SDS M2 Group Assignment

Andreas Jørgensen, Cathrine Olsen, Louise Christoffersen
Mette Møller

5. November 2020

1 Introduction

The movement Black Lives Matter had a resurgence in May this year after the death of George Floyd. The hashtag #blacklivesmatter went viral in order to support the movement and criticise the law enforcement's misuse of power against African-American people. A counter movement called Blue Lives Matter was started to support the police.

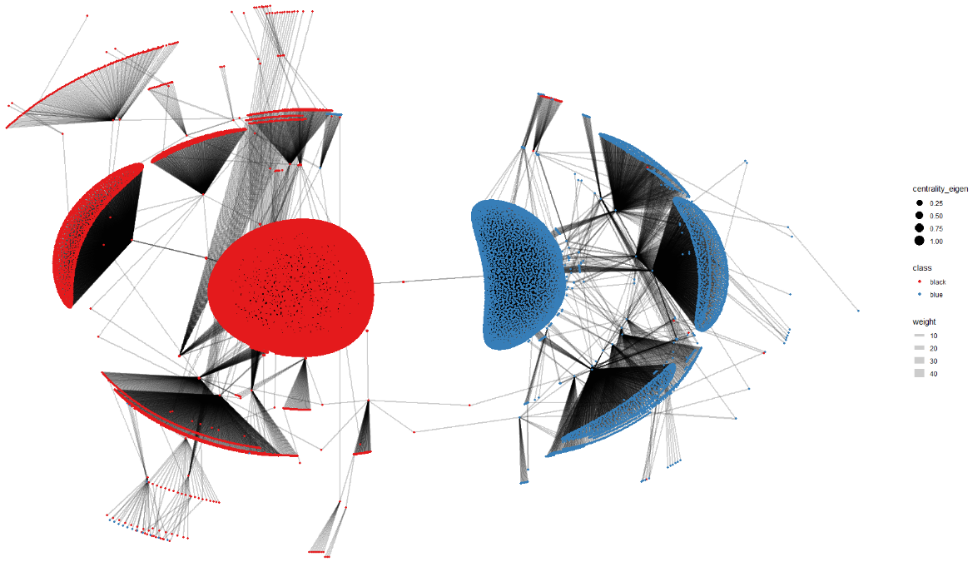
The purpose of this report is to be able to train a machine learning model to classify tweets as either being pro BlackLivesMatter or pro BlueLivesMatter. Such a model can be used to gauge public opinion regarding the ongoing debate about police brutality.

Firstly, a network analysis is performed showing the structure of retweets followed by analysis of individual words to investigate which words are most commonly used by the two classes. General topics are investigated using Latent Dirichlet Allocation (LDA). Lastly, a model is created to predict the classes of the tweets using Supervised Machine Learning.

The data used for this report is downloaded from Twitter 03-11-2020. Tweets containing #BlackLivesMatter are downloaded and afterwards tweets containing #BlueLivesMatter with a total amount of tweets of 23,559. Each tweet are then classified according to the hashtag.

2 Network analysis

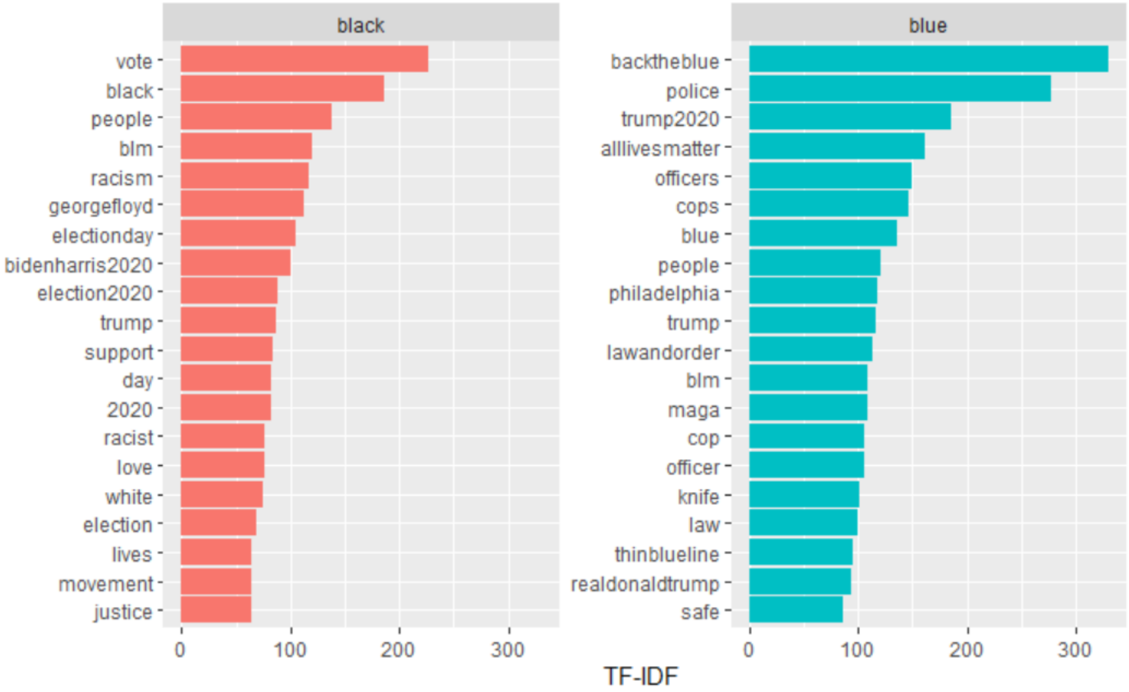
The network presentation of the tweets contain the individual people in the data and the connections between them in the form of retweets. The most central person of the network is the basis of the plot below where only people with a connection to that individual appear.



The plot shows two large separate networks here that are only connected by a few individuals that have retweeted things from both sides of the debate. Both networks follow the same structure with a few tweets being retweeted a lot. The broad method of classification used when collecting the data also has its drawbacks which can be seen in the networks. Some of the offshoots in the networks are the opposite class than the rest of the network they are in and this might be an indicator that those tweets have been classified wrongly. This could have been overcome by thoroughly going through the individual tweets, but that was unfortunately not possible in the time frame given. The plot furthermore shows that there are very few mutual connections between individuals, and that people most likely retweet tweets posted by people in the same class.

3 Natural Language Processing

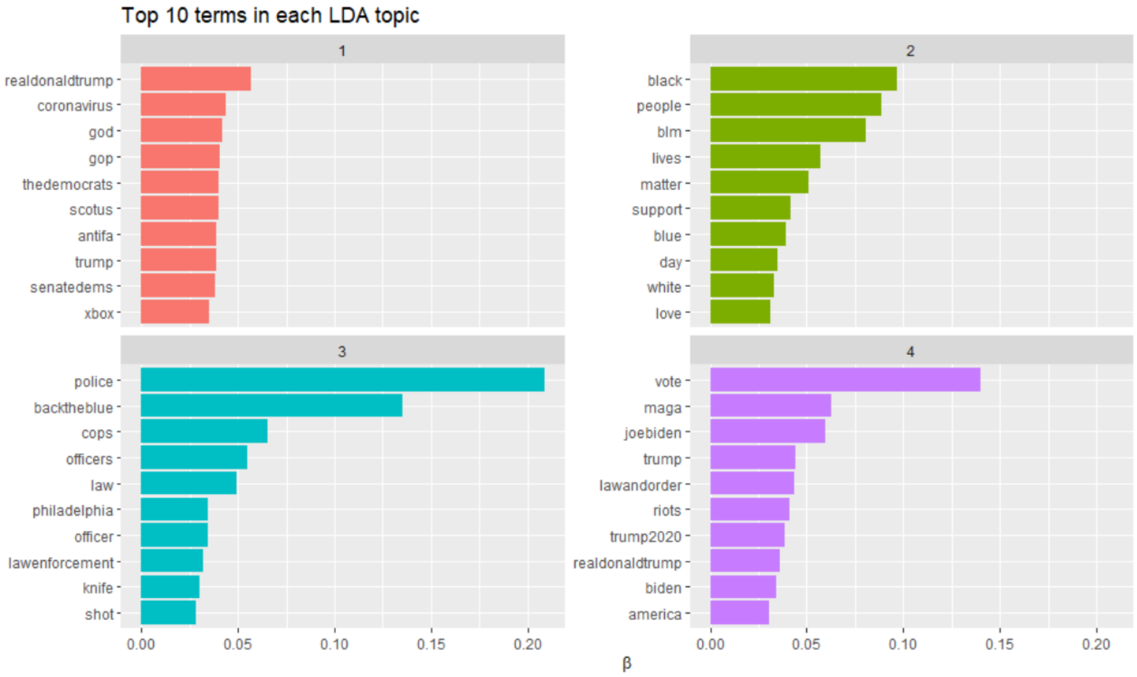
In this section all tweets have been separated into words and words like ”or”, ”is”, and ”the” that carry only little meaning are removed from the data. The aim of this is to investigate the most important words. The top 20 words in each class is presented below.



-The top words in class *blue* is mainly focused on law enforcement while *black* is more focused on race. Both classes has an element of politics related to the election where mentions of Joe Biden appear in *black* while Donald Trump appears in multiple versions in class *blue*.

3.1 Topic Modelling

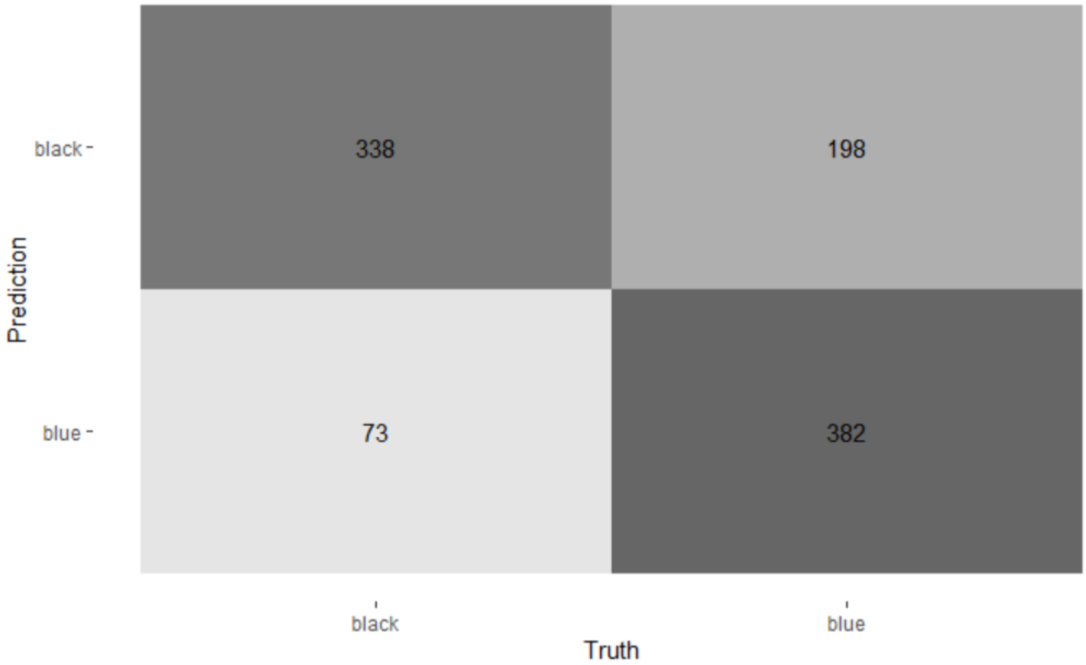
A Latent Dirichlet Allocation is used to divide the words into 4 different topics. The graph below shows the probability that certain words appear in each of the topics.



Topic 1 seems to consist of various words related to politics - yet some seem somewhat unrelated. *Topic 2* contain words related to race and beliefs, *topic 3* is related to law enforcement, and *topic 4* is mostly related to the election.

3.2 Supervised Machine Learning

Several machine learning models have been trained on 75 percent of the data to predict the classes of tweets and the best has been selected. The model is then tested on the remaining 25 percent of the data to evaluate its performance. The result is presented as a confusion matrix which shows the true classes compared to the classes predicted by the model.



The performance is also measured in overall accuracy and accuracy in classifying the two classes.

Class	**Estimate**
Black	82 %
Blue	66 %
Overall	73 %

The model performs well on the *black* class, meaning that it is able to correctly predict whether a tweet supports the *blacklivesmatter* cause. On the other hand the model performs poorly when dealing with tweets that support the *bluelivesmatter* cause. After an investigation into some of the misclassified tweets, it seems that the model struggles when tweets that are categorized as *blue* mentions politics and the election. This means that the current American election is likely causing some interference in the model. This might have been avoided if the data was collected from before the election began in earnest.