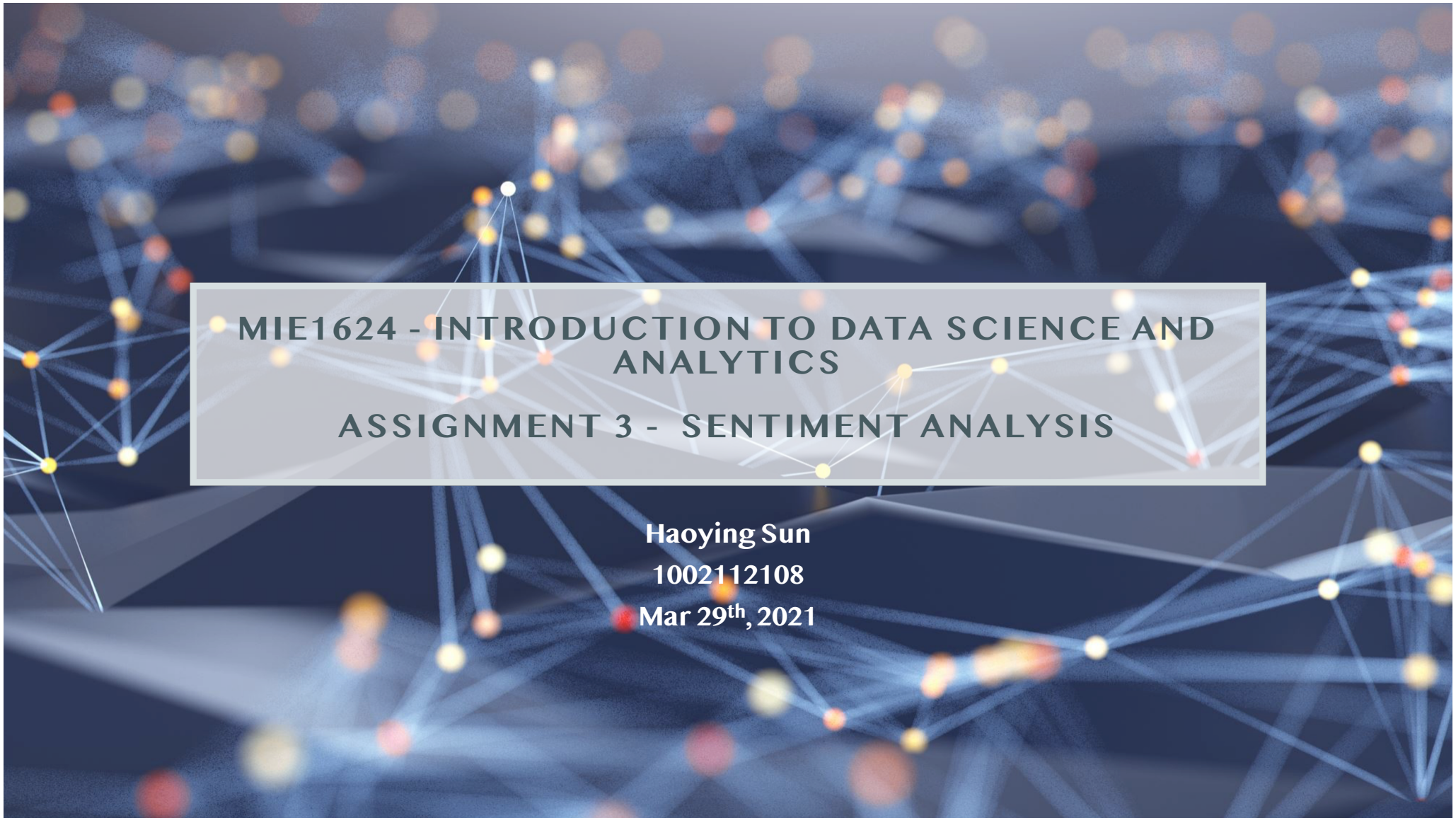
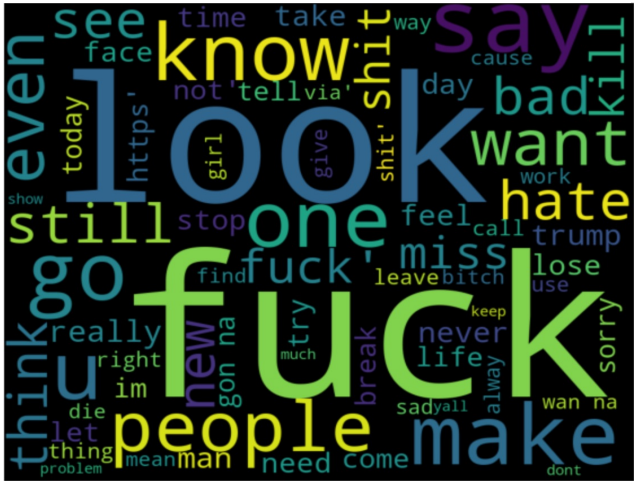


MIE1624 - INTRODUCTION TO DATA SCIENCE AND ANALYTICS

ASSIGNMENT 3 - SENTIMENT ANALYSIS

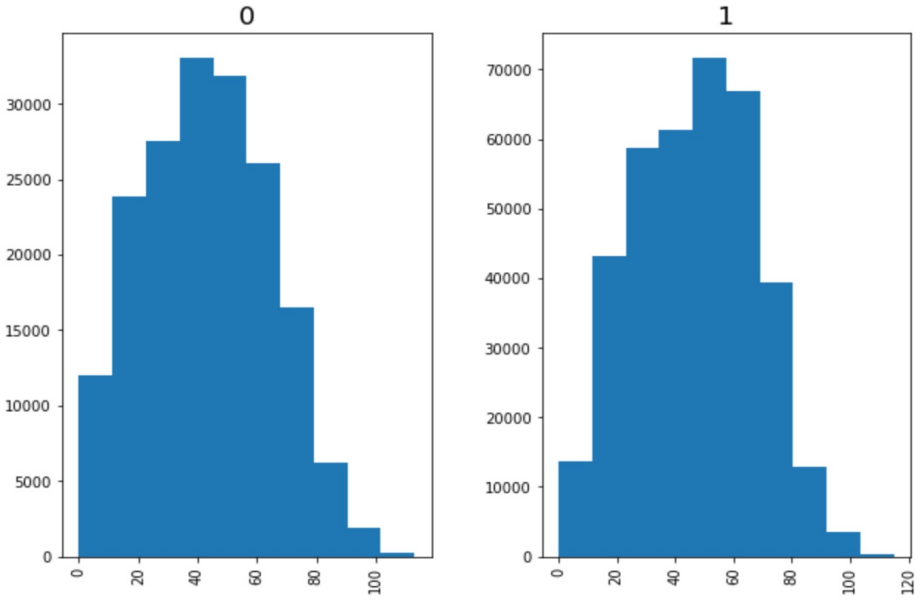
Haoying Sun
1002112108
Mar 29th, 2021

Exploratory Analysis of Generic Dataset



A wordcloud for the positive tweets in the generic dataset is shown on the top left in white. Some words with noticeable positive emotions are: happy, love, great, amaze, awesome, good, etc.

A wordcloud for the negative tweets in the generic data is shown on the bottom left in black. It can be observed that a lot of words express negative emotions, such as kill, die, problem, mean, bad, etc. Some words are even swearing words.



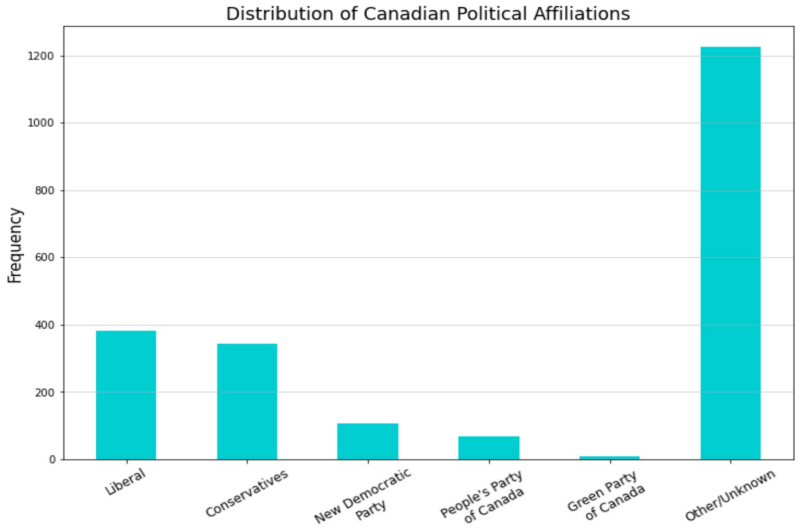
The length of the positive and negative tweets are examined and plotted as histograms shown above. It can be noticed that the distribution of the length of both positive and negative tweets are slightly skewed towards the left. In general, positive tweets are longer than negative tweets by about 20 words.

Exploratory Analysis of Canadian Election Dataset

The top 100 words for the positive tweets are shown on the top left with white background. It can be noticed that a lot of these words display positive emotions, such as "better", "love", "new", "great", etc.

The top 100 words for the negative tweets are shown on the bottom left with black background. It is noticed that the negative wordcloud has some overlap with the positive wordcloud. These words are mostly related to the political party and their candidates. One word with noticeable negative emotion is "lie".

Interestingly, the word "Monday" appears in the positive wordcloud. This is possibly because the election date is 10/21/2019, which is a Monday.



As shown above, the majority of the tweets have "Other/Unknown" as their political affiliation, possibly because the self-defined search words do not contain all the relevant hashtags used in the tweets. Another reason may be that some tweets do not contain any hashtags. Looking at the tweets that have an identified political affiliation, it is observed that Liberal Party has the highest frequency, followed by Conservative Party, New Democratic Party, People's Party and Green Party.

Many factors need to be considered during this analysis, such as: 1) Many tweets do not specifically express specific support for a certain political party but only opposition towards some parties. 2) The demographics of the twitter users will affect the frequency of each political affiliation. For example, supporters for Liberal Party mostly reside in urban areas and have more younger generation. These users will have a higher likelihood of using twitter and have a better understanding of technology (e.g. how to use hashtags and how to use the right hashtags).

Implementation of Base Models on Generic Data

Model	Attributes of the Base Model	Performance (f1 score)
Logistic Regression	solver='saga', random_state=0, max_iter = 1000, dual=False	WF: 97.09%, TF-IDF: 97.06%
Decision Trees	criterion='gini', random_state = 1	WF: 95.16%, TF-IDF: 95.37%
Random Forest	n_estimators = 100, criterion = 'gini', random_state = 42	WF: 96.39%, TF-IDF: 96.54%
KNN	n_neighbors = 620 (sqrt of number of samples)	WF: 85.81%, TF-IDF: 36.45%
Linear SVC	random_state = 0, max_iter = 10000, loss='hinge'	WF: 97.06%, TF-IDF: 97.08%
Complement Naïve Bayes	Default setting	WF: 95.49%, TF-IDF: 94.55%
XGBoost	max_depth=3, random_state = 0	WF: 92.21%, TF-IDF: 92.22%

Chosen model: logistic regression with word frequency => Resulting testing accuracy on election data sentiment: 53.4%

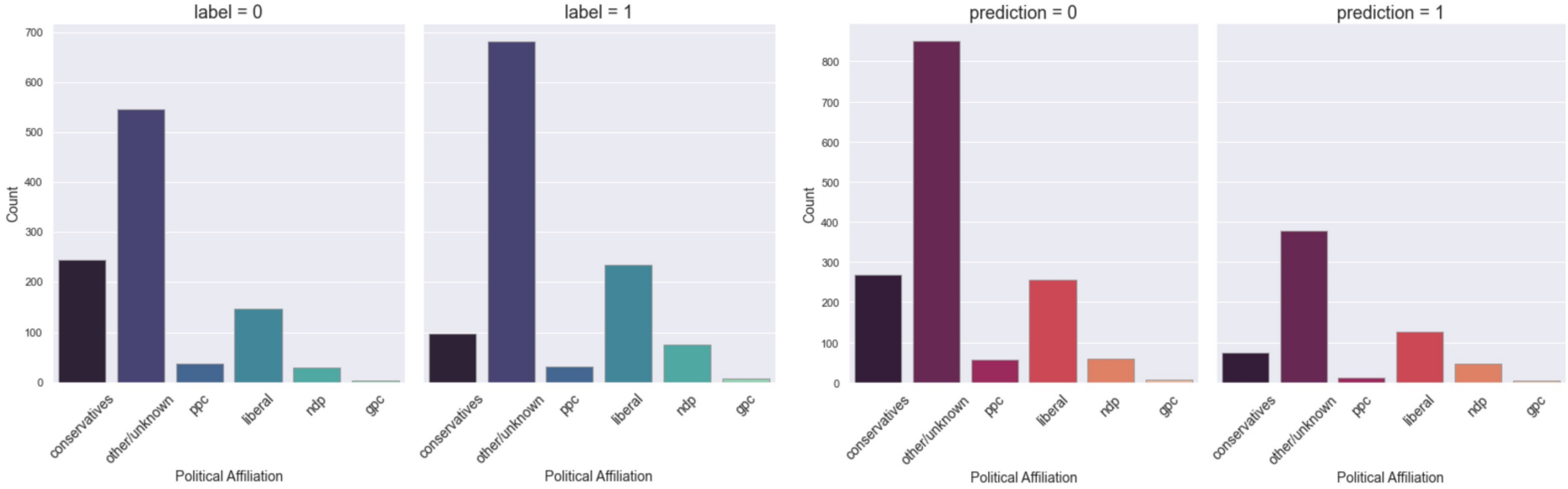
Implementation of Best 3 Models on Negative Election Data

Model	Attributes of the Base Model	Hyperparameters	Performance (Accuracy score)
Logistic Regression	solver='saga', penalty='elasticnet', random_state=0, max_iter=3000, multi_class='multinomial', class_weight='balanced', dual=False	C: 0.01, 0.05, 0.1, 0.5, 1, 5, 10 l1_ratio: 0.2, 0.4, 0.8, 1	WF: 86.65%, with C=1, l1_ratio = 0.4 WE: 87.93%, with C=10, l1_ratio = 0.2
Linear SVC	class_weight='balanced', loss='hinge', random_state = 0, max_iter = 10000, multi_class='ovr'	C: 0.001, 0.01, 0.05, 0.1, 0.5, 1,5, 10, 100	WF: 99.57%, with C=0.5 WE: 88.64%, with C=10
Random Forest	criterion = 'gini', class_weight='balanced', random_state = 42	n_estimators: 200, 400, 600, 800, 1000, max_depth: 20, 40, 60, 80, 100, None	WF: 93.61%, with n_estimators=600, max_depth=20 WE: 100%, with n_estimators=600, max_depth=20

Chosen model: Random Forest with word embedding => Resulting testing accuracy on negative election data negative reason: 51.32%

*WE: Word embedding

Visualization of Logistic Regression Model Performance on Election Data Sentiment

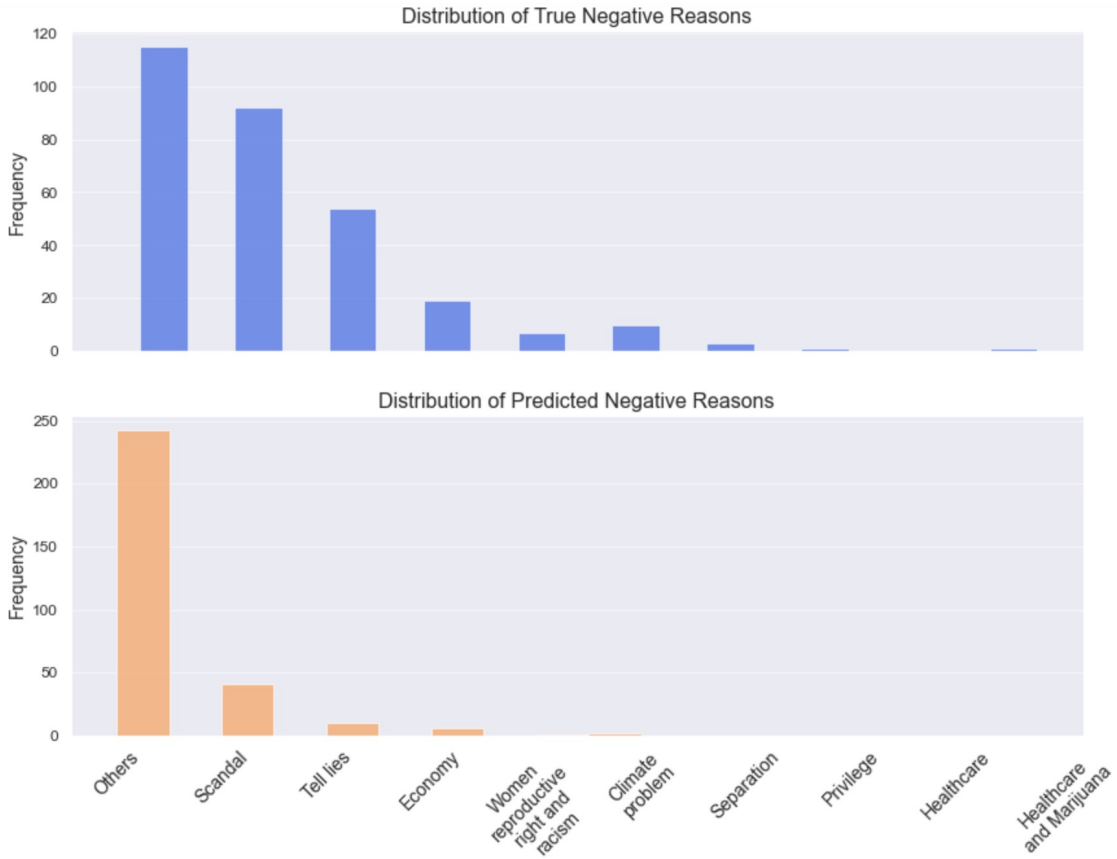


The validation accuracy is significantly lower than the training accuracy (train: 97.09%, test:53.4%), indicating that the model is overfitting. The plots above show that the model noticeably assigns more "Other/Unknown" tweets with negative sentiment, whereas in the true data, more "Other/Unknown" tweets have positive sentiment. Same results are observed with "ppc" tweets, "liberal" tweets, "ndp" tweets and "gpc" tweets - the model tends to assign negative sentiment to these tweets whereas in the true data, more of these tweets have positive sentiment.

Based on the model's perspective on the positive tweets, Liberal Party has the highest popularity in the public's eye among all 5 identified political parties. Following Liberal Party, Conservative Party and New Democratic Party have the second and third highest popularity, respectively. People's Party and Green Party have the second least and the least popularity.

To improve accuracy of the model the following recommendations could be considered:
1) Use an exhaustive/more complete list of stopwords to reduce potential noises
2) If possible, increase the amount of training data
3) Use more sophisticated feature engineering techniques such as N-grams

Visualization of Random Forest Model Performance on Election Data Negative Reason



The resulting validation accuracy (51.32%) is significantly lower than the training accuracy (100%), confirming that the model is overfitting. As shown in the distribution plot on the left, the model assigns most of the tweets the reason "Other" and "Scandal". This is because the original dataset is highly imbalanced, with "Others", "Scandal" and "Tell lies" taking up 80% of the entire dataset. Even "class_weight" of the model is set to "balanced", it did not do much to help adjust the prominent data imbalance.

Factors and recommendations to be considered include:

1) The number of maximum features of the model may not be optimal. Normally, decreasing the number of maximum features could help reduce likelihood of overfitting. Different settings for "max_features" is indeed worth exploring, and is likely to provide insights into improving the model accuracy.

2) The number of estimators of the model may be too small. The number of estimators determines the number of trees inside the model. In general, more trees will lead to better accuracy, however, will also lead to longer runtime.

3) The maximum depth of each tree may be too big. Decreasing the maximum depth helps with reducing complexity and decreasing bias. Starting with a smaller maximum depth (e.g. around 5) and then slowly increasing it may be helpful for improving the accuracy.

4) More sophisticated methods to deal with imbalanced data may be worth trying, such as boosted decision trees.

5) If possible, gather more data to reduce the imbalance between classes and provide more features for training the model.


THANK YOU!