Twitter Sentiment Classification

Chris Hollman
October, 2022

Overview

- An online publication group is exploring the use of twitter data to help their writers more efficiently generate content.
- The concept is to classify tweets by their sentiment and generate visuals that can narrow down a writer's field of research.
- This could cut down on prep time, identify subject matter that is generating interest, and give writers a tool to quickly write pertinent and well informed articles..

Project Objective:

The goal of this project is to build a model that is able to accurately classify whether a given tweet is negative, positive, or neutral, filtering each sentiment into individual groups.

Once organized, common words within a given sentiment group can be identified, giving insight as to what is causing these emotions or reactions among users.

Deliverables

Sentiment Distributions

These will give the writer an idea of the overall public opinion on a given subject.

Word Clouds/Frequencies

Once separated by sentiment, word clouds and frequency counts are a quick reference tool to hone in on why people feel a given way.

Example Tweets

Pulling examples of tweets containing interesting words from the visualizations will help develop context.

Examining Data

The dataset we are using today consists of 9,000 tweets collected by CrowdFlower. These tweets are related to the 2012 South by Southwest (SXSW) conference and are predominantly directed toward Google and Apple events and products.

Processing The Data

Cleaning:

- Removing punctuation
- Standardizing case
- Dropping irrelevant tweets

Processing

- Removing stopwords
- Stemming/Lemmatizing
- Vectorizing

Modeling

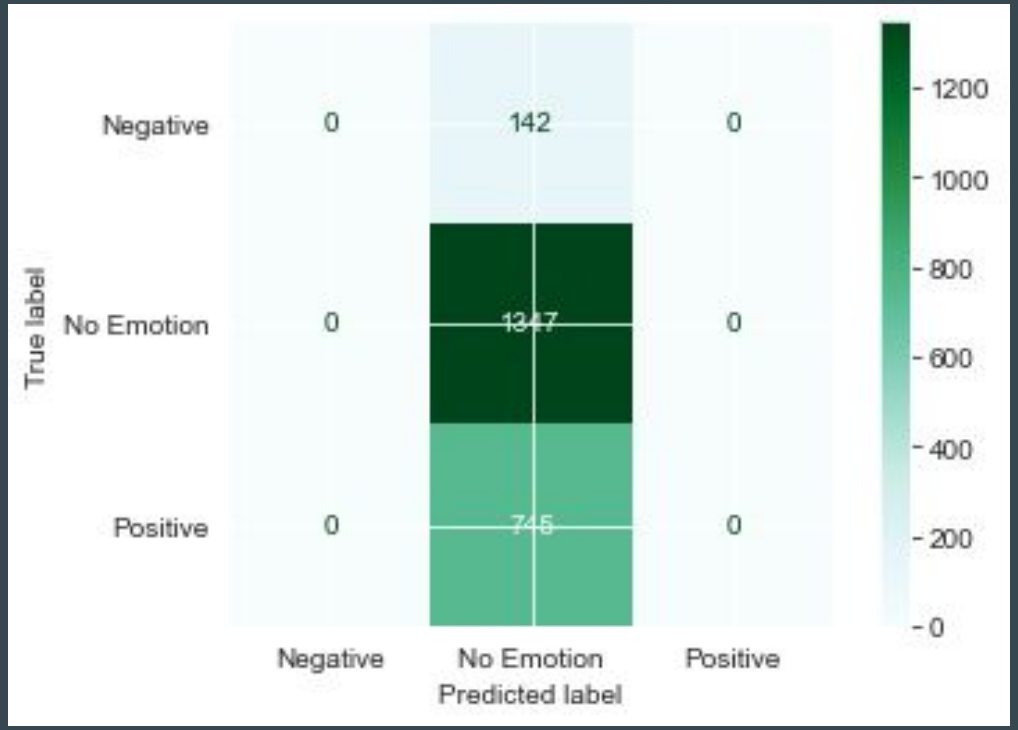
Strengths

- Majority class performance
- Overall accuracy
- Shows potential for ‘positive’ category.

Weaknesses

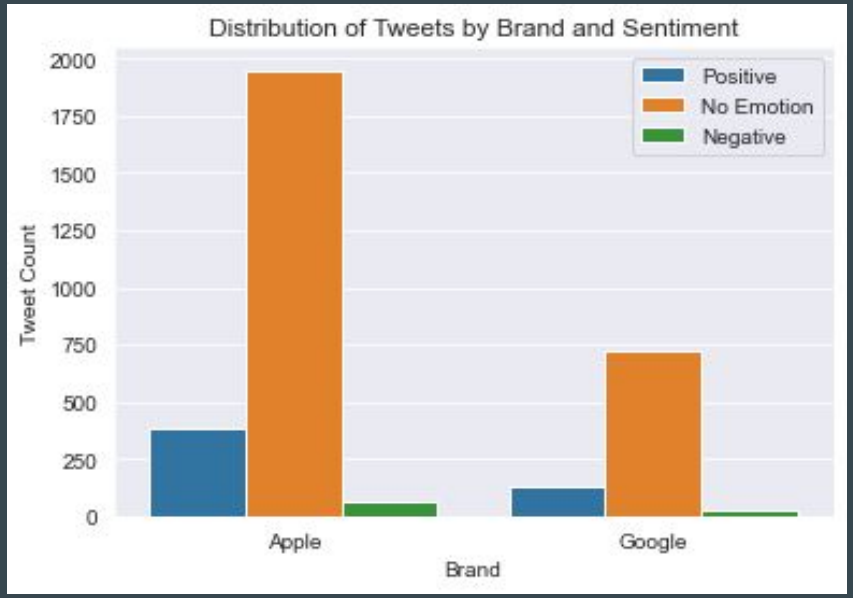
- Minority class performance
- Tendency to overfit
- Needs more examples of minority classes.

Baseline Confusion Matrix

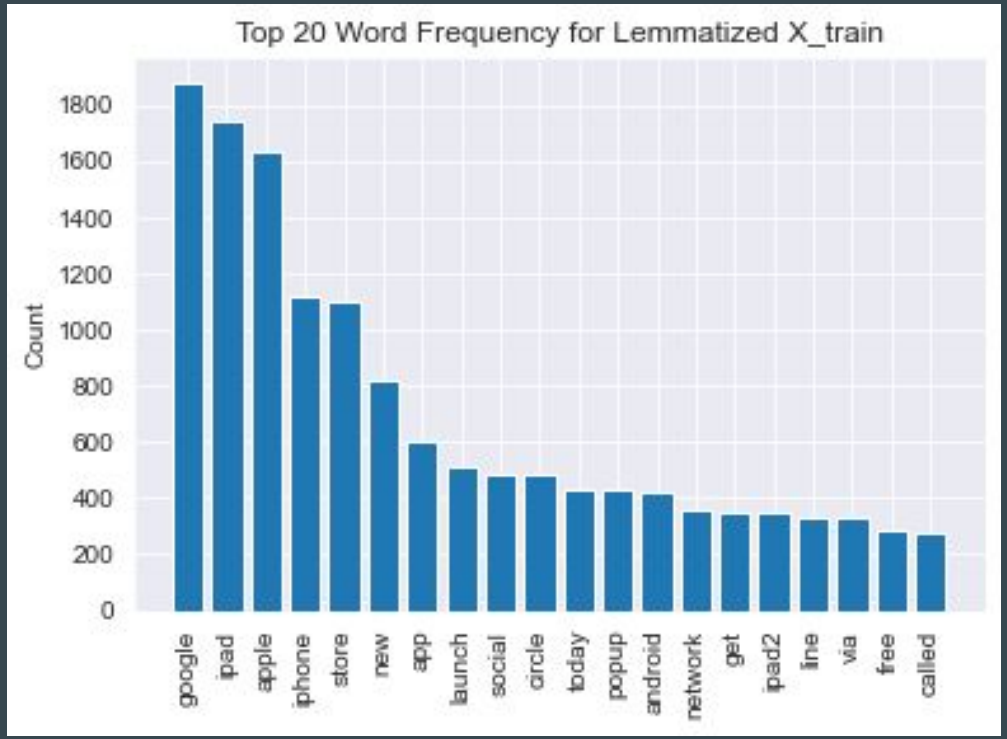


Final Confusion Matrix

Visuals/Results: Sentiment Distributions



Top 20 Word Frequency



Apple: Positive

Apple: Negative

SXSW
@User111

iPad design headaches: I hate the iPad 2 back button with the heat of a million suns.

4:05 PM · Jun 24, 2012

76 Retweets **14** Quote Tweets **619** Likes

Google: Positive

Google: Negative

Recommendations

Use of Deliverables:

The tools developed by this model should be used in the order shown in this presentation.

Potential Articles

- iPad breakdown/review
- Failed launch of Circles
- Festival overview

Next Steps

More Minority Data

The real weakness of this model is poor performance on minority classes. A possible solution is to find more examples to balance out distribution.

Other Subjects

The model also should be applied to other subject matter to test usefulness in other realms.

Recent Data

The tweets in this set are 10 years old. More recent examples should be obtained to see how our model reacts to modern day tweets.

Thank You!

Please feel free to ask any questions.

You may also reach me via email:

Chris Hollman
chollman91@gmail.com