Twitter Sentiment Analysis

Milestone 2

Avinash Alapati

Department of Master Data Science Program

Bellevue University

DSC680-T301 Applied Data Science (2243-1)

SENTIMENT ANALYSIS

DSC 680 -PROJECT2 MILESTONE 2

Avinash Alapati

Introduction

This project aims to conduct sentiment analysis on selected tweets from Twitter, with the specific tweets chosen as the project progresses. The primary objective is to unveil the underlying sentiment expressed in users' tweets, classifying opinions into two categories: positive and negative. Subsequently, an analysis will be conducted on the classified data to determine the percentage distribution within each category among the sampled population.

In the realm of data science, Natural Language Processing (NLP) is a focal point of research, and sentiment analysis stands out as one of its prevalent applications. This field has significantly transformed various aspects of business operations, influencing everything from opinion polls to the formulation of comprehensive marketing strategies. Consequently, it has become essential for data scientists to be well-versed in sentiment analysis.

NLP's capability to process thousands of text documents for sentiment in seconds is revolutionary, drastically reducing the time required compared to manual efforts. This project will adhere to a systematic approach to address a general sentiment analysis problem. The process commences with the preprocessing and cleaning of raw tweet text, followed by an exploration of the cleaned text to gain contextual insights. Subsequent steps involve extracting numerical features from the data, and

ultimately, utilizing these feature sets to train models that can discern the sentiments conveyed in the tweets.

Business Problem:

Sentiment analysis, facilitated through Natural Language Processing (NLP), is a process that automates the extraction of attitudes, opinions, views, and emotions from diverse sources such as text, speech, tweets, and databases. The classification of opinions into categories like "positive" or "negative" is integral to this process, which is alternatively known as subjectivity analysis, opinion mining, and appraisal extraction.

While the terms opinion, sentiment, view, and belief are often used interchangeably, there are nuanced differences:

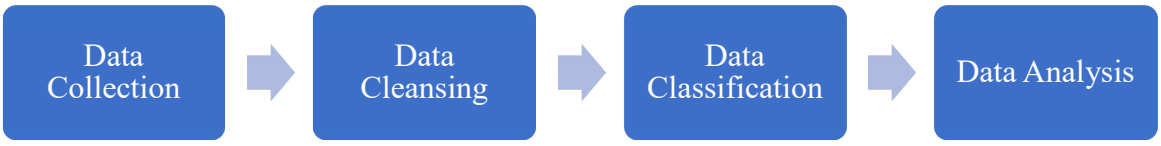
- Opinion: A conclusion open to dispute, influenced by varying perspectives among experts.
- View: A subjective opinion, reflecting personal perspectives.
- Belief: Deliberate acceptance and intellectual assent to a particular idea.
- Sentiment: An opinion that represents one's feelings.

In the contemporary landscape, sentiment analysis and NLP have become pivotal. The vast volume of information shared daily on social media platforms and blogs presents a challenge for computer comprehension. However, advancements in computer performance, aligned with Moore's law projections, and the introduction of distributed computing technologies like Hadoop or Apache Spark, have made processing large datasets feasible.

This technological progress holds immense potential for understanding textual data, significantly enhancing data analytics and search engines. A compelling use case for sentiment analysis lies in deciphering a customer's perception of a product. This valuable data empowers companies to identify product issues, discern trends ahead of competitors, enhance communication with their target audience, and gain insights into the effectiveness of marketing campaigns. Leveraging this knowledge, companies receive valuable feedback to inform the development of the next generation of their products.

Approach:

This section will highlight the technical approach that will be followed for this project and will include the system description.



Data Collection

The provided dataset is the Sentiment140 Dataset, encompassing 1,600,000 tweets obtained through the Twitter API. The dataset contains several columns, including:

- target: indicating the polarity of the tweet (positive or negative)
- ids: representing the unique identifier of the tweet
- date: indicating the date of the tweet
- flag: denoting the associated query; labeled as "NO QUERY" if no query is present
- user: specifying the name of the user who tweeted
- text: signifying the content of the tweet

Data Cleansing

A tweet encompasses a multitude of opinions expressed in diverse ways by various users. The Twitter dataset utilized in this project has already been categorized into two classes, namely negative and positive polarity. This categorization facilitates the ease of conducting sentiment analysis to observe the impact of different features. However, the raw data, accompanied by polarity labels, is prone to inconsistencies and redundancies. In the preprocessing of tweets, the following steps are undertaken:

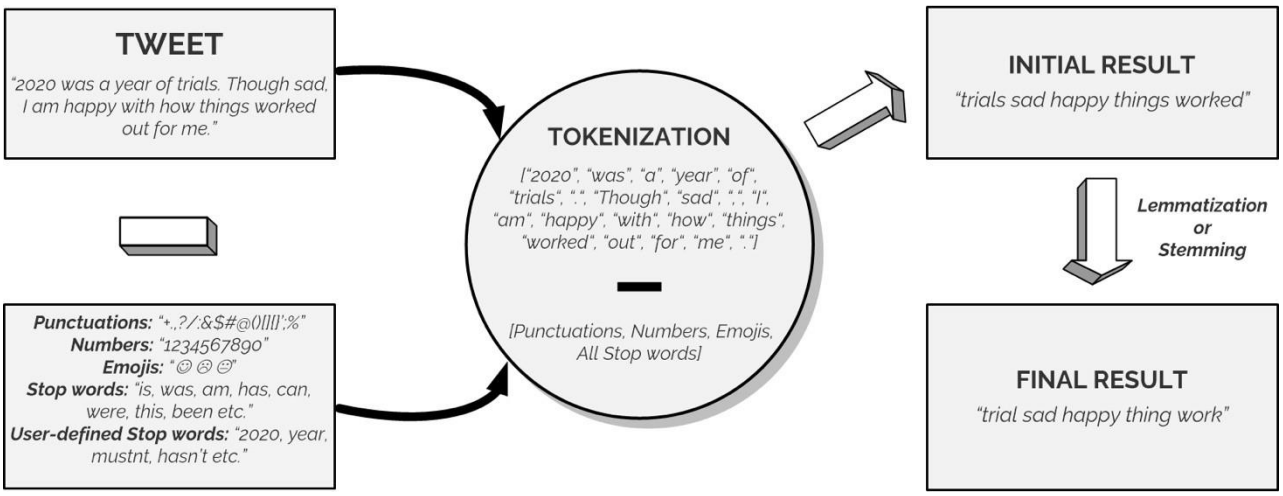
- Elimination of all URLs (e.g., www.xyz.com), hash tags (e.g., #topic), and targets (@username).
- Removal of stop words.
- Replacement of repeated characters.
- Elimination of all punctuations, symbols, and numbers.

The subsequent phase of the system involves cleansing the collected data, entailing the removal of punctuations and converting all text to lowercase. This preparatory step proves beneficial in the subsequent stages of the project, particularly in the "Bag of Words" approach. Lowercasing aids in reducing redundancy in the database used for storing words.

Classifying the data

In pursuit of the ultimate goal, the individual tweets required cleaning, a task accomplished through the application of the "Tokenization" concept in Natural Language Processing (NLP). Tokenization involves breaking a sentence into smaller units known as "tokens" to eliminate extraneous elements. Another noteworthy technique employed is "Lemmatization," a process that reverts words to their base form.

To illustrate, consider the following simple example.



Machine learning techniques necessitate the representation of key features in text or documents for processing. These features are regarded as feature vectors, crucial for the classification task. Several examples of features reported in the literature include:

1. Words and Their Frequencies:
    - Unigrams, bigrams, and n-gram models, along with their frequency counts, are considered as features.
2. Parts of Speech Tags:
    - Parts of speech, such as adjectives, adverbs, and specific groups of verbs and nouns, serve as indicators of subjectivity and sentiment. Syntactic dependency patterns can be generated through parsing or dependency trees.
3. Opinion Words and Phrases:
    - Beyond individual words, phrases and idioms conveying sentiments, like "cost someone an arm and leg," can be utilized as features.
4. Position of Terms:
    - The position of a term within a text can influence its impact on the overall sentiment of the text.
5. Negation:
    - Negation, while challenging to interpret, is a significant feature. The presence of negation often alters the polarity of the opinion.
6. Syntax:
    - Syntactic patterns, including collocations, are employed as features by many researchers to learn subjectivity patterns.

Addressing this aspect of the project is expected to be challenging, involving an examination of individual words or word groups in a tweet to assign sentiment. This task is complex, particularly in dealing with slang words and sarcasm, which are challenging for computers to comprehend.

The "Bag of Words" Model adopts an approach involving the creation of databases for positive, negative, and neutral words. Each tweet is disassembled into individual words, compared against these databases, and assigned sentiment based on matches. A counter is incremented or decremented

by a fixed amount, depending on the assigned weighting. The final counter value determines the sentiment classification—higher for predominantly positive words, for example.

To identify common words, the POS-tag (Parts of Speech tagging) module in the NLTK library was utilized. Additionally, the WordCloud library was employed to generate a Word Cloud based on word frequency, superimposing the results on the Twitter logo using Matplotlib. The Word Cloud visually represents words with higher frequency in larger text sizes and less common words in smaller text sizes.

Negative Sentiment Word Cloud

Positive Sentiment Word Cloud

Data Analysis

Upon classifying the data, subsequent analysis becomes imperative. This analysis could encompass straightforward metrics like customer satisfaction percentages or delve into more intricate examinations. For instance, a comprehensive investigation may involve comparing customer sentiment regarding two analogous products, seeking to discern correlations between favorable sentiments and elevated sales for those specific products.

Setting up the Classification Model

Upon completing the model training, we proceed to assess its performance using evaluation measures. Subsequently, we employ the following evaluation parameters to scrutinize the models' effectiveness:

- Accuracy Score

- Confusion Matrix with Plot
- ROC-AUC Curve

Conclusion

In conclusion, our analysis indicates that Logistic Regression stands out as the most effective model for sentiment analysis on the given dataset. This choice aligns with the Occam's Razor principle, which posits that for a problem statement lacking specific assumptions, the simplest model tends to perform the best. Given that our dataset doesn't carry specific assumptions, the simplicity of Logistic Regression aligns with the principles outlined by Occam's Razor and proves to be the optimal choice for the mentioned dataset.

Questions:

1. What other algorithms can be applied to the model?
2. What are the challenges in text mining from Twitter?
3. What are the challenges in data cleansing?
4. What kind of data needs to be filtered from the text?
5. Are there any available data sources to mine data for twitter analysis?
6. What is the confidence level in predicting the sentiments?
7. Which model provides the best prediction?
8. What ethical considerations were taken care while analyzing the data?
9. What are the future improvements that you can specify for this analysis?
10. Can this model be able to predict sentiment of any product which is tweeted?

References

11. Alec Go, Lei Huang, Richa Bhayani, 2009. Twitter Sentiment analysis, s.l.: The Stanford Natural Language Processing Group.
12. https://developer.twitter.com/en/docs/tutorials/how-to-analyze-the-sentiment-ofyour-own-tweets
13. https://textblob.readthedocs.io/en/dev/quickstart.html#sentiment-analysis

14. https://towardsdatascience.com/sentiment-analysis-of-tweets-167d040f0583
15. Alexander Pak, Patrick Paroubek, 2010. Twitter as a Corpus for Sentiment Analysis and Opinion Mining, s.l.: LREC.
16. Berry, N., 2010. DataGenetics. [Online]
17. Available at: http://www.datagenetics.com/blog/october52012/index.html [Accessed 14 04 2014].