International Journal of Electrical and Computer Engineering (IJECE)
Vol. 7, No. 1, February 2017, pp. 402~407
ISSN: 2088-8708, DOI: 10.11591/ijece.v7i1.pp402-407

Streaming Big Data Analysis for Real-Time Sentiment based Targeted Advertising

Lekha R. Nair, Sujala D. Shetty, Siddhant Deepak Shetty
Department of Computer Science, Birla Institute of Technology and Science (BITS) Pilani, Dubai Campus, United Arab Emirates

Article Info

Article history:

Received Jul 15, 2016
Revised Dec 25, 2016
Accepted Jan 8, 2017

Keyword:

Big data
Spark
Streaming big data processing
Targeted advertising
Tweet sentiment analysis

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. In this paper, a cloud based system for real time targeted advertising based on tweet sentiment analysis is designed and implemented using the big data processing engine Apache Spark, utilizing its streaming library. Application is meant to promote cross selling and provide better customer support.

Corresponding Author:

Lekha R. Nair,
Department of Computer Science,
BITS Pilani, Dubai Campus
P.O. Box: 345055, Dubai International Academic City, Dubai, United Arab Emirates
Email: lekharnair@gmail.com

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 now 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 [1] which can be applied in social marketing. Business market leaders have identified the potential of these sites to gather opinions about a product rather than conducting a market survey, as the data from former reflects recent opinions, mostly unbiased feelings which will be more realistic and comes in huge volumes representing a fair percentage of general public, though the data being largely unstructured. In brand competition, immediate actions taken based on customer feedbacks result in strategic advantage of one brand over another. Satisfied customers of a product are more likely to buy an associated product from the same brand if an effective marketing strategy targeting those customers is successfully implemented. While a contented customer can bring in more revenue, excessive negative sentiments regarding a product, spreading over social media, can adversely affect the sales and result in losing loyal and prospective customers.

1.1. The Problem: Striking While the Iron is Hot

Cross-selling, where an additional product or service is sold to an existing customer, is detailed in [2] which requires advertising the precise product to the exact customer at the correct time. In the current

highly competitive marketing scenario, cross-selling can bring in huge revenue and the strategy is very effective when the existing customer has positive sentiment towards the owned product while targeted advertising to these customers can increase return on investment. At the same time, for every brand it is necessary to tackle the issues raised by the unsatisfied customer and to pacify him at the earliest so as to regain his brand confidence.

Attaining social network sourced real time big data for analysis is not easy as most sites lack public application programming interface(API) for a third party to access, with Twitter being an exception. According to internet statistics more than 6000 tweets are posted per second which is huge enough in terms of volume and velocity to be handled by traditional data analytic system and hence necessitates the usage of a big data processing system.

In this work, an apache spark based big data application is modelled and implemented on cloud that processes real time tweets regarding a product *x* and identify its sentiment.If the sentiment is negative, customer support is offered instantly and feedback is requested through direct message, else, advertisement of an associated product *y* is targeted to the user. Location of the user is also collected to provide location specific services and to identify geographic areas where marketing or customer service section need to be concentrated. Since these prospective customers are targeted at the right time when they have expressed their sentiments, it is obvious that this could be a better marketing strategy.

1.2. Selecting Associated/Recommended Product

In market basket analysis, customer transactions are analysed to recognize their purchasing pattern. Association rule learning [3] is a method to identify relations among variables in a dataset which can be used to find related products in customer transactions leading to effective marketing decisions. By association analysis, for a product *x*, an associated product *y* can be identified which is bought together with or after buying product *x*.

Recommendation systems identify products to be recommended based on customer's past purchases and other users behavior. A plethora of work have been carried out in association analysis [4-5] and recommender systems [6-7] and it is not included in the scope of this paper where it is assumed that the associated product *y* and the product to be recommended *z*, had already been identified.

1.3. Related Works

Many research works have been carried out in sentiment analysis [8]. Finding customer sentiments towards a brand by mining social media text was the topic of [9] while usage of twitter data for sentiment analysis was discussed in [10]. Several works were done for revealing sentiments regarding persons or products that made use of twitter data [11-13] . In most of the works, analysis was performed on static data. Usefulness of social media in business is an active research area and marketing scope of social media is detailed in [14]. Relationship marketing via twitter is the topic of discussion of [15], while marketing helpfulness of twitter in hotel industry is explained in [16].

This work implements automated real time targeted advertising system based on real time sentiment analysis of twitter data. Done from a Big Data perspective, the system is highly scalable as it makes use of big data processing engine Spark, which takes into account of challenges and opportunities of big data [17]

2. RESEARCH METHOD

2.1. Dataset: Twitter Streaming Data

Twitter, the prevalent microblogging site with 320 million monthly active accounts as per company statistics, allows user to send 140 character limited messages termed tweets, visible to all. One can also send a direct message which is visible only to the intended user. Twitter's global stream of data can be accessed with the aid of Twitter streaming API. For this real time access to tweets, a persistent HTTP connection is required to be open. An application intended to use Twitter API need to obtain OAuth access token on behalf of a twitter account. Authorized requests to the Twitter Streaming API can be issued by the application making use of access token and secret keys. Once the connection is established, Spark Streaming built on the top of spark core takes care of the reception of real time tweets which then processed by spark core engine.

2.2. Tools: Apache Spark and Spark Streaming Library

Since traditional data processing systems have scalability issues and are not equipped to handle streaming data of immense volume, a scalable big data processing system is preferred for this application. Spark [18] is an open source computing engine meant for distributed data processing. Hadoop [19], the first generation big data processing engine is slowly being replaced by Spark which is considered as the second generation Big Data processing engine by [20].

Driver program of spark application runs the main function and performs parallel operations on various worker nodes in a spark cluster. Spark uses the concept of Resilient Distributed Dataset (RDD) [21], which is a collection of immutable objects segregated across the cluster nodes for performing parallel operations. RDDs can be persisted in memory for repetitive use and due to this in-memory analytics, spark performs faster than the Hadoop, especially in iterative applications. Though Spark is mainly a batch processing engine, Spark ecosystem is equipped with Spark Streaming that is destined for streaming data processing as given in Figure 1. In spark streaming, continuous stream of data is represented by discretized stream (Dstream) which is a sequence of RDDs. In this work, spark streaming receives and handles the real time tweets from the Twitter Streaming API after establishing the connection.

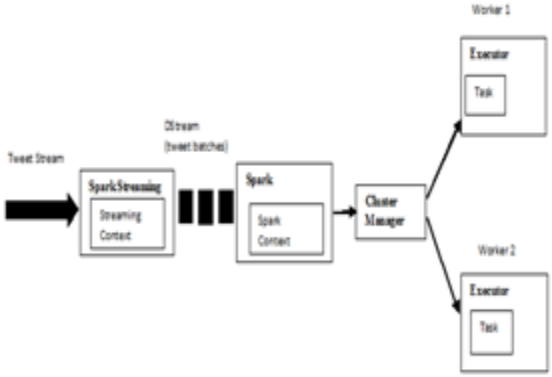

Figure 1. Spark with Streaming : Architectural diagram

2.3. The System Model

The work flow model of the system is given in Figure 2, which is built around Spark. Once the connection with Twitter streaming API is established, from among thousands of tweets posted per second, the application filter tweets regarding a particular product *x*. Spark Streaming handles this streaming data and pack these tweets into batches and hand over to underlying spark core engine for processing. Sentiment of each tweet is analyzed in real time and if found positive/neutral, advertisement of an associated product *y* or a recommended product *z* is targeted to the tweeter, while steps are taken to offer customer support and gather relevant information regarding dissatisfaction in case of negative tweets, so that remedial measures can be taken immediately to prevent losing prospective customers

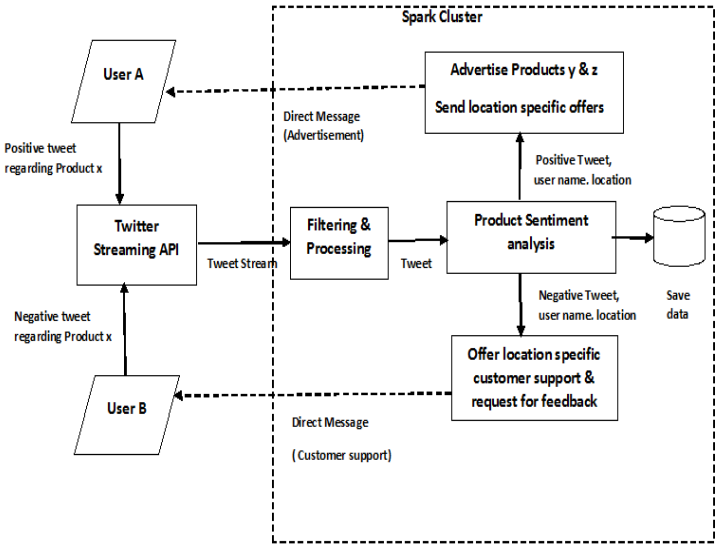

Figure 2. Work Flow Model of the Application

2.4. Product Sentiment Analysis

Many research works have been carried out in sentiment analysis. Stanford offers an open source sentiment analyzer library that can be used effectively to carry out sentiment analysis. There is no fool proof algorithm for sentiment analysis as many NLP algorithms stumble on accurately identifying sarcastic comments. For this prototype, we have used the most primitive type of sentiment analytic method of finding the relative count of positive and negative sentiment holding words in the tweet. A large collection of over 5000 words like good, amazing etc. commonly used to express positive sentiments are compiled in a text file to be used as a lookup table. Same is done for negative words as well. A counter is initialized to zero, assuming neutral sentiment, and for each word in the tweet, comparison is done with a set of positive words and negative words. If the word is associated with positive sentiment, counter is incremented or if negative, the counter is decremented and the sign of the final counter value determines whether the product is associated with positive, negative or neutral sentiment. Though the method is simple, it is having obvious drawbacks and can be replaced by any convenient sentiment analytic algorithm.

2.5. Location Specific Services

For location enabled tweets, user location is identified from the tweet, and location specific offers and services are targeted to these users. By mapping tweet locations based on sentiments, geographic areas where attention is required can be identified and appropriate actions can be taken.

2.6. Algorithm

Select a product *x*

a. Find associated product *y* using association analysis
b. Find a product *z* that can be recommended to user using recommendation system.
c. While (twitter API connection is true)
    a. Filter tweet stream regarding the product
    b. For each tweet (tweet(i))
        1. Get username(user(i) and location(loc(i))
        2. Find sentiment of the tweet senti(i)
            If(senti(i)==positive OR neutral)
            Advertise associated product *y* and *z* to the user(i)
            If (loc(i) is not null)
            advertise location loc(i) specific offers to the user(i) else
            Offer customer support to user(i) and request for user(i) feedback
        3. Save tweet(i), loc(i) and senti(i) for further analysis.

3. RESULTS AND ANALYSIS

Though many works regarding sentiment analysis of twitter data were done before, this work utilizes real time tweet sentiment analysis for real time targeted advertising making use of scalable open source spark streaming, which was not attempted before. The application was built using Simple Build Tool (SBT) and run on a Spark cluster with a master and two slave nodes configured on i5 processor, 4GB RAM and Ubuntu 14.04 operating system. It was also successfully deployed on Amazon Elastic Compute Cloud (EC2). Spark Cluster with t2.micro configuration was created and after testing the application, the cluster was destroyed. Spark ec2 script was utilized in launching and managing spark cluster in EC2 cloud.

Table 1. Received Tweets and Real Time Response based on Sentiment and Location

Real Time Tweets Received	Sentiment identified, Location	Direct Message Sent (Targeted Advertisement)
my xpad10 works fine	Positive, null	Limited period offer, 10% discount on all Orange mobile accessories
Price of XPad-10 is good but picture quality is poor	Neutral,Dubai	Limited period offer, 10% discount on all Orange mobile accessories Amazing offer: Clearance Sale at Orange i-stores at Deira City Center, Dubai
New xpad 10 sucks, dont buy	Negative, India	Please call toll free no 800-1234 for all your complaints or visit ww.orange.com/custcare to serve you better

The application was initially tested by filtering tweets regarding popular products available in the market and its sentiment were analyzed, location identified and saved in a file. Targeted advertising was disabled in this case. Thirty to sixty tweets per minute were observed regarding already established market products, but the number is expected to shoot up in the initial periods when a new product is launched..

The application was tested by sending positive and negative sentiment tweets from 5 different twitter accounts about a hypothetical product xPad-10 from company Orange. All the tweets were received in real time and its sentiments were identified and accordingly promotional offer messages or feedback request/customer support details were sent by the application as direct message to each tweeter as given in Table 1. Also the tweet details were recorded in a file for detailed analysis later.

4. DISCUSSIONS

In this paper a scalable spark application to perform real time targeted advertising to prospective customers based on the sentiments expressed on related products on twitter is implemented. Since no sentiment analysis algorithm gives a fool proof result, the observed sentiment may be different in some cases, but since the application is about real time targeted advertising, it will not have any negative effect on the performance.

Twitter users who are very much concerned about their privacy might disable location tracking, where location specific services becomes insignificant. Also, if the user disables the option of receiving direct messages from everyone, it will be hard to target that user for advertising.

5. CONCLUSION

The Big Data analytic system meant for real time targeted advertising where target identification is done on the basis of customer sentiments shared on twitter, was successfully built around the big data processing system Apache Spark and tested on Amazon EC2 cloud.

The same application with slight modification can be used in international politics for direct campaigning and to take corrective measures based on public opinions as well as to formulate winning strategy based on predictions in elections. In this work in addition to real time analysis, the individual tweet with its location and predicted sentiment is stored in a csv file which can be mined to gain insights towards a long term policy formulation.

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] S. Li, B. Sun and L. M. Alan, "Cross-selling the right product to the right customer at the right time", *Journal of Marketing Research*, vol. 48, no. 4, pp. 683-700, 2011.
[3] R. Agrawal, T. Imieliński and A. Swami, "Mining association rules between sets of items in large databases", in *ACM SIGMOD international conference on Management of data*, 1993.
[4] C.C. Aggarwal, C. Procopiuc and P.S. Yu, "Finding localized associations in market basket data", *IEEE Transactions on Knowledge and Data Engineering*, vol. 14, no. 1, pp. 51 - 62, 2002.
[5] M. Kubat, A. Hafez, V.V. Raghavan, J.R. Lekkala and W.K. Chen, "Itemset trees for targeted association querying", *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, no. 6, pp. 1522 - 1534, 2003.
[6] H.K. Kim, J.K. Kim and Y.U. Ryu, "Personalized Recommendation over a Customer Network for Ubiquitous Shopping", *IEEE Transactions on Services Computing*, vol. 2, no. 2, pp. 140 - 151, 2009.
[7] K.A. Almohsen and A.J. Huda, "Recommender Systems in Light of Big Data", *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 5, no. 6, 2015.
[8] B. Liu, "Sentiment analysis and opinion mining", *Synthesis lectures on human language technologies*, vol. 5, no. 1, pp. 1-167, 2012.
[9] M.M. Mostafa, "More than words: Social networks' text mining for consumer brand sentiments", *Expert Systems with Applications*, vol. 40, no. 10, pp. 4241-4251, 2013.
[10] A. Pak and P. Patrick, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining", *LREC*, vol. 10, pp. 1320-1326, 2010.
[11] S. Liu et al., "TASC:Topic-Adaptive Sentiment Classification on Dynamic Tweets", *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 6, pp. 1696 - 1709 , 2015.
[12] P.R. Cavalin et al., "A scalable architecture for real-time analysis of microblogging data", *IBM Journal of Research and Development*, vol. 59, no. 2/3, pp. 16-1, 2015.
[13] X. Chenyan, Y. Yang and H. Chun-Keung, "Hidden in-game intelligence in NBA players' tweets", *Communications of the ACM*, vol. 58, no. 11, pp. 80-89, 2015.

[14] M.S. Yadav et al., "Social commerce: a contingency framework for assessing marketing potential", *Elsevier Journal of Interactive Marketing*, vol. 27, no. 4, pp. 311-323, 2013.
[15] B.A. Watkins and R. Lewis, "*Twitter as Gateway to Relationship Marketing: A Content Analysis of Relationship Building via Twitter*", in Social Media and Strategie Communications, UK, Palgrave Macmillan , 2013, pp. 25-44.
[16] X.Y. Leung, B. Billy and A.S. Kurt, "The marketing effectiveness of social media in the hotel industry a comparison of facebook and twitter", *Journal of Hospitality & Tourism Research*, vol. 39, no. 2, pp. 147-169, 2015.
[17] H. Bagheri and A. Abdusalam, "Big Data: challenges, opportunities and Cloud based solutions", *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 5, no. 2, p. 340, 2015.
[18] [Online]. Available: https://spark.apache.org/docs/latest/. [Accessed 15 February 2016].
[19] T. White, "Hadoop: The Definitive Guide, 3rd Edition", O'Reilly Media, California, 2012.
[20] F. Gebara, H. Hofstee and K. Nowka, "Second-Generation Big Data Systems", *IEEE Computer*, vol. 48, no. 1, pp. 36-41, 2015.
[21] M. Zaharia, M. Chowdhury, M.J. Franklin, S. Shenker and I. Stoica, "Spark: Cluster Computing with Working Sets", in *USENIX conference on Hot topics in cloud computing*, 2010.