International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 1, Issue 3, ISSN No. 2455-2143, Pages 68-72
Published Online December – January 2016 in IJEAST (http://www.ijeast.com)

RECENT EMERGENT TRENDS IN SENTIMENT ANALYSIS ON BIG DATA

Bhupendra, Komal Varshney, Dhruv
GL Bajaj Institute of technology and Management
Greater Noida, UP India

ABSTRACT - Sentiment analysis of social networking site involves a lot of data preprocessing .The way Big data heavy volume, highly volatile and vast variety of data has entered our lives, it is becoming day by day difficult to manage and gain business advantages out of it .Basically Sentiment analysis is used for knowing voice or response of crowd for of our products, services, organizations etc. .The practical implication of our research work is that organizations can apply our big data stream analytics framework to analyses user specified requirements and so that we can develop more effective marketing and production strategies.

KEYWORDS *- Big data, Sentiment analysis, text mining, Hadoop tool.*

I. INTRODUCTION

With the emergence of social media news in every domain of today's digital age from algorithmic trading and product recommendations to politics, there is a huge amount of research work going on in the field of sentiment analysis and opinion mining which is taking us leaps and bounds with the advent of Big Data platforms and tools. The amount of data that could be collected, processed and stored cheaply and effectively is increasing at an exponential rate with the advent of Hadoop and other related parallel platforms and tools. Our work aims at studying the importance of preprocessing in the era of big data where storage and processing of unstructured data is as simple as processing structured data. So why do we have to preprocess the data if Hadoop and other big data tools support handling unstructured data effectively? If we want what kind of preprocessing are we talking about and how different it is from the preprocessing that we do in a regular KDD process? What kind of tools work well in such a scenario and how it is done effectively on such a tremendous volume of data? Since most of the tools used in the Hadoop Ecosystem fundamentally works on the basis of the Map Reduce paradigm (which is a batch model), how well do they handle the pre rate, like Twitter or posts from Facebook users? How do we handle the velocity part of the Big Data problem? Are the tools of the past no longer authenticate for these purposes due to the huge volume, variety and velocity of data? These are some of the questions that we are trying to answer. The main objective of this work is to identify the best framework or set of tools to pre-process the data from a social networking site like twitter. Even most of the algorithms for mining big data are found to be unstructured data and also found to be strong to variations in data formats and structure, we emphasis on the importance of preprocessing data as its importance are found to be many folded. First it lets us understand the unstructured data that we are dealing with a advance way. Second it helps in dimensionality reduction thereby eliminating a lot of unnecessary features from being handled. Third it allows answering the Value part of the Big Data paradigm. Finally it allows us to fine grip the data model according to the processing requirements making it much more reliable and accurate. The research paper is organized as follows. The second section is about the review of literature analyzing the existing techniques for preprocessing of social networking site data especially, twitter feeds. The next section (Section 3) discusses the various preprocessing techniques and the necessity of social media data and what type of output is expected from a preprocessing framework and also discusses the importance of different words and emotions as a special case with respect to twitter site data. Section 4 discusses about the Stanford Twitter Dataset, the Twitter API . Section 5 explains in detail the different parts of the preprocessing framework design and about the tools and techniques that are found to be suitable for it. processing of data that is arriving at a faster Section 6 deals with the different platforms and tools that could be employed to handle the pre-processing of twitter data and a brief discussion about their various effectiveness from a theoretical standpoint. It also deals with the experimental setup and the various parameters considered during the preprocessing phase. Section 7 deals with results and discussion .In today's scenario data is generated from every other known platform. The sensors used to gather information about climate ,posts to social media sites, digital images and videos, ecommerce transaction records, and mobile phone GPS signals to name a very minute amount of device that produce data and that can used for various significant applications. This data gathered is

International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 1, Issue 3, ISSN No. 2455-2143, Pages 68-72
Published Online December – January 2016 in IJEAST (http://www.ijeast.com)

termed as big data. The McKinsey Global Institute estimates that data amount is growing 40% per year, and will grow 44x between 2010 and 2020. But it's often the visible parameter; volume of data is not the only single factor that matters.

II. CHARACTERISTICS OF BIG DATA

2.1 Volume

The data generated is enormous in these systems, especially machine generated data. Machine generated data is produced in much larger quantities when compared to non-traditional data. For example, act eristic that matters. A single jet engine is responsible for generating 10TB of data in 30 minutes of its operation. With approximately more than 25,000 airline flights per day, the daily expected volume of just this single source of data can run into terabytes. In contrast, data generation social media is comparatively less over whelming, though still on the larger end of the spectrum.

2.2 Velocity

Social networking data streams – while not as massive as machine-generated datasets produce a large amount of opinions and relationships which are immensely valuable to the customer relationship management. Though it is limited to 140 characters per tweet and the high velocity (or frequency) of Twitter data ensures large volumes (over 8 TB per day as estimated).

2.3 Variety

Traditional data formats are known to be relatively well defined observe a data schema and change slowly. On the other non-traditional data formats are unstructured and change at a dizzying rate. As new services and devices are added, new sensors deployed, or new marketing campaigns adopted, new data types are needed immediately to capture the resultant information and analyze the information.

III. HADOOP

The Hadoop platform was designed to solve problems which had a lot of data for processing. It uses the divide and conquers methodology for processing. It is used to handle complex unstructured data which doesn't match into tables. Twitter data being relatively unstructured can be best stored using Hadoop. It also finds a lot of approach in the field of online retailing, search engines, finance domain for risk analysis etc.

IV. HDFS

Hadoop Distributed File System is a distributed file system which runs on commodity machines. It is highly fault for bearing and is designed for low cost machines. HDFS has a high throughput tackiest application and is suitable for requests with large amount of data. HDFS has a 1master server architecture which has a single name node Data replication is done for achieving wrong tolerance. The large data cluster is stored as a sequence of blocks. Block size and the replication factor are configurable. Replicationfactorissetto3inour project which means 3 copies of the same data block will be maintained at a time in the cluster which regulates the file system access. Data nodes handle read and write requests from the file system's clients (user). They also perform block creation, deletion, and replication upon instruction from the name node. Replication of data in the file system adds to the data integrity and the robustness of the system.

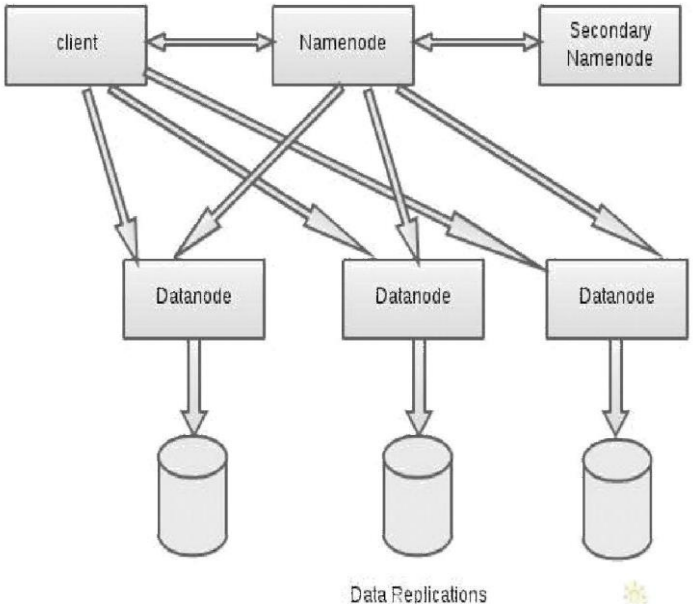


Figure 1.

V. SENTIMENT ANALYSIS

The field of sentiment analysis is a gigantic one and also popularly known as opinion mining. Sentiment analysis of natural language texts is currently a vast and growing field. Statistical methods are well suited for psycholinguistic analysis for inferring people's attitude, social -class, standards etc. It has led to many algorithms being developed which checks whether a given online text is

objective or subjective, and whether any opinion that has been expressed is positive or negative.
Such various methods have been applied on a large scale to study sentiment-related work. According to a survey across all of the studies described there, only 4% were emotional (adjectives) words used in written texts. This shows that evaluating the effective lexicons is not sufficient to recognize affective information from texts and raises the feelings that methods like machine learning or keyword spotting or lexical affinity may not perform well for this aim and these are strongly been criticized in. A similar approach has been used for Facebook status updates to observe changes in the mood over the entire year and to generally assess "the overall emotional health of the nation"and another such project evaluate six dimensions of overall emotions expressed in Twitter, showing that these six dimensions of emotions typically reflect significant offline events in the life of the people. Nevertheless, despite extensive research into the role played by sentiment in online communication, there are yet no investigations into the role what role sentiment plays in significant online events. To reduce this polarity and to effectively fill this gap, this paper assesses whether or not.

Twitter-based surges of interest in an event can be associated with increases in expressed strength of feeling. Within the segment of media, a well-known idea is that emotions play a significant role in engaging attention, as expressed for violence by the famous saying "if it bleeds, it leads", and through proven evidence that audiences emotionally engage with the news. It seems very logical and straightforward, therefore, to hypothesize that such events triggering large reactions in Twitter would be associated with increase in the strength of expressed sentiments, but there is no yet for this hypothesis. Sentiment analysis of natural language texts is currently a vast and growing field. Statistical methods are well suited for psycholinguistic analysis for concluding people's attitude, social-class, standards etc.

Fuzzy logic is known to assess input text by identifying regular verbs and adjectives in the sentences, not considering their semantic relationships that have pre-assigned affective category, centrality and intensity. Similarly to machine learning, it cannot produce a legitimate analysis for smaller text units such as sentences. Knowledge-based approach studied in what manner humans express emotions in face to face communications and based on this substantial study, two-dimensional affective lexicon database and a set of rules that describes dependencies between linguistic contents and emotions have been developed. In our personal opinion this approach is very much similar to keyword-spotting and therefore not appropriate for sentence-level sentiment recognition. The use of domain specific corpora for emotional grouping of text has shown very promising results regarding sentiment analysis of blogs, but it requires special tuning on data important to build category specific classifiers for human interest domains.

The system to be built involves processing of data from the user's feedback, comments or tweets and converting it into a form where it can be analyzed and studied. The huge deposits of data in the web world are difficult to extract and analyze. Sentiment analysis is an algorithm used to determine the mood of the public. Opinion has four basic parts named as topic, holder, claim and sentiment. The holder expresses a view on the topic and gives out a sentiment on it.

The main challenge in the words found in data stream is finding out sense of a word (meaning) in a given sentence, when the word has multiple meanings. The disambiguation misleads the opinion making system. People use different types of contradictory statements together which might change the meaning.

VI. LITERATURE SURVEY

6.1 Sentiment Based Time-Series Analysis

Time series analysis of online data is clubbed with sentiment analysis in order to make predictions in addition to the work reviewed above. For example, sentiment and frequency of a blog post can predict movie emoluments, with more positive reviews suggesting improved profits. The estimates of the everyday amount of fear, worry, and anxiety in live blogs may also predict overall stock market movement (Gilbert & Karahalios, 2010).

The relation between sentiment and spikes of online interest has been investigated through an analysis of live blog posts for a particular range of mood-related terms For example – excited, drunk, tired terms self -chosen by the bloggers to elucidate their current mood. To recognize current modifications in mood, the average number of mood-annotated posts was correlated with the average for the exact particular hour of the exact particular day of the week and over all the previous weeks for which the data was available (Balog, mishne , & Rijke ,2006).

To find the root cause of identified mood changes, frequent words for posts related to the change in mood were correlated with a reference set so that it can identify common words. Though the full evaluation was not conducted, the algorithm was able to identify major new stories. (THelwall & Prabowo, 2007; Thelwall, Prabowo & Fairclough, 2006).

The time series analyses of emotion have also been applied on a scope of online and offline texts to detect overall

International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 1, Issue 3, ISSN No. 2455-2143, Pages 68-72
Published Online December – January 2016 in IJEAST (http://www.ijeast.com)

trends in amount expressed (Dodds & Danforth, 2010). A distinct approach using Twitter data from 2008 and 2009 compared the polarity of tweets admissible to topics with the results of appropriately significant questions in published opinion polls. A strong correlation was observed between the results of an opinion poll and sentiment based twitter cores, showing that among the well known topics, Automatic sentiment detection of Twitter could audit public opinions. (O'Connor, lasubramanyan, Routledge, & Smith, 2010) .

The method used was time series method which was able to anticipate the outcome of the debate as well as particular points of interest in the debate that provoked emotions.

VII. OBSERVATIONS

7.1 Information Extraction From The Relevant Datasets

The crucial problem found in text extraction is automatic document sense extraction. Noun phrases are mostly found in a document. Extraction of such noun phrases is a difficult job which is solved to an extent by statistical methods. Blogs is a place where we can find a variety of problems such as lexical, syntactic and stylistic. The comments in the blogs are subjective i.e. the first comment (parent comment) might contain the subject of the conversation and thus we can find the link between comments following it .Automatic text summarization analyzes important information from texts and therefore it is used to understand the text. Blogs: Blogs is an online portal where users expresses his sentiments and is available to the world through the hyperlink. Daily events and feelings have been posted on blogs followed by comments of various users which are in fact their personal opinion and sentiments.

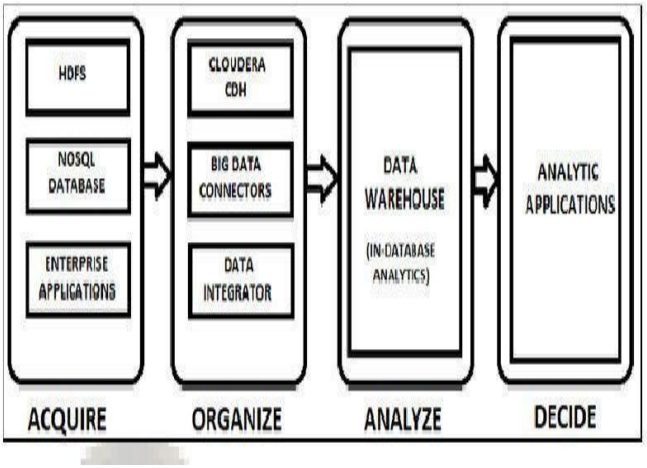
Reviews: Sites like zomato.com, amazon.com etc. are giants in review data. Here the users post their opinion about a product or an entity. These reviews in turn are useful for users referring them in decision making as whether to invest in a certain product or not.

7.2 How Is Sentiment Is Calculated

The sentiment which might be positive, negative or neutral is determined by certain constraints which determine the emotional content of a statement. This emotional content is actually bought into existence by adjectives or adverbs. Tonal sentiments (adjectives, nouns and auxiliary) are such that their sentiment score is considered to be neutral. In technical words, these parts are the emotion creators to form a sentiment chain. Such word chain recalculation can sometimes be a problem.

Therefore it is easier to secure lexemic defined sentiments strength for adverbs, tonally neutral adjectives, nonverbal collocations, and for predicative elements, namely, for verbs and verb collocations. The word chains in the sentence are used to determine sentence level sentiment strength. The sentence is represented as ‘subject chain’, ’predicative chain’, ‘object chain’. Then the system analyzes and calculates the polarity of each chain. A set of rules is used to define the sentiment of the whole sentence and the sum of polarity scores of the three chains determines the sentence polarity score. These two values are independent and therefore we do not look at its value. Output: The system output is as such that the font sizes determine the polarity of the sentences. If the polarity is higher than neutral then an increased font size is observed. Font size is proportional to the sentiment score. Depending on the sentiment score and sentiment value, a sentence can consist one of the following seven sentiment characteristics: weak, negative, medium negative, strong negative, neutral, weak positive, medium positive and strong positive.

Figure 2



7.3 Tools Used In Big Data Industry

Oracle was the first in the database industry to offer big data solutions. Its idea is centered on the idea that one can extend your current information enterprise architecture to acquire new big data. Hadoop and Oracle No SQL database run alongside of your oracle data warehouse to address your big data requirements. The following diagram represents the oracle big data solution. The requirements in a big data infrastructure span data acquisition, data organization and data analysis.

Acquire: Big data refers to higher data streams which relarge in terms of velocity and variety. No Sql is mostly used to acquire big data as they are well suited for dynamic data structures and can be scaled. The No Sql does not categorize its data or parse it, it simply captures it. Organize: Organizing the data in a proper form is called data integration. As there is such a large amount of data, relocating the data will be much time consuming as well as more of resources will be used. Thus, the data is organized at its initial destination location only. The organization system must be able to process the large amount of data at its initial location, as well as should be able to support a high throughput. It should also be able to handle various varieties of data formats, from structured to unstructured. Hadoop is a technology used to handle large data and organize it. Hadoop Distributed File System (HDFS) is a long term storage system usually used for web logs. Analyze: Since it's not always possible to move data during its organizing phase, analysis is introduced such that some data will stay at its original location, while some would be accessed from a data warehouse transparently. The analytic system must be able to support deeper analytics like statistical analysis and data mining. It should be able to scale to large data volumes, deliver faster response time decisions based on analytical model. New solutions just not come from new data, but the old data can also be analyzed to gain new perspective on old problems. Finally as it is evident, there are some practical implications from this research, the three major steps as proposed by oracle plays a very significant part in Big Data Implementations across all sectors of the industries. Also the analysis of sentiment of the spiking events in Twitter posts gives some strong evidence that important events in Twitter are associated with increases in average negative sentiment strength. The overall level of sentiment in social networking site seems to be typically quite low and so the importance of sentiment should not be exaggerated.

VIII. FUTURE WORK

In this it has shown the way for doing sentiment analysis for social networking site. We can visualize the word map i.e., the most frequent words that are used in positive, moderate and negative fields by using R language to visualize.

IX. REFERENCE

[1] Bernard J. Jansen, Mimi Zhang, Kate Sobel and AbdurChowdury,Micro-blogging as online word of mouth branding, 27th International Conference Extended Abstracts on Human Factors in Computing Systems, New York, 2009,pages 3859-3862.

[2]G. Forman, An extensive empirical study of feature selection metrics for text Classification, Journal of Machine Learning Research, vol.3, 2003, pages 27-35.

[3] 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 Intelligence, Sixteenth Conference on Innovative Applications of Artificial Intelligence, San Jose, 2529 July 2004, 761769

[4] S. Ghemawat, H. Gobioff and ST. Leung, "The Google File System," ACM SIGOPS Operating System Review, Vol. 37, Iss. 5, pp. 2943, December 2003.

[5] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The Hadoop Distributed File System," in the 26th IEEE Symposium on Mass Storage Systems and Technologies, pp. 110, May 2010.

[6] Bahrainian, S.A., Dengel, A., Sentiment Analysis using
Sentiment Features, In the proceedings of WPRSM Workshop and the Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence, Atlanta, USA, 2013.

[7] Valitutti, A., Strapparava, C. and Stock, O. (2004) Developing Affective Lexical Resources. Psychology.

[8] Ashraf M.Kibriya, EibeFrank, BerhardPfahringer and Geoffrey Holmes, Multinomial Naïve Bayes for text categorization revisited, Proceedings of the 17th Australian Joint conference on Advances in Artificial Intelligence,Australia,2004, pages 488-499.

[9] Bo Pang,LillianLee,ShivakumarVaithyanathan, Thumbs up? Sentiment Classification using Machine LearningTechniques, Proceedings Conference of Emperical methods in natural language, Association for Computational Linguistics Language Processing (EMNLP), University of Cornell ,Philadelphia, 2002, pages 79- 86

[10] N.Rogovschi and N.Grozavu, "Opinion retrieval through unsupervised topological learning", IEEE International Joint.
Conference on Neural Networks (IJCNN), 2014 2014 , pp. 3130 – 3134

[11]] A. Muangon, S.Thammaboosadee, C. Haruechaiyasak, "A Lexiconizing Framework Of Feature-Based Opinion Mining in Tourism Industry", Fourth IEEE International Conference on Digital Information and Communication Technology and it's Applications (DICTAP), 2014 , pp.169 – 173.

[12] V.Y.Karkare and S.R.Gupta, "Product Evaluation Using Mining and Rating Opinions of Product Features", IEEE International Conference on Electronic Systems, Signal Processing and Computing Technologies (ICESC), 2014 2014, pp. 382 – 385.