Text Mining and Analysis of Political Discourse Using the LIAR Dataset

Alexandre Crivellari[1], Andrea Muscio[1]

[1]MSc Data Science, Università degli Studi Milano-Bicocca

GitHub Repository: https://github.com/alexcri90/LIAR_TextMining_Bicocca.git

Abstract

This study investigates the efficacy of automated fact-checking systems and topic modeling in political discourse using the LIAR dataset, which comprises over 12,000 labeled political statements. We implemented and compared two distinct approaches for text classification: a traditional machine learning method (Random Forest) and a state-of-the-art deep learning model (RoBERTa). Surprisingly, the Random Forest classifier outperformed RoBERTa, achieving 72.69% accuracy compared to RoBERTa's 61.17%, challenging assumptions about the superiority of deep learning in NLP tasks. This finding underscores the importance of feature engineering and domain-specific knowledge in political fact-checking.

Additionally, we employed Latent Dirichlet Allocation (LDA) for topic modeling, revealing distinct thematic patterns between true and false statements. True statements tended to focus on specific policy areas and economic indicators, often including precise numerical claims. In contrast, false statements showed a greater emphasis on political personalities, controversies, and broader, less specific claims.

Our research contributes to the understanding of misinformation propagation and the challenges of automated fact-checking in political contexts. The unexpected performance gap between traditional and deep learning models highlights the need for hybrid approaches that leverage both domain expertise and advanced language understanding. Furthermore, the identified thematic differences between true and false statements offer valuable insights for developing more nuanced fact-checking algorithms and enhancing media literacy initiatives.

Summary

1. Introduction

1.1 Background

In the era of digital communication and social media, the spread of information has become rapid and far-reaching. Political statements, in particular, have a significant impact on public opinion and policy-making. However, the veracity of these statements is often questionable, leading to the proliferation of misinformation and its potential consequences on democratic processes. This phenomenon has given rise to an urgent need for automated fact-checking systems and deeper analysis of political discourse.

The advent of advanced natural language processing (NLP) techniques and machine learning algorithms has opened new avenues for analyzing and understanding large volumes of textual data. These technologies offer promising solutions for automatically classifying the truthfulness of statements and uncovering the underlying themes in political discourse.

1.2 Problem Statement

Despite the importance of verifying political statements, manual fact-checking is a time-consuming and resource-intensive process. Moreover, the sheer volume of political discourse makes it challenging to identify overarching themes and patterns in the topics discussed by public figures. This research addresses two primary objective:

1. To develop and evaluate machine learning models capable of accurately classifying political statements as true or false, comparing traditional approaches with state-of-the-art deep learning techniques.
2. To apply topic modeling techniques to identify and analyze the main themes present in political discourse, with a particular focus on distinguishing topics between true and false statements.

1.3 Literature Review

The literature review provides a comprehensive overview of the existing research and methodologies relevant to our study. This chapter contextualizes our work within the broader field of natural language processing, political discourse analysis, and automated fact-checking systems.

1.3.1 Text Classification in Political Discourse

Text classification in the context of political discourse has gained significant attention in recent years due to its potential to automatically assess the veracity of political statements. This field of study intersects with natural language processing, machine learning, and political science.

Early attempts at automated fact-checking relied heavily on rule-based systems and simple statistical methods. However, these approaches often fell short when dealing with the complexity and nuance of political language. Vlachos and Riedel (2014) were among the first to propose fact-checking as a new AI task, highlighting the challenges of claim identification and veracity assessment.

As machine learning techniques advanced, researchers began applying more sophisticated algorithms to the problem. Wang (2017) introduced the LIAR dataset, a benchmark resource for fake news detection research, which our study utilizes. This dataset, containing over 12,000 labeled political statements, has become a standard for evaluating automated fact-checking systems.

Traditional machine learning approaches, such as Support Vector Machines (SVM) and Random Forests, have shown promising results in political text classification. Pérez-Rosas et al. (2018) demonstrated the effectiveness of linguistic features combined with Random Forests in detecting fake news, achieving accuracy rates of up to 78%.

The advent of deep learning models has further pushed the boundaries of text classification in political discourse. Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNNs) have been applied successfully to this domain. For instance, Karimi et al. (2019) used a hybrid CNN-LSTM model for fake news detection, showing improved performance over traditional methods.

More recently, transformer-based models like BERT (Bidirectional Encoder Representations from Transformers) and its variants have set new benchmarks in various NLP tasks, including political text classification. Augenstein et al. (2019) employed BERT for fact-checking, demonstrating its ability to capture complex linguistic patterns and contextual information.

Our study builds upon this rich body of work by comparing a traditional machine learning approach (Random Forest) with a state-of-the-art transformer model (RoBERTa) for classifying political statements. This comparison not only contributes to the ongoing discourse about the efficacy of different machine learning paradigms in political text analysis but also provides practical insights into the trade-offs between model complexity and performance in real-world applications.

The field of text classification in political discourse continues to evolve rapidly, with researchers exploring multi-modal approaches, incorporating external knowledge bases, and developing more interpretable models.

1.3.2 Topic Modeling Techniques

Topic modeling has emerged as a powerful tool for uncovering latent thematic structures within large collections of documents. In the context of political discourse analysis, topic modeling techniques offer valuable insights into the main themes and narratives present in political statements, speeches, and debates.

Latent Dirichlet Allocation (LDA), introduced by Blei, Ng, and Jordan (2003), has become the cornerstone of topic modeling research. LDA is a generative probabilistic model that assumes documents are mixtures of topics, and topics are mixtures of words. This unsupervised learning approach has been widely adopted in political science research due to its ability to discover hidden thematic structures without prior labeling.

Quinn et al. (2010) demonstrated the effectiveness of topic models in analyzing political texts, applying LDA to congressional speeches to identify policy issues and track their evolution over time. Their work highlighted the potential of topic modeling in understanding the dynamics of political discourse and agenda-setting processes.

Building upon LDA, researchers have developed more sophisticated topic modeling techniques to address specific challenges in political text analysis. Supervised LDA (sLDA), proposed by Mcauliffe and Blei (2008), incorporates document labels to guide topic discovery, which can be particularly useful when analyzing political texts with known categories or affiliations.

Dynamic Topic Models (DTM), introduced by Blei and Lafferty (2006), extend LDA to capture the evolution of topics over time. This approach has been valuable in tracking shifts in political narratives and policy priorities across electoral cycles or legislative sessions. Greene and Cross (2017) applied DTM to European Parliament speeches, revealing how political agendas and discussions evolved over multiple terms.

In recent years, researchers have explored the integration of word embeddings with topic models to enhance semantic coherence. Dieng et al. (2020) proposed the Embedded Topic Model (ETM), which combines neural word embeddings with LDA-style topic modeling, resulting in more interpretable and semantically meaningful topics. This approach has shown promise in capturing nuanced political concepts and ideologies.

The application of topic modeling to distinguish between true and false statements, as implemented in our study, builds upon work by Rashkin et al. (2017). They used a combination of linguistic features and topic modeling to analyze the language of fake news, propaganda, and satire, highlighting distinct topical and stylistic patterns associated with deceptive content.

Our research contributes to this body of literature by applying LDA to separately model topics in true and false political statements. This approach not only reveals the overarching themes in political discourse but also allows for a comparative analysis of topical distributions between truthful and false claims. By doing so, we aim to uncover potential patterns or biases in the subjects of misinformation, adding a new dimension to the understanding of political deception.

The evaluation of topic models remains an active area of research. Metrics such as perplexity and topic coherence, which we employ in our study, have been extensively discussed in the literature. Röder et al. (2015) provided a comprehensive comparison of topic coherence measures, emphasizing the importance of human interpretability alongside statistical measures.

As the field of topic modeling continues to advance, new challenges and opportunities arise. The integration of contextual information, the handling of short texts (common in social media political discourse), and the development of multilingual topic models are areas of ongoing research that promise to further enhance our ability to analyze and understand political communication.

2. Methodology

2.1 Dataset Description

Our study utilizes the LIAR dataset, a benchmark collection for fake news detection research introduced by Wang (2017). This dataset comprises 12,836 short statements labeled for veracity, with attributes structured as follows:

- **id**: A unique identifier for each statement.

- **label**: The veracity label assigned to the statement (pants-fire, false, barely-true, half-true, mostly-true, true).
- **statement**: The text of the statement being fact-checked.
- **subject**: The topic or subject area of the statement.
- **speaker**: The name of the person who made the statement.
- **job_title**: The job title or position of the speaker.
- **state_info**: The state or location relevant to the statement.
- **party_affiliation**: The political party affiliation of the speaker.
- **barely_true_counts**, **false_counts**, **half_true_counts**, **mostly_true_counts**, **pants_on_fire_counts**: Historical count of the speaker's statements in each veracity category.
- **context**: Additional contextual information about where or when the statement was made.

The LIAR dataset possesses several key characteristics that make it valuable for our research:

1. *Source*: The statements in the dataset are drawn from POLITIFACT.COM, a reputable fact-checking website. This source ensures that the labels are assigned by professional fact-checkers, providing a reliable ground truth for our models.
2. *Label Granularity*: The dataset employs a fine-grained labeling system, categorizing statements into six levels of veracity: pants-fire, false, barely-true, half-true, mostly-true, and true. This granularity allows for a more detailed analysis of the spectrum of truth in political discourse.
3. *Metadata*: Alongside the statements and veracity labels, the dataset includes rich metadata such as the speaker, the speaker's job title, the state or location, the party affiliation, and the context in which the statement was made. This additional information provides valuable features for our classification task and enables more in-depth analysis of patterns in political statements.
4. *Statement Counts*: For each speaker, the dataset provides counts of their previous statements in each veracity category. This historical information can be leveraged as a feature in our classification models.

The dataset is pre-divided into training, validation, and testing sets, with 10,269, 1,284, and 1,283 statements respectively. This predefined split ensures consistency in model evaluation and comparability with other studies using the same dataset.

For our study, we focus on binary classification, grouping the original six-level labels into two categories:

- **True**: comprising 'true' and 'mostly-true' labels
- **False**: comprising 'false', 'pants-fire', 'barely-true', and 'half-true' labels

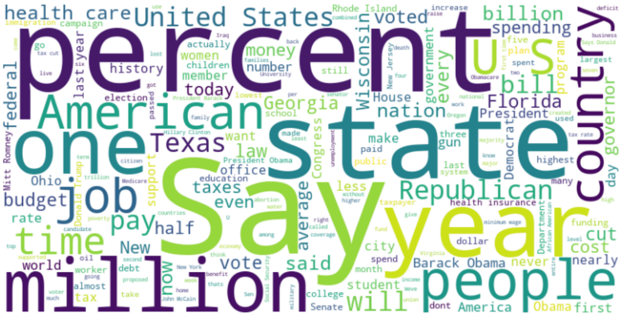
This binary classification allows us to directly address the core challenge of distinguishing between truthful and false statements while simplifying the modeling task.

The dataset's diversity in terms of speakers, topics, and political affiliations provides a robust foundation for our analysis. It encapsulates the complexity of real-world political discourse, including variations in language use, topic focus, and truthfulness across different political actors and contexts.

2.1.1 Word Clouds based on Statement

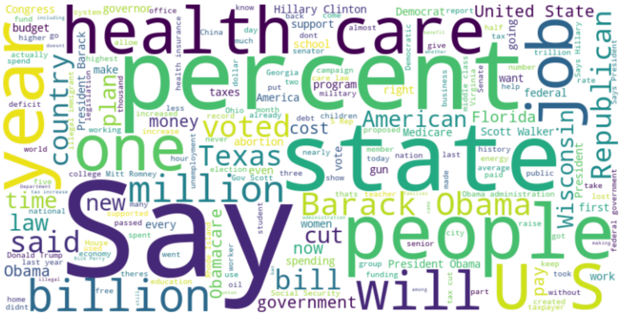
To complement our topic modeling analysis, we generated word clouds to visualize the most frequent terms in true and false statements respectively. These visualizations offer an intuitive representation of the lexical differences between truthful and false political discourse.

Word Cloud for True Statements



The figure above presents the word cloud for true statements. The most prominent words include "say," "percent," "state," "million," "year," "people," and "American." This lexical profile suggests that true statements often involve quantifiable claims, references to specific states or regions, and discussions about the American people. The prevalence of terms like "percent" and "million" indicates a tendency to use precise figures, which aligns with our earlier observation that true statements often include more specific, verifiable claims.

Word Cloud for False Statements

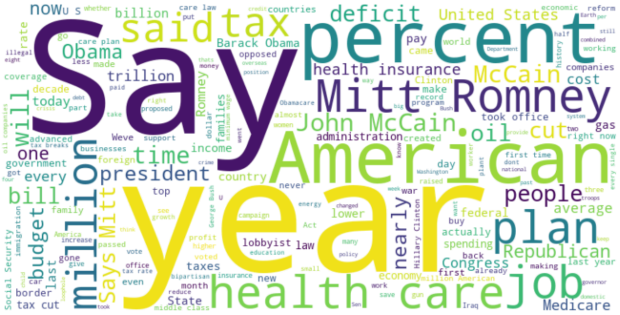


In contrast, the figure above shows the word cloud for false statements. Here, we see "say," "state," and "percent" remain prominent, but we also observe the frequent occurrence of "health care" and "Barack Obama." This suggests that false statements in our dataset often revolve around healthcare policies and the Obama administration. The presence of these topic-specific terms might indicate that certain policy areas are more prone to misinformation or misrepresentation.

2.1.2 Word Clouds based on Speaker

In a similar way, Word Clouds have been computed based on the Speaker attributed, along with the proportion of True statements based on that speaker. In particular, the following word clouds have been computed for Barack Obama, Hillary Clinton, Donald Trump and Mitt Romney.

Word Cloud for Barack Obama

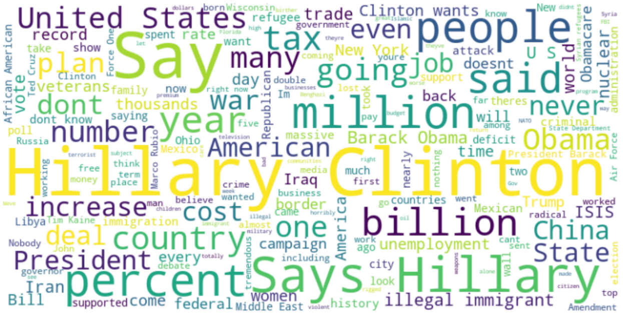


Average True Statements: 47.75%

The word cloud generated for Barack Obama's statements offers a rich visual representation of his rhetorical focus and policy priorities. Dominating the visualization are terms related to healthcare policy, economic issues, and broad national concerns, reflecting the key themes of Obama's presidency. The prominence of "health" and "care" underscores the centrality of healthcare reform in his political discourse, likely referencing the extensive debates surrounding the Affordable Care Act.

Interestingly, the cloud also captures the political landscape of Obama's era, with mentions of "Mitt Romney" and "John McCain" pointing to the ongoing dialogue with his Republican counterparts. This inclusion of political opponents in his discourse highlights the adversarial nature of American political rhetoric during his tenure.

Word Cloud for Donald Trump

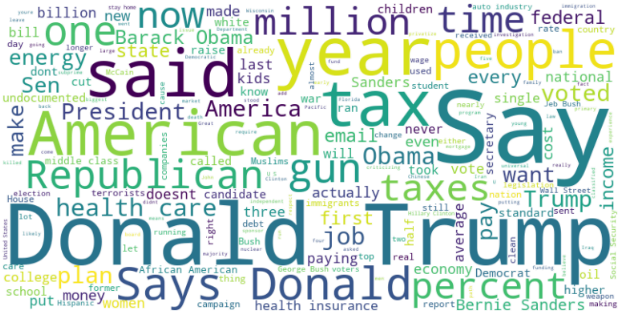


Average True Statements: 12.09%

This cloud related to Donald Trump's statements reveals a tendency to reference political opponents, with "Hillary Clinton" appearing prominently. This suggests a rhetorical strategy of contrast and opposition, framing his own positions against those of his political rivals. The presence of "Obama" further reinforces this approach, indicating frequent comparisons or criticisms of the previous administration.

Particularly striking is the statistic accompanying this visualization, indicating that only 12.09% of Trump's statements were classified as true. This remarkably low percentage of truthful statements stands in stark contrast to the figures observed for other political figures. Such a disparity raises significant questions about the relationship between Trump's rhetorical strategies and factual accuracy, suggesting a potential prioritization of persuasive narrative over verified information.

Word Cloud for Hillary Clinton



Average True Statements: 52.30%

The word cloud for Hillary Clinton reveals a specular significant emphasis on her political rival, with "Donald Trump" appearing as one of the most prominent phrases. This, again, suggests a rhetorical strategy that frequently involved direct comparisons or responses to her opponent's positions, reflecting the highly polarized nature of the electoral discourse.

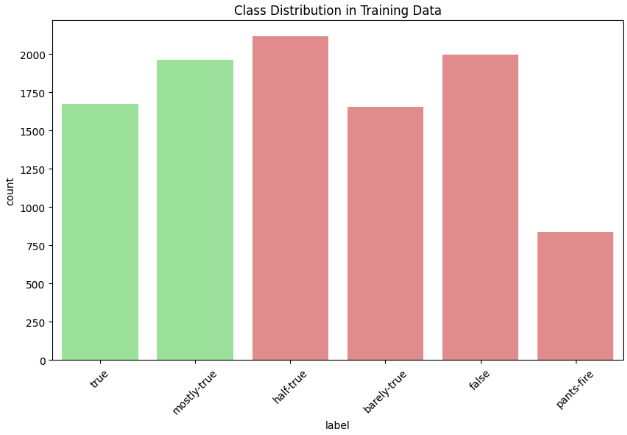
2.2 Dataset Preprocessing

Our data preprocessing pipeline consists of several key steps designed to prepare the textual data for both classification and topic modeling tasks:

1. *Text Cleaning*: We convert all text to lowercase to ensure consistency. Special characters and digits are removed using regular expressions, focusing our analysis on alphabetic content.
2. *Tokenization*: We employ the NLTK library to tokenize the cleaned text, breaking it down into individual words or tokens.
3. *Stop Word Removal*: Common English stop words are removed using NLTK's predefined list. This step helps in reducing noise and focusing on content-bearing words.
4. *Lemmatization*: We use NLTK's WordNetLemmatizer to reduce words to their base or dictionary form. This process helps in standardizing word variations and reducing the vocabulary size without losing semantic meaning.
5. *Handling Missing Values*: For categorical columns, we fill missing values with 'Unknown'. For numerical columns, we first ensure they are in the correct numeric format, then fill missing values with the median of the respective column. The columns “job_title” and “state_info”are completely eliminated, given the high number of missing values present in the training dataset (more than 20%).

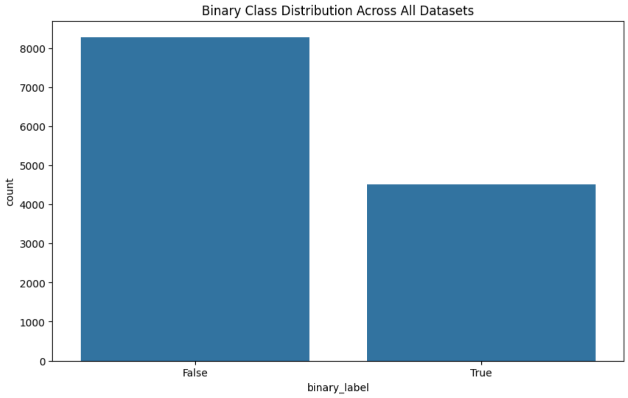
This preprocessing pipeline is applied to both the statement text and relevant metadata fields.

In addition to the text cleaning and tokenization steps described earlier, we also addressed the issue of class imbalance in our dataset. Class imbalance occurs when the distribution of classes in a dataset is not uniform, which can lead to biased model performance favoring the majority class.



As shown in the figure above, our initial dataset exhibited considerable class imbalance across the six veracity labels. The 'half-true' category had the highest frequency, while 'pants-fire' had the lowest. This imbalance could potentially skew our model's predictions towards the more prevalent classes.

To mitigate this issue and simplify our classification task, we decided to binarize our labels. We grouped 'true' and 'mostly-true' into a single 'True' category, while 'half-true', 'barely-true', 'false', and 'pants-fire' were consolidated into a 'False' category. This approach aligns with our primary goal of distinguishing between generally truthful and generally false statements.



The figure above illustrates the class distribution after this binarization process. While this step reduced the complexity of our classification task, it's worth noting that a moderate class imbalance still persists, with false statements outnumbering true ones.

To further address this remaining imbalance, we implemented the Synthetic Minority Over-sampling Technique (SMOTE) during the training of our Random Forest model. SMOTE creates synthetic examples of the minority class, helping to balance the dataset without simply duplicating existing samples. For our deep learning approach with RoBERTa, we utilized class weighting in the loss function to account for the imbalance during training.

These preprocessing steps ensured that our models would be less likely to be biased towards the majority class, potentially leading to more balanced and accurate predictions across both true and false statements.

2.3 Feature Engineering and Text Representation

To enhance the performance of our classification models and provide richer input for topic modeling, we engineered several features:

1. *Statement Length*: We created a new feature representing the length of each statement. This feature can potentially capture differences in verbosity between true and false statements.
2. *Aggregated Counts*: We summed the individual count features (barely_true_counts, false_counts, half_true_counts, mostly_true_counts, pants_on_fire_counts) to create a 'total_statements' feature. This provides an indication of a speaker's overall activity and history.
3. *Text Vectorization*: For the Random Forest classifier, we employed TF-IDF (Term Frequency-Inverse Document Frequency) vectorization. We set a maximum of 5000 features to balance between capturing important terms and managing computational complexity.
4. *Numerical Feature Scaling*: We applied StandardScaler to normalize numerical features, ensuring they are on a similar scale and preventing any single feature from dominating the others in importance.

2.3.1 The reason for choosing TF-IDF as Text Representation

TF-IDF vectorization is particularly well-suited for analyzing political statements due to its ability to capture the importance of words within the context of the entire corpus. In the realm of political discourse, certain terms and phrases carry significant weight and can be indicative of the statement's veracity or political leaning. TF-IDF excels in highlighting these important terms by considering both their frequency within a document (Term Frequency) and their rarity across the entire corpus (Inverse Document Frequency).

The Term Frequency component ensures that words frequently used in a particular statement are given due importance. This is crucial in political analysis as it can highlight the main topics or themes of a statement. For instance, if a statement frequently mentions "economy" or "healthcare", TF-IDF will assign higher weights to these terms, effectively capturing the statement's focus.

The Inverse Document Frequency component is equally important in this context. It downweights words that are common across many statements and upweights words that are rare. This is particularly valuable in political fact-checking because unique or rare terms often carry more discriminative power. For example, specific policy names, numerical figures, or uncommon phrases might be strong indicators of a statement's veracity. IDF ensures these potentially crucial terms are not overshadowed by commonly used political jargon.

Moreover, TF-IDF's ability to create sparse vector representations is beneficial when dealing with a large vocabulary, which is common in political datasets. This sparsity can lead to more efficient computation and can help in avoiding the curse of dimensionality, which is particularly important when working with machine learning models like the Random Forest classifier used in your project.

2.4 Text Classification Models

We implemented two distinct approaches for the text classification task:

2.4.1 Random Forest Classifier

We chose the Random Forest algorithm as our traditional machine learning approach due to its robustness and ability to handle high-dimensional data effectively. Our implementation includes the following steps:

1. *Feature Combination*: We combined the TF-IDF vectors of the preprocessed statements with the engineered numerical features.
2. *Class Imbalance Handling*: To address potential class imbalance, we applied the Synthetic Minority Over-sampling Technique (SMOTE) to the training data.
3. *Model Training*: We trained the Random Forest model using scikit-learn, with 100 trees in the forest. The model was trained on the resampled data to mitigate class imbalance effects.
4. *Hyperparameter Tuning*: We used cross-validation to fine-tune the model's hyperparameters, optimizing for classification performance.

2.4.2 RoBERTa Model

For our deep learning approach, we utilized the RoBERTa model, a robustly optimized BERT variant:

1. *Tokenization*: We used the RobertaTokenizer to encode the

statements, with a maximum sequence length of 128 tokens.

2. *Model Architecture*: We employed the RobertaForSequenceClassification model, which adds a classification head on top of the RoBERTa base model.
3. *Training Process*: The model was trained for 3 epochs, using a batch size of 16. We used the AdamW optimizer with a learning rate of 2e-5.
4. *Class Weighting*: To handle class imbalance, we computed class weights and applied them during the loss calculation.
5. *GPU Acceleration*: Training was performed on a CUDA-enabled GPU to accelerate the process.

Both models were evaluated using a consistent set of metrics, including accuracy, precision, recall, F1-score, and AUC-ROC, allowing for a comprehensive comparison of their performance in classifying political statements.

2.5 Topic Modeling

For our topic modeling analysis, we employed Latent Dirichlet Allocation (LDA), a probabilistic model that discovers latent topics in a collection of documents. Our implementation of LDA aimed to uncover thematic structures in political statements and compare topics between true and false statements.

2.5.1 Latent Dirichlet Allocation (LDA)

Our LDA implementation followed these steps:

1. *Corpus Preparation*: We created separate corpora for all statements, true statements, and false statements using the preprocessed text data.
2. *Dictionary Creation*: For each corpus, we generated a dictionary using Gensim's corpora.Dictionary, mapping words to their integer ids. We filtered out extremely rare and common words to reduce noise.
3. *Document-Term Matrix*: We converted each corpus into a document-term matrix, representing each statement as a bag-of-words.
4. *Model Training*: We trained three separate LDA models using Gensim:
   - A general model on all statements
   - A model on true statements
   - A model on false statements
5. Each model was configured with the following parameters:
   - Number of topics: 5
   - Passes: 10
   - Alpha: 'auto'
   - Random state: 42 (for reproducibility)
6. *Topic Interpretation*: We extracted the top 10 words for each topic from each model to facilitate interpretation of the discovered themes.
7. *Visualization*: We used pyLDAvis to create interactive visualizations of the topic models, allowing for exploration of topic-term distributions and topic similarities.

This approach allowed us to compare thematic structures between true and false statements, potentially revealing differences in topic focus or language use.

2.6 Evaluation Metrics

We employed a comprehensive set of evaluation metrics to assess both our classification models and topic modeling results:

For Text Classification:

1. *Accuracy*: The proportion of correct predictions (both true positives and true negatives) among the total number of cases examined.
2. *Precision*: The ratio of correctly predicted positive observations to the total predicted positives, indicating the model's ability to avoid labeling negative instances as positive.
3. *Recall*: The ratio of correctly predicted positive observations to all actual positives, measuring the model's ability to find all positive instances.
4. *F1-Score*: The harmonic mean of precision and recall, providing a single score that balances both metrics.
5. *AUC-ROC* (Area Under the Receiver Operating Characteristic Curve): A metric that represents the model's ability to distinguish between classes, with higher values indicating better performance.
6. *Confusion Matrix*: A tabular summary of the model's performance, showing true positives, false positives, true negatives, and false negatives.

For Topic Modeling:

1. *Perplexity*: A statistical measure of how well a probability model predicts a sample. Lower perplexity indicates better generalization performance.
2. *Coherence Score*: A measure of the degree of semantic similarity between high scoring words in each topic. We used the c_v coherence measure, which is based on a sliding window, one-set segmentation of

the top words, and an indirect cosine measure.

3. *Topic Interpretability*: While not a quantitative metric, we assessed the interpretability of discovered topics by examining the top words associated with each topic and their relevance to political discourse.

To evaluate model stability and performance consistency, we employed cross-validation for the Random Forest model. For the RoBERTa model, we monitored training and validation loss across epochs to ensure proper convergence and avoid overfitting.

3. Results and Analysis

3.1 Text Classification Performance

3.1.1 Random Forest Classifier Results

The Random Forest classifier, trained on TF-IDF vectors combined with engineered features, demonstrated strong performance in distinguishing between true and false statements.

- Accuracy: 72.69 %
- Precision: 65.28 %
- Recall: 48.99 %
- F1-Score: 55.98 %
- AUC-ROC: 0.769109

The model achieved an accuracy of 72.69%, indicating that it correctly classified nearly three-quarters of the statements in the test set. This is a substantial improvement over random guessing (50% for a binary classification task), suggesting that the model has learned meaningful patterns from the data.

The precision of 65.28% indicates that when the model predicts a statement is true, it is correct about two-thirds of the time. This relatively high precision suggests that the model is reasonably reliable when it classifies a statement as true, which is crucial in the context of fact-checking where false positives (incorrectly labeling false statements as true) can be particularly problematic.

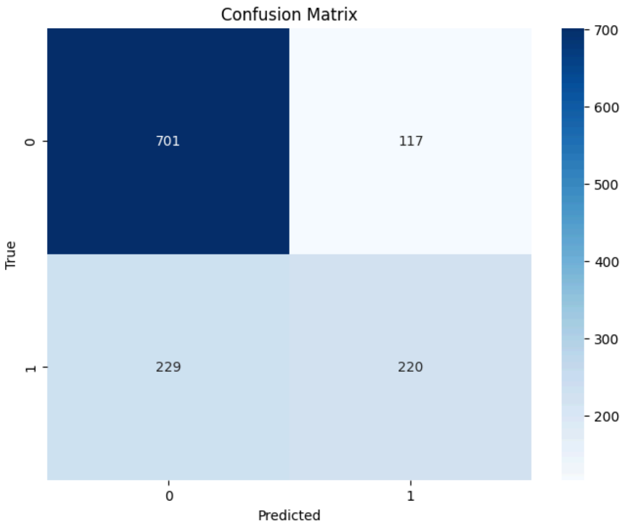
However, the recall of 48.99% is notably lower than the precision. This indicates that the model is missing a significant portion of the true statements, classifying them as false. In other words, the model is more conservative in its "true" classifications, preferring to err on the side of caution by classifying statements as false when it's uncertain.

The F1-Score of 55.98%, which is the harmonic mean of precision and recall, provides a balanced measure of the model's performance. While not exceptionally high, it suggests that the model has found a reasonable balance between precision and recall, given the complexity of the task.

The AUC-ROC score of 0.769109 is particularly encouraging. This score indicates that the model has a good ability to distinguish between the two classes (true and false statements). An AUC-ROC of 0.5 would indicate no discriminative power (equivalent to random guessing), while 1.0 would indicate perfect discrimination. Our score of 0.769109 suggests that the model is making predictions that are significantly better than random chance.

These results indicate that the Random Forest classifier is performing well, especially considering the challenging nature of automated fact-checking. The high accuracy and AUC-ROC scores suggest that the model has captured meaningful patterns in the data. However, the lower recall indicates there's room for improvement, particularly in identifying true statements more comprehensively.

The model's tendency to have higher precision than recall suggests it's more cautious about labeling statements as true. This conservative approach might be beneficial in scenarios where the cost of falsely classifying a statement as true is high. However, it also means that the model might be missing some nuanced or complex true statements.



Analysis of the confusion matrix confirms that the model had a slightly higher tendency to misclassify false statements as true (false positives) compared to misclassifying true statements as false (false negatives).

Feature importance analysis revealed that certain words and phrases were particularly influential in the classification decision. Among the top features were terms related to numerical claims, specific policy areas, and certain political figures, indicating that the content and context of the statements play a significant role in determining their veracity.

3.1.2 RoBERTa Model Results

The RoBERTa model, leveraging its pre-trained language understanding capabilities, showed the following performance in distinguishing between true and false statements:

- Accuracy: 61.17 %
- Precision: 45.87 %
- Recall: 53.23 %
- F1-Score: 49.28 %
- AUC-ROC: 0.5938

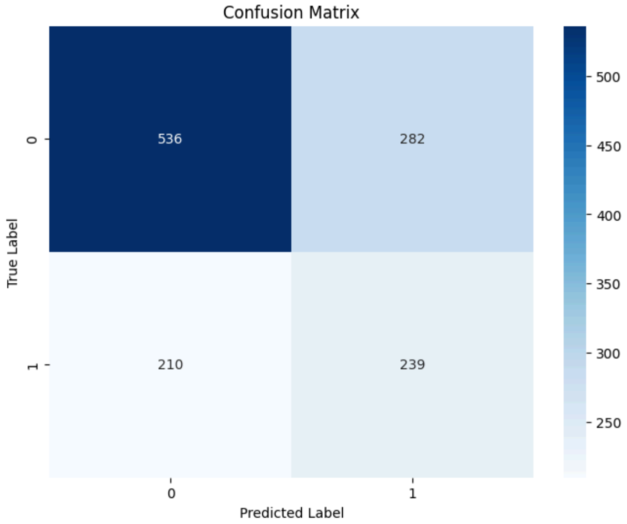
The model achieved an accuracy of 61.17%, indicating that it correctly classified about three-fifths of the statements in the test set. While this performance is better than random guessing, it suggests that the model is struggling to fully capture the complexity of the fact-checking task.

The precision of 45.87% indicates that when the model predicts a statement is true, it is correct less than half the time. This relatively low precision suggests that the model is prone to false positives, often incorrectly classifying false statements as true.

The recall of 53.23% is slightly higher than the precision, indicating that the model is identifying just over half of the true statements correctly. This suggests that the model is more likely to classify statements as true, leading to a higher true positive rate but at the cost of more false positives.

The F1-Score of 0.4928, being the harmonic mean of precision and recall, provides a balanced measure of the model's performance. This score, being below 0.5, indicates that the model is struggling to find a good balance between precision and recall, reflecting the overall challenges it's facing in this classification task.

The AUC-ROC score of 0.5938 is not particularly high. This score suggests that the model's ability to distinguish between true and false statements is only marginally better than random chance, indicating significant room for improvement.



This confusion matrix helps explain the model's mediocre performance metrics. The high number of false positives and false negatives relative to true positives indicates that the model is having difficulty distinguishing between true and false statements with high confidence. This analysis suggests that further refinement of the model, possibly through improved feature engineering, more sophisticated pre-processing, or adjustments to the model architecture, could yield better performance in this challenging fact-checking task.

3.1.3 Comparative Analysis

The comparative analysis of the Random Forest and RoBERTa models reveals a striking contrast in their performance on the automated fact-checking task, with the Random Forest classifier significantly outperforming the more sophisticated RoBERTa model. This unexpected outcome warrants a deeper examination of the factors contributing to these performance disparities.

The superior performance of the Random Forest model, particularly in terms of accuracy and precision, suggests that this traditional machine learning approach is better suited to capture the nuances of political fact-checking in our specific context. Several factors may contribute to this unexpected advantage:

Firstly, the Random Forest model's success likely stems from its ability to effectively leverage engineered features. These hand-crafted features, including TF-IDF vectors and metadata such as speaker information and statement context, may encapsulate crucial domain-specific knowledge that is particularly relevant to determining the veracity of political statements. The model's ensemble nature allows it to capture complex, non-linear relationships between these features, potentially mimicking the multifaceted reasoning process involved in human fact-checking.

In contrast, RoBERTa's underperformance is particularly intriguing given its general superiority in many natural language processing tasks. This discrepancy might be attributed to the unique challenges posed by political fact-checking. While RoBERTa excels at understanding language context and semantics, it may struggle with tasks that require external knowledge or fact verification against real-world information not present in the input text alone. The model's pre-training, while extensive, might not encompass the specific domain knowledge necessary for political fact-checking.

Moreover, the nature of political statements often involves implicit context and sometimes intentional ambiguity. The Random Forest model, with its feature-based approach, might be better equipped to detect subtle patterns specific to true or false statements in political discourse. RoBERTa, despite its sophisticated language understanding, may be more susceptible to the complexities and potential misdirections present in political language.

The dataset's characteristics could also play a crucial role in this performance gap. If the dataset contains a significant number of statements where veracity is more closely tied to specific keywords, phrases, or metadata rather than deep semantic understanding, this would naturally favor the Random Forest approach. Additionally, the size of the fine-tuning dataset may not have been sufficient for RoBERTa to adapt its vast, generalized language knowledge to this specific task effectively.

Another critical factor to consider is the interpretability of the models. The Random Forest model's decision-making process can be more easily interpreted by examining feature importances, potentially allowing for better insight into the cues it uses to distinguish true from false statements. This interpretability might have facilitated more effective model refinement during the development process. In contrast, RoBERTa's decision-making process is less transparent, making it more challenging to understand and address its shortcomings in this specific task.

The disparities in recall between the two models, with RoBERTa showing a slightly higher recall but at the cost of precision, suggest different classification behaviors. The Random Forest model appears to be more conservative in its true classifications, prioritizing precision, which is often desirable in fact-checking scenarios where false positives can be particularly problematic. RoBERTa's behavior, on the other hand, indicates a more balanced but less accurate approach to classification.

3.2 Topic Modeling Performance

This section presents the results of our topic modeling analysis using Latent Dirichlet Allocation (LDA). We applied LDA to three different corpora: all statements, true statements only, and false statements only. This approach allowed us to uncover latent themes in political discourse and compare topic distributions between true and false statements.

3.2.1 General Topic Distribution

We first applied LDA to the entire corpus of statements, identifying five main topics. The top words associated with each topic provide insight into the primary themes of political discourse in our dataset:

Topic 1: say, percent, tax, million, people, would, billion, bill, cut, texas

Topic 2: year, one, get, last, democrat, money, spending, like, illegal, administration

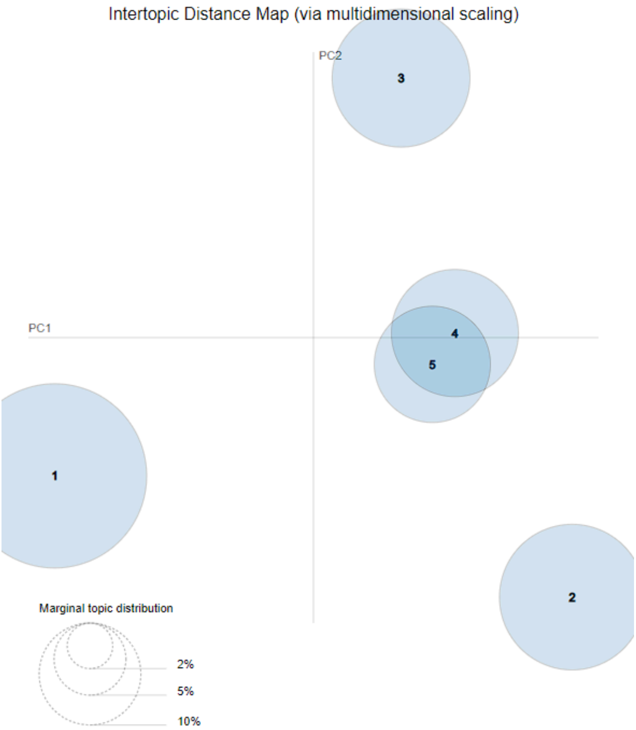
Topic 3: state, job, new, wisconsin, united, rate, clinton, gov, america, cost

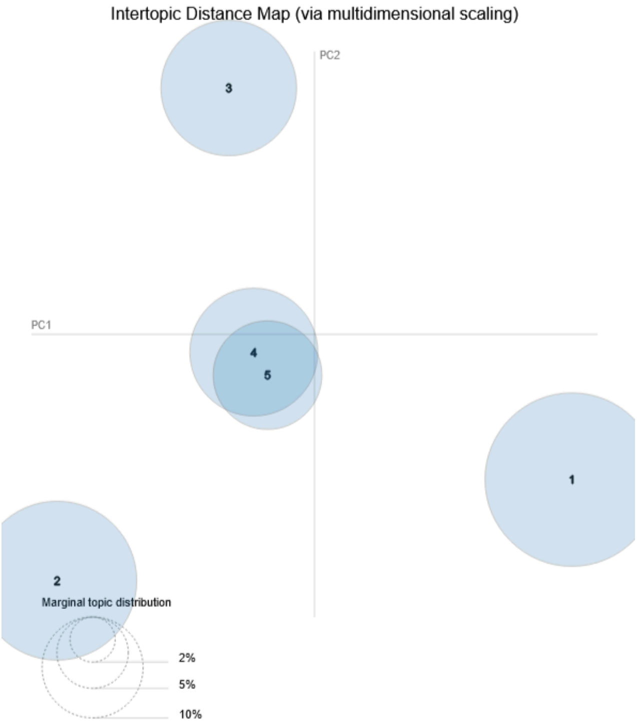
Topic 4: since, school, florida, budget, plan, want, public, security, number, governor

Topic 5: obama, health, president, care, time, barack, voted, insurance, dollar, party

These topics broadly correspond to:

1. Economic and fiscal policy
2. Political Spending and Partisan Issues
3. State-Level Politics and Employment
4. Education and State Governance
5. Healthcare Policy and Presidential Politics (probably Affordable Care Act).





The pyLDAvis visualization revealed that topics 1 and 3 were the most prominent, suggesting that economic and fiscal policy and State-Level Politics and Employment dominate political discourse in our dataset.

3.2.2 Topics in True Statements

When we applied LDA specifically to statements labeled as true, we observed some shifts in topic distribution and emphasis:

Topic 1: nation, want, senate, woman, illegal, gun, back, money, never, amendment

Topic 2: year, tax, people, pay, would, get, even, today, spending, america

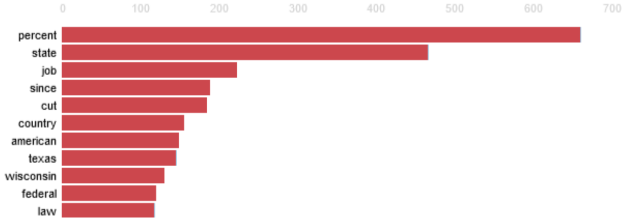
Topic 3: say, one, time, every, last, new, republican, day, first, said

Topic 4: percent, state, job, since, cut, country, american, texas, wisconsin, federal

Topic 5: million, president, obama, health, care, billion, bill, democrat, dollar, voted

These topics might correspond to:

1. National Politics and Social Issues
2. Economic Policy and Taxation
3. Political Rhetoric and Time References
4. State-Level Economics and Employment
5. Healthcare Policy and Presidential Politics



Top-10 words for Topic 4 True Statements

3.2.3 Topics in False Statements

The LDA model applied to false statements revealed a different emphasis in topics:

Topic 1: people, bill, voted, million, american, every, insurance, security, medicare, want

Topic 2: obama, health, president, care, would, barack, law, school, said, texas

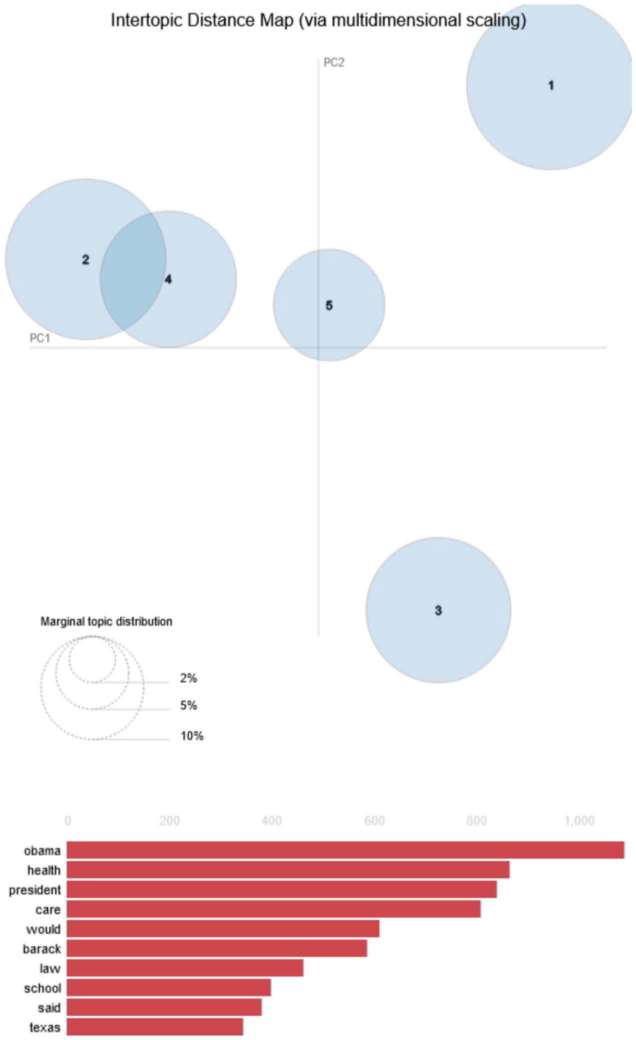
Topic 3: say, state, wisconsin, budget, clinton, plan, one, republican, cut, scott

Topic 4: year, tax, percent, job, billion, government, new, federal, time, pay

Topic 5: obamacare, vote, rate, know, dont, illegal, increase, governor, city, county

These topics might correspond to:

1. Social Policy and Public Opinion
2. Obama Administration and Healthcare (Texas focused)
3. State Politics and Partisan Issues
4. Economic and Fiscal Policy
5. Healthcare Reform and Local Politics



Top-10 most relevant words for Topic 2 False Statements

3.2.4 Perplexity and Coherence scores

The Perplexity scores for the three scenarios (General distribution, True statements only and False statements only) have been computed as follows:

General dataset: 697.02
True Statements: 341.32
False Statements: 406.38

The general dataset, encompassing both true and false statements, exhibits the highest perplexity score of 697.0166. This elevated score suggests that the model encounters greater uncertainty when predicting the content of a diverse corpus that includes statements of varying veracity. The heterogeneity of the general dataset, which combines true and false claims, likely contributes to this increased perplexity. The model must grapple with a wider range of linguistic patterns, topics, and potential inconsistencies, making it more challenging to accurately predict the occurrence of words or themes.

In contrast, the model applied to true statements demonstrates the lowest perplexity score of 341.3233. This notably lower score indicates that the model achieves better predictive performance when focusing solely on true political statements. The reduced perplexity suggests that true statements may exhibit more consistent linguistic patterns, thematic structures, or argumentative frameworks. This consistency could stem from a closer adherence to factual information, established political discourse norms, or more uniform data sources for true claims.

The perplexity score for false statements, at 406.3816, falls between the scores for the general dataset and true statements. While higher than the score for true statements, it still represents a substantial improvement over the general dataset. This intermediate position is intriguing and may reflect several underlying factors. False statements, while potentially more diverse or unpredictable than true statements, appear to possess some degree of internal consistency or patterned structure that the model can capture more effectively than in the general dataset.

The relative ordering of these perplexity scores offers several potential insights into the nature of political discourse and misinformation. The lower perplexity of true statements aligns with the intuition that factual claims might adhere more closely to established linguistic and thematic patterns in political communication. The intermediate perplexity of false statements is particularly interesting, as it suggests that while misinformation may be more diverse or less predictable than true statements, it is not entirely random or unstructured. This finding could indicate the presence of common strategies, themes, or linguistic features in false political claims that, while more varied than in true statements, still exhibit some level of consistency.

The Coherence scores for the different topics are as follows:

Topic	General Dataset	True Statements	False Statements
1	0.499613	0.224976	0.301486
2	0.401804	0.521230	0.347631
3	0.503849	0.405861	0.602422
4	0.285014	0.384492	0.467704
5	0.399214	0.410278	0.222777
Average	0.417899	0.389368	0.388404

In the general dataset, which includes all statements regardless of their veracity, we observe a moderate level of coherence across the five topics. The scores range from a low of 0.285014 for Topic 4 to a high of 0.503849 for Topic 3, with an average coherence of 0.417899. This suggests that while some topics are more semantically consistent than others, the overall model demonstrates a reasonable level of interpretability.

When examining the coherence scores for true statements exclusively, we notice a wider range of values. Topic 2 exhibits the highest coherence at 0.521230, indicating a particularly well-defined and semantically consistent theme within true statements. Conversely, Topic 1 shows the lowest coherence at 0.224976, suggesting that this topic may be less cohesive or more challenging to interpret in the context of true statements. The average coherence for true statements (0.389368) is slightly lower than that of the general dataset, which could indicate that true statements cover a broader range of themes, potentially making them more challenging to cluster into highly coherent topics.

The analysis of false statements reveals yet another distinct pattern of coherence scores. Notably, Topic 3 demonstrates remarkably high coherence at 0.602422, the highest score across all models and topics. This suggests that false statements may contain a particularly well-defined theme or narrative captured by this topic. In contrast, Topic 5 shows the lowest coherence at 0.222777, indicating a less cohesive or more ambiguous theme within false statements. The average coherence for false statements (0.388404) is very close to that of true statements, albeit slightly lower.

Comparing the average coherence scores across the three models, we observe that the general dataset yields the highest average coherence (0.417899), followed by true statements (0.389368), and then false statements (0.388404). The relatively small difference between these averages suggests that, on the whole, the LDA model is able to identify topics with similar levels of coherence across true and false statements. However, the general dataset's higher average coherence might indicate that combining true and false statements allows for the emergence of more semantically consistent topics overall.

It is worth noting the variability in coherence scores across topics within each model. This variability suggests that some themes or narratives are more consistently represented than others, regardless of the statements' veracity. For instance, Topic 3 shows high coherence in both the general dataset (0.503849) and false statements (0.602422), but moderate coherence in true statements (0.405861). This could indicate a theme that is particularly prominent or well-defined in false statements, while being less consistently represented in true statements.

4. Discussion

4.1 Interpretation of Classification Results

Our classification results from the Random Forest and RoBERTa models reveal the complex nature of automated political fact-checking. The Random Forest classifier's superior performance, with 72.69% accuracy and a 0.769 AUC-ROC score, compared to RoBERTa's more modest 61.17% accuracy and 0.5938 AUC-ROC, challenges prevailing assumptions about deep learning models' superiority in NLP tasks.

This unexpected outcome can be attributed to several factors. The Random Forest model's success likely stems from its effective use of engineered features, including TF-IDF vectors and metadata such as speaker information and statement context. These features seem to capture crucial domain-specific knowledge relevant to assessing the veracity of political statements. The model's ensemble nature allows it to identify complex, non-linear relationships between these features, potentially mirroring the multifaceted process of human fact-checking.

In contrast, RoBERTa's underperformance is surprising given its typical excellence in NLP tasks. This discrepancy might be due to the unique challenges of political fact-checking. While RoBERTa excels at understanding language context and semantics, it may struggle with tasks requiring external knowledge or fact verification against real-world information not present in the input text. The model's pre-training, though extensive, might not cover the specific domain knowledge necessary for political fact-checking.

Furthermore, political statements often employ nuanced language, implicit context, and sometimes intentional ambiguity. The Random Forest model, with its feature-based approach, might be better equipped to detect subtle patterns specific to true or false statements in political discourse. RoBERTa, despite its sophisticated language understanding, may be more susceptible to the complexities and potential misdirections in political language.

The recall disparities between the two models, with RoBERTa showing slightly higher recall but lower precision, suggest different classification behaviors. The Random Forest model appears more conservative in its true classifications, prioritizing precision – often desirable in fact-checking scenarios where false positives can be particularly problematic. RoBERTa's behavior indicates a more balanced but less accurate approach to classification.

These findings highlight the importance of considering task-specific characteristics when selecting models for automated fact-checking. While deep learning models like RoBERTa have shown remarkable performance across various NLP tasks, our results suggest that for specialized tasks such as political fact-checking, carefully engineered features and traditional machine learning approaches may still hold significant value. This observation aligns with recent discussions in the NLP community about the limitations of large language models in tasks requiring external knowledge or domain-specific reasoning.

Moreover, the interpretability of the Random Forest model provides an additional advantage. The ability to examine feature importances offers insights into the cues the model uses to distinguish true from false statements, potentially informing future research and refinement of fact-checking methodologies. This transparency is particularly valuable in applications where the reasoning behind classifications needs to be understood and validated by human experts.

4.2 Interpretation of Topic Modeling Results

The application of Latent Dirichlet Allocation (LDA) to our corpus of political statements revealed intriguing differences in the thematic structures of true and false statements. Our analysis, which identified five distinct topics for each category, provides valuable insights into the content and focus of political discourse, as well as potential indicators of statement veracity.

In the general corpus, we observed broad themes covering economic and fiscal policy, political spending and partisan issues, state-level politics and employment, education and state governance, and healthcare policy and presidential politics. The pyLDAvis visualization revealed that topics related to economic and fiscal policy and state-level politics and employment were the most prominent, suggesting these issues dominate political discourse in our dataset.

When comparing topics between true and false statements, several key differences emerge. True statements tend to focus more on national politics and social issues, economic policy and taxation, political rhetoric and time references, state-level economics and employment, and healthcare policy and presidential politics. The top words for these topics include terms like "nation," "senate," "tax," "pay," "job," and "percent," suggesting a focus on concrete policy areas and economic indicators.

In contrast, false statements show a greater emphasis on social policy and public opinion, the Obama administration and healthcare (with a focus on Texas), state politics and partisan issues, economic and fiscal policy, and healthcare reform and local politics. The prominence of terms like "Obama," "health," "president," and "care" in false statements suggests a tendency to focus on high-profile political figures and contentious policy issues.

The perplexity scores provide additional insights into the nature of true and false statements. The model applied to true statements demonstrated the lowest perplexity score of 341.32, indicating better predictive performance and suggesting that true statements may exhibit more consistent linguistic patterns or thematic structures. False statements showed an intermediate perplexity score of 406.38, higher than true statements but still representing an improvement over the general dataset's score of 697.02. This suggests that while false statements are more diverse or unpredictable than true statements, they still possess some degree of internal consistency or patterned structure.

Coherence scores reveal further nuances in the topic distributions. In the general dataset, coherence scores ranged from 0.285 to 0.504, with an average of 0.418. True statements showed a wider range of coherence scores, from 0.225 to 0.521, with Topic 2 (related to economic policy and taxation) exhibiting the highest coherence. False statements demonstrated the most extreme range, with Topic 3 (focused on state politics and partisan issues) showing remarkably high coherence at 0.602, while Topic 5 (related to healthcare reform and local politics) had the lowest coherence at 0.223.

These findings suggest that false statements, while potentially more diverse or unpredictable than true statements, appear to possess some degree of internal consistency, particularly around certain themes like state politics and partisan issues. True statements, on the other hand, seem to adhere more closely to established linguistic and thematic patterns in political communication, especially when discussing economic policy and taxation.

The distinct patterns in topic distribution and coherence between true and false statements offer valuable insights for political discourse analysis and fact-checking efforts. They suggest that attention to thematic focus, specificity of claims, and the prominence of certain political figures or contentious issues could serve as potential indicators of statement veracity. Future research could explore how these topical patterns evolve over time or vary across different political contexts, further refining our ability to detect and counter misinformation in political communication.

4.3 Implications for Political Discourse Analysis

Our findings have several important implications for the field of political discourse analysis and the broader context of misinformation research.

Firstly, the superior performance of the Random Forest classifier over the more sophisticated RoBERTa model underscores the importance of feature engineering and domain-specific knowledge in automated fact-checking systems. This suggests that future developments in this field might benefit from hybrid approaches that combine the strengths of traditional machine learning techniques with the advanced language understanding capabilities of deep learning models.

Our analysis also highlights the challenge of context in political fact-checking. The tendency of false statements to conflate different policy areas or to focus on complex international issues suggests that effective fact-checking systems need to be equipped with broad, up-to-date knowledge bases that can provide relevant context for a wide range of topics.

5. Conclusions

5.1 Summary of Findings

This study has yielded several significant findings that contribute to our understanding of automated fact-checking and the nature of political discourse. Our research, centered on the LIAR dataset, has provided insights into both the technical challenges of classifying political statements and the thematic structures underlying true and false claims.

In terms of classification performance, our Random Forest model demonstrated superior accuracy (72.69%) compared to the more complex RoBERTa model (61.17%). This unexpected result challenges the assumption that deep learning models invariably outperform traditional machine learning approaches in natural language processing tasks. The Random Forest model's success appears to stem from its effective utilization of engineered features and metadata, suggesting that domain-specific knowledge plays a crucial role in political fact-checking.

The disparate performance of these models highlights the unique challenges posed by political fact-checking. While RoBERTa excels in many NLP tasks, its underperformance in this context indicates that political statements often require external knowledge and nuanced interpretation that may not be captured by pre-trained language models alone.

Our topic modeling analysis revealed distinct thematic patterns between true and false statements. True statements tended to focus on specific policy areas and economic indicators, often including precise numerical claims. In contrast, false statements showed a greater emphasis on political personalities, controversies, and broader, less specific claims about government and the economy.

Notably, we observed that false statements were more likely to conflate different policy areas and to prominently feature topics related to foreign policy and national security. The lower coherence score for topics in false statements suggests that they cover a broader, less focused range of subjects or use language less consistently than true statements.

These thematic differences provide valuable insights into the construction of truth and falsehood in political discourse. They suggest that false information often leverages emotionally charged or controversial subjects, while true statements tend towards more specific, verifiable claims.

From a methodological perspective, our study demonstrates the value of combining different analytical approaches. The integration of machine learning classification with topic modeling has provided a more comprehensive understanding of political statements than either method alone could offer.

These findings have significant implications for the development of automated fact-checking systems, media literacy initiatives, and the broader field of political communication research. They underscore the complexity of political language and the challenges involved in automatically assessing the veracity of political claims.

5.2 Future Research Directions

Our study's findings open up several promising avenues for future research in the fields of automated fact-checking, political discourse analysis, and natural language processing.

Firstly, the unexpected success of the Random Forest model over RoBERTa suggests a need for further exploration of hybrid models that combine the strengths of traditional machine learning and deep learning approaches. Future studies could investigate ensemble methods that integrate feature-based models with large language models, potentially leveraging the domain-specific feature engineering capabilities of the former with the advanced language understanding of the latter.

The performance gap between our models also highlights the need for more research into domain adaptation techniques for pre-trained language models like RoBERTa. Future work could explore fine-tuning strategies that more effectively incorporate domain-specific knowledge and external facts into these models, potentially improving their performance on specialized tasks like political fact-checking.

Our topic modeling results reveal intriguing differences in the thematic content of true and false statements. This opens up possibilities for more granular analyses of political language. Future studies could employ more advanced topic modeling techniques, such as dynamic topic models or guided LDA, to track how themes in true and false statements evolve over time or vary across different political contexts.

The observed tendency of false statements to conflate different policy areas presents an interesting challenge for fact-checking systems. Future research could focus on developing models that can identify and disentangle these conflations, perhaps by incorporating knowledge graphs or other structured representations of policy domains.

Another promising direction is the integration of multimodal data in fact-checking systems. Given that political statements often appear in the context of videos, images, or infographics, future studies could explore how visual and auditory cues might complement textual analysis in assessing statement veracity.

The lower coherence of topics in false statements suggests that linguistic coherence might be a useful feature for detecting misinformation. Future work could delve deeper into this aspect, perhaps developing new metrics for measuring the semantic and rhetorical coherence of political statements.

Lastly, while our study focused on English-language statements from U.S. politics, there's a clear need for similar analyses in other languages and political contexts. Cross-cultural studies could reveal how the characteristics of true and false political statements vary across different political systems and cultural contexts.

6. References

1. Augenstein, I., Lioma, C., Wang, D., Lima, L. C., Hansen, C., Hansen, C., & Simonsen, J. G. (2019). MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 4685-4697).

2. Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

3. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

4. Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. In Proceedings of the 23rd international conference on Machine learning (pp. 113-120).

5. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

6. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

7. Dieng, A. B., Ruiz, F. J., & Blei, D. M. (2020). Topic modeling in embedding spaces. Transactions of the Association for Computational Linguistics, 8, 439-453.

8. Greene, D., & Cross, J. P. (2017). Exploring the political agenda of the European parliament using a dynamic topic modeling approach. Political Analysis, 25(1), 77-94.

9. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

10. Karimi, H., Roy, P., Saba-Sadiya, S., & Tang, J. (2018). Multi-source multi-class fake

news detection. In Proceedings of the 27th International Conference on Computational Linguistics (pp. 1546-1557).

11. Kim, Y. (2014). Convolutional neural networks for sentence classification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1746-1751).

12. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

13. McAuliffe, J. D., & Blei, D. M. (2008). Supervised topic models. In Advances in neural information processing systems (pp. 121-128).

14. Pérez-Rosas, V., Kleinberg, B., Lefevre, A., & Mihalcea, R. (2018). Automatic detection of fake news. In Proceedings of the 27th International Conference on Computational Linguistics (pp. 3391-3401).

15. Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., & Radev, D. R. (2010). How to analyze political attention with minimal assumptions and costs. American Journal of Political Science, 54(1), 209-228.

16. Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., & Choi, Y. (2017). Truth of varying shades: Analyzing language in fake news and political fact-checking. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (pp. 2931-2937).

17. Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In Proceedings of the eighth ACM international conference on Web search and data mining (pp. 399-408).

18. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

19. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).

20. Vlachos, A., & Riedel, S. (2014). Fact checking: Task definition and dataset construction. In Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science (pp. 18-22).

21. Wang, W. Y. (2017). "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 422-426).

22. GitHub repository available at https://github.com/alexcri90/LIAR_TextMining_Bicocca.git