JCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-6, June 2019 E-ISSN: 2347-2693

Frame Tone and Sentiment Analysis

S.V. Balshetwar[1*], R.M. Tuganayat [2], G.B. Regulwar[3]

[1]Dept of CSE, Satara College of Engineering and Management, Shivaji University, Kolhapur, India
[2]Shri Shankarprasad Agnihotri College of Engineering, Wardha Sant Gadge Baba Amravati University, Amravati India
[3]Dept, PLITMS, Buldana, Maharashtra, India

Corresponding Author: balshetwar.satara@gmail.com, Tel.: 9881697956

DOI: https://doi.org/10.26438/ijcse/v7i6.2440 | Available online at: www.ijcseonline.org

Accepted: 16/Jun/2019, Published: 30/Jun/2019

Abstract— Electronic text on internet can be used for many online activities similarly it can also be used for social movement activities. The electronic text through social movements can also be used to describe an issue, place blame, identify victims, propose a solution and appeal readers to take action on it. Texts such as these are framing documents. Framing is a unique concept in sociology & political science in which people interpret information and speak in favour or claim. Online communities are using frames on social media for their good or bad goals. Thus framing and contents in it have cumulative effect on sentiment of people which needs to be studied. Sentiment analysis explores attitudes, feelings, and expressed opinions regarding products, topics, or issues. The research presented here proposes a framework that applies statistical methods in text analytics to extend research in framing process to find sentiments expressed by people in frame.

In research work, first phase is to pre-process text; it uses supervised machine learning methods that create a tone based term matrix. Second phase discover distinct patterns that characterize prominent frames by classifying the corpus into frames and non-frames. last phase aims to classify frames more specific into motivational, investigative and predictive on the basis of sentiments expressed in them so as to find out threat, cause or solution for an issue. The research presented here aims to develop a tool that will help social movement organizations and concerned authorities to portray issue and helps in organizing activities properly.

Keywords— Sentiment, Tone, Frame, Context- Concept Quadruple.

I. Introduction

In this world of information technology the ever growing thing is data, it can be set of facts, observations anything in the form of digitized manner. Now a day the data has been in trillion gigabytes. Every part of this technological world is flooded of data today. Almost 80% of this data is unstructured; because the data comes from various new sources like device logs, server logs, twitter feeds, chat data, blogs, web pages, emails, and social media content. This makes a huge collection of text data which is created by humans to express themselves to others, so it has become an important source of data that may contain valuable information.

Every person making use of social media has abundant of data on internet, companies make use of this data to analyze many things right from sentiments of people in purchasing a product to fraud detection or technically for securing the companies data, this data is also used to drive social movements.

Social movements occur when caucus of people stand against injustice or issue which has a socio political background. Similar to but distinct from terrorism but can be named as intruder/extruder terrorism. It is nowadays used widely to mobilize people using social media. Online communities are using social media to raise complex societal issues for political goals or to really put forward social issues like public health and climate change, which is catching attention of researcher and media. Social movements play an important role in bringing changes in society in political, educational, health and other areas. Collective actions and emotions of people are portrayed through them.

Suppression of peoples demand through authorities or their agents using violence or other measures is the main reason to raise a social movement. Social movements are redemptive, resistance or revolutionary. Redemptive movements are religious and are total personal transformation, resistance movements are those which are used against legal bodies to disrupt civil order and stability

and revolutionary movements are those which are carried out for revolution.
In past few decades, with the rapid economic growth, our society has entered crucial period of economic reform but social conflicts continued to accumulate and spread over, with the transformation of economic and social, bought some new problems, and continued to show new conflicts. When some conflicts spread through the internet, users on internet used the internet to fully express their beliefs, attitudes, opinions and emotions, thus they become consent in a very short period of time. Fermented emotions and induced actions have affected our social life.

The strong electronic public opinion could push the government, the enterprise and even individuals to dangerous places in this process and could trigger a deadly crisis. Such text can raise serious challenges in social harmony and strength. Emergency response capacity for the government, enterprises and other relevant agencies can be tested. So many mass incidents are growing and mobilized through internet text, so there is a need of some tool/mechanism for crisis finding and assessment. When internet makes use of strategic devices for presenting prominent characteristics about an issue, through piece of text that makes use of specific words, images for conveying meaning about an issue, it is called framing.

"Framing refers to the way a story is told and to the way these cues or stories, in turn, trigger the shared and durable cultural models that people use to make sense of their world."

Framing provides the means for SMOs to inform others of the issue, change the way in which others think about the issue, and invite contribution to act on the issue. In this context, framing refers to these actions of SMOs. The goal is to change views of public towards social issues, and ultimately the manner in which they act over an issue.

Studying media or peoples interpretation on particular social issue can offer impend into public knowledge and their attitudes on the topic ascend. The social movement organization does make use of internet to mobilize people's sentiments and take opinion cum action on an issue. Thus to understand social communication and influence require the study of text.

Analysts are assailed with more text than they can possibly read. In response, research into the processing and analysis of text has flourished. The need to find information from raw text has fuelled the development of information retrieval. The need to discover meaning, themes, attitude in a corpus of documents has led to the development of area called sentiment analysis using text analytics techniques.

Some of the most promising and challenging areas of development in text analytics seek to use NLP to understand what people really mean when they use a set of words. Sentiment analysis, involve analysis of text for estimating how favourable a review is for an activity. Text has been used to detect emotions in the related area of affective computing. It is particularly necessary to understand what social influence and information is transmitted through online content about complex social issues.

Sentiment expressed by people on a particular issue carry a lot of hidden meaning and emotions, so it needs to be studied. "Until now researchers have worked on the sentiment analysis of sentiments expressed by people about products and services. This research focuses on sentiment analysis of social issues."

Paper is organized as follows, Section I contains the introduction of sentiment analysis, framing and its tone, Section II contain the related work of sentiment analysis and framing, Section III contain the methodology with architecture of system and flow of working, Section IV contain the results and discussion V Concludes research work with future directions.

II. RELATED WORK

In the paper we have done the literature review of sentiment analysis, concepts related to sentiment analysis like extraction, classification, comparisons, summarisation, visualisation, and some linguistic factors affecting sentiment analysis along with approaches and techniques for sentiment classification.

A. Sentiment analysis:

Up to now, researchers have focused on research that has been done on various related tasks to sentiment analysis. Although most of the research not all has been done on products [1], but some work is also done in the area sentiment analysis of social issues [2].

Various unsupervised techniques and supervised techniques such as SVM, Nave Bayes Maximum entropy are used by the researchers in the area of sentiment analysis. Bag of words is most common approach used for sentiment extraction by traditional classification techniques even work is done in area of sentiment clue and pattern extraction [2]. Some researchers have also focused on finding meaning or description of sentiment like feature based sentiment analysis performed by Hu and Liu but a less attention given on work related to defining structure for opinion.

B. Psychology of emotion

With few exceptions, current sentiment analysis methods aim to detect sentiment one-dimensionally, giving a score on a range from negative to positive sentiment. While this pragmatic approach proves useful, Jack et al. (2014) [77] speculated that there are four basic emotions., Ekman (1992) names six, and Plutchik (1991) identifies two additional basic emotions in humans, 90 such classifications being given over the past century, as noted by Plutchik (2001).

C. Sentence subjectivity Analysis

Subjectivity is the linguistic expression of belief, emotion, evaluation, or attitude (Wiebe 1994). The task of classifying a sentence as subjective or objective is often called subjectivity classification in existing literature.

Given a sentence s, two sub-tasks are performed:

1 Subjectivity classification: Determine whether sentence s is a subjective sentence or an objective sentence.

2 Sentence-level sentiment classifications: If sentence s is subjective, determine whether it expresses a positive, negative or neutral opinion.

Many papers have been published on subjectivity classification and sentence-level sentiment classification. For subjectivity classification, it applied supervised learning. For sentiment classification of each subjective sentence, but with many more seed words, and the score function was log-likelihood ratio. The same problem was also studied considering gradable adjectives, and using semi-supervised learning. In [3], researchers also built models to identify some specific types of opinions.

Extracting subjective sentences from reviews is major research area where Pang and Lee designed and subjectivity detector framework that uses min-cut graph technique which uses context information like sentence proximity to construct the graph and work states that it outperforms SVM and Naive Bayes in sentence subjectivity analysis.

More recently, (Wiebe & Riloff, 2015) introduced a bootstrapping method that learns subjective patterns from un-annotated documents. For this method, the authors needed to define an initial set of rules that were manually annotated and, to this end, required linguistic expertise (Scheible & Schütze, 2012). (Riloff et al., 2006) has also proposed a method that defines subsumed relationships between different elements (unigrams, n-grams and lexicon-syntactic patterns). The idea is that if an element is subsumed by another, the subsumed element is not needed, something that can remove redundant elements in the subjectivity classification (Liu, 2012).

Considering the effect of dynamic adjective and its feature like gradability and semantic orientation Wiebe and et. al. have identified that the adjectives are strong predictors of subjectivity.

D. Sentiment Classification:

Sentence can be classified as positive, negative or neutral depending upon the opinion words present in it. Sentiment classification is a special task of text classification whose objective is to classify a text according to the sentimental polarities of opinions it contains (Pang et al., 2002), e.g., favorable or unfavorable, positive or negative.

Sentiment classification actually recognizes sentiment orientation of a document on the target. Sentiment classification is about subjectivity, polarity, and polarity strength of an opinion text as per Binali et al.

Sentiment analysis focuses on evaluating attitudes and opinions on a topic of interest using machine learning techniques. The definition of sentiment analysis in data mining can be described from two perspectives: functional and operational. The functional aspects focus on practical uses of the method. For instance, Liu [4] describes sentiment analysis as a process that categorizes a body of textual information to determine feelings, attitude and emotions towards a particular issue or object. The definition points to the way sentiment analysis works and describes the outcome of polar classification.

Sentiment classification can be binary or multiclass. While in the binary sentiment classification, a text is classified as positive or negative, in the multiclass sentiment classification, the text can be classified in two ways. First, the classes will be negative, neutral, and positive; second, the classes will be strongly negative, negative, neutral, positive, or strongly positive. In the second way, the strength of sentiment affects the classes based on the application. Terms such as faintly, slightly, or stoutly are used to show the strength of opinions. One should have knowledge about syntax and semantic to extract and analyze the sentiment correctly. Subjectivity classification is more difficult than sentiment classification.

Techniques for sentiment classification can be categorized into two main approaches such as machine learning approach, lexicon based approach. We have used hybrid techniques for our system i.e. supervised learning technique and dictionary based approach

Methods to classify documents:

The main challenge of machine learning systems is to determine the distinction between the lexical level of “what actually has been said or written” and the semantic level of “what is intended” or “what was referred to” in the text or utterance. LSA, PLSA, LDA

Calomiris and et.al. [5] have develop a classification methodology for the context and content of news articles to predict risk and return in stock markets . Mao et al. has explained the sentiment categorization has allured growing interest from natural language processing. The goal of sentiment categorization is to recognize automatically whether a specified piece of text expresses positive or negative opinion towards a subject of appeal.

Heston and Sinha [6] investigated the usefulness of textual processing for predicting stock returns. They specifically used a neural network applied to a broad dataset of news stories.

Generally sentimental word dictionaries will be used for labelling of Small piece of data called “crunches”. These kinds of dictionaries contain certain threshold value for sentiment word and the defined value is used to decide sentiment of word is positive or negative for subjective sentences. SentiWordNet V3.0 or WordNet are the online available sentiment word dictionaries. We have employed sentiwordNLP and added few new words in context of social issue (water)

E. Sentiment tone:

It is sometimes important to understand the tone of the text besides understanding what people are talking about. In essence sentiment analysis it is the process of determining the emotional tone behind a series of words, used to gain an understanding of the attitudes, opinions and emotions expressed within a text.

Challenged by the question why some texts are more interesting than the others, in their paper, Anderson and McMaster concluded that the “emotional tone” of a story can be responsible for the reader’s interest. The results of their study suggest that a large-scale analysis of “emotional tone” of the collection of texts is possible with the help of a computer program. We have found tone of document using a combination of noun phrase and verb.

F. Sentiment extraction:

In sentiment analysis, sentiment extraction is performed when one wants to find out relationship between the opinion target and sentiment. It is not an individual step but performed under another linked step that connects opinion target and sentiment in some context. To evaluate the sentiment in text content such as news articles or blogs two main tasks are important: the identification of relevant sentiment words and its sentiment weight, which is nothing but sentiment extraction.

The orientation of sentiment words can be found using dictionary based approach or corpus based approach. Dictionary based approach uses seed of words with their synonym and antonym corpus approach uses concurrences of words.

Most of the dictionary-based algorithms for sentiment analysis consider word frequency in documents. Laura Cruz and et. al. [7] research has shown that collected corpus words with low frequencies can be useful to set polarities. Harb et al state that dictionary based approaches are unsupervised in nature. SentiWordNet dictionary and smiley dictionary is utilized by Wandhe et al.to score the sentiment into positive, negative and objective. Kar Kei Lo and et.al., [8] studied the relationship between social media and the financial performance of penny stocks and used the net proportion of positive words in stock articles in social media to help predict the future stock performance for penny stocks.

Bouchaib et.al. [9]have prepared a template for an open government portal that takes advantage of different technologies, linked data to discover the links between heterogeneous datasets, natural language processing to aggregate in a semantic level similar data-set on a dictionary approach and a feature learning approach. Mensikova1 et. al.[10], describe a series of experiments to apply sentiment analysis as an indicator for human trafficking. They applied existing binary e.g., Netflix and categorical e.g., Stanford Treebank sentiment models directly to subsets of web ads from our MEMEX human trafficking corpus for which ground-truth regarding human trafficking was available. Khodak et al[11] introduce the Self-Annotated Reddit Corpus (SARC), a large corpus for sarcasm research and for training and evaluating systems for sarcasm detection. Gaillat et.al.[12] build a Sentiment Analysis (SA) system dedicated to financial microblogs in English. The purpose of their work was to build a financial classifier that predicts the sentiment of stock investors in microblog platforms.

Stuart et.al. [13] research focuses on a corpus-driven approach to SA using semantics, language patterns and statistics. In Cruz et.al.[14]paper they describe methodology to integrate domain-specific sentiment analysis in a lexicon-based system initially designed for general language texts.

G. Sentiment evaluation and Text structure:

Structure of the text under Sentiment evaluation is the main ingredient in raising different challenges during sentiment analysis. Text structure under sentiment analysis is an important and essential factor to be considered. There are different types of text formats as: structured, Unstructured and semi-structured. Following table discusses the relationship between the type of structure of data and sentiment analysis challenges.

Table 1 shows relationship between the type of structure of data and sentiment analysis challenges.

Sr. No.	**Title of paper and publication year**	**Authors**	**Domain**	**SA challenge**	**Document structure type**
1	2015. Online paper review analysis.	Doaa, M.E., Hoda, M.O.M., Osama, I.,	Scientific papers	Lexicon	Structured
2	2016. Enhancement bag-of-words model for solving the challenges of sentiment analysis	Doaa, M.E.,	Scientific papers	Lexicon + negation	Structured
3	2014A scalable lexicon based technique for sentiment analysis.	Chetan and Atul	tweets	Huge lexicon	Unstructured
4	2014. Detecting spam review through sentiment analysis	Qingxi, P., Ming, Z.,	online customers reviews	Spam and fake detection	Unstructured
5	2016 Data mining of heterogeneous data with research challenges	Monika Kalra , Niranjan Lal	No	Heterogeneous data	structured, semi-structured unstructured
6	2017. Structured data extraction from emails	Ashraf Q. Mahlawi Sreela Sasi	project or product	keyword extraction	Semi-structured unstructured

H. Part of speech (POS) and sentiment analysis research

Parts of Speech (POS) explain how a word is used in a sentence, i.e whether it is a verb, noun, adjective and so on. In text processing, those POS (or word classes) are usually represented as their abbreviation and we call it tag. POS can be used in multiple applications in text analytics.

Farah Benamara and et.al to date, there is almost no work on the use of adverbs in sentiment analysis, nor has there been any work on the use of adverb-adjective combinations (AACs). They propose an AAC-based sentiment analysis technique that uses a linguistic analysis of adverbs of degree. Rahim Dehkharghani has proposed a hybrid approach for building adjective polarity lexicon, which is experimented on Turkish combines both lexicon based and corpus based methods.

The work presented by Kamps J and et.al. in their paper focused on the use of the synonymy relation between adjectives in WordNet to generate a graph, measures the shortest path between the adjective and two basic sentiment seeds. Xiaowen Ding and Bing Liu provides a good insight on the usage of Linguistic rules to identify the polarity of context-dependent opinion words at the sentence level. Linguistic rules at Intra-sentence and inter-sentence conjunction as well

I. Sentiment Visualization:

C. Wang presents SentiView, an interactive visualization system that aims to analyze public sentiments for popular topics on the Internet. Tat A et.al. gives a visualization technique for personal chat history. Hao M et. al. present four novel visualization techniques for customer feedback analysis. Wang Y presents a polarity classification and visualization of citations to represent a citation graph for paper review purposes.

J. Framing:

When internet sources use text as strategic devices for presenting noticeable features and views about an issue, using certain keywords and casted images and sentences for the purpose of conveying latent meanings about an issue, it is called framing. Public opinion, attitudes, beliefs, and behaviours can be influenced by how an issue is framed, particularly when framing comes from leaders who have the power to go beyond influence to manipulation [15].There are many social, environmental and political issues that can be studied through the lens of framing on internet sources because these issue are specific, complex and characterized by uncertainty. Newspapers regularly produce in-depth articles about socio-environmental, political issues that are published even online. Biased framing or terms are more likely to influence uninformed respondents or respondents with less exposure or interest in an issue therefore, citizens' attitudes and beliefs about an social issue are likely influenced by the way related people frame this issue. This source of unstructured text offers researchers a body of content to test our learning and machine-discovery approach on a relevant issue.

Framing is a tool used by media and politicians to make salient points that would direct their readers to a desired frame of mind. Frank Luntz was the first "professional pollster to systematically use the concept of framing as a campaign tool. Framing theory and frame analysis provide a broad theoretical approach that analysts have used in communication studies, news [15], politics, and social movements, but theoretical base is insufficient so the study here proposes a framework that discovers distinct frame with help of keywords that advances the process of machine learning of digital data for classification and clustering.

Eric P et al.[16]have presented a technique for identifying frame-invoking language. It described a human subject's pilot study that explored how individuals identify framing and informs the design of our technique. Antonio et al. have presented an approach to the representation, acquisition and analysis of frame evidence. Bjorn et al.[17] have explored the application of supervised machine learning (SML) to frame coding. Tsur et al. have implemented the automatic classification of four generic news frames: conflict, human interest, economic consequences, and morality.

K. How framing stimulate sentimental expression?

Teresa A. & et al has discussed on how framed news articles about climate change emphasizes on the risks to environment, public health or national security. They have also discussed the benefits of mitigation and adaptation related actions. Sujin choi & et al has studied how an online community has employed social media to mobilize people for a political goal The use of internet for social mobilization using frames has been found to have a major impact on social movement organization for fundraising and generating resources.

Thus framing can connect people logically or emotionally. In order to engage people in framing, behavior of the people on a particular issue is a task of research. The research presented here deals with sentiment [emotions].

L. Types of frames:

There are different types of frame like Conflict frame, Human interest frame, Economic interests frame, Moralization frame, Common/collective interest frame, Responsibility frame, Identification /rationalization frame, Individual and state frame, Troubled industry frame, Affirmative action frame, Fetal life frame, Women's rights frame, Migration autonomy frame, Refugee protection frame. Motivational framing text was obtained from a site promoting a July 2008 climate rally in Australia as shown below.

III. METHODOLOGY

A. Existing Methodologies and Analysis:

Subjectivity classification can presented as objective/subjective or fact/opinion. A sentence can be either subjective sentence or objective sentence. Objective sentence contains the facts. It has no judgement or opinion about the object or entity while subjective sentence has opinions Sentiment classification actually recognises sentiment orientation of a document on the target. Sentiment classification is about subjectivity, polarity, and polarity strength of an opinion text as per Binali et al.

Challenged by the question why some texts are more interesting than the others, in their paper, Anderson and McMaster concluded that the "emotional tone" of a story can be responsible for the reader's interest. The results of their study suggest that a large-scale analysis of "emotional tone" of the collection of texts is possible with the help of a computer program.

Sentiment of words may change in the context and different domains, for example the word carrier may have different meaning in different context and in different domain. So in order to recognize the correct sentiment of some words more knowledge about the domain or context in which it is use is required so that real sentiment in context can be recognized. Peter D. et. al has carried research to find out real sentiment of a word or term in a context.

Latent semantic analysis (LSA) is a technique in natural language processing, in particular in vectorial semantics, of analyzing relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms. LSA assumes that words that are close in meaning will occur in similar pieces of text. A matrix containing word counts per paragraph (rows represent unique words and columns represent each paragraph) is constructed from a large piece of text and a mathematical technique called singular value decomposition (SVD) is used to reduce the number of columns while preserving the similarity structure among rows. Words are then compared by taking the cosine of the angle between the two vectors

formed by any two rows. Values close to 1 represent very similar words while values close to 0 represent very dissimilar words.
The popularity of SVD in the field of text analytics is due to more than just its ability to reduce dimensionality. The truncation of the decomposition addresses, at least in part, the problem of synonymy. Truncated SVD reflects similar co-occurrences of terms in the dimension values and thus approximates the manner in which a human perceives similarity between words.

Generally sentimental word dictionaries will be used for labelling of Small piece of data called "crunches". These kinds of dictionaries contain certain threshold value for sentiment word and the defined value is used to decide sentiment of word is positive or negative for subjective sentences. SentiWordNet V3.0 or WordNet are the online available sentiment word dictionaries.

Academic research in computer science on hidden meanings in association of terms and documents to disclose relationships is found in works related to text summarization, information retrieval, and text data mining. The initial paper on text summarization is that of Hans Peter Luhn work done at IBM, Luhn proposed that the frequency of a particular word shows its importance. He contributed in the identification of the concept term frequency (TF), in this context he says that it is possible to find the important terms on the basis of its term frequency within that document.

Edmundson was the first to describe a system that produces document abstracts. His primary contribution was the development of a typical structure for an summary extraction which integrates features of word frequency and positional importance, lent from the works of Luhn and Baxendale.

Calomiris and et.al. [5] have develop a classification methodology for the context and content of news articles to predict risk and return in stock markets. They found that the meaning of news flow can be captured usefully through a small number of a theoretical measure (sentiment, frequency, and entropy). The meaning of those measures for stock market risk and return vary over time, and vary according to the topical context in which sentiment and frequency are measured. Nevertheless, they found that it is possible to construct a parsimonious and flexible forecasting model that maps usefully from these theoretical, context-specific measures of news flow into equity market risk and return.

Mao et al. has explained the sentiment categorization has allured growing interest from natural language processing. The goal of sentiment categorization is to recognize automatically whether a specified piece of text expresses positive or negative opinion towards a subject of appeal. The standpoints have been presented that utilizes a human model based on random process to determine text polarity categorization is presented. Experiment outcomes displayed that on movie review corpus, the human modeling approach has a relatively greater precision than SVMs and Naive bayes classifier. In the experiment, method to determine text polarity classification has many advantages (1) The method automatic extracts the semantic features without semantic word dictionary.(2) The accuracy of the method will be higher when more prior knowledge is added.

While performing sentiment analysis on twitter posts is done using Machine learning approaches, it comes with a drawback that it considers the whole post has a uniform statement and assigns a sentiment score for the entire post. To overcome this issue, Ontology-based technique is applied. It divides twitter post into a set of aspects and assigns sentiment score to each distinct aspect (Kontopoulos et al., 2013).

Heston and Sinha[6] investigated the usefulness of textual processing for predicting stock returns. They specifically used a neural network applied to a broad dataset of news stories.

Most of the dictionary-based algorithms for sentiment analysis consider word frequency in documents. Laura Cruz and et. al. [7] research has shown that collected corpus words with low frequencies can be useful to set polarities.

SentiWordNet dictionary and smiley dictionary is utilized by Wandhe et al. to score the sentiment into positive, negative and objective. Rule based and fuzzy logic approach is used to handle negation words. Fuzzy Intensity Finder Algorithm is used to find intensity of each word.

Stuart et.al. [13] research focuses on a corpus-driven approach to SA using semantics, language patterns and statistics. The basic data set was a corpus of patient narratives (www.patientopinion.org.uk) giving feedback about treatment for a medical condition within the NHS. As the data was collected over a six-year period, this kind of analysis is possible with Patient Opinion Corpus. Information gathered through sentiment analysis on a corpus of patient narratives could ultimately make a difference to the running of hospitals and the wellbeing of patients. The research describes the linguistic analysis of a corpus of patient narratives that was used to develop and test software to carry out sentiment analysis on the aforementioned corpus. Sentiment analysis can be used in different domain like education, marketing, politics etc for different purpose and each domain can have different applications for example marketing domain can use it to extract and visualise opinions about a company's product and its features. it can also be used in comparison, market and product and trend analysis.

All the methods are using machine learning techniques for subjectivity classification and sentiment classification to obtain at the best sentiment analysis of product, services, stock or patient narratives and etc but not on social issues.

B. System Architecture:

As shown in Fig 1. Classifier 1 is responsible to classify unstructured data as frame documents and non-frame documents and Classifier 2 is responsible to classify frame documents into 3 distinct types motivational, predictive (prognostic) and investigative (Diagnostic).

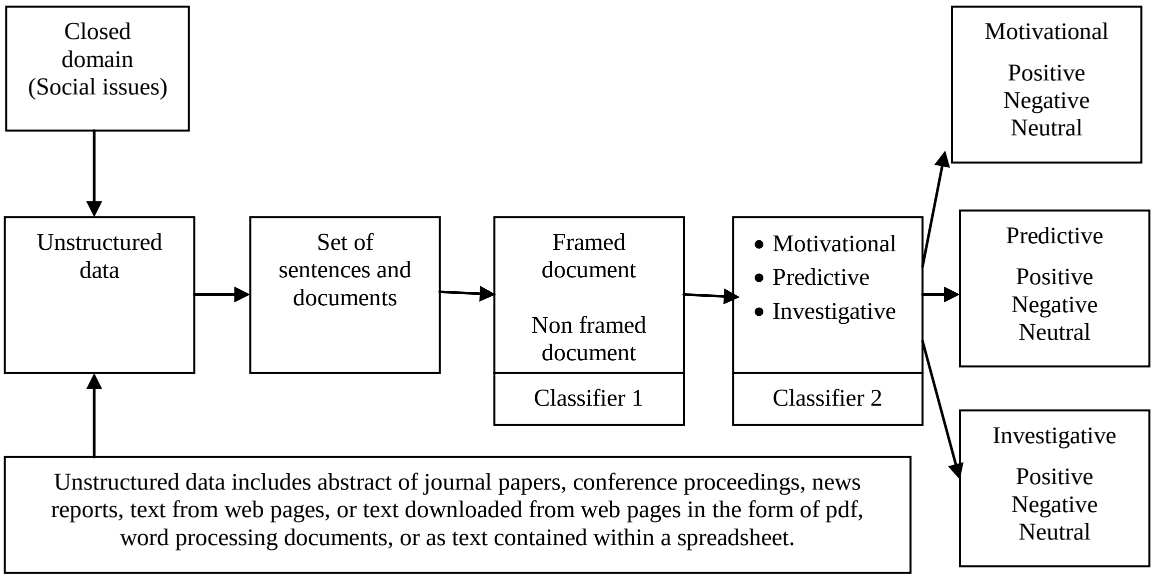

Figure1. Framework: Analysis of sentiments from frames.

C. Proposed Methodology design and implementation

The purpose of this research is to test a new framework for finding sentiments of people from distinct frame patterns. Supervised machine learning technique along with statistical algorithm is used for document classification and analyzing the sentiments in it because framing and sentiments have cumulative effect.

This research supports efforts concurrently taken in machine learning and data mining to find or discover patterns and hidden relationships in unstructured text. This method has several improvements over conventional methods of text mining:

First, unlike the traditional approach to processing unstructured texts by focusing on syntax and its meaning, these method views texts as frames that act as carter of information that has power to aware people about an issue that will interrelate with individuals existing beliefs.

Second, by studying framing in computer research, this method does difficult processing of text representations that has the possibility to extend bodies of research beyond bag-of-words like technique for bridging bag-of-words, classification.

Third, this technique improves current methods of text analysis and mining by using machine learning to look at larger datasets and also to explore many more topics that can't be evaluated by human coders. Moreover, this approach will allow for the automatic emergence of the frame from the online unstructured text at the end of analysis and then find sentiments from them and classify as motivational, diagnostic or prognostic.

In machine learning and statistics, feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features.

For automatic detection of frames in document from a corpus here we use a classifier1 where each document from corpus is classified as Frame or Non-Frame. Then further only frame documents are taken and classified with classifier2 using sentiment analysis into motivational, prognostic and predictive.

For this classification activity the main input is feature vector that is derived using a quadruple, this quadruple represents

the concept of document which is in some context of social issues represented in corpus.
Quadruple Q is representing concept C of a Document D as below:

$$Di = Ci = Qi = \{si, npi, vi, oi\} \quad (3)$$

Where si is subject, npi is noun phrase, vi is verb and oi is object.

This way all quadruple are calculated for all documents in the corpus and this is taken further for evaluation. While a concept is derived the words used in it carry some meaning in context of some issue.

Two documents are of same concept but different words are used to represent the same concept, under such situation if context in which these words are used is same then both the documents will fall under one cluster and this would be very helpful in classification. From this we derive contextual concept which is used in accurate classification of documents.

Contextual concept is derived by finding synonyms of Quadruple given in equation (3) from the derived dictionary and online wordnet.

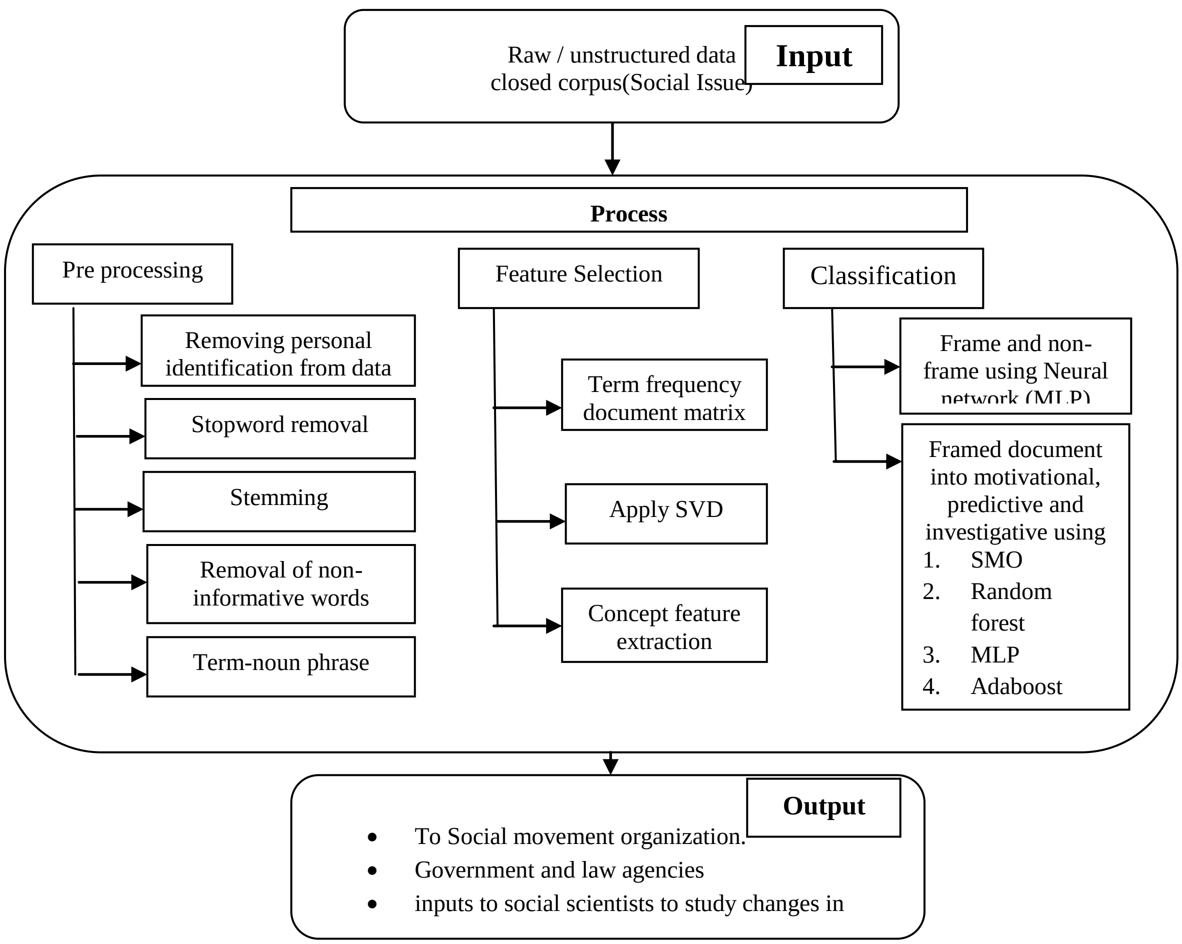

Figure 2 shows detailed steps of working approach

IV. Results and Discussion

A. The major contributions of this paper are as follows:
The basic idea is to define a proper structure for opinion and then extract frame and sentiments in them.
Structure of opinion is defined using a new design of quadruple that has importance to formulate concept from frame in some social context.

Feature selection vector uses noun phrase to find tone in document and combination of it with verb to find sentiment.
The primary data mining task is classification of text documents, resulting in two models. The first is dichotomous, classifying text documents as being either framing or non-framing. The second, a polychotomous model, classifies text documents as one of four types: diagnostic, prognostic, motivational, or non-framing.

For the purpose of calculating evaluation measures for this model, a classification of "Framing" is considered to be "positive," and a classification of "Non-Framing" is "negative."
In this thesis, all of the documents address one topic, water problem, and the job is to detect those documents that were written for the purpose of influencing opinion and actions regarding the topic. This task becomes more difficult when one considers the fact that both non-framing and diagnostic framing texts can define water problems and its effect on our planet. The slight difference is that the non-framing document may have been written for the purpose of educating the reader, while the diagnostic framing document is intended to not only educate, but also influence the reader's perception of events and personal experience.

In the same manner, elements of prognostic framing text may logically be found in non-framing text. Like different people use the same words for water related problems but are discussing very different ideas.

If a term frequently appears in the many documents then that term will not be useful in distinguishing the documents from one another. This is reflected in a lower term weight. So another important relation to get concept properly is to use the global frequency, it gives discrimination value to each term and it is observed that less frequently a term appears in whole document the more discriminating it is. So here global frequency of every term is obtained and incorporated in term weight.
For the whole task the input was as below:

Table1

Total documents in corpus	2566
Frame	690
Non-Frame	1876

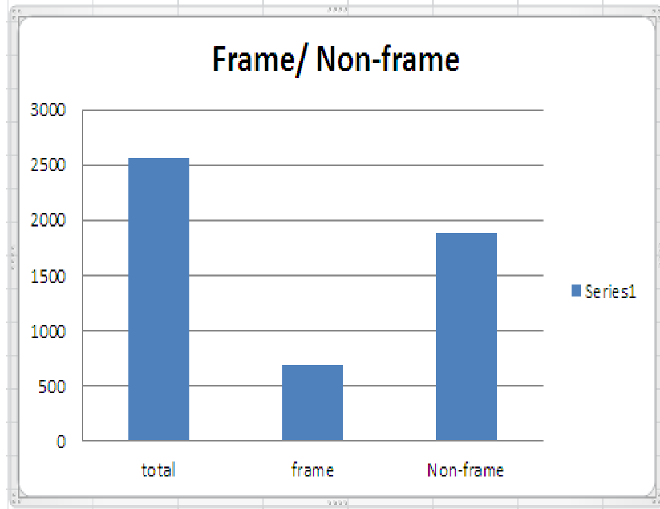

Figure 3 % Frame documents in corpus

Confusion matrix is used for performance measurement for machine learning classification problem where output can be two or more classes. It is a table with 4 different combinations of predicted and actual values.
The unbalanced test data set was scored by model. The resulting confusion matrix is shown in table 1. The rows are the true framing and non-framing classifications for the documents in the test data set. The columns are the classifications generated by the classifier 1 using MLP. Each cell contains the cross-tabulated number of documents.

Table 2 Confusion Matrix for Classifier 1

	Model classification		
True classification	Framing	Non framing	total
Framing	96	594	690
Non framing	594	1282	1876
Total	690	1876	2566

Table 3 Evaluation Metrics for Classifier 1

Evaluation Metric	Model at classifier1 using MLP	Model at classifier1 using (Dummy keyword)Dkw
Precision	0.6603	0.8673
Recall	0.9765	0.9202

F1 Measure	0.7879	0.8929
Accuracy	0.9785	0.9684

Once the documents in corpus are labeled with Frame and Nonframe. We utilized Algorithm 2 to find the sentiments in frame document only and classified them into motivational, investigative and predictive using noun phrase for tone in frame using quadruple designed in equation 3. We experimented with SMO, Random Forest, Multilayer perceptron and Adaboost1 and found following results.

Table 4 Showing evaluation metrics of classifier 2 using different algorithms based on contextual concept quadruple.

Algorithm used for experimenting Classifier2	Precision	Recall	F-measure
Multilayer Perceptron (MLP)	0.483	0.479	0.479
Sequential Minimal Optimization (SMO)	0.507	0.507	0.503
Random Forest	0.91	0.91	0.91
Adaptive Boosting (Adaboost M1)	1.00	1.00	1.00

Correctly classified instances means if we use 2 fold technique; we had given 50 % records for training and 50% for testing. On the basis of training, records are correctly classified.

y axis % of records correctly classified after training and x axis means no. of folds used for training and testing of input data.

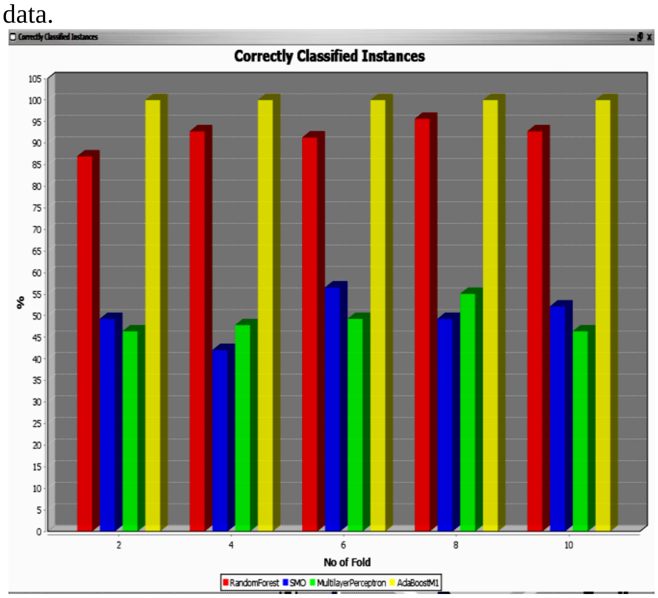

Figure 4 % correctly classified instances for all 4 classifier algorithms

On the basis of limited set of data and applying cross fold validation we obtained following output:

Kappa statistics: The Kappa statistic is used as a means for evaluating the prediction performance of classifiers. The Kappa statistic (or value) is a metric that compares an Observed Accuracy with an Expected Accuracy (random chance). The kappa statistic is used not only to evaluate a single classifier, but also to evaluate classifiers amongst themselves. In addition, it takes into account random chance (agreement with a random classifier), which generally means it is less misleading than simply using accuracy as a metric.

Mean absolute error – quantity used to measure how close forecasts or predictions are to the eventual outcomes.

Root mean square error - measure of the differences between values (sample and population values) predicted by a model or an estimator and the values actually observed. It represents the sample standard deviation of the differences between predicted values and observed values.

Relative absolute error: The mean absolute error, divided by the corresponding error of the ZeroR classifier on the data (i.e. the classifier predicting the prior probabilities of the classes observed in the data).

Root relative squared error: The root mean squared error, divided by the corresponding error of the ZeroR classifier on the data (i.e. the classifier predicting the prior probabilities of the classes observed in the data).

Table 5 Showing different measure used to calculate performance of classifier 2

	Algorithm	Output parameters	values
Classifier2	Random forest	Correctly classified instances	94.02%
		In Correctly classified instances	5.97%
		Kappa statistics	86%
		Mean absolute error	7.36%
		Root mean squared error	15%
		Relative absolute error	44.88%
		Root relative squared error	56.80%
		Total number of instances	67
Classifier2	Sequential Minimal Optimization (SMO)	Correctly classified instances	61.19%
		In Correctly classified instances	38.80%
		Kappa statistics	34%
		Mean absolute error	20%
		Root mean squared error	31%
		Relative absolute error	99.58%
		Root relative squared error	99.30%
		Total number of instances	67
	Multi layer	Correctly classified	50.74%

	perceptron	instances	
		In Correctly classified instances	49.25%
		Kappa statistics	21%
		Mean absolute error	16%
		Root mean squared error	36%
		Relative absolute error	73.50%
		Root relative squared error	110.07%
		Total number of instances	67
Classifier2	AdaBoost 1	Correctly classified instances	100%
		In Correctly classified instances	0%
		Kappa statistics	100%
		Mean absolute error	0
		Root mean squared error	0
		Relative absolute error	0
		Root relative squared error	0
		Total number of instances	67

B. Model Suitability:

Using features generated, we perform ten-fold cross validation for measuring classifiers accuracy.

A way to measure a model's suitability is to assess the model against a set of data that was not used to create the model. To measure the suitability of a binary regression model, one can classify both the actual value and the predicted value of each observation as either 0 or 1.

The predicted value of an observation can be set equal to 1 if the estimated probability that the observation equals 1 is above, and set equal to 0 if the estimated probability is below. There are four possible combined classifications:

Prediction of 0 when the holdout sample has a 0 (True Negatives, the number of which is TN)

Prediction of 0 when the holdout sample has a 1 (False Negatives, the number of which is FN)

Prediction of 1 when the holdout sample has a 0 (False Positives, the number of which is FP)

Prediction of 1 when the holdout sample has a 1 (True Positives, the number of which is TP)

$$Precision = TP/(TP + FP)$$

Precision = TP/(TP + FP) = Fraction of predicted positives that are correct

$$Recall = TP/(TP + FN)$$

Recall = TP/(TP + FN) = fraction of observations that are actually 1 with a correct predicted classification

$$Fmeasure = \frac{2 * Precision * recall}{(Precision + recall)}$$

F-measure = 2*Precision*recall/(Precision+ recall)

$$Accuracy = \frac{TP + TN}{(TP + TN + FP + FN)}$$

Accuracy = TP + TN /(TP + TN + FP + FN)= fraction of observations with correct predicted classification.

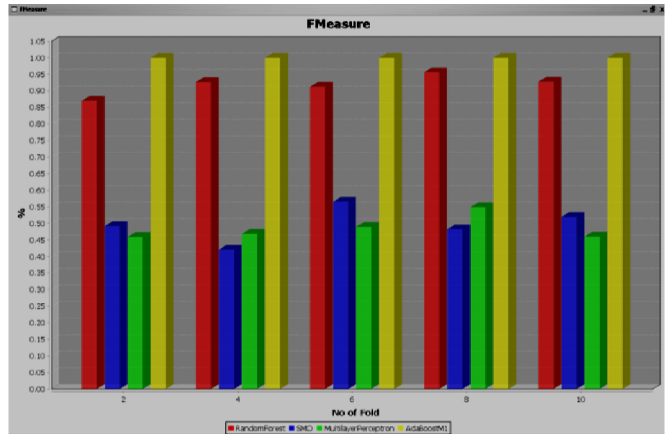

Figure 5 Model suitability: Fmeasure for all the four algorithms

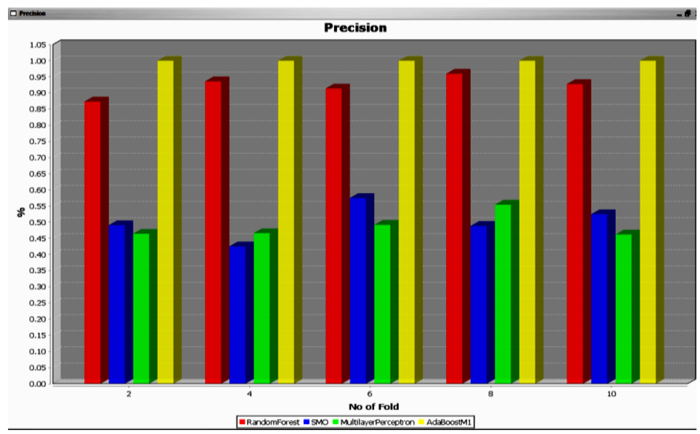

Figure 6 Model Suitability: Precision for all the four algorithms

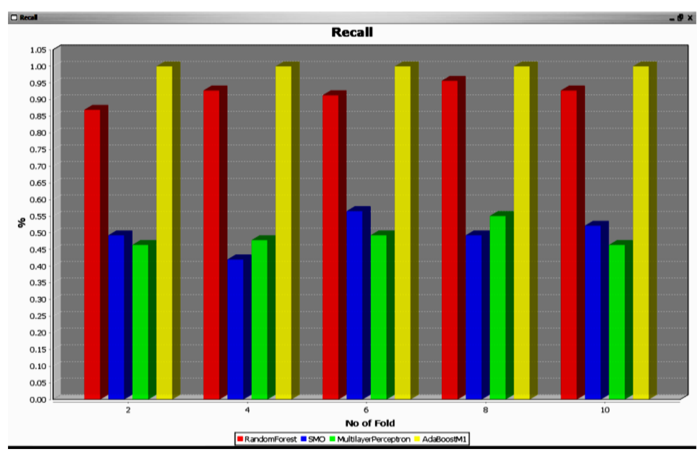

Figure 7 Model Suitability: Recall for all the four algorithms

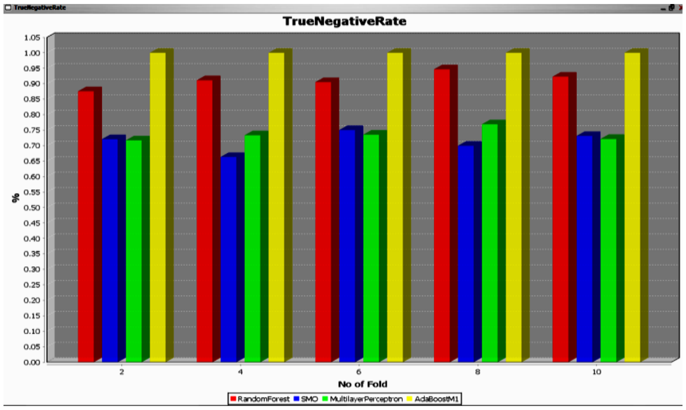

Figure 8 Model suitability: True Negative for all the four algorithms

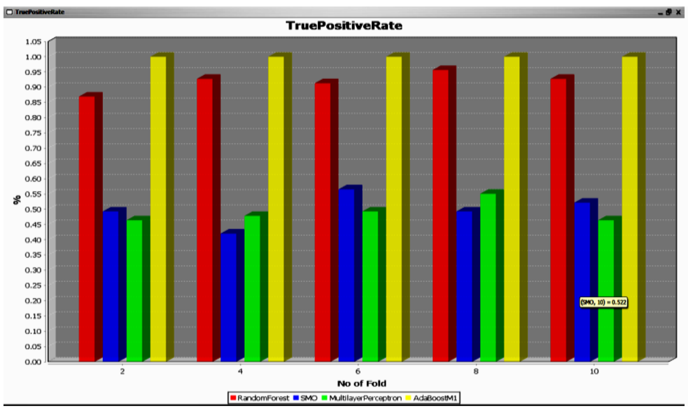

Figure 9 Model suitability: True Positive rate for all the four algorithms

After performing classification of sentiments, using the scores obtained from sentiment analysis the output after applying every algorithm of classifier 2 is as shown below.

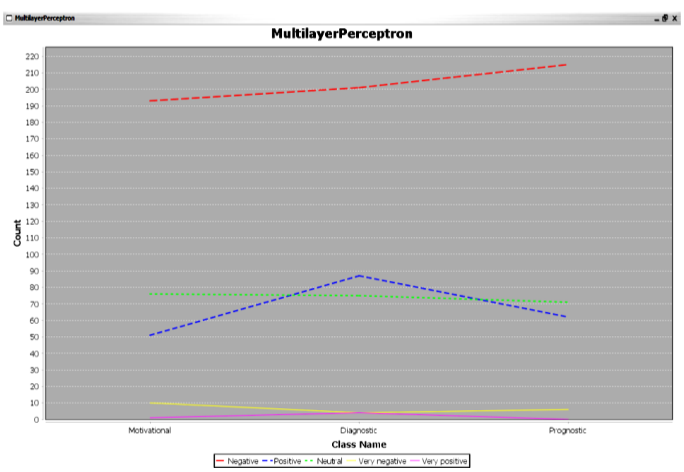

Figure 10 Graph shows count of positive, negative, neutral, very negative and very positive words in documents after classification using MLP classifier.

Graph shows that count of positive words is more in diagnostic class whereas negative words are same for motivational and diagnostic class and increases for prognostic. Neutral word count is same for all three classes whereas true positive rate is more for motivational class that indicates that MLP is suited to diagnose problem more correctly.

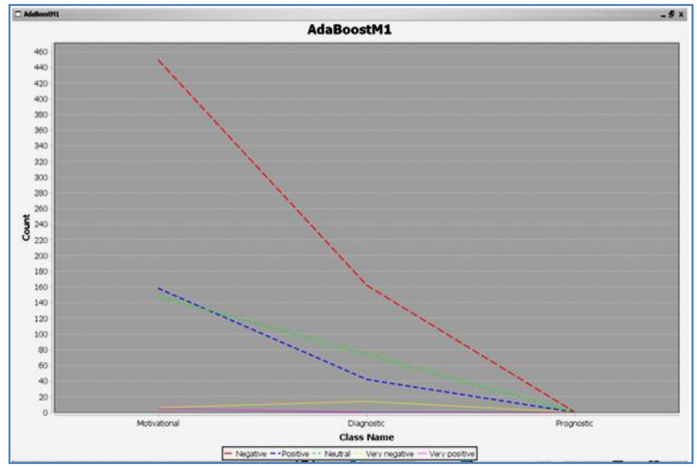

Figure 11 Graph shows count of positive, negative, neutral, very negative and very positive words in documents after classification using Adaboost M1 classifier. Graph shows that adaboostm1 is suited to diagnose motivational and diagnostic type of frames but not for prognostic.

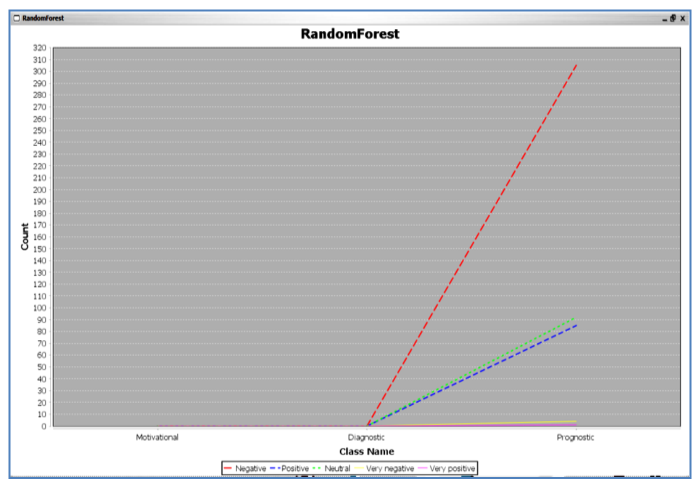

Figure 12 Graph shows count of positive, negative, neutral, very negative and very positive words in documents after classification using Random Forest classifier. Graph shows that random forest is suited to diagnose prognostic frames and very less diagnostic type of frames but not for motivational.

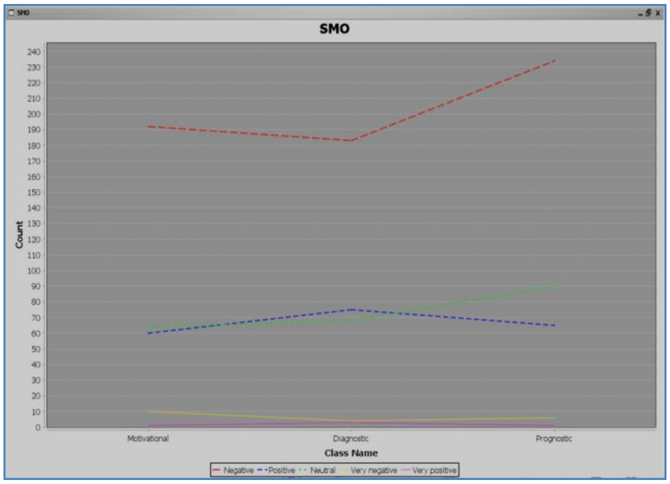

Figure 13 Graph shows count of positive, negative, neutral, very negative and very positive words in documents after classification using SMO classifier. Graph shows that SMO is suitable for all the types of frames classification. Its recall, precision, fmeasure is better compared with MLP but not with adaboostm1 and randomforest.

The question after all sentiment analysis remains how to make use of the available output, concerned authority can make use of the graphs shown below to see if the problem which is taken as a social issue is motivational, prognostic or diagnostic.

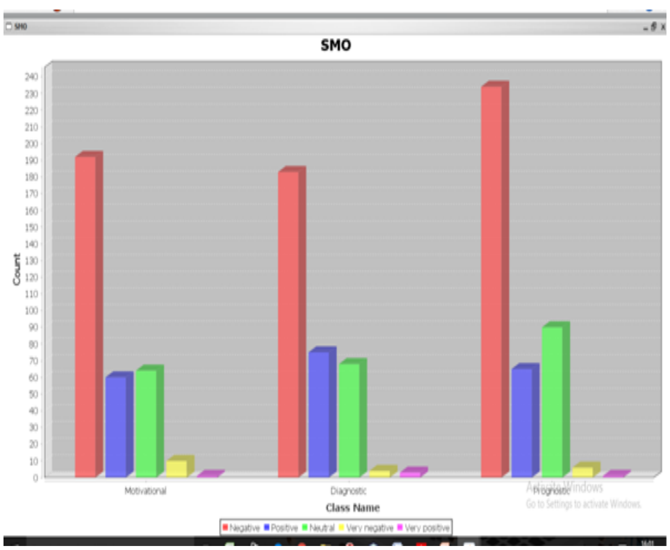

Figure 14 With SMO as classifier algorithm the graph depicts that the contents shows motivational, diagnostic as well as prognostic but positive score is more for diagnostic type of class so the problem needs some diagnosis and solution is needed to handle the problem.

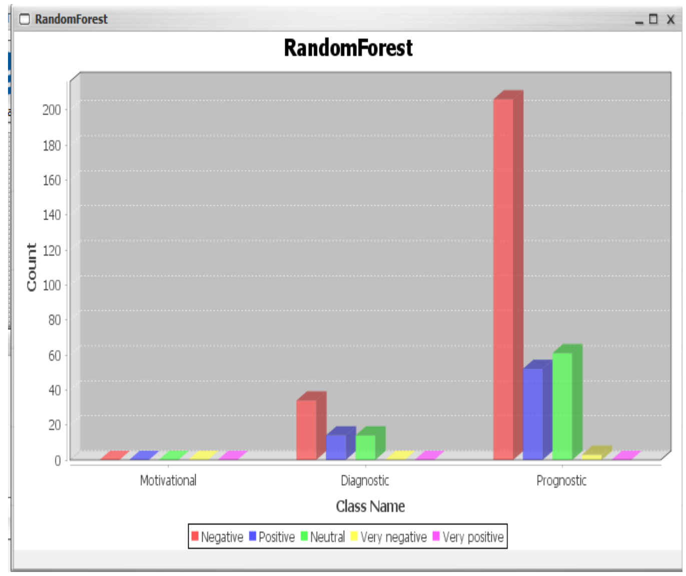

Figure 15 With Random forest as classifier algorithm the graph depicts that the contents shows motivational, diagnostic as well as prognostic but positive score is more for prognostic type of class so the prediction of the problem related to social issues is negative needs some immediate action and solution is needed to handle the problem.

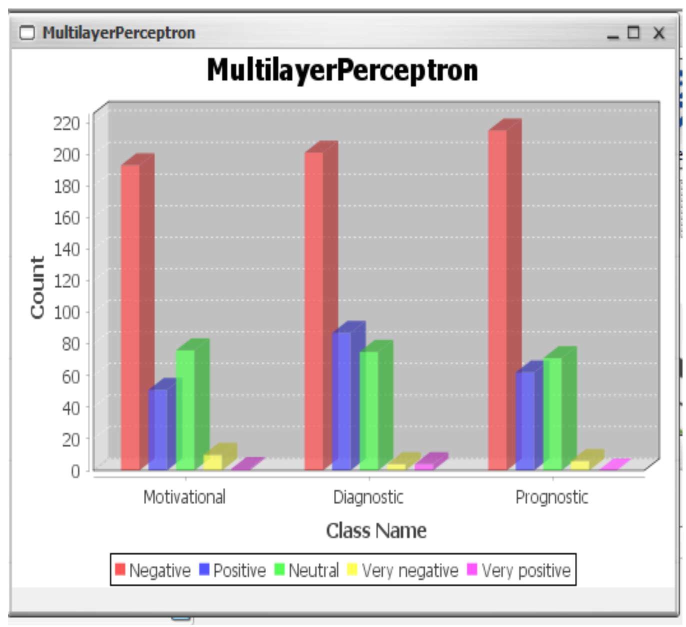

Figure 16 With Multi layer perceptron as classifier algorithm the graph depicts that the contents shows motivational, diagnostic as well as prognostic but positive score is more for diagnostic type of class so the prediction of the problem related to social issues is positive and needs some diagnosis and solution is needed to handle the problem.

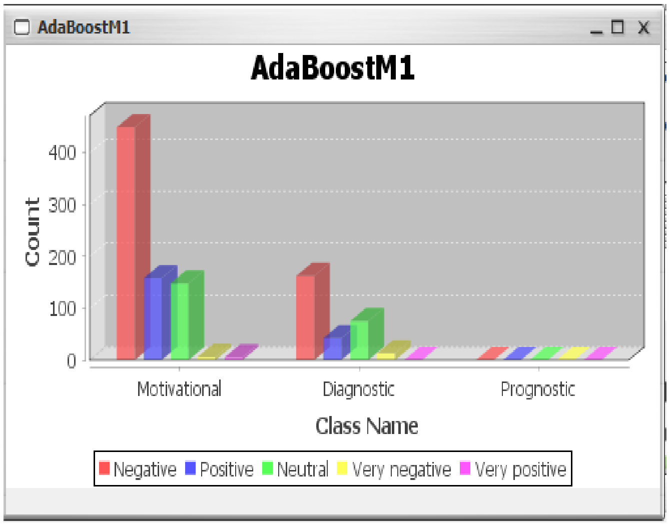

Figure 17 With Ada Boost M1 as classifier algorithm the graph depicts that the contents shows motivational, diagnostic but not prognostic but negative score is more for motivational type of class, but the class is motivational so the prediction of the problem related to social issues is positive and it can be used as motivational contents for people and solution can be achieved through this.

Result Analysis: The results from figure 7 to figure 20 clears that the model designed using quadruple q of equation 3 for feature selection helps in proper analysis of sentiments. In comparison with four classifier algorithms, adaboostm1 gives 100 % accuracy and shows tone based sentiment classification.

V. Conclusion and Future Scope

A. Conclusion

From all above discussed contributions for the tool designed for sentiment analysis of social issues concerned authorities can make use of it for various purposes, we conclude that

Quadruple has combination of noun phrase and verb for identification of tone in framing documents.

Proposed tool will help the concerned authorities to secure the people from unwanted situation.

The classifier1 model using MLP has accuracy of 0.9784 is higher than the accuracy of 0.9685 for the dummy keyword model. The dummy keyword model had high recall, but low precision. It found all but five of the framing documents in the test data set, but one-third of the documents that it identified as framing were not. The dummy variable model had slightly lower recall than the MLP model, but returned a small number of false positives resulting in higher precision. The differences in precision and recall for these two models are reflected in the F1 measure, which is higher for the dummy variable. Therefore, the MLP model is selected as the best model at classification level 1in the proposed framework.

The structure of the quadruple used for feature extraction performs well with all the four classification algorithms in classifier 2.

To achieve the overall performance, structure of opinion and feature selection must be proper and we have achieved it up to some extend using the quadruple structure that captures the concept in context of social issue.
Results show the performance of our tool in comparison with other tools available in market.

Table 6 Performance of proposed tool in comparison with other tools

	The proposed approach	**Rapid Miner**	**IBM Watson**	**Meaning cloud**	**Stream Crab**
Automatically constructed dictionary	Yes	No	No	No	No
Context based dictionary	Yes	No /customisable	No /customisable	No	No
Classification degree	positive, negative, neutral, very negative and very positive	Positive / negative / and Neutral customisable	Positive/ negative/ and neutral	Positive/ negative / and neutral	Positive/ negative
Intensifier	Yes	No	No	No	No
Visualisation	Yes	No	No	No	No

B. *Future Scope*

- The work presented here is applied only for corpus (writings on single issues) of text document. It can be made available for different issues creating context dictionary for that particular issue
- The work presented here is not applicable for real time data which may be worked out as future scope.

ACKNOWLEDGEMENT

I would like to thank Prof. R M. Tuganayat for his encouragement and comments on earlier drafts of this paper. Without the support of my loving husband V.P. Balshetwar (B.E. Electronics) and our kids Siddhesh and Katyayani it would have been impossible to complete this work.

REFERENCES

[1] Xiaowen Ding, Bing Liu, and Philip S. Yu “A holistic lexicon-based approach to opinion mining” Proceedings of the international conference on Web Search and web Data Mining, ACM, , pp. 231-240 2008.

[2] Alexander Conrad, Janyce Wiebe, and Rebecca Hwa, “Recognizing arguing subjectivity and argument tags, Proceedings of the Workshop on Extra Propositional Aspects of Meaning in Computational Linguistics” ACM,pp. 80-88 2012.

[3] Jack, R. E., O. G. Garrod, and P. G. Schyns “ Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time.” Current biology 24 (2), 187–192 2014.

[4] Liu, Bing. “Sentiment Analysis and Subjectivity. Handbook of Natural LanguageProcessing”, 2nd ed. Chapman and Hall: Florida, 2010.

[5] Calomiris, C. W., & Mamaysky, H. “How News and Its Context Drive Risk and Returns around the World” SSRN Electronic Journal 2017.

[6] Heston, S., and Sinha, N.. “News versus Sentiment: Predicting Stock Returns from News Stories.” Financial Analysts Journal, 73 2017.

[7] Laura Cruz, José Ochoa, Mathieu Roche, Pascal Poncelet. Dictionary-Based Sentiment Analysis Applied to a Specific Domain. SIMBIg: Symposium on Information Management and Big Data, Sep 2016.

[8] Kar Kei Lo and Michael Chau. “A Penny Is Worth a Thousand? Investigating the Relationship Between Social Media and Penny Stocks” ACM Trans. Manage. Inf. Syst. 9, 4, Article 14 35 pages (March 2019).

[9] Lamiaa Sinif and Bouchaib Bounabat. Approaching an optimizing open linked government data portal. In Proceedings of the 2nd International Conference on Smart Digital Environment (ICSDE'18), Faissal El Bouanani and Ahmed Habbani (Eds.). ACM, New York, NY, USA, 135-139. 2018.

[10] Anastasija Mensikova1, Chris A. Mattmann1,2. Ensemble Sentiment Analysis to Identify Human Trafficking in Web Data. In Proceedings of ACM Workshop on Graph Techniques for Adversarial Activity Analytics (GTA3 2018). ACM, New York, NY, USA 2018.

[11] Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. 2017. A large self-annotated corpus for sarcasm. arXiv preprint arXiv:1704.05579.

[12] Gaillat, Thomas, Sousa, Annanda, Zarrouk, Manel, & Davis, Brian “FinSentiA: sentiment analysis in English financial microblogs” Paper presented at the 25th French Conference on Natural Language Processing - TALN2018, Rennes, France, 14-18 May, 2018.

[13] Keith Stuart, Ana Botella, and Imma Ferri “ A Corpus-Driven Approach to Sentiment Analysis of Patient Narratives “ CILC 2016 (EPiC Series in Language and Linguistics, vol. 1), 2016.

[14] Moreno-Ortiz, J Fernández-Cruz “Identifying polarity in financial texts for sentiment analysis: a corpus-based approach“ Procedia-Social and Behavioral Sciences, 2015.

[15] https://en.wikipedia.org/wiki/Framing_(social_sciences)

[16] Eric P. S. Baumer, Elisha Elovic, Ying “Crystal” Qin, Francesca Polletta, Geri K. Gay1,"Testing and Comparing Computational Approaches for Identifying the Language of Framing in Political News",Human Language Technologies: The Annual Conference of the North American, Page No. 1472–1482, 2015.

[17] Bjorn Burscher, DaanOdijk, RensVliegenthart, Maarten de Rijke and Claes H. de Vreese, "Teaching the Computer to Code Frames in News: Comparing Two Supervised Machine Learning Approaches to Frame Analysis",Communication Methods and Measures, Volume 8, Page No. 190–206, 2014.

[18] Daan Odijk Bjorn Burscher Rens Vliegenthart Maarten de Rijke, "Automatic Thematic Content Analysis: Finding Frames in News", International Conference on Social Informatics, Springer, Page No. 333-345, 2013.

[19] Myers, T.A., Nisbet, M.C., Maibach, E. W. & Leiserowitz, A “A public health frame arouses hopeful emotions about climate change: A letter. Climatic Change” 113, 1105–1112 2012.

[20] Choi, Sujin, and Han Woo Park. "An exploratory approach to a Twitter-based community centered on a political goal in South Korea: Who organized it, what they shared, and how they acted." New Media & Society 16.1 2014.

Authors Profile

Mrs. Balshetwar S V pursed Master of Technology from University of shivaji, Kolhapur in 2008. She is currently pursuing Ph.D. and currently working as Assistant Professor in Department of Computer Science and engineering, Satara college of engineering and management, shivaji university, Satara since 2011. She is a life member of ISTE since 2013. She has published more than 5 research papers in reputed international journals and conferences including IEEE and it’s also available online. Her main research work focuses on

Big Data Analytics, Data Mining, Data Security, Text Analytics, and Sentiment Analysis. She has 8 years of teaching experience and 4 years of Research Experience.

Mr. Tuganayat R M pursed Ph.d. in engg. & tech from RSTM Nagpur university in 2012. Master of Technology from University of Nagpur, in 1997. He is currently currently working as Principal in Aghnihotri engineering college, Wardha, Amravati university, Amravati since 2014. He is a IEE USA member and life member of ISTE, IETE, CSI bengalaru, Member of IEI. He has published more than 40 research papers in reputed international journals and conferences including IEEE and it's also available online. His main research work focuses on Big Data Analytics, Data Mining, Data Security, Text Analytics, and Sentiment Analysis. He has30 years of teaching experience and 10 years of Research Experience.He has four Ph.D. Scholars working under him and one scholar awarded Ph.D.

