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 attributes can be tags like “mergers and acquisitions”, “IPO” for financial news. The 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 sentiment extraction) would give a ‘sentiment score’ to each story depending on the 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 the problem of news clustering. In Chapter 3 we use some of the existing text clustering methods. In Chapter 4 and Chapter 5 domain knowledge has been incorporated into the unsupervised methods in order to provide a direction to the clustering process for improved performance. In Chapter 4 we have proposed a competitive learning method suitable for text data where we have introduced the concept of spherical k-means in place of k-means for constrained text clustering. In this chapter it has been found that although k-means is not suitable for text clustering, it performs quite well when the concept of rival penalized competitive learning is incorporated into it. However the performances of spherical k-means and k-means for text data come closer when the concept of rival penalized competitive learning is included into both of them. Part of Chapter 4 has been published as Chakrabarti and Pal (2014).

The next chapter focuses on topic modeling method Latent Dirichlet Allocation (LDA). Here we have tried to improve the performance of LDA by adding domain knowledge in the form of constraints. In this chapter we propose two modifications over an existing constrained LDA method. The improvement suggested is in two ways- incorporation of non-linearity in the change in importance of constraints, and using the reinforcement learning concept to both encourage satisfaction of constraints and penalize violation of constraints.

The next two chapters of this thesis mainly focus on the applications where news can help in taking financial decisions. In Chapter 6 we focus on sentiment extraction from financial news and its applications. Finally in Chapter 7 we look into social media,

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 has been accepted for publication as "Twitter and Financial Markets" in WEI International Academic Conference in Athens, 2015.

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