South by Southwest Tweet Content

Elizabeth Webster
October 2022

Introduction

This project explores the use of a classification model in order to predict the sentiment of a tweet.

We will be using natural language processing to increase the accuracy of our model.

Outline

- Business Problem
- Data
- Methods
- Models
- Text Patterns
- Recommendations
- Next Steps

Business Problem

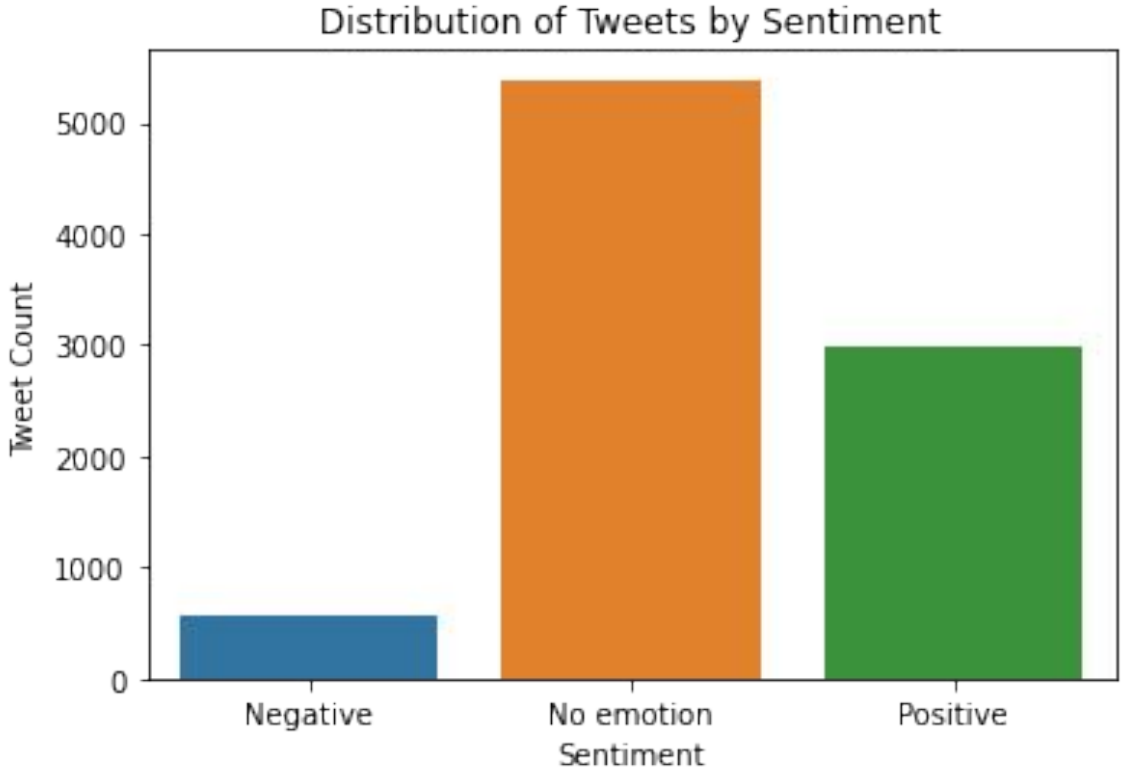
- South by Southwest - SXSW
- Enhance experience for future attendees
- Ability to assess how a current conference is progressing

Data

Dataset from CrowdFlower

- Over 9,000 tweets pertaining to SXSW
- Three columns:
  - Tweet Text
  - Product
  - Sentiment
- Sentiments:
  - Positive
  - Negative
  - No Emotion
  - I Can’t Tell

Distribution of Data By Sentiment



Methods

Data Cleaning:

- Finding and filling missing data
- Removing unnecessary data
- Train-test split

Natural Language Processing:

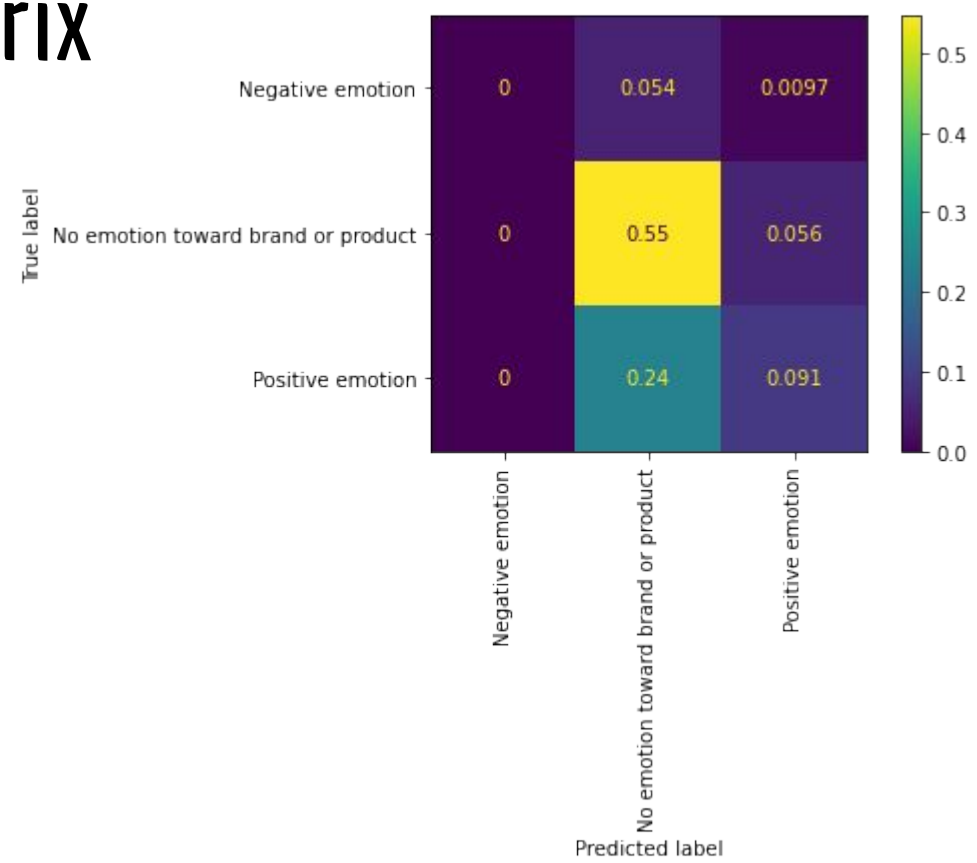
- Removing stopwords and punctuation
- Distilling words to their root - Lemmatization
- Adding features

Models - Multinomial Naive Bayes

- Baseline Model - 60.2% accuracy
  - As accurate as assigning ‘No Emotion’ to all
- Removing stopwords - 60.8% accuracy
- Lemmatizing - 60.7% accuracy
- Increasing Features - 64% accuracy
- Final Model:
  - 66% accuracy for train set
  - 64% accuracy for test set

Models - Confusion Matrix

- Model is overwhelmingly predicting ‘No Emotion’
- Model is not predicting ‘Negative Emotion’ at all



Text Patterns

Positive Word Cloud

Negative Word Cloud

Recommendations

- Utilize Word Clouds to identify commonly used positive and negative words
- Collect more data labelled as positive or negative
- Identify and remove SXSW specific stopwords

Next Steps

- Explore sentiments for specific products
- Part of speech tagging for more accurate lemmatization

Thank You!

Email: eaw524@gmail.com

GitHub: https://github.com/elizabeth524