International Journal of Engineering &Technology, 7 (3.12) (2018) 466-473

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Stress Analytical Modelling Based on People's Views on Social Networks

John[1], Vivia Mary[2], Prashar[3], Aastha[4], Khamesra[5], Garvit[6], V.R. Pratyusha[7]

[1,2] Assistant Professor, Department of Information Technology, SRM University
[3,4,5,6,7]Student, Department of Information Technology, SRM University

Abstract

With the growing world, the human mind has grown too much with its own complexities. Gone are the days where people used to express themselves through speech or by verbal contact. Now, the era of social media has brought an interface to the world where they can convey their opinions as well as their inner most thoughts through various social networks. People are more comfortable to express their emotions on these social media rather in the real world. This all has led to the need of Sentiment analysis. It has a major role in detecting stress in humans and how surrounding environment is affecting the population of the world. The project analyses the stress among people through tweets. Self-report questionnaires face to face interviews wearable sensors is the main basis of psychological stress that is caused traditionally. The project covers all possible aspects of interactions on social media. Firstly, by fetching tweets from twitter dynamically based on keyword entered by user and segregating them into positive, negative and neutral categories using Naive Bayes algorithm. Secondly, performing sentiment analysis on a dataset containing movie reviews and thirdly, on a very large dataset containing 5 million tweets using Hadoop and an added algorithm of logistic regression for improved performance and efficiency. The entire project was carried out using a distinct step by step procedure consisting of data collection, data cleaning, training of data, data modelling, algorithm application and visualization. Experiments were conducted on an extensive basis to verify the superior theory algorithms and credibility of the project.

1. Introduction

People have found an easy way to escape from the real world by using social media. They have started expressing themselves on various social platforms. Social Network is the easiest way through which people convey their emotions. Whether introverts or extroverts, all express their views on twitter or other social media platforms. With growing competition, human race have taken up a lot of stress and have disturbed their mental balance. Increasing stress has made them vulnerable which has resulted in no communication. So, people let their emotions out on social media like Twitter. A population is divided into various groups on the basis of age, gender, financial background, cultural background etc. So their stress also differs in different ways. Human stress is becoming a potential threat. In 2011 new business conducted a worldwide survey and reported that they was the prominent rise stress levels of around half the population in the previous 2 years. Stress itself is a very common and non clinical threat in life. But rather harmful stress fracture through the mental and Physical health of people is the excessive and chronic stress. People's life are been changed by the rise in social media. There is also a change in the research areas of Wellness and Healthcare. Social networks like Twitter when faced with the development, encourage people more and more to share their daily moods and events which there by enabling them to have an interaction with their friends through the help of social networks. It is common to categorize data based on keywords into positive or negative. In this document, we have effectively collected cleaned, trained and modeled data. Very large data sets have been worked upon using multiple using Hadoop. The algorithms of Naive Bayes and Logistic Regression have helped us achieve new accuracy and precision levels through relentless experimentation.

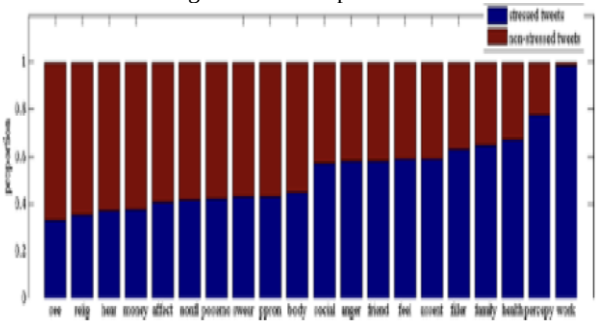

Fig. 1: Proportion of stressed and unstressed tweets in a population

2. Literature Survey

Mobile phone data, research on individual traits and weather conditions have sought to derive a way of of way of of a way of of way of of recognising the daily stress. This has proven a very important fact that the quality of a good good life has been reduced by stress and this is also the cause of many diseases. For the reason mentioned previously many researchers have been devising certain systems for detecting stress for the response to parameters that are physiological. However the systems mentioned above obligate a very important fact that certain sensors which are obtrusive in nature are constantly borne by human users. The aforementioned paper does recommend another procedure that provides information of accurately perceiving stress on a daily basis by certain physiological metrics that can be derived from the the social activity of the users users users and

many other increased indicators like date, frequency of occurrences, polarity, etc. The multifactorial demographic model proposed in this paper, is user independent and can obtain an efficient score of 63 percentage percentage to recognise the problem of stress on a daily basis. Most of the multimedia applications can be enforced on this competent model because of the remarkable feature space that is diminished in low dimension.

1. Daily Stress Recognition from Mobile Phone Data, Weather Conditions and Individual Traits

It is proved by extensive research, many recurring diseases are a result of stress which also results in a reduction in reduction in the quality of life. A method of detecting stress by the use of systems has been devised by some researchers for this reason based on the physical or physiological parameters. However, it is required by the user that they constantly carry around the necessary obtrusive sensors for the aforementioned systems to work. There is also an alternative procedure proposed by this paper that gives some information depicting that according to behavioural metrics behavioural metrics metrics, the stress on a daily basis can be perceived accurately accurately which in turn can imitate imitate can imitate imitate turn can imitate imitate the activity originating on the user's mobile phone and further indicators can be indicators can be derived out search forecast of weather of weather forecast of weather of weather (data that carries properties that are transitory of our surroundings), the Identity and traits of an individual's temperament temperament (data that refers to the perpetual dispositions of people). An efficiency of 72.28 percentage of 72.28 percentage percentage is obtained for the 2 class stress recognition problem on a daily basis on the multifactorial statistical model developed in this paper. The model of the system has high efficiency in implementation in many multimedia applications which is the result of A Remarkable feature space space that is diminished in low dimension. Moreover, there are indications that the user talks and policies have an intense predictive power.

2. Retrieval Evaluation with Incomplete Information

The Cranfield interpretation technique is examined in this aforementioned paper as being physically powerful for violations in extreme levels of the completeness hypothesis. It is presumed that all of the of the relevant data contained in documents that are within a test collection can be identified and are said to be be to be be present in the collection. It can be depicted that evaluation measures used presently lack the efficiency and robustness for the fragmented and incomplete important conclusions considerable. A new method method of measurement was introduced which can remarkably correlate with the procedures of measurements that are existent. There is an availability of total or complete conclusions which are more powerful and show high efficiency to judgement to judgement judgement sets that are fragmented. The suggestion that can be made on this decision is that and extensively large, bulky or aggressive test collections which are said to be built by the use of the pooling procedures that are existent or techniques that use the tools and equipments which belong in a lab, can even suppose the fact that information and data that is applicable maybe imperfect and partial.

3. Semantic Concept Discovery for Large-Scale Zero-Shot Event Detection

Intricate events in internet videos that are uncontrolled are detected detected are detected using the technique on which the aforementioned paper is centered. While the profusion of the training data that is labels can be placed confidence in by most of the existing works, this can be actually considered a challenging zero shot framework in which there is no provision of training input data. There is an initial pre training of a variety of of a variety of concept classifiers in in this paper that makes use of the input data that is originating as alternative alternative sources. The semantic correlation of every single concept with the respective development of interest is then calculated calculated and evaluated. The concept classifiers that are discovered on all testing videos are enforced and a couple of score vectors are acquired after performing exaggeration of prediction with further precision and taking into consideration, the discriminating management. There is a conversion of the divergent the divergent vectors into correlation matrices in pair wise order. Unanimity is then searched for by adopting a framework that is the nuclear norm rank aggregate structure. An effective and remarkably expandable algorithm is suggested here, to deal with the increase in demand of Optimisation installation comma which has a capacity order that is faster than the alternatives that are existing.

4. LTP: A Chinese Language Technology Platform

A Unified processing platform platform of the Chinese that can incorporate a suite of Natural Language Processing Processing Language Processing Processing or NLP module of high-performance and and relevant corpora is the Language Technology platform (LTP). CoNLL and SemVal are some relevant relevant evaluations that where achieved as good results especially for the the parsing modules that were syntactic and semantic. From an analysis on an internal representation of data based on XML, it is found that it is found that the modules and corpora can be easily used by human users by the process of invoking a dynamic link library (DLL) or web service service application program interface (API).

5. Chih chung Chang and Chih-Jen Lin. Libsvm: a Library for Support Vector Machines

A library for the support vector machine, SVM for the support vector machine, SVM support vector machine, SVM is the LIBSVM. This package has been on an active development says since 2000. The goal of this library is are helping users in the the method of applications of SVM to the techniques easily. LIBSVM has racked up widely up widely known popularity in the field of machine learning and many other areas. In the aforementioned article, all the details of implementation of LIBSVM are presented. theoretical convergence, SVM problems on Optimisation, estimates on probability, multiclass classification and selection of parameters are some of the issues of the issues some of the issues that are discussed in detail in this paper this paper in this paper this paper.

3. System Analysis

Software engineering and Systems engineering often go through a process of requirements analysis. Determining the needs or conditions conditions that should be met for a new or altered project or product are the tasks that encompass the fees of requirement analysis. It also takes into account the requirements that are possibly conflicting amongst the various stakeholders. it also does the task of of documentation, analysis, management of software or system requirements and validation.

Existing system: Analysis on a data set that is already stored belonging to a particular field is offered by the existing system. There are not many algorithms that are sufficient enough to be employed that enable the improvements improvements the improvements employed that enable the improvements improvements the improvements that enable the improvements improvements of the competency and accuracy of the results. The results take high time periods to be achieved. The feature selection that is appropriate for the system and the fetching of the data dynamically and the processing which uses keywords entered by the user is not allowed.

Proposed system: Multiple analytical models are produced by the proposed the proposed system. In one such system, tweets tweets system, tweets tweets such system, tweets tweets system, tweets are extracted dynamically using keywords and the Naive Bayes algorithm is applied. in the second system, there is a categorisation of data into positive, negative and neutral by searching for the polarities of the the words in the text provided provided text provided. The third system employs its techniques techniques on a data set set that is magnanimous which contains tweets. This data set undergoes cleaning, training and modelling for pertinent results to be produced. A technique known as Logistic regression is used.

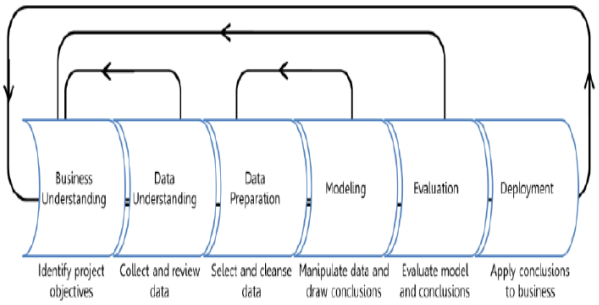

Fig. 2: Business Model

4. System Design

In the aforementioned paper, the system generated or developed for social network and sentiment analysis is described. This system is used for operation on Twitter data. Twitter social network is a platform that's contains thoughts, opinions, references to images, facts and other media. Recently, users have good online videos that are streamed live or filmed live. In this real channel of communication, a user can choose the topics and the note of reference to these topics according to the interest and culture of his choice. The communities which are part of the network and the way the propagation of information is done are able to be highlighted by a study of the number of interconnections in a node and the network topology in which the interconnections are made. Nothing can be said about the degree of agreement or the coalition that happens between the members of a community through the study. An investigation should be carried out to the semantic contents of these messages in order to solve this task. In terms of effectiveness, the sentiment analysis will be showing many difficulties compared to the problems shown by classic data mining. This problem is mainly caused because of the subtle distinction that is existing between a positive or a negative sentiments or that is existing between a neutral and a positive one. Let us take an example of a sentence that contains sarcasm or irony. Yeah, the interpretation in the meaning of the sentence is very subjective in a strict manner. In such a case, it can be a disagreement between two human beings about a real feeling that is being expressed. Furthermore, not always the expression of opinions is done through the use of words expressing opinion. In many of the cases, the construction of a special language comes into play. These are also known as the figures of speech. There are many difficulties that arise from the usage of slang or non formal expressions as this will not be belonging to the vocabulary structure of a language. A particular opinion or a mood at a certain point can be expressed in an intensive way often through the usage of search terms. There are many additional problems which arise due to the domain of the subject. In some particular cases, we can notice that certain feelings are expressed by certain words that are often topic dependent. For example let us take this sentence: "It's quiet!". If the topic of discussion was a car engine, this will be rendered as a positive opinion. But if this was the case of a phone discussion, this sentence reveals a disapproval. As a service that provides microblogging, Twitter is used to the publishing of very short messages which have account of 140 characters or tweets in maximum. On one site, this characteristic may seem easier as it can force people to take a position. Taking another side, the usage of very few words does not allow the repetition of concepts or emotions by the user. The user would rather use slang which are shared by the community or punctuations or emoticons as part of the speech structure. There is an east of retweeting which can increase the difficulty in understanding what the user is really feeling. The actual sentiment enclosed in the tweets can also be distorted by the intense use of citations. However, by the combination of information some actual cases can be hoped to disambiguate and the efficiency of the machine learning algorithms for the system can be improved by getting an opportunity to understand the slang of the channel under examination.

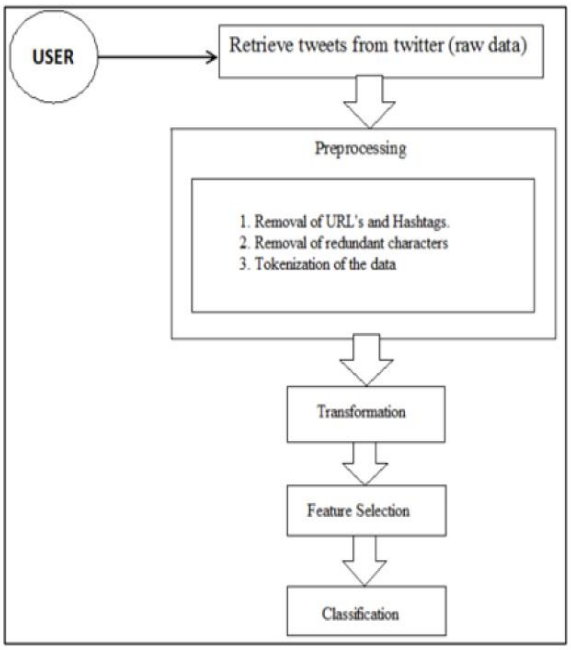

Fig. 3: System Architecture

5. Materials and Methodology

5.1. Naive Bayes Algorithm

It is calculated using the formula P(d1|h) * P(d2|H) which goes on. There can be a substantial assumption which is of the most unhappening situation in real-world data. The input data can be static or dynamic. The dynamic method is executed using the twitter api. It will return all the tweets in the category of keyword entered. It also gives the accuracy of the algorithm for classifying tweets. The accuracy achieved in this project is that of 63 percent. We took many keywords to fetch tweets, such as padmavat, karnataka elections, narendra modi etc. The concepts used for better efficiency are, in the method of machine learning, the interest of users are generally in opting the hypothesis, (h) that is the best and given data, (d). Taking into account a problem involving classification, the class for the assignment of a new data instance (d) maybe our hypothesis, (h). The easiest way of selection of the hypothesis that is highly probable, is knowledge of the data had in prior. We can also use the knowledge acquired in prior about the example in question. A method is provided by the Bayes Theorem through which the user can evaluate the probability of the hypothesis based on our knowledge acquired in prior.

Bayes Theorem can be put forward as: P(h|d) = (P(d|h) * P(h)) / P(d), where,

P(h|d), is said to be the probability, given the data d, of hypothesis h. This is also known as the posterior probability.

P(d|h), is said to be the probability, given the hypothesis h was true, of data d.
P(h) is said to be the probability of hypothesis h, regardless of the data input, being true. This process is known as the prior probability of h.
P(d) is said to be the probability, regardless of the hypothesis, of the data.
It can be seen that the algorithm is interested in the calculation of the posterior probability of P(h|d) by knowing the prior probability p(h) with P(D)and P(d|h). After the evaluation of the posterior probability for various numbers of different hypotheses known, the selection of the highest probability hypotheses can be done by the user. This is the hypothesis that is highly probable and maybe in a formal manner, be known as the maximum a posteriori(MAP) hypothesis. This can be written as: MAP(h) = max(P(h|d)) or MAP(h) = max((P(d|h) * P(h)) / P(d)) orMAP(h) = max(P(d|h) * P(h)). The probability can be calculated by the normalizing the term, P(d). We can discard of it when we are keen in the constant and the highly probable hypothesis as there is a constancy hypothesis which can be to normalize. Coming back to the process of classification, each class of the training data has an even number of instances. Then there will be a equal probability in each class (e.g. P(h)). Repeatedly, this would become a constant term that is contained within our equation and we could have limited it down so that we have been end up with: MAP(h) = max(P(d|h)) Naive Bayes is a classification algorithm for binary(two-class) and multi-class classification problems. The technique is easiest to understand when described using binary or categorical input values. It is called Naive Bayes or idiot Bayes because the evaluations of the probabilities for each and every hypothesis are simplified to make their calculation tractable. Rather than attempting to calculate the values of each attribute value P(d1, d2, d3|h), they are assumed to be conditionally independent given the target value interact. Nevertheless, the approach performs surprisingly well on data where this assumption does not hold.

5.2. Data Cleaning Algorithms

The quality of the input data is the prime concern in quality of the management of information. Problems in the quality of data in the information systems. The process focuses on removing redundant, inaccurate and incomplete data to remove errors and improve quality of data. Data cleaning though shown to be a time consuming and a tedious task, cannot be avoided. The parameters to be ensured in data cleaning are uniqueness, consistency, integrity, validity, completeness, schema conformance etc. We have used two techniques for data cleaning-
• Using Association Rules
• Using Functional Dependencies

A. Mining the quality of data using association rules

It is used to quantify, detect, correct and explain the deficiencies in the quality of data in very large databases. The quality of the data is include in addition to finding relationship between the items in huge databases. All transactions have to be checked for the level of confidence by the rule generated for all transactions known as the association rule.
B. Data quality mining using functional dependencies
In tuples, the relationship between the candidate key and the attributes is reference to by an important feature known as functional dependency. It decreases the number of functional dependencies. Though, it is a time consuming process.

5.3 Logistic Regression Algorithm

Logistic Regression is a classification technique which is preferred above linear regression. Some of the factors leading to this choice are- The regression line can take any value between negative and positive infinity. In most cases, the Y axis takes values only among 0 and 1. Therefore, the regression line almost always predicts the wrong value of Y in classification problems. The probability of the default class can be modeled by using the technique of Logistic regression. For example, if we are to model people's sex as male or female according to their height, then the first class would be given to the male and the model of Logistic regression can be written using the probability of male when given a person's height. In a more formal manner, P(sex=male|height). When written in another manner, the model of the probability that an input (X) belongs to (Y=1) which is the default class, this can be written in a formal manner as: P(X)=P(Y=1/X). It should be taken into note that the probability of the prediction has to be transformed into binary values of zero or one in accordance to actually enable the prediction of a probability. The method of Logistic regression in linear, but Logistic function is used to transform predictions. The result of this procedure is that we cannot furthermore evaluate the estimation like we do with linear regression as the combination of the inputs in a linear order. For instance, the model can be stated as: p(X) =e(b0 + b1 X)=(1 + e(b0 + b1 X)), we can transform the equation given as follows: “ln(p(X) / 1 -p(X)) = b0 + b1* X“.The output calculation that we see on the right side is linear again just like linear regression, and the input on the left side is a log of the probability of the default class and hence this is useful. This ratio on the left is called the odds of the default class (it's historical that we use odds, for example, odds are used in horse racing rather than probabilities). While calculating odds we take the ratio of the probability of the event and divided by the probability of not the event, e.g. 0.8/(1-0.8) which has the odds of 4. So we could instead write: ln(odds) = b0 + b1 X. The exponent can be moved back to the right and we write it as: odds = e(b0 + b1 X). This can help us for more clear decision that's the model is developed from a linear synthesis still of the input data, but the log odds of the default class is what the linear combination of inputs relate to. Making Predictions with Logistic Regression: The development of predictions using Logistic regression model can be deemed as simple as the plugging in of numbers into the equation of Logistic regression, and getting a result calculated.

6. Experimentation

1. Analytical model based on Twitter feeds
2. Analytical model on movie reviews
3. Analytical model on large dataset of twitter feeds on Hadoop clusters
4. Implementation of Data Cleaning

```
class TwitterClient(object):
    def __init__(self):
        # keys and tokens from the Twitter Dev Console
        consumer_key = 'ysMIgjJkTG6jiPFDPzHJuJwny'
        consumer_secret = 'gKOtlqswDUXzQYhfbcfOwtidFzTyWfOcSZNDES7ouPkpAJHIlh'
        access_token = '313175303-NEdaimn6oVgjwePGmOXPi6MkTuwqjbYDfH05dNzD'
        access_token_secret = 'leKVcqoUzvgcwxhnBauQZELUCKQQ2EgGvRNAGNJ8MMkW1'

        # attempt authentication
        try:
            # create OAuthHandler object
            self.auth = OAuthHandler(consumer_key, consumer_secret)
            # set access token and secret
            self.auth.set_access_token(access_token, access_token_secret)
            # create tweepy API object to fetch tweets
            self.api = tweepy.API(self.auth)
        except:
            print("Error: Authentication Failed")

    def clean_tweet(self, tweet):
        return ' '.join(re.sub("(@[A-Za-z0-9]+)|([^0-9A-Za-z \t])|(\w+:\/\/\S+)", " ", tweet).split())

    def get_tweet_sentiment(self, tweet):
        # create TextBlob object of passed tweet text
        analysis = TextBlob(self.clean_tweet(tweet))
        # set sentiment
        if analysis.sentiment.polarity > 0:
            return 'positive'
        elif analysis.sentiment.polarity == 0:
            return 'neutral'
        else:
            return 'negative'

    def get_tweets(self, query, count = 10):
        # empty list to store parsed tweets
        tweets = []
```

Fig. 4: Fetching and cleaning of tweets

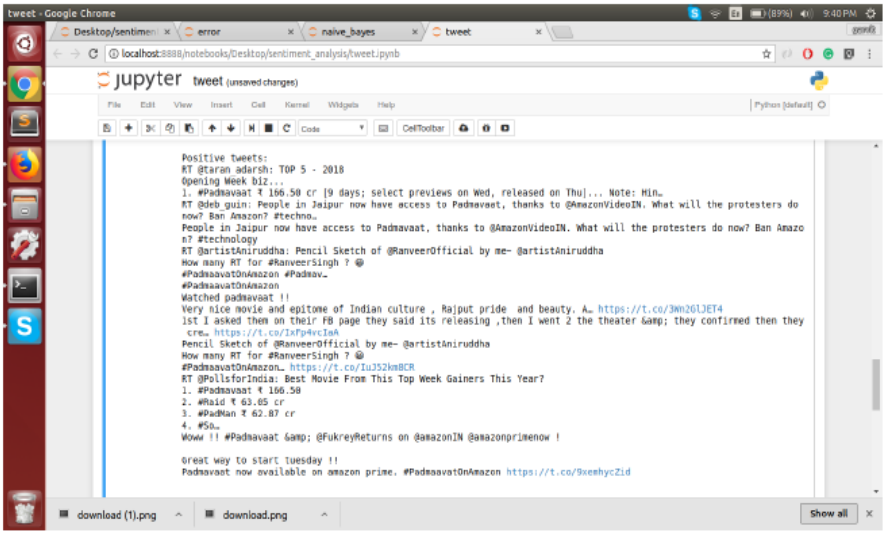

Fig. 5: Results of Tweet Analysis

```
    print "Total Negative found in positive reviews %s" % len(actual_pos_set['negative'])

    rt_negative_reviews = open(NEGATIVE_REVIEWS_FILE, 'r')

    expected_neg_set = collections.defaultdict(set)
    actual_neg_set = collections.defaultdict(set)

    for index, review in enumerate(rt_negative_reviews.readlines()):
        expected_neg_set['negative'].add(index)
        actual_sentiment = predict_sentiment(review, 'negative')
        actual_neg_set[actual_sentiment].add(index)

    print "Total Positive found in negative reviews %s" % len(actual_neg_set['positive'])

    print 'accuracy: %.2f' % nltk.classify.util.accuracy(classifier, test_data)
    print 'pos precision: %.2f' % precision(expected_pos_set['positive'], actual_pos_set['positive'])
    print 'pos recall: %.2f' % recall(expected_pos_set['positive'], actual_pos_set['positive'])
    print 'neg precision: %.2f' % precision(expected_neg_set['negative'], actual_neg_set['negative'])
    print 'neg recall: %.2f' % recall(expected_neg_set['negative'], actual_neg_set['negative'])

run_sentiment_analysis_on_rt()
```

```
First 5 positive words [({'a+': True}, 'positive'), ({'abound': True}, 'positive'), ({'abounds': True}, 'positive'),
({'abundance': True}, 'positive'), ({'abundant': True}, 'positive')]
First 5 negative words [({'2-faced': True}, 'negative'), ({'2-faces': True}, 'negative'), ({'abnormal': True}, 'negat
ive'), ({'abolish': True}, 'negative'), ({'abominable': True}, 'negative')]
Number of positive words 2006
Number of negative words 4783
Total number of words 6789
Total Negative found in positive reviews 1123
Total Positive found in negative reviews 2837
accuracy: 0.63
pos precision: 1.00
pos recall: 0.79
neg precision: 1.00
neg recall: 0.47
```

Fig. 6: Results of Naive Bayes classifier

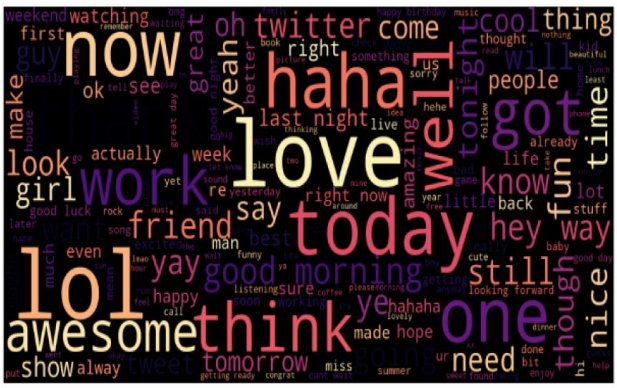

Fig. 7: View of wordcloud

1. Analytical Model on Real Time Twitter Feed Processing

Twitter is one of the biggest social networking site with more than 2 million tweets every single day. These tweets are about a lot of trending topics from all around the world. The first step becomes data extraction of relevant tweets by classifying the tweets by categories with high accuracy for improved information retrieval. In this model, the tweets are acquired in real time and the training dataset is obtained using twitter4j api. Lists of all the possible categories to which the tweets are are sent as input belongs to are returned. It also gives an accurate measurement of the efficiency of the algorithm that is used for tweets classification. The keyword is to be entered by the user in realtime, for instance Karnataka Elections, and it takes into consideration while categorizing data. A very crucial part of the analytical modeling includes data cleaning. Data cleaning is crucial for better performance. Its concepts are-

• Removal of outliers- Low or high frequency words are removed using an approach called the bag of words.
• Removal of Stop words- Most common words like the, at, is, such are removed.
• Stemming of Keyword- Reduction of the words back to the basic form, root or stem by the usage of porter stemming.
• Cleaning of the crawl data.
• Spelling correction- to correct spellings.
• Recognition of the named entity – The most appropriate word is found out along with the result category for ranking.

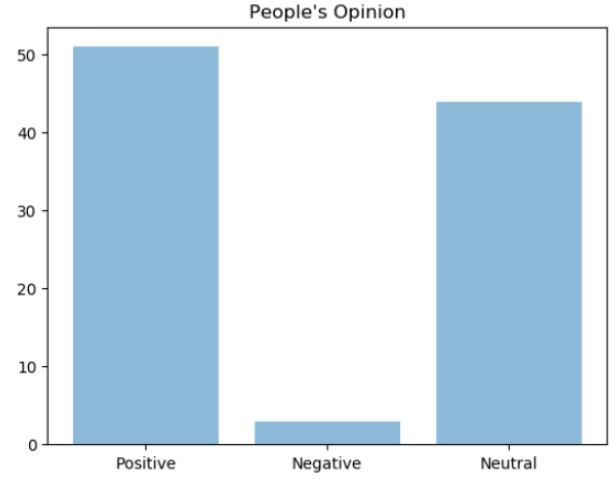

Fig. 8: Statistics of tweets

The tweet classification algorithm used in Naive Bayes. It is a supervised learning technique, of which we have discussed earlier. Validation of precision is done through cross validation. Here, the data is split equally sized into k number of subsets. Each and every subset can be used for the purpose of testing and the remainder can be used for training purposes. Often, cross validation is performed after the stratification of subsets. An overall estimate of the error can be estimated by the averaging of the errors. We have successfully achieved an accuracy of 63 percent.

2. Analytical Model on Movie Reviews

The movie reviews' prediction content is an issue of classification often known as sentiment analysis. The sentiment analysis models can be developed by using a bag-of-words model which is a very popular method which can transform the document into vectors in which every single word inside the document is associated with a score. The dataset is a collection of movie reviews from Rotten Tomatoes. There are two types of methods used for bad of words-

A. Discriminative methods

• Nearest neighbor classification
• Support vector machines

B. Generative methods

• Naive Bayes
• Probabilistic Latent Semantic Analysis

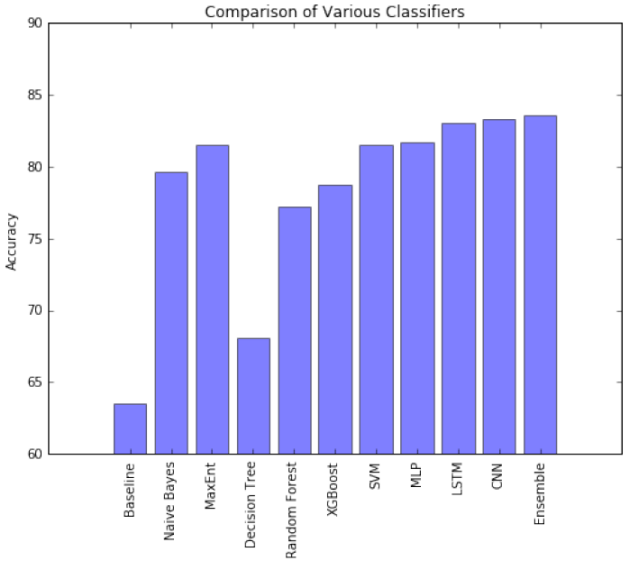

Fig. 9: Comparison of various classifiers

The inputs provided to the algorithm consist of reviews only in English. Each and every single text has undergone conversion to lowercase. Around certain punctuations like commas, periods and brackets, there is white space. There is a splitting of text that result in having one sentence occurring in a line. A few natural language processing tasks have used this data. The performance of classical models for classification (such as Support Vector Machines) on the information has a range of 70% as the high limit and 80% as the low limit. This could mean that cetain estimations have to be made after the building of the analytical model on new reviews including the ones with text. This leads to a requirement of the performance on all new reviews, of same preparation of data, as was performed on the model of the training data. This constraint has to be built, by the splitting of the training and test datasets, into evaluating the models prior to any preparation of data. Any sort of information found in the test set can indeed help in a more advanced preparation of the input data which is turn may not be available during the process of data preparation and model training. A test set is generated by a mixture of the last 100 reviews that were positive and the last 100 reviews that were negative. The remaining 1800 reviews will be kept as the dataset for training purposes. This is a 90-10 split of the data for training and testing respectively. The filenames of the reviews are used to impose the splitting easily where, the reviews are labeled as 000 to

899 which become the training dataset and those labeled 900 onwards become the dataset used to test the model. As the text data will be almost clean, much preparation won't be required. Without diving much into details, The data preparation is done using the following method:

• Splitting of tokens based on whitespaces.
• Removal of all punctuation between words.
• Removal of all words that have non-alphabetical characters.
• Removal of all words which are called stop words.
• Removal of all words with a length less than 1 character.

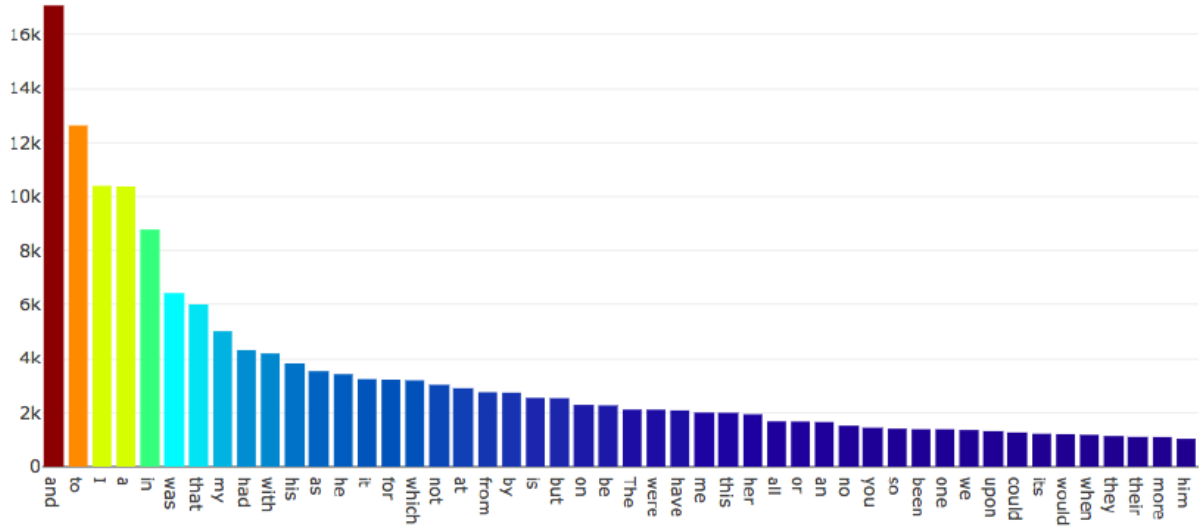

Fig.10: Frequency of occurrence of words

3. Analytical Model Based on Twitter Feeds Using Hadoop Clusters

Hadoop is a file system that is distributed and used for the storage and processing of huge volumes of data that is semi-structured, structured and unstructured. The most important aspects of tweets are-

• Tweets
• Followers
• Retweets.

Hadoop is useful to process complex data which are deeply nested and have variable schema. Also, while loading data into HDFS, files can easily be separated in creation time.

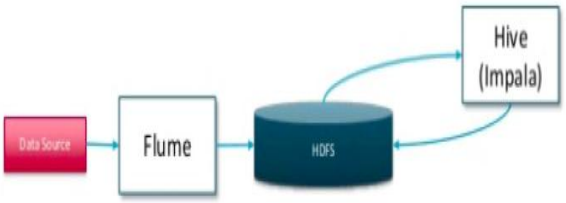

Fig. 11: A canonical form of Hadoop Architecture

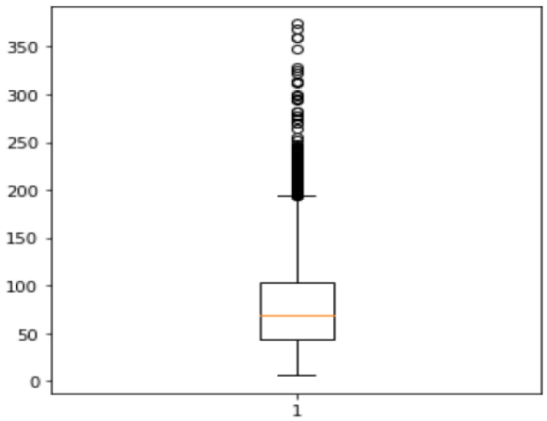

Fig. 12: Shows the distribution of words in uncleaned dataset.

7. Results and Conclusion

Mental stress is hazardous to the health of the people. It is difficult for the timely detection of stress for dedicated care. Therefore, we presented multiple analytical models for detecting, understanding and presenting the levels of users' physical stress states from humans' social media data at various levels. The project focuses on leveraging relevant content through dynamic and static means. Employment of real-world social media input data as the basis; we analysed the interrelation between users' levels of psychological stress and their behavior on social media online. It also visualizes the differences in different techniques employed. This project focussed on various aspects such as large datasets of twitter feeds, segregating twitter feeds on the basis of hashtags to extract content pertaining to a particular interest and understanding and predicting stress from long textual data obtained from film reviews by critics from the internet. The project, in the process very well engages competent algorithms such as Naive Bayes and Logistic Regression. Naive Bayes was used to sort data into positive, negative and neutral by using combinations and synthesis of words in dictionary and datasets. Logistic Regression was used to predict the outcome based on independent variables as input. Usage of Logistic Regression took us one step closer towards our stress prediction goal. The data cleaning algorithms used w, they included stripping, tokenizing, visualizing, calculation of frequency in occurrence of words, interpolated etc. to reduce the size of files to less than half of it and improve performance. We also visualized the results for better interpretation. All in all, this project was a wholesome approach to analysis of stress based on social interactions in social networking sites.

References

[1] Yuan Zhang, Jie Tang, Jimeng Sun, Yiran Chen, and Jinghai Rao(2016). "Moodcast: Emotion prediction via dynamic continuous factor graph model." IEEE International Conference on Data Mining.

[2] Andrey Bogomolov, Bruno Lepri, Michela Ferron, Fabio Pianesi, and Alex Pentland (2014). "Daily stress recognition from mobile

phone data, weather conditions and individual traits." In ACM International Conference on Multimedia, 477–486.

[3] Chris Buckley and EllenM Voorhees(2004). "Retrieval evaluation with incomplete information." In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, 25–32.

[4] Xiaojun Chang, Yi Yang, Alexander G Hauptmann, Eric P Xing, and Yao-Liang Yu(2015). "Semantic concept discovery for large-scale zero-shot event detection." In Proceedings of International Joint Conference on Artificial Intelligence, 2234–2240.

[5] Wanxiang Che, Zhenghua Li, and Ting Liu(2010). "LTP: A Chinese language technology platform." In International Conference on Computational Linguistics, 13–16.

[6] Chih chung Chang and Chih-Jen Lin(2001). "LIBSVM: a library for support vector machines." ACM International Conference on Intelligent Systems And Technology, 2(3):389–396.

[7] Frank R. Kschischang, Brendan J. Frey(2015). "Factor Graphs and the Sum- Product Algorithm." IEEE Transactions.

[8] Xiao jun Chang, Yi Yang1, Alexander G. Hauptmann, Eric P. Xing and Yao- Liang Yu(2015). "Semantic Concept Discovery for Large-Scale Zero-Shot Event Detection." Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence.

[9] Jennifer Golbeck, Cristina Robles, Michon Edmondson, and Karen Turner(2011). "Predicting personality from twitter." In Passat/socialcom 2011, Privacy, Security, Risk and Trust, 149–156.