YouTube Sentiment Analysis

By Andie Donovan, Shon Inouye, and Matthew Peterschmidt

Agenda

An overview of what we are going to talk about today

1. Introduction
2. Data Modeling Steps
3. Results
4. Live Demo
5. Conclusions

Introduction

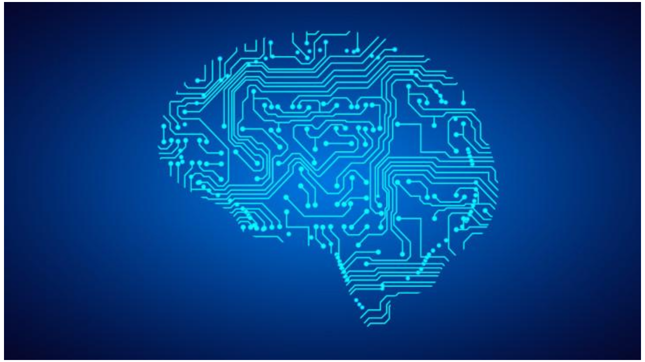
Basic overview of NLP and our business problem

- Analysis of YouTube comments
  - Machine Learning in Python
- Sentiment Analysis: Determining emotions and attitudes from text
- Insights from social data are extremely valuable

Goals

Why study YouTube comments?

- Perform sentiment analysis
  - Positive, Negative, and Neutral
- Create a user-friendly application



Data Modeling Steps

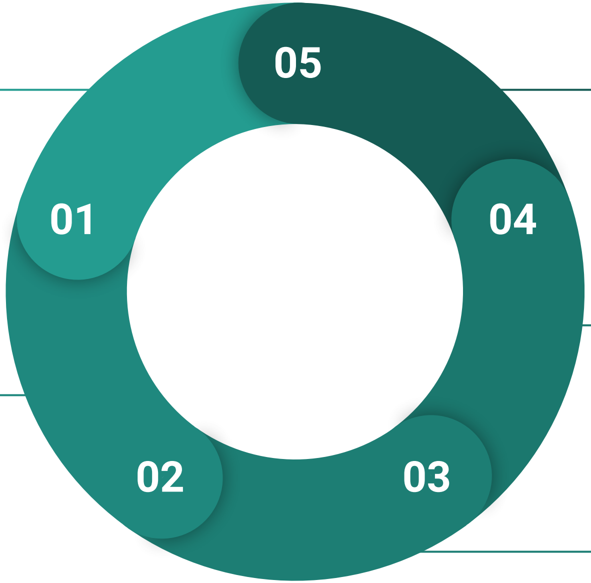
Steps in the data science process

Data Collection

Choosing pre-labeled social media data sets and extracting comments from select videos from the API

Data Cleaning and NLP

Cleaning up the comments and performing natural language processing to reduce noise and redundancy in the data



Data Visualizations

Plotting graphs, creating word clouds, and building a dashboard to showcase results

Making Predictions

Using the models to make predictions on the classification of the comments based on fitted models

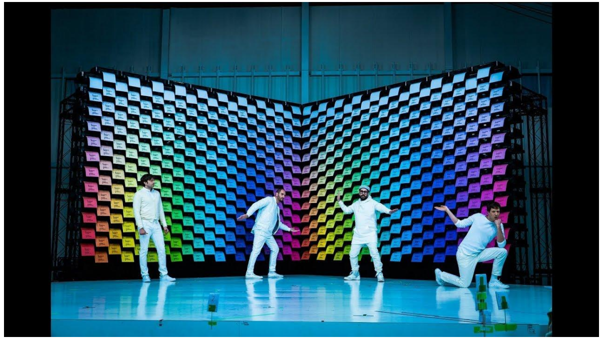
Data Transformation and Modeling

Transforming the textual into numeric format and fitting machine learning algorithms on labeled training data

The Data

Sources and labelling mechanisms

- Manually labeled comments: 2,633
  - OkGo's "Obsession"
  - Trump Inauguration
  - Logan Paul in Japan
  - Taylor Swift
  - 2018 Royal Wedding
- Obtained pre-labeled data: 12,198
  - Twitter Dataset
  - Social Media Blogs
- User Entered Video



The Data

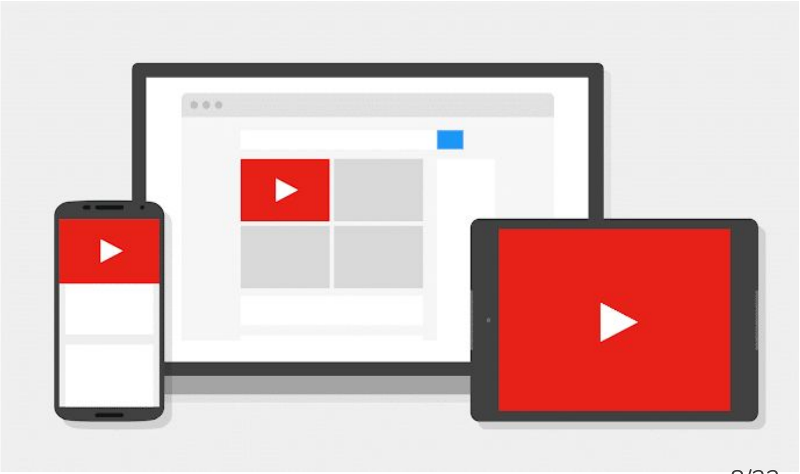
Sources and labelling mechanisms

Label	Comment
-1	Everyone knows brands of papers, but no one knows about welfare
0	Your paper cut balance is ...
1	Made me smile. Great work
1	Blowing my mind yet again
0	Should have gone with Dunder Mifflin
1	The mad methodical geniuses do it again
-1	Waste of ink and paper

Data Collection

Pulling data out of Google's YouTube API

- YouTube API (Application Program Interface)
- Real-time comments from videos



Data Cleaning & NLP

Cleaning up the data and performing Natural Language Processing on texts

- Removing non-alphanumeric characters (ex: %, &, *, $, #, @)
- Natural Language Processing:
  - Removing stop words (ex: a, that, at, this)
  - Lemmatization
  - Stemming

- Data Transformations
  - N-grams
  - TF-IDF

Natural Language Processing

Teaching the computer how to process human language

Comments

I am really great at commenting on videos. Yep--that video was a great one!

Comments are cool!

Tokenizing

~~I~~, ~~am~~, really, great, ~~at~~ commenting, ~~on~~, videos, yep, ~~that,~~ video, ~~was~~, ~~a~~, great, one

comments, are, cool

Lemmatization + Stemming

really, great, comment, video, yep, video, great, one

comment, cool

Transformations

0.038, 0.075, 0.000, 0.075, 0.038, 0.075, 0.075, 0.038

0.000, 0.150

Models

Machine Learning models used in predicting classification outcomes

- Models
  - Multinomial Naive Bayes Model *
  - Multinomial Logistic Regression *
  - Kth Nearest Neighbor
  - Linear Support Vector Machine
  - Random Forest
  - Gradient Boosting *

- Randomized Grid Search
  - Hyperparameter Tuning
- 5 Fold Cross Validation
  - Model Validation

Analysis of Results

Estimated accuracy of models and sentiment ratio results

- Training and testing on YouTube data only:

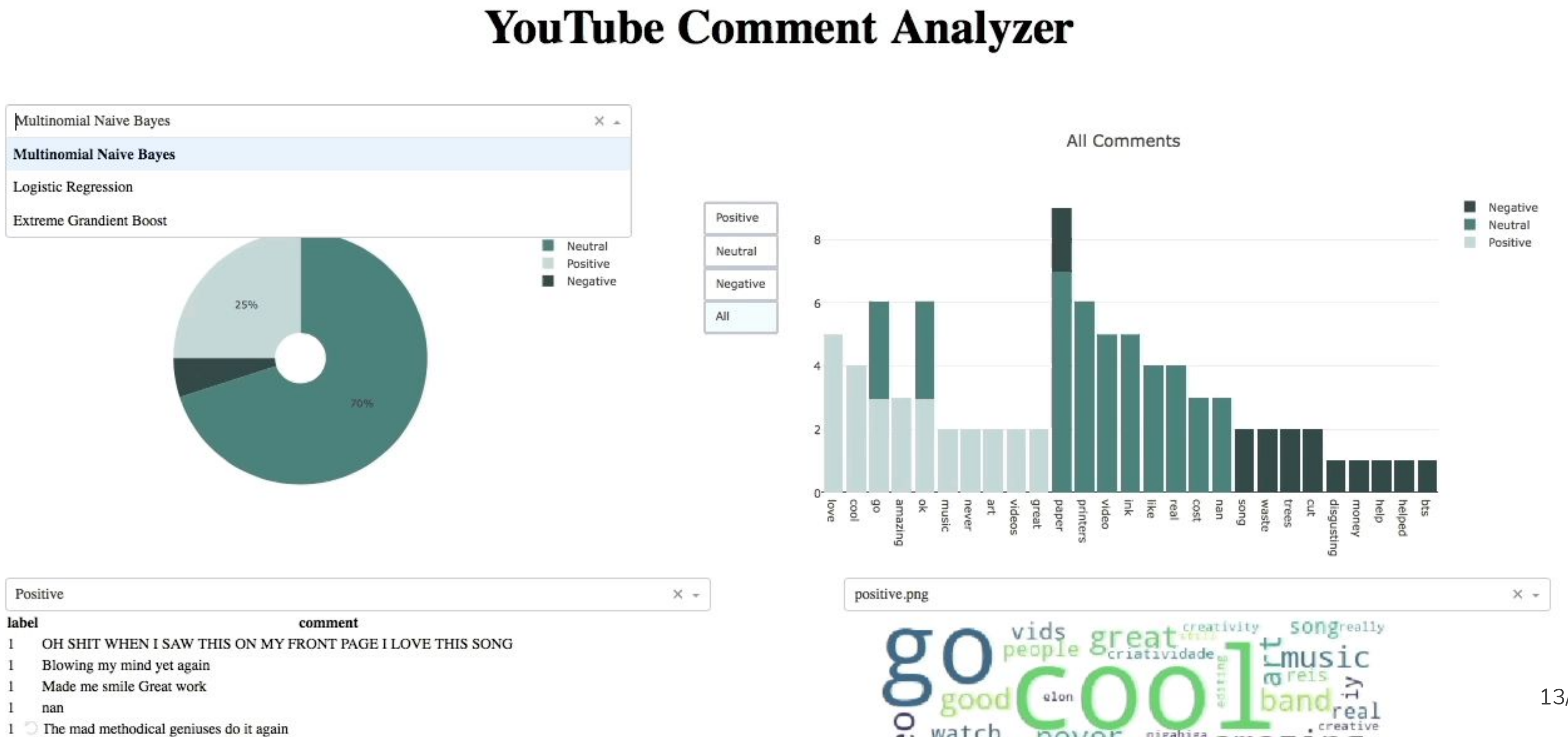
	MNB	LR	KNN	RF	GB
Accuracy	0.64	0.65	0.43	0.58	0.62

- Training on blog, twitter, & YouTube data and testing on YouTube Data:

	MNB	LR	KNN	RF	GB
Accuracy	0.58	0.53	0.45	0.57	0.56

Dashboard Demo

Using Dash to create interactive visualizations of results



Hurdles

Blockers in the project + troubleshooting

- Manually classifying data
- Comments in different languages
- Emojis, spam, and misspellings
- Sarcasm and long or mixed sentiment comments

Conclusions

Concluding remarks and recap of our findings

- Able to classify comment sentiment with about 65% accuracy
- Model performed better when training on just YouTube data
    - YouTube comments are a unique form of data and communication
- Models had a difficult time predicting negative comments
- Models had a relatively easier time predicting neutral comments
- Some videos had comments that were very content-specific
    - Our models performed worse on these types of videos

Future Work

Next steps in the project and NLP areas to look into

- Vader, another way to do sentiment analysis
- Compare video like-dislike ratio to ratio of comment sentiments
- Analyze comment sentiment for videos over time
  - Ex: Election Debates before and after controversial event

Credits

Giving thanks for support, involvement, and resources

- Data Science at UCSB
- Conor O’Brien
- Raul Eulogio (our troubleshooting guru)

Thanks for Listening

Appendix

What the data looks like and statistical models

Shon Inouye

Andie Donovan

Matthew Peterschmidt

Conclusions

Concluding remarks and recap of our findings

- Best Model for Training on YouTube Data
  - Multinomial Naive Bayes
    - Accuracy: **68%**
    - Precision: **0.62, 0.67, 0.75**
    - Recall: **0.48, 0.76, 0.71**

- Best Model for Training on Social Media and YouTube Data
  - Multinomial Logistic Regression
    - Accuracy: **67%**
    - Precision: **0.50, 0.58, 0.86**
    - Recall: **0.20, 0.84, 0.69**

Sources

Open source resources we used in this project

- YouTube/ Google API

- Two sources for outside datasets:
  - Sanders Analytics Twitter: https://github.com/zfz/twitter_corpus
  - Social Media Blogs: https://www.kaggle.com/c/si650winter11

- Python, NLTK, Scikit-Learn software, packages, and documentation