ENSE 480 Project Proposal

“Conducting a sentiment analysis on Taylor Swift to understand public opinion and perception”

Mikayla Peterson (200425538)

Introduction

The research problem I want to address is how has the public opinion and perceptions of Taylor Swift changed over time. I plan to do this by performing sentiment analysis on historical tweets containing her name or her Twitter handle, @taylorswift13, to see if tweets about her are generally positive, negative, or neutral and if the overall sentiment has changed over the years.

Scope

The goals of this project are to:

1. Collect and pre-process Twitter data/tweets in preparation for sentiment analysis.
2. Perform sentiment analysis (classifying the sentiment of the tweets into positive, negative, and neutral) on the Twitter data.
3. Create graphs and other assorted graphical visualizations of the results of the sentiment analysis as well as other relevant statistical information about the data set.

If I have time, I would like to compare multiple classification methods (Naïve Bayes, support vector machine (SVM), and bidirectional encoder representations from transformers (BERT)) and compare their classification performances against several metrics such as accuracy, precision, recall, and F1 score. I would also like to explore making an interactive dashboard to display my results with Google Looker Studio if I have the time.

Analyzing the responses and mentions to Ms. Swift's fan management team, Taylor Nation (Twitter handle: @taylornation13), is outside the project scope. If I were doing an analysis of Ms. Swift's brand, it would be appropriate to include Taylor Nation's tweets since they directly handle Ms. Swift's marketing, merchandise, and overall fan engagement on social

media. Analyzing the sentiment of her fans vs. the public vs. anti-fans would be difficult, so this is outside the project scope.

Representation and Data Structures

Input:

Tweets containing the key term *Taylor Swift*, mention *@taylorswift13*, or are a reply to one of her tweets. The final number of tweets and how far back they go is going to rely on how many tweets I can collect from the API according to the call and size limits.

Output:

The tweets will be classified as positive, negative, or neutral according to the model I make. The Jupyter notebook will have several graphs to visualize the data and the model's performance. Time permitting, I would like to make an interactive dashboard to view these graphs and show other relevant statistical information.

Techniques and Algorithms

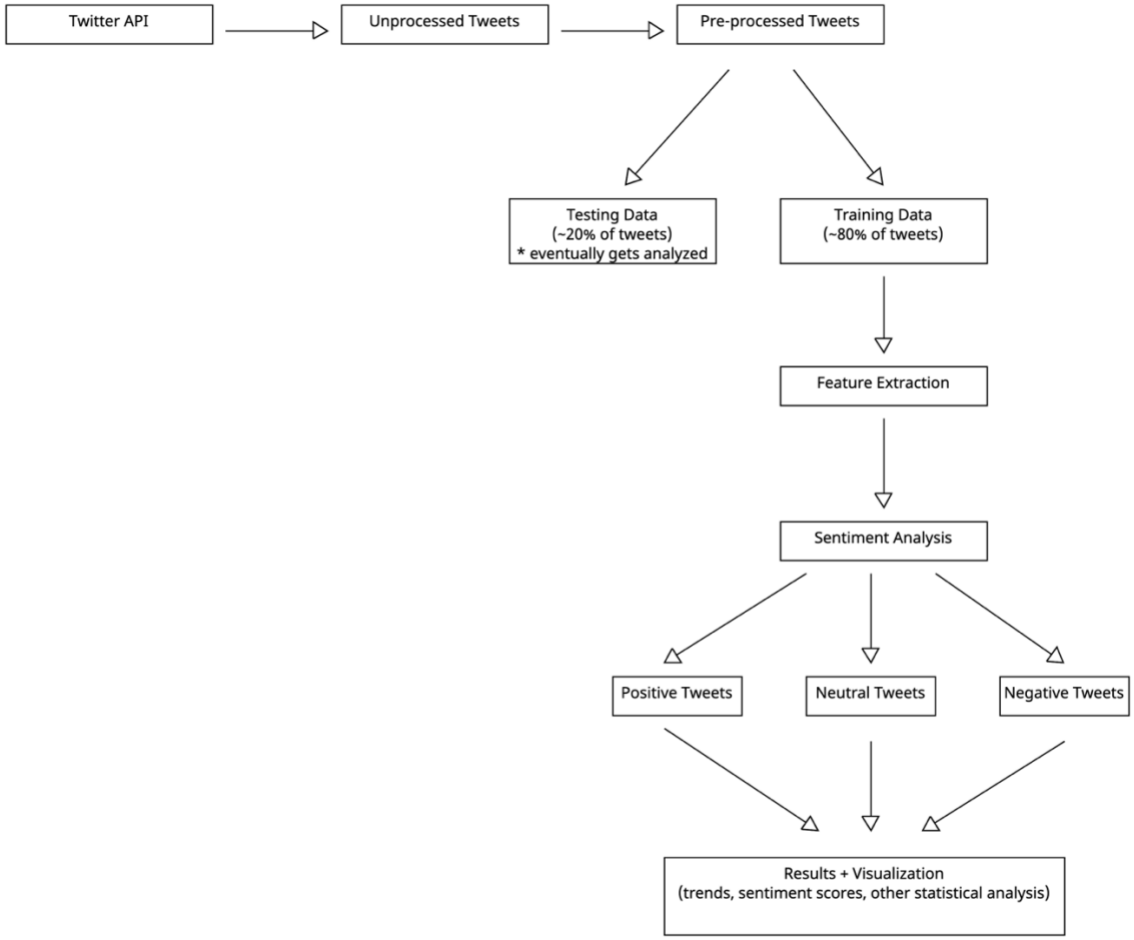
The general steps I will be following are:

1. Collect Twitter data
    a. I will be accessing the official Twitter API through Tweepy API. Other alternative libraries I can use include Python Twitter Tools and Twython. They all appear to offer the functionality I need, so picking the library will just be a matter of personal preference.
2. Pre-process and convert the pre-processed data into numerical features
    a. The Natural Language Toolkit has functions for pre-processing and feature extraction (i.e. tagging and parsing) built in. I am using this library for this specific task since it was meant to be used for general natural language processing

tasks, whereas other libraries such as SpaCy and Scikit-learn are meant to do other tasks in this process.

3. Classify a portion of the tweets for training
    a. 80% of the data will be used for training. The other 20% will be for testing.
4. Train the Naïve Bayes model with the training data*
    a. I am using Naïve Bayes since it requires a small amount of training data, and it assumes independent variables. I plan to implement it using TensorFlow or scikit-learn. Alternative models include support vector maps.
5. Evaluate the model
    a. This will be done by evaluating a series of metrics such as accuracy, precision, and F-score.
6. Perform sentiment analysis on the test data*
7. Visualize results
    a. I am planning to use matplotlib. If I have time, I would like to make an interactive dashboard using Google Looker studio. I have some experience using matplotlib, so it would be relatively simple for me to use in my project. I want to use Google Looker over other interactive dashboards such as Power BI and Tableau since Looker is entirely cloud-based and is not as complex as the other two.

A Structure Diagram



References

1.9. naive Bayes. scikit. (n.d.). Retrieved February 8, 2023, from https://scikit-learn.org/stable/modules/naive_bayes.html

Kamperis, S. (2020, December 31). *How to implement a naive Bayes classifier with tensorflow*. Let's talk about science! Retrieved February 8, 2023, from https://ekamperi.github.io/machine%20learning/2020/12/31/naive-bayes-classifier-in-tensorflow.html

NLTK. (2023, January 2). Retrieved February 8, 2023, from https://www.nltk.org/

Tweepy. (n.d.). Retrieved February 8, 2023, from https://www.tweepy.org/