USA Nov.2020 Election 20 Mil. Tweets (with Sentiment and Party Name Labels) Dataset

This dataset includes 24,201,654 tweets related to the US Presidential Election on November 3, 2020, collected between July 1, 2020, and November 11, 2020. The related party name and sentiment scores of tweets, also the words that affect the score were added to the data set.

Election-related Tweets were searched by using the names of the parties, their abbreviations, the names of the party candidates, and the election slogans. The tweets that contain “#USAelection” or “#NovemberElection” hashtag or at least one of the following keywords about four parties that received the most votes in the last election, were collected.

Keywords about Democratic Party: @DNC OR @TheDemocrats OR Biden OR @JoeBiden OR "Our best days still lie ahead" OR "No Malarkey!"

Keywords about Republican Party: #MAGA2020 OR @GOP OR Trump OR @POTUS OR @realDonaldTrump OR Pence OR @Mike_Pence OR @VP OR "Keep America Great"

Keywords about Green Party: @GreenPartyUS OR @TheGreenParty OR “Howie Hawkins” OR @HowieHawkins OR “Angela Walker” OR @AngelaNWalker

Keywords about Libertarian Party: @LPNational OR “Jo Jorgensen” OR @Jorgensen4POTUS OR “Spike Cohen” OR @RealSpikeCohen

The dataset contains more than 20 million tweets with 11 different attributes of each of them. The data file is in comma-separated values (CSV) format and its size is 3,48 GB. It is zipped by WinRAR to upload and download easily. It is zipped file size is 766 MB. It contains the following information (11 Column) for each tweet in the data file:

Created-At: Exact creation time of the tweet [Jul 1, 2020 7:44:48 PM– Nov 12, 2020 5:47:59 PM]

From-User-Id: Unique ID of the user that sent the tweet

To-User-Id: Unique ID of the user that tweet sent to

Language: Language of tweets that are coded in ISO 639-1. [%90 of tweets en: English; %3,8 und: Unidentified; %2,5 es: Spanish].

Retweet-Count: number of retweets

PartyName: The Label showing which party the tweeting is about. [Democrats] or [Republicans] if the tweet contains any keyword (that are given above) related to the Democratic or Republican party. If it contains keywords about two parties then the label is [Both]. If it doesn’t contain any keyword about two major parties (Democratic or Republican) that the label is [Neither].

Id: Unique ID of the tweet

Score: The sentiment score of the tweets. A positive (negative) score means positive (negative) emotion.

Scoring String: Nominal attribute with all words taking part in the scoring

Negativity: The sum of negative components

Positivity: The sum of positive components

The VADER algorithm is used for sentiment analysis of tweets. The VADER (Valence Aware Dictionary and sEntiment Reasoner) lexicon and rule-based sentiment algorithm to score a text. it is specifically attuned to sentiments expressed in social media and produces scores based on a dictionary of words. This operator calculates and then exposes the sum of all sentiment word scores in the text. For more details about this algorithm: https://github.com/cjhutto/vaderSentiment

This data can be used for developing election result prediction methods by social media. Also, It can be used in text mining studies such as understanding the change of feelings in tweets about parties; determining the topics that cause positive or negative feelings about the candidates; to understand the main issues that Twitter users concern about the USA election.