KAGGLE-TRUMP TWEETS

Do you tweet like Donald Trump?

Report

Anna-Mai Allikmäe, Iris Luik, Susanna Mett

GitHub repository: https://github.com/maiallikmae/IDS_project_trump_tweets

Business understanding

In the modern society, social media has a major influence in the workings of social and cultural processes which is why it is essential to thoroughly investigate its larger trends and notable cases. In this project, we concentrate on Donald Trump, a former president of the United States of America, whose social media usage — namely, his usage of Twitter — is an exceptional example of what power can a virtual public platform have in world politics. In the case study, however, we do not qualitatively look into influential tweets in relation to world events (e.g. the possibility of starting a war with North Korea in the physical world), rather we quantitatively explore the wider trends in Donald Trump tweets online. Examining Trump's tweets and their popularity can provide a valuable insight into popular and effective discourses on social media.

Although we recognise the specificity of the tweets of Trump, we still believe that our case study does not interest only social scientists and historians who are specialised in the research of historic discourses etc. Understanding the phenomenon of the tweets of Trump is useful for wider audiences like governments and their PR-teams for investigating potential communication tactics, discussing how to respond to unexpected negotiation situations, and what is effective among social media users. Likewise, social media users and businesses alike can see what kinds of topics and ways of communicating are the most successful in terms of gaining popularity on social media.

In order for the project to be successful, the case study must identify what causes the popularity of Trump's tweets both thematically (i.e. what topics are the most popular) and discursively (i.e. what ways of communicating are the most popular). For thematic insight we examine the popularity of different topics mentioned in tweets and the investigation of

discursive means includes sentiment analysis and exploration of the usage of capitalised words.

The data for the project is fully available – the tweets of Donald Trump are gathered in a csv-file which is downloaded from Kaggle. The project will be conducted using Jupyter Notebook interface and various Python tools, such as natural language processing (NLP) for text classification and the basis of machine learning. By natural language processing we mean software that is able to identify, classify, and manipulate written text in English language. Text classification is a process in which text is divided into categories, in particular, we assign labels to the tweets of Trump. The categories will be created by algorithms and labels assigned manually. By sentiment analysis we mean assigning negative, neutral or positive sentiment value to each tweet using previously developed sentiment analysing tools in Python NLP. Data mining in this project includes combining different parameters (such as number of retweets and sentiments or different categorisations) to find the most precise prediction of the popularity of the tweets.

The project will be finished by the deadline set by the course and completed by team members. The potential risks of successfully finishing the project include getting stuck with some parts of coding, especially in the machine learning section, and too boldly interpreting the results. However, there are enough resources available online to properly conduct the research. The costs and benefits are not relevant in terms of the project since the outcome is not quantifiable and has an effect in more subtle ways such as a change in discourses that show their consequences later.

Data understanding

Data that we need for our project is the tweets of Donald Trump and its likes and retweets accompanied by dates. We found the data that meets the requirements from Kaggle. Kaggle had two datasets which were about the same size and from which we will use the second dataset because it has been cleaned from non-textual tweets like pictures since we are not going to simultaneously analyse pictures in this particular project. The dataset we are using was uploaded in 2020 and it consists of 41 122 rows and 9 columns. Columns are:

id - the tweet's id (integer)

link - the URL of the tweet (string)
content - the tweet itself, the text of the tweet (string)
date - date and time when the tweet was posted (string)
retweets - number of retweets the tweet got (int), ranges between 0 and 309892
favorites - number of favorites the tweet got (int), ranges between 0 and 857678
mentions - accounts mentioned in the tweet (string), usually NaN
hashtags - hashtags used in the tweet (string), usually NaN
geo - geotag of the tweet (float), usually NaN

From these columns we will only use content, retweets and date, so the other metadata will be removed during the data cleaning. The tweets are from 2009 to 2020 and we have 41 122 tweets from this period of time. There are no signs of data quality problems since the most valuable data that we will work with is tweet content and its popularity with date which have no NaN values. The data is good enough to support our main goal — studying the case of former president Donald Trump's communication manners.

During data exploration we gained valuable information and planned how we will clean and prepare our data for model building. Firstly, we want to use simple text cleaning processes for tweets which includes removing punctuation, special characters, web pages URLs, extra tabs and also correcting or removing typos. We proceed by filtering out stop words (most common words in any language), after which we will be using stemming and lemmatization.

For model building we want to split our data — so we would have pre-presidential and presidential (20th of January 2017 till 20th of January 2021) time because the content of the tweets and amount of attention they got are noticably different and using those time periods together would probably result in erraneous model. Our plan is to build a model using presidential time tweets and later comparing the results to pre-presidential time tweets.

We also want to categorize our retweets and favourites as most popular, popular, slightly popular and unpopular. We found during data exploration that we have an imbalanced dataset for this categorization so we need to balance it before modelling. Besides building models we also want to visualize and analyse Trump's tweets. For initial exploring we only need to remove unnecessary columns. Afterwards we think that it would be interesting to compare

pre-presidential and presidential time most popular tweets and topics. Also explore the tweets during the election campaigns.

Planning your project

Tasks

1. Exploring data - in this task we are getting familiar with our data. We will examine our data and see how much and what kind of cleaning needs to be done. Work hours: 3 hours of work for each of us.
2. Cleaning data - we will remove 3 columns out of 9 and will work with content, date, retweets and favorites. We will remove all punctuations, special characters, web pages URLs, extra tabs and correct or remove typos. Work hours: 6 h (Iris)
3. Constructing data - we will find the most popular words used in Donald Trump's tweets and illustrate those, this will give us a better understanding of data. On the basis of retweets and favorites we will create 4 classes for tweets: most popular, popular, slightly popular and unpopular and then we will encode favorites and retweets into values 0-3. Furthermore, we will assign sentiment value for each tweet. We will use natural language processing (NLP). Work hours: illustrating 15 h (Susanna), further cleaning and categorization 6 h (Anna-Mai)
4. Building models - we will set the parameters and try different models on our data. This is the most important and time-consuming task. Work hours: 35 h (Susanna, Anna-Mai)
5. Evaluating results - in this task we will evaluate our results and compare it to the goals we set before data mining. Work hours: 10 h (Iris)
6. Presenting the final results - we are going to make a poster and prepare our presentation for the poster session on 16th of December. Work hours: 2 of us are going to make a poster (3 hours of work for Susanna, Mai) and one of us is going to present our work (3 hours of work for preparing the presentation and presenting it, Iris).