Engineering Multifaceted Features for Robust Fake News Classification

Uthkarsh Reddy Junuthula
uthkarshreddyjunuthula@my.unt.edu

Jashia Mitayeegiri
JashiaMitayeegiri@my.unt.edu

Mohammed Abdul Azeem
abdulazeemmohammed@my.unt.edu

Mohammed Sabith Abul Faiz
mohammedsabithabulmohammedsabithabul@my.unt.edu

October 15, 2023

1 Idea Description

The goal of this project is to leverage techniques pertaining to effective feature engineering, Natural Language Processing (NLP) and machine learning to develop a robust system capable of accurately distinguishing between fake and genuine news articles. This initiative is particularly critical given the rising threat of misinformation, which has the power to mislead the public and sway opinions. In summary, the project is committed to building a reliable and versatile tool for the effective identification of fake news.

2 Goals and Objectives

The main goal of the Fake news detection project is to apply advancements in Natural Language Processing (NLP) and machine learning to build a sys-

tem capable of accurately identifying and categorising fake news from news articles.

Before starting the project, it is mandatory to understand what things and features constitute fake news, providing a base framework for the classification process. Data Acquisition here is collecting the news articles from various resources and labelling it as fake or not fake labels. Cleaning,removing stopwords that could be predefined or user defined,tokenizing and standardising the textual data, preparing it for analysis and feature extraction. Using word embeddings or Bert for transforming the textual data to numerical data. Selecting Machine Learning and Deep Learning algorithms based on further research for training purposes. Performing various performance measures in order to validate their reliability and effectiveness. Later fine tuning the model for further proper predictions.

3 Motivation

In this age of social media, we are consuming news at a much higher rate than ever. In our day to day lives, we come across a wide range of news articles, however, the challenge is to analyse which stories are true and which are fabrications. The alarming rise in the broadcasting of fake news poses a grave threat to society at large, influencing public opinion and even shaping political landscapes. The urgency to tackle this issue is further amplified by data revealing a considerable increase in the spread of misinformation in recent years.

4 Significance

Fake news has been prevalent for ages, but we are currently new to online fake news which is virtual and has got no bounds to it as the user is not physically present and no one can be found as the main source of it. It has got high influences on the people and has shifted the locus of interest related to the contemporary issue. The dataset we have chosen will help to detect the rumours and fake news used in the articles. This mainly helps us to gain insight into the fake news and the latent pattern involved in it which promotes us to implement a safer internet space. Analytics will enable us to recognize the vulnerable communities and the influence of fake

news on people and also to protect these communities by providing them with advocacy support. The better detection of algorithms will permit us to capture the facts and the fakes used and authorised to filter content and limit such kind of content generated. We can enforce better community guidelines and effective regulations which comply with various legal policies.

5 Literature Survey

The paper [1] focuses on the extraction and verification of the fake news detection system which involves generation of dataset such that the actual claims are modified, and the system is developed such that it automatically verifies if the claim is modified are not. It has used textual entailment models which help us to determine the logical relationship between the claim and the modified claim. Further classification is done using the support vector machines and random forest. It also uses the rule based heuristic approach for verification of facts. The paper [2] involves the use of convolution neural networks (CNNS) to know the perspective of the articles towards a given topic and use the Recurrent Neural Networks (RNN) to capture the temporal dependencies and sequential patterns in the articles. It has also addressed various challenges of fake news detection which involves the changing nature of the news, and the emerging fake news pattern might make the model obsolete. The [3] deals to detect the fake news using the consistency of the multimodal data which involves many data types and structures, the authors propose that the use of the alignments and coherence of text to uncover the inconsistencies that are present in the fake news. It uses different fusion techniques like early fusion and late fusion to detect combining various sources of information. The paper [4] uses a hybrid approach to tackle the fake news detection called as the CSI which involves two main components: the content which is the textual component and the social interactions which are the social context components. The CNN architecture is used to extract the local features of the text and the LSTM network is used to extract the interactions and patterns of that specific article. Both architectures are fused to form a fusion layer which helps us predict if the news is fake or a fact.

6 Objectives

Data Collection: Use a pre-compiled dataset of tweets with labels. Data Preprocessing: Clean and preprocess the data. Feature Engineering: Apply techniques like TF-IDF, feature scaling, and dimensionality reduction. Modelling: Use multiple machine learning models like Logistic Regression, Random Forest, and Support Vector Classifier or models such as tensor.flow models or neural networks Evaluation: Assess the models using metrics like accuracy, precision, recall, and F1-score.

7 Features

Text and Title Length Word Count in Text and Title Average Word Length in Text and Title Text and Title Variance Word Density TF-IDF Vectorization Sentiment Analysis Score N-grams Analysis Topic Modeling

8 Expected Outcome

The end goal is to attain high accuracy in the fake news detection process. We will analyse various textual attributes, such as length of the article and its title, word count, and even sentiment scores, to offer users a reliable tool to filter out news that they can rely on. In doing so, we aspire to elevate the level of safety on the internet and protect individuals from falling victim to deceptive information.

9 GitHub

The project can be found at https://github.com/Uthkarshh/Feature$_{E}ngineering_{P}roject_{G}roup_{8}$.

References

[1] J. V. A. C. C. & M. A. Thorne, "The fact extraction and verification (FEVER) shared task.," in *23rd Conference on Computational Natural Language Learning*, 2019.

[2] Z. W. Z. & C. H. Zhang, "Deep learning for fake news detection: A comprehensive review.," in *WIREs Data Mining and Knowledge Discovery*, 10(4), e1376, 2020.

[3] Y. W. T. L. S. W. Junxiao Xue, "Detecting fake news by exploring the consistency of multimodal data," *ACM Information Processing and Management*, vol. 58, no. 5, 2021.

[4] N. S. S. & L. Y. Ruchansky, "CSI: A hybrid deep model for fake news detection.," in *IEEE International Conference on Data Mining (ICDM)*, 2020.