

 **rcauvy / product-sentiment-analysis-nlp** Public

generated from [jirvingphd/osemn-project-template](#)

 GPL-3.0 License

☆ 0 stars 🍴 0 forks


☆ Star ▾

👁 Unwatch ▾


- <> Code
- 🔍 Issues
- 🔗 Pull requests
- 🎬 Actions
- 📁 Projects
- 📖 Wiki
- 🛡 Security

🔗 master ▾

...

 **rcauvy** Merge branch 'master' of [https://github.com/rcauvy/prod...](https://github.com/rcauvy/product-sentiment-analysis-nlp) ... 36 seconds ago 🕒 43

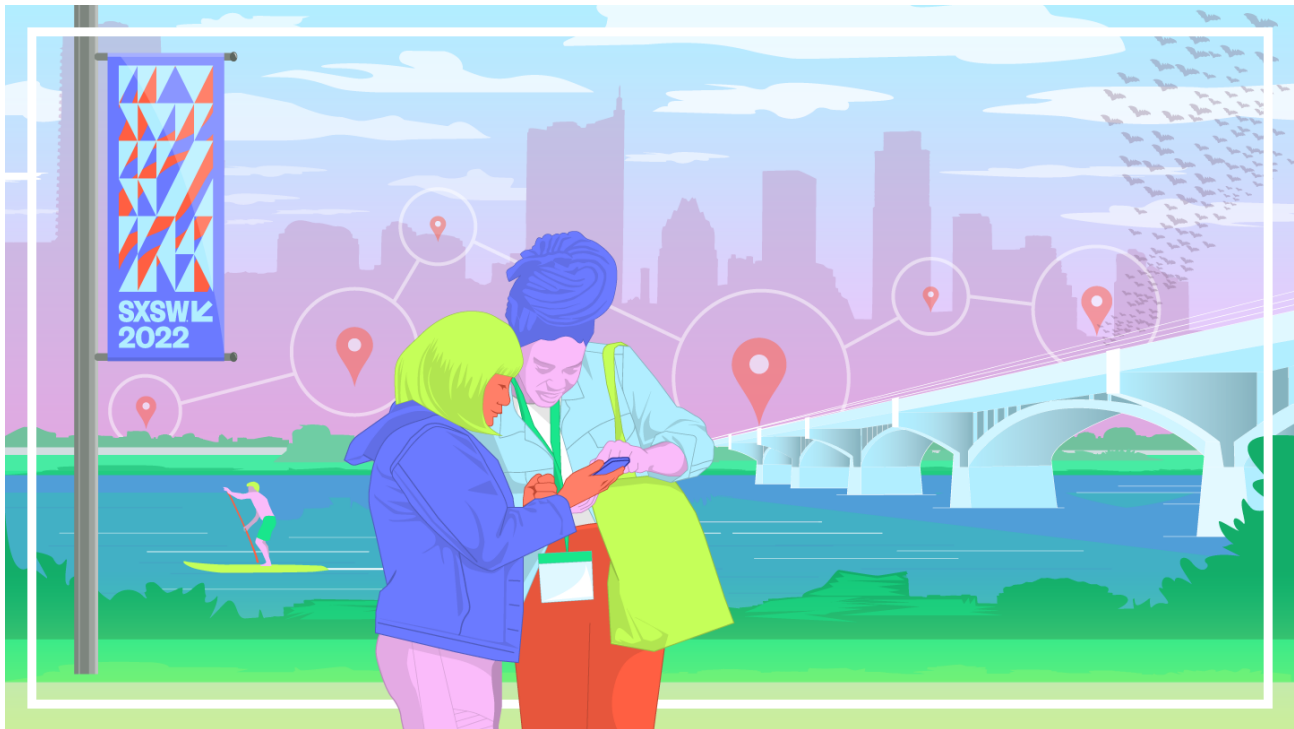
[View code](#)

☰ README.md 

Public Opinion From Social Media

Brand and Product Sentiment From Twitter Using NLP and Machine Learning

Authors: Robert Cauvy



Introduction:

Over the past decade conversations have increasingly shifted towards social media. Businesses across all industries could stand to benefit from listening to these conversations about themselves and how their products and brand are perceived by their users and prospective customers. Understanding what it is that customers enjoy the most and the least about your company's products and brand is crucial to retaining your loyal customers as well as attracting new ones. In addition to analyzing tweets various machine learning models will be trained and tested to classify tweets as either positive or negative sentiments towards the companies products and services.

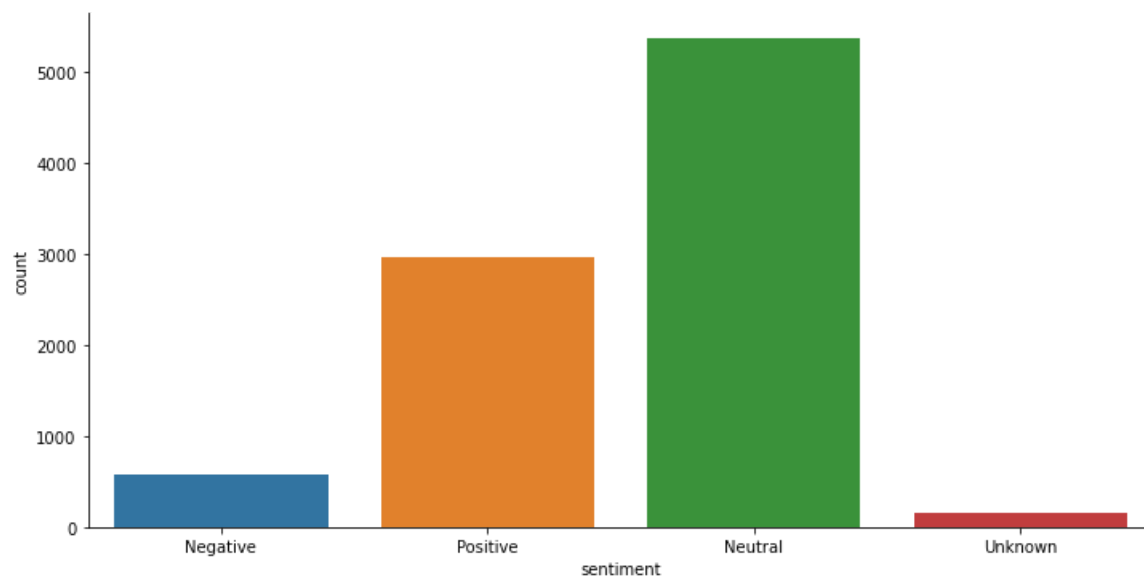
Business problem:

Understanding public opinion about the products and services can provide valuable insights. Applying human capital to track social networks is simply not a scalable solution which makes the application of Natural Language Processing and Machine Learning classifiers well suited for this business problem. The objective of this project is provide the businesses stakeholders from (Apple and Google) a model that identifies which tweets hold either a positive or negative sentiment about their brand or products from a corpus of tweets collected at the SXSW Conference. Furthermore, this project will provide the stakeholders with a list of topics and keywords that most affect public perception, leaving actionable insights for future marketing and product design decisions.

Data:

This project utilizes a dataset provided by CrowdFlower to from data.world. The dataset contains over 9,000 tweets from SXSW(South by Southwest) Conference about new product releases from Apple and Google. The tweets have been labeled as to which emotion they convey towards a particular product category or company brand based off of the language contained in the tweet. According to the provider of the dataset, humans that were tasked with labeling the sentiments of each tweet by evaluating which brand or product the tweet was about and if the tweet expressed positive, negative, or no emotion towards a brand and/or product.

Distribution on Twitter sentiment



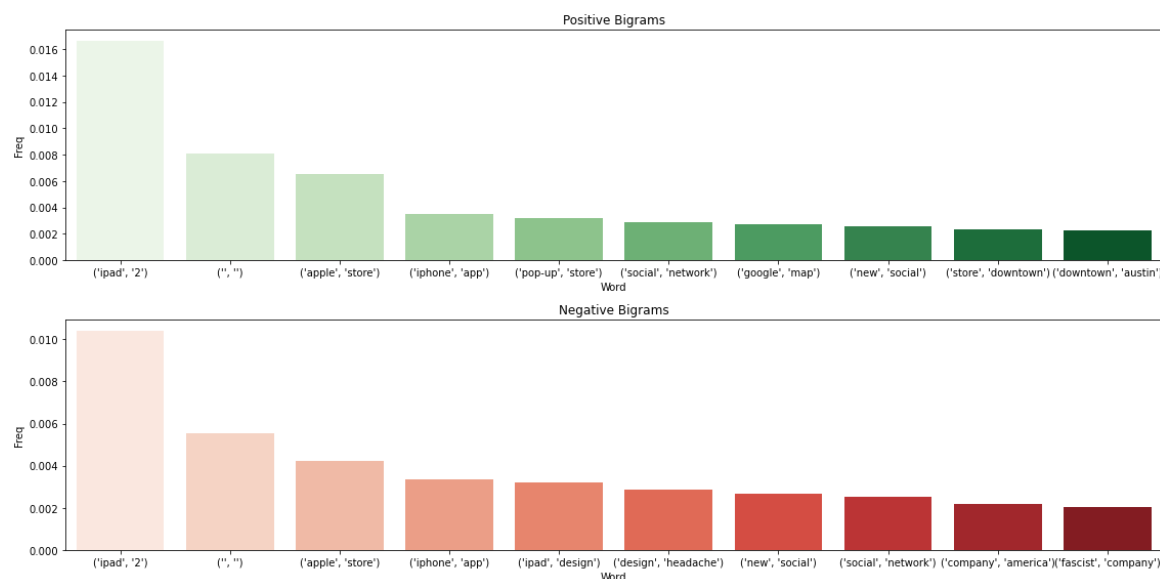
A breakdown of the sentiment distribution from the Twitter data

Exploratory Analysis

Text Preprocessing

Using NLTK's tweettokenizer url links, @mentions, punctuations and non-ASCII characters were removed. There was also a customized list of stopwords based off the NLTK default for English. Some of the added stopwords are relevant to the conference generating the tweets. Next a Document Term Matrix and Term Frequency-Inverse Document Frequency were evaluated.

Word Bigrams



This plot show the more common bigrams from positive and negative tweets.

Machine Learning Modeling

After preprocess the text data, it was first trained on Multinomial Naive Bayes classifier. The model was then hyperparameter tuned with GridsearchCV optimized for recall macro and then modeled again with RandomOverSampling to address the class imbalance. The next model tested was Logistic Regression and finally the Random Forest Classifier. The models were evaluated with sci-kit learns classification report, confusion matrix and roc curve.

Results

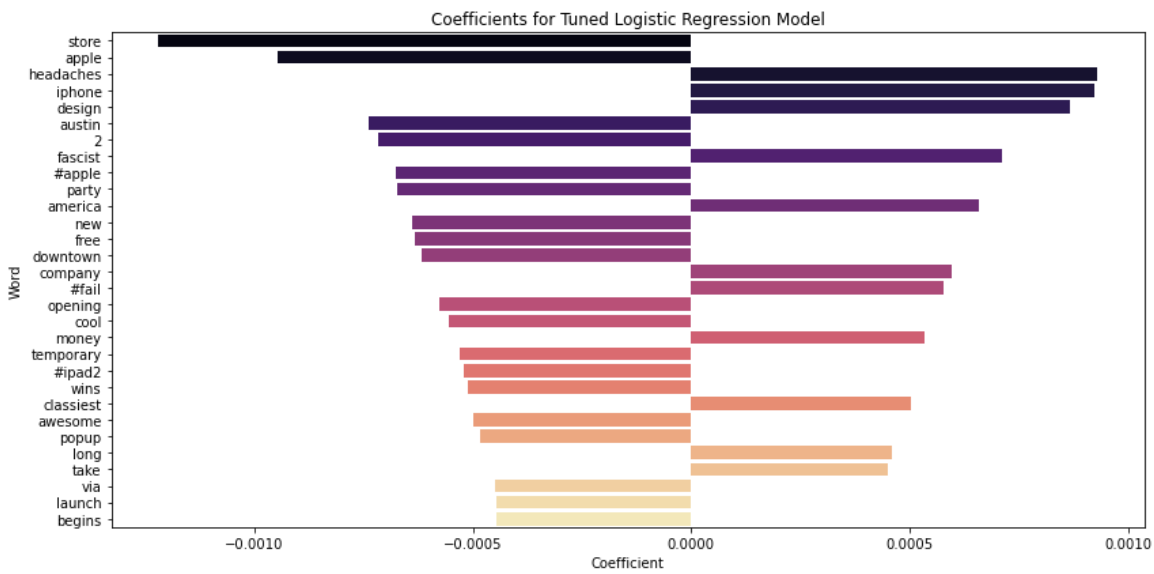
The best performing models were the hyperparameter tuned RandomOverSampled Multinomial Naive Bayes model based off an average recall macro score of 0.75. The next best performance was from the tuned logistic regression model with a recall macro score of 0.72. Ultimately it can determined that the tuned logistic regression model can be selected as the best even though the accuracy score was 0.79 compared to the MNBayes 0.84. A 0.79 accuracy score on the testing data shows that it correctly classified the tweets as having positive or negative sentiments at a rate of 79%. Not a bad score for this metric, however this model (along with all the other classifiers and iterations) was better at predicting the majority class (Positive tweets) than it was at identifying the tweets with negative sentiment. 38% of the negative tweets were incorrectly categorized as positive and 19% of the positive tweet were misclassified as being negative. All of these metric scores outperform the baseline model.

Word Cloud Comparisons



Here you can see a the word counts visualized into wordclouds comparing the tweets with positive and negative sentiments. There is a lot of crossover but some subtle differences.

Logistic Regression Coefficients



The graph plotted above shows how the words in the tweet affects the model's classification. Words such as store, Apple, and pop-up were contained in the tweet it was more likely to be categorized as positive while on the other hand if the tweet contained words such as iphone, headache, design and battery led the model to predict it as having a negative sentiment.

Recommendations:

Discovered in both the exploration and modeling phase a lot of the negative sentiment were focused on the headaches from the design, the battery of the iPhone and the associated prices. The recommendations from the negative feedback is to improve the battery life and improve on the product design. Alternatively, it would appear that the pop-up store in the downtown Austin area was very well received and should be further looked into for generating buzz at other locations during new product releases. The terms 'party' and 'free' were also linked to positive tweets about the brands. The marketing team should look to plan other events with giveaways at future conferences.

Limitations & Next Steps

The greatest limitation to this project was the size of the dataset. The data started with 9,092 records, which is not the largest mount of data to begin with. It was then later reduced down to 3,537 after dropping the tweets with neither a positive or negative sentiment that was needed for classify into a binary target variable. There was also a significant class imbalance, where only 569 tweets or about 16% of the remaining data were labeled as having negative sentiment. I would imagine that the business stakeholders of this project would be more interested in the tweets labeled as negative from both their brand and products and that if their competitors since it leads to more actionable insights.

After the size of the dataset, another limitation of this project was the target variable was a binary classification where in the real-world a multi-class model could be more useful to identify whether tweets have a positive, negative or neutral sentiment, even though the neutral tweets will not be as useful for extracting insights. The next step to this project after collecting more labeled data would be to train more complex models and other deep NLP techniques like a word2vec vecotrizer, and neural networks. Even though those models may be less interpretable, it should have a higher performance score that can be used in tandem with the successful models used here to extract coefficients.

For further information

Please review the narrative of the analysis in [our jupyter notebook](#) or review our [presentation](#)

For any additional questions, please contact [**rcauvy@gmail.com](mailto:rcauvy@gmail.com)

Repository Structure:

└─ README.md
this project.

<- The top-level README for reviewers of

└─ index.ipynb
jupyter notebook

<─ narrative documentation of analysis in

└─ presentation.pdf

<─ pdf version of project presentation

└─ images
sourced

<─ Images generated from code and externally

└─ data
analysis

<─ Tweet data from CrowdFlower used in the

Releases

No releases published

[Create a new release](#)

Packages

No packages published

[Publish your first package](#)

Languages

● Jupyter Notebook 100.0%