TILBURG UNIVERSITY

ANALYSING TEXT REPRESENTATIONS AND CLASSIFIERS FOR PREDICTING THE ECONOMIC POLITICAL LEANINGS OF REDDIT USERS

BEĀTE ABĀŠINA

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

STUDENT NUMBER

2123525

COMMITTEE

dr. Chris Emmery
dr. Sharon Ong

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

June 24th, 2024

WORD COUNT

8413

ACKNOWLEDGMENTS

First and foremost, I would like to thank my family. To my parents for supporting me both financially and mentally and to my sisters for always being there when needed. I would like to thank my partner and best friend, Clara, for standing by me and supporting me throughout the process. Last but certainly not least, I would like to acknowledge and thank my supervisor, Dr. Chris Emmery, for his guidance and insightful advice that have genuinely helped this research.

ANALYSING TEXT REPRESENTATIONS AND CLASSIFIERS FOR PREDICTING THE ECONOMIC POLITICAL LEANINGS OF REDDIT USERS

BEĀTE ABĀŠINA

Abstract

Author profiling is an area within natural language processing (NLP) that focuses on predicting attributes of text authors. While existing research has mainly focused on Twitter datasets for predicting features, such as age and gender, this research aims to broaden the field by exploring the Reddit dataset. Specifically, the paper focuses on predicting the economic political leaning of Reddit users, recognizing the importance of analysing political discourse for society's safety and enhancing understanding between citizens and leaders. To achieve this objective, the study employs three distinct word embeddings – Term Frequency-Inverse Document Frequency (TF-IDF), Word2Vec, and Big Bird – along with three different classifiers – Logistic Regression (LR), Support Vector Classifier (SVC), and Gradient Boosted Decision Trees (GBDT). These algorithms are compared against majority and feature engineering baselines. Notably, the best-performing models based on the F1-score entail a feature engineering baseline trained on LG and TF-IDF embeddings trained on SVC, revealing scores of 46.00% and 45.00%, respectively. Moreover, SHapley Additive exPlanations (SHAP) values indicate the significant importance of character trigrams in the best-performing model. Conversely, the worst-performing models are the majority baseline with an F1-score of 27.79% and Big Bird word embeddings trained on SVC with an F1-score of 37.00%. The paper also recognizes that the results could have been influenced by factors such as stepping away from binary classification and involving the "center" class.

1 DATA SOURCE, ETHICS, CODE, AND TECHNOLOGY STATEMENT

The Reddit dataset has been acquired from Emmery et al. (2024), and a non-disclosure agreement was signed. The obtained data was anonymised - Reddit usernames were deleted after splitting the data, and only users' political leaning and textual posts were used for learning algorithms. Work on this thesis did not involve collecting data from human participants or animals. The original owner of the data and code used in this thesis retains ownership of the data and code during and after the completion of this thesis. However, the authors are aware of the use of this data for this thesis and potential research publications. All the figures were created by this paper's author. The reused/adapted code fragments are clearly indicated in the notebook. In terms of writing, the author used assistance with the language of the paper. Grammarly[1] was used to improve the author's original content for paraphrasing, spell-checking, and grammar. The code is available at GitHub repository https://github.com/abasinaB/Predicting_political_leaning_Reddit

2 INTRODUCTION

2.1 *Problem Statement*

The data science community has recognized the advantages of employing social media as a valuable data source. Its value lies in its abundance of data and the dual nature of social media posts and interactions.

On the one hand, social media has proven to be a powerful tool in pivotal events, such as the Arab Spring (Wolfsfeld et al., 2013), as well as in raising awareness for social movements, for example, Black Lives Matter and MeToo (Greene et al., 2019). On the other hand, it is also a place that fosters hate speech, racism, and echo chambers (Proferes et al., 2021).

While much attention in author profiling tasks has been directed towards Twitter (now known as X) as a dataset (Conover et al., 2011; Pennacchiotti & Popescu, 2021; Preoţiuc-Pietro et al., 2017), this research seeks to extend the use of other social media platforms, such as Reddit. By employing this dataset, this paper explores different types of text representations and one of the polarizing topics on digital platforms – economic political discourse. This research will utilize natural language processing (NLP) techniques to engage in author profiling and predict Reddit posts authors' economic political leaning.

[1] https://app.grammarly.com/

2.2 *Social and Scientific Relevance*

The social implications of this research goal are fundamental. As social media can be a place that fosters echo chambers and political radicalisation (Atari et al., 2022), the ability to predict an individual's political leanings holds significant promise for society, its stability, and security. Early detection of radicalisation could significantly contribute to mitigating potential harm and providing a safer digital environment. Moreover, understanding and analysing political discourse greatly value democracy, creating a more informed and engaged society (Chambers, 2018).

Beyond the social impact, this research also holds scientific relevance. Reddit, a platform not extensively explored in author profiling studies, presents a rich source of longer texts compared to Twitter for investigating the capabilities of different word embeddings and classifier models. Moreover, the longer texts present an opportunity to explore the transformer model Big Bird and its word embeddings, an area that previous research has yet to explore fully. This would extend author profiling research and provide researchers with valuable insights for future investigations.

2.3 *Research Question*

To gain value for the society and research community and fulfil the research objective, this paper proposes the main research question:

> *To what extent can machine learning classifiers, leveraging various text representations, accurately predict the economic political leaning of Reddit users based on their textual posts?*

Three sub-questions will be addressed to implement a concise and structured approach. First, previous research on predicting political leaning has demonstrated the effectiveness of Term Frequency-Inverse Document Frequency (TF-IDF) and Bag of Words (BoW) as valuable techniques for extracting word embeddings (Conover et al., 2011; Kosiv & Yakovyna, 2022). Furthermore, Kramp et al. (2023) has shown the potential of Big Bird embeddings compared to engineered text features for longer texts. Therefore, the first sub-questions is:

RQ1 *How are the accuracy and F1-score of classifiers influenced by TF-IDF, Word2Vec, and Big Bird word embeddings compared to the feature engineering baseline?*

Second, various classifiers have been adopted in NLP tasks. One of the most prominent papers in predicting political leaning has used Gradient

Boosted Decision Trees (GBDT) (Pennacchiotti & Popescu, 2021). At the same time, other researchers advocate for Support Vector Machine (SVM) algorithms (Conover et al., 2011), such as Support Vector Classifier (SVC) (Kosiv & Yakovyna, 2022). However, when using Big Bird embeddings, Kramp et al. (2023) opt for Logistic Regression (LR). Thus, the second sub-question is:

RQ2 *To what extent do LR, SVC, and GBDT, combined with different text representations, accurately predict economic political leaning compared to the baselines?*

Last, to increase the explainability and uncover the most crucial features of the most successful model, this paper analyses SHapley Additive exPlanations (SHAP) values. Hence, the third sub-research question is:

RQ3 *Which features are the most important in predicting economic political leaning based on SHAP values?*

2.4 *Main Findings*

After addressing all the research questions, the main findings reveal the continued significance of computing engineered features (without word embeddings) alongside the success of TF-IDF word embeddings. However, the results of classifiers present a less clear picture, with SVC demonstrating the highest and lowest outcomes.

In order to show how the findings came about, this research follows a certain structure. The paper starts with an exploration and synthesis of background literature. Subsequently, the methodology is explained and justified. Following this, the results are presented and then discussed. Finally, conclusions are drawn.

3 LITERATURE REVIEW

3.1 *Predicting Author's Characteristics and Author Profiling*

Author profiling and its related fields, including authorship verification, originate from stylometry, a branch of linguistics focused on the author and characteristic attribution by analysing text writing style and quantitative features (Neal et al., 2017). The development of stylometry has been widely credited to Lutoslawski (1898) and his work on Plato's dialogues. Exploring word frequency, importance, and other text features, the scholar analysed the chronology of the dialogues.

Over time, stylometry techniques have been adopted in author profiling where to infer authors' characteristics not only linguistic features of the text are examined, but also others, such as content-based and stylistic features (Ouni et al., 2023). With the rise of the internet and the abundance of data from online platforms, author profiling has shifted towards analysing social media users rather than focusing on written texts, such as works from classical authors like Charles Dickens and William Makepeace Thackeray (Mendenhall, 1887) or federal papers (Mosteller & Wallace, 1963).

Research on predicting authors' characteristics using internet data has explored various datasets, predicting a range of features through diverse methodologies. However, significant focus has been on the Twitter dataset to predict features, such as gender and age (Ouni et al., 2023). Diversity in methods is more noticeable. Some scholars have emphasised network behaviour analysis. This encompasses examining retweets (Stefanov et al., 2020; Wong et al., 2021), as well as studying retweeted and followed accounts (Volkova et al., 2014). Such methods have demonstrated effectiveness in predicting authors' attributes and related areas, such as stance prediction (Magdy et al., 2015).

Other researchers have directed their focus toward analysing individual tweets or posts. This involves examining the hashtags (Conover et al., 2011; Pennacchiotti & Popescu, 2021), mentions of other profiles, and the textual content itself. Various techniques from NLP have been applied to analyse the posts, such as studying the text's stylistic features (Ouni et al., 2023), including prototypical words via different n-grams, word embeddings and grammatical and spelling mistakes (Goldin et al., 2018), and statistical features, such as average sentence length. However, many authors opt for a hybrid approach that integrates multiple methods and features.

3.2 *Predicting Political Leaning*

These methods and features have also been adopted to explore the author's political leaning. As with other characteristics also here, the Twitter dataset has been widely used, and the focus often centres on the political context of the United States of America (USA) or social politics, where individuals are categorized as either Republican or Democrat or liberal or conservative.

A prominent work in this area was introduced by Pennacchiotti and Popescu (2021). The researchers leveraged the Twitter data set to develop algorithms that predict political leaning (Democrat or Republican), ethnicity, and Starbucks enthusiasm. Pennacchiotti and Popescu (2021) analysed users' profiles, networks, posting patterns, and posts. To examine users' posts, the authors considered prototypical words, hashtags, sentiment, and different Latent Dirichlet Allocation (LDA) variations to identify prevalent topics in the tweets. The authors utilised GBDT to achieve results that exceeded 80% in precision, recall, accuracy, and f-score, underscoring the significance of linguistic features.

A similar approach was undertaken by Conover et al. (2011). Just as Pennacchiotti and Popescu (2021), Conover et al. (2011) focused on the Twitter dataset and considered the tweet's author's network, hashtags, and linguistic features. While sharing similarities with the work of Pennacchiotti and Popescu (2021), Conover et al. (2011) delved into the 2010 USA midterm elections, categorizing political leaning not as binary variable, but as left, right, or ambiguous. The scholars employed SVM and specified the TF-IDF algorithm to extract prototypical words and create word embeddings. Compared to Pennacchiotti and Popescu (2021) Conover et al. (2011) research highlighted the significance of hashtags and networks over word embeddings, achieving an accuracy rate exceeding 90%.

In recent years, a shift away from USA-centric politics has been more noticeable (Kosiv & Yakovyna, 2022; Preoţiuc-Pietro et al., 2017). For example, Preoţiuc-Pietro et al. (2017) viewed political leaning not as a categorical variable but as a scale ranging from conservative to liberal. By combining Twitter data with survey responses, the researchers considered a range of linguistic features such as unigrams, Linguistic Inquiry and Word Count (LIWC), Word2Vec, and sentiment analysis. Preoţiuc-Pietro et al. (2017) utilized LR and observed about an 8% increase in accuracy from their baseline.

Furthermore, Kosiv and Yakovyna (2022) considered context other than the USA and explored the Ukraine-Russia war. By employing various classifiers and word embeddings, the authors categorized individuals as either pro-Russian or pro-Ukrainian. The findings showed the effectiveness of BoW and TF-IDF word embeddings while using SVC, LR, Long Short-Term Memory (LSTM), and Convolutional Neural Network (CNN).

3.3 *Transformer Models*

The transformer models, introduced by Vaswani et al. (2017), have emerged as a significant state-of-the-art algorithm in machine learning tasks, including predicting political stances from text. These models have been widely used in computer vision and NLP tasks, such as translation and answer generation (Lin et al., 2022), and less in creating word embeddings that are passed on to other algorithms (Ahuja & Sharma, 2022; Pericherla & Ilavarasan, 2024). Transformers' architecture is characterized by their decoder, encoder, and self-attention mechanism. This structure has been proven beneficial in keeping long-range dependencies and allowing parallelization.

Bidirectional Encoder Representations from Transformers (BERT) has gained significant attention among other transformer models. Baly et al. (2020) leveraged this transformer to predict news articles' left, center, or right political leaning. The researcher used LSTM and BERT. Baly et al. (2020) results showed that BERT significantly increased f-score and accuracy compared to the majority baseline. Building upon the BERT, Jiang et al. (2023) adopted their transformer model - Retweet-BERT, tailored for predicting political leaning. The authors achieved a macro f-score exceeding 95% by combining social and textual features. However, both papers utilized transformer algorithms as classifiers and did not explore their ability to create word embeddings.

3.4 *Big Bird*

Although different transformer models, such as BERT, have demonstrated success in specific tasks, they have disadvantages when working with longer texts due to memory constraints (Zaheer et al., 2020). Hence, Zaheer et al. (2020) introduced transformer Big Bird. This transformer model utilizes sparse instead of full attention. Big Bird's sparse attention incorporates random, window, and global attention. This has proven to be an effective and accurate transformer in various tasks involving longer text data, such as document translation (Li & Chan, 2019) and Native Language Identification (NLI) (Kramp et al., 2023).

Despite its success in tasks involving longer text data, the use of Big Bird in predicting political leaning and creating word embeddings has not been extensively explored. On the one hand, Nikolaev et al. (2023) employed a Big Bird classifier to predict the political leaning of manifestos from the MARPOR dataset. On the other hand, Kramp et al. (2023) leveraged Big Bird word embeddings for the NLI task. The researchers extracted 768-word embedding features from the last hidden state of the Big Bird transformer. This approach proved to be efficient and fast compared to more traditional methods that rely on features like spelling mistakes and n-grams. Nevertheless, utilizing Big Bird embeddings to predict political leanings has not yet been explored in existing research.

3.5 *Research Gap*

Explored literature indicates potential areas for enriching author profiling and predicting political leanings. First, Reddit has not been widely studied compared to Twitter as a dataset. Second, while political leaning prediction often revolves around binary categories linked to the USA or social politics, there is a lack of investigation into the economic political leaning scale.

Third, although transformer models, such as Big Bird, have been utilized as classifiers, their application for generating word embeddings has been side-lined.

Therefore, in this research, the aim is to compare various word embeddings alongside different classifiers to predict economic political leaning. The effectiveness of well-established word embeddings in political leaning prediction, such as Word2Vec and TF-IDF, will be compared with non-word-embedding features and the embeddings generated by Big Bird. The analysis will involve classifiers LR, SVC, and GBDT, which have been shown effective in previous studies. Consequently, to gain insights into how the model works, SHAP values will be computed.

4 METHODOLOGY & EXPERIMENTAL SETUP

In order to compare different text representations and build machine learning models for predicting political leaning using the Reddit dataset, this research leveraged the Python programming language and utilized Google Colab, which employs Jupyter notebooks. The visualization of experimental steps can be seen in Figure 1.

The libraries used were Pandas (Mckinney, 2011), Scikity-Learn (Pedregosa et al., 2011), NumPy (Harris et al., 2020), Matplotlib (Hunter, 2007), Langdetect (Shuyo, 2021), SymSpell[2], Levenshtein[3], Spacy (Honnibal & Montani, 2017), Transformers[4], Torch (Paszke et al., 2017), Tqdm[5], Gensim (Rehurek & Sojka, 2011), Seaborn (Waskom, 2021), and SHAP (Lundberg & Lee, 2017).

4.1 *Dataset Description*

This study utilizes the Reddit dataset to address the research questions and assess the performance of classifiers when using various word embeddings to predict political affiliation.

Reddit is a social media platform that facilitates user discussions through posts and comments subject to community voting mechanisms (Proferes et al., 2021). Compared to Twitter, it allows users to create longer posts and contribute to different subreddits covering diverse topics.

The data was collected in 2020-2022 by Emmery et al. (2024), making it temporally relevant. As the dataset is still under review, a non-disclosure agreement was signed. It contains usernames, corresponding political

[2] https://github.com/wolfgarbe/SymSpell

[3] https://github.com/ztane/python-Levenshtein

[4] https://github.com/huggingface/transformers

[5] https://github.com/tqdm/tqdm

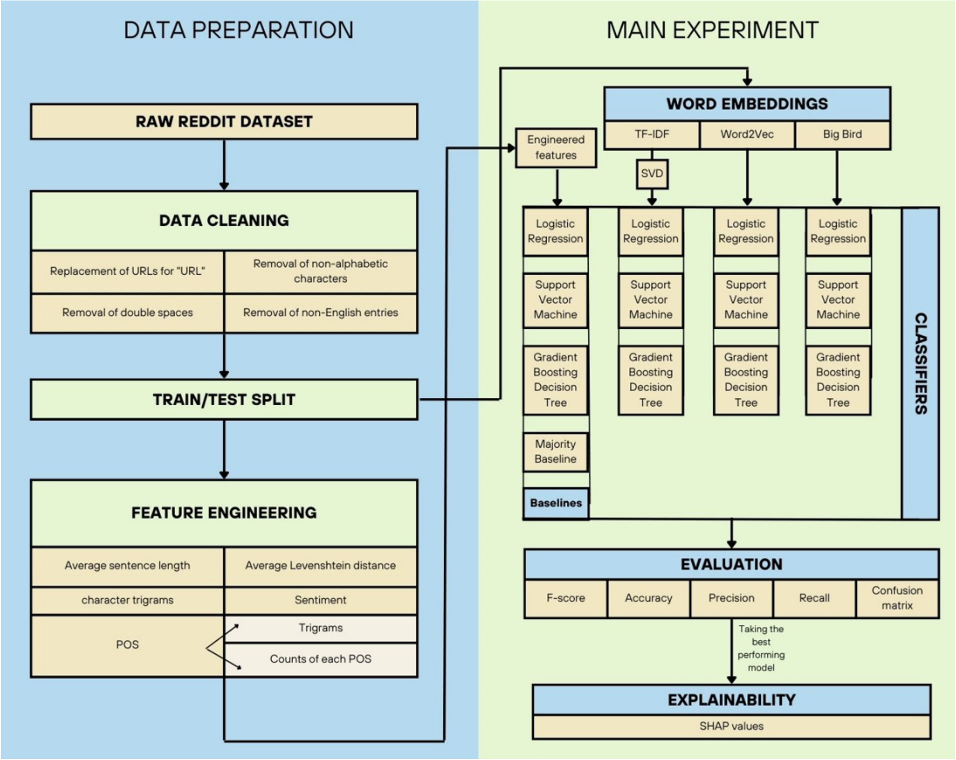

Figure 1: Flowchart of the Research, Showing Data Preparation Process and Main Experiment of Evaluating Four Different Text Representations Combined with Three Learning Models and Explaining the Best-Performing Model

Note. POS stands for Parts of speech, TF-IDF for the Term Frequency-Inverse Document Frequency, and SVD for Singular Value Decomposition.

leanings, and posts sourced from users active in the r/PoliticalCompass subreddit. This includes their posts within the subreddit and all their other posts on Reddit. Each user's posts were aggregated and then split so each entry would have 1,500 words.

Compared to most works on political leaning, the dataset is labelled based on the Political Compass economic side of political stance. The dataset contains three distinct classes: "left," representing advocates of a government-led economy, "center," and "right," reflecting people that opt for a completely free market (The Political Compass, n.d.). Prior to any data cleaning, the dataset contains 57231 entries - 25201 "center" (44.03%), 17454 "right" (30.50%), and 14576 "left" (25.49%).

4.2 *Preprocessing*

4.2.1 *Cleaning the Data*

In order to clean data and reduce noise, several data-cleaning steps were implemented. Repeating white spaces were substituted with one and URLs were replaced by string "URL"[6].

Only letters, numbers, spaces, and apostrophes were kept to filter out unnecessary characters [7].

Entries with non-English text were detected and removed to keep data more consistent between datasets, and as some techniques applied were focused on the English language, such as sentiment detection. This excluded 2082 entries, resulting in the dataset with 55149 entries. The distribution of classes slightly shifted to 44.22% for "center," 30.42% for "right," and 25.35% for "left."

Other techniques were considered, such as lemmatization (Aydin, 2023). However, it was decided not to employ this technique to avoid eliminating valuable information indicative of political leaning through writing style.

4.2.2 *Splitting the Data*

The data was split into two datasets – the train and test sets. The training dataset was utilized for training the classifiers and cross-validation, whereas the training dataset was kept for out-of-sample evaluation.

A specific splitting approach was employed to ensure balanced datasets and keep independence between them. Each author was assigned to either the train or test dataset (Kramp et al., 2023) while stratifying on political leaning. Subsequently, individual author entries were added up within

[6] https://stackoverflow.com/questions/11331982/how-to-remove-any-url-within-a-string-\in-python

[7] https://github.com/python/cpython/tree/3.12/Lib/re/

their respective datasets. Although the goal was to get an 80/20 split between train and test data, this method resulted in 80,41% train data and 19,59% test data. The distribution of classes within train data was 44.06% for "center," 32.22% for "right," and 24.71% for "left," while in test data, the distribution was 44.86% for "center," 27.98% for "right," and 27.15% for "left."

4.3 *Text Representations and Answering Research Question 1*

To address RQ1, *How are the accuracy and F1-score of classifiers influenced by TF-IDF, Word2Vec, and Big Bird word embeddings compared to the feature engineering baseline?* four text representations were considered. These encompassed baseline as feature engineering and three types of word embeddings.

4.3.1 *Baseline: Feature Engineering*

The baseline of feature engineering was primarily drawn from Kramp et al. (2023) as they are the pioneers in Big Bird word embeddings, and their baseline used various text features.

Average sentence length. Computed for each entry by dividing the total number of words in an entry (1500 as previously specified) by the sentence count.

Spelling mistakes. Made by first correcting the spelling of user texts using a maximum edit distance of 2[8] and the 'en-80k' dictionary[9]. The average Levenshtein distance was then calculated by comparing the original post with the corrected one on a word-by-word basis. All distances were summed and divided by the total number of words in the entry (1500).

Character trigrams. The top 1000 character trigrams were computed as new features.

Parts of speech (POS). Words, numbers, symbols, or punctuations from the non-processed entries were transformed in their respective POS tags. The 300 most frequently occurring POS trigrams were made into features representing their number of appearances in each entry. Moreover, to account for the possibility that people from left, right, or center political spectrum might use different amounts of verbs, nouns, pronouns, or any other POS, the total number of each POS category was recorded for each entry (Bugdani, 2022), adding nine other features that were not present in Kramp et al. (2023) work.

[8] https://symspellpy.readthedocs.io/en/latest/examples/lookup.html#basic-usage

[9] https://github.com/wolfgarbe/SymSpell/blob/master/SymSpell.FrequencyDictionary/en-80k.txt

Sentiment. Another feature not included in Kramp et al. (2023) used in this research was sentiment. As other researchers on political leaning (Pennacchiotti & Popescu, 2021) have used it, and literature shows that some political ideologies, such as conservatives, tend to use more negative language (Sterling et al., 2020), it was used as a feature in this research. This was done by using positive and negative word dictionaries[10] and updating each entry score by +1 for positive word and -1 for negative word.

Therefore, feature engineering resulted in 1312 features.

4.3.2 *Term Frequency-Inverse Document Frequency*

The adoption of TF-IDF started by Sparck Jones (1972) in the 1970s. This word embedding method is composed of two components. Term Frequency (TF) represents the frequency of a word divided by the total number of words in an entry. Inverse Document Frequency (IDF) assesses the significance of each word by measuring how much the word is seen in all the entries, assigning smaller values to words that appear more in the whole text.

This word embedding technique has been successfully adopted by scholars in the predictions of political leaning (Conover et al., 2011; Kosiv & Yakovyna, 2022) as it has the advantages of dealing with large datasets and accounts for the importance of terms in the text. Therefore, it was also implemented in this research. Due to the limits of computational power and Random Access Memory (RAM) and to reduce noise, the max_features parameter was set to 10000, and stop words were excluded. Moreover, to further decrease dimensionality, Singular Value Decomposition (SVD)[11] based on Elbow method (Falini, 2022) was applied that resulted in 70 features.

4.3.3 *Word2Vec*

Word2Vec, introduced by Google researcher Tomas Mikolov in 2013 (Mikolov et al., 2013), is a method that learns representations of words in a vector space. Compared to TF-IDF, Word2Vec considers context, meaning that words closer in context are represented as closer in the vector space.

Due to this advantage, other researchers have also adopted this approach in author profiling. For instance, Kosiv and Yakovyna (2022) utilized Word2Vec as BoW, where a word is predicted based on its surrounding

[10] https://github.com/shekhargulati/sentiment-analysis-python/tree/master/opinion-lexicon-English

[11] https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.TruncatedSVD.html

words. Therefore, this research also utilized this method and employed a pre-trained Google model[12] that has been trained on news articles. The decision to use a pre-trained model was driven by computational constraints and the benefit that a pre-trained model has been trained on large data that could potentially enhance out-of-sample performance. The text entries were tokenized, and a pre-trained model was applied to transform words into vectors.

4.3.4 *Big Bird*

As mentioned in section 3.4, Big Bird was introduced by Zaheer et al. (2020) and is a transformer model that uses a sparse attention mechanism, which, compared to full attention, has a better ability to handle longer text sequences.

Given the often long entries on Reddit, leveraging this transformer model could be beneficial. While the exploration of using Big Bird for word embeddings is scarce, promising results have been demonstrated by Kramp et al. (2023), thus, was explored in this research. Following the approach outlined in Kramp et al. (2023) code[13], this research employed a pre-trained Big Bird model with the roberta base[14] and an input size of 2048, selected for its best performance in Kramp et al. (2023). Using the last hidden state of the Big Bird model, 768 features were extracted.

4.4 *Classifiers and Answering Research Question 2*

To evaluate different types of text representations and answer RQ2 (*To what extent do LR, SVC, and GBDT, combined with different text representations, accurately predict economic political leaning compared to the baselines?*), three classifiers were selected and compared to baseline models.

Classifiers, LR, SVC, and GBDT, were fed the baseline features, TF-IDF, Word2Vec, and Big Bird word embeddings. All models were trained to predict classes of "left," "center," and "right," representing different economic political leanings. To ensure model stability and avoid overfitting, 5-fold cross-validation was employed during testing. No hyperparameter tuning was conducted. Although it could have hampered the classifier performance and their robustness, it was done using the same models for word embedding comparison (Kramp et al., 2023).

[12] https://code.google.com/archive/p/word2vec/

[13] https://github.com/SergeyKramp/mthesis-bigbird-embeddings/blob/master/feature_extractors/transformer_feature_extractor.py

[14] https://huggingface.co/google/bigbird-roberta-base

4.4.1 *Majority Baseline Model*

In addition to baselines with engineered features tested on each classifier, a majority baseline was also included in the analysis. Including the majority baseline provided a sanity check to evaluate whether the models were learning from the data and outperforming a simple majority-class prediction model.

4.4.2 *Logistic Regression*

The origins of LR can be traced back to the early 19th century (Cramer, 2002), but its development has been credited mainly to Cox (1958). LR is utilized for binary categorization, where each category is assigned a probability through a logistic function.

Kramp et al. (2023) employed LR and highlighted it as one of the best-performing algorithms on the Reddit-L2 dataset (Goldin et al., 2018). However, given that LR is designed for binary prediction, this research employed a variant of it - multinomial logistic regression to classify more than two categories by utilizing logits and assigning probabilities to each class (Kwak & Clayton-Matthews, 2002). Similar to Kramp et al. (2023), also in this study, the default parameters were adopted with max_iter set to 1000.

4.4.3 *Support Vector Classifier*

SVM was first developed by Cortes and Vapnik (1995) in the 1990s as a supervised learning algorithm designed to draw a hyperplane separating distinct categories in the data. When the data cannot be linearly separable by decision boundary, SVM can leverage the Kernel trick to map features into higher-dimensional spaces. SVM can be applied for regression and classification. For classification purposes, SVC is utilized.

In the area of predicting political leanings, both Conover et al. (2011) and Kosiv and Yakovyna (2022) have successfully employed SVM and SVC, respectively, showing positive results. Given SVC's ability to handle non-linear classification and high-dimensional feature spaces, it emerges as a fitting algorithm for this study. Therefore, it was applied in this research using default parameters.

4.4.4 *Gradient Boosted Decision Trees*

Gradient Boosting Machine (GBM) is an ensemble learning algorithm introduced by Friedman (2001). This algorithm applies the method of boosting, where weak learners are added after one another to the model, with each learner correcting errors made by its predecessors. When decision trees are used as the learner in this algorithm, it is referred to as GBDT.

A work by Pennacchiotti and Popescu (2021) applied GBDT to predict political leaning, achieving an accuracy and f-score exceeding 80%. By combining multiple weak learners into a single strong learner, GBDT demonstrates the ability to deliver accurate results; thus, this study employed the algorithm with its standard hyperparameters.

4.5 *Evaluation*

The performances of each learning algorithm on the test set for out-of-sample evaluation were compared based on metrics such as accuracy, f-score, recall, and precision, both in relation to each other and against the baseline models.

Accuracy scores were computed as it is equally important to predict all classes correctly (Pennacchiotti & Popescu, 2021). Additionally, the F1-score was utilized to take into account recall and precision equally and address the class imbalance within the dataset. While precision and recall were not the primary focus of this research, they were presented for the benefit of future researchers. The evaluation process also included an analysis of errors through the examination of the confusion matrix. This matrix showed specific types of misclassifications made by the learning algorithms.

4.6 *SHapley Additive exPlanations Values and Answering Research Question 3*

In order to increase transparency and gain deeper insights into the best-performing algorithm and answer RQ3 (*Which features are the most important in predicting economic political leaning based on SHAP values?*), SHAP values were computed. Introduced by Lundberg and Lee (2017), SHAP values are grounded in game theory and provide local and global interpretability for features. Because these values can be visualized for comprehension and are model-agnostic, they can be utilized across various learning algorithms, rendering them a valuable tool for enhancing explainability in this research. Given the computational demands of SHAP values, particularly in high-dimensional feature spaces, a random subset comprising 20% of the test cases (2161 entries) was explained in this research. Moreover, the model-agnostic KernelExplainer was employed to get global interpretability, and SHAP values were computed for each class separately.

5 RESULTS

Throughout this paper, a total of 13 models were developed - a majority baseline as well as three distinct types of word embeddings (TF-IDF, Word2Vec, Big Bird) and feature engineering baseline, each coupled with three diverse classifiers (LR, SVC, GBDT). Additionally, the best model was explained using SHAP values. This section presents the performances and results of these models.

5.1 *Text Representations' Results*

Table 1 presents average performance metrics - accuracy, F1-score, precision, and recall – on test data of different text representation techniques. The averages were composed by taking the mean performance of each text representation method among the different classifiers.

Table 1: Average Text Representation Performances on the Test Dataset Across Classifiers: Logistic Regression, Support Vector Classifier, and Gradient Boosting Decision Trees.

	Feature Engineering Baseline	TF-IDF	Word2Vec	Big Bird
Accuracy	46.33	**48.33**	47.00	**46.00**
F1-score	43.67	**44.00**	41.67	**40.00**
Precision	45.00	**47.00**	44.33	**43.67**
Recall	46.33	**48.33**	47.00	**46.00**

Note. TF-IDF stands for Term Frequency-Inverse Document Frequency. All values are represented in percentages. The values in black bold represent the best average performance across text representations, while the values in grey bold represent the worst performances.

The TF-IDF achieved the highest accuracy score of 48.33%, and Big Bird embeddings yielded the lowest accuracy at 46.00%. Similarly, TF-IDF exceeded other text representation techniques in precision and recall with scores of 47.00% and 48.33%, respectively, while the Big Bird embeddings were the worst with scores of 43.67% and 46.00%.

The lowest overall scores can be seen in the F1-score. However, the trend is also sustained here, and TF-IDF presents the best score of 44.00%, while Big Bird obtains the lowest score of 40.00%. Hence, these results underscore the effectiveness of the TF-IDF and reveal limitations in the performance of the Big Bird embeddings.

5.2 *Classifiers' Results*

Table 2 highlights the performance metrics for the majority baseline and the average scores for each classifier across different text representation techniques. Compared to the word embeddings, here results are more mixed.

Table 2: Average Classifier Performances on the Test Dataset Across Text Representations: Feature Engineering Baseline, Term Frequency-Inverse Document Frequency, Word2Vec, and Big Bird

	Majority Baseline	LR	SVC	GBDT
Accuracy	**44.86**	47.00	**47.25**	46.5
F1-score	**27.79**	41.75	42.25	**43.00**
Precision	**20.12**	**45.75**	45.25	44.00
Recall	**44.86**	47.00	**47.25**	46.50

Note. LR stands for Logistic Regression, SVC for Support Vector Classifier, and GBDT for Gradient Boosting Decision Trees. All values are represented in percentages. The values in black bold represent the best average performance across different models, while the values in grey bold represent the worst performances.

The majority baseline demonstrates the poorest performance across all metrics, with an accuracy of 44.86%, F1-score of 27.79%, precision of 20.12%, and recall of 44.86%. Among the classifiers, SVC scores the highest accuracy score of 47.25% and recall of 47.25%, while LR achieves the highest precision at 45.75%. As with word embeddings, the lowest scores are on the F1-score, but GBDT achieves the highest F1-score of 43.00%.

5.3 *Models' Results*

While the average results highlight the highest performance of TF-IDF word embeddings and the best F1-score achieved by the GBDT algorithm, it is essential to consider the best performances across various combinations of word embeddings and classifiers, as seen in Table 3.

Although the majority baseline delivers the worst results, it is worth noting that its accuracy and recall are only 4.14% lower than the best-performing model, which features the LR classifier with TF-IDF word embeddings (49.00%). Nevertheless, the majority baseline's F1-score and precision are significantly lower than other models, showcasing a difference of 9.21% to 18.21%, pointing to the classifiers' learning ability. The majority bassline's F1-score is 27.79%, while the best F1-score of 46.00% is achieved by the LR classifier trained without word embeddings (feature engineering baseline). Across different word embeddings, the best F1-score is recorded

Table 3: Performance Results on the Test Dataset for Classifiers in a Combination with Different Text Representations

	Classifier	Accuracy	F1-score	Precision	Recall
	Majority Baseline	**44.86**	**27.79**	**20.12**	**44.86**
Feature Engineering Baseline	LR	46.00	**46.00**	46.00	46.00
	SVC	47.00	43.00	46.00	47.00
	GBDT	46.00	42.00	46.00	46.00
	Average	46.33	43.67	45.00	46.33
TF-IDF	LR	**49.00**	43.00	**49.00**	**49.00**
	SVC	48.00	**45.00**	46.00	48.00
	GBDT	48.00	44.00	46.00	48.00
	Average	48.33	44.00	47.00	48.33
Word2Vec	LR	47.00	38.00	44.00	47.00
	SVC	48.00	44.00	46.00	48.00
	GBDT	**46.00**	43.00	**43.00**	**46.00**
	Average	47.00	41.67	44.33	47.00
Big Bird	LR	**46.00**	40.00	44.00	**46.00**
	SVC	**46.00**	**37.00**	**43.00**	**46.00**
	GBDT	**46.00**	43.00	44.00	**46.00**
	Average	46.00	40.00	43.67	46.00

Note. TF-IDF stands for Term Frequency-Inverse Document Frequency, LR for Logistic Regression, SVC for Support Vector Classifier, and GBDT for Gradient Boosting Decision Trees. All values are represented in percentages. The values in black bold represent the best average performance across text representations, while the values in grey bold represent the worst performances.

for SVC combined with TF-IDF (45.00%). Repeatedly showcasing the success of TF-IDF features.

Big Bird embeddings exhibit the lowest results across all word embeddings. Its accuracy for all classifiers is 46.00%, the lowest across the results. Perhaps more important, it also displays the poorest F1-scores, with the lowest recorded at 37.00% when paired with SVC.

Interestingly, even though GBDT demonstrated the best average F1-score, it does not consistently achieve the best F1-score performance when looking at each algorithm separately. Additionally, it shows the lowest recall and accuracy when paired with Word2Vec and Big Bird embeddings.

5.4 *Error Analysis*

All confusion matrixes reveal consistent patterns and can be seen in Appendix A. The best and worst performing models based on the F1-score

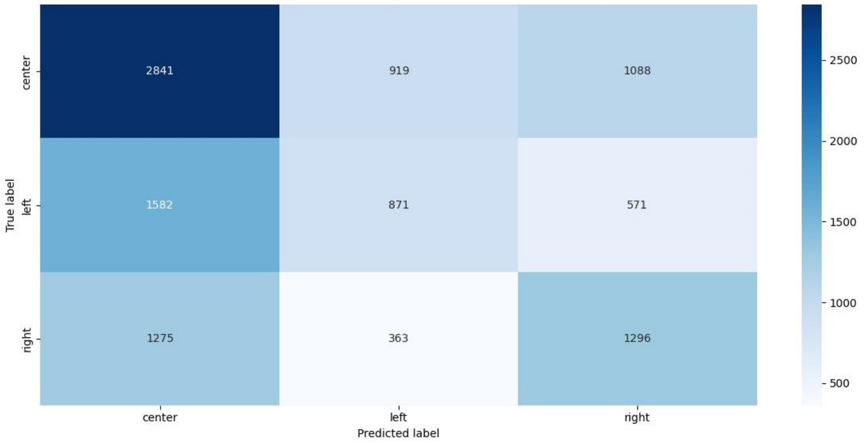

Figure 2: Confusion Matrix for Multiclass Classification of Logistic Regression Trained on Engineered Features

are feature engineering baseline with LR and Big Bird word embeddings with SVC. Their confusion matrixes are presented in Figure 2 and Figure 3.

Both algorithms are the most successful in accurately predicting the "center" class but struggle with correct predictions for the "left" class. Generally, both algorithms are more prone to predict any of the classes as "center" and least willing to assign any entry as "left." Notably, LR trained on feature engineering baseline demonstrate better results for correctly identifying individuals as left or right-leaning than Big Bird embeddings, resulting in superior performance in predicting the "left" and "right" classes. However, SVC with Big Bird embeddings is more successful in predicting the "center" class correctly, but this could be explained by its willingness to predict any entry as "center."

5.5 *SHapley Additive exPlanations values*

The best-performing model in terms of F1-score is the LR model trained on the feature engineering baseline. Therefore, this model is utilized to address RQ3 and explore the importance of features through SHAP values. Beeswarm graphs illustrating the nine most important features for the "left" category can be observed in Figure 4, "center" in Figure 5, and "right" in Figure 6.

The most important features for predicting the "left" class are the character trigrams " th", followed by the count of verbs and punctuations within the entry. Increased usage of the character trigram " th" and punctuations tends to correlate with a higher likelihood of being categorized as left-

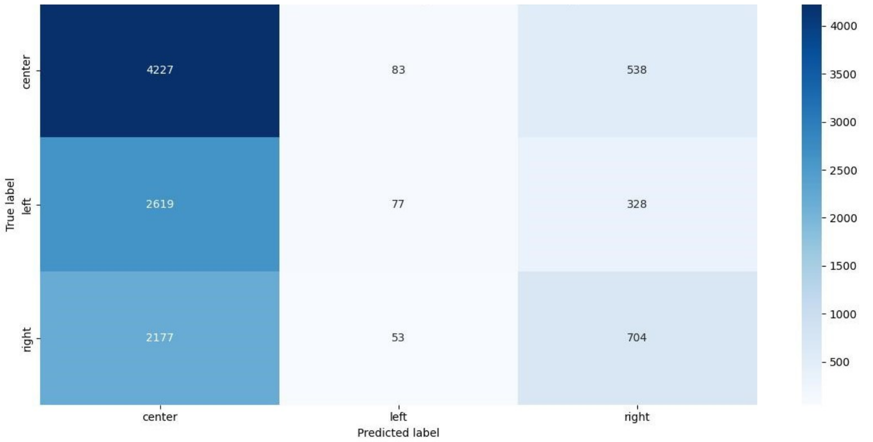

Figure 3: Confusion Matrix for Multiclass Classification of Support Vector Classifier Trained on Big Bird Word Embeddings

leaning, whereas a higher verb count indicates a lower probability of falling into this class.

Within the "center" class, the most important features based on their absolute SHAP values are character trigrams "’s ", "you", and " th". Smaller occurrence of these character trigrams in entry is associated with a higher probability of being assigned to the "center" category. In contrast to the "left" class, POS trigrams, such as "cconj propn punct" (coordinating conjunction, proper noun, and punctuation) and "noun punct cconj" (noun, punctuation, and coordinating conjunction), play a more significant role in predicting the "center" category.

Regarding the prediction of right economic political leaning, the key features include the character trigram "you", followed by the POS trigram "cconj propn punct" and the character trigram "the". Higher use of "you" and "the" and lower presence of the POS trigram "cconj propn punct" makes the learning algorithm more likely to predict the person as right-leaning. Compared to the "center" class, a higher frequency of character trigrams ", a", "ou ", and "t’s" indicates a reduced likelihood of being predicted as "right," while a higher occurrence of these features suggests a more center-leaning orientation.

6 DISCUSSION

The aim of this study was to evaluate various text representation techniques for predicting the political leaning of text authors using a Reddit dataset and three distinct classifiers. To ensure a concise and organized paper, the main research question was divided into three sub-questions focusing

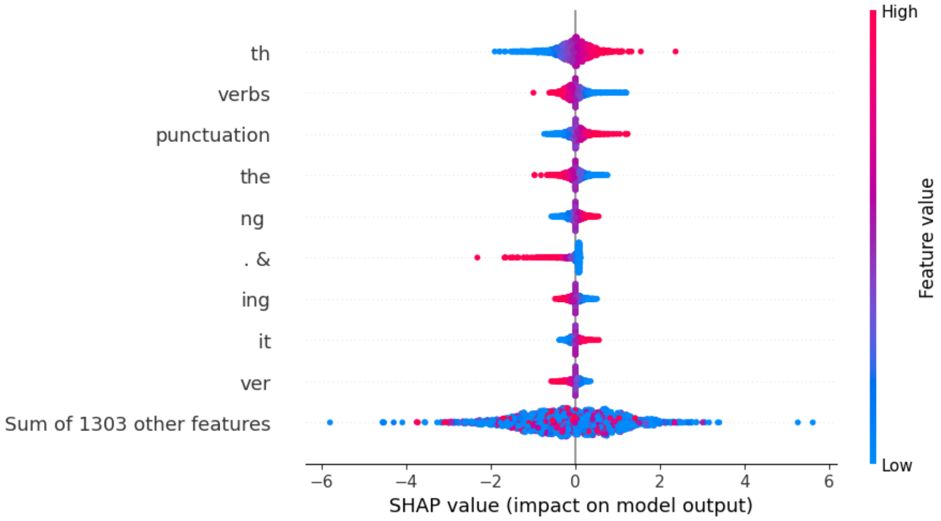

Figure 4: Beeswarm Graph of SHAP Values for the "Left" Class Taken from Logistic Regression Model Trained on Engineered Features

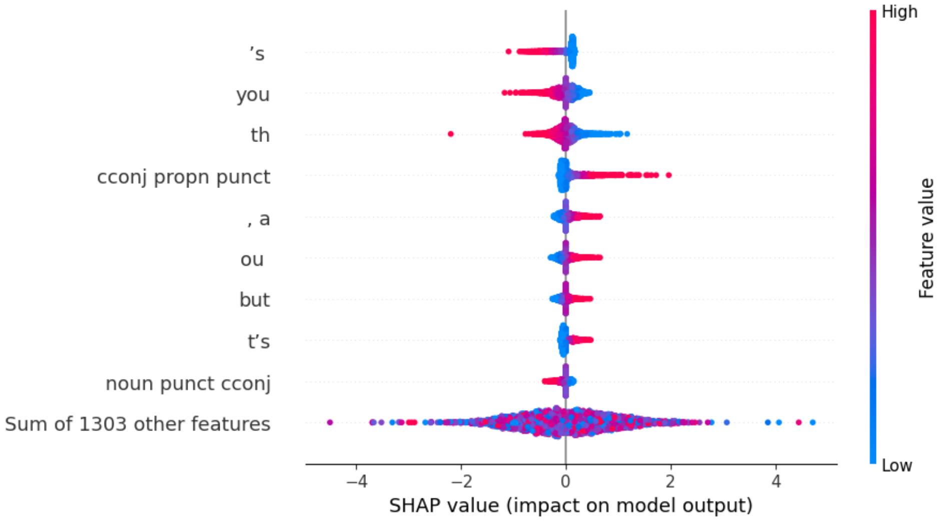

Figure 5: Beeswarm Graph of SHAP Values for the "Center" Class Taken from Logistic Regression Model Trained on Engineered Features

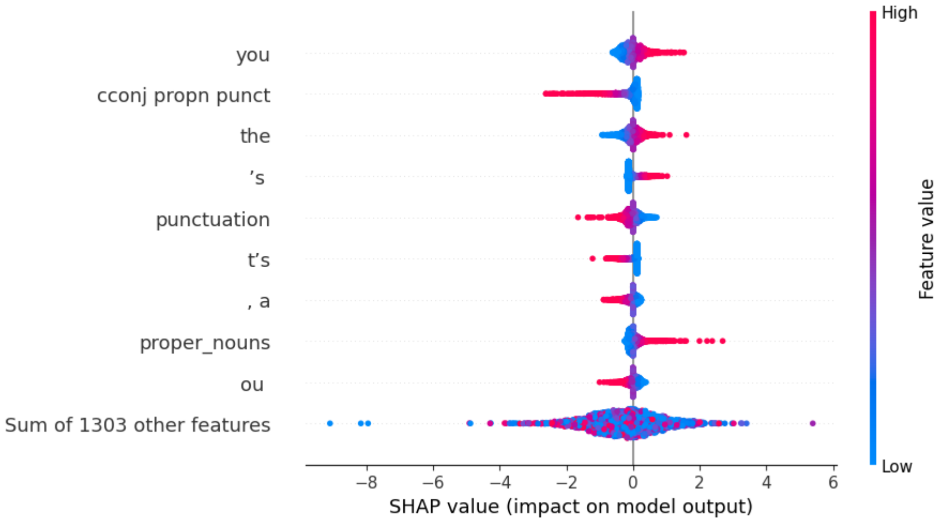

Figure 6: Beeswarm Graph of SHAP Values for the "Right" Class Taken from Logistic Regression Model Trained on Engineered Features

on word embeddings, classifiers, and model explainability. Following the methodology, the results were presented in the previous section. The discussion of the results and overall process has been conducted below.

6.1 *Results Discussion*

6.1.1 *Text Representations' Results Discussion*

The results of various word embeddings reveal that the lowest results are achieved on F1-scores. Big Bird embeddings show the lowest overall performance, followed closely by Word2Vec, while TF-IDF demonstrates the best performance. Explanations of varying text representations' performances are analysed below.

Firstly, Kramp et al. (2023) highlighted the potential of Big Bird embeddings in NLI tasks; however, this study did not validate their superiority. It is important to note that Kramp et al. (2023) worked on the NLI task, which differs from predicting political leaning and may require different techniques and yield different results. Furthermore, their study did not compare Big Bird embeddings against Word2Vec and TF-IDF. It is possible that TF-IDF could have also shown positive results in their research. Furthermore, the use of pre-trained roberta base[15] and not engaging in fine-tuning might have impacted the performance of the Big Bird embeddings.

[15] https://huggingface.co/google/bigbird-roberta-base

Secondly, employing pre-trained Word2Vec embeddings trained on a news dataset[16] could have hampered their performance. News and Reddit posts can contain diverse language styles, with news content often characterized by formal writing and language styles. The presence of slang and different rhetoric in Reddit posts may require different vectors for specific words, influencing the embeddings' performance.

Thirdly, on average, TF-IDF emerges as the best-performing text representation method across all performance metrics. TF-IDF was trained on the dataset employed in this research, allowing the algorithm to adapt dataset-specific patterns. This is likely to cause the better performance.

Fourthly, the feature engineering baseline model closely approaches TF-IDF performance. This underscores that more classical methods of text representation are still valuable. However, it is essential to acknowledge that this feature engineering method requires significantly longer time and greater computational power.

Hence, the analysis of the first research question underscores TF-IDF as the most successful word embedding for predicting the political leaning of Reddit users. However, its performance does not significantly differ from the feature engineering baseline.

6.1.2 *Classifiers' Results Discussion*

The average performance of classifiers shows more mixed results than word embedding results. Here, LR exhibits greater precision, while SVC excels more in recall and accuracy. Notably, GBDT scores the highest F1-score, which combines metrics of precision and recall and accounts for class imbalance that accuracy fails to achieve.

Nevertheless, the majority baseline displays the lowest performance across all metrics. Its accuracy is only 2.39% worse than the best classifiers' accuracy, while its F1-score is 15.21% worse than the best average performance. The baseline's reliance on assigning all instances to the majority class ("center") improved its accuracy due to class distribution imbalance. However, due to this imbalance, the F1-score showed worse results. The increase of the F1-score from the majority baseline to other classifiers suggests that other models are learning.

Therefore, when exploring the second research question, it remains ambiguous which classifier performs the best. However, all the algorithms outperform the majority baseline.

[16] https://code.google.com/archive/p/word2vec/

6.1.3 *Models' Results Discussion*

While it is important to identify patterns in word embeddings' and classifiers' performances, analysing the combinations of these elements holds significance as some interesting aspects appear.

Despite that, on average, TF-IDF outperforms other text representation techniques. The highest overall F1-score is achieved by the feature engineering baseline when used with LR. This furthermore highlights the success of using more traditional methods in prediction tasks.

Another interesting pattern to notice is that when averaging the results, GBDT stands out with the highest F1-score; however, this learning algorithm does not perform best with any of the text representation techniques when looking at separate results. LR without word embeddings and SVC with TF-IDF displays better F1-scores. Moreover, while SVC performed well with TF-IDF, it showcased the lowest F1-score when coupled with Big Bird word embeddings. This highlights the robustness of GBDT as a learning algorithm compared to SVC and the importance of analysing various combinations of text representation and classifiers.

6.1.4 *SHapley Additive exPlanations Value Result Discussion*

SHAP values reveal the importance of features for the best-performing model on the F1-score – LR trained on the feature engineering baseline. Also, here, interesting patterns emerge.

Examining the nine most important features across all classes, a mix of character trigrams, POS trigrams, and POS counts emerge as significant predictors. Different character trigrams appear to be the most important feature across all categories. Notably, often character trigrams appear to represent complete words, such as "you," "the," and "but." This suggests that word embeddings (used in other models) should play a significant role in predictions.

Certain features such as sentiment, average sentence length, and spelling mistakes do not rank among the nine most important features across all classes. This absence implies that there are no significant differences in these attributes across different economic political leanings.

However, differences between classes can be observed in other features. The model suggests that right-leaning individuals tend to use less punctuation and more word "the" than left-leaning. Furthermore, when comparing the "center" and "left" categories, "left" are less likely to use the character trigram " th" than center-leaning individuals. It is important to note that this would likely not indicate the beginning of the word "the" because the "left" category is not likely to use this trigram.

Moreover, the model captures more distinct differences between the "right" and "center" categories, particularly in the usage of character trigrams such as "you", ", a", "ou ", and "t's". This pattern could be attributed to the greater representation of the "right" and "center" categories compared to the "left," leading the algorithm to draw more comparisons between these two groups.

Even though SHAP values reveal interesting findings, they have to be analysed carefully as the performance metrics achieved in this study fall below those reported by other researchers (Conover et al., 2011; Jiang et al., 2023; Pennacchiotti & Popescu, 2021). Several aspects may have contributed to this.

Firstly, prior research primarily focused on other datasets, particularly emphasizing the Twitter dataset. In contrast, the dataset used in this paper has not yet been fully reviewed. This difference in datasets poses a challenge for comparison, as varied datasets may require diverse methods.

Secondly, a considerable amount of research has been concentrated on binary classification, such as NLI by Kramp et al. (2023), and republican or democrat political leaning prediction by Pennacchiotti and Popescu (2021). In contrast, besides left and right political leaning, this research also employed center class to acknowledge that political leaning is more of a spectrum and a significant middle exists. The inclusion of this additional class made the difference between classes less noticeable and, therefore, is likely to cause poorer performance. This was also seen in confusion matrixes, where all algorithms were more prone to predicting any class as "center."

Thirdly, previous research also pointed out the importance of incorporating features other than text representation, for instance, Conover et al. (2011) integrated hashtags and Volkova et al. (2014) explored at network patterns. Even though this could be more significant for the Twitter dataset, it is possible that analysing similar features in the Reddit dataset could reveal more about one's economic political leaning.

Finally, this research did not engage in hyperparameter tuning (also further discussed in 6.3). While optimizing hyperparameters could potentially enhance the results, it is unlikely that they would reach the levels reported in the existing literature that mostly exceeded 80% in performance metrics (Conover et al., 2011; Jiang et al., 2023; Pennacchiotti & Popescu, 2021). Moreover, if algorithms reached this accuracy or F1-score measurement after tuning, it could potentially indicate that the learning algorithms are less robust.

6.2 *Scientific and Social Impact*

The results of this research may appear statistically modest; however, they carry significant value with implications in both the scientific and social realms.

As mentioned before, analysing political discourse is relevant and vital for society and its security. Understanding the stance of society is crucial for politicians, leading to deeper understanding and fostering a more harmonious political environment (Chambers, 2018). Additionally, this research shows the diverse nature of political ideologies, emphasizing that political leanings exist along a spectrum rather than as a binary classification. While incorporating more categories may make the prediction harder, it more accurately resembles the diversity within society, where individuals often align with various points along the political spectrum based on different topics, with a substantial center community.

This also adds to the scientific community and points out that other researchers should also take into account this detail. Taking into account the "center" class increases the understanding of the distribution of classes and broadens the research on economic political leaning prediction. Moreover, this research contributes to the research of Reddit as a dataset and diverse word embeddings, especially the under-researched field of Big Bird word embeddings.

6.3 *Limitations*

Even though this research has addressed three research questions and implies social and scientific significance, it is important to acknowledge several limitations.

Firstly, the research on English machine learning models has been vast, and there is a lack of exploration of other languages. Also, this research adopted a data cleaning method that filtered out non-English entries and applied dictionaries for sentiment and spelling mistakes that are based on English. This also led to the choice of pre-trained Word2Vec and Big Bird models that are trained on English datasets. Even though the choices were made because most of the dataset was in English, this makes the developed algorithms not suitable for other languages.

Secondly, the decision not to engage in hyperparameter tuning while trying to facilitate a clearer comparison of different word embeddings, as already mentioned, may have hampered the results. Hyperparameter tuning could have optimized the parameters for each word embedding and classifier, potentially enhancing performance. However, this would have made comparison of word embeddings less obvious.

Thirdly, previous research showed promising results for Big Bird embeddings (Kramp et al., 2023); however, in this research, Big Bird embeddings showed the worst results. Kramp et al. (2023) used and compared different Big Bird models, but this research only chose the best model of Kramp et al. (2023) research. Additional exploration of different Big Bird models, perhaps tailored to this specific research context, could have potentially shown better results.

Fourthly, both Word2Vec and roberta base for Big Bird embeddings utilized in the study were pre-trained. As mentioned before, training models on this dataset or using other pre-trained models could have been more suitable for this research. Both of the pre-trained models were trained on news[17] [18], while Big Bird also used stories, books, and Wikipedia. The language of these sources differs from the informal, user-generated language typically found on Reddit, potentially leading to a mismatch in language styles and affecting the performance of the embeddings on Reddit data.

Lastly, while the dataset used in the research has been developed by established scholars (Emmery et al., 2024), it could have limitations. The dataset is still under review and has limited utilization in other studies. This introduces challenges in comparing the findings with existing research, underscoring the need for caution in drawing conclusions based on this dataset alone.

6.4 *Future Research*

Despite the limitations of this research, it paves the way for future research opportunities.

The worst results were observed with Word2Vec and Big Bird embeddings, which the use of pre-trained models may have influenced. One promising field for further investigation is the potential of Reddit-specific Word2Vec and Big Bird models. This research would capture Reddit users' diverse slang and language nuances, offering valuable insights for future Reddit-related research.

Moreover, despite Big Bird embeddings being the worst results, this area remains relatively unexplored and warrants further exploration. By fine-tuning the algorithms and experimenting with different model bases, Big Bird embeddings could potentially emerge as a fast and efficient word embedding technique.

Additionally, as this research shows that a more traditional method of feature engineering baseline still can compete with more advanced

[17] https://huggingface.co/google/bigbird-roberta-base
[18] https://code.google.com/archive/p/word2vec/

techniques and SHAP values suggest that character trigrams, often being words, play an important role in predictions, it would be valuable to explore the combination of feature engineering baseline and different word embeddings. For example, combining well-performing engineered features with TF-IDF word embeddings has the potential to boost the algorithm's F-score. Moreover, exploring diverse combinations of word embeddings could capture various aspects of textual information, thereby enriching the overall analysis.

7 CONCLUSION

This research explored algorithms' predictive capabilities in determining Reddit users' political leanings. The main research question was "To what extent can machine learning classifiers, leveraging various text representations, accurately predict the economic political leaning of Reddit users based on their textual posts?"

Based on previous research and identified research gaps, a methodology was employed. This involved the exploration of three distinct word embedding types – TD-IDF, Word2Vec, and Big Bird – across three different classifiers: LR, SVC, and GBDT. The algorithms were benchmarked against both the majority baseline and the feature engineering baseline, trained on all three classifiers, resulting in the development of 13 unique models.

The main experiment results reached F1-scores between 37.00% and 45.00%. Notably, the most promising results were achieved with the feature engineering embeddings on LR, where SHAP values highlighted the importance of character trigrams and POS, and the TF-IDF embeddings on SVC. While TF-IDF embeddings demonstrated the best performance compared to other embeddings, the classifier results exhibited a more varied performance.

While the results may appear modest in comparison to previous studies, it is essential to consider the contextual disparities between research methodologies, including variations in datasets and the incorporation of a "center" class. Hence, through the exploration of diverse word embeddings and classifiers, this research has made a meaningful contribution to the study of author profiling.

REFERENCES

Ahuja, R., & Sharma, S. C. (2022). Transformer-based word embedding with cnn model to detect sarcasm and irony. *Arabian Journal for Science and Engineering*, 47(8), 9379–9392.

Atari, M., Davani, A. M., Kogon, D., Kennedy, B., Ani Saxena, N., Anderson, I., & Dehghani, M. (2022). Morally homogeneous networks and radicalism. *Social Psychological and Personality Science*, *13*(6), 999–1009.

Aydin, A. (2023). Stemming & lemmatization in nlp: Text preprocessing techniques. https://ayselaydin.medium.com/2-stemming-lemmatization-in-nlp-text-preprocessing-techniques-adfe4d84ceee

Baly, R., Da San Martino, G., Glass, J., & Nakov, P. (2020). We can detect your bias: Predicting the political ideology of news articles. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)* (pp. 4982–4991). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.404

Bugdani, T. (2022). Pos tagging in nlp using spacy. https://www.askpython.com/python/examples/pos-tagging-in-nlp-using-spacy

Chambers, S. (2018). *Reasonable democracy: Jürgen habermas and the politics of discourse*. Cornell University Press.

Conover, M. D., Gonçalves, B., Ratkiewicz, J., Flammini, A., & Menczer, F. (2011). Predicting the political alignment of twitter users. *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*, 192–199.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*, 273–297.

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *20*(2), 215–232.

Cramer, J. S. (2002). The origins of logistic regression [Tinbergen institute working paper].

Emmery, C., Miotto, M., Kramp, S., & Kleinberg, B. (2024). SOBR: A corpus for stylometry, obfuscation, and bias on reddit. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources, and Evaluation*.

Falini, A. (2022). A review on the selection criteria for the truncated svd in data science applications. *Journal of Computational Mathematics and Data Science*, *5*, 100064. https://doi.org/10.1016/j.jcmds.2022.100064

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189–1232. http://www.jstor.org/stable/2699986

Goldin, G., Rabinovich, E., & Wintner, S. (2018, October). Native language identification with user generated content. In E. Riloff, D. Chiang, J. Hockenmaier, & J. Tsujii (Eds.), *Proceedings of the 2018 conference* *on empirical methods in natural language processing* (pp. 3591–3601). Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1395

Greene, L. S., Inniss, L. B., Crawford, B. J., Baradaran, M., Ben-Asher, N., Capers, I. B., James, O. R., & Lindsay, K. (2019). Talking about black lives matter and metoo. *Wisconsin Journal of Law, Gender & Society*, *34*.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., . . . Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Honnibal, M., & Montani, I. (2017). *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Jiang, J., Ren, X., & Ferrara, E. (2023). Retweet-bert: Political leaning detection using language features and information diffusion on social networks. *17*, 459–469. https://doi.org/10.1609/icwsm.v17i1.22160

Kosiv, Y., & Yakovyna, V. (2022). Three language political leaning text classification using natural language processing methods. *Applied Aspects of Information Technology*, *5*, 359–370. https://doi.org/10.15276/aait.05.2022.24

Kramp, S., Cassani, G., & Emmery, C. (2023). Native language identification with big bird embeddings. https://api.semanticscholar.org/CorpusID:261705924

Kwak, C., & Clayton-Matthews, A. (2002). Multinomial logistic regression. *Nursing research*, *51*(6), 404–410.

Li, L., & Chan, W. (2019). Big bidirectional insertion representations for documents. https://arxiv.org/abs/1910.13034

Lin, T., Wang, Y., Liu, X., & Qiu, X. (2022). A survey of transformers. *AI Open*, *3*, 111–132. https://doi.org/https://doi.org/10.1016/j.aiopen.2022.10.001

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems 30* (pp. 4765–4774). Curran Associates, Inc. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

Lutoslawski, W. (1898). Principes de stylométrie appliqués a la chronologie des œuvres de platon. *Revue des Études Grecques*, *11*(41), 61–81. Retrieved June 8, 2024, from http://www.jstor.org/stable/44283556

Magdy, W., Darwish, K., & Weber, I. (2015). # Failedrevolutions: Using twitter to study the antecedents of isis support. https://arxiv.org/abs/1503.02401

Mckinney, W. (2011). Pandas: A foundational python library for data analysis and statistics. *Python High Performance Science Computer*.

Mendenhall, T. C. (1887). The characteristic curves of composition. *Science*, *9*(214), 237–249. Retrieved June 8, 2024, from http://www.jstor.org/stable/1764604

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. 26). Curran Associates, Inc. https://proceedings.neurips.cc/paper%5C_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf

Mosteller, F., & Wallace, D. L. (1963). Inference in an authorship problem. *Journal of the American Statistical Association*, *58*(302), 275–309. https://doi.org/10.1080/01621459.1963.10500849

Neal, T., Sundararajan, K., Fatima, A., Yan, Y., Xiang, Y., & Woodard, D. (2017). Surveying stylometry techniques and applications. *ACM Comput. Surv.*, *50*(6). https://doi.org/10.1145/3132039

Nikolaev, D., Ceron, T., & Padó, S. (2023). Multilingual estimation of political-party positioning: From label aggregation to long-input transformers. https://arxiv.org/abs/2310.12575

Ouni, S., Fkih, F., & Omri, M. N. (2023). A survey of machine learning-based author profiling from texts analysis in social networks. *Multimedia Tools and Applications*, *82*(24), 36653–36686.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., & Lerer, A. (2017). Automatic differentiation in pytorch. *NIPS-W*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pennacchiotti, M., & Popescu, A.-M. (2021). A machine learning approach to twitter user classification. *Proceedings of the International AAAI Conference on Web and Social Media*, *5*(1), 281–288. https://doi.org/10.1609/icwsm.v5i1.14139

Pericherla, S., & Ilavarasan, E. (2024). Transformer network-based word embeddings approach for autonomous cyberbullying detection. *International Journal of Intelligent Unmanned Systems*, *12*(1), 154–166.

Preoţiuc-Pietro, D., Liu, Y., Hopkins, D., & Ungar, L. (2017, July). Beyond binary labels: Political ideology prediction of Twitter users. In R. Barzilay & M.-Y. Kan (Eds.), *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 729–740). Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-1068

Proferes, N., Jones, N., Gilbert, S., Fiesler, C., & Zimmer, M. (2021). Studying reddit: A systematic overview of disciplines, approaches, methods, and ethics. *Social Media+ Society*, *7*(2), 20563051211019004.

Rehurek, R., & Sojka, P. (2011). Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, *3*(2).

Shuyo, N. (2021). Langdetect - language detection library ported from google's language-detection. https://pypi.org/project/langdetect/

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, *28*(1), 11–21.

Stefanov, P., Darwish, K., Atanasov, A., & Nakov, P. (2020). Predicting the topical stance and political leaning of media using tweets. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 527–537.

Sterling, J., Jost, J. T., & Bonneau, R. (2020). Political psycholinguistics: A comprehensive analysis of the language habits of liberal and conservative social media users. *Journal of personality and social psychology*, *118*(4), 805.

The Political Compass. (n.d.). Politicalcompass.org. https://www.politicalcompass.org/about

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper%5C_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Volkova, S., Coppersmith, G., & Van Durme, B. (2014). Inferring user political preferences from streaming communications. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 186–196.

Waskom, M. L. (2021). Seaborn: Statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021

Wolfsfeld, G., Segev, E., & Sheafer, T. (2013). Social media and the arab spring: Politics comes first. *The International Journal of Press/Politics*, *18*(2), 115–137.

Wong, F. M. F., Tan, C. W., Sen, S., & Chiang, M. (2021). Quantifying political leaning from tweets and retweets. *Proceedings of the International AAAI Conference on Web and Social Media*, *7*(1), 640–649. https://doi.org/10.1609/icwsm.v7i1.14422

Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L., & Ahmed, A. (2020). Big bird: Transformers for longer sequences. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (pp. 17283–17297, Vol. 33). Curran Associates, Inc. https://proceedings.neurips.cc/paper%5C_files/paper/2020/file/c8512d142a2d849725f31a9a7a361ab9-Paper.pdf

APPENDIX A

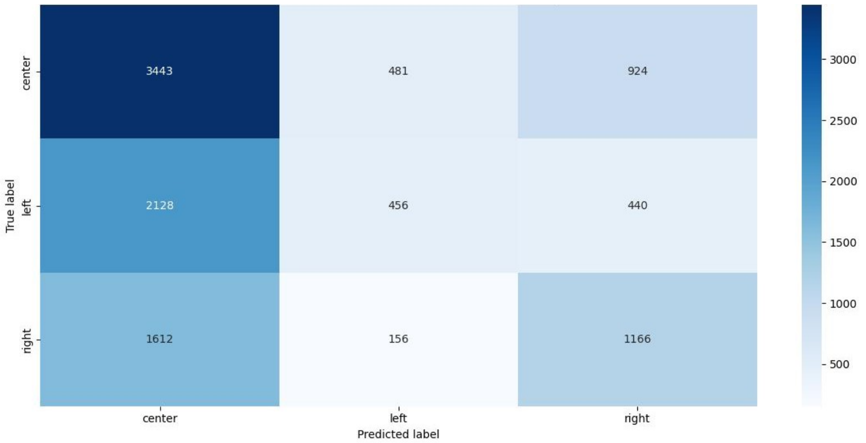

Figure 7: Confusion Matrix for Multiclass Classification of Support Vector Classifier Trained on Engineered Features

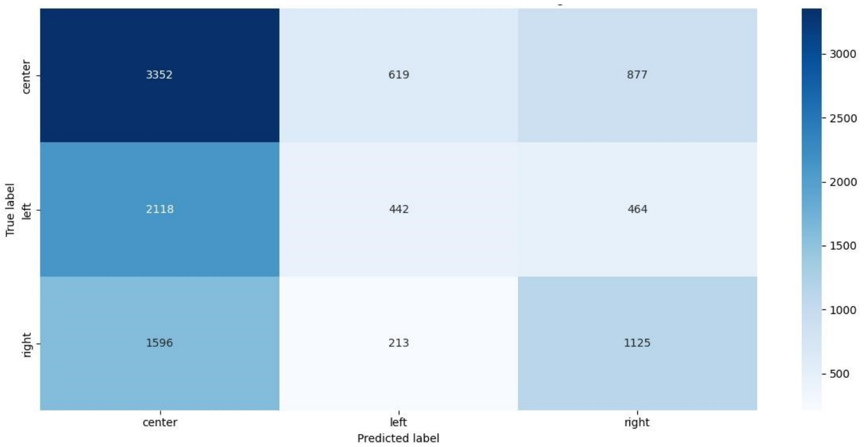

Figure 8: Confusion Matrix for Multiclass Classification of Gradient Boosted Decision Trees Trained on Engineered Features

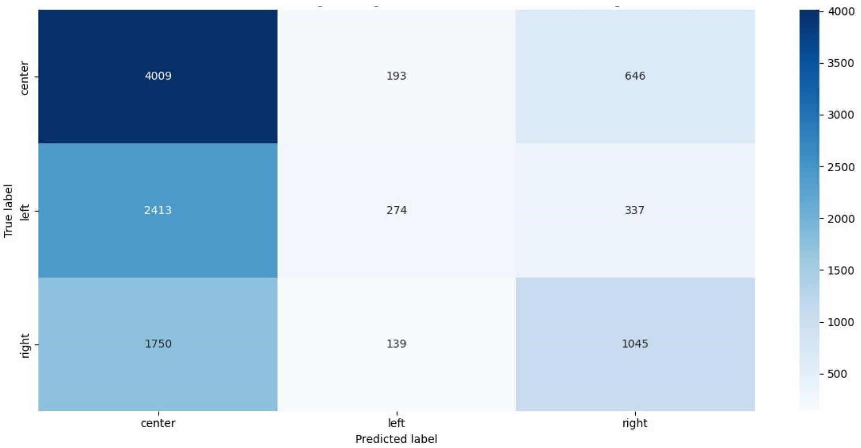

Figure 9: Confusion Matrix for Multiclass Classification of Logistic Regression Trained on Term Frequency-Inverse Document Frequency Word Embeddings

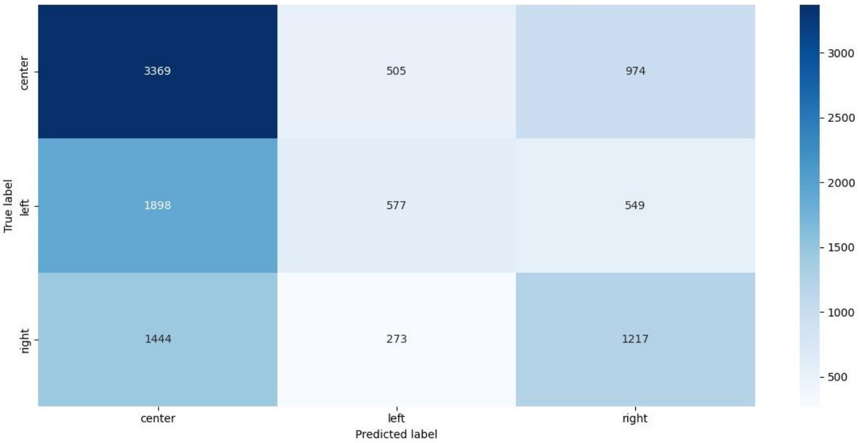

Figure 10: Confusion Matrix for Multiclass Classification of Support Vector Classifier Trained on Term Frequency-Inverse Document Frequency Word Embeddings

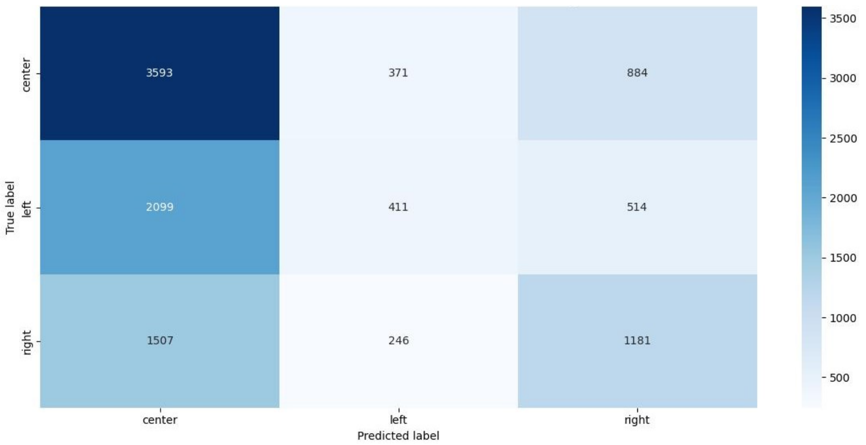

Figure 11: Confusion Matrix for Multiclass Classification of Gradient Boosted Decision Trees Trained on Term Frequency-Inverse Document Frequency Word Embeddings

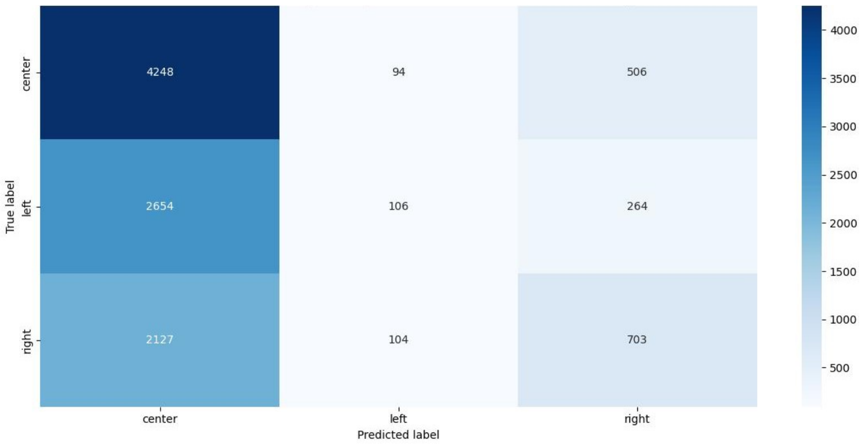

Figure 12: Confusion Matrix for Multiclass Classification of Logistic Regression Trained on Word2Vec Word Embeddings

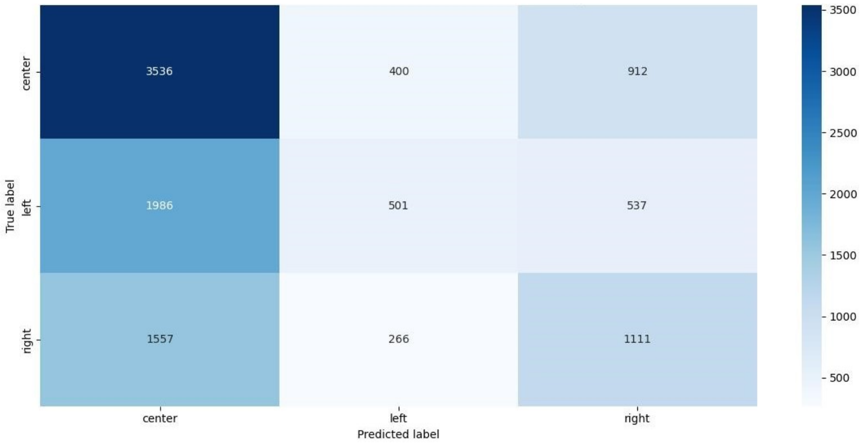

Figure 13: Confusion Matrix for Multiclass Classification of Support Vector Classifier Trained on Word2Vec Word Embeddings

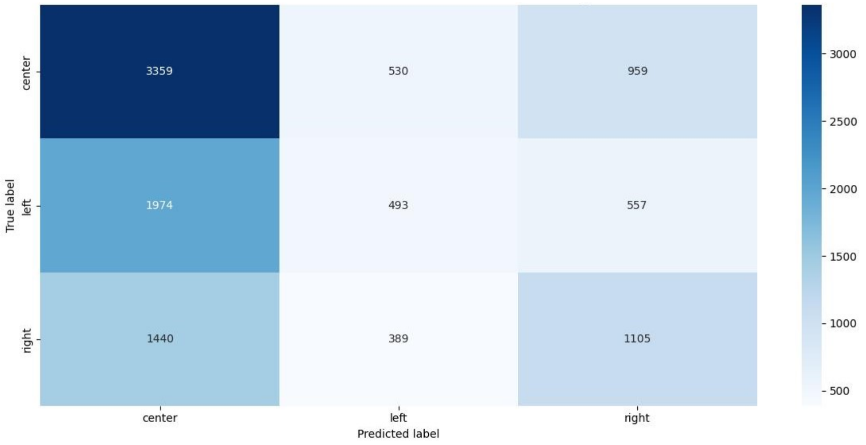

Figure 14: Confusion Matrix for Multiclass Classification of Gradient Boosted Decision Trees Trained on Word2Vec Word Embeddings

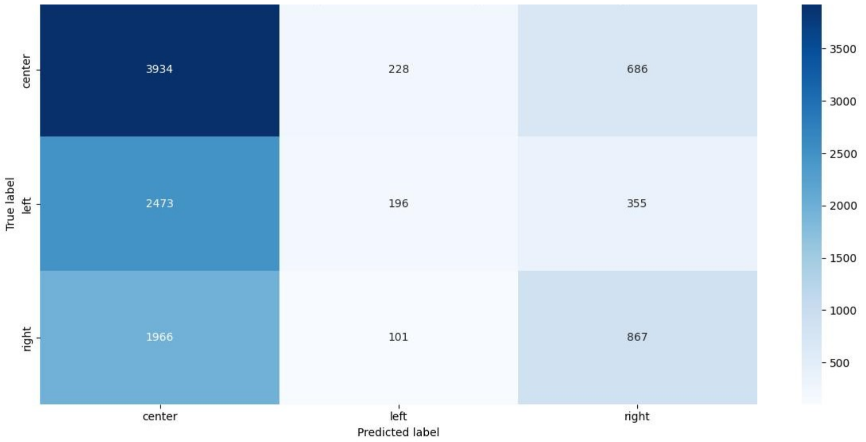

Figure 15: Confusion Matrix for Multiclass Classification of Logistic Regression Trained on Big Bird Word Embeddings

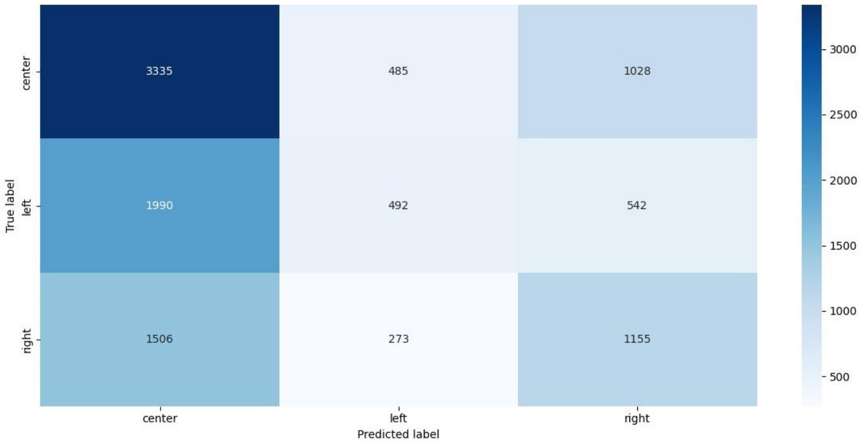

Figure 16: Confusion Matrix for Multiclass Classification of Gradient Boosted Decision Trees Trained on Big Bird Word Embeddings