International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 07 | July-2018 www.irjet.net p-ISSN: 2395-0072

SENTIMENTAL ANALYSIS ON RAW SOCIAL MEDIA DATA

C.KANNAKI

Student, Department of Computer Science, Government Arts College, Coimbatore, Tamil Nadu, India.

Abstract - *The study of these structures uses social network raw data to analysis, identifies and computes local and global people's sentiment opinions using ANN- Data Dictionary. The social network perspective provides a set of methods for analyzing the structure of whole social entities as well as a variety of theories explaining the patterns observed in these structures. Millions of users share their opinions on Social Networks, making it a valuable platform for tracking and analyzing public sentiment. Such tracking and analysis can provide critical information for decision making in various domains. The proposed work tries to analyze and interpret the public sentiment variations in micro blogging services. Using Pre-processing method we analysis and remove redundant data and compute the actual score of sentiment words. Next, we compute the number of positive and negative words in a comments using Deep Learning method. We propose a sentimental data analysis model using Neural Networks. To further enhance the readability of the mined reasons, we select the most representative tweets for foreground topics and develop another generative model called Reason Candidate and Background LDA (RCB-LDA) to rank them with respect to their popularity within the variation period. The Performance chart and Timing chart shows the Existing and Proposed methods performance enhancement.*

Key Words: ANN-Data Dictionary, Pre-processing, Deep Learning Method, Sentimental data Analysis, Neural Network, Reason Candidate and Background LDA (RCB-LDA).

1. INTRODUCTION

A social network is a social structure made up of a set of social actors (such as individuals or organizations) and a set of the dyadic ties between these actors. The social network perspective provides a set of methods for analyzing the structure of whole social entities as well as a variety of theories explaining the patterns observed in these structures.

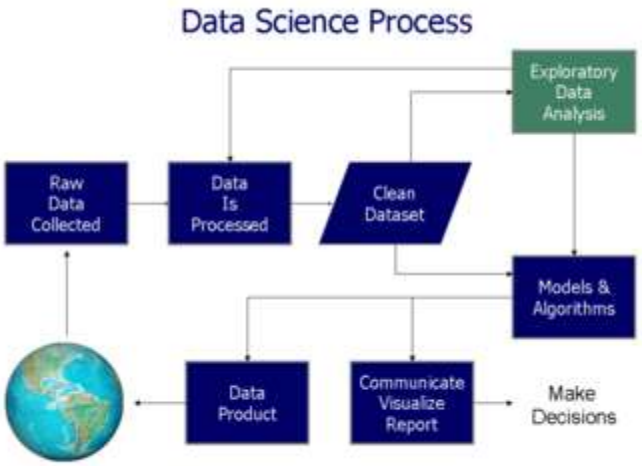
Millions of users share their opinions on Social Networks, making it a valuable platform for tracking and analyzing public sentiment. Such tracking and analysis can provide critical information for decision making in various domains. Therefore it has attracted attention in both academia and industry. Previous research mainly focused on modelling and tracking public sentiment. In this work, we move one step further to interpret sentiment variations. We observed that emerging topics (named foreground topics) within the sentiment variation periods are highly related to the genuine reasons behind the variations. To the best of our knowledge, our study is the proposed work that tries to analyze and interpret the public sentiment variations in micro blogging services.

2. RELATED WORKS

Social media sites (e.g., Flickr, YouTube, and Facebook) are a popular distribution outlet for users looking to share their experiences and interests on the Web. These sites host substantial amounts of user-contributed materials (e.g., photographs, videos, and textual content) for a wide variety of real-world events of different type and scale. By automatically identifying these events and their associated user-contributed social media documents, which is the focus of this paper, we can enable event browsing and search in state-of-the-art search engines.[1].Presents parameter estimation methods common with discrete probability distributions, which is of particular interest in text modelling. Starting with maximum likelihood, a posterior and Bayesian estimation, central concepts like conjugate distributions and Bayesian networks are reviewed. As an application, the model of latent Dirich let allocation (LDA) is explained in detail with a full derivation of an approximate inference algorithm based on Gibbs sampling, including a discussion of Dirichlet hyper parameter estimation.[2]. Web advertising (Online advertising), a form of advertising that uses the World Wide Web to attract customers, has become one of the world's most important marketing channels. This paper addresses the mechanism of Content Oriented advertising (Contextual advertising), which refers to the assignment of relevant ads within the content of a generic web page, e.g. blogs. As blogs become a platform for expressing personal opinion, they naturally contain various kinds of expressions, including both facts and comments of both a positive and negative nature. In this paper, we propose the utilization of sentiment detection to improve Web-based contextual advertising. The proposed SOCA (Sentiment-Oriented Contextual Advertising) framework aims to combine contextual advertising matching with sentiment analysis to select ads that are related to the positive (and neutral) aspects of a blog and rank them according to their relevance [3].

3. RESEARCH METHODOLOGY

3.1 Data Analysis



Data analysis is a process for obtaining raw data and converting it into information useful for decision-making by users. Data is collected and analyzed to answer questions, test hypotheses or disprove theories. There are several phases that can be distinguished. The phases are iterative, in that feedback from later phases may result in additional work in earlier phases.

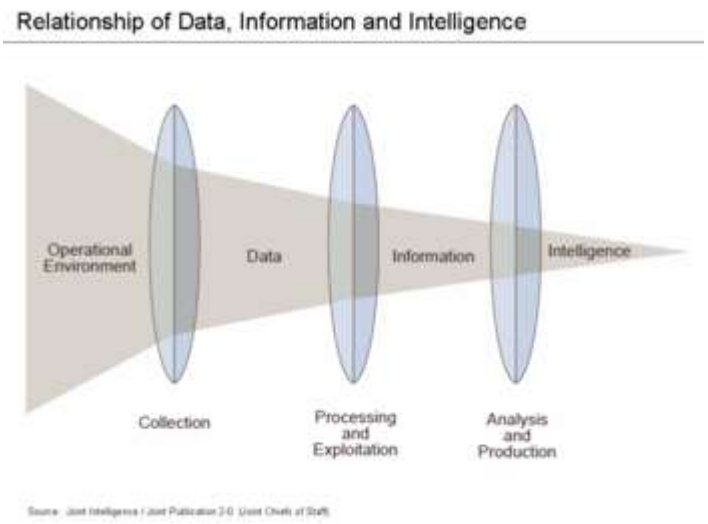
3.2 Data Requirements

The data necessary as inputs to the analysis are specified based upon the requirements of those directing the analysis or customers who will use the finished product of the analysis. The general type of entity upon which the data will be collected is referred to as an experimental unit (e.g., a person or population of people). Specific variables regarding a population (e.g., age and income) may be specified and obtained. Data may be numerical or categorical (i.e., a text label for numbers).

3.3 Data Collection

Data is collected from a variety of sources. The requirements may be communicated by analysts to custodians of the data, such as information technology personnel within an organization. The data may also be collected from sensors in the environment, such as traffic cameras, satellites, recording devices, etc. It may also be obtained through interviews, downloads from online sources, or reading documentation.

3.4 Data Processing



The phases of the intelligence cycle used to convert raw information into actionable intelligence or knowledge are conceptually similar to the phases in data analysis.

Data initially obtained must be processed or organized for analysis. For instance, this may involve placing data into rows and columns in a table format for further analysis, such as within a spreadsheet or statistical software.

3.5 Data Cleaning

Once processed and organized, the data may be incomplete, contain duplicates, or contain errors. The need for data cleaning will arise from problems in the way that data is entered and stored. Data cleaning is the process of preventing and correcting these errors. Common tasks include record matching, de duplication, and column segmentation. Such data problems can also be identified through a variety of analytical techniques. For example, with financial information, the totals for particular variables may be compared against separately published numbers believed to be reliable. Unusual amounts above or below pre-determined thresholds may also be reviewed. There are several types of data cleaning that depend on the type of data. Quantitative data methods for outlier detection can be used to get rid of likely incorrectly entered data. Textual data spellcheckers can be used to lessen the amount of mistyped words, but it is harder to tell if the words themselves are correct.

3.6 Exploratory Data Analysis

Once the data is cleaned, it can be analyzed. Analysts may apply a variety of techniques referred to as exploratory data analysis to begin understanding the messages contained in the data. The process of exploration may result in additional data cleaning or additional requests for data, so these activities may be iterative in nature. Descriptive statistics such as the average or median may be generated to help understand the data. Data

visualization may also be used to examine the data in graphical format, to obtain additional insight regarding the messages within the data.

3.7 Modelling and Algorithms

Mathematical formulas or models called algorithms may be applied to the data to identify relationships among the variables, such as correlation or causation. In general terms, models may be developed to evaluate a particular variable in the data based on other variable(s) in the data, with some residual error depending on model accuracy (i.e., Data = Model + Error).

Inferential statistics includes techniques to measure relationships between particular variables. For example, regression analysis may be used to model whether a change in advertising (independent variable X) explains the variation in sales (dependent variable Y). In mathematical terms, Y (sales) is a function of X (advertising). It may be described as Y = aX + b + error, where the model is designed such that a and b minimize the error when the model predicts Y for a given range of values of X. Analysts may attempt to build models that are descriptive of the data to simplify analysis and communicate results.

4. EXISTING MODEL

4.1 Mapping Function

The main strategy of mapping the words and documents to the space is to first compute the word embeddings, and then derive the document embeddings based on the word embeddings by considering the word occurrences. Linear projection is assumed to transform the original feature representation of words to their embedding presentation. Specifically, adk projection matrix PA issued to map words in domain A to a k-dimensional embedding space R k, while a dh projection matrix PB issued to map words in domain B to the same embedding space. Given in total M+MA words in domain A including the M pivots appearing in both domains and MA non-pivot words only appearing in domain A, we let n e z

I oM+MA.

Denote their corresponding word embeddings stored in an (M+MA)k embedding matrix e ZA computed by the linear projection mapping given as

$$\widetilde{\mathbf{Z}}_A^T = \left[\mathbf{P}_A^T\mathbf{U}_A^T,\ \mathbf{P}_A^T\mathbf{A}^T\right].$$

Similarly, ne z(B)ioM+MBi=1denotes the embeddings forward s in domain B, which results in an(M+MB)k embedding matrix e ZB computed by

$$\widetilde{\mathbf{Z}}_B^T = \left[\mathbf{P}_B^T\mathbf{U}_B^T,\ \mathbf{P}_B^T\mathbf{B}^T\right].$$

5. PROPOSED METHODOLOGY

All the problems and requirements were overcome in the proposed architecture, through developing a dedicated application for this analysis model. This application will make a major impact among the market researchers like data analyst, data reporter, business analyst, business growth predictors and etc. This application is easy, simple and user friendly to use all types of sentimental analysis model. Here raw data will be given as the input and out the output will a tremendous data reporting model. Some marketers prefer leaving the analysis to dedicated methods, the methods behind sentiment analysis is nothing short from fascinating the various levels of analysis, the detail and the intricacy that make this analysis more accurate when performed by another human rather than a machine. Nowadays, sentiment analysis is an integral part of social listening, although it can also be performed on its own. Sentiment analysis is more than just a feature in a social analytics method. This is a field that is still being studied. While this comment is general, it can be broken down into sentences. This comment has a number of opinions around Simply Measured, both positive and negative. Sentiment refers to the emotion behind a social media mention. It's a way to measure the tone of the conversation is the person happy, annoyed, and angry or neutrals. Sentiment adds important context to social conversations. Without it, measurement of mentions alone could be misleading. If the requirement is to measuring mentions for a company's new product, user might assume a surge in mentions meant it was being well received. After all, more mentions more people talking about the product. Measuring sentiment will help you understand the overall feeling surrounding a particular subject, enabling you to create a broader and more complete picture of the social conversations that matter to you.

5.1 Algorithm

DECLARATION

DS = data set

V = Vocabulary (Extracted from dataset)

C = Categorization :sen count

N = Occurrence

A(Pj | Ni) = Array Deceleration: Pj denotes positive and Ni denoted negative

R = Result

1. Let initiate the process : class index : System.Web.UI.Page, Open database access to receive the dataset: OleDbConnection con;
2. Declare Array : ArrayList reccomment = new ArrayList(); Received data will be stored in an array;
3. String.IsNullOrEmpty (GridView1.Rows[i].Cells[j].ToString())) Confirming the data grid is empty to receive the data;
4. Now LET dataset will be DS;
5. Updating positive word : posi.Add((string)red[0].ToString()); Positive word will be stored as the separate string collection : cmd1.ExecuteReader();
6. Updating Negative word : nega.Add((string)red[0].ToString()); Negative word will be stored as the separate string collection : cmd2.ExecuteReader();
7. Grid updating: GridView2.Rows[i].Cells[j].Text == posi and nega;
8. Merging as a neural Networks = A(Pj | Ni) array deceleration for Pj and Ni;
9. OleDbDataReader rd1 = cmd1.ExecuteReader();
10. postedby = rd[0].ToString();

    pdate = rd[1].ToString();

    ptime = rd[2].ToString();

    shareto = rd[3].ToString();

    post = rd[4].ToString();

    memname = rd[5].ToString();

    comment = rd[6].ToString();

    memdate = rd[7].ToString();

11. Display the DS grid value in the data reader : GridView4.Visible = false; cmd1 = new OleDbCommand(query, con);
12. Data analysis process: Let V = Vocabulary: toterr.Add(rd1[0].ToString()): Spliting word as vocabulary
13. Also Split word will be stored as the C = Categorization : string[] a = toterr [i1] .ToString () .To Upper().Split(' ');
14. Now Analysis will produce the result: Compare Srting[DS,V,C] with A(Pj | Ni)
15. loadarr();

    pcount = posi.Count;

    ncount = nega.Count;

    scount = sen.Count;

16. (sen[i].ToString() == posi[j].ToString());
17. (sen[i].ToString() == nega[j].ToString());
18. Display Result R : input sen = rcount
19. rcount = rank.Count; for ranking: (sen[i].ToString() == rank[j].ToString())

    stat = rd[8].ToString();

20. Display ranking result : string[] a = toterr[i1].ToString().ToUpper().Split(' ');
21. Retrieve data from the DB
22. Display results in charts

6. RESULTS AND FINDINGS

This chapter deals with all the result and the obtain values from the available dataset. According to this thesis, initially all the data will be considered as the input data and processing data. But as per proposed method we need to preprocess the data for a fine tuned result.

Table – 1: Comparison Table

Data	Before Processing	After Processing
Total data	199	140
Number of positive data	100	59
Number of negative data	39	31
Number of moderate data	60	50

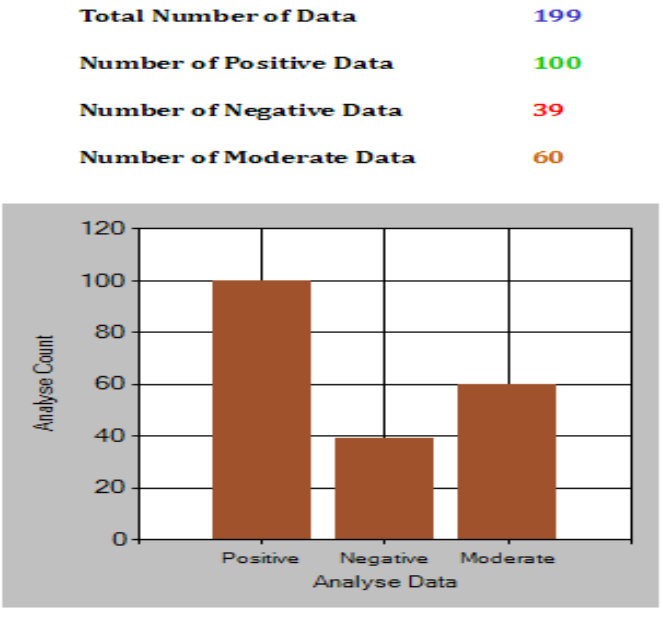


Chart 1 - Analyse Chart

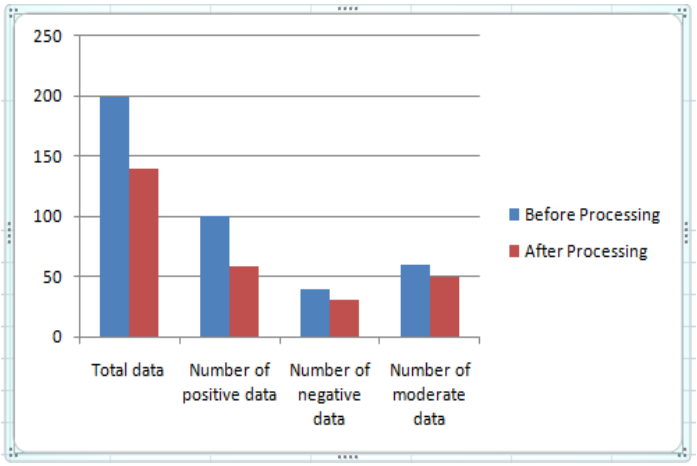


Chart 2 - Comparison Chart

The above comparison chart shows the data change before and after pre-processing.

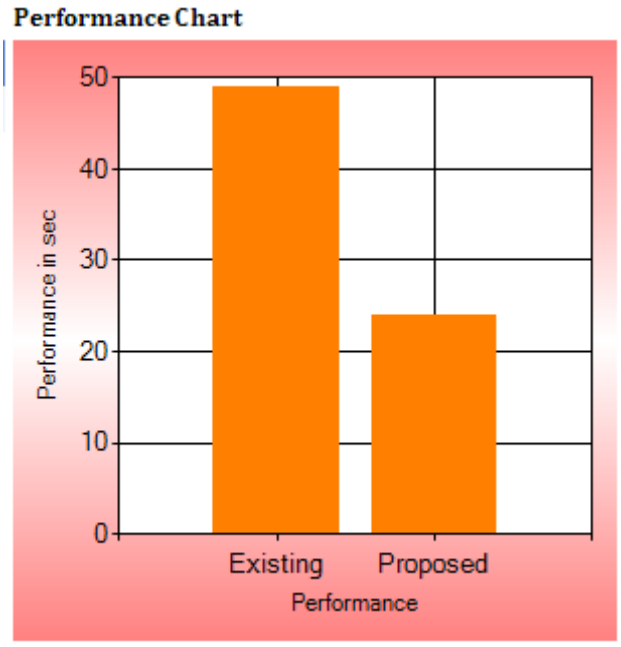


Chart 3 – Performance Chart

While comparing with the existing system, the proposed system has been improved a lot by adding more techniques. Performance chart shows the timing difference between Existing and Proposed system performance.

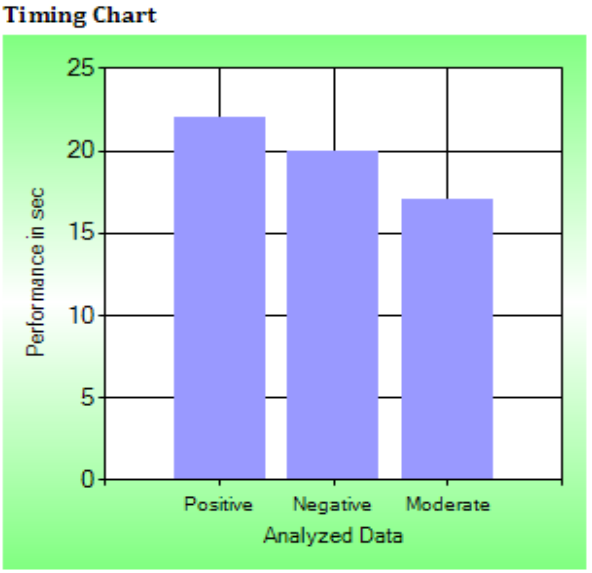


Chart 4 - Timing Chart

Timing chart shows how much time it takes to process each data to separate in sentiment basis.

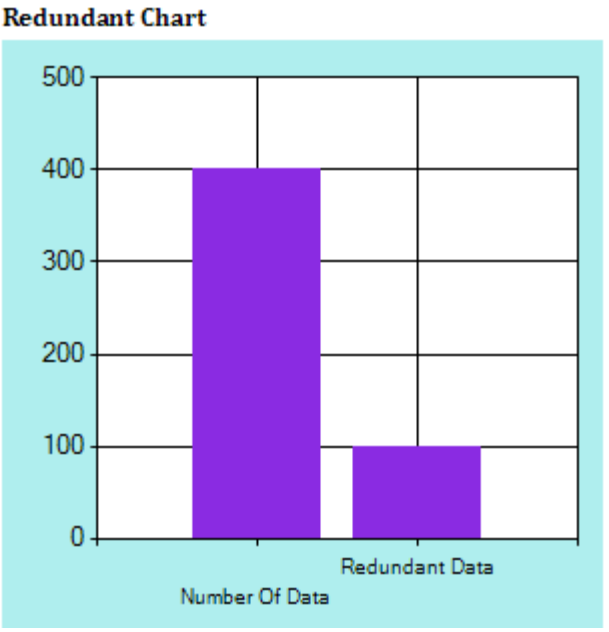


Chart 5 - Redundant Chart

Redundant chart gives number of repeated data occurs. Repeated data affect the accuracy of the process. Removing the redundant data improve our process.

7. CONCLUSION

Classification is very essential to organise data, retrieve information correctly and swiftly. Implementing machine learning to classify data is not easy given the huge amount of sheterogeneous data that's present in the web. Text categorization algorithm depends entirely on the accuracy

of the training data set for building its decision trees. The text categorization algorithm learns by supervision. It has to be shown what instances have what results. Due to this text categorization algorithm, it cannot be successfully classify documents in the web. The data in the web is unpredictable, volatile and most of it lacks Meta data.

REFERENCES

[1] H. Becker, M. Naaman, and L. Gravano, "Learning similarity metrics for event identification in social media," in Proc. 3rd ACM WSDM, Macau, China, 2010.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," J. Mach. Learn. Res., vol. 3, pp. 993–1022, Jan. 2003. TAN ET AL.: INTERPRETING THE PUBLIC SENTIMENT VARIATIONS ON TWITTER 1169.

[3] J. Bollen, H. Mao, and A. Pepe, "Modelling public mood and emotion: Twitter sentiment and socio-economic phenomena," in Proc. 5th Int. AAAI Conf. Weblogs Social Media, Barcelona, Spain, 2011.

[4] Kyunglag Kwon, Yunwan Jeon, Chanho Cho, Jongwoo Seo and In-Jeong Chung, "Sentiment Trend Analysis in Social Web Environment" – in IEEE Transaction on Knowledge and Data Engineering, VOL 28, No 7,March 2017.

[5] D. Chakrabarti and K. Punera, "Event summarization using tweets," in Proc. 5th Int. AAAI Conf. Weblogs Social Media, Barcelona, Spain, 2011.

[6] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," CS224N Project Rep., Stanford: 1–12, 2009.

[7] T. L. Griffiths and M. Steyvers, "Finding scientific topics," inProc. Nat. Acad. Sci. USA, vol. 101, (Suppl. 1), pp. 5228–5235, Apr. 2004.

[8] D. Hall, D. Jurafsky, and C. D. Manning, "Studying the history of ideas using topic models," inProc. Conf. EMNLP,Stroudsburg, PA, USA, 2008, pp. 363–371.

[9] G. Heinrich, "Parameter estimation for text analysis," Fraunhofer IGD, Darmstadt, Germany, Univ. Leipzig, Leipzig, Germany, Tech. Rep., 2009.

[10] Z. Hong, X. Mei, and D. Tao, "Dual-force metric learning for robust distracter-resistant tracker," in Proc. ECCV, Florence, Italy, 2012.

[11] M. Hu and B. Liu, "Mining and summarizing customer reviews," in Proc. 10th ACM SIGKDD, Washington, DC, USA, 2004.