Tweets_Analysis_DD

The project's objective

The aim of this project is to help the social media manager of **Dunkin Donuts** to predict online engagement for his next.

The Project's steps

Retrieving Data

The first thing we started with, is to get tweets posted by **Dunkin Donuts** official Twitter account while excluding **Retweets & Replies**.
Using the Pagination, we ended up having a total of `793 tweets` with the following columns: `author_id`, `conversation_id`, `source`, `possibly_sensitive`, `id`, `text`, `reply_settings`, `lang`, `created_at`, `referenced_tweets`, `in_reply_to_user_id`, `public_metrics.retweet_count`, `public_metrics.reply_count`, `public_metrics.like_count`, `public_metrics.quote_count`, `entities.annotations`, `entities.urls`, `entities.mentions`, `entities.hashtags`, `attachments.media_keys`, `attachments.poll_ids`.

Feature Engineering

After binding all the tweets together in the same dataframe, we thought about creating new variables based on the one we got from Twitter. These tweets will bring more insights to our analysis and will help in creating our prediction model.

The main features we came up with in this step are:

-`Engagement variable` based on weighted average of likes, replies, quotes and retweets.
-`Engagement categories` based on quantiles [1,4].
-`Tweet Length` of original tweet with all hashtags and urls.
-`Clean Tweet Length` after removing all hashtags and urls.
-`Is_weekend` precising whether the tweet was posted during a weekday or not.
-`Upper_count`giving the number of uppercase words per tweet.
-`Exclamation_count`showing the frequency of exclamation marks inside the tweets.
-`Hashtag_ct` counting the number of hashtags per tweet.
-`Photos_count` precising the number of photos existing as an attachment by tweet.
-`Videos_count`precising the number of videos existing as an attachment by tweet.
-`Gifs_count`precising the number of gifs existing as an attachment by tweet.
-`Other_media_count`precising the number of media types other than photos, videos and gifs, existing as an attachment by tweet.
-`Emojis_count` giving the number of emojis per tweet.

Sentiment Analysis

In the meantime, we worked on assigning sentiment for our corpus.
For this step, we tried different **dictionary-based approach** libraries, citing mainly:
- `Sentimentr`: which is designed to quickly calculate text polarity sentiment in the English language at the sentence level, by attempting to take into account valence shifters (i.e., negators, amplifiers (intensifiers),

de-amplifiers (downtoners), and adversative conjunctions) while maintaining speed.
To analyze sentiment using this library we had to make some changes to our corpus by remove numbers, punctuation, URLs, hashtags, mentions, controls, special characters, leading and trailing white spaces and finally converting all the text to lowercase.
- `Vader`: which is developed to deal with the language patterns used in social media such as: words that increase the sentiment like **great** or **love**, words that are all caps are often used to amplify emotion, marks such as **!** or **!!!** will increase the sentiment, social media slang words and emojis.
As for this method, we did not need to bring any changes to the corpus so we applied the function directly to the raw tweets' text. After comparing these two methods' results, we chose to consider the **Vader** results since they were more accurate when doing random check on the data to see if the sentiment assigned goes with the content of the tweet or not.

Topics' modeling

Model creation & evaluation

Insights

Overview

- Dunkin Donuts had a the highest peak of tweets for the period of analysis, during **September 2020**, with tweets about the **National Dunkin day** that took place on Tuesday, September 29, 2020, calling customers to visit DD's locals and announcing different offers on that occasion (`60 tweets`).

- The majority of Dunkin Donuts tweets were publishing during the afternoon `64%`, followed by tweets posted during the evening `28%`, with a dominance for tweets published in the **weekdays** rather than **weekends**.
- **Diverse tweets** are the one dominating the coverage with `34%` of the share of voice, followed by tweets announcing **special offers** `14%`, and **special celebrations** posts `13%`.

- Most of the tweets are with length around the **ideal tweet's length** which is **100 characters**, with the range [60-80] characters being the one with the highest number of tweets `114 tweets`.

- Positive tweets are dominating the coverage with `59%` of overall tweets.