Rules Based Text Analysis to Monitor and Control Multiparty Access in Online Social Networks

R.Prem Kumar, A.Rengarajan , S.Hariharan

Abstract: ***Text Analytics always depend on following the rules. Computers follow rules correctly. They do as they commanded continuously and efficiently. This is essential because a good text analytics program implements a large set of rules. People are affected from their discriminations grew in culture, education, age, gender and other environmental factors. People are flexible; they learn, adapt and change. A computer may be accurate to the rules and these rules can be evaluated on occasion and peoples can change these rules to redefine the results. The most in-discriminated person in the society is still human and therefore make mistakes. The word "hat" cannot be interpreted as "cat" and it is not easy to correct due to the error's inconsistency and not able to predict all of the factors that lead to this error. One cannot recall everything - rules, definitions, cross references and relationships Discoveries in text analytics are more complex because of the problems of people such as consistency, objectivity, subjectivity and depth. But computers have been redefined to discover patterns in data. The amount of text created every day in social media needs monitoring and evaluation of text. A text analytics engine can immediately reveal the relationships between terms of feeling for or expressions of hostility against a given brand or product and estimate the significance of any change in those relationships.***

I. INTRODUCTION

1.1 Text Analytics

Testing public conversations about a brand or a product needs more lifting from text analytics tools, and recognizing the technical features of text analytics is essential for modern marketing executives. Text analytics attempt to interpret meaning from the written word. This is difficult because human communication depends on context. The inability to maintain with the level of communication is surrounded by the emerging of social media. The need to recognize conversations happening in the society is hard. Blog posts, tweets, YouTube comments represents what is on the minds and in the hearts of the public. But these options are not possible. Text Analytics is a promising solution and useful for decades

Text Analytics apply to classify or extract information from text with respect to context. The objective of Text Analytics is to form a subject-matter framework based on the differences in opinion, meaning and attitude of the contexts. It begins through answering a few questions that are not able to answer about the text that have to be analyzed. A comment on a blog post is not a separate message and should be perceived in context with the original post. Sentiment Analysis is a category of Text Analytics that identifies the polarity (positive, negative or neutral) of the content.

Revised Manuscript Received on June 14, 2019

R.Prem Kumar, Research Scholar , Saveetha University ,Department of CSE, , Chennai ,India.

A.Rengarajan Department of Computer Science and Engineering, Vel Tech Multi Tech Dr.RR Dr.SR Engineering College,Chennai ,India.

S.Hariharan, Research Scholar, Saveetha University ,Department of CSE, , Chennai ,India.

Text Analytics is a category of datamining , that tries to find textual patterns from large non-structured sources. Text Analytics, also known as Intelligent Text Analysis, Text Data Mining or Knowledge-Discovery in Text (KDT), is the process of retrieving non-trivial information and knowledge from unstructured text. Text Analytics is similar to data mining. Text Analytics can cover unstructured or semi-structured data sets such as emails, full-text documents and HTML files, blogs, newspaper articles, academic papers, etc. Text Analytics is an interdisciplinary field that is used on information extraction, data mining, machine learning, statistics and computational linguistics.

In machine learning context is necessary. Therefore the machine needs an ontology. An ontology is the combination of relationships and definitions within a subject that allows to manipulate the machine. With context and meaning, ontology management renders formal labelling and information categorization.

The computer system does not know about the texts and the relationship between words. So the computer system should be trained about the relationship between words by the humans. Ontology management allows a computer to manipulate content in context.

1.2 Trusting Machines to Understand Peoples

Sentiment analysis is the most critical task in text analytics. Peoples frequently have problems in recognizing each other even when communicating face to face. Textual misunderstanding is more frequent due to the absence of facial expressions or vocal clues.

Because of this legal writing is excessively heavy and attempts to avoid misunderstanding. The solution for this is the peoples working closely with machines. Computers can manipulate whether a given phrase is positive or negative with some degree of confidence. Peoples with particular domain expertise can analyze low confidence results and teach the machine the way of grading these low confidence results. Over time, the computer grasps the expert's aspect and tend to more precise and useful.

1.3 Needing Machines to Understand Peoples

Text Analytics still had a drawback, even if Text Analytics is performed by intelligent and well trained people.

1. Consistency

Text Analytics always depend on following the rules. Computers follow rules correctly. They do as they commanded- continuously and efficiently. This is essential because a good text analytics program implements a large set of rules.

2. Objectivity

People are affected from their discriminations grew in culture, education, age, gender and other environmental factors. People are flexible; they learn, adapt and change. A computer may be accurate to the rules and these rules can be evaluated on occasion and peoples can change these rules to redefine the results.

3. Subjectivity

The most in-discriminated person in the society is still human and therefore make mistakes. The word "hat" cannot be interpreted as "cat" and it is not easy to correct due to the error's inconsistency and not able to predict all of the factors that lead to this error.

4. Depth

One cannot recall everything - rules, definitions, cross references and relationships. Text Analytics for a particular domain may cover many designations, several classifications and unknown situational definitions.

5. Discovery

Discoveries in text analytics are more complex because of the problems of people such as consistency, objectivity, subjectivity and depth. But computers have been redefined to discover patterns in data.

The amount of text created every day in social media needs monitoring and evaluation of text. A text analytics engine can immediately reveal the relationships between terms of feeling for or expressions of hostility against a given brand or product and estimate the significance of any change in those relationships.

1.4 Social Media and Text Analytics

Several communications have provided the public more influential power. Public thought is more believed and given importance than a company's promotional declarations. Text Analytics is the only way for monitoring, comprehending and involving in public discussions. While the various technologies may be confusing, the use of these technologies are essential for measuring the success of social media investments and keeping better relations and a competitive edge.

It is more tedious to measure how many people is exposed to a particular brand than multiplying the number of the commercials by the number of users. The number of people who read about a particular brand cannot be measured

Several names for the same thing is difficult to understand. Because of this referential disambiguation is required to parse the names. Peoples are top at this activity, but with sufficient readers, computers should be utilized to understand, find and tabulate this new form of brand experience.

Once the number of comments for a given product or brand has been tallied, the total number of exposures should be separated by a subject and then by sentiment. This is the situation in which text the analytics in social media becomes hard for customer care and customer service.

With the increase in responses of the survey, comments and online feedback of the customer service representative, customer service departments are the first to understand the importance of text analytics. Each customer service employee has a view about the most often problems noted and their relative severity based on frequency and customer tone of voice. But only a complete analysis can validate those feelings

The aim of a company is to create a brand image in the public minds through advertising alone. Advertising is only a category of brand-building process. Consumers are performing more considered purchases as opposed to impulse purchases. That research necessarily includes expert reviews with blogs, friends with recommendations and customers with suggestions

The online and written communications' polarity about any product or service have become essential statistically and influence purchases. The people saying about a person online is becoming more crucial than advertising through believing friends, experienced customer. An organization should stand on the top of public thought to answer quickly and correctly.

II. EXISTING WORK

2.1 An Empirical Study on Text Analytics The big data has real issues of usability. Especially, information retrieval's major part is the great experience in big data. The main aim is finding or developing the best cost effective and reliable techniques to extract the values from more terabytes and peta bytes of available data. So big data analytics is important. Conventional analytics focuses on structured data but not suitable for unstructured data with large volume. Text analytics is the technique of extracting significance from the unstructured text for identifying transformations and patterns. [2]

2.2 Visual Analytics of Text Streams through Multiple Dynamic Frequency Matrices

A Visual Analytics tool which provides exploration tasks and situation awareness for text streams is presented. For attaining this objective, a data model for encoding the streaming text is designed in multiple dynamic frequency matrices. The visualizations comprises two dynamic Theme Rivers. These two dynamic Theme Rivers enables real time expedition of most of the aspects retrieved from texts stored in long term and short term buffers. Also the messages' geographical location can be visualized. [3]

2.3 Multi-Model Semantic Interaction for Text Analytics

Semantic Interaction provides an inherent communication technique between difficult statistical models and human users. Semantic interaction focus instead on manipulating the spatialization directly through restricting the users that manipulate the model parameters. But this semantic interaction technique is not considerably scalable for several text documents. To address this issue, the multi-model semantic interaction concept is presented in which semantic interactions can trigger several models at different levels of data scale, allowing the users to manage large data problems. [4]

2.4 DSSM with Text Hashing Technique for Text Document Retrieval

Text documents are the source of saving the information that may be personal or general. Nowadays, text documents are producing at enormous speed and the requirement is to process the data immediately for upgrading the search engine. A system which enhances the process of retrieving the information from text documents in search engine from unstructured data is presented. [5]

2.5 Integrated Visual Analytics Tool for Heterogeneous Text Data

A Java-based Visual Analytics Tool which reads a set of various text data sources and retrieves the relations, main keywords and events from the text data through the use of ontology and language processing techniques. Finally this Java-based Visual Analytics Tool provides a combined and intuitive search interface to users to support efficient and strong investigation for huge and difficult data set. [6]

High Performance Text Analytics

High Performance processing of text data is significant. A library for high performance Text Analytics is presented. This library facilitates programmers for mapping text data to a heavy numeric representation that can be managed efficiently. The library incorporates three optimizations for performance i)text data's effective memory management ii) parallel computation on relative data structures which map text to values iii) optimization of the relative data structure's type depending on the program context. [7]

2.6 Word Cloud Explorer: Text Analytics based on Word Clouds

Word clouds have become a direct and visually appealing visualization technique for text. They are used in different aspects to provide an overview through extracting down the text to words which emerge with highest frequency. Especially this is performed in a stable way as pure text summarization. For tasks with general text analysis, the application of word clouds are described. [8]

2.7 Text Analytics of Web Posts' Comments Using Sentiment Analysis

Social networking websites are the most general platform to share one's suggestion. The posts and the comments on one's wall such as facebook makes people to decide under several situations based on other peoples suggestion. These comments affects one's thought while deciding because of more likes and comments. The outcome is that the user likes a specific post and puts a negative comment for it. To address the issue, the sentiment analysis approach is presented. This sentiment analysis identifies the original popularity of a post on social networking websites. This sentiment analysis approach gives the original statistics to support the decision i.e. if the idea or thought suggested by the user facilitates the post or not. With the help of a lexicon based sentiment analysis approach, the comments are analyzed. [9]

2.8 Short Text Opinion Detection using Ensemble of Classifiers and Semantic Indexing

An Ensemble system of Classifiers and Semantic Indexing techniques are applied to Short Text Opinion Detection. Two major problems make difficult to demonstrate the classification algorithmic rule. The low number of characteristic could be removed per massage & that messages are filled with the symbols and abbreviations and idioms. Estimate the classifiers ensemble along with the nine public, non-encoded datasets and real datasets. [65]

III. PROPOSED METHODOLOGY

Text Analytics is the method of extracting useful information from the unstructured text input. In case of web domain, text analysis is related to the process of performing web content mining by utilising the text and hypertext from the provided document.

3.1 Text Analytics

To generate knowledge from the unstructured data, initially the input data is converted into the text data in a manageable format. Feature selection techniques are used to store only the significant features. In certain criteria, the information conveyed by the phrase posses more than one word, and then the text representation models like N-gram and LSTMN are used to sequence the data as features. Text analytics is the process that utilises text-mining applications over the collected data set. Due to the development of big data approach the text analytics approach, become more common among the data scientists.Deep learning algorithms are introduced to explore the unstructured data from the big data platform.

3.2 Text mining and analytics

Text mining and analytics provides valuable insights for the organisations from the customer emails, call logs, survey comments, social media post and other sources of text related data. The text analytics are widely used for virtual agents and AI chat bots to generate automated reply to the queries raised by the customers. The process of text mining is similar to that of data mining, the text mining focuses over the text instead of structured data. The initial step of text analytics is organised and structure the data in specific format so that it can be applied to quantitative and qualitative analysis.

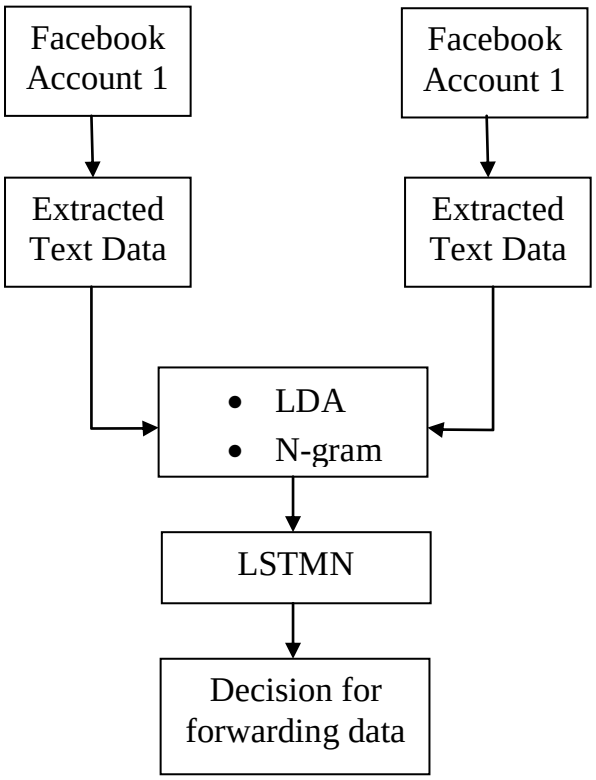


Figure 3.1 Block Diagram

Natural Language processing (NLP) is used to perform this typical function.

In traditional NLP models are developed based on the rule based or statistical methods.

These methods provided a flow to analyse the data set. The researches on data analysis focuses on less supervised deep learning models and handling bigger data sets. Neural network is used to perform deep learning over larger data sets. It processes the data sets in iterative way. This provides more supporting functions than other traditional machine learning techniques. **Figure 3.1** shows the block diagram to perform text analysis over data collected from Facebook account.

4. N-Grams

In document processing, text classification is a basic task followed to handle tremendous amount of document in digital form. Presence of textual errors like grammar errors and spelling errors in the digitalised document makes data handling more difficult.

4.1 Document processing

The character recognition errors affect the performance of the text classification algorithm. Thus, the reliability of the text classification algorithm should be high to overcome the errors. To overcome this drawback N-gram based text categorisation algorithm is implemented to overcome the textual errors in the document. N-gram is represented as a n-character slice of the larger string. In general the N-gram can be framed for any set of word pair in the string. For example N-gram can be developed using first and third words of the string. The string was sliced into several overlapping N-grams. The N-grams can be developed with various lengths simultaneously. The beginning of the word and ending of the word is represented with a blank space. The string with the length K can be classified as K+1 bi-grams, tri-grams and quad-grams and so on. N-gram based modes are very useful in processing the ASCII inputs with noises in domains like text retrieval, interpreting the postal addresses and other applications involve natural language processing. The efficiency of implementing N-gram algorithm is achieved due to its natural approach of this model. The input string is divided into several parts. The error presented in the document will affect only the smaller part classified by the N-gram algorithm. The textual errors are avoided by comparing the parts with each other to identify the similarities of each parts.

IV. LSTM

Long Short-Term Memory network is developed as an extension of recurrent neural networks used to extend the limit of their memory capacity. Thus it is useful to perform neural network functions over the data set with larger time lags. The fundamental units of LSVM is utilised in the layers of Recurrent Neural Network (RNN) thus it is generally represented as LSTM network. LSTM is used to store the inputs of the RNN for longer time period. LSTM maintains a memory cell. It reads, write and delete the data from the memory cells similar to the computer memory. The Memory cells are developed as a gated cells thus the cell decide whether to save the data or delete data from the memory cell. The gated cell makes decision based on the importance of the data set.

5.1 Long Short Term Memory Ttext classification:

A social media post about weather in particular location is processed. The data's about weather are sequential. The dependencies between the words classify with Recurrent Neural Network (RNN) model. Initially, the data is converted to numeric with word embedding. The word embeddings expresses words, which describe the same things as similar vectors. IN addition, the relationship between the words express through arithmetic vector. In LSTM model,

i) The data are imported and preprocessed.
ii) The word embedding converts word to numeric vectors to form word vector.
iii) The word vector apply as input to train LSTM network.
iv) The LSTM network classifies the texts.

The **figure 5.1** shows the histogram of word vectors class distribution with LSTM network.

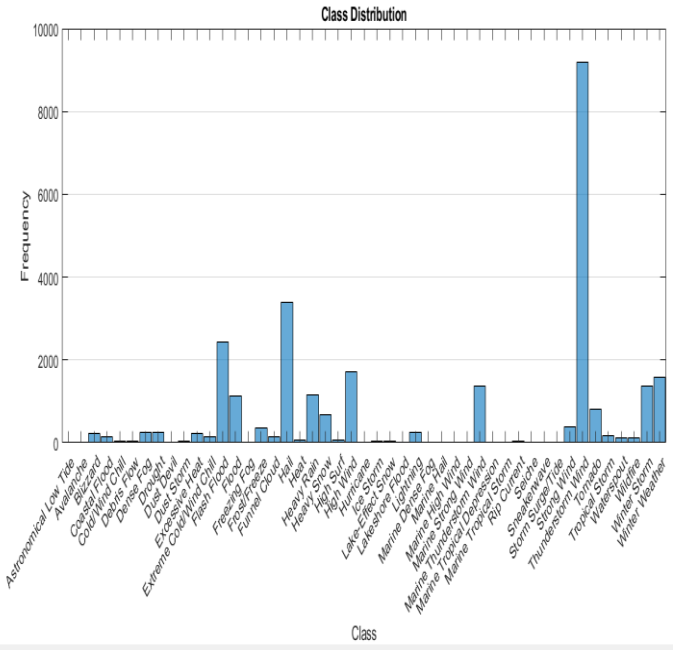


Figure 5.1: Class distribution for different words in social media post.

From the **Figure 5.2** it is clear that certain classes of words are imbalanced. The imbalanced data produce output below the acceptance rate. The balanced data form by removing minimal instances of word occurrences in class distribution. The resultant text datas are extracted and represented as word cloud as shown in figure 6.5.

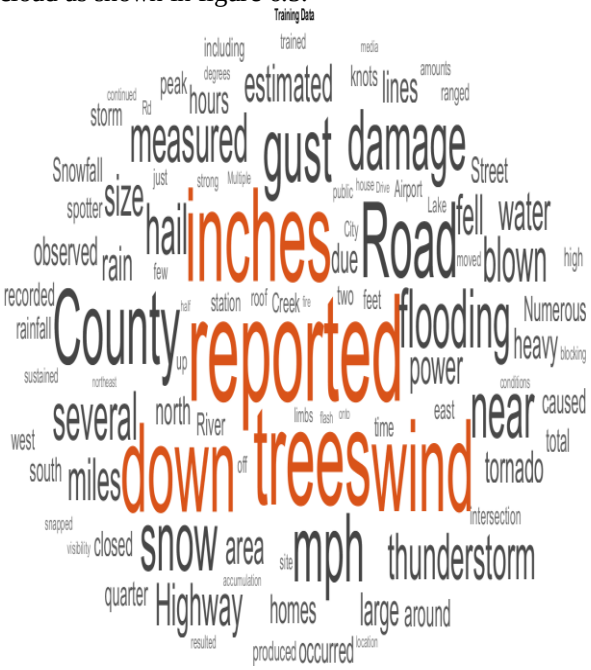
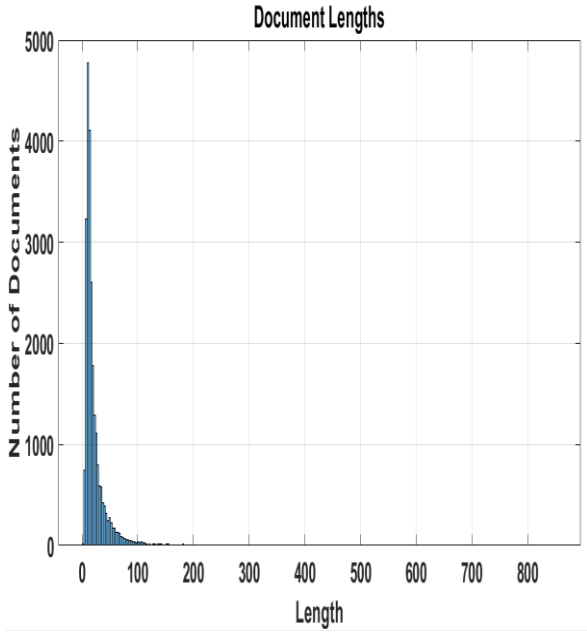


Figure 5.2 Word cloud for weather report post.

Similarly, the **figure 5.3** shows the output of a document processed with LSTM network. In document processing, the words vectors are formed by truncating to produce similar length of word sequences.



N-Gram 5.3: Number of documents and length analysis.

6. N-gram Counts

6.1 Generating N-gram Counts from facebook details:

The traditional methods to generate the frequency of the N-gram are to measure the number of occurrence of N-gram in the provided facebook details. Different data structure formats like trigram and 4-gram, and several models are developed to perform the N-gram counts. The size of the memory in the processing unit is limited thus the methods used to count N-gram generates limited amount of N-gram which can fit in the provided memory space. Measuring the counts from N-gram in a larger database is more important for major algorithms. The accuracy of the results of the algorithms increases based on the amount of available data sets.

Utilising the facebook details as a data set is widely discussed to measure the N-gram from the provided data set. The amount of data set collected from facebook details is higher thus, a memory efficient model is to be used to identify the frequency count from the hit count which is returned as a result for every request. The hit count is utilised to generate joint frequency and marginal counts for the bigram model. The concept can be carried to generate marginal count and joint frequency for the trigrams and 4-grams generated from the data set of facebook details. The Log Likelihood Ratio is performed to measure marginal count and frequency count.

V. CONCLUSION

Text analytics or Natural Language processing are referred as Artificial Intelligent methods which helps the user to extract key contents from the text data set. Text analytics promotes machine-learning capability for the research areas like clinical and drug discovery sectors. The Text analytics is also referred as text mining which process large text data sets and provide the keywords to understand the information provided by the data set. In addition to it text analytics convert the unstructured data set into a structured form which can be used for further analysis. The text analytics identifies the facts, relationship, asserts from the larger data sets. This information is extracted and converted into a structured data for visualizing, analysing and integrating to form a structured data and refine the information using machine-learning systems.

Text analytics was performed through two methods they are Linguistic rules and Machine learning systems. Linguistic Rules functions based on the rule based pattern-matching model based on simple Boolean keywords. And also performed by creating a complex model developed by the field experts. Linguistic rules are applied to perform faster analysis. Machine Learning approach applies patterns over the text dataset. Statistical methods are applied to compare the documents to one another and generate the most important text information's from the larger text corpus or database. The Machine learning approaches are ranged from simple to complex implementations, it gather the valuable informations and distinct patterns from the data set provided. The data set collected from social media like facebook was used to perform text analytics to avoid unwanted sharing of public post to the unknown users. LDA, N-Gram, and LSTM algorithms are used to extract informations and generate patterns to compare the relationship between the post and the user account. The performance of the algorithms were analysed and decision-making was performed to forward the shared data to the user based on the details collected from the users face book details.

REFERENCES

1. B. Liu, "Handbook Chapter: Sentiment Analysis and Subjectivity. Handbook of Natural Language Processing," Handbook of Natural Language Processing. Marcel Dekker, Inc. New York, NY, USA, 2009.
2. K. Dave, S. Lawrence, and D. M. Pennock, "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews," in Proceedings of the 12th international conference on World Wide Web, 2003, pp. 519–528.
3. J. A. Richmond, "Spies in ancient Greece," Greece and Rome (Second Series), vol. 45, no. 01, pp. 1–18, 1998.
4. J. Thorley, Athenian democracy. Psychology Press, 2004.
5. D. D. Droba, "Methods used for measuring public opinion," American Journal of Sociology, pp. 410–423, 1931.
6. "Public Opinion Quarterly." [Online]. Available: //poq.oxfordjournals.org. [Accessed: 02-Dec-2016].
7. R. Stagner, "The cross-out technique as a method in public opinion analysis," The Journal of Social Psychology, vol. 11, no. 1, pp. 79–90, 1940.
8. L. Knutson, "Japanese opinion surveys: the special need and the special difficulties," Public Opinion Quarterly, vol. 9, no. 3, pp. 313–319, 1945.