Natural Language Processing

Liberals vs Conservatives

Matteo Marturini

MAT. SM3500484

M.Sc. Data Science and Scientific Computing

University of Trieste

July 22, 2022

Contents

1 Problem statement

Social media platforms like Facebook (2.936 billion users), Instagram (1.21 billion users) and Twitter (229 million million users) have become the predominant way people, businesses and organizations communicate: companies nowadays invest a lot of resources into social media advertising, while institutional organizations and public figures share information and opinions that can influence millions of people all around the world with just a couple of sentences. Group of researchers suggest that social medias foster political polarization by creating “echo chambers” that insulate people from opposing views about current events.

The objective of this project is to apply Natural Language Processing techniques to social media posts labelled as "Liberal" or "Conservative" to analyze whether major differences in the topics of discussion and the sentiment related to it exist between the two political leans.

Being able to automatically identify the political lean of a post may be useful for several reasons:

- to check whether a "neutral" organization is instead following a political agenda.
- together with hate speech analysis to identify if some political parties are associated to persecutions of specific groups.
- together with social media policies to guarantee free speech and avoid censorship just based on the political lean.

2 Dataset

The dataset consists of 12854 posts from Liberal and Conservative leaning subreddits (Kaggle). Reddit is a social media platform where users submit content to the site such as links, text posts, images, and videos, which are then voted up or down by other members. Posts are organized by subject into user-created boards called "communities" or "subreddits". Submissions with more upvotes appear towards the top of their subreddit and, if they receive enough upvotes, ultimately on the site’s front page.

2.1 Exploratory Data Analysis

A reddit post is composed of a title section and a body section where an image, video, link or text can be published. The posts of the dataset are organized into rows of a table with the following columns: ’Title’, ’Political Lean’, ’Score’, ’Id’, ’Subreddit’, ’URL’, ’Num of Comments’, ’Text’, ’Date Created’. Apart from the initial exploratory data analysis phase, the Natural Language Processing techniques are applied to the ’Title’ column appended with the ’text’ one when present and the ’Political Lean’ as a label category; all the remaining columns are not considered.

The ’Political lean’ label was assigned by the creator of the dataset according to the subreddit the post belongs to. In figure 1 are reported the percentages of ’Liberal’ vs ’Conservative’ posts; in figure 2 and 3 are reported respectively the name of the subreddits associated to the ’Liberal’ and ’Conservative’ label, and the corresponding percentage of posts.

2.2 Preprocessing

To mantain relevant information encoded in the posts while minimizing the variance, the following pre-processing techniques are applied:

- graphical emojis removal.
- textual emojis removal.
- URL links removal.
- punctuation removal.
- lemmatization.
- digits removal.
- non-aplhanumeric symbols removal.
- lower casing.
- stop words removal.

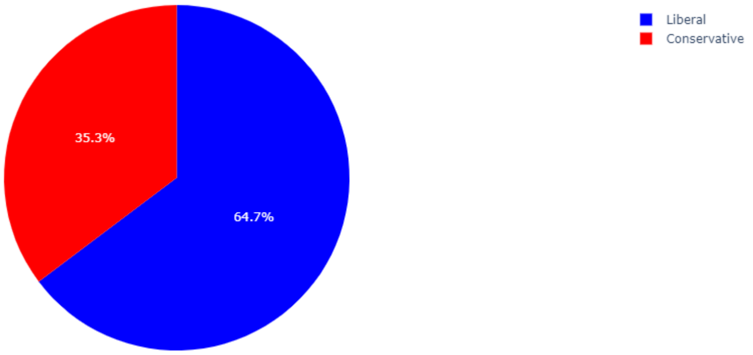


Figure 1: Percentages of 'Liberal' vs 'Conservative' posts.

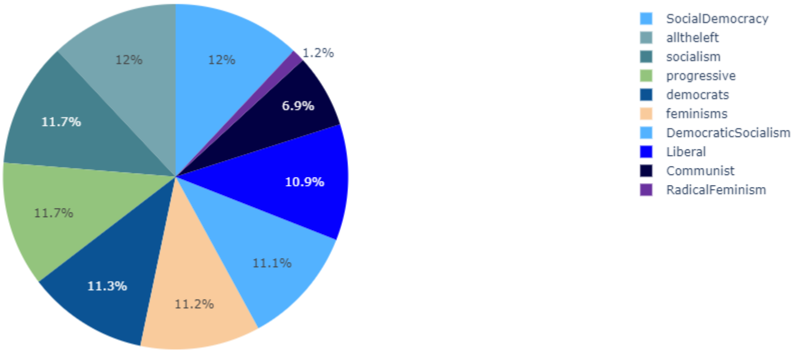


Figure 2: 'Liberal' subreddits.

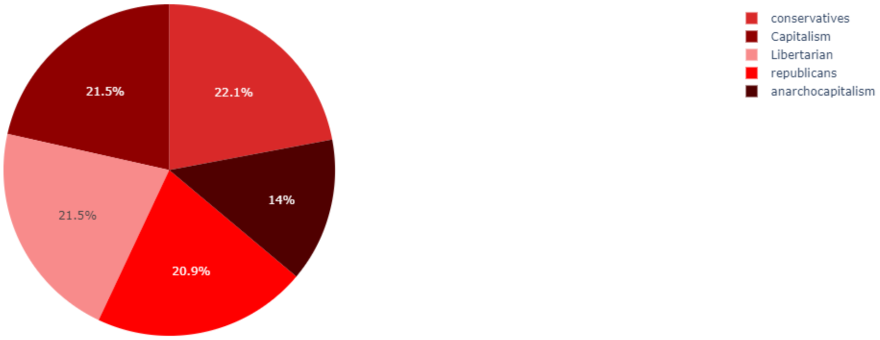


Figure 3: 'Conservative' subreddits.

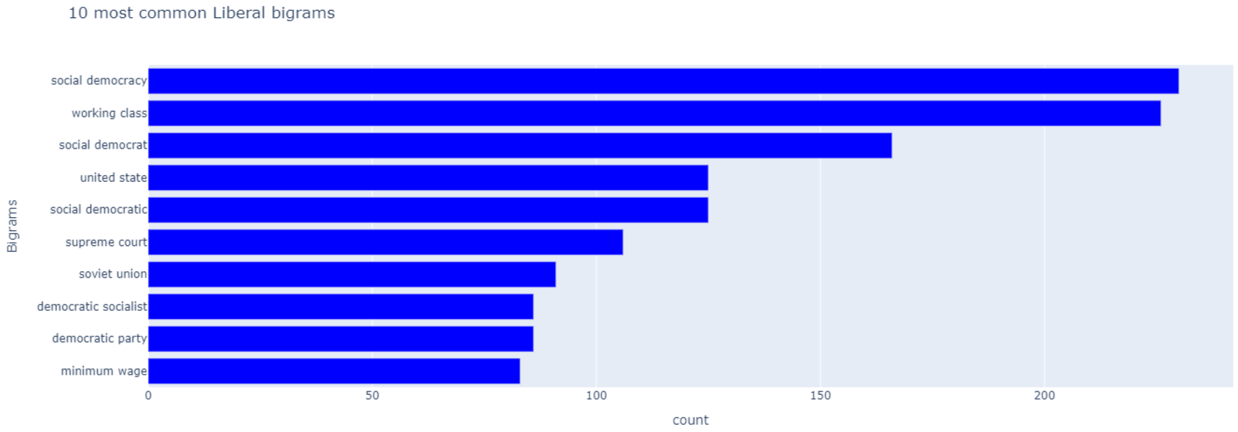


Figure 4: 'Liberal' Bigrams.

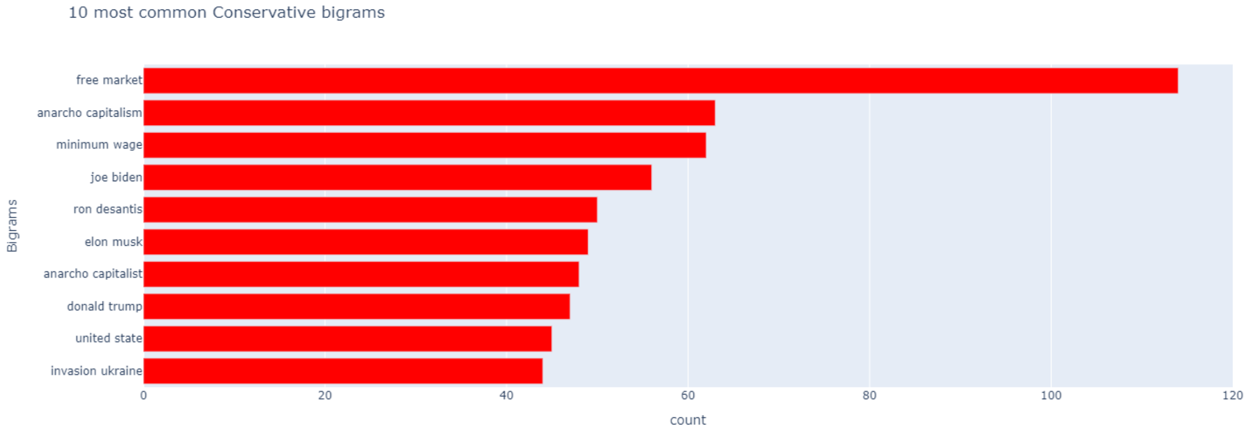


Figure 5: 'Conservative' Bigrams.

2.3 Bigrams analysis

In figure 4 and figure 5 are reported respectively the 10 most common Bigrams for liberals and conservatives.

The bigrams slightly differ, as for the 'Conservative' label it can be noticed that the 2 most common bigrams are 'free market' and 'anarcho capitalism', which are strongly associated to the Republican party, while for the 'Liberal' label the 2 most common ones are 'social democracy' and 'working class', which are historically associated to the Democrats party. However, similarities are also present: namely, 'minimum wage' and 'united state' are frequent bigrams which are present in both groups. This preliminary result suggest that even though the 2 political leans share some topics of discussion and words used, differences do exist and are noticeable.

In the 10 most common liberal bigrams, it can also be noticed that there are different bigrams that convey the same meaning, but differ just because of the morphology of the word, such as 'social democracy', 'social democrat', 'social democratic' and 'democratic socialist'. Moreover, in the conservative bigrams, it is common to find names like 'Elon Musk', 'Donald Trump' or 'Joe Biden'. These two outcomes suggest that two additionally preprocessing techniques are needed: stemming and the joining of collocations.

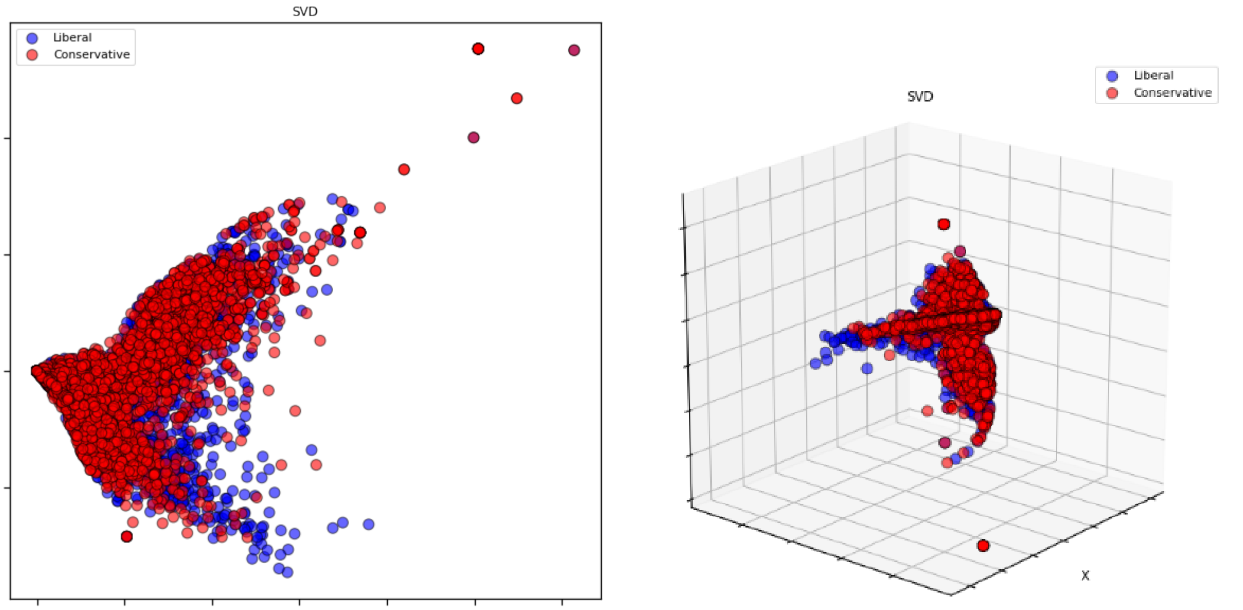


Figure 6: SVD applied to the TF-IDF data matrix for visualization in 2D and 3D space.

3 Visualization

After the preprocessing phase, a data matrix is obtained by applying TF-IDF vectorization. Then, dimensionality reduction techniques are applied to the data matrix in order to visualize the documents in the 2D and 3D space, and analyze differences and similarities between the two sets of documents.

The first technique applied is Singular Value Decomposition (SVD): the goal is to decompose the data matrix of n documents and m terms into 3 matrices:

$$X_{n\times m} = U_{n\times k} S_{k\times k} (V_{m\times k})^T \tag{1}$$

where k represents the number of latent concepts. If k is chosen to be 2 or 3, the rows of the matrix U can be plotted respectively in 2D and 3D space, as done in figure 6 for our dataset.

From these visualizations it can be observed that the 2 sets of documents strongly overlap, suggesting that to identify meaningful differences, an higher dimensional analysis may be needed. Nevertheless we can infer from these plots that there exist important latent factors which are in common between the 2 political leans, and that the words used have a high degree of similarity.

The second dimensionality reduction technique applied is Non-Negative Matrix Factorization (NMF), according to which the data matrix is decomposed as the following:

$$X_{n\times m} = W_{n\times k} (H_{k\times m})^T \tag{2}$$

where W gives us a lower-dimensional view of the documents, while H a lower-dimensional view of the terms. Following an anlogous reasoning as for SVD, we can plot the documents in 2D and 3D space, as reported in figure 7.

Consistently to what obtained with SVD, the 2 sets of documents strongly overlap, with minor differences among the groups.

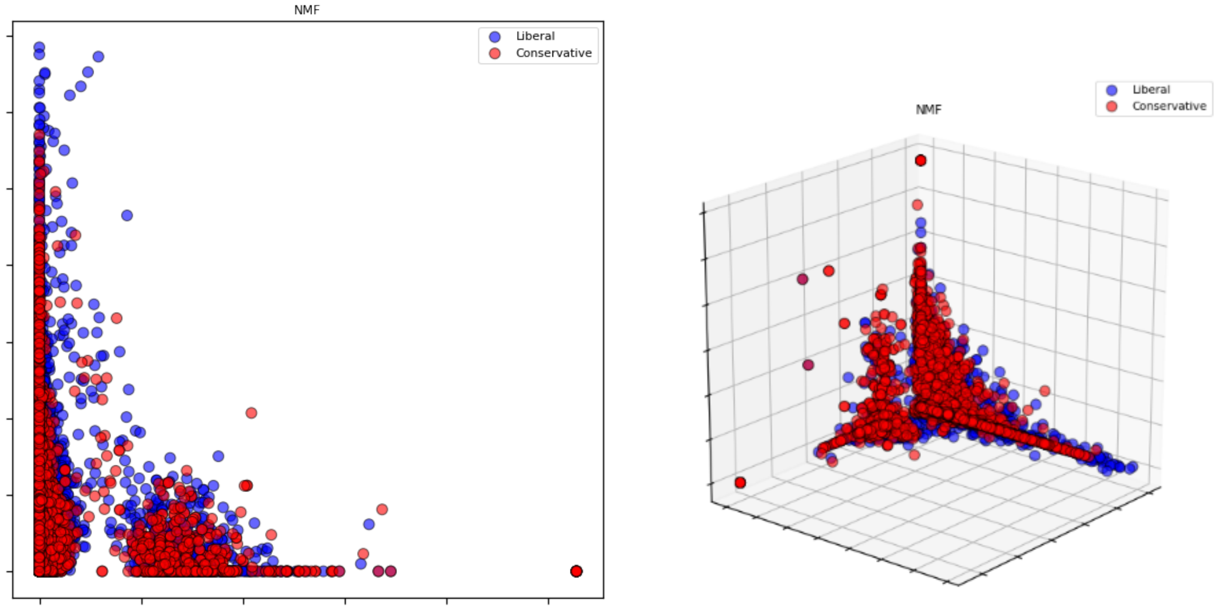
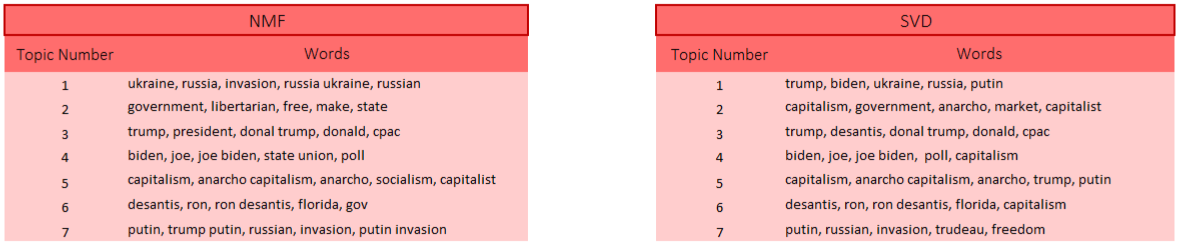


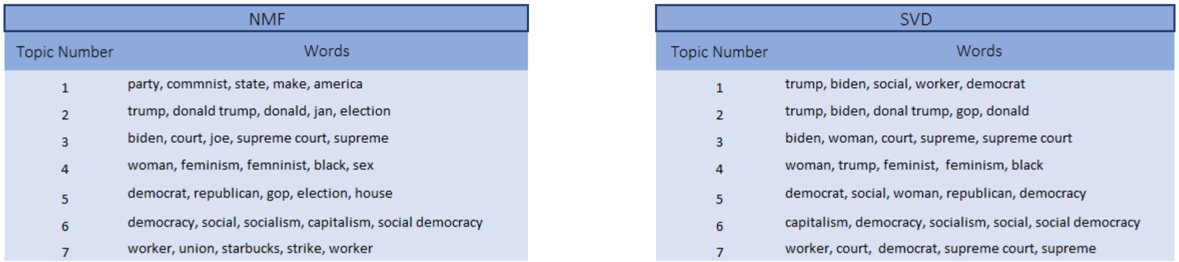
Figure 7: NMF applied to the TF-IDF data matrix for visualization in 2D and 3D space.



NMF	
Topic Number	Words
1	ukraine, russia, invasion, russia ukraine, russian
2	government, libertarian, free, make, state
3	trump, president, donal trump, donald, cpac
4	biden, joe, joe biden, state union, poll
5	capitalism, anarcho capitalism, anarcho, socialism, capitalist
6	desantis, ron, ron desantis, florida, gov
7	putin, trump putin, russian, invasion, putin invasion

SVD	
Topic Number	Words
1	trump, biden, ukraine, russia, putin
2	capitalism, government, anarcho, market, capitalist
3	trump, desantis, donal trump, donald, cpac
4	biden, joe, joe biden, poll, capitalism
5	capitalism, anarcho capitalism, anarcho, trump, putin
6	desantis, ron, ron desantis, florida, capitalism
7	putin, russian, invasion, trudeau, freedom

Figure 8: 'Conservative' Matrix Factorization latent topics.



NMF	
Topic Number	Words
1	party, commnist, state, make, america
2	trump, donald trump, donald, jan, election
3	biden, court, joe, supreme court, supreme
4	woman, feminism, femninist, black, sex
5	democrat, republican, gop, election, house
6	democracy, social, socialism, capitalism, social democracy
7	worker, union, starbucks, strike, worker

SVD	
Topic Number	Words
1	trump, biden, social, worker, democrat
2	trump, biden, donal trump, gop, donald
3	biden, woman, court, supreme, supreme court
4	woman, trump, feminist, feminism, black
5	democrat, social, woman, republican, democracy
6	capitalism, democracy, socialism, social, social democracy
7	worker, court, democrat, supreme court, supreme

Figure 9: 'Liberal' Matrix Factorization latent topics.

4 Topic Modeling

4.1 Matrix Factorization

To get a first general idea of the topics discussed in the 2 political leans, matrix factorization is applied: the top 5 words associated to a latent topic are extracted from matrix V in case of SVM and matrix H in case of NMF. The results are reported in figure 8 and figure 9 respectively for the 'Conservative' and 'Liberal' label.

This first result corroborates the idea visualized in section 3 and suggested in section 2.3 according to which the two political leans use similar words and share some topics of discussion; for example, two common subjects between the groups are a topic related to Donald Trump (topic 3 in conservative NMF and topic 2 in liberal NMF) and a topic related to Joe Biden (topic 4 conservative NMF and topic 3 lliberal NMF). However, this outcome also highlights the difference in group-specific topics, like 'woman, feminism, femninist, black, sex' which may refer to human rights for gender and ethnicity related groups, specific of the 'Liberal' group, and the topic 'desantis, ron, ron desantis, florida, gov' which refers to Ron DeSantis, the political governor of Florida, specific to the 'Conservative' label.

An unexpected result is that only conservatives seems to talk about the russian invasion of Ukraine, as suggested by topic 1 of both the conservative NMF and SVD, while this topic is not present in the liberal topics. A possible reason may be be the fact that the creator of the dataset obtained posts relative to the different groups in different time periods, influencing the relevant subjects people were discussing about.

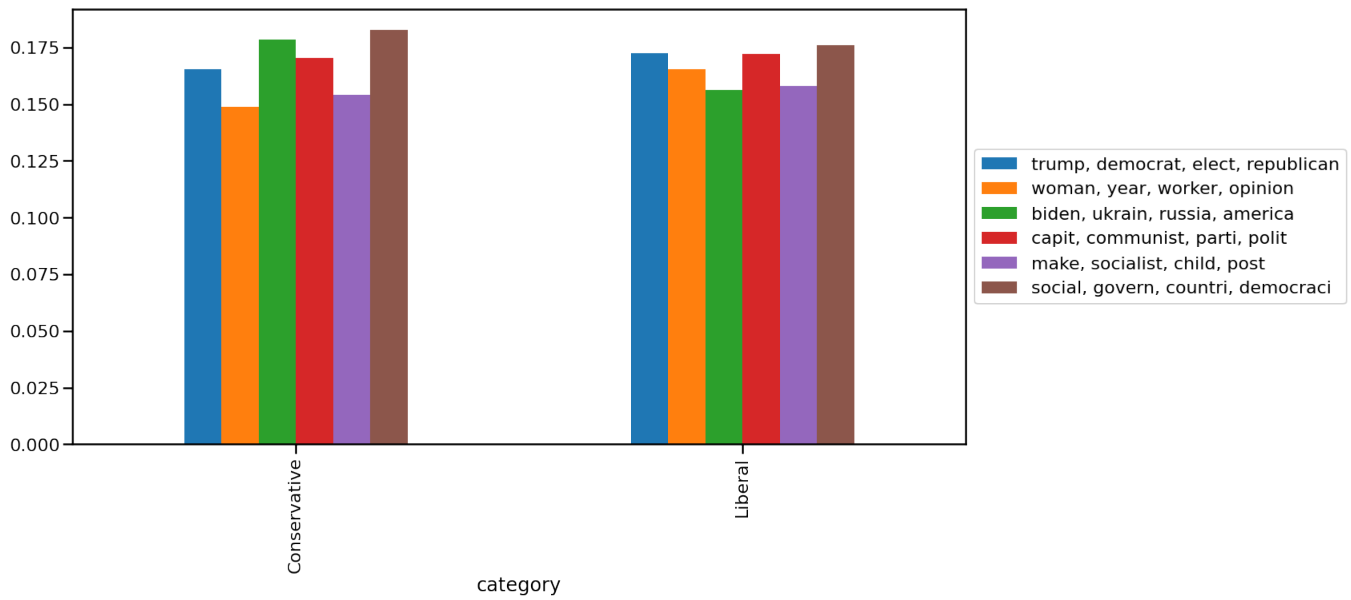
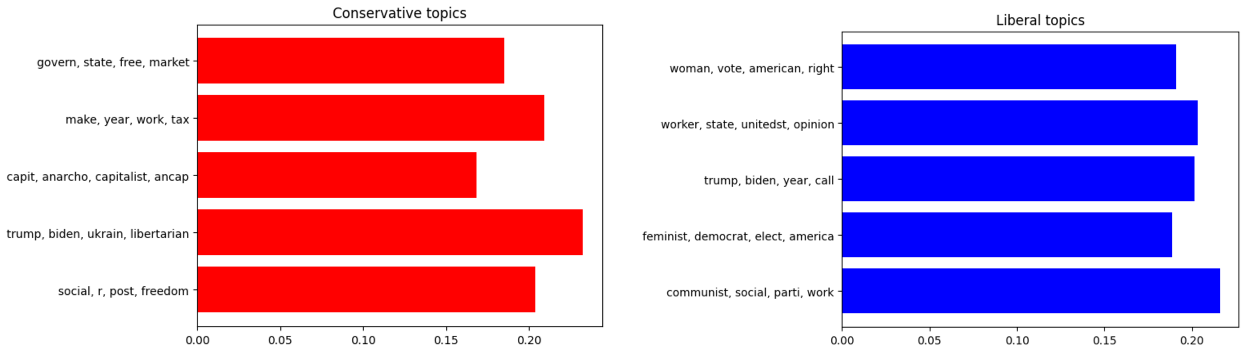


Figure 10: LDA topic analysis.



(a) Conservative.

(b) Liberal.

Figure 11: LDA topic modeling on the separate groups.

4.2 LDA

One of the most common technique for topic modeling is Latent Dirichlet Allocation (LDA). Since from the matrix-factorizaion topic analysis it can be observed that some words are redundant for single topics, 2 additional preprocessing steps are applied: stemming and joining of the most common collocations, such as ’Joe Biden’, ’Donald Trump’, or ’Supreme Court’.

First, LDA is applied to the whole dataset, and the topics obtained are reported in figure 10 grouped by the political lean. Then LDA is applied to the set of posts labelled ’Liberal’ and ’Conservative’ separately; the results are reported in figure 11.

Applied to the whole dataset, LDA does not show any significant differences among the 2 political leans; however, when applied to the 2 subsets separately, analogous differences found by the matrix factorization technique can also be found in this case. For example, the Conservative topic ’govern, state, free, market’ may refer to policies adopted by the government (or state) to regulate the market, in contrast to the more capitalistic view of a free market that self-regulates. On the other hand, the liberal topic ’work, state, unitedst, opinion’ may refer to a different view of the government, one that interprets it under the lens of the working class.

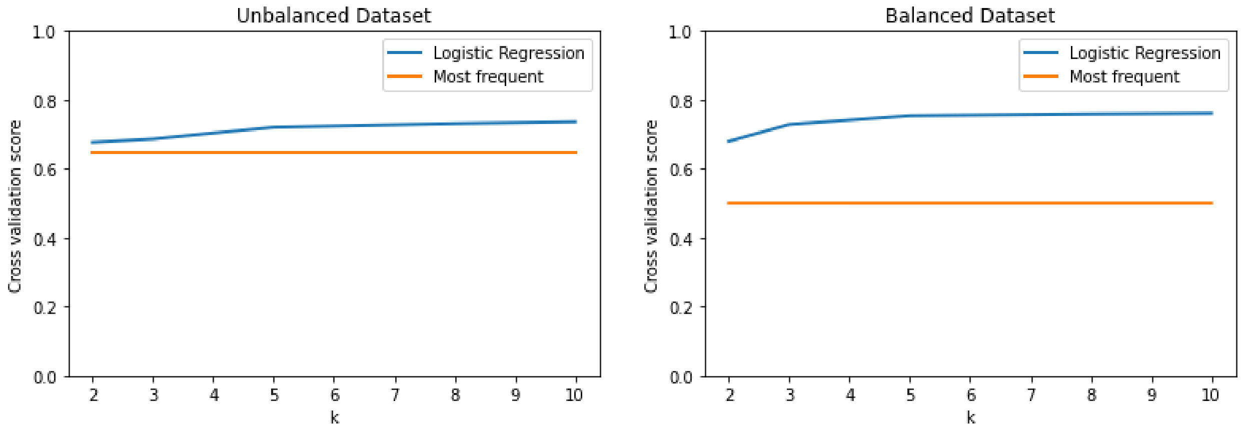


Figure 12: Cross validation applied to Logistic Regression and a Dummy Classifier for the balanced and unbalanced dataset.

5 Text Classification

In the following section, 2 text classification techniques are applied in order to understand if it is possible to create a model which is able to automatically assign a post as 'Liberal' or 'Conservative' based on the words used. The models considered are Logistic Regression and Neural Networks.

5.1 Logistic Regression

Logistic Regression is a popular statistical model used for binary classification. As reported in figure 1, the dataset is clearly unbalanced, hence SMOTE (Synthetic Minority Over-sampling Technique Chawla (2002)) is used to create a more balanced dataset; in figure 12 is shown the difference with respect to a classifier which always classify a new element as belonging to the most frequent class in the cross validation score when using the balanced and unbalanced dataset.

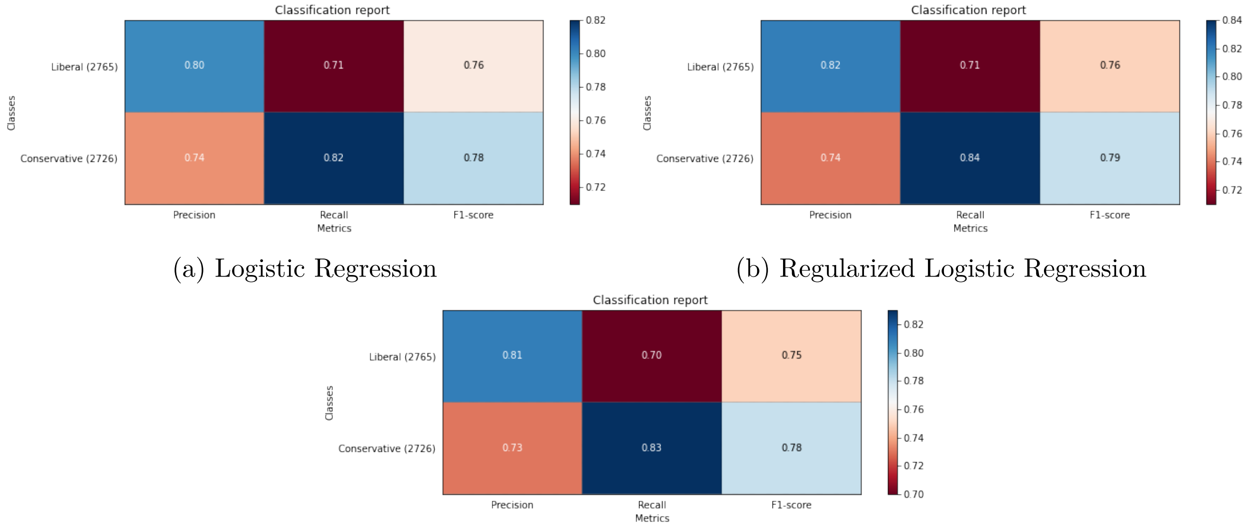
To evaluate the performance of the classifier, the precision, recall and F1-score are reported in figure 13 for standard Logistic Regression, Logistic Regression with regularization, and Logistic Regression applied to the dataset considering just the most relevant features according to the chi-squared feature selection test.

As depicted in the images, there is no significant improvement by using regularization or feature selection, and the performance metrics are consistent throughout the 3 techniques: the precision for liberals is always higher than that for conservatives, meaning that the number of false positives for liberals is less than the number of false positive for conservatives, while the recall is higher for the conservative label, meaning that the number of false negatives for conservatives is less than that of liberals. In other words, the model tends to classify incorrectly posts as 'Liberal' while they are actually 'Conservative'.

5.2 Neural Networks

Neural Networks have become increasingly more powerful in several Machine Learning tasks and Natural Language Processsing; in the following, a simple Feed-Forward Artificial Neural Network is employed for classifying posts as 'Liberal' or 'Conservative'. The ANN consists of:

- 1 input layer representing a Reddit post.
- 4 linear hidden layers.
- 1 output layer with 2 elements, each representing the probability of the post of belonging to the 'Liberal' or 'Conservative' label.



(a) Logistic Regression

(b) Regularized Logistic Regression

(c) Logistic Regression with feature selection

Figure 13: Classification report for different Logistic Regression based techniques.

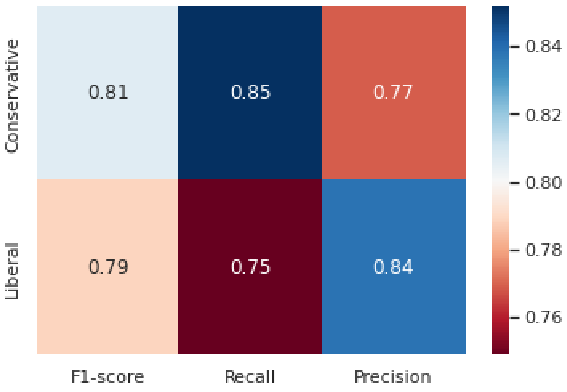
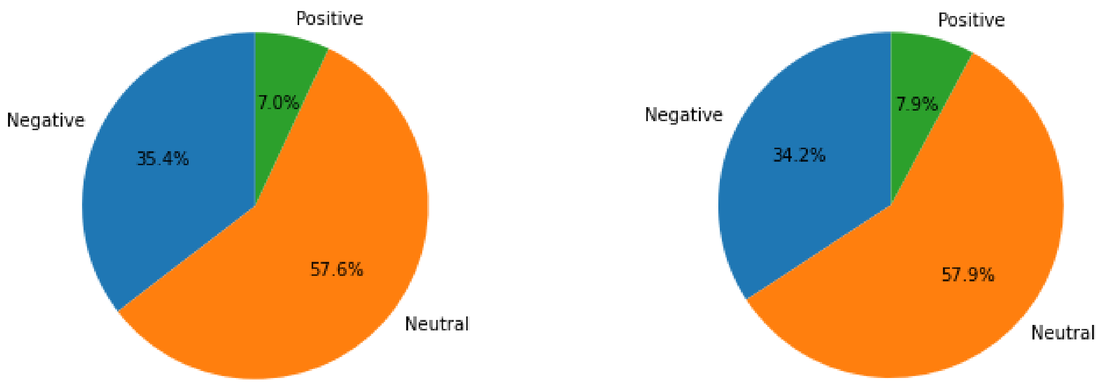


Figure 14: Feed-Forward Neural Network classification performance.

The post is then labelled according to the highest probability in output.

In figure 14 are reported the results of the Performance metrics. The results show the same pattern of performance in the precision and recall of Logistic Regression, but with a small improvement in the overall performance.

The F1-score in both cases is approximately 75-80%; this result suggest that the classifiers are able to discover and learn the major differences in the words used and topics discussed between the 2 political leans, but the performance does not reach higher values because, as we found in previous sections, there exist really similar posts among liberals and conservatives, and these posts are the one that the classifier is not able to label correctly, penalizing the performance.



(a) Liberal

(b) Conservative

Figure 15: Sentiment Analysis with pre-trained Deep Learning model.

6 Sentiment Analysis

Lastly, Sentiment Analysis is performed on the 2 political leans separately to inspect whether there is a tendency of one group towards positive or negative emotions with respect to the other.

The analysis is performed with a pre-trained model (cardiffnlp/twitter-roberta-base-sentiment-latest) available on the HuggingFace website (HuggingFace), a community and data science platform that provides access to pre-trained Deep Learning models and datasets. The model is a roBERTa-base model trained on ≈124M tweets and finetuned for sentiment analysis; roBERTa Liu (2019) is a robustly optimized method for pretraining natural language processing systems that improves on Bidirectional Encoder Representations from Transformers, or BERT, the self-supervised method released by Google in 2018.

The underlying assumption by applying this model is that Reddit posts are considered really similar to Twitter posts, being both a social network with a limited amount of characters for post, and the possibility of typing emojis and posting images or links to other websites and articles.

The model receives as input the text, and outputs a probability distribution over 3 categories: 'Negative', 'Neutral' and 'Positive'; the class is then assigned as the max among these 3 values. The results are reported in figure 15.

No significant difference is found among the 2 labels, with the majority of the posts labelled as neutral or negative, and very little percentage of positive posts. The higher percentage of negative over positive posts is as expected, because people tend to share content that moves them emotionally, especially negative emotions such as anger and despise towards opposite political views of certain topics.

A surprising result is the high percentage of neutral posts; this may be due to the fact the majority of Reddit posts have a short title which just introduces the actual and meaningful content of the post, which may be an attached image or a link to an article. Another plausible reason may be the fact that in Reddit, a common format for posting is that an user post a question or share something with the aim of receiving explanation agreed among a large group of people about specific topics, while the actual discussion can be protracted in the comments of that post. Hence, being the title just an introduction or a question, it does not connote a strong sentiment, resulting in a high percentage of the neutral label.

7 Conclusions

This project aimed at analyzing differences and similarities between Reddit posts labelled as ’Conservative’ or ’Liberal’ based on the Subreddit they were extracted from.

The results obtained suggest that the two groups share a common pool of words used in the posts and the relative topics of discussion. They also suggest that, apart from the similarities, significant differences do also exist. In particular, liberals seem to talk about subjects which are historically related to the Political Left, such as the rights of workers, rights of women and rights of minority groups; they also seem to talk more about their own form of government, such as social democracy and communism. On the other hand, conservatives seem to talk about subjects historically associated to the political Right, such as free market, capitalism and freedom. This result may be of support of the idea that social media tend to increase polarization among the people, because liberals engage in "liberal" topics ignoring the ones more correlated to conservatives, strengthening their own biases and world views, and vice-versa for conservatives.

Furthermore, the results obtained in section 5 support the fact that computational models can be build in order to classify a post as ’Liberal’ or ’Conservative’ just based on the words used. Further improvements can be obtained by employing more complex models (CNN, RNN, attention-based network) and a larger dataset.

Lastly, no significant differences are found in the Sentiment Analysis of the two groups, and the result according to which the majority of posts is labelled as neutral indicates that Reddit posts may not be ideal for capturing the sentiment of people because of the format and the way the platform is used.

References

N. V. Chawla. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research, 321-357*, 2002.

Y. Liu. Roberta: A robustly optimized bert pretraining approach. *https://arxiv.org/abs/1907.11692*, 2019.