Kaggle Trump tweets

Team D17: Anna-Mai Allikmäe, Iris Luik, Susanna Mett, Kevin Telliskivi

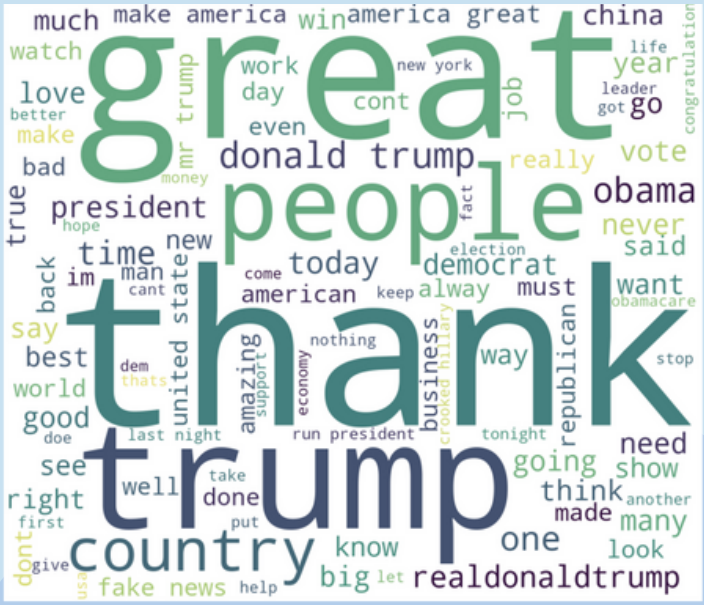
INTRODUCTION

Social media is an influencial platform and the twitter account of former president Donald Trump is an exceptional example of what power a virtual public platform can have in world politics. In this project, our first goal was to analyse and visualise the content of Trump's tweets and the second goal was to build a model which could predict which tweets gain more popularity than others.

DATA

The dataset from Kaggle contains 41122 tweets of Donald Trump including the content and metadata of all his tweets from 2009 to 2020. As the basis of popularity, we used the count of a tweet marked as a favourite.

100 most used words in Trump's tweets



METHODS

We removed tweets that contain only mentions and hyperlinks, lemmatised the content and divided the data into pre-presidential and presidential time categories. The number of favourites was divided into four categories. We analysed the tweets including exploring most used words overall, comparing pre-presidential and presidential time popular words and generating word clouds.

For building a model, we used topic modelling and sentiment analysis to be used as inputs in the model predicting the popularity of tweets. We experimented with models containing 20, 50, 75, and 100 topics generated with LDA model. Sentiment analysis using Flair divided all tweets into either positive or negative based on their content. The final result was a random forest classifier model.

Sentiment of tweets in pre-presidential time and presidential time

Pre-presidential time

Presidential time

RESULTS

The most frequent words used in Trump's tweets are "great" and "trump" with over 6000 instances, "president", "thank", and "people". The use of word "trump" decreased significantly after becoming a president, the usage of "great" increased and the usage of "thank" decreased. Trump's tweets were more positive before he became the president. The random forest model remained imprecise but the model containing 50 categories and sentiment values proved to be the most accurate. The model lacks precision either because topics and sentiment are not the best features to predict the number of favourites or because the generated input models or used random tree model are not accurate enough.

0 – unpopular (< 70 000 favorites)
1 – slightly popular (70 000 - 135 000)
2 – popular (135 000 - 200 000)
3 – very popular (> 200 000)

Confusion matrix of a model predicting popularity based on 50 topics and sentiment values