Your report needs to have the following four sections:

1) Introduction: A summary of what you expected and did, and two-three of your most significant

findings (please use some numerical results here)

2) Descriptive Analysis: Introduce your descriptive findings about the dataset here

3) Classification Methods: Provide a description of your strategy and the steps you took to improve

your classification model (this includes the steps you followed for data-preprocessing, setting up

the model, and checking the strength of the model)

4) Classification Results: A detailed discussion on the results you obtained. What is your accuracy score? Evaluate and criticize yourself / your team.

The widespread adoption of social media for political communication creates unprecedented opportunities to monitor the opinions of large numbers of politically active individuals in real time. However, without a way to distinguish between users of opposing political alignments, conflicting signals at the individual level may, in the aggregate, obscure partisan differences in opinion that are important to political strategy. In this article we describe several methods for predicting the political alignment of Twitter users based on the content and structure of their political communication in the run-up to the 2010 U.S. midterm elections. Using a data set of 1,000 manually annotated individuals, we find that a support vector machine (SVM) trained on hashtag metadata outperforms an SVM trained on the full text of users’ tweets, yielding predictions of political affiliations with 91% accuracy. Applying latent semantic analysis to the content of users’ tweets we identify hidden structure in the data strongly associated with political affiliation, but do not find that topic detection improves prediction performance. All of these content-based methods are outperformed by a classifier based on the segregated community structure of political information diffusion networks (95% accuracy). We conclude with a practical application of this machinery to web-based political advertising, and outline several approaches to public opinion monitoring based on the techniques developed herein.

—---------------------------------------------------------------------------------------------------------------------------

Entering the peak of the the 2020 election season, social media platforms are firmly entrenched as a venue for Americans to process campaign news and engage in various types of social activism. But not all Americans use these platforms in similar ways. A new Pew Research Center analysis of U.S. adults’ Twitter behaviors finds that Democrats and Republicans have notable differences in how they use the site – from how often they tweet to the accounts they follow or mention in their own posts.The goal of this project was to use modelling methods to predict the political alignment of Twitter users. The training data consisted of 588243 unique tweets, tweet stats such as the number of times the tweet was favourited, number of retweets the tweet received, the hashtags included in the tweet, tweet year, etc. These are the explanatory variables we used to predict whether a politician who posted the tweet is a Democrat or Republican. The partisanship of the politician is predicted using a Logistic Regression model.

—-------------------------------------------------------------------------------------------------------------------------

### **Abstract:**

The goal of this project is to use modelling methods to predict the political alignment of Twitter users. The training data consists of over five hundred thousand unique tweets, tweet stats such as the number of retweets and favourites the tweet received, the hashtags included in the tweet, tweet year, etc. These are the explanatory variables used to predict whether a politician, who posted the tweet is a Democrat or Republican. The partisanship of the politician is predicted using a Logistic Regression model.

### **Introduction:**

During the peak of election season, social media platforms are adopted by highly active users with noteworthy thoughts and opinions about the elections. Twitter, especially, is burgeoning with tweets from users with vastly different points of view. We use several methods for predicting the political alignment of Twitter users based on the content and structure of their political communication in the run-up to the elections.

### **Descriptive Analysis:**

**Most relevant terms**

Term frequency-inverse document frequency is a text vectorizer that transforms the text into a usable vector. It combines 2 concepts, Term Frequency (TF) and Document Frequency (DF).

The term frequency is the number of occurrences of a specific term in a document. Term frequency indicates how important a specific term in a document.

TF represents text data as a matrix where the rows are the number of documents and columns are the total number of distinct terms in all the documents.

Document frequency is the number of documents that contain a particular term. It indicates how common the term is. Inverse document frequency (IDF) is the weight of a term, it’s purpose is to reduce the weight of a term if its occurrences are scattered throughout all the documents.

The higher the TF-IDF score the more important or relevant the term is; as a term gets less relevant, its TF-IDF score approaches 0.

Following are the visual representations after performing TF-IDF vectorization on our congressional tweets dataset.

**Word cloud of text data**

In order to understand the data better, exploratory analysis and investigation of the data were performed. It helped determine how to best manipulate data sources to get the answers we were looking for, making it easier to check assumptions, discover patterns, spot anomalies, and test a hypothesis.

The prime assumption that can be made is that a vast majority of Republicans would have conservative views and the majority of Democrat followers would have liberal views.

For discovering patterns in unstructured text data a word cloud is efficacious. It is a collection or cluster of words depicted in different sizes. The bigger and bolder the word appears, the more often it is selected/voted for by an audience member. Following is the word cloud obtained for political tweets.

**Coherence Score**

We used the coherence score to measure how interpretable the topics are to people. We used the C\_v measure for evaluating the coherence score of our Twitter data. C\_v measure is based on a sliding window, one-set segmentation of the top words and an indirect confirmation measure that uses normalized pointwise mutual information (NPMI) and the cosine similarity. The chart below outlines the coherence score, C\_v, for the number of topics across the dataset.

### **Classification Methods:**

**Text Processing**

Text preprocessing is traditionally an important step for natural language processing (NLP) tasks. In order to transform Twitter data into a more digestible form so that machine learning algorithms can perform better. This process involves the following:

Removing links and special characters - Since the tweets were web scraped, they contained some URLs and special characters that were mistranslated in the dataset. Since these are not useful for our NLP tasks, they are considered as noise and hinder the performance of the algorithm. It is beneficial to remove them. Regex was used to eliminate all this noise from the text data.

Removing stop words - Stopwords are very common words like “we”, “are”, “as”, etc. do not assist in NLP tasks such as sentiment analysis or text classifications. Hence, we removed stopwords like these to save computing time and resources.

Lemmatization - Lemmatization is the process of converting a word to its root form. It considers a language’s full vocabulary to apply a morphological analysis to words, aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma. For example - “meeting” would be lemmatized and converted to “meet”, “was” to “to be”, and so on. We used the NLTK library’s WordNetLemmatizer function. Wordnet is a publicly available lexical database of over 200 languages that provides semantic relationships between its words. It is one of the earliest and most commonly used lemmatization techniques.

POS Tagging - Part of Speech tagging was executed using the NLTK library. POS Tagging in NLTK is a process to mark up the words in text format for a particular part of a speech based on its definition and context. Some NLTK POS tagging examples are CC, CD, EX, TO, etc. POS tagger is used to assign grammatical information to each word of the sentence.

After preprocessing the text and converting it into a form that is predictable and analyzable for our task, we created a new column that stores the cleaned and processed version of each tweet in a new column in a dataset. The next step was to vectorize the processed text.

**Classification algorithms**

We tried several algorithms in order to determine the algorithm with the best trade-off between the resulting accuracy metric and computational resource consumption. Following are the algorithms we tested for text classification:

BERT (Bidirectional Encoder Representations from Transformers): It uses Transformer which is an attention mechanism that learns contextual relations between words (or sub-words) in a text. In its basic form, Transformer comprises two separate mechanisms — an encoder that reads the text input and a decoder that produces a prediction for the task.

Logistic regression

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression).

### **Classification Results:**

**Confusion matrix**

A confusion matrix is an N X N matrix, where N is the number of classes being predicted. For the problem in hand, we have N = 2, and hence we got a 2 X 2 matrix. In the confusion matrix, 0 stands for cases belonging to the “Democrat” class and 1 stands for cases belonging to the “Republican” class.

The structure of a confusion matrix returned by sci-kit learn is as follows:

The confusion matrix obtained for our Logistic Regression model is as follows:

The following interpretations can b e made from the

confusion matrix obtained -

• 56629 examples were correctly labeled as "Democrat"

• 8217 examples should have been classified as "Demo-

crat" but were labeled by the model as being "Republi-

can"

• 7804 examples should have been classified as "Repub-

lican" but were falsely labeled by the model as being

"Democrat"

• 45911 examples were correctly labeled as "Republican

However, in order to understand the performance of our logistic regression model better, a classification report containing precision, recall, F-1 score and support would be more relevant.

**Classification report**

Accuracy : the proportion of the total number of predictions that were correct. Our logistic regression model had an accuracy of 0.86. This means that 86% of the examples were correctly classified as Democrat or Republican.

Positive Predictive Value or Precision : It denotes the proportion of cases that were correctly identified for each class. Precision is calculated as the sum of true positives across all classes divided by the sum of true positives and false positives across all classes. Our model had a precision of 0.88 for Democrats and 0.85 for Republicans. This means that 88% of the examples which indeed belonged to the Democrat class were correctly labeled as Democrat and 85% of the examples with the true label, Republican were correctly classified as Republican by our Logistic Regression model.

Sensitivity or Recall: It denotes the proportion of actual positive cases which are correctly identified. Recall is calculated as the sum of true positives across all classes divided by the sum of true positives and false negatives across all classes. In our case, the recall values we obtained were 0.87 and 0.85 for the Democrat and Republican classes respectively.

F-1 score: F-Measure provides a way to combine both precision and recall into a single measure that captures both properties. For this scenario, neither precision nor recall can tell the entire story. F-measure provides a way to express both metrics in a single score. It is calculated as follows: F-Measure = (2 \* Precision \* Recall) / (Precision + Recall). Our logistic regression model has an F-1 score of 0.88 for Democrat classification task and 0.85 for Republican classification task.

Support: Support is the number of actual occurrences of the class in the specified dataset. So, this means that there were 64846 total examples for the “Democrat” class label and 53751 total examples for the “Republican” class.

**AUC-ROC curve**

This curve is an performance metric for binary classification problems. It is a probability curve that plots the True Positive Rate v/s False Positive Rate at differnt threshold values and is useful in separating the ‘signal’ from the ‘noise’. The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve.

We got an AUC value of 0.9443. This means that there is a high chance that the model is able to correctly label the data as belonging to the “Democrat” or “Republican” class.

References:

<https://www.pewresearch.org/politics/2020/10/15/differences-in-how-democrats-and-republicans-behave-on-twitter/>

<https://cnets.indiana.edu/wp-content/uploads/conover_prediction_socialcom_pdfexpress_ok_version.pdf>

<https://www.ibm.com/cloud/learn/exploratory-data-analysis>

<https://towardsdatascience.com/nlp-text-preprocessing-a-practical-guide-and-template-d80874676e79>

<https://www.geeksforgeeks.org/python-lemmatization-approaches-with-examples/>

<https://www.guru99.com/pos-tagging-chunking-nltk.html#:~:text=POS%20Tagging%20in%20NLTK%20is,each%20word%20of%20the%20sentence.>

<https://towardsdatascience.com/tf-idf-simplified-aba19d5f5530#:~:text=Term%20frequency%2Dinverse%20document%20frequency,specific%20term%20in%20a%20document.>

<https://www.baeldung.com/cs/topic-modeling-coherence-score#:~:text=We%20can%20use%20the%20coherence,words%20are%20to%20each%20other.>

<https://en.wikipedia.org/wiki/Catboost#cite_note-catboost-categorical-handling-7>

<https://en.wikipedia.org/wiki/Logistic_regression#:~:text=Logistic%20regression%20is%20a%20statistical,a%20form%20of%20binary%20regression).>

<https://machinelearningmastery.com/precision-recall-and-f-measure-for-imbalanced-classification/#:~:text=binary%20classification%20problems.-,In%20an%20imbalanced%20classification%20problem%20with%20more%20than%20two%20classes,false%20negatives%20across%20all%20classes.>