Content Analysis of YouTube Data

CS229 Autumn 2017 Category: Natural Language

Afshin Moin, Abhishek Bharani, Peng Seng Kuok
{afshinm, abharani, pkuok}@stanford.edu

Motivation

- Popularity of social media is increasing rapidly
- YouTube is the most popular video sharing website

We do:

- Sentiment analysis of the comments
- Category prediction based on comments
- Investigate the effect of dimensionality reduction

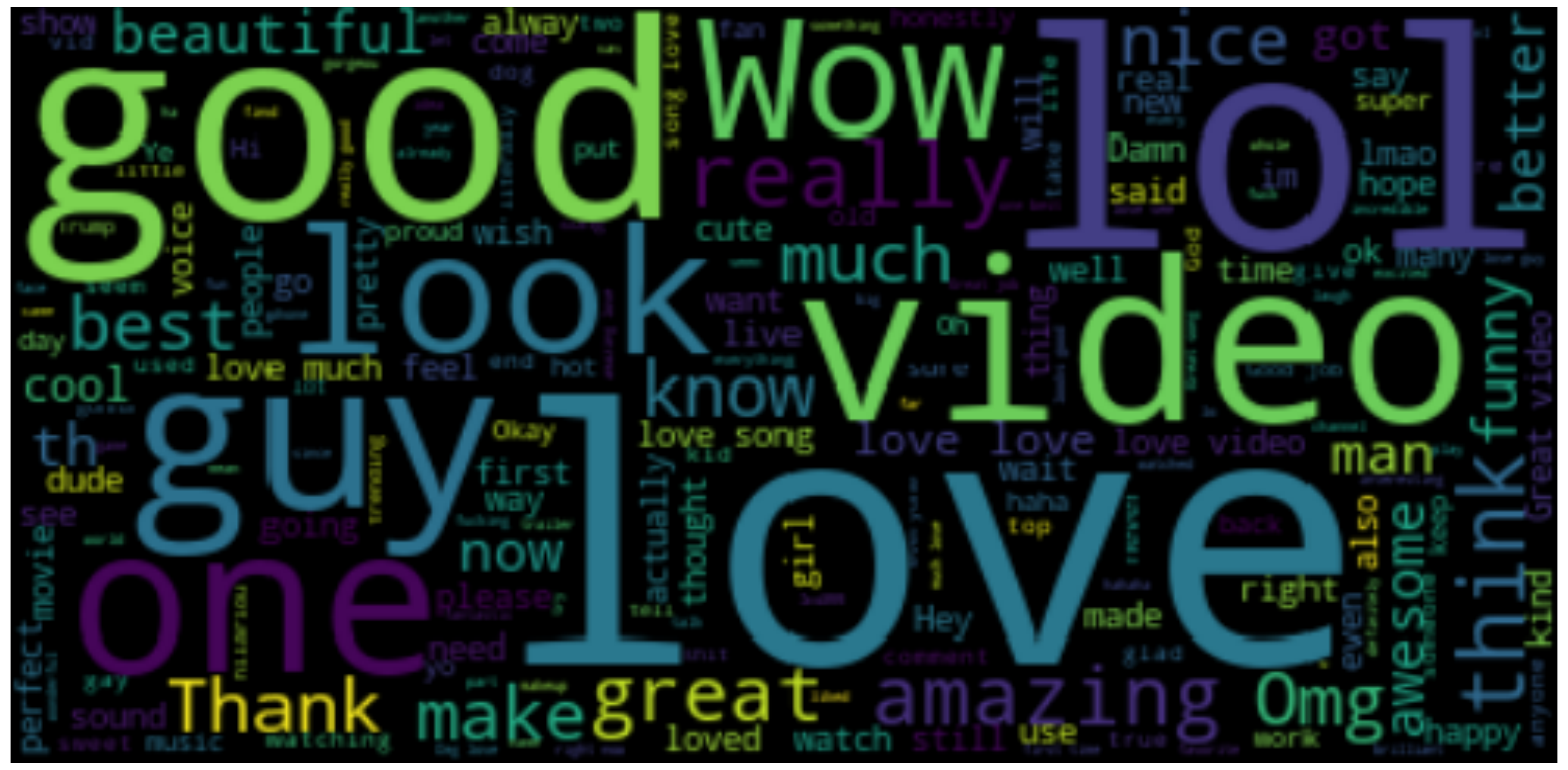
Data

- YouTube data acquired from Kaggle competition
- Top 200 videos per day along with their comments, category and tags
- 2354 videos
- 691408 comments

Feature Extraction

- Generate sentiment labels using TextBlob to estimate comment polarity
- Comments were vectorized using TF-IDF (Term Frequency – Inverse Document Frequency)

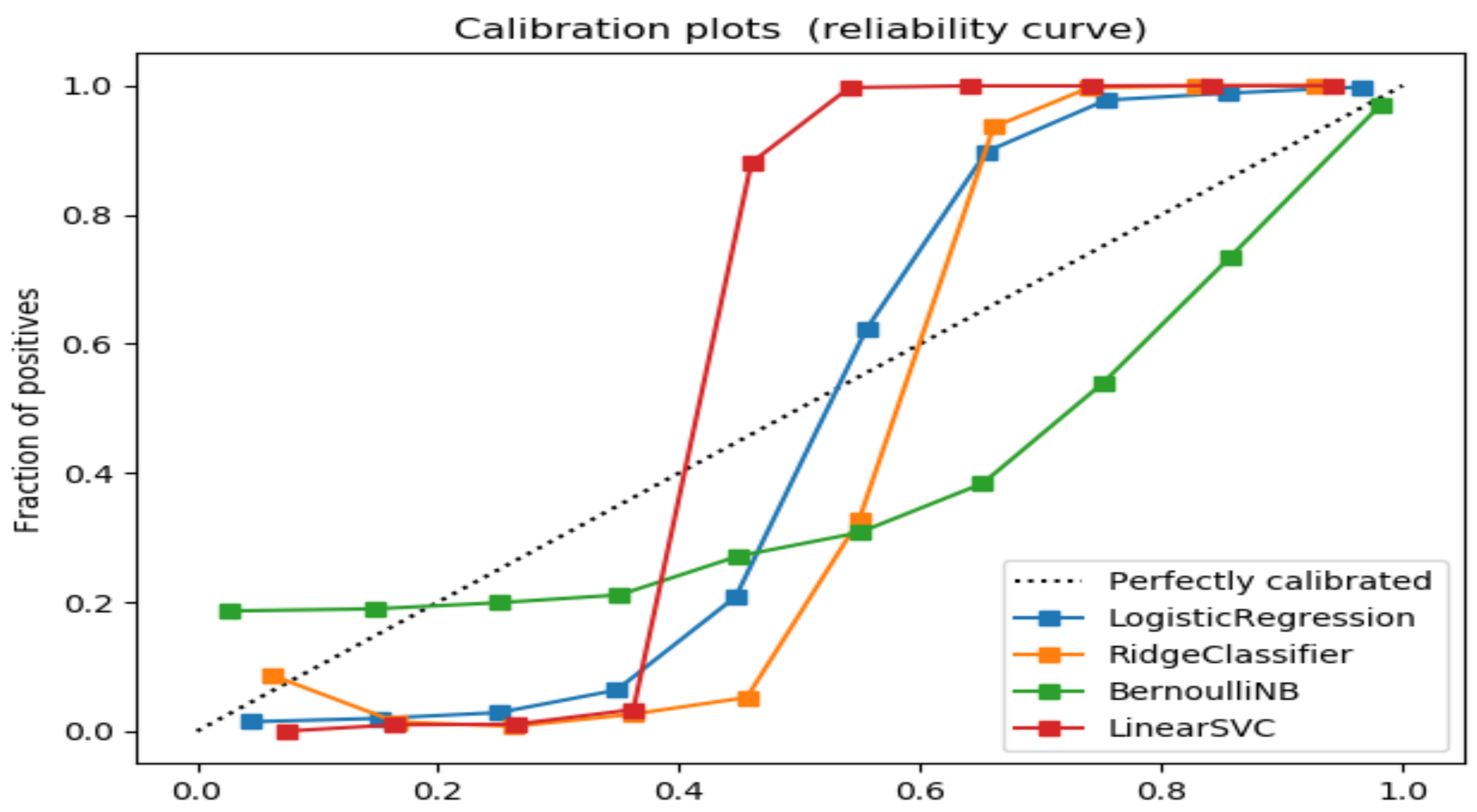
Common Words in Positive Comments



Sentiment Classification

Model	Train Accuracy	Test Accuracy
Logistic Reg.	95.64	94.47
Ridge Classifier	93.98	90.46
Bernoulli NB	72.15	71.53
Linear SVC	96.89	95.14

Calibration Curves

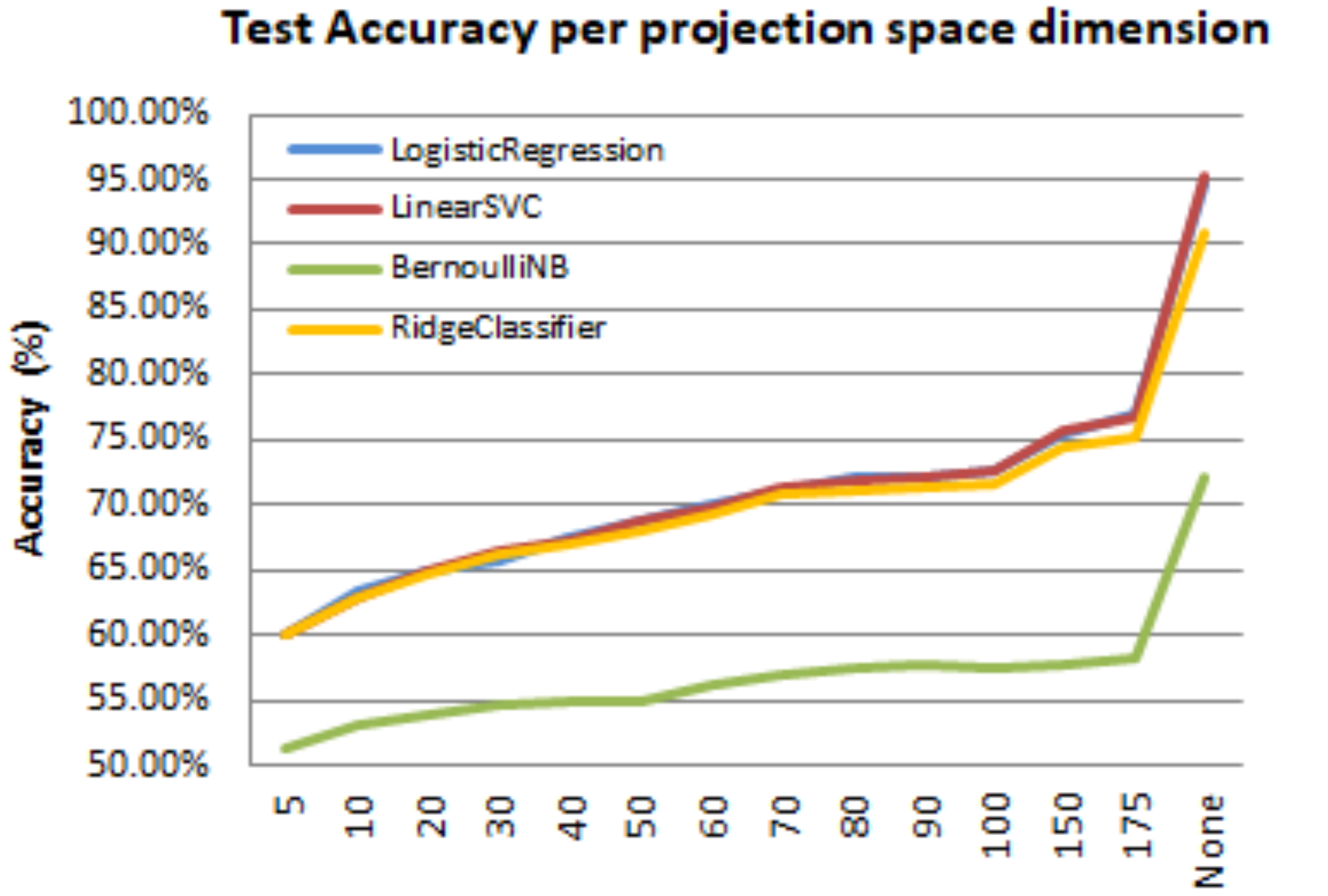


- Calibration curve is computed using two categories
- Naïve Bayes Model is not well-calibrated in [0,1]
- Logistic Regression is well-calibrated in [0,1]

Category Classification

Model	Train Accuracy	Test Accuracy
Logistic Reg.	60.70	53.21
Ridge Classifier	65.77	55.91
Bernoulli NB	45.53	41.61
Linear SVC	68.95	57.79

Effect of Dimensionality Reduction



- Truncated SVD was used to project TF-IDF feature vectors to lower dimensional spaces

Future Work

- Category classification based on tags
- Automatic tag generation given category
- Automatic comment generation given category, sentiment and comment length

References

- YouTube Database For Kaggle Competitions. https://www.kaggle.com/datasnaek/youtube
- TextBlob Library For Python. https://textblob.readthedocs.io/en/dev
- G. Toderici, H. Aradhye, M. Pasca, L. Sbaiz, and J. Yagnik. Finding meaning on youtube: Tag recommendation and category discovery. In Computer Vision and Pattern Recognition, 2010