Tell IF Fake: Fake News Detection on Twitter

Bartu Koksal[*]
Kaan Yigit[*]
bartuk@acm.org
kyigit@acm.org
University of Illinois at Urbana-Champaign
Champaign, Illinois, USA

</tell_if_fake: Fake News Detection on Twitter (Sorry Elon)>

ABSTRACT

The goal of this project is to be able to identify and classify news in Twitter that are fake without the use of any advanced deep architectures that require lots of compute and resources while not being interpretable such that the algorithm is accessible and understandble by all.

KEYWORDS

Fake News Detection, Machine Learning, Information Retrieval, Topic Modeling

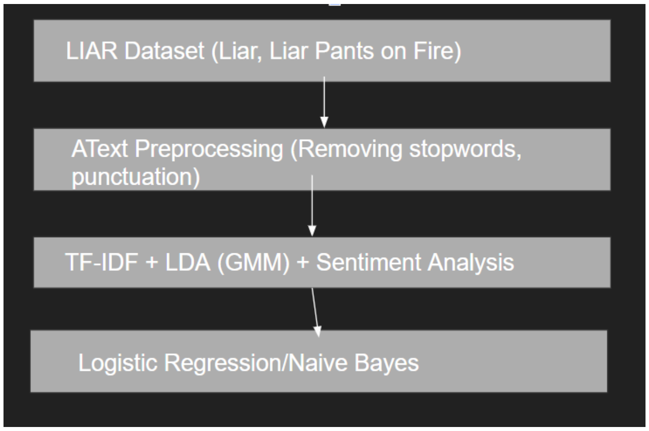
ACM Reference Format:
Bartu Koksal and Kaan Yigit. 2024. Tell IF Fake: Fake News Detection on Twitter . In . ACM, New York, NY, USA, 4 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

The detection of fake news is primarily centered around leveraging sophisticated deep neural network architectures to model complex semantics. However, the implementation of deep neural architectures such as Convolutional Neural Networks and Transformers involves significant computational resources and engineering efforts. These requirements often pose challenges for deployment on accessible platforms like web and mobile, as exemplified by applications such as Twitter. In this project, we explore an alternative approach by utilizing advanced retrieval and ranking methods, including TF-IDF, alongside unsupervised learning techniques such as Latent Dirichlet Allocation and Gaussian Mixture Modeling. Additionally, we incorporate sentiment analysis using a pretrained roBERTa model. By integrating these methods, we aim to classify news articles effectively. Our approach culminates in the application of Logistic Regression on the LIAR dataset, facilitating a streamlined and computationally efficient classification process. dataset to classify the news.

2 PROJECT PIPELINE

As outlined in the Introduction, our project pipeline consists of four key stages: data collection from the LIAR dataset, text preprocessing, integration of TF-IDF ranking with LDA (GMM) and sentiment analysis, and finally, classification using Logistic Regression. Below, we detail each step of the pipeline to highlight the engineering and mathematical challenges addressed in pursuit of our project objectives.


LIAR Dataset (Liar, Liar Pants on Fire)
AText Preprocessing (Removing stopwords, punctuation)
TF-IDF + LDA (GMM) + Sentiment Analysis
Logistic Regression/Naive Bayes

2.1 Data Collection

The initial phase involves sourcing data from the LIAR dataset, a comprehensive collection of labeled statements in the political arena, curated to facilitate the study of misinformation. This dataset provides a robust foundation for training our models and testing their efficacy in distinguishing between truthful and deceptive content.

[*] Both authors contributed equally to this research.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

2.2 Text Preprocessing

In this step, we refine the raw data by removing stop words, punctuation, and other non-informative elements. This normalization is crucial for reducing noise and improving the efficiency of subsequent analytical processes.

2.3 TF-IDF Ranking + LDA (GMM) + Sentiment Analysis

Combining TF-IDF ranking with Latent Dirichlet Allocation (LDA) and Gaussian Mixture Models (GMM) allows us to capture both the statistical and thematic structures within the data. Additionally, sentiment analysis, powered by a pretrained roBERTa model, provides insights into the emotional tone of the texts, which can be indicative of bias or manipulative intent.

2.4 Classification Models

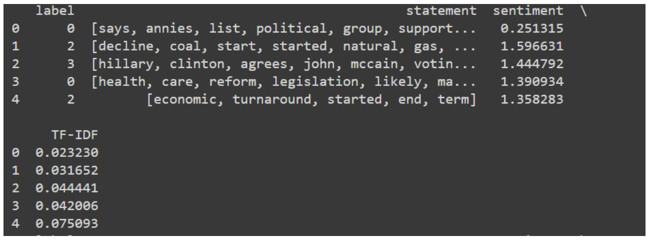
The final stage of our pipeline employs Logistic Regression to classify news articles as either 'true' or 'false'. This method was selected for its effectiveness in binary classification tasks and its interpretability, which is essential for analyzing which features most strongly influence the determination of news authenticity.

3 DATA COLLECTION AND PRE-PROCESSING

We utilized the LIAR dataset, known as "Liar, Liar Pants on Fire," which categorizes Twitter statements with labels ranging from 0 (false) to 5 (pants-fire), including intermediate values such as 1 (half-true), 2 (mostly-true), 3 (true), and 4 (barely-true). For our preprocessing, we simplified this schema by reassigning the label 5 (pants-fire) to 0 (false), because our modeling approach—distinct from more complex deep learning architectures—does not effectively differentiate the nuanced semantic and syntagmatic relationships necessary to discern such specific categories of misinformation. Moreover, we processed our training data using standard preprocessing techniques for this course, which include removing stopwords and punctuation. We chose not to implement stemming, as we believe our modeling approach is capable of discerning some level of nuance among true and false news items.

4 TF-IDF

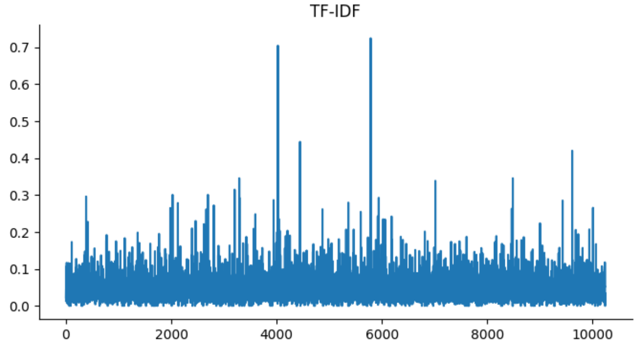
We applied the Term Frequency-Inverse Document Frequency (TF-IDF) technique to weight and normalize words in our dataset, prioritizing words that are unique to individual documents and thus potentially more informative. We've built a vocabulary of the most common 500 words using the statements/tweets, and we've run the each statement/tweet in training set as if queries and taken their TF-IDF score compared to vocabulary. We've added a 'TF-IDF' feature to our table to write the each score to the table. You can see an example of the table from the code snippet as well:



```
   label                                          statement  sentiment  \
0      0  [says, annies, list, political, group, support...   0.251315
1      2  [decline, coal, start, started, natural, gas, ...   1.596631
2      3  [hillary, clinton, agrees, john, mccain, votin...   1.444792
3      0  [health, care, reform, legislation, likely, ma...   1.390934
4      2          [economic, turnaround, started, end, term]   1.358283

     TF-IDF
0  0.023230
1  0.031652
2  0.044441
3  0.042006
4  0.075093
```

These TF-IDF scores served as crucial features in our machine learning models, particularly aiding in differentiating between fake and real news by accentuating the most distinctive terms in each statement. To enhance the accuracy of our fake news detection further, we integrated TF-IDF scores with additional features in an SVM classifier, demonstrating the effectiveness of employing text-specific statistical measures in text classification tasks. Moreover, our implementation includes document length normalization and leverages the BM25Okapi model, which optimizes the relevance calculations for different document lengths and term frequencies, aligning with best practices in information retrieval.



Above is the frequency distribution of TFIDF scoring of the training corpus.

5 LATENT DIRICHLET ALLOCATION (LDA)

Latent Dirichlet Allocation (LDA) is utilized for sophisticated topic modeling, serving as a foundational technique in our approach to creating a nuanced mixture model. This model integrates separate language models for topic analysis, distinctly categorizing news items as true—including categories such as barely true and slightly true—and false. To achieve this differentiation, we have incorporated a lambda factor, $\lambda = 0.7$, to optimally balance the influence of each category in the mixture model.

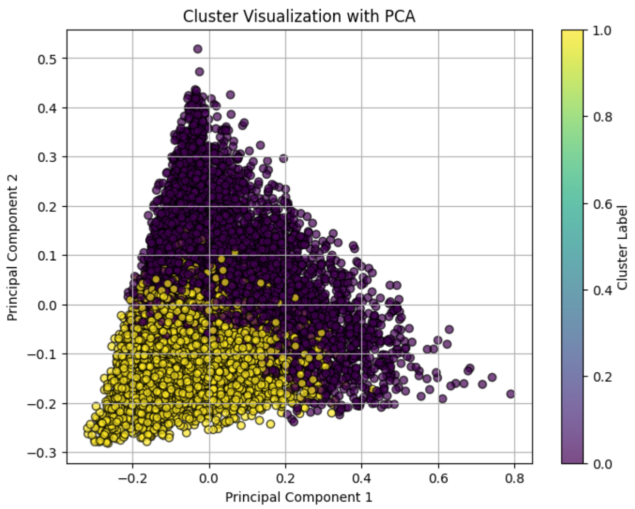
5.1 Implementation of LDA Topic Models

In practical terms, we implemented LDA to construct two differentiated topic models. These models were meticulously blended using the aforementioned lambda factor, ensuring that each retains its distinct characteristics while contributing to a comprehensive understanding of the dataset. This methodological choice allows us to map out the semantic landscapes of both true and fake news effectively.

5.2 Gaussian Mixture Modeling

Following the integration of the topic models, we employed Gaussian Mixture Modeling (GMM) to assign cluster labels to the sentences in our dataset. GMM is particularly adept at identifying latent groupings in data, which enables us to classify the textual content based on its proximity to the characteristics of true or fake news. This classification is pivotal, as it provides a measure of the likelihood that a given piece of news falls into one of the two main categories.

5.3 Visualization with Principal Component Analysis



To visually substantiate our findings, we conducted a Principal Component Analysis (PCA) on the clustered data. This analysis is detailed in the Cluster Analysis section of this paper and offers a graphical representation of the distribution between true and fake news categories. The PCA visualization not only clarifies the clustering effects but also highlights the distinct areas where true and fake news topics converge or diverge, providing further insights into the semantic and thematic structures within the data.

6 UTILIZATION OF THE ROBERTA MODEL FOR SENTIMENT ANALYSIS

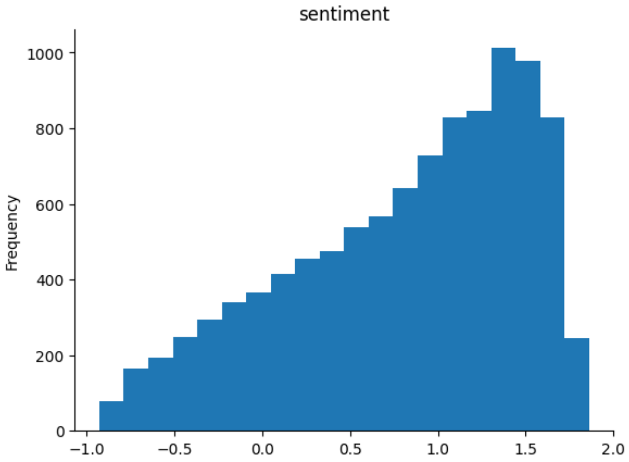
6.1 Deployment of RoBERTa

We deployed RoBERTa, a pre-trained transformer model renowned for its high accuracy in sentiment analysis, to assess the emotional tone behind textual statements in our dataset. This model, developed from the transformer architecture, is particularly adept at understanding and interpreting the nuances of language semantics, making it an ideal tool for evaluating the sentiments conveyed in text.

6.2 Calculation of Sentiment Scores

Sentiment scores were calculated by analyzing the output probabilities for the Negative, Neutral, and Positive categories provided by the RoBERTa model. This method allows for a nuanced quantification of sentiment, reflecting the complex emotional undercurrents present in the statements being analyzed.

6.3 Insights Derived from Sentiment Scores



The sentiment scores derived from RoBERTa were utilized to investigate potential biases in the statements. Specifically, we examined whether extreme sentiments indicated sensationalism or misinformation—factors that are highly relevant in the context of fake news detection. By understanding the sentiment profiles of statements, we can better identify those that may be intentionally skewed to deceive or mislead.

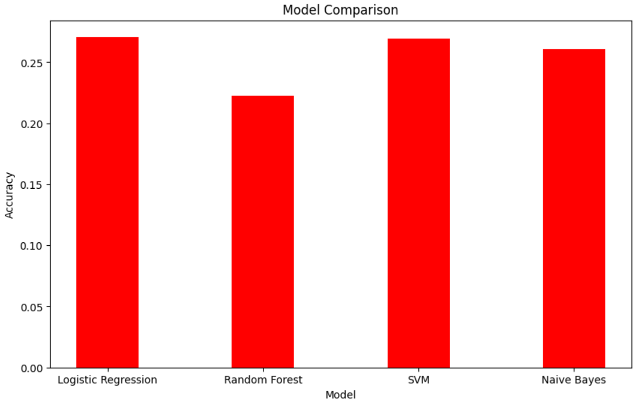
6.4 Augmentation with Synthetic Variables

To enhance our analysis, we augmented the data with a synthetic variable derived from the formula:

$$2 \times \text{Neutrality Rate} - (\text{Positive Rate} + \text{Negative Rate})$$

This formula was devised to provide a consolidated metric that captures the overall sentiment balance of a statement, further enriching our dataset for more robust machine learning modeling and insights.

7 CLASSIFICATION MODELS



article array

Classification Model	Accuracies
Logistic Regression	0.2707
Support Vector Machine	0.2691
Naive Bayes	0.2604
Random Forest	0.2225

Table 1: Accuracies of different classification models

It is seen that our accuracies are low, however, it must be noticed that no deep modeling is done here except the use of pretrained models like roBERTa. We adduce this to the low silhoutte score we've gotten from the Gaussian Mixture Model prediction with 0.12, showing that our unsupervised learning model wasn't successful in learning the parameters/clusters for the individual texts. Further improvement of the unsupervised model of Latent Dirichlet Allocation (LDA) making it closer to a silhoutte score of 1 would make the overall running of our classification models better. Complex engineering and algorithmic feat that we've experimented here costed the accuracy here. However, further improvement of the Gaussian Mixture Model will show in time that our methodology will work and enable a great classification without the use of deep neural networks and architectures that require intensive resources and computation. We believe that the way we've tried to connect different methodologies across different sections of course made our project worthwhile and a conveying of our interest and improving technical proficiency of the course.

8 ACKNOWLEDGMENT

We'd like to thank our Professsor, Prof. Pablo Robles-Granda for all his helps and great insight as a mentor and a teacher who is always interested in the development of his students.

9 REFERENCES

Truică, Ciprian-Octavian, and Elena-Simona Apostol. 2023. "It's All in the Embedding! Fake News Detection Using Document Embeddings" Mathematics 11, no. 3: 508. https://doi.org/10.3390/math11030508

Wang, William Yang. “‘liar, Liar Pants On Fire’: A New Benchmark Dataset for Fake News Detection.” arXiv.Org, 1 May 2017, arxiv.org/abs/1705.00648.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009