News Analytics: Categorization and Sentiment Extraction from Financial News

by

Muktamala Chakrabarti

Thesis Advisory Committee

Prof Asim K. Pal (Joint Thesis Advisor)

Prof Ashok Banerjee (Joint Thesis Advisor)

Prof Partha Sarathi Dasgupta

Abstract

With the popularity of electronic and social media, there are innumerable news stories available across the web. Faster analysis and summarization of these news stories can help in taking business decisions. Financial analysts need to go through news stories about companies to take trading decisions. Companies also analyze millions of tweets or blogs of general people to understand customer sentiment. News analytics is a broad area which includes various text mining methods to analyze news stories which can be used for further applications. It includes methods from Machine Learning, Natural Language Processing, Information Retrieval etc. Basically news analytics can help in finding out some of the qualitative and quantitative attributes from the news stories. The qualitative

quantitative attributes can be the sentiment score which can describe the tone (positive / negative) of the story.

In the present thesis we have looked into two problems in the domain of news analytics; one is text categorization and the other is sentiment analysis. For text categorization we have clustered the news stories into categories in two ways: unsupervised learning which does not use any domain knowledge (i.e. expert knowledge) and semi-supervised learning where partial domain knowledge has been used to direct the clustering. The former one uses clustering (i.e. unconstrained clustering) and the latter one uses constrained clustering methods. The sentiment analysis (also referred as

overall tone, positive or negative, of the story.

The first three chapters after Introduction and Literature Survey, Chapters 3 through 5, of this thesis focus mainly on th

specifically twitter. We extract the sentiment or tone of the tweets and measure the correlation of the sentiment scores extracted from twitter messages with return, volatility and liquidity. Part of Chapter 7

Papers -

Chakrabarti, M., & Pal, A. K. (2014). Competitive Learning with Pairwise Constraints for Text. In *PRICAI 2014: Trends in Artificial Intelligence* (pp. 370-382). Springer International Publishing.

Chakrabarti, M., & Pal, A. K. (2014). Constrained Latent Dirichlet Allocation. Working paper

Chakrabarti, M., Pal, A. K. and Banerjee, A.(2014). Twitter and Financial Markets. Accepted for publication at WEI International Academic Conference