COVID vaccines analysis

Goal : Classify tweets sentiment towards covid19

Use cases

1. Detecting people's opinion towards vaccine
2. Identify overall customer ratings for various vaccines from different Providers
3. Investigate potential social problems about vaccine

Data Collection

- Create developer accounts for TwitterAPI
- Investigate JSON response and filter out relevant attributes for classification
- Use python Tweepy to connect the API
- Scrape data related to Covid19 Vaccine. Pfizer, Moderna, Astrazeneca Sinovac
- Combine all sources of data.

Data Annotation

1. Target: pos, neu, neg
2. Evaluate if the tweet can be classified, if no, put 'del in the target column
3. Decision-making:
    - How to deal with questions?
    - Do we treat subjective and objective tweets differently?
    - How to deal with tweets with several sentences and compound sentiments?
    - How to deal with modality, eg, should, could have been etc.?

- Tweets with political standpoint might have personal sentiment tendency?
- How to deal with wish or command?
- How to deal with Sarcasm?

4. Criteria regarding to annotation:

- Questions will depend, determine based on the tone.
- Usually subjective tweets have a strong sentiment and objective tweets have a neutral sentiment such as news headlines or ads Classify subjective ones as pos or neg except they are obviously neutral and objective ones as neutral unless they are obviously pos or neg
- If tweets have more than two negative or positive words, tweets should be positive or negative. If it has a tie, classify as neutral.
- Modality will depend, but cases like could have been are usually negative
- Do not pick standpoints, stand in the middle point to classify the sentiments.
- Wishes and hope are classified as positive. Command usually is negative Sarcasm should be classified as negative.
- If we have to read more than two times to determine, put 'del or neu

Data Cleaning

1. Remove duplicates tweets based on text
2. Extract location from user column, discard coordination column
3. Extract hashtags from entities column
4. Split location into the format of city, province(two capitalized letters).
5. Preprocessing the full text column Exclude URLS a Exclude mention sign

- Exclude HTML reference characters to keep entities column clean, convert it to the number of hashtags
- Delete the untidy columns such as entities, users, location.

EDA

- Missing values analysis
- Univariate
- Numeric features frequencies (bar chart)
- Bivariate
- Relationship between predictors against target
- Multivariate
- Relationship among predictors(collinearity)
- WordCloud

Feature Extraction

- Subjectivity/Objectivity
  - Number of first-person pronouns
  - Number of second-person pronouns
  - Number of third-person pronouns
- Part of speech
  - Number of past-tense verbs
  - Number of future-tense verbs
  - Number of future-tense verbs
  - Number of common nouns
  - Number of proper nouns
  - Number of adverbs
  - Number of wh- words
  - Number of coordinating conjuctions

Corpus statistics

- Number of commas
- Number of multi-character punctuation tokens

- Average length of sentences, in tokens
- Average length of tokens, excluding punctuation-only tokens, in characters
- Number of sentences
- Number of words in uppercase(>= 3 letters long)
- Number of hashtags
- Sentiment words and phrases:
- Positive words ratio per tweet o Negative words ratio per tweet

- Other features

  - Location(city, province, country)
  - Retweet count
  - Favourite count

Building Pipeline

- Further cleaning of text for machine learning algorithm:

  - Exclude punctuations and digits
  - Convert emojis to text description
  - Lowercase all words
  - Lemmatization

Text Vectorization

Statistic approach:

- Bag of Words (hyperparameter tunning)
- TF-IDF(How relevant a word is associated to a doc)

Neural network:

- word embeddings