**Abstract**

By offering a platform for the distribution of knowledge, social networking has taken over the entire planet. Most of the time, people pass along information without verifying its accuracy. Social media platforms are now utilized to influence decisions in a variety of areas, including politics, advertising, and more. It is hardly unexpected that misinformation is being circulated on social media as a means of swaying public opinion.

One methodical and intentional strategy used to persuade people for political or religious ends is propaganda. In this study, machine learning and deep learning algorithms were used to attempt to distinguish between propaganda and non-propaganda text.

**Introduction**

Computer science is crucial in today's technological age for offering answers to practically all new sectors. Computer science has advanced dramatically since the 1970s, when the internet first appeared. It is now used in a variety of multidisciplinary fields, including remote sensing, technical diagnosis, traffic control systems, criminology, medical imaging, image processing, data mining, and automatic surveillance. The market is seeing a huge increase in the number of hardware and software products because of these applications. Today, data analytics is a key area of study for finding patterns in massive data sets. It integrates with several disciplines, including bioinformatics, natural language processing (NLP), machine learning, and others.

Important information is extracted from text, images, videos, and other sources during data mining. Data mining can perform both descriptive and predictive tasks. Descriptive tasks are used to characterize data, while predictive jobs use historical data to estimate the future. Some of the tasks involved in data mining include clustering, correlation, and pattern finding. Online social network analysis is a difficult procedure because of the massive utilization, variety, volume, validity, and real-time data that these networks generate. Online social networks (OSN) communicate using computer-mediated tools that facilitate the creation and dissemination of knowledge, ideas, business benefits, and novel communication techniques through online communities and links.

**Literature Review**

The increasing allure and beauty of using social networks has an impact on our daily lives, whether directly or indirectly, by making us more likely to rely on other people's opinions and suggestions when making small or major decisions, such as whether to buy inexpensive small items or cast an online vote in elections to elect a new government. It is not surprising that social media has evolved into a tool for swaying public opinion by disseminating false information in line with the times. The widespread use of false information and propaganda on social media calls for their detection and opposition. In their investigation of the traits of fraudulent accounts, H. Gao concentrated on the text-based URLs that appear in messages. Due to the rise in social media users, political and government-related events receive the most attention, and social media abuse has become widespread. An approach for identifying and following political misuse in social networks was put forth by J. Ratkiewicz the network structure at the time of the election deviates somewhat from the regular pattern. A. Halu investigated the social network structure throughout the election season. Wide region parties survive by accumulating a finite fraction of the votes at election time, according to a model for opinion dynamics that was provided. N. Ramakrishnan and colleagues researched and mined data from online social networks pertaining to civil disturbance. They extracted data pertaining to the incident, determined the contributing elements, and examined the course of the event. J. S. Liu investigated how the chosen political groups develop and change over time and identified the inner circles of government political power holders under formal work relations. For this, they carried out three key processes: network development, community identification, and community evolution tracking.

A semantic graph-based method for radicalization detection in social media was proposed by M. Ashcroft Pro-ISIS users frequently discuss religion, historical figures, and ethnicity, whereas anti-ISIS users concentrate more on politics, physical regions, and counter-ISIS initiatives.

**Proposed Methodology**

Based on the thorough literature review, an extended analytical and experimental methodology is applied. Real-time data was gathered from the news items and presented to the model, which then produced the findings based on the numerous aspects that affect the model's performance. The five main steps of the overall process are as follows:

Corpus/data collection, annotation of the data, feature extraction, pre-processing, application of machine Learning and deep learning algorithms, evaluation, accuracy metrics and validation are calculated.

**About the Dataset**

For the analysis, dataset is taken from the link-<https://zenodo.org/record/3271522#.Yxg-FuxN2SV>. The dataset is a publicly available text dataset. The corpus contains 52k articles from 100+ news outlets. Each article is labelled as either “propagandistic” (positive class) or “non-propagandistic” (negative class). The labelling was done indirectly using a technique known as distant supervision, i.e., an article is considered propagandistic if it comes from a news outlet that has been labelled as propagandistic by human annotators.

The whole dataset is split into training, test and validation datasets which are of size (35986, 15), (10159, 15) and (5125, 15) respectively.

Output Data Analysis

The training dataset is highly imbalanced with 11.17% of propaganda labels present in the data set. The output class is binary variant.

The validation dataset is highly imbalanced with 11.22% of propaganda labels present in the data set. The output class is binary variant.

Chart, pie chart

Description automatically generated Chart, pie chart

Description automatically generated

**Data Cleaning and Text-Pre-Processing**

Below are the steps being performed as a part of text preprocessing.

1.Lowercasing each row

2.Deconstruction of English words (ex: can’t – cannot)

3.Removal of special characters and punctuations

4.Removal of numbers and digits

5.Removal of stop words

6.Lemmatization of each word in each row

**Word cloud analysis**

A word cloud, also known as a tag cloud, is a graphic representation of text data in the form of tags. These tags are usually single words, and the size and color of the words indicate how important they are. The need to analyze the enormous amounts of text produced by these systems is growing as unstructured data in the form of text continues to see unprecedented development, particularly within the field of social media. By showing the word frequency in the text as a weighted list, a word cloud is a fantastic tool for aiding in the visual interpretation of literature and is helpful in swiftly getting insight into the most important elements in a particular text.

**Primary Unit of Analysis**

Propaganda based news detection and classification is one of the challenging tasks.

(1) Because it is written for a particular purpose and particular group of people and

(2) Because of similar use of context and linguistic features written for normal news as well. To solve these problems, this project aims to build an efficient web application where user can do a check whether the article or news is based on propaganda or not.

**Unit of Analysis:**

Our unit of analysis is to classify the propaganda and non-propaganda articles and news present in the dataset using NLP techniques.

The type of variable present in the dataset is news, articles-based text. We are going to use various NLP techniques to handle text data, such as pre-processing using removal of stop words, stemming, etc., pre-processing based on pre-trained embeddings such as fasttext, glove, word2vec, vectorization using tfidf, count and word encoding.

There will be various machine learning models going to be built on the vectorized dataset. Models are:

1. Naïve Bayes
2. Logistic Regression
3. Decision Tree
4. Random Forest
5. AdaBoost
6. XGBoost

We are also going to build deep learning-based models and architectures which involves the algorithms below:

1. Recurrent Neural Network
2. Convolutional neural Network
3. Attention Models
4. BERT

As the problem statement is a classification problem below are the evaluation metrics will be used.

1. Classification Report
2. Confusion Matrix
3. Roc Curve
4. AUC Score
5. Precision
6. Recall
7. F1 Scores

**Machine Learning Models**

**1.a Gaussian Naïve Bayes-with Count Vectorizer**

Graphical user interface, application

Description automatically generated

**1.b Gaussian Naïve Bayes-with tfidf Vectorizer**

Graphical user interface, application

Description automatically generated

**2.a** Bernoulli Naive Bayes- **with Count Vectorizer**

Chart, line chart

Description automatically generated

**2.a** Bernoulli Naive Bayes- **with tfidf Vectorizer**

Graphical user interface, application

Description automatically generated

3 a. Weighted Logistic Regression- **with Count Vectorizer**

Graphical user interface, application

Description automatically generated

**3 b. Weighted Logistic Regression-** with Tfidf Vectorizer

Graphical user interface, application

Description automatically generated

4 a. Non-Weighted Logistic Regression-- **with Count Vectorizer**

Graphical user interface, chart

Description automatically generated

**4.b Non-Weighted Logistic Regression--** with Tfidf Vectorizer

Graphical user interface, application

Description automatically generated

**5 a.** Decision Tree Classifier- **with Count Vectorizer**

Graphical user interface, application

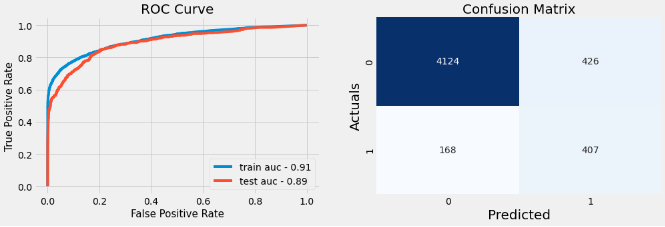
Description automatically generated

**5 b. Decision Tree Classifier- with Tfidf Vectorizer**

Graphical user interface, chart

Description automatically generated

**6.a Random forest- with Count Vectorizer**

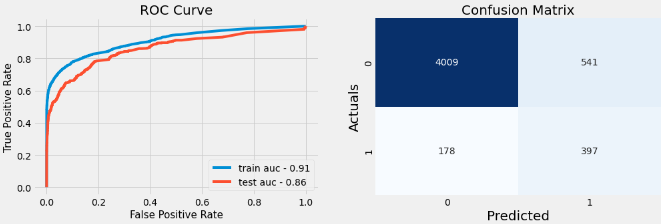


**6.b Random forest- with Tfidf Vectorizer**

Graphical user interface, application

Description automatically generated

7 a.Ada Boost Classifier- **with Count Vectorizer**



**7 b. Ada Boost Classifier-** with Tfidf Vectorizer

Graphical user interface, application

Description automatically generated

**8 a.**XGBOOST- **with Count Vectorizer**

Graphical user interface, application

Description automatically generated

**8 b.**XGBOOST- **with Tfidf Vectorizer**

Graphical user interface, application

Description automatically generated

**Output Comparison**

Table

Description automatically generated

**Deep Learning Models**

**1.LSTM Model-**

Chart, line chart, scatter chart

Description automatically generated

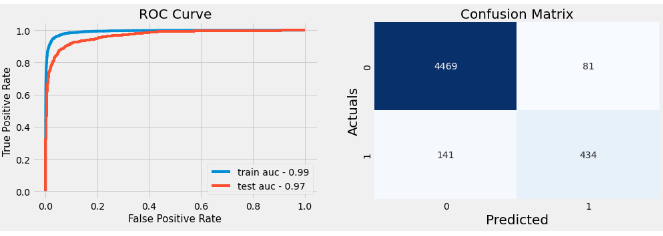
Graphical user interface, application

Description automatically generated

**2.GRU Modelling**

Chart, line chart

Description automatically generated



**3.CONV-LSTM Model**

Chart, line chart, scatter chart

Description automatically generated

Graphical user interface

Description automatically generated

**4.CONV-GRU Model-**

Chart, line chart

Description automatically generated

Graphical user interface, application

Description automatically generated

**5.Pre-Trained BERT Modelling**

Chart, line chart

Description automatically generated

Graphical user interface, application

Description automatically generated

**Output Comparison**

Graphical user interface

Description automatically generated with medium confidence

**Visualization of top 10 Models**

Chart, bar chart

Description automatically generated

**Conclusion and Future works**

Given the impact this unregulated content has on society, the significance of spotting propaganda is clear. This study intends to develop a system for identifying propaganda utilizing a convolution LTSM Bases Deep Learning Model network. To anticipate accurately, the neural network trains itself over many iterations. This model's efficiency in testing was 97.58% after processing a dataset of more than 52k+ news articles. To demonstrate that the proposed system is superior to the n-gram models, a comparison study is conducted.

The analysis of propaganda project has a promising future. It can be extended to include categorizing the test piece according to the kinds of propaganda strategies employed. It will improve the model's specificity, broadening its application. In social media and other online platforms where there is a high risk of fake news, propaganda detection can be employed to prevent swaying public opinion for private advantage. To preserve original viewpoints in society and uphold the democratic ideal, news policing will be automated.