**Alternus Vera (draft)**

Asha Aher     Monica Dommaraju     Swati Narkhede                             Sruthi Chilukuri

San Jose State University     San Jose State University     San Jose State University               San Jose State University

ashabalshiram.aher@sjsu.edu    monica.dommaraju@sjsu.edu    swatiganesh.narkhede@sjsu.edu      srisruthi.chilukuri@sjsu.edu

***Abstract- The sources for news have become so many from online news and social media to news articles and newspapers.In this paper, we as a team would like to put forward a machine learning model to make a prediction on the fake news and differentiate it from the real news based on various aspects. The data has been collected from kaggle and other sources which contained texts along with the speakers and authors; giving us good scope to build a separate model that can classify the truthful news content from the fake news.***

**Keywords**-Data enrichment, Tf-idf, lemmatization, stemming, stop words, tokenization, LDA, Topic Modelling, Cosine Similarity, spell check.

1. INTRODUCTION

The importance of detecting fake news has become prominent in today's world where the outgrowth and abundance of news content is growing day by day. With all the online news, social media, newspapers and articles, it's important to differentiate between the fake news from real news.

1. ABOUT DATASETS
2. Liar-Liar

Liar-Liar Pants on Fire is a multi-class classification dataset having 3 files train, test and valid; has 14 columns: ID, label, statement, subjects, speaker’s job title, state information, etc.

1. Fake News

Fake News dataset is selected from kaggle containing Id, author, text and label.

1. ISOT Fake News

This dataset consists of both Fake news and Real news. Real news was obtained by crawling articles from Reuters. Fake news was collected from unreliable websites flagged by Politifact. These datasets mostly cover topics on World news and political news.

III. DATA CLEANING

Data cleaning and preprocessing of raw contents, containing the following steps:

* Tokenization
* Stemming
* Stop words removal
* Lemmatization
* Spell check

IV. DATASET VISUALIZATION

The Liar Liar dataset has been visualised to understand how the distribution of text has been in the 6 classes and also of how all words have the highest frequency in the dataset. The suitable pairplots, word clouds and bar plots have been generated and inferences were made based on those visualizations.

     V. WHAT WE TRIED

As a part of data amalgamation, We tried to merge the kaggle fake news dataset with the liar-liar dataset using the Label column initially. To do it, We first cleaned the liar-liar dataset and changed it from a six class-multiclass dataset to a binary data set. Later, We realised that merging based on the Y label has not contributed to any enrichment. We later amalgamated using the author column of the dataset as the "speaker" and "author" features in both datasets had common values.

On amalgamated dataset each team member applied one feature followed by  classification and regression algorithms. Comparison accuracy table is listed in section  VII.

VI. FEATURE ANALYSIS:

***FACTOR1: Content Statistics***

 For any text based applications; it is important to understand the statistics or metrics of any content/data that is undergoing processing.

Applications like textual analysis, information retrieval or text summarization use these metrics to better understand the term's frequency , their importance in the given document.

When it comes to exploring data using statistical methods, we use ***Exploratory graphics.***

These are the preliminary steps we apply to the data in hand to:

* feel the data- see their distributions and shape
* check assumptions-and dependencies; correlations
* check if the data assumptions match the methods/models we got to use
* check for anomalies-outliers and unusual distributions and errors
* get suggestions-help us pursue a different angle or different method of analysis

But why do we have to use Exploratory graphics? It is because the graphical representations are information dense and are often the best way to check for shape, gaps or outliers.

Depending on the data, we use univariate distributions like bar charts when there is one variable or swarm plots and scatterplots for multi-variable data.

Similarly, we use box plots for qualitative variables and histograms to detect outliers. scatterplot

Overall, content statistics help us in answering the following questions:

* Do we have what you need?
* Are there clumps or gaps?
* Are there any exceptional cases?
* Are there any errors in data?

When it comes to exploring data using numerical methods, we use ***Exploratory statistics:***

What do we do when we say we are dealing with numerics in the data? We explore it to get what are called the empirical estimates.

In simple words, we manipulate data in many ways like the following:

* Transform variables
* Check sensitivity of results
* Use robust statistics like resampling technique or  cross validation
* Transforming operations like smoothing functions fix skewed dataset

Then comes the ***descriptive statistics.*** To tell the  data story; we use little data to stand in for a lot of data. These include mean, mode, & median to represent the center of the data and range, percentiles, interquartile range, variance and standard deviation to represent the spread of the data.

**Application of the factor:** After performing data amalgamation of the three datasets: Kaggle Fake news, Prolificit News Dataset and the Liar-Liar dataset; I have applied various cleaning techniques like Stemming, Lemmatization and tokenization.

This was followed by feature engineering based on the content statistics factor.

 The lengths of the textual columns like “Content Length” and “Title Length” have been taken to create new features and normalize them.

Later, the median, mean and variance of these textual columns have been calculated to generate new features called “applicability\_text” and “applicability\_title” which is formulated as the following expression which closely represents the standard deviation:

Applicability

 = normalized\_length\*mean(actual\_text\_length)

**Model Evaluation:**

The features that we so generated have been fed to various machine learning algorithms to check how impactful these new features have been on the performance of the model.I have applied Gaussian Naive Bayes, K-NN, Support Vector machines, Decision Tree classifier, XGBoost Classifier & MLP classifier.

As far as evaluation metrics are concerned; Have used cross validation, precision score, recall score, ROC-AUC, F1 score, accuracy and confusion matrix.

**Interpretation of Results:**

XGBoost classifier has given the best results for the given data with an accuracy score of 73%. The Type I and Type II errors generated from the confusion matrix have been the least. Looking at the F1 score; which is a combined representation of both precision and recall is 73% ; showcasing close consistency.

**FACTOR2: Misleading Intentions**

Mis- leading intentions is described as the diverted or false sentence/text i,e misleading someone when you point them in the wrong direction, metaphorically or literally.Below are some examples:

Example: If  you give a stranger/passenger direction away from the place he’s looking for,then you are misleading him.

Predicting whether the sentence/test is misleading or not. This could help to save from the false intentions. Using this feature, we can predict if the given content is honest or dishonest.Below are steps followed.

* Adding labels to the dataset.
* Removing Special Characters and Punctuations.
* Converting characters to Lowercase.
* Removing Stop Words
* Checking for null values and dropping the non-required columns.
* One hot encoding for converting categorical values to numeric values.
* Checking for null values and dropping the non-required columns.
* Lemmatization for grouping together the different inflected forms of a word so they can be analyzed as a single item.
* Stemming for producing morphological variants of a root/base word.

FACTOR3: Writing Style

Writing style of any article depicts many things about the content. We can determine whether the content is True or False.

We can identify fakeness of any article by examining the Sentiments generated by the content. Usually, a fake content contains strong sentiments. True articles have a neutral tone. To identify the fakeness of any article we can examine the grammatical mistakes, Writing quality and professionalism of the article.

Below are the steps followed:

* Cleaning the text
* Stemming
* Lemmatization
* Topic modeling
* Sentiment Analysis

FACTOR 4: Title vs Body

Title vs Body is another major factor which is really useful in identifying if the news is fake or real. Title may be used as a click bait with strong words to lure users into reading the content of the article.

The idea of performing this factor analysis is to correlate the title with the body of the article. This can be done by initially cleaning both the title and text to remove special symbols, punctuations etc, removing stop words and later by performing lemmatization and stemming. Similarities between the title and text are then calculated by performing TF-IDF followed by cosine similarity. TF-IDF is performed to filter the strong words and cosine similarity is then applied for better results.

This similarity score thus can be used as a new feature for classification modelling. Another latent feature can also be extracted by calculating the number of common words between the TF-IDF vectors of title and text. So by performing title vs body factor analysis, we were able to extract two important features which can be used with additional features generated by other factor analysis methods to classify the final dataset.

VII.    Evaluation Matrix:

|  |  |  |
| --- | --- | --- |
| Feature | Classifier | Accuracy |
| Misleading Intent | Random Forest TF-IDF | 86.14% |
| Misleading Intent | Naive Bayes - TF-IDF | 48.44% |
| Misleading Intent | Logistic Regression - TF-IDF | 51.30% |