amandabgaeta / NLP-SXSW-Tweet-Sentiment

Code | Issues | Pull requests | Actions | Projects | Wiki | Security | Insights | Settings

master | Go to file | Add file | Code

amandabgaeta Create presentation.pdf ... 1 minute ago 8

data	First commit	10 hours ago
images	updated images, titling wordclouds, pdfs	26 minutes ago
scratchWork	First commit	10 hours ago
.gitignore	First commit	10 hours ago
Preprocessing_V1.ipynb	updated images, titling wordclouds, pdfs	26 minutes ago
Preprocessing_V1.pdf	updated images, titling wordclouds, pdfs	26 minutes ago
Preprocessing_V2.ipynb	updated images, titling wordclouds, pdfs	26 minutes ago
Preprocessing_V2.pdf	updated images, titling wordclouds, pdfs	26 minutes ago
README.md	grammar edits	4 hours ago
presentation.pdf	Create presentation.pdf	1 minute ago

README.md

Introduction: Business Problem

About

No description, website, or topics provided.

Readme

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

Jupyter Notebook 100.0%

A marketing agency specializing in brand representation at large scale events/festivals (like SXSW) is building an internal tool to help manage social presence of major brands.

This project is the first step in building a base model to fuel auto-responses and amplification of positive tweets for their clients that will aid in social presence and positive interactions with the brand.

The agency invested in manually labelling these tweets positive/negative/neutral to support building the first model with the intention to build this dataset and predictiveness over time.

The Data

The modeling iterations have been created using 2011 SXSW tweet sentiment dataset of ~9000 tweets. The file provided utilized can be found in the data folder of this repo ('judge-1377884607_tweet_product_company.csv').

Link to data source: https://data.world/crowdflower/brands-and-product-emotions

Data Processing

Preprocessing_V1.pdf and Preprocessing_V2.pdf includes cleansing, editing, functions, and feature engineering that was utilized in model iterations.

The /scratchWork folder has variations of data discovery, data analysis, model approaches, and notebook iterations that were used to get to the final Preprocessing_V1 and Preprocessing_V2 notebooks.

Model Iteration and Results

There were two different approaches to preprocessing that data that impacted model performance.

Preprocessing_V1.pdf documents the first approach to preprocessing utilizing a tweet specific processing package along with nltk to create the dataset input for modeling. Additionally this notebook has iterations upon the dataset where feature engineering was applied to denote brand (Apple versus Google) and type of tweet subject (product, service, app, none) that was appended to the sparse matrix for the vocabulary utilized in each tweet.There was limited ability to get movement in model score results beyond the baseline in regards to accuracy and precision metrics. In recognizing that preprocessing decisions may have played a part in this, the Preprocessing_V2 notebook was started.

Preprocessing_V2.pdf documents the most updated approach to this classification model. This V2 notebook applies the SpaCy package in the preprocessing step and leads to improvement in the goal metric of precision using an SVM.

The final model using SVC produced an accuracy score slightly higher than a baseline random guess at 69%, and more importantly a very high precision score at 91% with a mean 3-fold cross validation score of 87%. The F1 score was extremely low at 14%, further proving this model's pure focus on the precision metric versus recall.

Data Analysis

The baseline dataset shows that:

- 61% of the tweets are neutral, 33% are positive and 6% are negative in sentiment
- 36% are directed at a brand
- of the tweets that are directed at a brand, 73% are directed at Apple and 37% are directed at Google Note: this analysis can be found in Preprocessing_V1.pdf

Additionally WordCloud iterations in Preprocessing_V2.pdf can show us:

- In positive sentiment tweets:
    - More Apple product representation (ipad, iphone, ipad2, iphone app)
    - Apple products largely mentioned over brand versus Google which was more brand heavy
- In neutral/negative sentiment tweets:
    - Google social network has high representation as Circles was announced and launched in 2011. A particular tweet around the news was shared ~250 times and impacted the dataset

- Due to low proportion in negative tweets in complete dataset, "news" like Circles announcement was large representation of these wordclouds

Conclusion

Recommendations

On the business side, the team should get clear on vision for internal platform development in order to help guide specific model development beyond this test run. Additionally there should be consideration of investment in resources and timing for R&D around additional uses of this upfront data investment.

From the data perspective, as seen in the wordclouds of unique tokens of positive versus neutral/negative, 2-5 tweets with a unique word made it more influential. This points to the need for more data to train the model accurately. Thus, this model should be scaled by scraping Twitter utilizing hashtags of the events to get relevant tweets for SXSW in years past. The business will need to invest in the same service to label these tweet sentiments for model training purposes as well.

Future Work

This model baseline can be further invested in through more expansive scaling which could be accomplished by scraping tweets for additional events our clients/brands are represented at. Testing could include building models by event and/or as a whole depending on fit for business goals of internal platform

If business decided to prioritize and invest additional R&D resources mentioned in the recommendations section, future iterations of modeling still aiming for high precision, but with a focus on utilizing a more interpretable model type. This approach would aid in understanding the specific characteristics of positive sentiment within tweet language. The ability to deep dive into these results could be applied better at a strategic level to plan what to invest in for future events for positive attendee experiences and positive public press for brands.