CS-GY 6053: FOUNDATIONS OF DATA SCIENCE PROJECT PROPOSAL

A Machine Learning approach to Fake News Detection

Dhanesh Baalaji Srinivasan (**ds7636**)

Vaishnavi Chellappa (**vc2495**)

Venkataramanan Venkateswaran (**vv2265**)

What is the problem?

Detecting and analysing the accuracy of different machine learning and deep learning models on fake news detection, and to increase the efficiency of classifying fake news.

Motivation

In today's age where information is mainstream and available for free, the spread of fake news has become a growing cause for concern in our society. This deliberate misinformation has put individuals and specific communities at risk. As Data Scientists, we are keen on applying the Data Science workflow to solve a ubiquitous problem.

Ideal Outcome

Short-term goal: Achieve significant accuracy on differentiating fake news from real news.

Long-term goal: Create an app that efficiently performs fake news detection and classification on the go, and shows users only real, authentic articles.

How will you learn the background?

While we have established a foundational perception of the fake news detection process, we propose to obtain more domain knowledge by:

1. Clearly articulating the goals and boundaries of the project by understanding the problem we intend to address and the insights we aim to uncover.
2. Gathering and thoroughly examining the dataset so that we can unveil subtle nuances and plausible data-related issues.
3. Exploring academic literature, publications, and previous research pertinent to the project's field and acquiring a comprehensive understanding of the existing knowledge

and methodologies in the domain. For instance, Shu et al. (2017) explored the dynamics of fake news on social media, utilizing data mining techniques to dissect real-time information dissemination. Further, Ma et al. (2015) introduced a novel approach using tree-structured recursive neural networks to detect rumors on Twitter, addressing the unique challenges of analyzing information propagation on this platform.

4. Using fact checking websites like PolitiFact and FactCheck.org to verify news veracity

What kinds of data will you use?

The LIAR dataset was introduced in a 2017 paper titled "Fake News: A Survey of Research, Detection Methods, and Opportunities" by William Yang Wang. It was designed to provide researchers with a substantial and diverse collection of labeled data to help advance the development of fake news detection models. The LIAR dataset has already been split into three classes - train, test and validation. More details of the number of data in each split is given in the table below.

Description of the dataset

Features	Description	Type
Column 1 (ID)	the ID of the statement in the format ([ID].json), used to uniquely identify each record.	Nominal (unique identifier).
Column 2 (Label)	the label which we would use to classify each news. It has 5 values (barely true, false, half true, mostly true, pants on fire)	Ordinal (categorical, with six levels from "True" to "Pants on Fire").
Column 3 (Statement)	the actual news which we would want to classify	Textual data (natural language text).
Column 4 (Subject)	the topic in which the news belongs to	Nominal (categorical, representing the topic).
Column 5 (Speaker)	the speaker involved in the news	Nominal (categorical, identifying the speaker).
Column 6 (Speaker's Job Title)	the speaker's job title.	Nominal (categorical, describing the speaker's job).
Column 7 (State)	The state at which the news was published	Nominal (categorical, U.S. state).
Column 8 (Party Affiliation)	the party affiliation either democrat or republican	Nominal (categorical, political party).
Column 9 (Barely True Counts)	The count of barely true class	Numerical (integer).
Column 10 (False Counts)	The count of false class	Numerical (integer).

Column 11 (Half True Counts)	The count of half true class	Numerical (integer).
Column 12 (Mostly True Counts)	The count of mostly true class	Numerical (integer).
Column 13 (Pants on Fire Counts)	The count of pants on fire class	Numerical (integer).
Column 14 (Context)	Textual context of the statement	Textual data (natural language text).

Dimensions

	Total	Train Data	Test Data	Validation Data
Size	≈12000	10269	1283	1284

Visualization of the data

Train data:

The below graph shows the count of each label across the input texts in train data. We can see that we have a fairly even spread of each outcome.

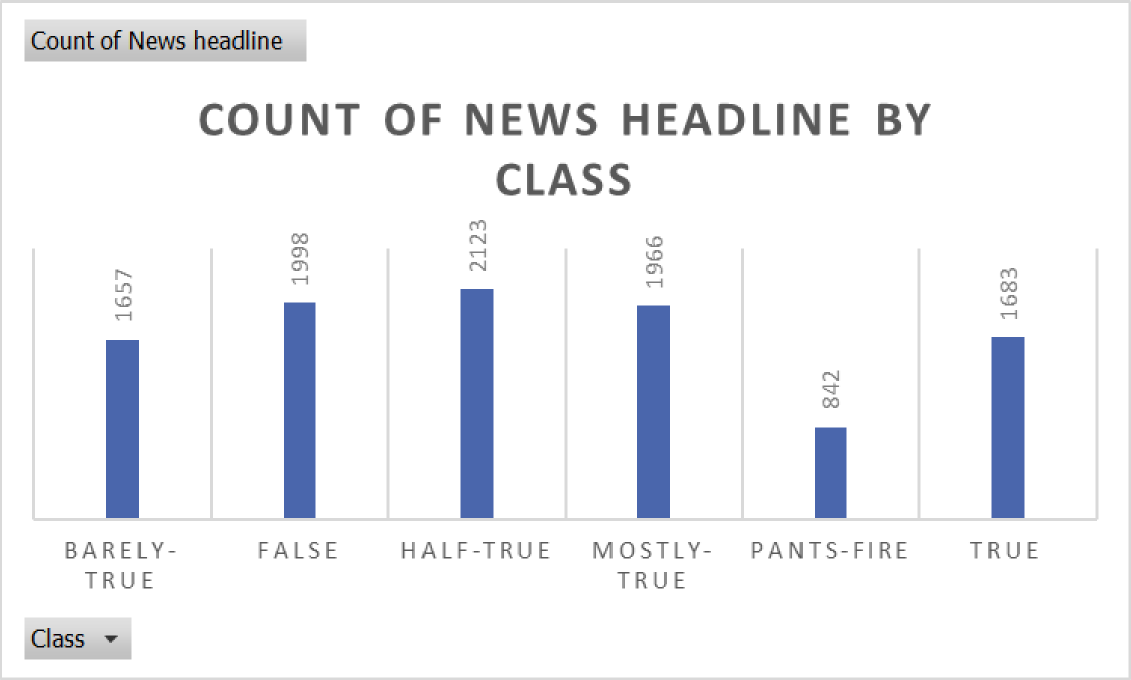

Fig 1: Count of data in each class in train set

Test and validation data:

The below graph shows the count of each label across the input texts in test and validation data. We can see that we have a fairly even spread of each outcome.

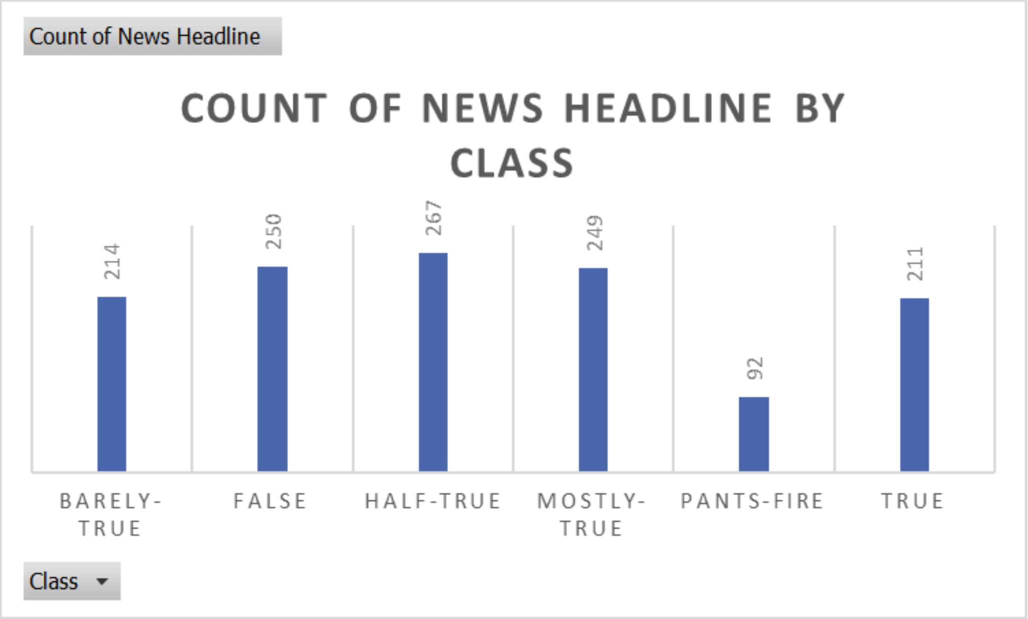

Fig 2: Count of data in each class in test set

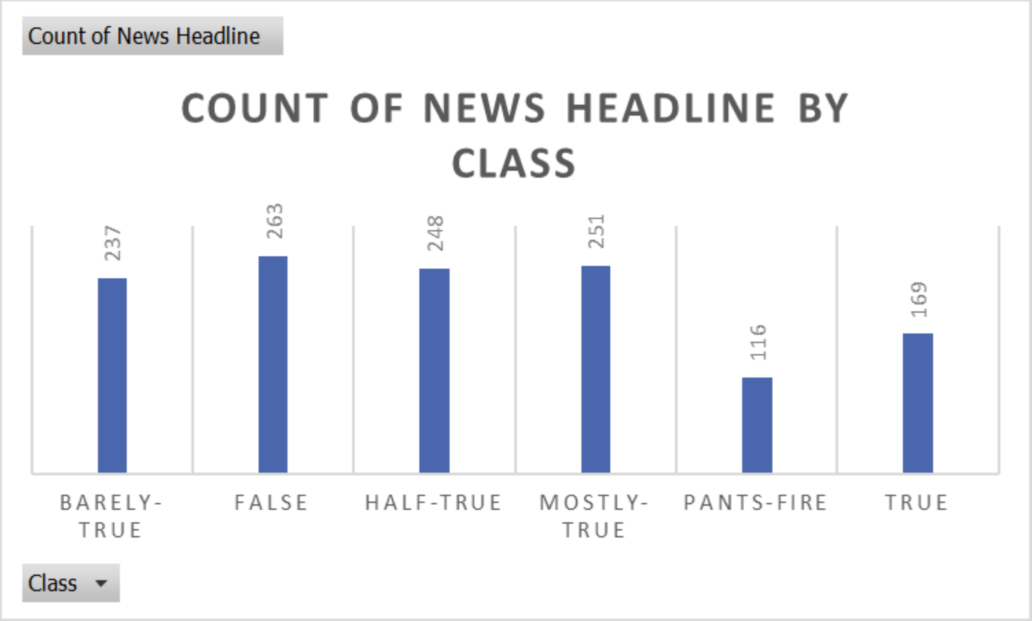

Fig 3: Count of data in each class in validation set

Featuring a well-balanced distribution of data across all possible outcomes, the LIAR dataset is an excellent choice for fake news detection, providing an ideal foundation for model development and evaluation in this domain.

Data Preprocessing

One caveat in the data distribution we observe is that we have very few data for ‘pants-fire’ class. In order to make it an even more evenly distributed dataset, we plan to reassign the ‘pants-fire’ class into the false class.

Exploratory Data Analysis

Text data is often high-dimensional and sparse because any given dataset may only contain a small subset of the entire vocabulary. Furthermore, we will extract insights from the data answering which news domain is prone to fake news articles and identifying words that are most prevalent in fake news articles.

How to process natural language into features for the model?

NLP models traditionally need numerical inputs to learn and classify the data. We plan to use Word2Vec to create word embeddings that can be used as numerical inputs to the machine learning models. The word embeddings contain 1s for the words that are actually present in the input, and 0s for all other words in the corpus.

What kind of models will you build?

Machine Learning

We’ll be using Support Vector Machines (SVM), and Random Forest for the text classification task because they are well-suited to work with textual data. SVMs and Random Forest are capable of capturing non-linear relationships in the data, which can be valuable when dealing with fake news, data in LIAR dataset which has considerable word ambiguity (i.e., many words have multiple meanings, and the correct interpretation often depends on the context) and semantic compositionality, (i.e. the meaning of the sentences is often not a simple linear combination of the meanings of its individual words).

Deep Learning

Deep Learning Models like BERT and LSTM networks can be employed for fake news detection as they can be used to capture complex patterns in text data. Both BERT and LSTM models are adept at handling long-range dependencies in sequences. BERT's attention mechanism and the inherent resistance of LSTMs to the vanishing gradient problem enable them to effectively capture and understand long-term dependencies present in the LIAR dataset. We plan to employ these two models as it is crucial to understand relationships and dependencies that extend throughout the entirety of the news article.

Metrics

Since the LIAR dataset has uniform data across all classes, we can use Accuracy as a measure to evaluate our model. Accuracy will give us a measure of how good our model is in classifying the news article.

We would also be using Precision as it gives the ability of a classification model to identify only the relevant data points. Precision is important as it is crucial for us to not classify real news as fake news and hence we would want our false positive rates to be low.

To get an overall picture of how the model is performing, we plan to use a confusion matrix, which summarizes the results of a classification problem, such as detecting fake news (positive class) and real news (negative class)

What assumptions are safe to make?

Features: The primary features used for fake news detection predominantly revolve around the textual content found in news articles or statements. The core feature typically under scrutiny is the "statement" itself.

We assume the data we get from the LIAR dataset is clean and reliable as the data is obtained from fact-checking websites like PolitiFact and GossipCop and the dataset is provided in a well-defined .tsv file. We assume that the fact-checking websites used in the LIAR dataset are credible and also assume that the dataset has a sufficient distribution of data to build a model that will generalize well to other news articles. This assumption is justified by the fact that the LIAR dataset spans a decade and contains 12.8k records, providing a substantial representation of news over that period. Consequently, we assume that there won’t be any confounding variables.

Problem of overfitting

Within the LIAR dataset, there are a few variables that can influence the classification of whether a news is fake or real. For example, the "Speaker" and their "Job Title" can impact whether a statement is considered true or false because they affect the speaker's credibility and potential biases. Similarly, the "Subject" of a statement can serve as another variable that can create overfitting. Different subject matters may exhibit varying tendencies to contain fake news, and this factor should be accounted for in the analysis. We have to take into account the fact that the dataset may have a lot of fake news from a particular speaker or subject, and the model could overfit on that.