Big Data Stream Analytics for Real Time Sentimentality Analysis

Dr.M.Murugesan

Professor in Computer Science&Engineering, Anurag Engineering College (Autonomous),

Ananthagiri (V&M), Suryapet (Dt) ,Telangana (TS)-508206.

murugeshvim@gmail.com

Abstract

Big Data constituting from the information shared in the various social network sites have great relevance for research to be applied in diverse fields like marketing, politics, health or disaster management. Social network sites like Facebook and Twitter are now extensively used for conducting business, marketing products and services and collecting opinions and feedbacks regarding the same. Since data gathered from these sites regarding a product/brand are up-to-date and are mostly supplied voluntarily, it tends to be more realistic, massive and reflects the general public opinion. Its analysis on real time can lead to accurate insights and responding to the results sooner is undoubtedly advantageous than responding later. The main contribution of this paper is the illustration of the development of a novel big data stream analytics context named BDSAC that leverages a probabilistic language model to analyze the consumer sentiments embedded in hundreds of millions of online consumer reviews. In particular, an inference model is embedded into the classical language modeling framework to enhance the prediction of consumer sentiments. The practical implication of our research work is that organizations can apply our big data stream analytics framework to analyze consumers' product preferences, and hence develop more effective marketing and production strategies.

Keywords Big Data, Data Stream Analytics, Sentiment Analysis, Online Review

1.Introduction

Social network sites have become a prominent platform to express opinions and feedbacks. With widespread use of smartphones and ever growing popularity of social network sites, most people own share their sentiments and experience about any new market product almost instantly in the social networks and these posts have great influence in the buying patterns of prospective customers. A model for knowledge transfer from social networks to predict human behavior is given in which can be applied in social marketing. 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 edges 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 analytics 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 7 V's: Volume, Velocity, Variety, Variability, Veracity, Value, and Visibility [1]. Currently, there are two common approaches to deal with big data, namely batch-mode big data analytics and streaming-based big data analytics. The distinguished characteristic of a big data stream is that data ontinuously arrive at high speed. Accordingly, effective big data stream analytics methods should process the streaming data in one go, and under very strict constraints of space and time. Currently, research about big data analytics algorithms often focuses on processing big data in batch mode, while algorithms de- signed to process big data stream in real-time or near real-time are notabundant.

Figure 1 depicts a taxonomy of the common approaches (tools) for processing big data. Big data analytics approaches can be generally divided into distributed or single host approaches. For distributed big analytics methods, there can be then further classified into batch mode processing or streaming mode processing. Even though batch mode big data analytics methods (e.g., MapReduce) are the current dominated method, online incremental algorithms that can effectively process continuous and evolving data stream are desirable to address both the “volume” and the “velocity” issue of big data pasted on online social media. MapReduce and big data stream analytics are two different classes of analytical approaches although they are related for certain theoretical perspectives. Recently, researchers and practitioners have tried to integrate streaming-based analytics and online computation on top of the MapReduce batch mode analytics framework. Sample tools of that kind include the Hadoop Online Prototype. However, more research should be conducted for the development of next generation of big data stream analytics methods that inherit the merits from both batch mode analytics and streaming analytics. The main contribution of this paper is the design and development of a novel big data stream analytics framework that provides the essential infrastructure to operationalize a probabilistic language modeling ap- proach for near real-time consumer sentiment analysis. There is significant research and practical value of our work because organizations can apply our framework to better leverage the collective social intelligence to de- velop effective marketing and product design strategies. As a result, these organizations become more competi- tive in the global marketplace, which is one of the original promises of big data analytics.

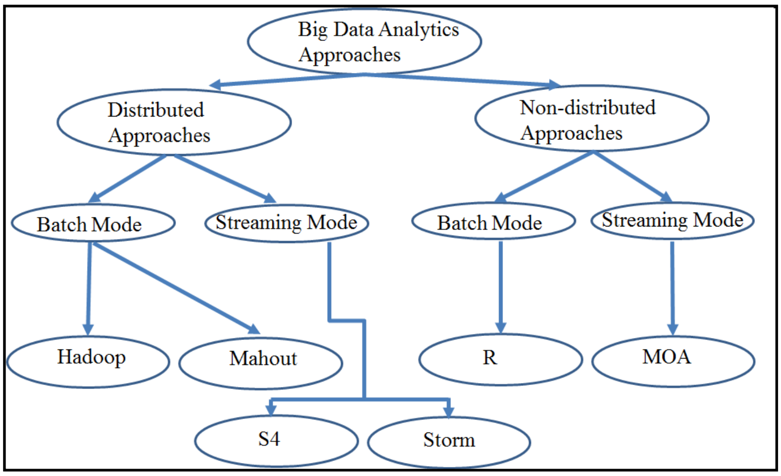

Figure 1 Taxonomy of Big Data Analytics

1. The Big Data Stream AnalyticsContext

An overview of the proposed framework that leverages Big Data Stream Analytics for online Sentiment Analysis (BDSAC) is depicted in **Figure 2**. The BDSAC context 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 Topsy API is used to retrieve product related comments fromTwitter.

The Storage Layer leverages Apache HBase and HDFS for real-time storage and retrieval of big volume of consumer reviews discussing products and services. 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 statis- tical 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 re- lated 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 potential product or service acquisitions. The Sentiment Extractor utilizes well-known sentiment lexicons such as OpinionFinder to extract the sentiment words embedded in consumer reviews. Finally, overall sentiment polarity prediction for consumer reviews 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 product or a product category is communicated to the user of the

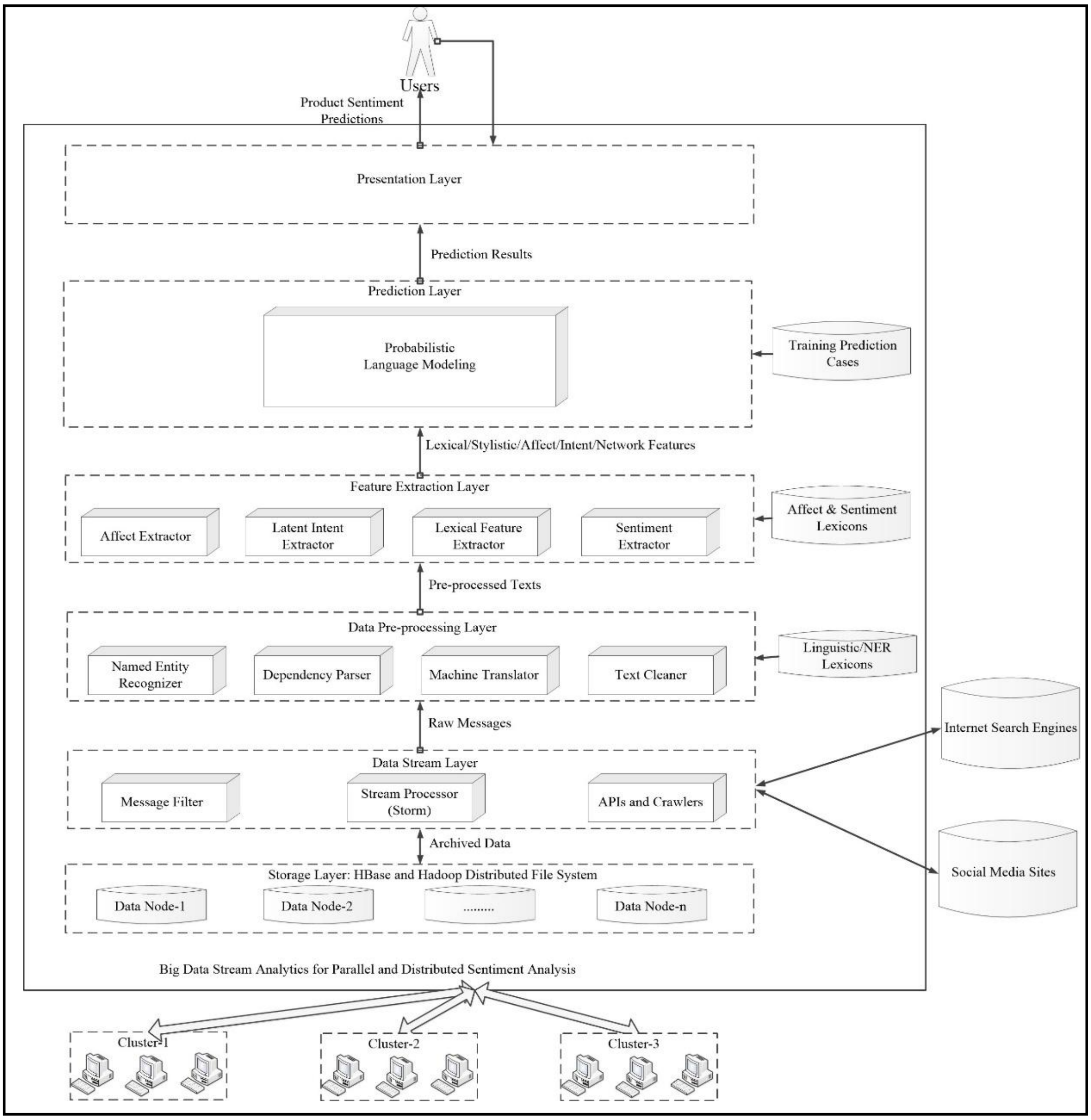

Fig 2. Big Data Stream Analytics Context

Discussions andSummary

While some research work has been devoted to big data analytics recently, very few studies about big datastream analytics are reported in the literature. The main theoretical contributions of our research include the de- sign and development of a novel big data stream analytics framework, named BDSASA for thereal-time analysis of consumer sentiments. Another main contribution of this paper is the illustration of a probabilistic inferential language model for analyzing the sentiments embedded in an evolving big data stream generated from online social media. The business implication of our research is that business managers and product designers can apply the proposed big data stream analytics framework to more effectively analyze and predict consumers’ about products and services. Accordingly, they can take proactive business strategies to streamline the marketing or product design operations.

One limitation of our current work is that the proposed framework has not been tested under an empirical set- ting. We will devote our future effort to evaluating the effectiveness and efficiency of the BDSASA framework based on realistic consumer reviews and social media messages collected from the Web. On the other hand, we will continue to refine the proposed inferential language model for better sentiment polarity prediction. For in- stance, a consumer may connect to other consumers via a social network. We may incorporate such connection features in the inferential language model when the sentiment polarity of a review is analyzed. Moreover, the prediction thresholds for probabilistic opinion scoring will be fine-tuned using the proposed PCGA. Finally, we will conduct a usability study for the proposed big data stream analytics service in a real-world e-Business environment.

References

[1] E. Zhong, W. Fan, J.W.L. Xiao and Y. Li, "ComSoc: Adaptive Transfer of User Behaviors over Composite Social Network", in 18th ACM SIGKDD international conference on Knowledge discovery and data mining, 2012.

[2] Lau, R.Y.K., Xia, Y. and Ye, Y. (2014) A Probabilistic Generative Model for Mining Cybercriminal Networks from Online Social Media. IEEE Computational Intelligence Magazine, **9**, 31-43.

[3] Turney, P.D. and Littman, M.L. (2003) Measuring Praise and Criticism: Inference of Semantic Orientation from Asso- ciation. ACM Transactions on Information Systems, **21**, 315-346.

[4] Wilson, T., Wiebe, J. and Rwa, R. (2004) Just How Mad Are You? Finding Strong and Weak Opinion Clauses. In: McGuinness, D.L. and Ferguson, G., Eds., Proceedings of the Nineteenth National Conference on Artificial Intelli- gence, Sixteenth Conference on Innovative Applications of Artificial Intelligence, San Jose, 25-29 July 2004,761-769.

[5] Archak, N., Ghose, A. and Ipeirotis, P.G. (2007) Show Me the Money!: Deriving the Pricing Power of Product Fea- tures by Mining Consumer Reviews. In: Berkhin, P., Caruana, R. and Wu, X., Eds., Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, 12-15 August 2007, 56-65.

[6] Turney, P.D. (2002) Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics,417-424.

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

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

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

[10] Lau, R.Y.K., Tang, M., Wong, O., Milliner, S. and Chen, Y. (2006) An Evolutionary Learning Approach for Adaptive Negotiation Agents. International Journal of Intelligent Systems, **21**, 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, **32**, 859.

[12] Ponte, J.M. and Croft, W.B. (1998) A Language Modeling Approach to Information Retrieval. Proceedings of the21st

[13] Zhai, C.X. and Lafferty, J. (2004) A Study of Smoothing Methods for Language Models Applied to Information Re- trieval. ACM Transactions on Information Systems, **22**, 179-214.

[14] Nie, J.-Y., Cao, G.H. and Bai, J. (2006) Inferential Language Models for Information Retrieval.

ACM Transactions on Asian Language Information Processing, **5**, 296-322.

[15] Lau, R.Y.K., Song, D., Li, Y., Cheung, C.H. and Hao, J.X. (2009) Towards a Fuzzy Domain Ontology Extraction Me- thod for Adaptive E-Learning. IEEE Transactions on Knowledge and Data Engineering, **21**, 800-813.