broad amount of content under categories such as science, politics, world, and education covering both real news sources and inaccurate content. Lacking values within some fields such as category or publication date simulate the variability and noise that naturally occur in datasets in the real world.

- *Kaggle*: It provided labeled datasets containing fake and real news articles, which served as the ground truth for model training and validation.

	category	title	weburl	pub_date	body	label
364	science	Gene-editing crops to be colourful could aid w...	https://www.theguardian.com/science/2024/apr/1...	2024-04-17 20:30:05	Genetically engineering crops to be colourful ...	1
1224	NaN	Super Tuesday : The Politics Daily - The Atlantic	https://www.theatlantic.com/politics/archive/2...	NaT	white socialism might have once viscerally con...	0
1696	NaN	Corona ... Report sparks controversy over "con...	https://www.albiddu.net/world/36029883.html	NaT	the controversy over the origin of the new cor...	1
373	science	Private moon lander lifts off aiming for first-...	https://www.theguardian.com/science/2024/feb/1...	2024-02-15 13:49:22	A solar-powered lunar lander designed by a for...	1
1408	NaN	Japan slides towards recession after typhoon	https://www.thetimes.co.uk/edition/business/ja...	NaT	japan slides towards recession after typhoon j...	0
142	world	India summons envoy after US criticises Delhi ...	https://www.theguardian.com/world/2024/mar/20/...	2024-03-29 02:27:02	The chief minister of Delhi has been remanded ...	1
851	Education	University of Buckingham's Vice-Chancellor Cle...	https://www.fakenews.com/university-buckingham...	2021-10-15 17:30:00	In a shocking turn of events, the University o...	0
761	NaN	Coronavirus : Shoppers bulk - buy hand sanitis...	https://www.thetimes.co.uk/article/online-shop...	NaT	coronavirus shoppers bulk buy hand sanitiser b...	0
922	NaN	Australian composer Brett Dean hospitalized wi...	https://apnews.com/23603c42a404a9dd8a04783cc72...	NaT	australian composer brett dean hospitalized wi...	0
1467	NaN	Notable & Quotable : Wisconsin	https://www.wsj.com/articles/notable-quotable-...	NaT	opinion commentary notable quotable notable-q...	0

Fig. 2. Illustrative example of news data pipeline

- *TruthSeeker Dataset:*: This dataset was sourced from the CIC TruthSeeker 2023 collection and contains over 134,000 records. Each entry includes a news statement, associated tweet, author, manually extracted keywords, and a binary label indicating whether the news is real (1) or fake (0). This structured dataset was chosen for its

size, relevance, and well-labeled format, which is ideal for text classification and sentiment analysis tasks.

- *Politifact Dataset:*: We used two CSV files sourced from Politifact: one containing fake news headlines and another with real ones. Each record includes: unique ID, URL pointing to the original article, the headline or statement, related tweet IDs (in some cases). Although minimal in structure, this dataset offers enough to explore linguistic trends and build a strong classifier.

B. Description about data processing;

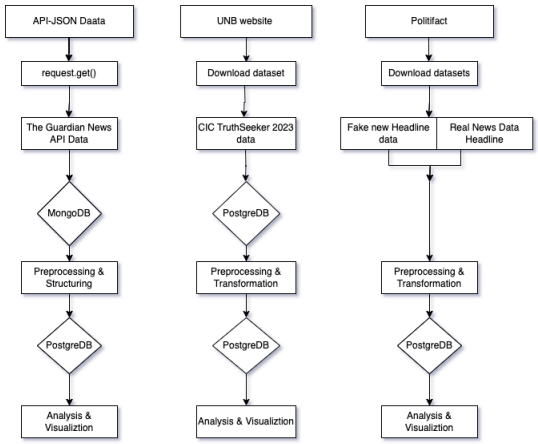

Fig. 3. Data pipeline

In this study, the ETL (Extract, Transform, Load) methodology was employed to manage the collection, preprocessing, transformation, and storage of fake news data. Initially, raw news articles in JSON format were extracted from sources such as The Guardian News API and Kaggle and loaded into MongoDB. This NoSQL database was selected for its flexibility in handling dynamic data structures, ideal for articles with varying metadata and structure. MongoDB also offers scalability and real-time ingestion capabilities due to its auto-update support.

In another instance, a raw CSV dataset was also read into Python for preprocessing. Preprocessing steps across both data sources involved identifying and handling null values, missing publication dates, and inconsistencies in article categories. Where necessary, mean imputation was applied to fill missing numerical values. For the text data, duplicate entries were removed, text was converted to lowercase, punctuation was eliminated, and stopwords were filtered out using NLTK. Tokenization was also applied to prepare the data for analysis.

Once transformed, the cleaned datasets were stored in PostgreSQL, a structured relational database chosen for its reliability, ACID compliance, and seamless Python integration via the psycopg2 library. PostgreSQL's support for client-server architecture also makes it suitable for team-based and networked environments. In addition, TF-IDF (Term Frequency-Inverse Document Frequency) vectorization was applied to convert the cleaned text into numerical form for machine learning analysis. The final vectorized data was again stored in PostgreSQL, completing the ETL pipeline.

The entire ETL process was designed with automation in mind using Dragster, ensuring that updates to the dataset could flow from ingestion to a model-ready state smoothly and without manual intervention.

IV. RESULTS AND EVALUATIONS

1) Number of words in Fake vs. Real News:: A detailed analysis was carried out to identify trends in word length of news articles categorized as true or false. The KDE plot (Fig. 4) indicates a comparison of word length distributions for both categories. The graph distinguishes between two groups: true news (blue curve) and false news (red curve). The steep peak of the red curve representing false news between 100–150 words indicates that false news stories are significantly shorter and more uniformly sized. The blue curve for real news is more spread out, with word counts peaking between 300–400 and extending up to over 1000 words in some cases. This wide spread suggests that real news stories are generally longer and more detailed. The visual theme clearly focuses on a tendency by manufactured news to be short, perhaps for easier reading and going viral, but proper news is well covered. The disparity forms the foundation for the use of word count as a critical factor in news article credibility determination.

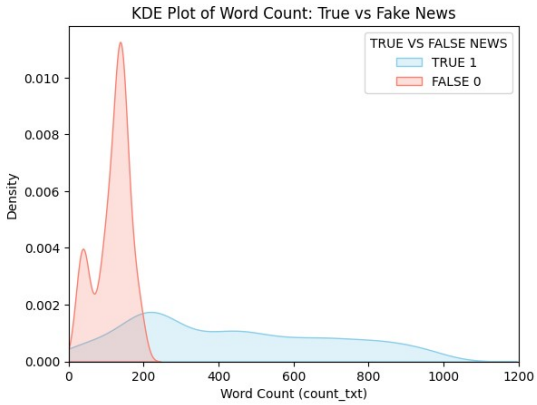

Fig. 4. Data pipeline

2) Word Frequency Patterns in Fake vs. Real News: The first word cloud (Fig. 5) consists of fake news articles. The prominent words are *"said," "will," "china," "corona," "virus,"* and *"india."* Interestingly, this cloud includes a mix of political, global health-related, and even sports terms such as *"ball," "wicket,"* and *"game,"* indicating a broader but less targeted engagement. Words such as *"government," "death," "plan,"* and *"world"* are also prevalent, but the presence of entertainment or sports terminology can suggest attempts to merge content for emotional or viral appeal at the expense of factual content. The cloud also suggests repeated and vague keywords, which can be employed to mislead or sensationalize.

In contrast, the second word cloud (Fig. 6) depicts real news stories. More precise and fact-oriented keywords are

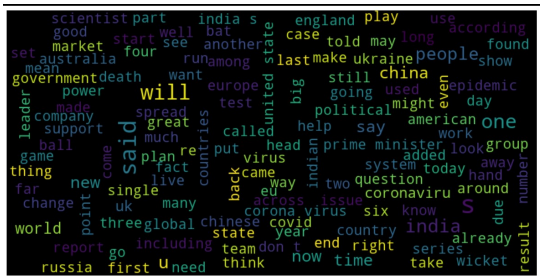

Fig. 5. Word cloud of fake news articles

used, e.g., *"coronavirus," "covid," "people," "government," "hospital," "testing," "death," "patient,"* and *"pandemic."* Such terms reflect an evidence-based news style with specific content, especially in terms of public policy and public health. The prevalence of the words *"public health," "confirmed," "symptom," "testing," "economy,"* and *"quarantine"* shows more concern with informing readers about known facts and providing detailed insights into current events.

Fig. 6. Word cloud of real news articles

This contrast shows how false news relies on general or emotionally charged language, while true news employs precise, topic-specific vocabulary. These observations can inform the development of text classification models and offer a linguistic perspective for identifying misinformation.

3) Psycholinguistic Features: The LIWC (Linguistic Inquiry and Word Count) lexicon categorizes words into psychological, linguistic, and emotional dimensions. It was used in this study to identify the most prominent features in both fake and real news content [1]. The top 20 most frequent features were extracted and compared across the two categories, as illustrated in Figures 7 and 15.

a) Analysis.: The most dominant feature in both fake and real news is *health*, suggesting that health-related topics are widely covered, regardless of the article's authenticity. However, fake news demonstrates a significantly higher frequency for categories such as *medical_emergency*, *business*, *government*, and *negative_emotion*, indicating a possible tendency toward fear-inducing or emotionally charged content [1].

On the other hand, real news articles show a more balanced distribution, with high frequencies for *technology*, *programming*, *internet*, and *science*, which reflect an informative and factual tone. The presence of *communication*, *social_media*,

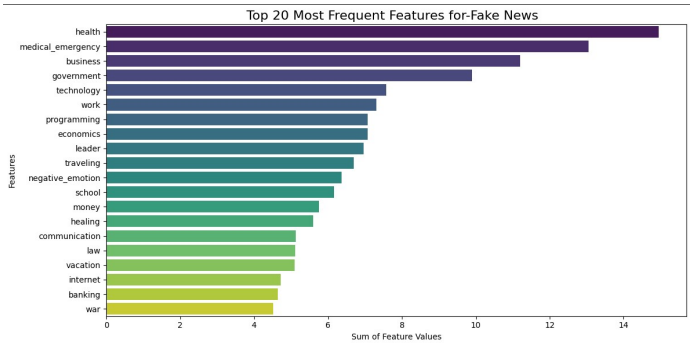

Fig. 7. Top 20 Most Frequent Topics for Fake News

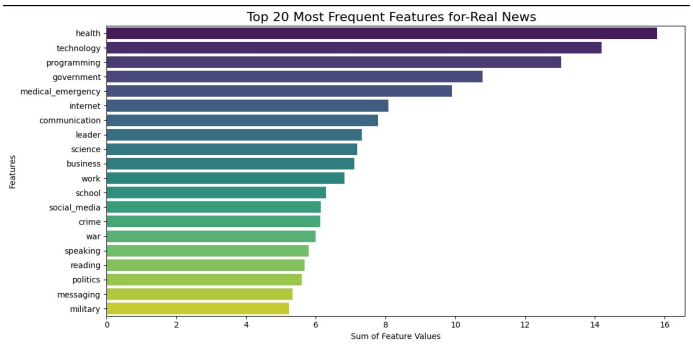

Fig. 8. Top 20 Most Frequent Topics for Real News

and *politics* in real news also points to a stronger emphasis on contextual and socio-political discourse [1].

This divergence in feature prominence highlights the utility of psycholinguistic attributes in distinguishing between fake and real news. The LIWC categories provide insight into the psychological framing of content, which can be valuable for automated detection models.

4) Keyword Frequency Comparison in Real vs Fake News: The bar chart (Figure 9) shows the top 10 keywords in real and fake news. It clearly highlights the distinct word usage patterns between both categories. Real news statements commonly used terms that were linked with official institutions and current events, while fake news often included more emotionally charged or controversial terms. This visual difference in word frequency gave us useful insights into the kinds of language used in fake vs real news.

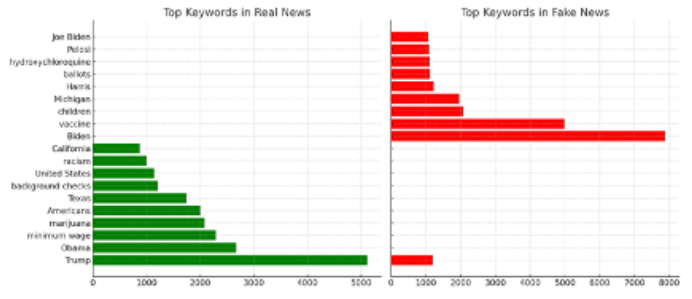

Fig. 9. Top 10 word Usage Patterns in Real and Fake News

5) Sentiment Polarity Distribution in Real vs Fake News: The sentiment polarity distribution (Figure 10) gives us a deeper understanding of the emotional tone present in fake and real news statements. We found that fake news had a broader range of polarity values, including more extreme negative and positive scores. Real news statements, on the other hand, were often more neutral. This pattern supports the idea that fake news relies more on emotionally loaded language to grab attention [6].

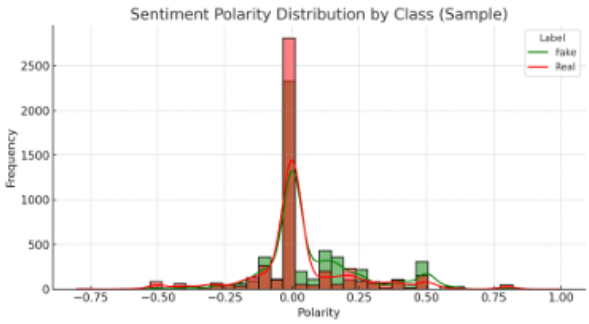

Fig. 10. Sentiment Polarity Distribution by Class (Sample)

6) Subjectivity Distribution in Fake vs Real News: The subjectivity distribution (Figure 11) shows how opinionated or factual the statements were. Fake news articles scored higher in subjectivity, indicating a greater frequency of personal opinions and emotive language. Genuine news articles were more objective, focusing on facts rather than subjective opinions [7]. This difference can help in distinguishing between fake and real content.

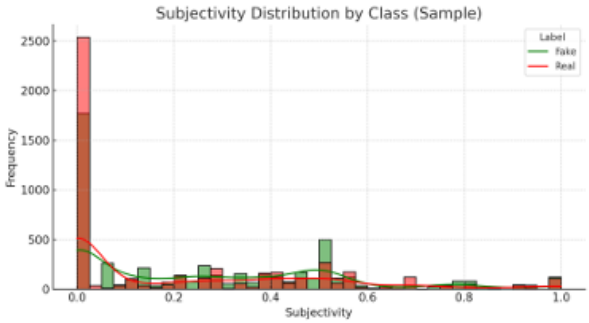

Fig. 11. Subjectivity Distribution by Class (Sample)

7) Statement Length Comparison in News Content: The boxplot (Figure 12) shows the length of statements for both fake and real news. As observed, fake news statements are usually shorter, with lower word counts, while real news tends to be longer and more detailed. This suggests that fake news uses shorter and possibly more sensational terms in an attempt to attract attention right away, while real news can provide more information and depth [6].

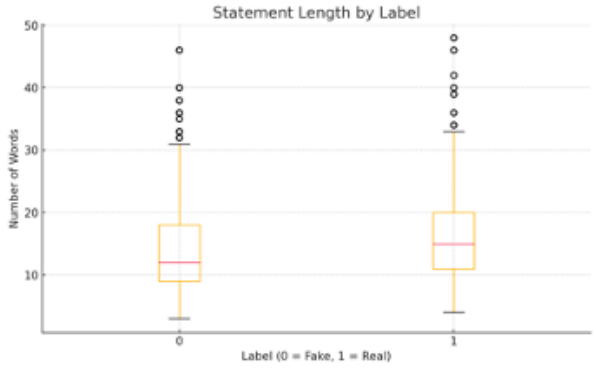

Fig. 12. Statement Length by Label

Readability Features

Readability features(Fig. 13) are crucial in distinguishing between fake and real news. These features evaluate the complexity, clarity, and structure of written text using standard readability metrics. Fake news often prioritizes emotional manipulation over logical reasoning and may exhibit linguistic patterns that differ from legitimate journalism [1].

Specifically, fake news articles may employ either overly simplistic language to appeal to a broader audience or overly complex wording to create an illusion of credibility. These manipulative strategies can mislead readers by reducing transparency or overwhelming them with jargon.

By analyzing readability metrics such as Flesch Reading Ease, Flesch-Kincaid Grade Level, Gunning Fog Index, and SMOG Index, we can assess whether a news article is more likely to be fake or real. These scores offer insight into how digestible the content is and reveal stylistic cues often associated with disinformation.

The incorporation of these features into a classification model enhances its ability to detect subtle cues in writing style that go beyond content, making them a valuable addition to fake news detection systems.

Metric	Indicates...
Flesch-Kincaid Grade Level	Years of education needed to understand the text
Flesch Reading Ease	Higher = easier to read
Gunning Fog Index	Estimate of years of formal education required
ARI (Automated Readability)	Based on characters per word and words per sentence
Character / Word Counts	General size and complexity indicators
Long/Complex Words	Words with >6 letters or >3 syllables

Fig. 13. Readability Features

A. Interpretation:

The distribution of the Automated Readability Index (Fig. 14) for this collection of made-up news articles shows a bias towards simpler language. The bias towards lower ARI

scores suggests that a lot of the fake news is composed to be easily digestible, perhaps engaging a broad audience with straightforward messaging. The presence of some articles with higher ARI scores indicates that not all fake news trends this way, but overall the trend is in the direction of lower reading levels. The threshold at ARI = 10 once again illustrates the pervasiveness of easily readable fake news in this data set. This result might have implications in understanding how fake news is disseminated and consumed.

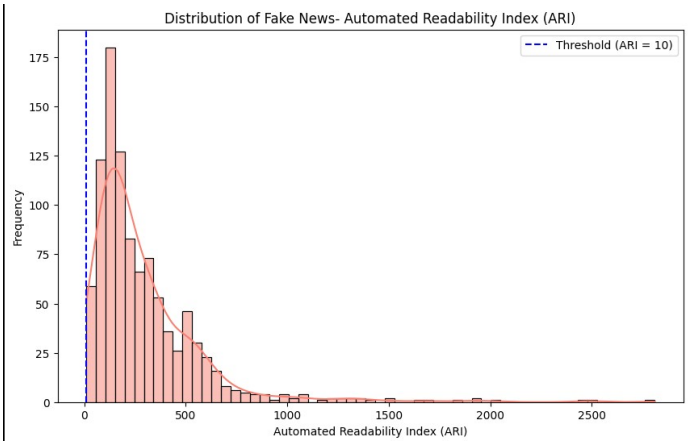

Fig. 14. Fake News readbility

Whereas real news also had a high concentration of articles with lower ARI scores (indicating accessibility issues), it also had a longer and higher tail in the direction of greater ARI values. This indicates that there were more articles with higher-level vocabulary, longer sentence lengths, and perhaps greater levels of specialist terms. This can be accounted for by the expertise of detailed reporting, discussing issues of complexity, or articles for a more informed or specialist audience.

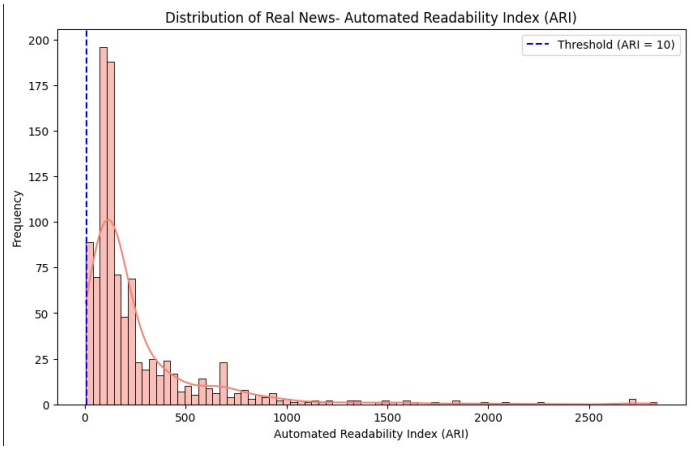

Fig. 15. Fake News readbility

B. Conclusions and Future Work

This project showed that machine learning algorithms, combined with simple natural language processing techniques like TF-IDF, are extremely helpful in distinguishing fake news from genuine news. The Random Forest classifier used here was shown to be nearly perfectly accurate, and the supporting visual analysis gave additional insight into how the fake and real news articles use different language, sentiment, and authors.

The visual findings, such as sentiment polarity, subjectivity, and keyword frequencies, are in support of our findings that there is more emotive and subjective content in fake news and more neutral and objective content in real news. The identification of common authors and statement length differences also improved our understanding of the dataset.

One of the benefits of this project was having the ability to work with a structured and semi-structured dataset. This allowed us to work with a more comprehensive data pipeline, encompassing different sources of data formats, and illustrating the flexibility of our ETL process. It further facilitated more general insights comparing results between structured and semi-structured news data sources.

There are lots of scopes to find more work in these areas, especially because false news (Fig. 15) remains a prominent problem currently. Machine learning and deep learning techniques can be effectively used for the prediction and detection of false news using the features I have proposed. Such an approach has the ability to significantly boost detection efficiency through utilizing several information sources, instead of relying on a single factor. This way, the system can provide more solid results and possess stronger grounds for the claims that it evaluates

REFERENCES

[1] V. Pérez-Rosas, B. Kleinberg, A. Lefevre, and R. Mihalcea, "Automatic Detection of Fake News," in *Proc. 27th Int. Conf. Comput. Linguist. (COLING)*, Santa Fe, NM, USA, 2018, pp. 3391–3401. [Online]. Available: https://aclanthology.org/C18-1287/. [Accessed: Apr. 3, 2025].

[2] N. Ahmed, D. Traore, and E. Saidi, "Detecting Fake News Using Machine Learning: A Survey," Procedia Computer Science, vol. 170, pp. 385–392, 2020. doi: 10.1016/j.procs.2020.03.071

[3] A. Sharma, V. Sharma, and M. S. Atrey, "Detecting Fake News Using Multimedia Content," Multimedia Tools and Applications, vol. 79, no. 3–4, pp. 3129–3150, Jan. 2020. doi: 10.1007/s11042-019-08142-1

[4] H. Shu, S. Wang, and D. Liu, "Fake News Detection on Social Media: A Data Mining Perspective," ACM SIGKDD Explorations Newsletter, vol. 19, no. 1, pp. 22–36, 2017. doi: 10.1145/3137597.3137600

[5] Y. Zhou and R. Zafarani, "A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities," ACM Computing Surveys, vol. 53, no. 5, pp. 1–40, 2020. doi: 10.1145/3395046

[6] M. Horne and S. Adali, "This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News," in Proceedings of the 2nd International Workshop on News and Public Opinion, 2017.

[7] R. Oshikawa, J. Qian, and W. Y. Wang, "A Survey on Natural Language Processing for Fake News Detection," in Proceedings of the 12th Language Resources and Evaluation Conference (LREC), 2020, pp. 6086–6093.