Project 2: Sentiment Analysis from X data – Milestone:3

Author: Venkat Jagadeesh Jampani

Date: 05/24/2024

Table of Contents

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.

Background/History

To perform sentiment analysis on Twitter data, you need a comprehensive understanding of the entire process, including setting up access to Twitter data, preprocessing, feature extraction, model training, and evaluation. Below, I provide detailed steps and code snippets to guide you through each phase.

Data Explanation

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:

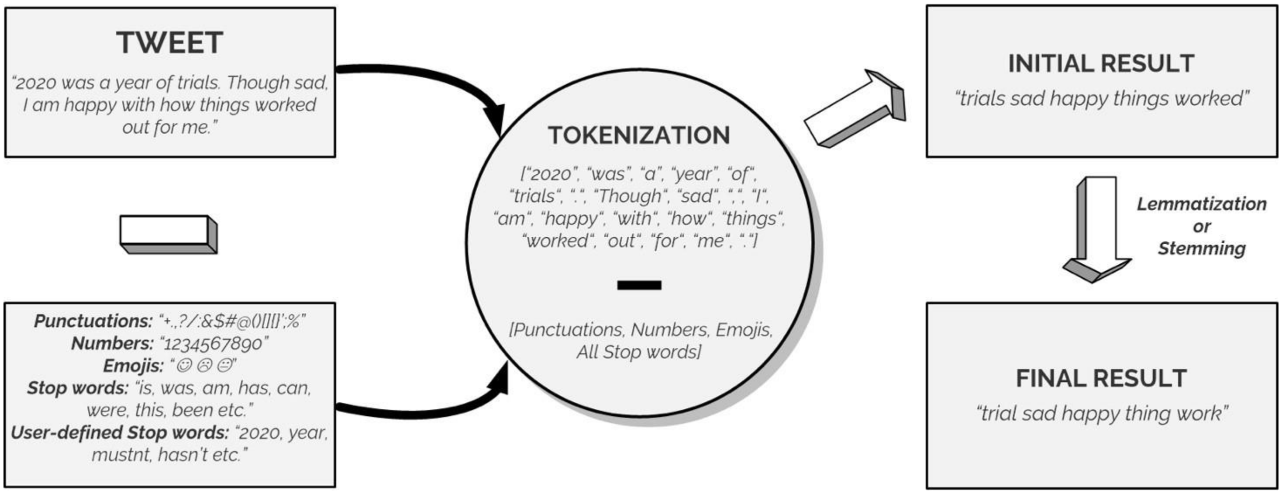
- Elimination of all URLs (e.g., www.xyz.com), hash tags (e.g., #topic), and targets (@username).
- Removal of stop words.
- Replacement of repeated characters.
- Elimination of all punctuations, symbols, and numbers.

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.


TWEET
"2020 was a year of trials. Though sad, I am happy with how things worked out for me."
Punctuations: "+.,?/:&$#@()[]{}';%"
Numbers: "1234567890"
Emojis: "☺ ☹ 😑"
Stop words: "is, was, am, has, can, were, this, been etc."
User-defined Stop words: "2020, year, mustnt, hasn't etc."
TOKENIZATION
["2020", "was", "a", "year", "of", "trials", ".", "Though", "sad", ",", "I", "am", "happy", "with", "how", "things", "worked", "out", "for", "me", "."]
−
[Punctuations, Numbers, Emojis, All Stop words]
INITIAL RESULT
"trials sad happy things worked"
Lemmatization or Stemming
FINAL RESULT
"trial sad happy thing work"

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:

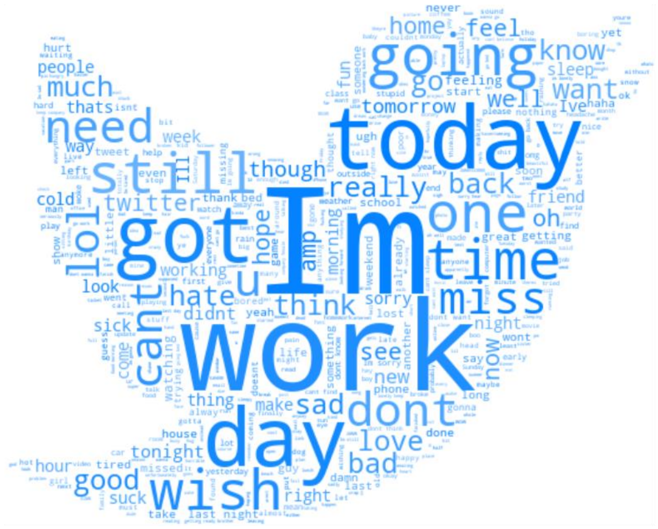
1. Words and Their Frequencies:
   - Unigrams, bigrams, and n-gram models, along with their frequency counts, are considered as features.
2. Parts of Speech Tags:
   - Parts of speech, such as adjectives, adverbs, and specific groups of verbs and nouns, serve as indicators of subjectivity and sentiment. 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.

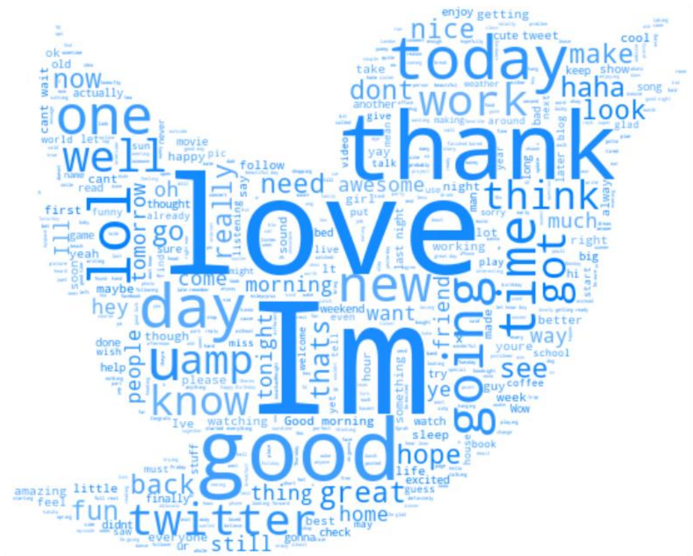
Addressing this aspect of the project is expected to be challenging, involving an examination of individual words or word groups in a tweet to assign sentiment. This task is complex, particularly in dealing with slang words and sarcasm, which are challenging for computers to comprehend.

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

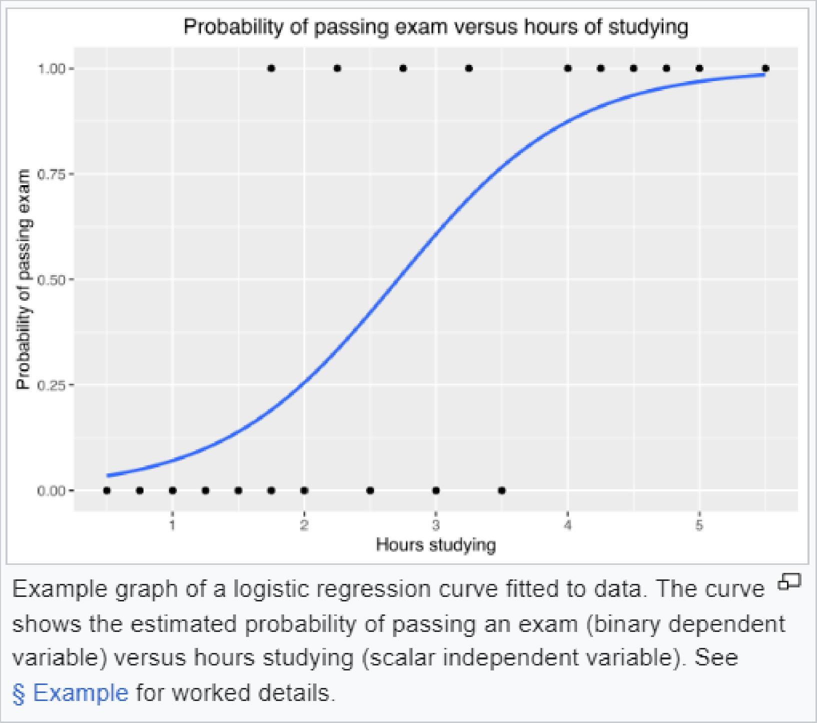
Upon classifying the data, subsequent analysis becomes imperative. This analysis could encompass straightforward metrics like customer satisfaction percentages or delve into more intricate examinations. For instance, a comprehensive investigation may involve comparing customer sentiment regarding two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

Methods

Below algorithms or model techniques will be utilized on the dataset to determine which features are related to our target variable "winner". Since the output of winner prediction is a categorical value, the problem which we are trying to solve is a Classification problem.

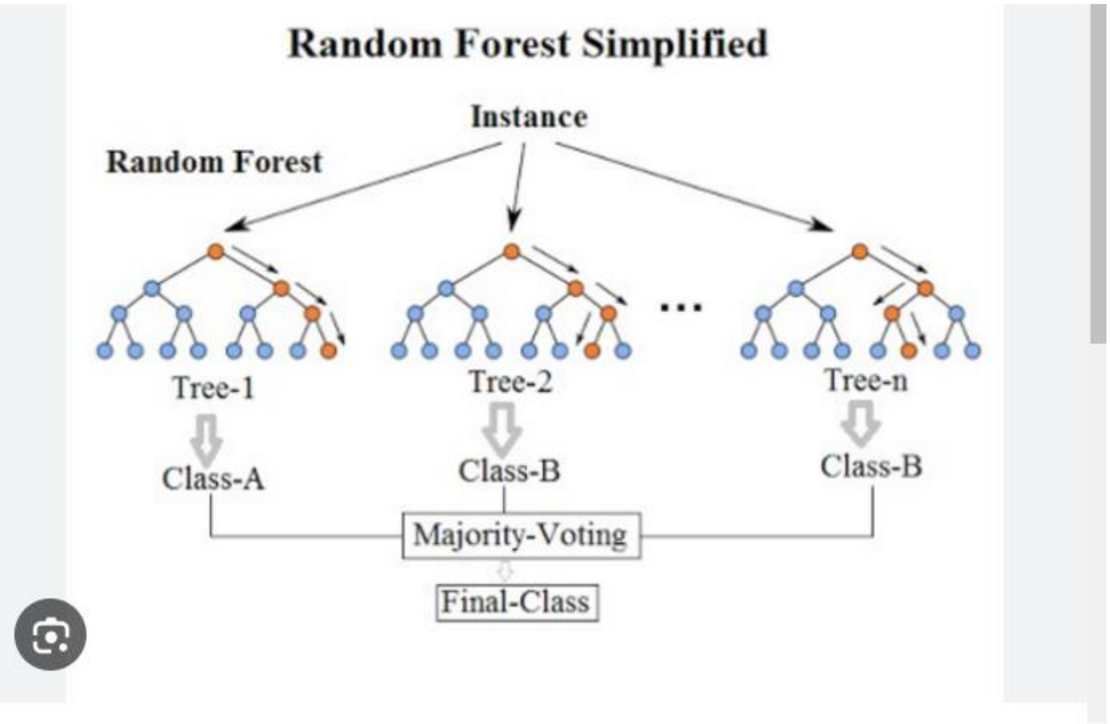
1. Logistic Regression
2. Random Forest
3. Decision Tree
4. Support Vector Machine (SVM) Classifier
5. K-fold (if required)
6. ROC-AUC Curve

Logistic regression is a statistical analysis method used to predict a binary outcome such as yes or no based on prior observation of the data set. Here, "Purchase" feature present in the dataset has only binary values and will be used as target for the model. This model falls under supervised learning as the data is well labelled and has a target variable, a column in the data representing values to predict from other columns in the data. Under supervised learning, this dataset falls under classification model as it reads the input and generates an output that classifies the input into two categories: one having purchase as "Yes" and "No". Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes.

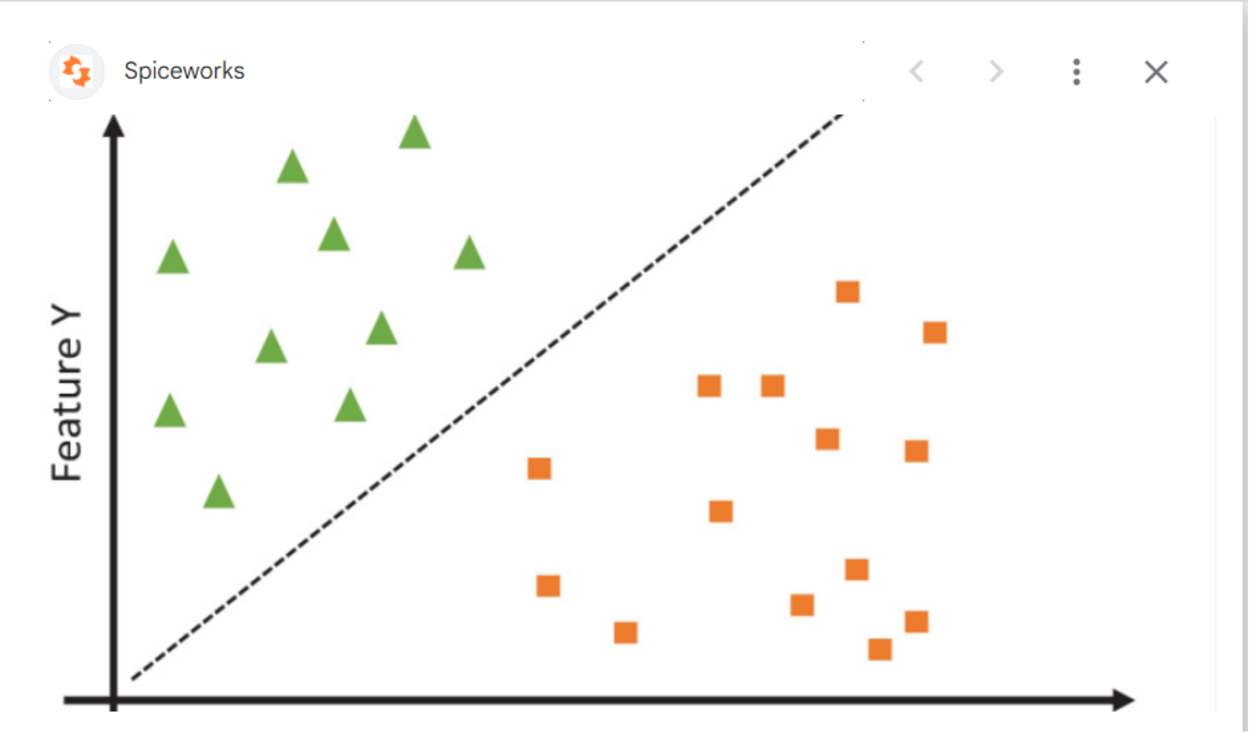


Example graph of a logistic regression curve fitted to data. The curve shows the estimated probability of passing an exam (binary dependent variable) versus hours studying (scalar independent variable). See § Example for worked details.

Random forests or random decision forests is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time.



Support vector machines so called as SVM is a supervised learning algorithm which can be used for classification and regression problems as support vector classification (SVC) and support vector regression (SVR). It is used for smaller dataset as it takes too long to process. In this set, we will be focusing on SVC.



Analysis

Since this is a classification problem, multiple algorithms can train the classifier according to the data being fed and using the pattern, we would predict the outcome. We will be trying Decision Tree Classifier, Random Forest Classifier, Logistic Regression and possibly SVM or K-fold and finally choose the algorithm best suited for this use case data.

This should include the modeling analysis with accuracy score calculated for all the models.

Accuracy: Accuracy represents the number of correctly classified data instance over the total number of data instances.

This should also include feature analysis using various methods to find the best features from the dataset. Best feature in the dataset which shows higher impact to the target variable “Match Winner” compared to others present in the dataset.

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.

Assumptions

The datasets being considered may not have all the required features to support the model. I have taken the best possible dataset from the available sources. Also, some other features which may not be relevant will also be excluded.

Limitations

The dataset considered for this prediction model is to be considered a fictional dataset as it may not represent real-world or factual data. The same goes for the supplemental datasets as well. Also, the number of features included in the current dataset may not be enough for an accurate prediction model although good enough to build one.

Challenges

Although the datasets taken from Kaggle have great deal of information, we can only assume that this is not an accurate dataset and not being fact checked.

The models employed may not be highly accurate but given the data, anything more than 50-60% can be reasonably considered to be good.

More features might be required to enhance this model.

Future Uses/Additional Applications

While this may not exactly represent the real-world data, this model is still similar and can be run against real-world datasets to all other similar social media applications around the world to gain useful insights.

Recommendations

This model predicts the sentiment analysis and relevant useful features that impact the prediction with better accuracy with a caveat that the model should be regressed when more or better real-world data is available.

Implementation Plan

As stated in the recommendations, this model can be implemented to predict the user sentiment, trends of users, along with evaluation of other useful features that may impact the same outcome. While evaluating other features from the datasets, model must be ensured to re-evaluate for no slippage.

Ethical Assessment

There are no possible ethical aspects to this model as the data is public info and doesn't really include any consumer or personal related information.

References

1. Alec Go, Lei Huang, Richa Bhayani, 2009. Twitter Sentiment analysis, s.l.: The Stanford Natural Language Processing Group.
2. https://developer.twitter.com/en/docs/tutorials/how-to-analyze-the-sentiment-ofyour-own-tweets
3. https://textblob.readthedocs.io/en/dev/quickstart.html#sentiment-analysis
4. https://towardsdatascience.com/sentiment-analysis-of-tweets-167d040f0583
5. Alexander Pak, Patrick Paroubek, 2010. Twitter as a Corpus for Sentiment Analysis and Opinion Mining, s.l.: LREC.
6. Berry, N., 2010. DataGenetics. [Online]

7. Available at: http://www.datagenetics.com/blog/october52012/index.html [Accessed 14 04 2014].