Noonum Twitter Sentiment Analysis

Team member: Huong Thai Jared Lord Danti Li | Bill Howe | University of Washington - Information School

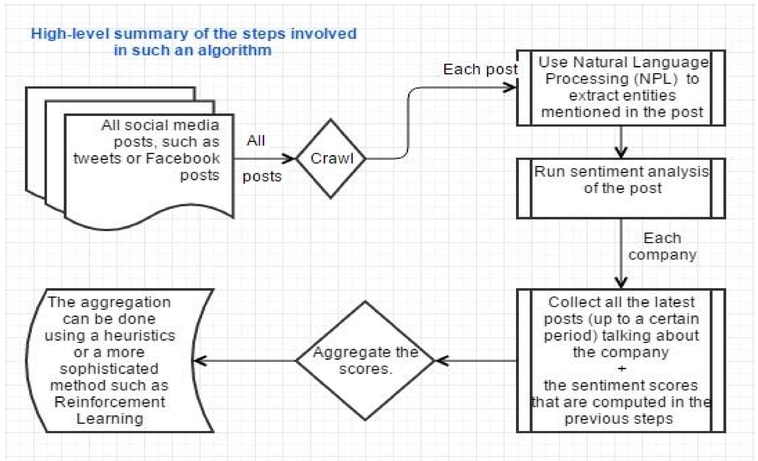
Research Question

To mine data from companies' tweets in Twitter, which one is the best method to build a tweet sentiment analysis model with limited supervised data?

Project Overview

One of the current hot topics in data analysis, both academic and industrial, is analyzing the relationship between stock price movement and general public sentiment of companies on social media. While some academics claim that the stock market is largely unpredictable, the past five years have had a proliferation of studies that show that social media can be used to predict stock motion ahead of time. This is a powerful tool for making informed decisions in the marketplace, and increase the return on investments. The techniques used to automatically acquire emotional and sentiment data from social media posts are thus a rapidly evolving field, and are coming under rapidly increasing demand (Bollen et al (2010) pegging sentiment data to the Dow Jones).

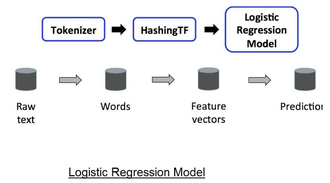
The circumstances led us to dial back and focus more on the particulars of technology to solve one of the current problems. because technology to predict stock movement from social media is highly complex.



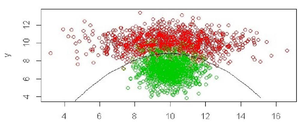
We've found that solving this sentiment analysis problem is itself a challenging problem, due to a lack of public labeled Twitter data, and the fact that Twitter data is highly noisy. That is, someone saying something good or bad about a company does not necessarily indicate that the mentioned company is doing good or bad. In this Capstone, we focus on the sentiment analysis aspect, experimenting the machine learning technologies.

Methodology

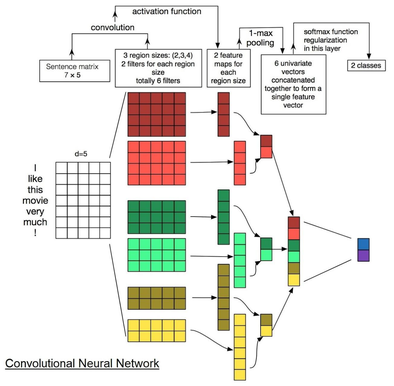
- Logit Regression & Naive Bayes. Tuning parameters with Grid search
- Convolutional Neural Network. Fine tuning with Transfer learning



Logistic Regression Model



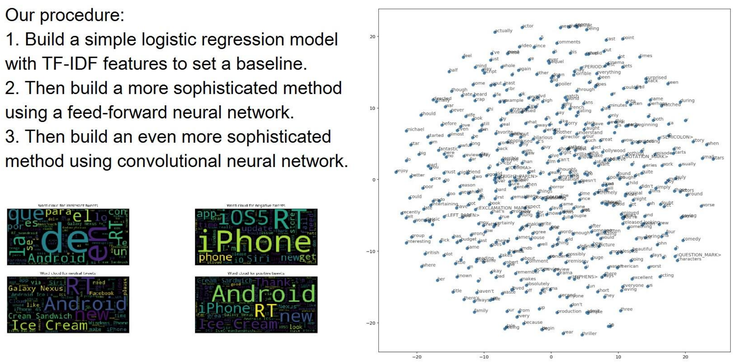
Naive Bayes classifer



Convolutional Neural Network

Data / Observations

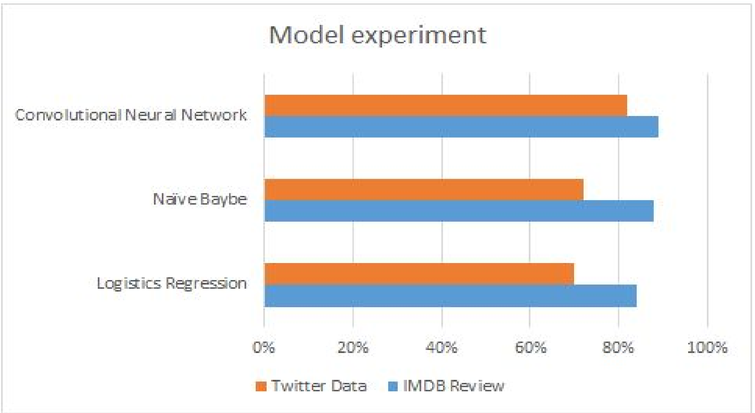
Our procedure:
1. Build a simple logistic regression model with TF-IDF features to set a baseline.
2. Then build a more sophisticated method using a feed-forward neural network.
3. Then build an even more sophisticated method using convolutional neural network.



The datasets we use to learn and train model : IMDB Movie Review and Twitter Sentiment from Sander group

- Data is labeled into positive, neutral, negative
- Processed data is perform to get clean and usable for analytic data.

Results



Conclusion

Logistics Regresstion and Naive Bayes are the popular models for sentiment analysis classification as for their efficiency

Convolutional Neural Network is newly applied for text classification and it shows a big improvement in accuracy comparing to previous models.

The models are trained better in IMDB dataset than Twitter dataset.

The CNN model accuracy is 82%, which is really good, considering that tweets are very different from movie reviews and that our training tweet data is small. This illustrates the power of transfer learning.

Although the model is not perfect, it is sufficiently useful in algorithmic trading. The reason is that traders can filter out the tweets which have more than 1 entity or the ones for which the prediction doesn't clearly indicate positive or negative. Given the huge volume of Twitter data, this filtered data is still big and traders can run the model to extract reliable and useful signals.

Future work: with additional amount of labeled Tweet sentiment data, we would expect that the results would be much better.

Works Cited

- Kim, Y. (2014). Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882.
- Yang, Q., Pan, S.J.: Transfer learning and applications. In: Intelligent Information Processing, p. 2 (2012)
- Johan Bollen,Huina Mao,,Xiao-Jun Zeng (2010) Twitter mood predicts the stock market. https://arxiv.org/pdf/1010.3003&embedded=true

Information School
UNIVERSITY of WASHINGTON