INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 3 ISSUE 2 (2017) PAGES 2153 - 2166
RECEIVED: 01-04-17. PUBLISHED: 26-04-17.

April 26, 2017

ANALYSING THE PUBLIC SENTIMENTAL DATA VARIATIONS ON SOCIAL NETWORKS USING SENTIMENTAL PATTERN RECOGNITION TECHNIQUE

[1]R.deenadevi, [2]Dr.V.Krishnapriya,

[1]Research Scholar, Department of Computer Science, Sri Ramakrishna college of Arts and science for women, Coimbatore,

[2]Head of theDepartment, Department of Computer Science, Sri Ramakrishna college of Arts and science for women, Coimbatore.

1. INTRODUCTION

1.1 OVERVIEW OF DATAMINING

Data mining for software engineering consists of collecting software engineering data, extracting some knowledge from it and, if possible, uses this knowledge to improve the software engineering process, in other words "operationalize" the mined knowledge. For instance, researchers have extracted usage patterns from millions of lines of code of the Linux kernel in order to find bugs. There is a wealth of data-mining and machine learning techniques. Not only they exist, but mature implementations are available and powerful hardware enables techniques to scale to large datasets. There are initiatives, such as centralized datasets and contests to enable scientific comparisons of appropriateness and performance.

The software systems work with inherently complex and difficult to conceptualize. This complexity lead to faults and defects as result increases the cost of software. Software metrics have long been a standard tool for assessing quality of software systems and the processes that produce them. But there are several drawbacks using the metrics as managers mostly rely on metrics which they can easily obtain and work with. Valuable metrics are difficult to obtain and are unavailable. The data generated in software system is huge and not easy to work with. If proper harnessing is done, it can be useful for various software engineering processes and phases.

Due to large and complex data generated day by day at quite a high rate data mining is introduced in software engineering. Software engineers are extensively applying data mining algorithms to various software engineering tasks so as to improve software productivity and quality. However mining software engineering data have several challenges and thus require number of algorithms to effectively mine text, graphs and sequences from such data. Software engineering data includes execution traces, historical code changes, code bases, mailing lists and bug data bases.

Software engineering data contains a wealth of information about a project's status, progress, and evolution. Using well-established data mining techniques, engineers and researchers have started exploring the potential of this valuable data to better manage their projects and to produce higher quality software systems that are delivered within budget and specified time period. Data mining is used by software engineers to previously unknown and unique data statistics within a set of collected data. Data mining tools are useful in predicting the future trends and behaviours which are helpful for engineers to take proactive knowledge driven decisions

1.2 PROBLEM SPECIFICATION

Unsupervised Cross-domain Sentiment Classification is the task of adapting sentiment classifier trained on a particular domain (source domain), to a different domain (target domain), without requiring any labelled data for the target domain .By adapting an existing sentiment classifier to previously unseen target domains, can avoid the cost for manual data annotation for the target domain. To model this problem as embedding learning, and construct here objective functions that capture: (a) distributional properties of pivots (i.e. common features that appear in both source and target domains), (b) label constrains in the source domain documents, and,(c) geometric properties in the unlabelled documents in both source and target domains.

Unlike prior proposals that first learn a lower-dimensional embedding independent of the source domain sentiment labels and next a sentiment classifier in this embedding, our joint optimisation method learns embeddings that are sensitive to sentiment classification. Experimental results on a benchmark datasets how that by jointly optimising the three objectives to obtain better performances in comparison to optimising each objective function separately, there by demon starting the importance of task-specific embedding learning for cross-domain sentiment classification. Among the individual objective functions, the best performance is obtained by (c).Moreover, the proposed method reports cross-domain sentiment classification accuracies that are statistically comparable to the current state-of-the-art embedding learning methods for cross-domain sentiment classification.

1.3 OBJECTIVE

- To develop a social engine to analyze the sentimental data using sentimental data analysis.
- Sentimental data can be analyzed with negative and positive words given in the statement. Word analysis can be done through training the word match using artificial neural network.
- Data Training to be done for data analysis.
- Both trained data and dataset will be analysed using text categorization
- Non hitting positive and negative words are categorized into moderate Category
- Pre-process to be done for removing unwanted data from the dataset

INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 3 ISSUE 2 (2017) PAGES 2153 - 2166
RECEIVED: 01-04-17. PUBLISHED: 26-04-17.
April 26, 2017

1.4 SPECIFIC OBJECTIVE

- Charts and graphs will be generated to the related information
- Instead of reading all the comments commented by the users, chart will represent the graphical information of the effectiveness of the discussion.
- Repeated comments will be separated
- Using Clustering and Classification method, repeated contents will be removed

1.5 MOTIVATION OF THE THESIS

The main motivation of this project is Sensitivity analysis. This is the study of how the uncertainty in the output of a mathematical model or system (numerical or otherwise) can be apportioned to different sources of uncertainty in its inputs. A related practice is uncertainty analysis, which has a greater focus on uncertainty quantification and propagation of uncertainty. Ideally, uncertainty and sensitivity analysis should be run in tandem.

Sensitivity analysis can be useful for a range of purposes, including

- Testing the robustness of the results of a model or system in the presence of uncertainty.
- Increased understanding of the relationships between input and output variables in a system or model.
- Uncertainty reduction: identifying model inputs that cause significant uncertainty in the output and should therefore is the focus of attention if the robustness is to be increased (perhaps by further research).
- Searching for errors in the model (by encountering unexpected relationships between inputs and outputs).
- Model simplification – fixing model inputs that have no effect on the output, or identifying and removing redundant parts of the model structure.
- Enhancing communication from modellers to decision makers (e.g. by making recommendations more credible, understandable, compelling or persuasive).
- Finding regions in the space of input factors for which the model output is either maximum or minimum or meets some optimum criterion (see optimization and Monte Carlo filtering).

Taking an example from Social Networks, in any census based process there are always variables that are uncertain. Like Comments, posts, feeds and other variables may not be known with great precision. Sensitivity analysis answers the comment, "if these variables deviate from expectations, what will the effect be (on the business, model, system, or whatever is being analyzed), and which variables are causing the largest deviations.

INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 3 ISSUE 2 (2017) PAGES 2153 - 2166
RECEIVED: 01-04-17. PUBLISHED: 26-04-17.
April 26, 2017

1.6 SCOPE OF THE THESIS

The Scope of the project is Sentiment Analysis is to define automatic tools able to extract subjective information from texts in natural language, such as opinions and sentiments, in order to create structured and actionable knowledge to be used by either a decision support system or a decision maker. Sentiment analysis has gained even more value with the advent and growth of social networking. Sentiment Analysis in Social Networks begins with an overview of the latest research trends in the field. It then discusses the sociological and psychological processes underling social network interactions.

The research explores both semantic and machine learning models and methods that address context-dependent and dynamic text in social networks, showing how social network streams pose numerous challenges due to their large-scale, short, noisy, context- dependent and dynamic nature.

1.7 SUMMARY

This thesis concentrates on analyzing the public sentimental data variations on social networks using reduced the performance time and the comments have been moderated by sentimental pattern recognition technique.

2. EXISTING METHODOLOGY

2.1 EXISTING SYSTEM

The existing system of the prediction are charts, here the charts will be in the normal format to understand the data. In classification, one is concerned with assigning objects to classes on the basis of measurements made on these objects. There are two main aspects to classification: discrimination and clustering, or supervised and unsupervised learning. In unsupervised learning (also known as cluster analysis, class discovery and unsupervised pattern recognition), the classes are unknown a prior and need to be discovered from the data. In contrast, in supervised learning (also known as discriminate analysis, class prediction, and supervised pattern recognition), the classes are predefined and the task is to understand the basis for the classification from a set of labelled objects (training or learning set). This information is then used to classify future observations. The present article focuses on the unsupervised problem, that is, on cluster analysis, but draws on notions from supervised learning to address the problem.

In cluster analysis, the data are assumed to be sampled from a mixture distribution with K components corresponding to the K clusters to be recovered. Let (X1, ..., Xp) denote a random 1 ×p vector of explanatory variables or features, and let Y∈ {1, ..., K} denote the unknown component or cluster label. Given a sample of X values, the goal is to estimate the number of clusters K and to estimate, for each observation, its cluster label Y. To have data X = (xij) on p explanatory variables (for example, genes) for n observations (for example, tumor mRNA samples), where xij denotes the realization of variable Xjfor observation i and xi = (xi1,...,.xip) denotes the data vector for observation i, i = 1,...,n, j = 1,...,p. to consider clustering procedures that partition the learning set $\mathcal{L}$ = {x1,...,xn} into K

INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 3 ISSUE 2 (2017) PAGES 2153 - 2166
RECEIVED: 01-04-17. PUBLISHED: 26-04-17.
April 26, 2017

clusters of observations that are 'similar' to each other, where K is a user-prespecified integer. More specifically, the clustering $\mathcal{P}(\cdot;\mathcal{L})$ assigns class labels $\mathcal{P}$ (Xi;$\mathcal{L}$) = $\hat{y}_i$ to each observation, where $\hat{y}_i \in \{1,\ldots,K\}$.

Clustering procedures generally operate on a matrix of pair wise dissimilarities (or similarities) between the observations to be clustered, such as the Euclidean or Manhattan distance matrices. A partitioning of the learning set can be produced directly by partitioning clustering methods (for example, k-means, partitioning around medoid (PAM), self-organizing maps (SOM)) or by hierarchical clustering methods, by 'cutting' the dendrogram to obtain K 'branches' or clusters. Important issues, which will only be addressed briefly in this article, include: the selection of observational units, the selection of variables for defining the groupings, the transformation and standardization of variables, the choice of a similarity or dissimilarity measure, and the choice of a clustering method. Our main concern here is to estimate the number of clusters K.

When a clustering algorithm is applied to a set of observations, a partition of the data is returned whether or not the data show a true clustering structure, that is, whether or not K = This fact causes no problems if clustering is done to obtain a practical grouping of the given set of objects, as for organizational or visualization purposes (for example, hierarchical clustering for displaying large gene-expression data matrices. However, if interest lies primarily in the recognition of an unknown classification of the data, an artificial clustering is not satisfactory, and clusters resulting from the algorithm must be investigated for their relevance and reproducibility. This task can be carried out by descriptive and graphical exploratory methods, or by relying on probabilistic models and suitable statistical significance tests.

To validating the results of a clustering procedure can be done effectively by focusing on prediction accuracy. Once new classes are identified and class labels are assigned to the observations, the next step is often to build a classifier for predicting the class of future observations. The reproducibility or predictability of cluster assignments becomes very important in this context, and therefore provides a motivation for using ideas from supervised learning in an unsupervised setting. Resembling methods such as bagging and boosting have been applied successfully in the field of supervised learning to improve prediction accuracy. to propose here a novel re sampling method, Clest, which combines ideas from discriminant and cluster analysis for estimating the number of clusters in a dataset. Although the proposed resampling methods are applicable to general clustering problems and procedures, particular attention is given to the clustering of tumors on the basis of gene-expression data using the partitioning around methods (PAM) procedure.

2.2 Mapping Function

The main strategy of mapping the words and documents to the space is to first compute the word embedding's, and then derive the document embedding's based on the word embedding's by considering the word occurrences. Linear projection is assumed to transform the original feature representation of words to their embedding presentation. Specifically, ad k projection matrix PA issued to map words in domain A to a k-dimensional embedding

INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 3 ISSUE 2 (2017) PAGES 2153 - 2166
RECEIVED: 01-04-17. PUBLISHED: 26-04-17.
April 26, 2017

space R k, while a d h 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 embedding's 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 embedding's forward s in domain B, which results in an(M+MB) k embedding matrix e ZB computed by

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

2.3 DATA ANALYSIS MODEL

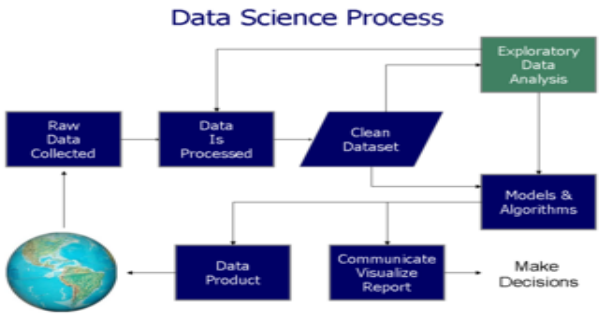


Fig: 2.2 Data Science Process

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.

2.3.1 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

INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 3 ISSUE 2 (2017) PAGES 2153 - 2166
RECEIVED: 01-04-17. PUBLISHED: 26-04-17.
April 26, 2017

population (e.g., age and income) may be specified and obtained. Data may be numerical or categorical (i.e., a text label for numbers).

2.3.2 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.

2.3.3 Data processing

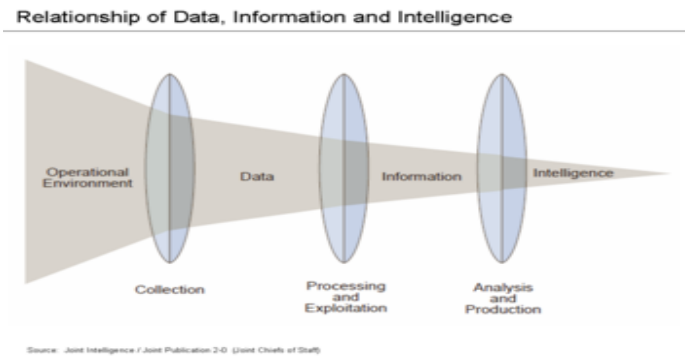


Fig 2.3 Data processing

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 spread sheet or statistical software.

2.3.4 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.

INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 3 ISSUE 2 (2017) PAGES 2153 - 2166
RECEIVED: 01-04-17. PUBLISHED: 26-04-17.
April 26, 2017

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

2.3.6 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 result.

3. PROPOSED METHODOLOGY

3.1 PROPOSED METHOD

Users of decision support systems often see data in the form of data cubes. The cube is used to represent data along some measure of interest. Although called a "cube", it can be 2-dimensional, 3-dimensional, or higher-dimensional. Each dimension represents some attribute in the database and the cells in the data cube represent the measure of interest. For example, they could contain a count for the number of times that attribute combination occurs in the database, or the minimum, maximum, sum or average value of some attribute. Queries are performed on the cube to retrieve decision support information.

In case a database that contains transaction information relating company sales of a part to a customer at a store location. The data cube formed from this database is a 3-dimensional representation, with each cell (p,c,s) of the cube representing a combination of values from part, customer and store-location. The contents of each cell are the count of the number of times that specific combination of values occurs together in the database. Cells that appear blank in fact have a value of zero.

The cube can then be used to retrieve information within the database about, for example, which store should be given a certain part to sell in order to make the greatest sales.

INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 3 ISSUE 2 (2017) PAGES 2153 - 2166
RECEIVED: 01-04-17. PUBLISHED: 26-04-17.
April 26, 2017

A data cube built from m attributes can be stored as an m-dimensional array. Each element of the array contains the measure value, such as count.

The array itself can be represented as a 1-dimensional array. For example, a 2-dimensional array of size x x y can be stored as a 1-dimensional array of size x*y, where element (i,j) in the 2-D array is stored in location (y*i+j) in the 1-D array. The disadvantage of storing the cube directly as an array is that most data cubes are sparse, so the array will contain many empty elements (zero values). Rollup or summarization of the data cube can be done by traversing upwards through a concept hierarchy.

A concept hierarchy maps a set of low level concepts to higher level, more general concepts. It can be used to summarize information in the data cube. As the values are combined, cardinalities shrink and the cube gets smaller. Generalizing can be thought of as computing some of the summary total cells that contain ANYs, and storing those in favour of the original cells.

If the source data is already in a star or snowflake schema, then it already have the elements of a dimensional model:

- Fact tables correspond to cubes.
- Data columns in the fact tables correspond to measures.
- Foreign key constraints in the fact tables identify the dimension tables.
- Dimension tables identify the dimensions.
- Primary keys in the dimension tables identify the base-level dimension members.
- Parent columns in the dimension tables identify the higher level dimension members.

Columns in the dimension tables containing descriptions and characteristics of the dimension members identify the attributes. Also get insights into the dimensional model by looking at the reports currently being generated from the source data. The reports will identify the levels of aggregation that interest the report consumers, as well as the attributes used to qualify the data. While investigating your source data, you may decide to create relational views that more closely match the dimensional model that you plan to create.

4. RESULT AND DISCUSSION

4.1 Compares the results with the existing systems

Even thou the existing system deal with huge data, the obtaining results are very less. No different types of results are produced in the existing system.

According to the sentimental analysis, there are various sentiments are available like

- Positive
- Negative
- Moderate

In the existing system, the system categorizes only the major positive and negative categories only. It dint not look up in depth for the data analysis. The major categories will not give a

refined result and clarity information from the data set. Also in the existing system major techniques are not implemented.As per the existing system:

While comparing with the existing system, the proposed system has been improved a lot by adding more techniques. The techniques are discussed in the abstract and objectives.

Positive Category	Negative Category
+ This product is so **good**	- I hate that product after delivery
+ I am **Happy** to get such a **good** product	- I **won't buy** this product any more

Data set	**Positive**	**Negative**	**Moderate**
This product is more **excellent** and **good**	Yes	No	No
I am more **disappointed** in this product	No	Yes	No
I am more **disappointed** in this product. I **won't buy** this anymore	No	Yes	No
I am **ok** with this product	Yes	No	ok
This is not worthy for 1000 rupees	No	Not	No
After buying the mobile phone, there are many scratches in the panel.	No	No	No sentimental words found. Scratches is the common word.

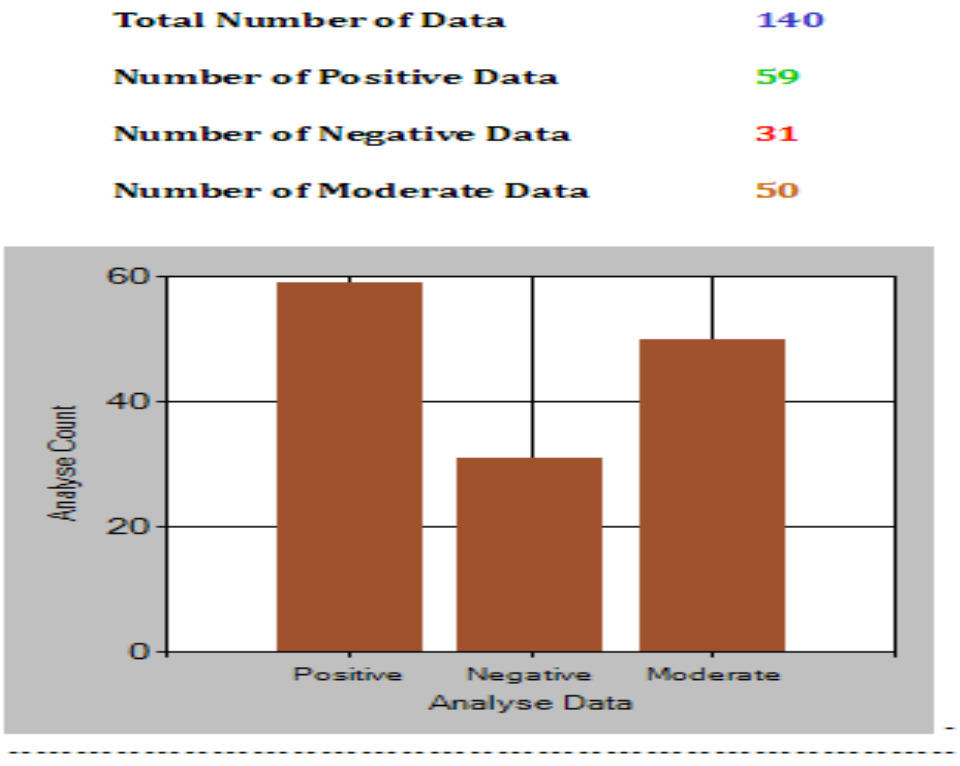
INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 3 ISSUE 2 (2017) PAGES 2153 - 2166
RECEIVED: 01-04-17. PUBLISHED: 26-04-17.
April 26, 2017

Total Data	140
Number of positive data	59
Number of negative data	31
Number of moderate data	50

4.2 DATA DISCUSSIONS

Data 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.

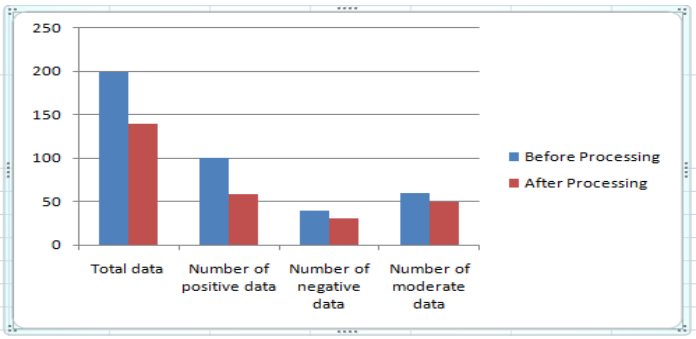
The above mentioned graph shows the unprocessed data,



The above mentioned figure shows the actual result after the preprocessing. The preprocessing has been done using clustering and classification methods. It removes all the repeated users and repeated comments. After the removal of data, the analyses will we executed again in the same execution process.

INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 3 ISSUE 2 (2017) PAGES 2153 - 2166
RECEIVED: 01-04-17. PUBLISHED: 26-04-17.
April 26, 2017

Comparison Chart



4.3 comparison chart for existing and proposed method

5. CONCLUSION AND FUTURE ENHANCEMENT

5.1 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 heterogeneous 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. The way forward for information retrieval in the web, in the future opinion would be to advocate the creation of a semantic web where algorithms which are unsupervised and reinforcement learners are used to classify and retrieve data. Thus the thesis explains the trends, threads and process of the text categorization algorithm which was implemented for finding the sensitive data analysis

5.2 FUTURE ENHANCEMENT

Inductive learning algorithms have been suggested as alternatives to knowledge acquisition for expert systems. However, the application of machine learning algorithms often involves a number of subsidiary tasks to be performed as well as algorithm execution itself. It is important to help the domain expert manipulate his or her data so they are suitable for a specific algorithm, and subsequently to assess the algorithm results. These activities are often called pre-processing and post processing.

The future enhancement discusses issues related to the application of the text categorization algorithm, an important representative of the inductive learning family. A prototype workbench which has been developed to provide an integrated approach to the application of text categorization is presented. The design rationale and the potential use of the system are justified. Finally, future directions and further enhancements of the workbench are discussed.

- Can implement for web based application
- Handshakes with inductive learning algorithm

- Improvisations can be done in the performance evaluation
- Prediction can be done for all kind of diseases
- In case of huge range of data set, data load balancing can be done

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, "Modeling public mood and emo-tion: Twitter sentiment and socio-economic phenomena," inProc. 5th Int. AAAI Conf. Weblogs Social Media, Barcelona, Spain, 2011.

[4] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market,"J. Comput. Sci., vol. 2, no. 1, pp. 1–8, Mar. 2011.

[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.

[12] Y. Hu, A. John, F. Wang, and D. D. Seligmann, "Et-lda: Joint topic modeling for aligning events and their twitter feedback," inProc. 26th AAAI Conf. Artif. Intell., Vancouver, BC, Canada, 2012.

[13] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao, "Target-dependent twitter sentiment classification," in Proc. 49th HLT, Portland, OR, USA, 2011.

[14] J. Leskovec, L. Backstrom, and J. Kleinberg, "Meme-tracking and the dynamics of the news cycle," in Proc. 15th ACM SIGKDD, Paris, France, 2009.

[15] C. X. Lin, B. Zhao, Q. Mei, and J. Han, "Pet: A statistical model for popular events tracking in social communities," in Proc. 16th ACM SIGKDD, Washington, DC, USA, 2010.

[16] F. Liu, Y. Liu, and F. Weng, "Why is "SXSW" trending? Exploring multiple text sources for twitter topic summarization," inProc. Workshop LSM, Portland, OR, USA, 2011.

[17] T. Minka and J. Lafferty, "Expectation-propagation for the gener-ative aspect model," inProc. 18th Conf. UAI, San Francisco, CA, USA, 2002.

[18] G. Mishne and N. Glance, "Predicting movie sales from blogger sentiment," inProc. AAAI-CAAW, Stanford, CA, USA, 2006.

[19] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, "From tweets to polls: Linking text sentiment to public opinion time series," inProc. 4th Int. AAAI Conf. Weblogs Social Media, Washington, DC, USA, 2010.

[20] B. Pang and L. Lee, "Opinion mining and sentiment analysis," Found. Trends Inform. Retrieval, vol. 2, no. (1–2), pp. 1–135, 2008