Twitter-based Stock Market Prediction

Erin Oerton, Richard Everett, Philipp Boeing

UCL

Abstract

We present a method to predict stock price movement based on Twitter sentiment. Our method incorporates three features – use of emoticons, use of emotional words, and use of hashtags – which are indicative of sentiment for each tweet. We investigate how this information can be used together with tweet volume and past stock data to predict future stock price change. We hypothesize that the utility of this method will vary by industry sector, and thus apply the method to three Dow Jones companies (McDonald's, Microsoft, and JPMorgan) as representative of the different sector types.

Companies

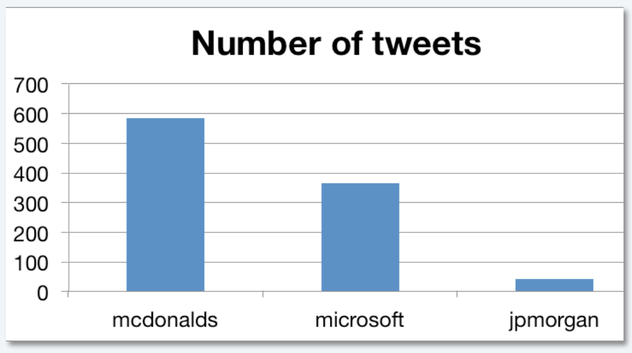
We chose to examine sentiment across tweets about **Microsoft, McDonalds,** and **JPMorgan**. These companies represent different ends of what we call the '**expertise spectrum**'. We hypothesize that the 30 companies in the Dow Jones dataset can be ordered according to the likelihood of tweets containing information that has any relation to the financial position of the company.

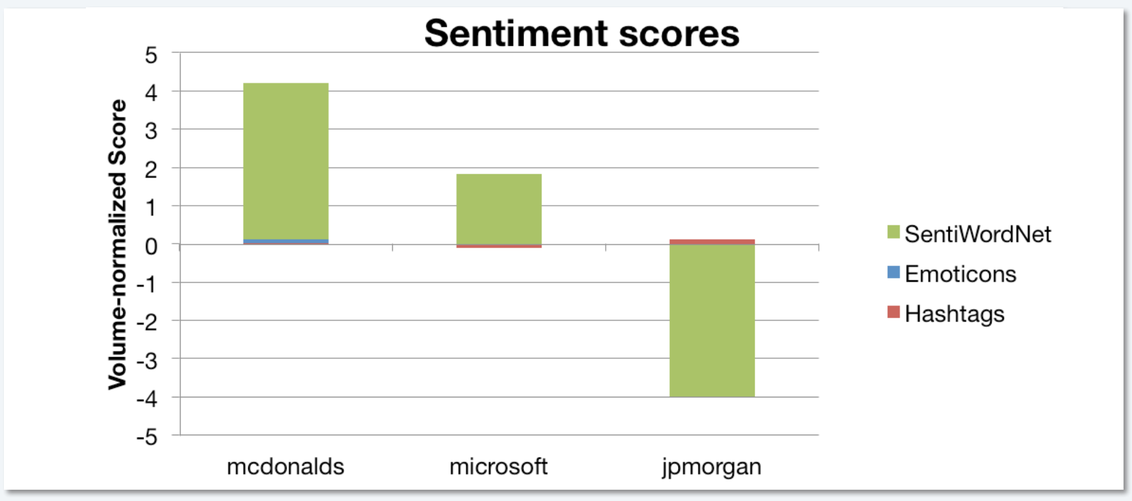
Current progress

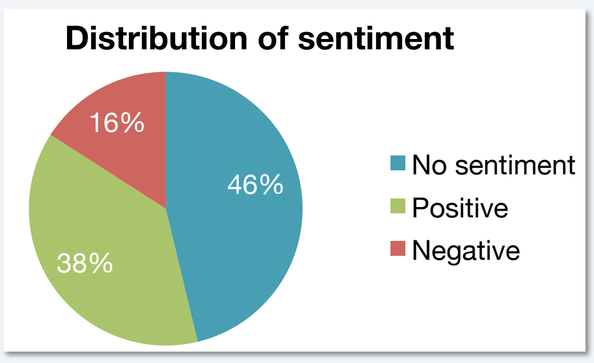
We have so far developed a pipeline to extract useful features from the tweets:

- the three sentiment metrics: **hashtags, emoticons**, and **SentiWordNet** scores
- **number of followers** for each tweet ('influence' of each tweet)
- **volume of tweets** about a particular brand ('influence' of each brand)

Representative scores were calculated over a subset of the data, which match our initial expectations about Twitter sentiments for each of our three brands. Scores were normalized to take into account volume of tweets for each brand.







In our preliminary investigation, we found that over half of the tweets in the dataset contained some form of sentiment according to our metrics. Of these, around 66% are positive, matching the expected distribution identified by Jansen et al[6].

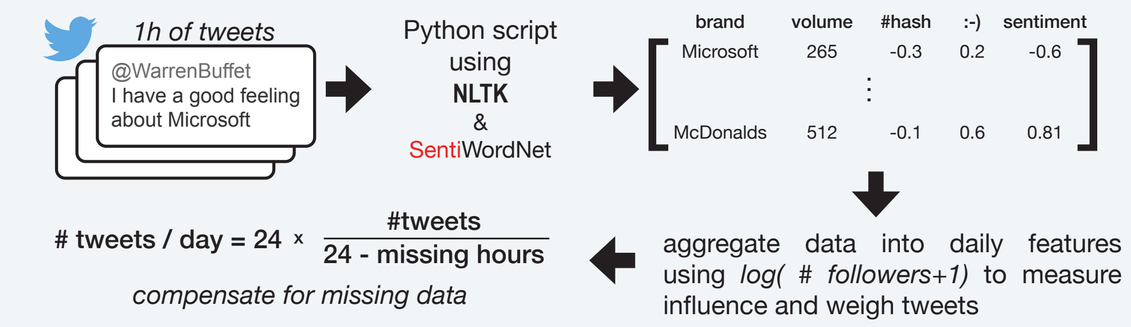
Method

Stage 1: Sentiment Extraction

Feature Extraction

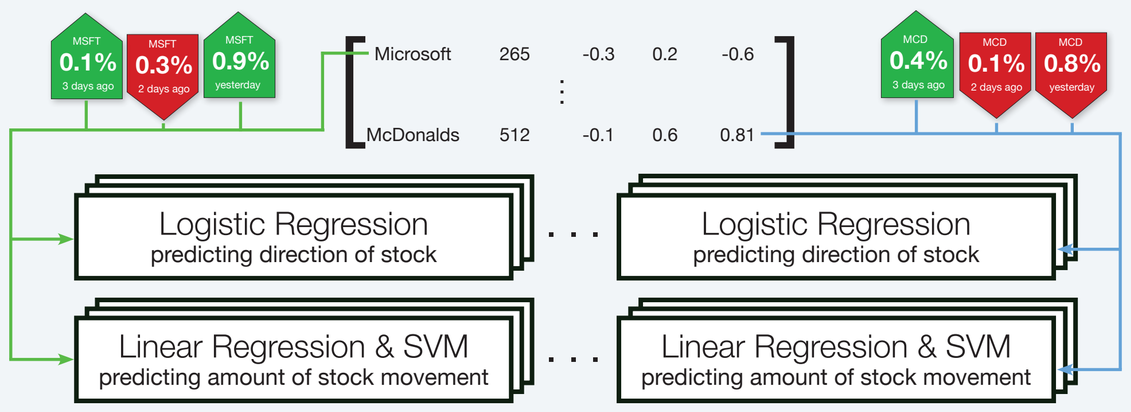
$$\text{Sentiment[Brand]} = \alpha_1 \text{LangSentiment} + \alpha_2 \text{Emoticon} + \alpha_3 \text{Hashtag}$$

Tweet Analysis and Aggregation

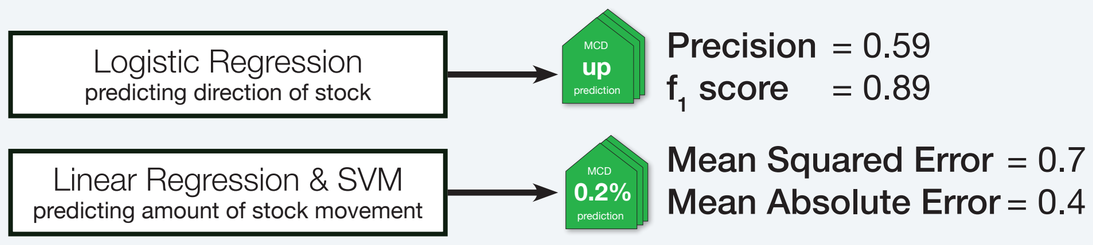


Stage 2: Training Stock Market Classifiers

For each brand we train a separate classifier, passing in the sentiment training vectors generated in Stage 1 supplemented with additional finance-based features.



Stage 3: Evaluation



Classifiers are trained by using **5-fold cross-validation**: 80% of days form training and validation sets, and the remaining days form an independent test set to calculate the prediction error.

Related work

Traditional sentiment mining techniques do not work well in the micro-blogging domain. There are a number of reasons for this: the use of **slang, misspelled words**, or the absence of the traditional **grammatical features** which tend to indicate subjective text[1]. Other authors have identified Twitter-specific features such as the use of **emoticons**[2] or **hashtags**[3] which are more useful indicators of subjectivity in this domain, which we make use of along with the popular sentiment analysis tool **SentiWordNet**[4]. Another feature was identified by Tumasjan et al.[5], whose examination of the use of Twitter to predict political sentiment identified a correlation between **daily volume of tweets** and electoral result. Our system applies this feature to the stock market domain.

References

1. Khuc, VN; Shivade, C; Ramnath, R; Ramanathan, J (2012) Towards Building Large-Scale Distributed Systems For Twitter Sentiment Analysis . In: Proceedings of the 27th Annual ACM Symposium on Applied Computing
2. Pak, A; Paroubek, P (2010) Twitter As A Corpus For Sentiment Analysis And Opinion Mining. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation
3. Kouloumpis, E; Wilson, T; Moore, J (2011) Twitter Sentiment Analysis: The Good, The Bad, and the OMG! In: Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media
4. Baccianella, S; Esuli, A; Sebastiani, F (2011) SentiWordNet 3.0: An Enhanced Lexical Resource For Sentiment Analysis And Opinion Mining. In: Seventh conference on International Language Resources and Evaluation
5. Tumasjan, A; Sprenger, TO; Sandner, PG; Welpe, IM (2010) Predicting Elections With Twitter: What 140 Characters Reveal About Political Sentiment. In: Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media
6. Jansen, BJ; Zhang, M; Sobel, K; Chowdhury, A (2009) Twitter power: Tweets as electronic word of mouth. In: Journal of the American Society for Information Science and Technology