Fake News Detection Report

(Dilip Chauhan and Gaurav Sahoo)

Objective:

The objective of this assignment is to develop a Semantic Classification model. We are using Word2Vec method to extract the semantic relations from the text and developing a basic understanding of how to train supervised models to categorise text based on its meaning, rather than just syntax. We will explore how this technique is used in situations where understanding textual meaning plays a critical role in making accurate and efficient decisions.

Business Objective:

The spread of fake news has become a significant challenge in today's digital world. With the massive volume of news articles published daily, I find it increasingly difficult to distinguish between credible and misleading information. This has motivated me to work on building systems that can automatically classify news articles as true or fake, with the aim of reducing misinformation and supporting public trust.

In this assignment, we developed a **Semantic Classification model** that leverages the **Word2Vec** method to capture recurring patterns and semantic themes in news articles. Using supervised learning models, our goal was to build a system that enables us and others to effectively identify whether a given news article is fake or true.

Pipelines that are performed

We have performed the following tasks to complete the assignment:

1. Data Preparation
2. Text Preprocessing
3. Train Validation Split
4. EDA on Training Data
5. EDA on Validation Data
6. Feature Extraction
7. Model Training and Evaluation

1. Data Preparation

As part of data preparation, we have performed below steps

- Loaded True.csv and Fake.csv using pd.read_csv().
- Used df.head(), df.info() to understand dataset structure.
- Added new column “news_label” in both dataframes and combined datasets and added labels (1 for real, 0 for fake).
- Checked for null values and handled it by dropping the null values using df.dropna().
- Combined the relevant columns into a new column “news_text” and then dropped irrelevant columns (i.e. 'title', 'text', 'date') from the DataFrame.

2. Text Preprocessing

As part of the Test Preprocessing below steps are performed

- Converted text to lowercase.
- Removed punctuation, special characters, and stopwords.
- Applied lemmatization using spaCy or NLTK.
- Applied POS tagging and lemmatization function to cleaned text and stored it in a separate column (i.e. “lemmatized_news_text”) in the new DataFrame
- Saved the cleaned data in new csv file as “clean_df.csv”

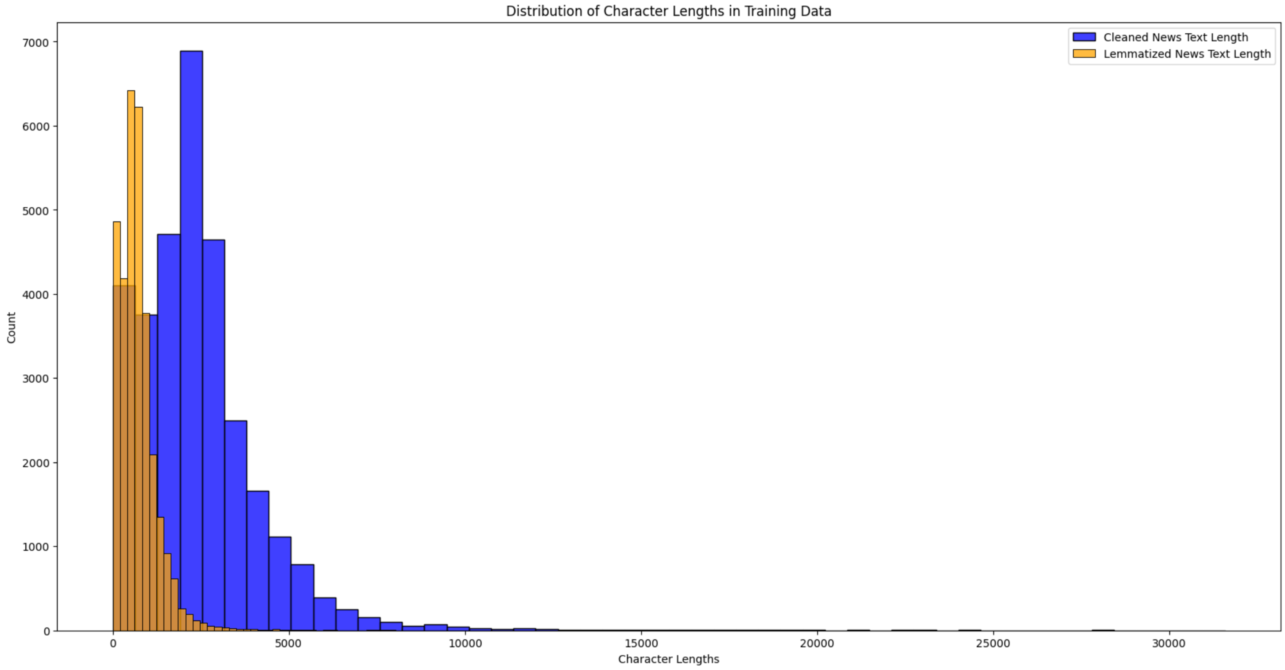
3. Train Validation Split:

Splitted the data into 70% train and 30% validation data as “X_train”, “X_val”, “y_train”, and “y_val”.

4. EDA on Training Data

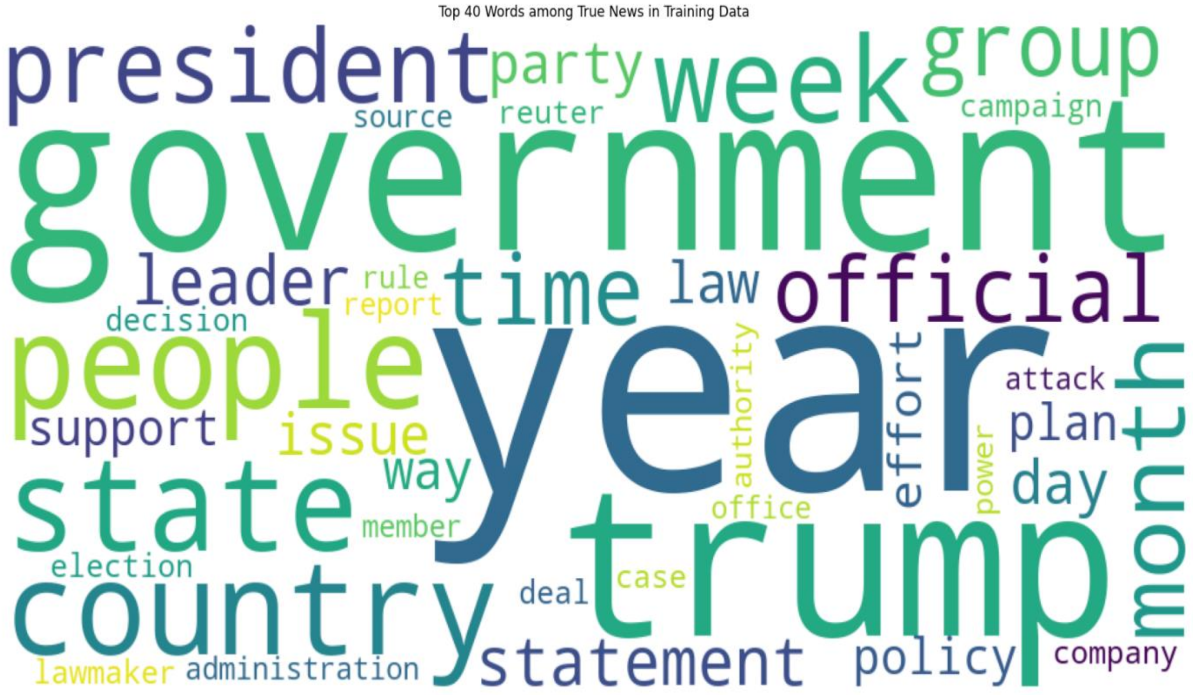
Insights from the Histogram Plot: Distribution of Character Lengths in Training Data

- Most Articles are Short to Moderate in Length:
    1. The bulk of both distributions lies in the 500 - 3000 character range.
    2. This suggests that the majority of news articles in the dataset are relatively concise.
- Lemmatization Reduces Text Size:
    1. The orange distribution (lemmatized text) is shifted left compared to the blue one.
    2. This confirms that lemmatization and removal of less informative words (like stopwords) reduces character count, making the text more compact.
- Tail Behavior:
    1. The cleaned text shows a longer tail, with some articles reaching lengths above 10,000 characters.
    2. These may be unusually verbose articles, outliers, or aggregated news summaries.
- Consistency Across Samples:
    1. Both distributions follow a similar right-skewed pattern, which is typical for text length distributions.
    2. This skew suggests a few articles are very long, but most are concentrated around the median.



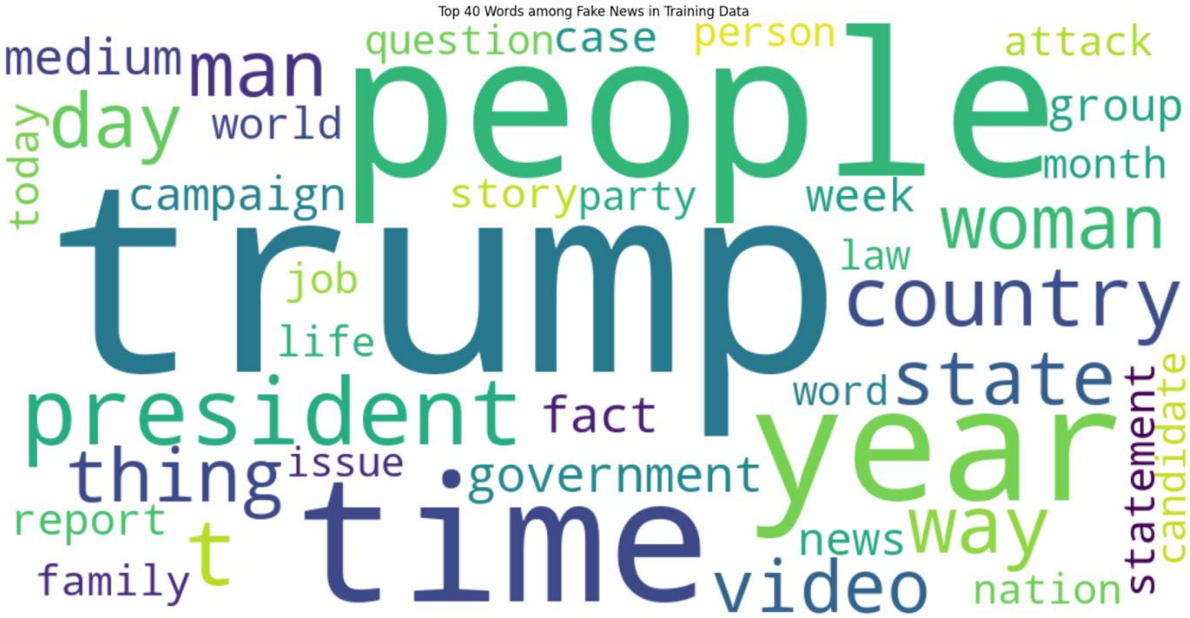
Insights from Word Cloud – Top Words in True News

- Frequent Focus on Governance and Policy:
    1. Words like “government”, “state”, “official”, “authority”, and “administration” dominate.
    2. This indicates that true news often covers formal institutions, governance, and public policy matters.
- Temporal References Are Common:
    1. Words such as “year”, “month”, “week”, “day”, and “time” are prominent.
    2. This reflects typical journalistic practice of providing context, timelines, or updates on ongoing events.
- Leadership and Political Entities:
    1. Terms like “president”, “leader”, “lawmaker”, “party”, and “trump” highlight a focus on individuals in power and political reporting.
    2. The presence of “Trump” reflects coverage during a time when he was a dominant political figure.
- Public Interest and Civic Themes:
    1. Words like “people”, “country”, “policy”, and “issue” suggest that the articles address national-level concerns relevant to society.
- Neutral and Factual Tone:
    1. The word usage overall reflects a more factual, institutional, and policy-driven vocabulary, consistent with mainstream reporting.



Insights from Word Cloud – Top Words in Fake News

- Frequent Focus on Individuals and Personal Stories:
    1. Words like “man”, “woman”, “person”, “family”, and “life” suggest a tendency toward storytelling and emotional framing.
    2. This contrasts with institutional or policy-oriented language often seen in true news.
- Dominance of Politicized Terms:
    1. Terms like “trump”, “president”, “campaign”, and “state” appear prominently.
    2. While similar to true news, the context here often leans toward sensational or polarizing narratives.
- Sensational and Vague Vocabulary:
    1. Words like “thing”, “story”, “fact”, “question”, and “video” are highly generic.
    2. This reflects a pattern of clickbait-style or unverified content, often lacking specificity or formal reporting standards.
- Narrative-Oriented and Conversational Tone:
    1. Phrases such as “way”, “job”, “today”, “word”, and “report” hint at casual storytelling or opinion-style writing, common in misinformation or blogs.
- Less Institutional Language:
    1. Compared to the true news word cloud, there's noticeably less use of terms like “authority”, “policy”, or “administration” reinforcing the lack of formal reporting structure.



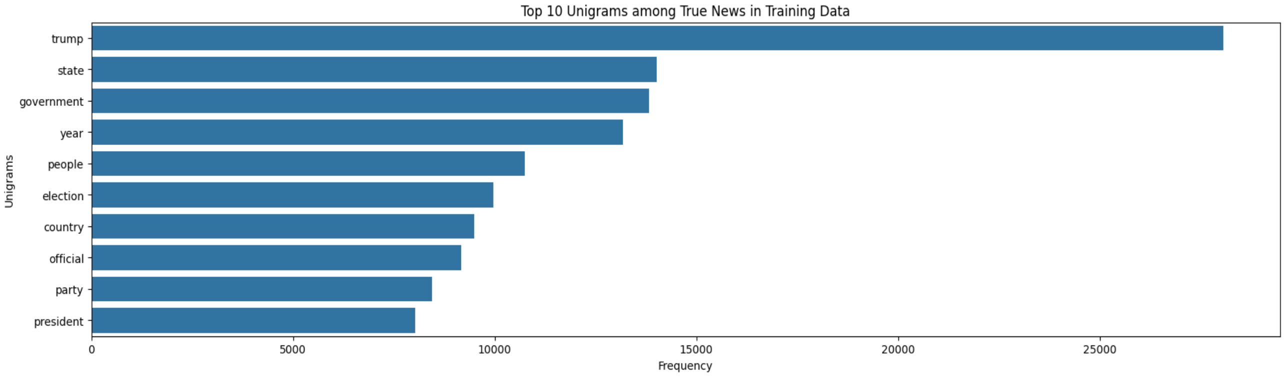
Insights Unigrams, Bigrams and Trigrams for True News:

1. Unigrams (Single Words):

Most frequent observed words of Unigrams are "trump", "state", "government", "year", "people", "election", "country", "official", "party", "president".

Key Insights:

- These terms suggest that true news focuses on politics, governance, and public administration.
- The high frequency of "trump" reflects his relevance in verified political coverage during the dataset's timeframe.
- Words like "official", "government", and "election" indicate institutional reporting and formal discourse.

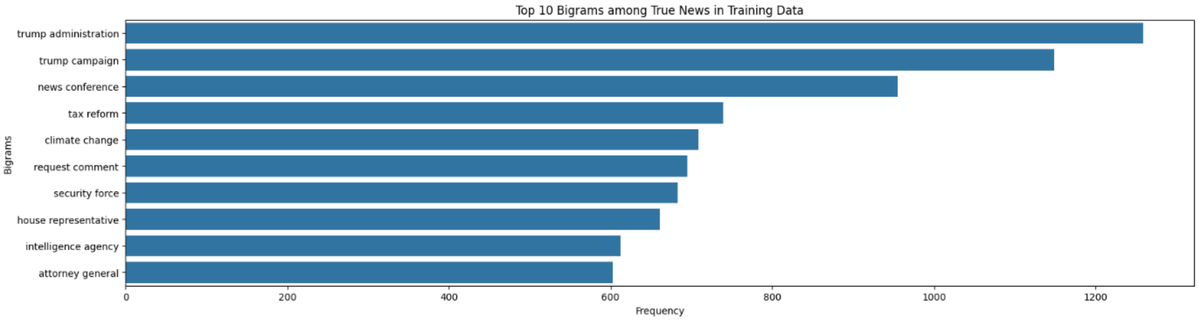


2. Bigrams (Two Word Phrases):

Most frequent observed words of Bigrams are "trump administration", "trump campaign", "news conference", "tax reform", "climate change", "security force", etc.

Key Insights:

- These bigrams reflect structured coverage of specific policies like tax reform, climate change and official entities like trump administration, intelligence agency.
- This emphasizes the formality and specificity in true news reporting often linked to verifiable events, press briefings, or policy actions.

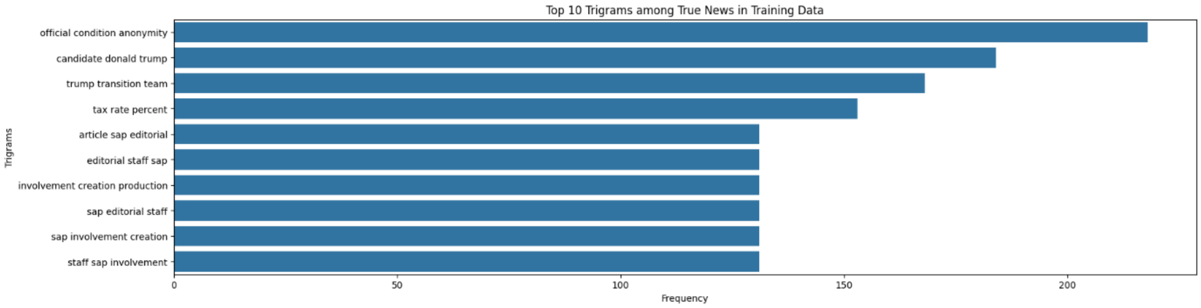


3. Trigrams (Three Word Phrases):

Most frequent observed words of Trigrams are "official condition anonymity", "candidate donald trump", "trump transition team", "tax rate percent", etc.

Key Insights:

- Trigrams often reflect journalistic phrasing like "official condition anonymity", a standard way to cite unnamed sources.
- Phrases like "trump transition team" and "candidate donald trump" point to election-related factual coverage.
- Repeated mentions of "editorial staff" or "sap editorial" may indicate structured publishing credits or syndicated content.



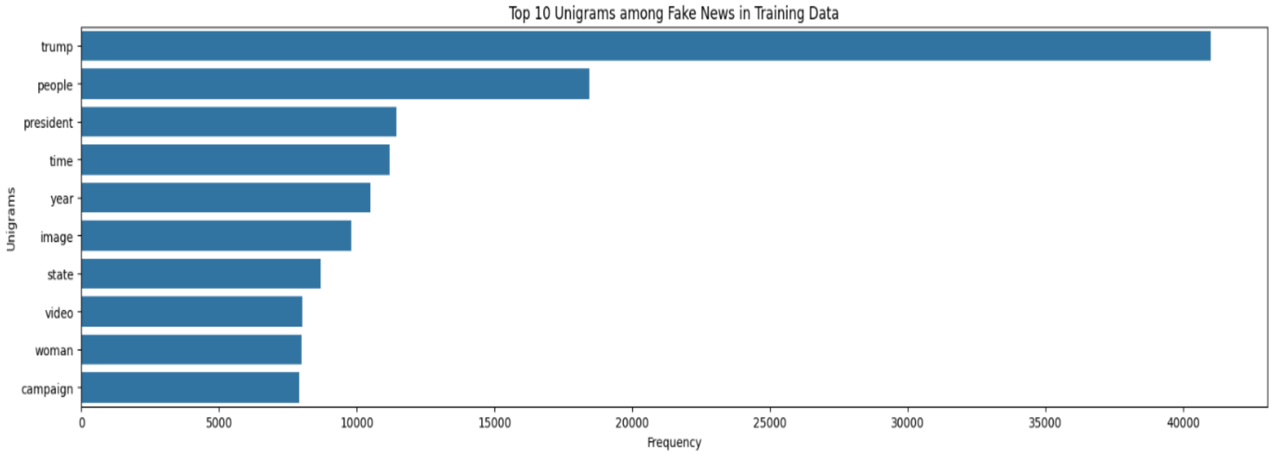
Insights Unigrams, Bigrams and Trigrams for Fake News:

1. Unigrams (Fake News):

Most frequent observed words of Unigrams are "trump", "people", "president", "time", "year", "image", "state", "video", "woman", "campaign".

Key Insights:

- The word "trump" dominates by a wide margin, indicating a heavy emphasis on politically charged or sensational content involving public figures.
- Words like "image", "video", and "woman" suggest a strong reliance on visual media, viral stories, or potentially emotionally driven content.
- Compared to true news, fake news unigrams tend to include more informal or visually suggestive elements.

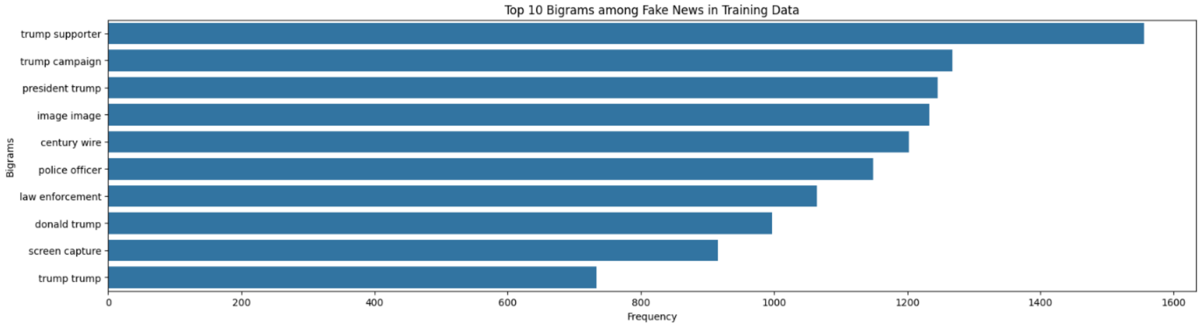


2. Bigrams (Fake News):

Most frequent observed words of Bigrams are "trump supporter", "trump campaign", "president trump", "image image", "century wire", "police officer", "law enforcement", "donald trump", "screen capture", "trump trump".

Key Insights:

- Repetitive terms like "image image" and "trump trump" may suggest redundancy or poor text generation, common in manipulated or auto-generated articles.
- The presence of "century wire" hints at repeated references to fringe or less credible media sources.
- High mention of "law enforcement" and "police officer" could relate to emotionally triggering topics like crime or conspiracy.

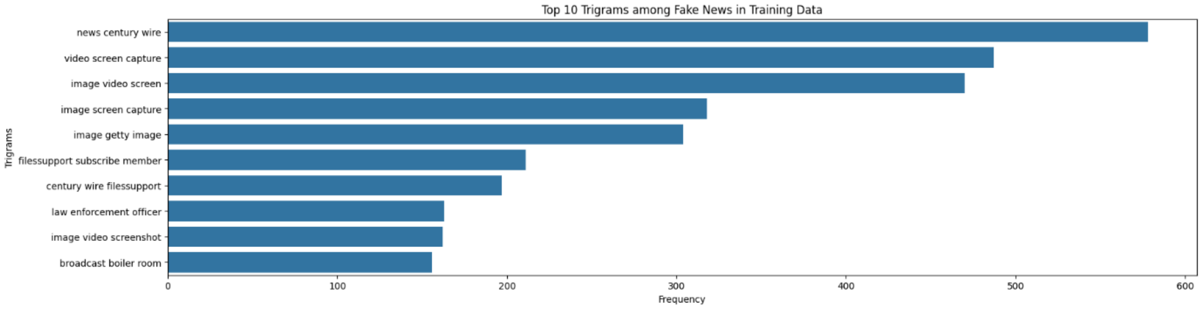


3. Trigrams (Fake News):

Most frequent observed words of Trigrams are "news century wire", "video screen capture", "image video screen", "image screen capture", "image getty image", "filessupport subscribe member", "law enforcement officer", etc.

Key Insights:

- Many top trigrams contain media descriptors (e.g., screen capture, getty image, video screenshot), showing that fake news often includes viral visuals or clickbait thumbnails.
- Phrases like "news century wire" and "filessupport subscribe member" suggest promotion of alternative or spammy news outlets, often linked to misinformation.
- The repetition of visual-related trigrams reinforces the strategy of using images/videos to boost engagement or credibility.



Overall Observation of Unigrams, Bigrams and Trigrams (True Vs Fake News):

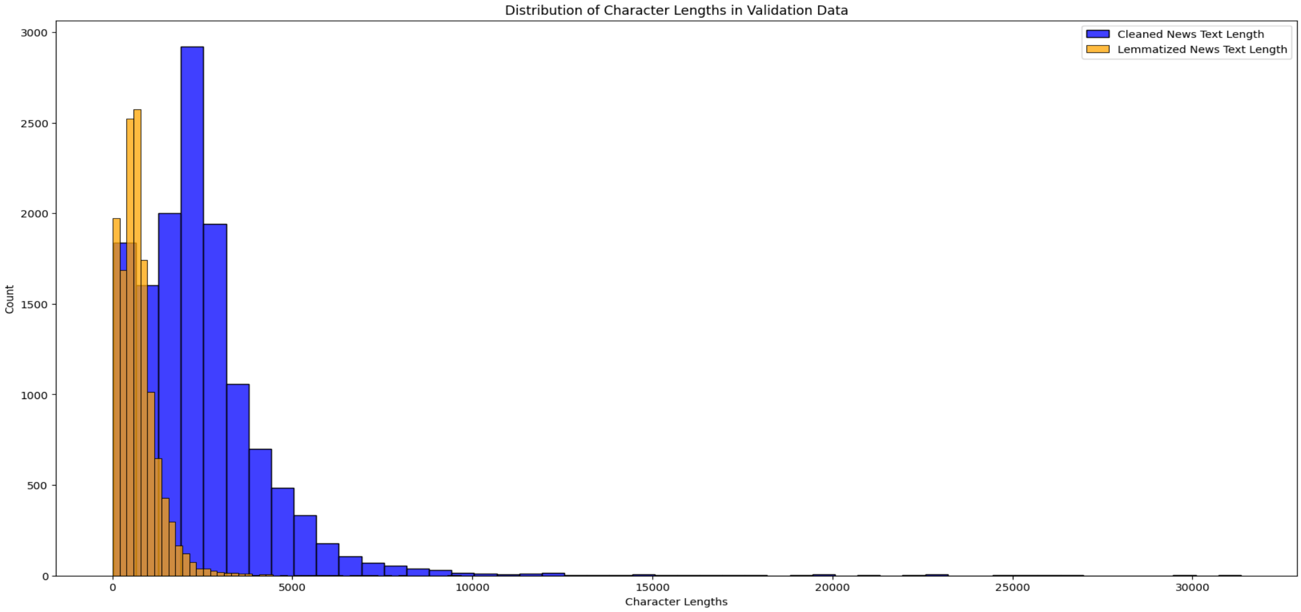
Compared to true news:

- Fake news relies more on visual and repetitive media language.
- It features fewer institutional terms and instead shows signs of emotional framing, redundancy, and heavy media narratives.
- Named entities and viral topics like Trump, police, campaign dominate, often lacking depth or structured reporting.

5. EDA On Validation Data:

Insights from histogram plot on validation data set:

- Most articles in the validation set have character lengths between 0 and 4000.
- A smaller number extend beyond 10,000 characters, and very few go up to or beyond 30,000.
- The orange bars (lemmatized text) show consistently shorter character lengths than the cleaned version. This confirms that lemmatization significantly reduces text size, making the input more compact while preserving semantics.
- Both cleaned and lemmatized text lengths exhibit a right skewed distribution, meaning:
    1. Most articles are short to moderately long.
    2. Only a few outliers are very lengthy.
- The mode for lemmatized text is around 500 - 1000 characters, with over 2500 articles in this range.
- The cleaned text peaks between 2500 - 3000 characters, with nearly 3000 articles.
- This analysis helps in padding/truncation decisions for model inputs.



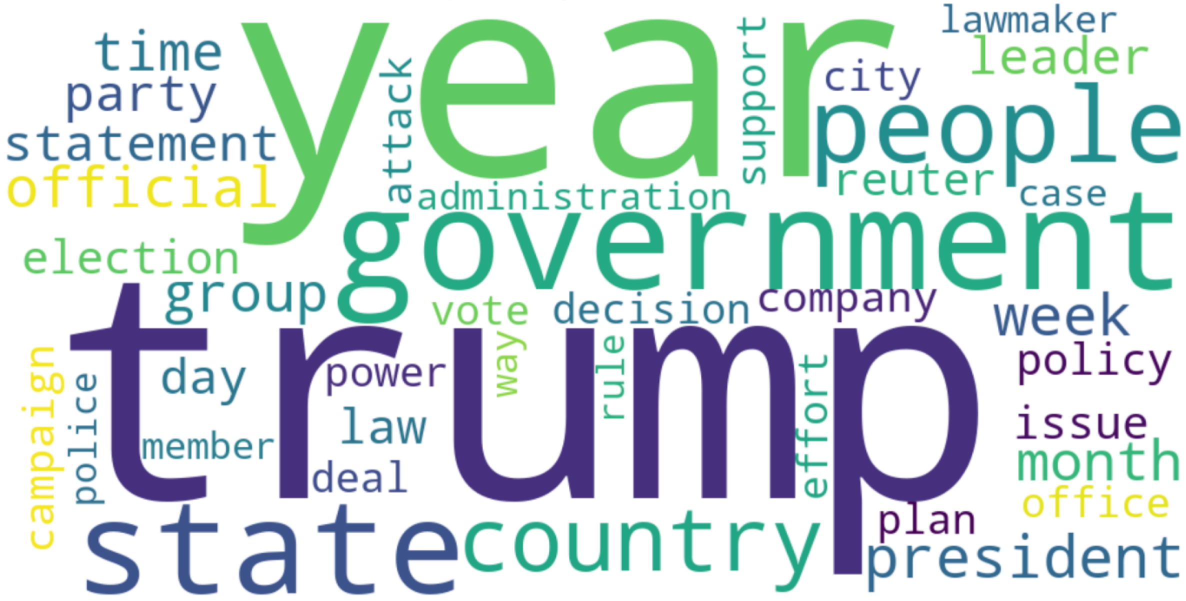
Insights from Word Cloud – Top Words in True News (Validation Data)

The most prominent and frequently occurring words in true news validation data include “year”, “trump”, “government”, “state”, “country”, “people”, “president”, “official”, and “policy”.

- **Dominance of Political Themes:**
  1. Words such as “trump”, “government”, “state”, “president”, and “official” suggest a continued focus on political and governmental reporting, just like the training set.
- **Institutional Reporting:**
  - Terms like “administration”, “policy”, “law”, and “decision” reflect structured news related to governance, legislation, and official communications.
- **Temporary Reference:**
  1. The prominence of “year”, “month”, “week”, and “day” highlights that timely reporting and chronological framing are central to true news articles.
- **Entities and Actions:**
  1. Words such as “people”, “leader”, “vote”, “campaign”, and “support” indicate reporting on democratic processes and civic participation.

The word cloud of true news in the validation set mirrors patterns from the training data, emphasizing institutional coverage, political affairs, and time-referenced reporting. This consistency supports the reliability of semantic patterns used in classification and model generalization.

Top 40 Words among True News in Validation Data



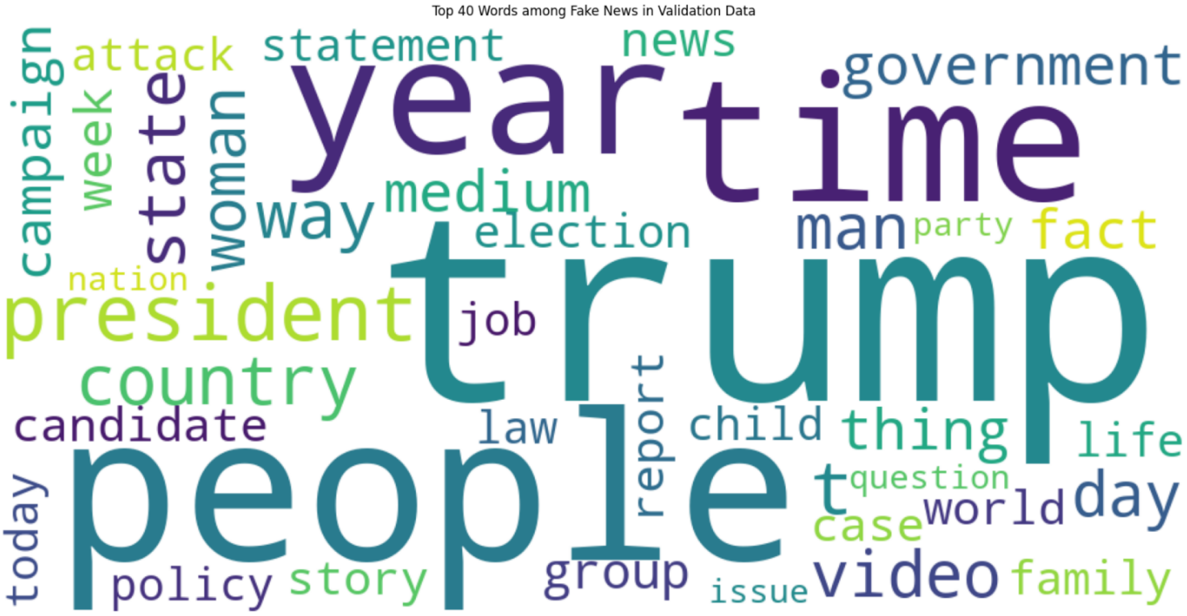
Insights from Word Cloud – Top Words in Fake News (Validation Data)

The most prominent words in fake news validation data include “trump”, “people”, “president”, “year”, “time”, “video”, “country”, and “story”.

- **Focus on Individuals and Sensational Topics:**
  1. Frequent words such as “trump”, “president”, “woman”, “man”, “child”, and “family” suggest that fake news often revolves around personal stories or characters possibly to draw emotional reactions or attention.
- **Informal and Broad Language:**
  1. Words like “thing”, “job”, “story”, and “way” reflect a more conversational or subjective tone, differing from the formal vocabulary in true news.
- **Multimedia Elements:**
  1. The presence of “video”, “news”, and “report” may indicate a heavy reliance on visual or viral content, which is commonly used in fake news to gain traction on social media.
- **Temporal and Contextual Framing:**
  1. Terms like “year”, “day”, “today”, and “time” show that temporal framing is also common in fake news, perhaps used to make stories feel urgent or current.
- **Keywords Indicating Speculation or Claims:**
  1. Words like “fact”, “question”, and “story” could suggest that fake news often makes assertions, raises doubts, or presents anecdotes rather than grounded reports.

The word cloud reveals that fake news articles often focus on individuals, emotional themes, and viral media, contrasting with the more institutional, policy-oriented language seen in true news. This distinction can be leveraged in semantic classification to differentiate between credible and misleading content.

Top 40 Words among Fake News in Validation Data

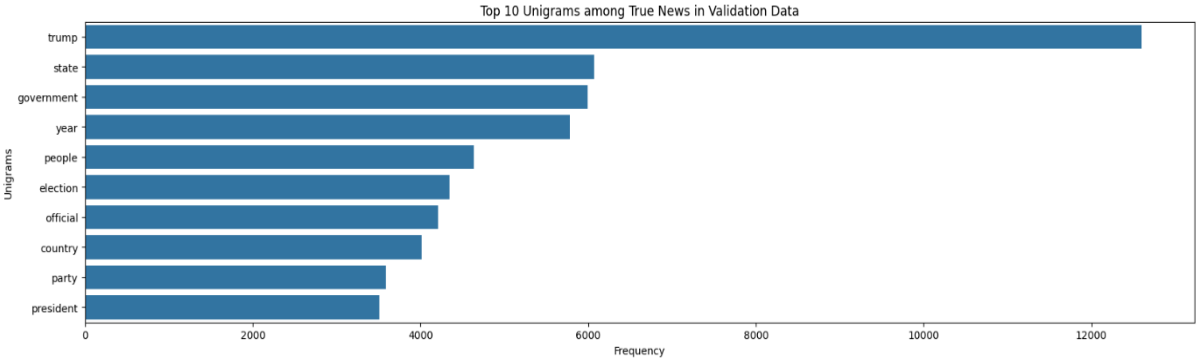


Insights Unigrams, Bigrams and Trigrams for True News:

1. **Unigrams in True News (Validation Data)**

   Key Insights:

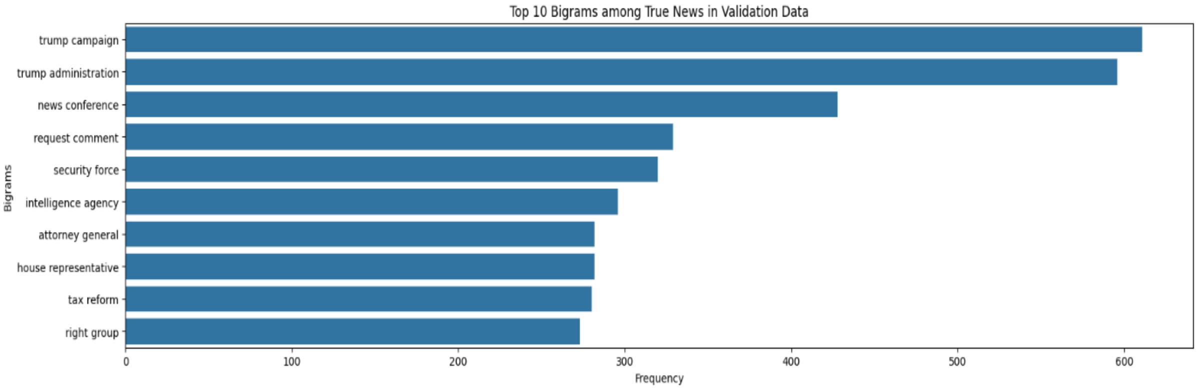
   - Most frequent observed words of Unigrams are "trump", "state", "government", "year", "people", "election", "country", "official", "party", "president".
   - The term “trump” dominates with more than double the frequency of the next most common word, reflecting the major political news coverage.
   - Words like “government”, “state”, and “official” suggest a strong focus on institutional and policy related reporting.
   - Presence of “election”, “party”, and “president” reinforces the political orientation of true news articles in the dataset.



2. **Bigrams in True News (Validation Data):**

   Key Insights:

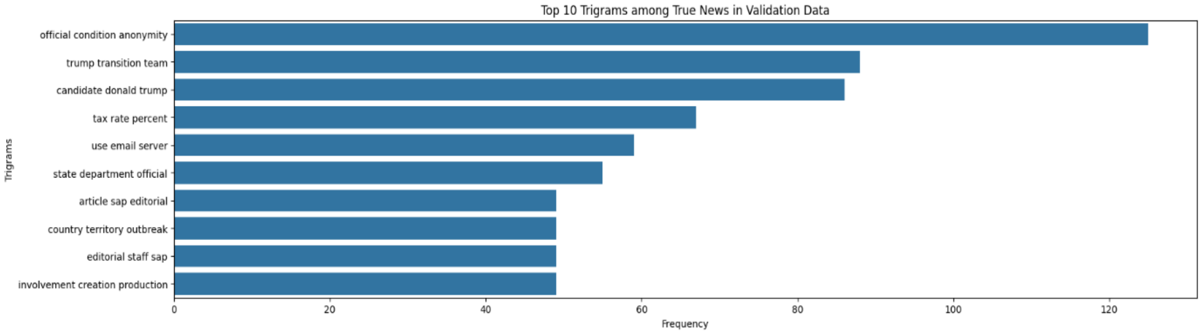
   - Frequent mention of “trump campaign” and “trump administration” aligns with the heavy political content.
   - Terms like “intelligence agency”, “security force”, and “attorney general” indicate coverage of governance and security-related topics.
   - “request comment” and “news conference” show journalistic reporting elements, reflecting factual, quote-driven content.



3. **Trigrams in True News (Validation Data):**

   Key Insights:

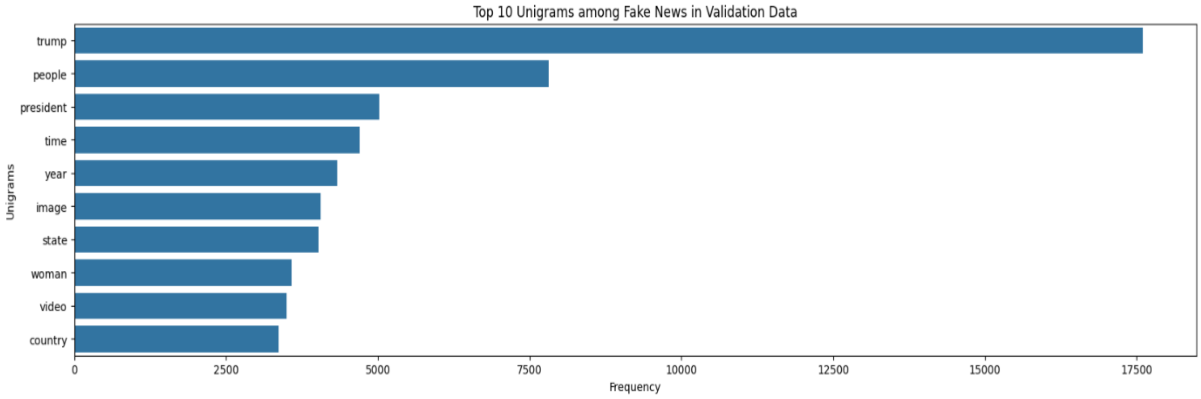
   - The top trigram “official condition anonymity” suggests credible journalistic sources with unnamed officials a hallmark of true news reporting.
   - Multiple trigrams reference real political figures or events, such as “candidate donald trump” and “trump transition team”.
   - Terms like “use email server” and “state department official” reflect detailed factual content from government or investigative reporting.
   - Mentions of “sap editorial staff” could indicate publisher metadata or structured attribution practices.



Insights Unigrams, Bigrams and Trigrams for Fake News:

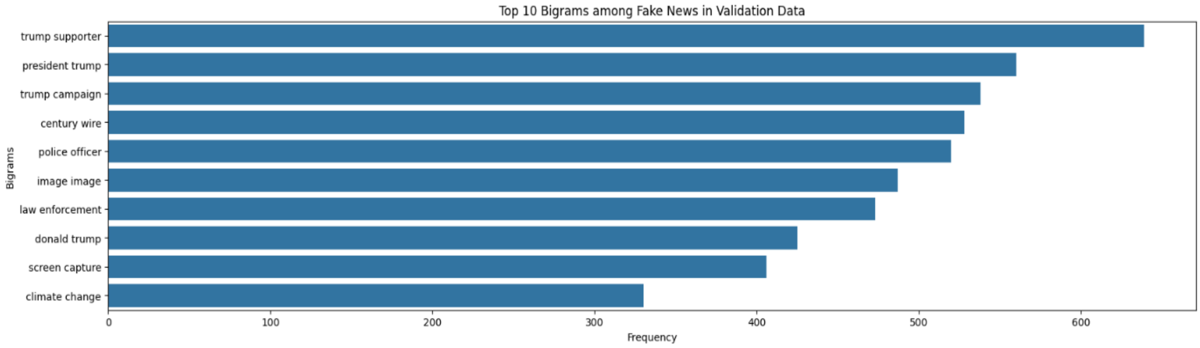
1. **Unigrams in Fake News (Validation Data):**

   - Dominant terms in this are trump, people, president, time, and year.
   - The word "trump" appears with a significantly higher frequency, indicating a strong presence in fake news.
   - Words like "image", "video", and "woman" may hint at sensational or media-heavy fake content.
   - Fake news often focuses on high-profile figures and media-based narratives. The frequent use of terms like image and video implies the use of visual content to manipulate or attract readers.



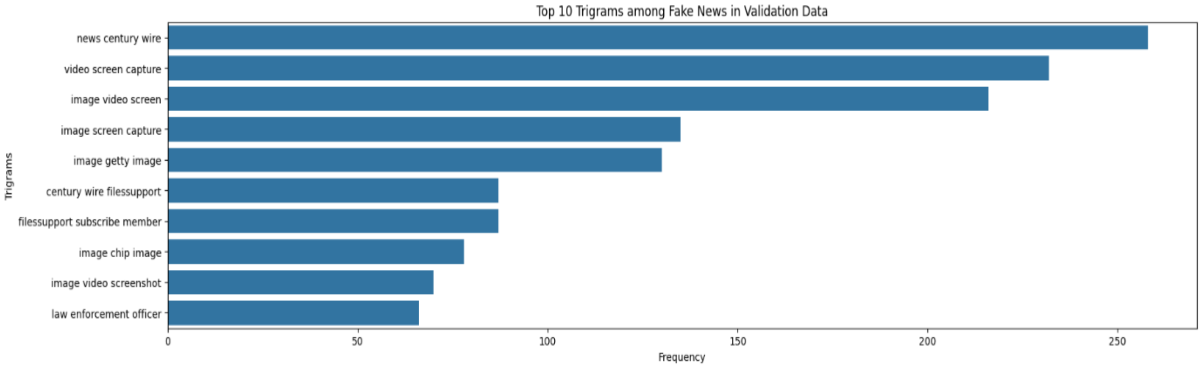
2. Bigrams in Fake News (Validation Data):

- Frequent pairs in this are trump supporter, president trump, trump campaign, police officer, century wire.
- Heavy usage of political bigrams, particularly involving Trump.
- Repeated terms like image image or screen capture suggest reliance on visual media or duplicated phrases.
- "Century wire" might refer to a specific source commonly associated with fake articles.
- These bigrams indicate that fake news leans heavily on political themes, often repeating names or using vague official-sounding sources to gain credibility.



3. Trigrams in Fake News (Validation Data):

- Frequent three-words in this are news century wire, video screen capture, image video screen, law enforcement officer.
- Trigrams further show that fake news may reuse phrases involving media or images, often in combinations like "video screen capture" or "image getty image".
- The presence of "law enforcement officer" and "filesupport subscribe member" suggests attempts to emulate authentic news or embed call-to-action phrases.
- Trigrams reveal that repetitive and media-related phrases are a key feature in fake news, possibly aimed at simulating official tone or misleading through familiarity.



6. Feature Extraction:

We have used Word2Vec for the feature extraction technique to convert news article texts into numerical vector representations before feeding them into supervised machine learning models.
Below are the implementation steps:
Tokenization: The preprocessed and lemmatized news text was first tokenized (split into individual words).

1. Word2Vec Training:
   - The model was trained using the gensim.models.Word2Vec implementation.
2. Key parameters included:
   - vector_size=100 the dimensionality of the word vectors.
   - window=5 the maximum distance between the current and predicted word.
   - min_count=2 ignores words that appear less than twice.
   - workers=4 number of threads used during training.
3. Vector Representation:
   - For each news article, average word vectors were calculated to represent the full document.
   - This results in a 100-dimensional vector per article (average of word vectors in the article).

Key Observations

- o Word2Vec allowed models like Logistic Regression and Random Forest to learn from semantic patterns in articles rather than just keyword frequency.
- o Using the average of word vectors ensured that even long articles were compressed into fixed-size input vectors, making them suitable for model training.

7. Model Training and Evaluation:

Used Model:

- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier

Evaluation Metrics:

Each model’s performance was evaluated using:

- Accuracy: Proportion of correct predictions.
- Precision: True Positives / (True Positives + False Positives) – useful to minimize false alarms.
- Recall: True Positives / (True Positives + False Negatives) useful to catch more fake/true news.
- F1 Score: Harmonic mean of precision and recall balances both metrics.

Performance Summary:

- Here are the general trends observed

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Logistic Regression	High	High	High	High
Decision Tree	Moderate	Moderate	Moderate	Moderate
Random Forest	Highest or Similar to Logistic	High	High	High

Insights by Model:

1. Logistic Regression:
   - Strong baseline model.
   - Performs well with Word2Vec embeddings due to its simplicity and generalization.
2. Decision Tree:
   - Tends to overfit on training data.
   - Performance was weaker on validation/test sets.
3. Random Forest:
   - Improved generalization due to ensemble voting.
   - Slightly better or comparable to logistic regression in terms of F1 score and recall.

8. Final Conclusions:

Summarising the findings by discussing patterns observed in true and fake news and how semantic classification addressed the problem. Highlighted the best model chosen, the evaluation metric prioritised for the decision, and assess the approach and its impact.

- True news emphasize institutional and policy-related terms such as "election", "country", "tax reform", "tax rate percent", "official condition anonymity" etc. Whereas fake news articles frequently contain multimedia references like "image", "video", "screen capture", "video screen capture", "image video screenshot" etc.
- Stop words were removed, and lemmatization was applied to reduce words to their root forms. Additionally, only nouns (NN and NNS POS tags) were retained, which significantly reduced the training text length while preserving essential information.
- The Word2Vec model was used to extract semantic relationships, representing each word as a vector. Sentence embeddings were obtained by averaging the vectors of the constituent words, which were then used as input features for model training.
- The Random Forest model delivered the best performance with an F1 score of 90% on the validation set.
- The F1 score was selected as the primary evaluation metric as it balances both precision and recall. This ensures that the model minimizes both false positives and false negatives effectively.
- Semantic classification captures the underlying meaning and context of news articles, making the model resilient to keyword manipulation. This robustness enhances its suitability for real-world deployment scenarios.