International Journal of Business and Management Study
Volume 3 : Issue 2 *[ISSN 2372-3955]*
Publication Date : 31 August, 2016

Leveraging Big Data to Predict Firms' Performance

Otto K.M. Cheng and Raymond Lau

***Abstract*—Though some research studies about big data have been reported in literature recently, few studies about applying big data analytics to analyze firms' performance are performed and reported in existing business literature. Empowered by a novel framework of big data analytics, this paper illustrates our preliminary work of apply the proposed big data analytics framework to predict firms' performance in stock market. In particular, the proposed framework leverages big data and probabilistic language modeling to analyze the stock performance of firms in near real-time. Our empirical test demonstrates the merits of the proposed framework. The practical implication of our research work is that firms can apply our big data analytics framework to analyze their internal and external performance, and hence develop more effective business strategies in advance.**

***Keywords*—Big Data, Big Data Analytics, Knowledge Management, Firm Performance**

I. Introduction

In the era of the Social Web, user-contributed contents have become the norm. The amounts of data produced by individuals, business, government, and research agents have been undergoing an explosive growth – a phenomenon known as the data deluge. For individual social networking, many online social networking sites have between 100 and 500 million users. By the end of 2013, Facebook and Twitter had 1.23 and 0.64 billion active users, respectively. The number of friendship edg-es of Facebook is estimated to be over 100 billion. The stream of huge amounts of user-contributed contents such as online consumer reviews, online news, personal dialogs, search queries, and so on have called for the research and development of a new generation of ana-lytics methods and tools to effectively process them, preferably in real-time or near real-time. Big data is often characterized by three dimensions, named the 3 V's: Volume, Velocity, and Variety [1]. Currently, there are two common approaches to deal with big data, namely batch-mode big data analytics and streaming-based big data analytics.

Otto K.M. Cheng
Department of Information Systems, College of Business, City University of Hong Kong
Hong Kong SAR

Raymond Lau
Department of Information Systems, College of Business, City University of Hong Kong
Hong Kong SAR

Most data originally produced from the Social Web is streaming data. For example, the data representing actions and interactions among individuals in online social media, or the data denoting some events captured by sensor networks is the typical kind of streaming data. Other types of big data perhaps are just a snapshot view of the streaming data generated from a specific point of time. For big data streams, data arrives at high speed, and the methods of big data stream analytics must process it in one pass and under very strict constraints of space and time. Currently, algorithms of big data analytics often deal with big data in batch model, while algorithms designed to process big data stream in real-time or near real-time are rare.

In general, big data analytics can be classified as distributed or single host approaches. For distributed methods, there are batch mode and streaming mode of processing. Although batch mode big data analytics approach (e.g., MapReduce) is the dominated method to date, online incremental algorithms that can effectively process evolving data stream are desirable to address both the "volume" and the "velocity" issue of big data on the Web. MapReduce and big data stream analytics are two fundamentally different programming paradigms though they are related from a theoretical perspective. Recently, there are some attempts to incorporate streaming and incremental computation on top of MapReduce such as the Hadoop Online Prototype. However, more research should be conducted toward big data stream analytics. The main contribution of this paper is the design and development of a novel big data analytics framework that provides the essential infrastructure to operationalize a business network based inference model incorporating evolving big data stream from online social media for correlated firm performance (e.g., stock performance) prediction in near real-time.

II. Literature Review

With the rapid growth of the Social Web, increasingly more Web users have posted and extracted viewpoints about products, people, or political issues via a variety of online social media such as Blogs, forums, chat-rooms, and social networks. The big volume of user-contributed contents opens the door for automated extraction and analysis of the sentiments or emotions referring to the underlying entities such as consumer products. Sentiment analysis is also referred to as opinion analysis, subjectivity analysis, or opinion mining [2,3]. Sentiment analysis aims to extract subjective feelings about some subjects rather than simply extracting the objective facets about these subjects [4]. Analyzing the sentiments of messages posted to social networks or online forums can generate countless business values for the organizations which aim to extract timely business intelligence about how their products or services are perceived by their customers [5]. Other possible applications of sentiment analysis include the analysis of the propaganda and activities of cybercriminal groups who pose

serious threats to business or government owned web sites [2]. In addition, sentiment analysis is often applied to predict stock price movements of firms.

Sentiment analysis can be applied to a phrase, a sentence, or an entire message [4]. Most of the existing sentiment analysis methods can be divided into two main camps. The first common paradigm utilizes a sentiment lexicon or heuristic rules as the knowledge base to locate opinionated expressions and predict the polarity of these opinioned expressions [3,6]. The second common approach of sentiment analysis is based on statistical learning methods [4]. Nevertheless, each camp has its own limitations. For instance, for the lexicon-based methods, common sentiment lexicons may not be able to detect the context-sensitive nature of opinion expressions. For example, while the term "small" may have a negative polarity in a hotel review that refers to a "small" hotel room, the same term could have a positive polarity such as "a small and handy notebook" in consumer reviews about computers. In fact, the token "small" is defined as a negative opinion word in the well-known OpinionFinder sentiment lexicon.

In contrast, statistical learning techniques such as super-vised machine learning method usually requires a large number of labeled training cases in order to build an effective classifier to identify the polarity of opinionated expressions. Unfortunately, it is not practical to assume the availability of a large number of human labeled training examples, particularly in a big data environment. On the other hand, both approaches may not be scale up to analyze a huge number of opinioned expressions as found in nowadays Social Web. There is an obvious research gap to develop new methods to be able to analyze big social media data in real-time or near real-time by leveraging a parallel and distributed system architecture. Our research work reported in this paper just tries to fill such a research gap. The business implication of our research is that firms can apply the proposed big data analytics framework to more effectively and promptly analyze/predict firms' performance (e.g., stock performance) in near real-time.

III. The Big Data Analytics Framework

An overview of the proposed framework that leverages Big Data Stream Analytics for online Sentiment Analysis (BDSASA) is describe in this section. The BDSASA frame-work consists of seven layers, namely data stream layer, data pre-processing layer, data mining layer, prediction layer, learning and adaptation layer, presentation layer, and storage layer. For these layers, we will apply sophisticated and state-of-the-art techniques for rapid service prototyping. For instance, Storm, the open-source Distributed Data Stream Engine (DDSE) for big data is applied to process streaming data fed from dedicated APIs and crawlers at the Data Stream Layer. For instance, the Twitter API is used to retrieve public comments about firms from Twitter.

The Storage Layer leverages Apache HBase and HDFS for real-time storage and retrieval of big volume of comments discussing products and services of targeting firms. The Stanford Dependency Parser and the GATE NER module [7] are applied to build the Data Pre-processing Layer. Our pilot tests show that the size of the multilingual social media data streams is within the range between 0.2 and 0.4 Gigabytes on a daily basis, and this volume is steadily growing. For the feature extraction layer, the Affect Miner utilizes a novel community-based affect intensity measure to predict consumers' moods towards products. Among the big six classes i.e., anger, fear, happiness, sadness, surprise, and neutral commonly used in affect analysis, we focus on the anger, fear, sadness, and happiness classes relevant for product sentiment analysis. The WordNet-Affect lexicon [8] extended by a statistical learning method is used by the Affect Miner. Since social media messages are generally noisy, one novelty of our framework is that we reduce the noise of the "affect intensity" measure by processing messages really related to consumers' comments about products or services.

Previous research employed the HMM method to mine the latent "intents" of actors [9]. We exploit a novel and more sophisticated online generative model and the corresponding distributed Gibbs sampling algorithm to build our Latent Intent Extractor that predicts the intents of consumers for firms' products or services. The Sentiment Extractor utilizes well-known sentiment lexicons such as OpinionFinder to extract the sentiment words embedded in public comments about firms. Finally, overall sentiment polarity prediction for public comments is performed based on a novel inferential language modeling method. The computational details of this inferential language modeling method for context-sensitive sentiment analysis will be explained in the next section. The overall sentiment polarity against a firm is communicated to the user of the system via the presentation layer. Different modes of presentations (e.g., text, graphics, multimedia on desktops or mobile devices) are supported by our framework.

In addition, a novel parallel co-evolutionary genetic algorithm (PCGA) is designed so that the proposed prediction model is equipped with a learning and adaptation mechanism that continuously tunes the whole service with respect to possibly changing features of the problem domain. The PCGA can divide a large search space into some subspaces for a parallel and diversified search, which improves both the efficiency and the effective-ness of the heuristic search process. Each subspace (i.e., a sub-population) is hosted by a separate cluster. Three fundamental decisions are involved for the design a genetic algorithm (GA), that is, a fitness function, chromo-some encoding, and a procedure that drives the evolution process of chromosomes [10]. First, the fitness function of our PCGA is developed based on a performance metric (e.g., accuracy of sentiment polarity prediction). Second, since various components of the proposed service should be continuously refined, there are multiple sub-populations of chromosomes to be encoded and co-evolved simultaneously. During each evolution cycle, the best chromosome of a sub-population (e.g., prediction features, social media sources, system parameters) is exchanged with that of other sub-populations. Armed with all the essential information, each chromosome of a sub-population represents a feasible prediction, and its fitness can be assessed accordingly.

IV. Probabilistic Language Modeling for Sentiment Analysis

Originally, the term "language model" has been widely explored in the speech recognition community, and it refers to a probability distribution which represents the statistical regularities for the generation of a language [11]. In other words, a language model is a probabilistic function that assigns a probability mass to a string drawn from some vocabulary. In the context of Information Retrieval (IR), a language model is used to estimate the probability that a document generates a query [12]. In particular, such a probabilistic inference is used to mimic the concept of document "relevance" of respect to . The basic unigram language model is defined according to the following formulas [12,13]:

$$P(q|d) \propto P(q|M_d) = \prod_{t \in q} P(t|M_d) \quad (1)$$

$$P(t|M_d) = (1-\lambda)P_{ML}(t|M_d) + \lambda P_{ML}(t|M_D) \quad (2)$$

$$P_{ML}(t|M_d) = \frac{tf(t,d)}{|d|} \quad (3)$$

where M_d is the language model of the document d. With Jelinek-Mercer smoothing [13], the probability of the document generating a query term t (i.e., $P(t|M_d)$) is estimated according to the maximum likelihood model $P_{ML}(t|M_d)$, and the maximum likelihood model of the entire collection $P_{ML}(t|M_D)$. λ is the Jelinek-Mercer smoothing parameter [13]. The smoothing process is used to alleviate the problem of over-estimating the probabilities for query terms found in a document and the problem of under-estimating the probabilities for terms not found in the document. The function $tf(t,d)$ returns the term frequency of term t in the document d, and $|d|$ is the document length measured by the number of tokens contained in the document.

However, previous studies found that applying the probabilities of query related terms of a relevant context instead of the probabilities of the individual query terms estimated based on the entire document collection (i.e., a general product review context) to a document language model will lead to a more effective smoothing process, and hence lead to good IR performance [14]. Following the similar kind of idea, we develop an inferential language model to compute the probability that a document d (e.g., a product review) will generate a term t found in a Sentiment Lexicon (SL). In order to ensure a more robust and effective smoothing process, the inferential language model can take into account terms (opinion evidences) associated with the opinion indicators in a relevant online review context. In particular, the associated opinion evidences are discovered based on the context-sensitive text mining process over an online review context. The inferential language model for context-sensitive opinion scoring is then defined as follows.

$$P(SL|d) \propto P(SL|M_d) = \prod_{t \in SL} P(t|M_d) \quad (4)$$

$$P(t|M_d) = (1-\lambda)P_{ML}(t|M_d) + \lambda P_{INF}(t|M_d) \quad (5)$$

$$P_{INF}(t|M_d) = \tanh\Big(\sum_{(t \to t')\in OE} P(t \to t') \cdot P_{ML}(t'|M_d)\Big) \quad (6)$$

where $P(SL|d)$ is the document language model for estimating the probabilities that the document d will generates the opinion indicators defined in a sentiment lexicon (SL). However, to address the common problem that sentiment lexicons may not capture all possible sentiments of a problem domain (e.g., context-sensitive opinion evidences are missing), the proposed language model can take into account other opinion evidences contained in the document by means of the inferential language model $P_{INF}(t|M_d)$. The set of context-sensitive opinion evidences OE is dynamically generated according to a context-sensitive text mining technique.

The term association (term inference) of the form $t \to t'$ is applied to the inferential language model to compute the probability that a document generates a term (e.g., an opinion indicator) which is contextually associated with another opinion indicator captured in a sentiment lexicon [15]. For easy of implementation, we only include the top χ term associations captured in OE for each opinion indicator t. It should be noted that the inference that d generating t' involves a certain degree of uncertainty. As a result, the maximum likelihood estimation of $P_{ML}(t'|M_d)$ is moderated by a factor $P(t \to t')$. The hyperbolic tangent function is applied to moderate the probability function $P_{INF}(t|M_d)$ such that its values fall in the unit interval.

V. Experiment and Results

To examine the effectiveness of the proposed Adjusted Influence Model, firms in 4 different sectors, Information Technology (IT), Utilities (UT), Financials (FN) and Health Care (HC), mainly from S&P500 are selected as research targets. Five years stock performance data , from Jan, 2008 to Dec, 2012 totally 292,880 records, are collected from yahoo finance. In our experiment, the public sentiments of targeting firms are extracted via Twitter in the corresponding periods. Based on the proposed big data analytics framework, we tried to predict firms' performance (e.g., their stock price movements). In addition, we also employed a Support Vector Machine (SVM)-based prediction model as a baseline to predict firms' performance. By comparing a system predicted firm performance with the actual firm performance collected from Yahoo Finance, we could calculate the system's prediction accuracy, recall, precision, and F-score. A higher

accuracy (F-score) indicates a better prediction performance. Our experimental results are depicted in Table I.

TABLE I. EXPERIMENTAL RESULTS

Table 1. Result of each model					
		IT	UT	FN	HC
Big Data	Precision	0.671	0.539	0.614	0.692
	Recall	0.704	0.563	0.615	0.633
	F-measure	0.704	0.581	0.612	0.659
	Accuracy	0.685	0.593	0.617	0.662
SVM	Precision	0.524	0.501	0.543	0.577
	Recall	0.571	0.489	0.542	0.544
	F-measure	0.549	0.493	0.542	0.560
	Accuracy	0.533	0.483	0.563	0.581

From Table I, we can see that about 65% average accuracy was obtained by the proposed big data analytics framework. It is more than 10 percent improvements than using a classical machine learning prediction model (SVM). Hence, the effectiveness of our novel big data analytics framework for firm performance prediction was verified, which in turn proved the correctness of our assumption that public's sentiments about firms are possible predictors for their business performance such as stock performance movements.

VI. Conclusions

While some research work has been devoted to big data recently, very few studies about applying big data analytics to predict firms' performance are reported in existing literature. The main theoretical contributions of our research include the design and development of a novel big data analytics framework for firms' performance prediction (e.g., directional stock price movement prediction). Another main contribution of this paper is the illustration of a probabilistic language modeling based sentiment inference method, which can infer the sentiments of firms induced from evolving big data stream generated by online social media such as Twitter. The business implication of our research is that business managers and financial analysts can apply our proposed framework to more effectively analyze and predict the business performance of targeted firms based on dynamic sentiments mined from big data stream. Accordingly, they can take proactive business strategies to streamline the operations of these firms.

One limitation of our current work is that the proposed framework has not been tested under a large empirical setting. We will devote our future effort to further evaluate the effectiveness and efficiency of the proposed framework based on real-world financial data and social media messages collected from multiple sources. On the other hand, we will continue to refine the proposed probabilistic language model for sentiment analysis. Moreover, we will incorporate firms' relationships into our prediction model. By way of illustration, Samsung is the largest rival of Apple Inc. for smartphones and tablet PCs. At the same time, Samsung is the largest supplier of Apple Inc. for display and battery devices. Our future work will incorporate business relationships into the prediction model by capturing entity-based (e.g., products, events, or locations) relationships among firms. Finally, the efficiency and scalability of the proposed framework will be under scrutiny.

Acknowledgment

This research work was partially supported by a grant from the Shenzhen Municipal Science and Technology R&D Funding—Basic Research Program (project: JCYJ20140419115614350).

References

[1] Gupta, R., Gupta, H., Mohania, M. 2012. "Cloud Computing and Big Data Analytics: What Is New from Databases Perspective?" in Big Data Analytics, Volume 7678 Lecture Notes in Computer Science, S. Srinivasa and V. Bhatnagar (eds.), New York: Springer-Verlag, pp. 42-61.

[2] Abbasi, A., Chen, H., and Salem, A. 2008. "Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums," ACM Transactions on Information Systems, 26(3).

[3] Turney, P. and Littman, M. 2003. "Measuring praise and criticism: Inference of semantic orientation from association," ACM Transactions on Information Systems, 21(4):315–346.

[4] Wilson, T., Wiebe, J. and Hwa, R. 2004. "Just how mad are you? finding strong and weak opinion clauses," in Proceedings of the Nineteenth National Conference on Artificial Intelligence, Sixteenth Conference on Innovative Applications of Artificial Intelligence, pp. 761–769.

[5] Lau, R.Y.K., Li, C., Liao, S. 2014. "Social Analytics: Learning Fuzzy Product Ontologies for Aspect-Oriented Sentiment Analysis," Decision Support Systems, (65), pp.80-94.

[6] Wilson, T., Wiebe, J. and Hoffmann, P. 2005. "Recognizing contextual polarity in phrase-level sentiment analysis." in Proceedings of the Human Langugae Technology Conference and Conference on Empirical Methods in Natural Language Processing, pp. 347–354.

[7] Maynard, D., Tablan, V., Ursu, C., Cunningham, H., and Wilks, Y. 2001. "Named Entity Recognition from Diverse Text Types," in Proceedings of the 2001 Conference on Recent Advances in Natural Language Processing, Tzigov Chark, Bulgaria.

[8] Valitutti, A., Strapparava, C., and Stock, O. 2004. "Developing Affective Lexical Resources," Psychology (2:1), pp. 61-83.

[9] Zhang, Q., Man, D., Wu, Y. 2009. "Using HMM for Intent Recognition in Cyber Security Situation Awareness," in Proceedings of the Second IEEE International Symposium on Knowledge Acquisition and Modeling, pp. 166-169.

[10] Lau, R.Y.K., Tang, M., Wong, O., Milliner, S., Chen, Y. 2006. An Evolutionary learning approach for adaptive negotiation agents. International Journal of Intelligent Systems 21(1), pp.41-72.

[11] Nadas, A. 1984. "Estimation of probabilities in the language model of the IBM speech recognition system," IEEE Transactions on Acoustics, Speech and Signal Processing, ASSP-32(4):859.

[12] Ponte, J. and Croft, B. 1998. "A language modeling approach to information retrieval," in Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 275–281.

[13] Zhai, C. and Lafferty, J. 2004. "A study of smoothing methods for language models applied to information retrieval," ACM Transactions on Information Systems, 22(2):179–214.

[14] Nie, J., Cao, G., and Bai, J. 2006. "Inferential language models for information retrieval," ACM Transactions on Asian Language Information Processing, 5(4):296–322.

[15] Lau, R.Y.K. 2003. "Context-Sensitive Text Mining and Belief Revision for Intelligent Information Retrieval on the Web," Web Intelligence and Agent Systems An International Journal, 1(3-4):1–22.