Emotional Geography Using Social Media Data

Ms. Twishi Saran, Ms. Roshani Chede, Ms. Humera Khalif, Mr. Akshay Volvoiker

Abstract— Deciphering the Emotional frame of mind of a human being, from a text written by him is an upcoming field of research and can be considered as an advanced form of Sentiment Analysis. As a boon of the rise in social networking and utilization of microblogging tools, human beings resort to social media platforms like Twitter for expressing their viewpoints on localized events and global subject matters. The enormous volumes of text collected from these social media platforms are prolific in feelings, opinions, and emotions of people, making it a reservoir of information that can be harnessed to gauge the thoughts, ideas, and behavior of individuals all over the world. However, obtaining, investigating, and then classifying the text is the key challenge. In this paper, we propose methods to classify textual data mined from Twitter into six different emotion states namely - Happy, Sad, Anger, Fear, Disgust and Surprise as identified by Paul Ekman [3] using Machine Learning and Natural Language processing approaches and then moving on to visualizing the result with a user interface showing statistics and the exact emotion state. Through this, we can dive deeper into understanding the exact emotion the person is trying to put across by just using his words. The proposed approach can be brought into play in fields such as Opinion mining, Understanding Human Mental Health, and psychological behavior, learning how external or social events impact a community's mood thereby aiding Urban Planning.

Index Terms— Emotion Analysis, Machine Learning, Mining, Natural Language Processing, Sentiment Analysis, Twitter.

1. Introduction

THIS paper takes notice of the advancement of Machine Learning techniques and integrates it with Emotion Analysis of user generated data that is obtained from Social Media Platforms. Our approach utilizes the micro-blogging and social networking site- Twitter, to gather textual data, on which emotion analysis can be executed, through the medium of machine learning techniques. The fundamental goal is to scrutinize the results of the machine learning classification techniques that are applied on to textual data and represent a finite set of emotions.

'Emotional geography' is a subtopic within human geography, dealing with the relationships between emotions and geographic places and their contextual environments. Emotional geography specifically focuses on how human emotions relate to, or affect, the environment around them. Human beings generally express widely and openly their opinions and thoughts that often portrays their personality, well-being and even their behaviour towards the community or any external event.

With the dawn and ensuing escalation of social media and networking, there has been a surge of social media users who exercise expressing their feelings and emotions on such widespread social interaction platforms. The users generate monumental size of data by expressing their day to day happenings, which if tapped can avail an accurate analysis of an individual user's emotion. Microblogging has become a popular form of online communication in today's digital era. The bloom of social media platforms has provided its users a new channel to generate and grasp information. Twitter is a widely used Microblogging platform. Today if any online user wants to gain insights on what is happening around the world or what anybody else is talking about and wishes to spark a global conversation, they resort to Twitter. It is the Twitter tweets in the form of short texts that provide us a diverse and freely accessible accumulation of emotions, expressed openly by users on numerous topics.

Social media analytics is the process of gathering and analysing data from social networks such as Facebook, Instagram, and Twitter. It is commonly used by marketers to track online conversations about products and companies. Social media analytics is also defined as, the art and science of extracting valuable hidden insights from vast amounts of semi-structured and unstructured social media data to enable informed and insightful decision making. In our case, Emotion Analysis of the data generated via the social media platforms is a method of gaining valuable insights to the mood of the individual user. For example, according to a research reference paper [1] the tweet: “Great Christmas spent with my amazing family” expresses a happy mood and the tweet: “Feelings Hurt Tonight!” expresses sadness. Another example from a reference paper states that a tweet- "I spent a fabulous evening with my family at the Taj hotel! ", has the word ‘fabulous’. This is an indicative of the happy state the individual is in. [2] Thus, we can infer the emotions of the individual from their tweet. Since the textual data is produced by the user firsthand, it is very humane and thus intriguing to learn about how people feel.

Emotions can be expressed in many ways namely facial expressions, gestures, speech, and written text. The written form of expression purely relies on usage of words for communicating the emotion. With the practice of Sentiment Analysis the main problem is that the analysis only informs whether the text is positive, negative or neutral but fails to predict the exact feel or intensity of the user's emotion. Under Emotion analysis we can categorize the text into multiple emotion categories namely: happy, sad, anger, fear, disgust, surprise, etc. thereby giving a deeper understanding of the emotion being expressed.

Our approach aims to make use of various Machine Learning Techniques and Natural Language Processing (NLP) Methods to create a model to classify the emotional states expressed in tweets into six standard emotions identified by Paul Ekman [3]

TABLE 1
EXAMPLE OF TWEETS WITH THEIR EMOTION

Tweet	Corresponding Emotion
Current social conditions of our country, sadden me	Sad
This is a serious misuse of power!	Anger
Waiting for my dreams to come true.	Anticipation
What a pleasant surprise it was to him, that day.	Joy

Our analysis is established on the tweets crawled from Twitter and our training datasets are created utilizing the same. The paper is organised as follows: Section I provides an Introduction to our approach. In Section II we have presented a comprehensive and comparative study of the Related Work and Background Research where we have studied the various emotional models and classification models providing a base for our study. Following this, in Section III we set the Objectives for our study and Propose a Methodology that leads to Section IV of Design. Our Design highlights our commission that includes Creating an efficient and well-labelled training dataset that has tweets representing all the stated emotion classes which can be exploited by the classifiers. Performing Pre-processing techniques on the data to remove the hindrances and do away with the social media jargon to make the text appropriate for usage. Developing a model that can label the tweets based on its textual features using Machine Learning techniques and NLP methods. Comparing the results obtained and the accuracy of the classifiers. Visualizing our results through a user interface that encapsulates the working and processing involved. Section 5 stages our Experiments and Section 6 exhibits the Results and Discussions. We Conclude with the future in Section 7 and have cited the References in Section 8 with useful links.

2. Related Work

Emotional states of individuals, also known as moods, are central to the expression of thoughts, ideas, and opinions, and in turn impact attitudes and behaviour. With the advent and the subsequent rise of social networks, there has been a surge of users expressing their emotions and daily feelings leveraging the social media platform. In the future for a true smart city a more humane component is needed to understand how the population of cities actually interact with and feel about their surroundings. While measuring objective data like traffic congestion or air quality is important, it does not tell the whole story of how people live in the cities or how cities should be developed to make them more liveable. However, analysing and classifying text on the basis of emotions is a big challenge and can be considered as an advanced form of Sentiment Analysis.

Sentiment is general feel or impression people get from consuming a piece of content. It is a binary system of "positive" and "negative" responses whereas emotions are described as intense feelings, it relies on a deeper analysis of human emotion and sensitivities. For example inside the positive category it detects specific emotions like joy, happiness, excitement etc.

Emotional analysis goes a step further into target customer motives and impulses. It gives precious and exact insights that can be easily transformed into actions.

2.1. Background

Emotion classification is an evolving field of research and a considerable amount of work has been done in this field using various techniques and methodologies. Paper [5] presents an approach to analyse sentiments of social media data. The paper mentions the need and practicality of analysing sentiments of people and specifies the way to classify a given tweet into two categories i.e. positive and negative. The results are visualised in a suitable way with emotion hotspots at locations in which tweets were collected. But the done work does not give a strong and thorough classification of the tweet rather just gives sentiments over emotion.

The paper [6] describes a rule based approach, which detects the emotion or mood of the tweet and classifies the twitter message under appropriate emotional category. The authors give briefing over differences between sentiments and emotions and how their model gives a deeper information of a particular tweet. The system consists of four classes of emotions namely Happy-Active, Happy-Inactive, Unhappy-Active, and Unhappy-Inactive.

Another work with the same four categories was carried out by paper [1]. They proposed Emotex, a method of classifying

Twitter messages into the distinct emotional classes they express. It describes a new approach for automatically classifying text messages of individuals to infer their emotional states. To label Tweets twitter hash-tags were used to indicate emotion expressed by tweets. This paper uses emotion-indicative categories such as positive emotions, negative emotions, anxiety, anger and sadness to build their domain-specific dictionary. Support Vector Machine (SVM), Naïve Bayes, K-Nearest Neighbour (KNN), Decision tree classification algorithms were used to identify emotions.

Paper [4], addresses the problem of detection, classification and quantification of emotions of text in any form. This paper proposes a hybrid method involving Machine learning algorithms and natural language processing method to classify textual data into six emotion categories. The NLP method attempts to classify and score text according to the emotion words present in it. The second approach uses standard classifiers like Sequential minimal optimization (SMO) and J48 to classify tweets. Then they combine both these approaches to propose a Hybrid approach to detect emotions in text more effectively. They also introduced the concept of Surety Factor to suggest the reliability of their output and the degree of usefulness and correctness of the results. Paper [2] aims at detecting a person's mental/emotional state by using twitter tweets, keywords in each tweet are used to express what the tweet's emotion is. The emotions are categorized into 8 classes: 'Anger', 'Anticipation', 'Disgust', 'Fear', 'Joy', 'Sad', 'Surprise', 'Trust'. They followed the method of vectorization for each tweet changing its qualitative value to quantitative. Classification done using the SVM, KNN, Decision Tree & Naive Bayes algorithms. They took 20 batches of 20 tweets of varied emotions in each case and calculated accuracy for each.

2.2. Analysis of Emotion Models

Emotion classification is the means by which one may distinguish one emotion from another. Classification of emotions from one of two fundamental viewpoints:

1. That emotions are discrete and fundamentally different constructs
2. That emotions can be characterized on a dimensional basis in groupings

2.2.1. Basic Emotion Theory

Basic emotion theories propose that humans have a limited number of emotions. Ekman's theory also postulates that emotions should be considered discrete categories [3] rather than continuous. Paul Ekman proposed seven basic emotions, fear, anger, joy, sadness, contempt, disgust and surprise. He later changed it to 6 categories namely, anger, fear, sadness, disgust, surprise and joy.

2.2.2. Plutchik's Model

Plutchik developed a psych evolutionary theory of emotion and created a model of emotions where he proposed distinction between basic and complex ones [Plutchik, 2001]. Different researchers define different basic effects and Plutchik in his model proposed eight basic, fundamental emotions: joy, trust, fear, surprise, sadness, anticipation, anger, and disgust.

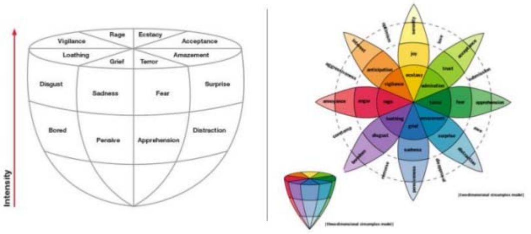


Figure 1: Plutchik's Wheel of emotion

Robert Plutchik also created a wheel of emotions to illustrate different emotions. Each primary emotion has a polar opposite such as: Joy is the opposite of sadness, Fear is the opposite of anger, Anticipation is the opposite of surprise and Disgust is the opposite of trust. A mixture of any two primary emotions may be called dyad. The emotions with no colour represent an emotion that is a mix of the 2 primary emotions. For example, anticipation and joy combine to be optimism, joy and trust combine to be love. Emotions intensify as they move from the outside to the centre of the wheel, and decrease in the other direction which is also indicated by the colour: The darker the shade, the more intense the emotion. For example, anger at its least level of intensity is annoyance. At its highest level of intensity, anger becomes rage or, a feeling of boredom can intensify to loathing if left unchecked, which is dark purple.

2.2.3. Circumplex Model

The Circumplex model of emotion was developed by James Russell. It proposes that all affective states arise from two fundamental neurophysiologic systems, one related to valence (pleasure–displeasure continuum) and the other to arousal, or alertness.

This model suggests that emotions are distributed in a two-dimensional circular space, Arousal represents the vertical axis and valence represents the horizontal axis, while the centre of the circle represents a neutral valence and a medium level of arousal. In this model, emotional states can be represented at any level of valence and arousal, or at a neutral level of one or both of these factors (for example, joy in an emotional state that is the result of strong positive valence and moderate arousal). Affective states other than joy likewise arise from the same two neurophysiological systems but differ in the degree or extent of activation. [7]

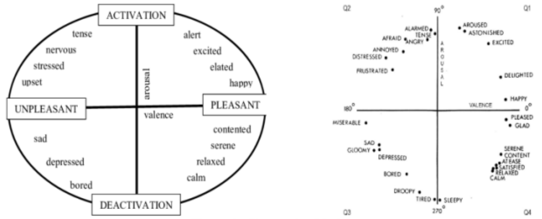


Figure 2: Circumplex Model

Circumplex models have been used most commonly to test stimuli of emotion words, emotional facial expressions, and affective states. This model postulates that the underlying structure of affective experience can be characterized as an ordering of affective states on the circumference of a circle (see Figure 2). The similarity between any two affective states is believed to be a function of their distance from one another on the perimeter of the circle with the dissimilarity between any two states increasing as the distance between them on the circle increases.

A sample of 28 words was chosen to represent the domain of affect. These were categorized into 8 categories labelled arousal, contentment, depression, distress, excitement, misery, pleasure, and sleepiness. These categories were then represented in a circular order such that (1) words opposite each other on the circle describe opposite feelings (1) words closer together on the circle describe feelings that are more similar. [8]

2.2.4. Vector Model

This two dimensional model consists of vectors that point in two directions, representing a "boomerang" shape. The model assumes that there is always an underlying arousal dimension, and that valence determines the direction in which a particular emotion lies. This results in two vectors that both start at zero arousal and neutral valence and proceed as straight lines, one in a positive and one in a negative valence direction. [9]

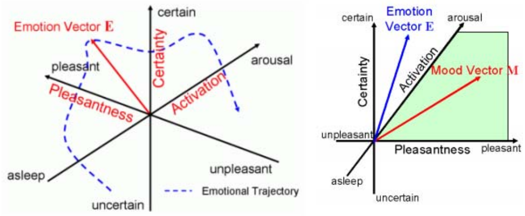


Figure 3: Vector Model

For example, a positive valence would shift the emotion up the top vector and a negative valence would shift the emotion down the bottom vector. In this model, high arousal states are differentiated by their valence, whereas low arousal states are more neutral and are represented near the meeting point of the vectors. Vector models have been most widely used in the testing of word and picture stimuli.

2.2.5. Lövheim Cube of Emotions

Lövheim proposed a cubic model that relates three levels of neurotransmitters and eight basic emotions labelled according to Tomkins. The most important monoamine neurotransmitters are serotonin, noradrenaline and dopamine, are essential in the control of behaviours and emotions.

A three-dimensional model, the Lövheim cube of emotion, was presented where the signal substances forms the axes of a coordinate system (serotonin is represented on the x-axis, noradrenaline on the y-axis and dopamine on the z-axis), and eight basic emotions are placed in the eight corners. The origin represents a situation where no signal substances at all are released. The other end of each arrow represents the maximum effect of the specific neurotransmitter system. The corners of the cube thus represent the combination of the extreme values, either low or high on the three axes respectively. An infinite number of combinations of different levels of the three neurotransmitters are possible, but all lie within this space, and within the eight ‘’extreme values’’, defined by the eight possible combinations of either zero or maximum effect of the three monoamine systems respectively.

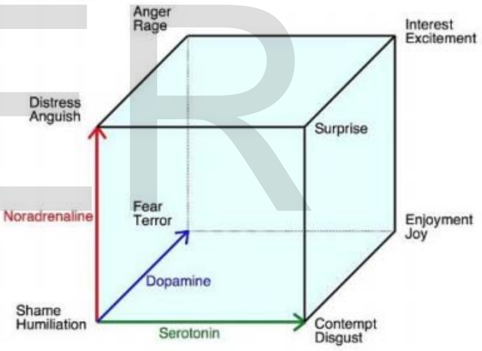


Figure 4: Lovheim cube of Emotions

According to psychologist Silvan Tomkins theory, the eight basic emotions are: Two positive: Interest/excitement and enjoyment/joy, one neutral: Surprise/startle, and five negative: Distress/anguish, fear/terror, shame/humiliation, contempt/disgust and anger/rage. Anger is, according to the model, for example produced by the combination of low serotonin, high dopamine and high noradrenaline. [10]

2.2.6. Positive Activation – Negative Activation (Pana) Model

The positive activation – negative activation (PANA) or "consensual" model of emotion, originally created by Watson and Tellegen, suggests that emotional experience can be reduced to two large, bipolar dimensions: Positive and Negative Affect. Similar to the vector model, states of higher arousal tend to be defined by their valence, and states of lower arousal tend to be more neutral in terms of valence. [9]

In the PANA model, the vertical axis represents low to high positive affect and the horizontal axis represents low to high negative affect. Terms within the same octant are highly positively correlated, whereas those in adjacent octants are moderately positively correlated. Words 90° apart are essentially unrelated to one another, whereas those 180° apart are opposite in meaning and highly negatively correlated.

Positive Affect, represents the extent to which a person avows a zest for life whereas Negative Affect is the extent to which a person reports feeling upset or unpleasantly aroused. These two factors have been characterized as "descriptively bipolar but effectively unipolar dimensions" to emphasize that only the high end of each dimension represents a state of emotional arousal whereas the low end of each factor is most clearly and strongly defined by terms reflecting a relative absence of affective involvement (e.g., calm and relaxed for Negative Affect, dull and sluggish for Positive Affect). Although the terms Positive Affect and Negative Affect might suggest to some readers that these mood factors are opposites (i.e., negatively correlated), they are in fact independent, uncorrelated dimensions.

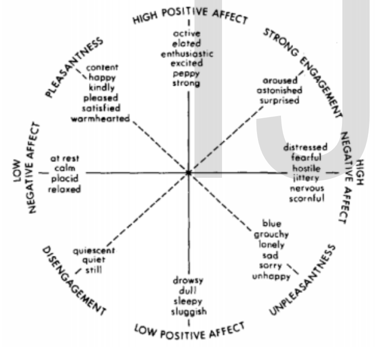


Figure 5: PANA Model

An alternative rotational scheme is indicated by the dotted lines in the figure. The first factor (the second principal component) is the Pleasantness-Unpleasantness factor (first dimension in studies of facial expressions and mood words). Similarly, the second factor (the second principal component) is the Arousal factor (called Strong Engagement-Disengagement). The Pleasantness octant contains terms representing a mixture of high Positive Affect and low Negative Affect; conversely, Unpleasantness includes words combining high Negative Affect and low Positive Affect. Terms denoting Strong Engagement have moderately positive loadings on both mood factors, whereas those representing Disengagement load negatively on each dimension.

2.2.7. Pleasure, Arousal, Dominance (Pad) Model

The PAD emotional state model is a psychological model developed by Albert Mehrabian and James A. Russell to describe and measure emotional states. PAD model of emotion views emotion as a space described with a three dimensions: pleasure / displeasure, arousal / non-arousal, and dominance / submissiveness. [11]

The Pleasure-Displeasure Scale measures how pleasant or unpleasant one feels about something. Anger and fear are unpleasant emotions (displeasure), however joy is a pleasant emotion. The Arousal-No arousal Scale measures how energized or soporific one feels. Rage has a higher intensity or a higher arousal state than anger, however boredom, which is also an unpleasant state, has a low arousal value. The Dominance-Submissiveness Scale represents the controlling and dominant versus controlled or submissive one feels. Fear and anger are unpleasant emotions, anger is a dominant emotion, while fear is a submissive emotion. [12]

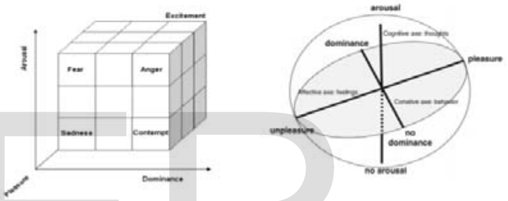


Figure 6: PAD Model of Emotion

The PAD model is a useful tool in understanding the environments that individuals may find themselves in and the three components of the PAD model are likely to impact each other. Dominance has a direct and positive impact on both arousal and pleasure. Arousal also has a direct and positive impact on pleasure. [13] The PA part of PAD was developed into a Circumplex model of emotion experience, and those two dimensions were termed "core affect". The D part of PAD was re-conceptualized as part of the appraisal process in an emotional episode (a cold cognitive assessment of the situation eliciting the emotion).

2.3. Classification Techniques

Classification of emotions is a task wherein the aim is to detect and recognize types of feelings/emotion through the expression of texts. There are different ways in which various researchers have classified sentiments or emotions of text.

2.3.1. Rule Based Approach

In this approach, classification is done by using a collection of "if…then" rules. The Left Hand Side is rule antecedent or condition and the Right Hand Side is rule consequent. Coverage of a rule is the fraction of records that satisfy the antecedent of a rule and accuracy of a rule is the fraction of records that satisfy both the antecedent and consequent of a rule. There are various advantages such as they are highly expressive as decision trees, easy to interpret, easy to generate, can classify new instances rapidly and their performance is comparable to decision trees.

Rule extraction is done to build a rule based classifier in the following manner. To extract a rule from a decision tree, one rule is created for each path from the root to the leaf node. To form a rule antecedent, each splitting criterion is logically passed through an AND operation. The leaf node holds the class prediction, forming the rule consequent.

2.3.2. Bag Of Words Approach

In this approach, the text is turned into fixed length vectors. This approach is often used for Natural language processing tasks like text classification. Each document is taken and the count or frequency of words in that document is recorded in a form of vector i.e. each word count acts as a feature. Any information about order of the words is entirely discarded and a bag of words with its frequency score is recorded.

Bag-of-words is used as a tool for feature extraction. It is simple to understand and implement but can lead to high dimension feature vectors due to large amounts of data.

2.3.3. Knowledge Based Approach

In this approach, classification is done as the function of some keywords. The main task is the construction of emotion discriminatory-word lexicons that indicate a particular class. The polarity of the words in the lexicon is estimated prior to the analysis work. There are various methods available for lexicon creation. For example, lexicons can be created by starting with some seed words and then using some linguistic heuristics to add more words to them, or starting with some seed words and adding to these seed words other words based on frequency in a text. For certain applications, there are publicly available discriminatory word lexicons for use in sentiment analysis.

2.4. Machine Learning Models

2.4.1. Naïve Bayes

Naive Bayes is a classification algorithm based on the application of Bayes theorem. The principle of naive Bayes classification algorithm is that, for the given content to be classified, calculating the emergence probability of each label under the conditions of the content features appearing, and then taking the label with largest probability as this content's label. For text classification, it is to see the text feature words.
The probability whose feature word appears relatively large will match the characteristics corresponding label. The following is the normal definition of Naive Bayes classification:
Set $X = \{x_1,\ x_2, x_3\ , \ldots \ldots . x_m\ \}$. X is the given content to be classified. The elements in X are the features of the content. $Y = \{y_1, y_2, y_3, \ldots \ldots y_n\}$. Y is the collection of different labels.

Calculating $P(X), P(X) \ldots \ldots P(X)$. If $P(X) = \{P(X), P(X), \ldots \ldots . . P(X)\}$ then X ∈ y_k [14]

The Naive Bayes algorithm assumes that all the features are independent of each other (hence the name Naive Bayes). It is represented by a document as a bag of words. This is a disturbingly simple representation: it only knows which words are included in the document and how many times each word occurs, and throws away the word order. Naïve Bayes classification is nothing more than keeping track of which feature gives evidence to which class. [6]

2.4.2. Support Vector Machine (SVM)

SVM is a supervised learning algorithm that analyses the data and recognizes patterns used for classification. SVM takes the set of training data and marking it as part of a category then predicts whether the test document is a member of an existing class. [15]

SVM model represents the vectors or points in space, mapped so that the vectors are grouped as a cluster and the division gap in between two clusters is widest possible. Then, classification is done by mapping the test vector; the vector is predicted to belong to that particular class based on which side of the gap it is. SVMs are also capable of performing non-linear classifications. [6] When using SVM, the high dimensional space does not need to be dealt with directly. To obtain the hyperplane, SVM does not deal with all data. It only considers support vectors. This avoids overfitting. [16]

2.4.3. Sequential Minimal Optimization (SMO)

Is an algorithm for solving the quadratic programming problem that arises during the training of SVM. SMO is widely used for training support vector machines and is implemented by the popular LIBSVM tool. It can quickly solve the SVM QP problem without any extra matrix storage and without using numerical QP optimization steps at all. SMO decomposes the overall QP problem into QP sub-problems, using Osuna's theorem to ensure convergence.

Unlike the previous methods, SMO chooses to solve the smallest possible optimization problem at every step. For the standard SVM QP problem, the smallest possible optimization problem involves two Lagrange multipliers, because the Lagrange multipliers must obey a linear equality constraint. At every step, SMO chooses two Lagrange multipliers to jointly optimize, finds the optimal values for these multipliers, and updates the SVM to reflect the new optimal values.

SMO is an iterative algorithm for solving the optimization problem described above. SMO utilizes the smallest possible QP problems, which are solved quickly and analytically, generally improving its scaling and computation time significantly. [17]

2.4.4. K Nearest Neighbor

K-Nearest Neighbors (KNN) is a classification algorithm that uses a distance function between the train data to test data and the number of nearest neighbours to determine the classification results. Distance function used in this experiment is the cosine similarity. Cosine similarity is one of the functions that are widely used in the document classification to find similarity between some documents. Determining document class is done by voting on K nearest neighbour. The nearest neighbour is the K- document with the highest similarity value. [15]

The main idea of this classification algorithm is that there is a data set has been categorized, when adding some new data into it, K nearest data from the new should be found in the data set; and then observing which category most of these k data belong to, then putting the new data in that category. [14] When this classification algorithm is used for text categorization, the best parameters should be tested in the classifier to make the effect of categorization best.

2.4.5. Decision Tree

A decision tree is a tree/flow chart structure. It will return an appropriate label based on the value of a given input in the flow chart. It has decision nodes and leaf nodes. The decision nodes check features of the input value and the leaf nodes match label for each input value according to the feature. To choose the label for an input value, the flowchart begins at the initial decision node known as root node of a decision tree. This node contains a condition that is used to check the one of input value's features and selects a branch according to that feature's value. Following the branch that describes the input value, it arrives at a new decision node with a new condition. Then this flow path goes on until it arrives at a leaf node which will provide a label for the input value. [14]It helps to predict what will be the output, given certain input variables. It represents all the variables involved in the decision and what are the consequences of each case. The decision tree is a greedy algorithm and it follows a top down approach. It partitions the data recursively at each step based on some input conditions.

ID3 is one of the Decision tree algorithms which generates a decision tree using a dataset. Initially there is a dataset and for each iteration and for every attribute, the entropy and information gain is calculated. It then selects the attribute which has the smallest entropy (or largest information gain) value. The set S is then split by the selected attribute to produce subsets of the data. The algorithm continues to recurse on each subset, considering only attributes never selected before. [6]

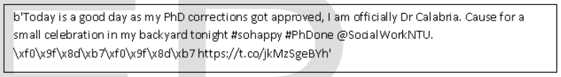
2.5. Dataset And Pre-Processing

2.5.1. Dataset Collection

In order to collect data from twitter one is required to create a developer account on twitter and generate a secret and consumer key, which allows the user to extract the tweets. Once done, Twitter's Streaming API is used with the help of Python programming language's tweepy library to gather the dataset. Tweepy is an easy to use python library for accessing the Twitter keys, API straight from the Tweepy website https://www.tweepy.org/. The Twitter API exposes dozens of HTTP endpoints that can be used to retrieve, create and delete tweets, retweets and likes.

Based on the authentication tokens program streams the data from twitter and stores it in a csv file namely with column details like date of creation, the text(tweet), retweet count, hashtags, followers count, friends count, location of the user and more. The textual data in the tweets is enclosed within b' or b'', starts with RT if it's a retweet, followed by mentioning the user id of the original author

The textual information contents: User id's (@handle_name) this provides reference to user handles, URL, escape sequence, Unicode characters and hashtags.



```
b'Today is a good day as my PhD corrections got approved, I am officially Dr Calabria. Cause for a
small celebration in my backyard tonight #sohappy #PhDone @SocialWorkNTU.
\xf0\x9f\x8d\xb7\xf0\x9f\x8d\xb7 https://t.co/jkMzSgeBYh'
```

2.5.2. Getting Labelled Data

The twitter data that we collect does not come with pre-defined labels. In order to get tweets with pre labels there are multiple approaches.

Twitter message features such as hash-tags and emoticons are likely to be useful features for sentiment and emotion classification. The usage of hashtags in tweets is very common, and Twitter dataset contains millions of different user-defined hash-tags [1].

Different hashtags related to Happiness, Sadness, Anger, Disgust, Fear and Surprise were collected. To collect the hashtags, emotion words mentioned in Circumplex Model and Plutchik's model were used. Each word from these models were looked up for their synonyms in thesaurus. Also emotion specific hashtags such as #veryhappy were added to the list [1] 40 hashtags for each emotion class was collected.

These tags were then fed to twitter API as keywords to get the tweets. Each tweet was also simultaneously labelled as an emotion category to which the hashtags belong.

2.5.3. Need For Pre-Processing

Pre-processing is the first step in text classification and choosing the right pre-processing techniques can improve classification effectiveness. [18] Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often not complete or consistent, and also lacks in certain trends or behaviours, and is bound to have some errors. It is a proven method of resolving such issues. Data pre-processing prepares raw data for further processing [19]

2.5.4. Pre-Processing Techniques

1. Stop word Removal

Stop words are a division of natural language. The motive that stop-words should be removed from a text is that they make the text look heavier and less important for analysts. Removing stop words reduces the dimensionality of term space. The most common words in text documents are articles, prepositions, and pronouns, etc. that do not give the meaning of the documents.

2. Incorrect Words and Slang Replace

Social media users usually write in an informal way and their texts contain a lot of slang and abbreviations. These words, in order to be interpreted correctly, have to be replaced to impute their meaning.

3. Lemmatization

It is the process of grouping together the different inflected forms of a word so they can be analysed as a single item. It does morphological analysis of words.

4. Replace URLs and User Mentions

In Twitter texts, almost every sentence contains a URL and a user mention. Their presence does not contain any emotions. We can either use the tags 'URL' and 'USER' to replace the urls and user mentions in the tweet text or completely remove them from the text.

5. Unicode

The textual data is decoded from the uft-08 standard thus may contain Unicode for special characters this code have to be looked up and converted for better understanding of the tweet text.

6. Hashtags

The tweet text sometimes contains many hashtags, we can use the hashtags by removing the '#' from the tags and using the textual part only.

7. Tokenization

This setting splits the documents into words/terms, constructing a word vector, known as bag-of-words.

8. Negation handling

Negations are those words which affect the sentence and change its polarity. Words like not, never, can't and so on are negation words.

3. OBJECTIVES

Our model aims at classification and analysis of human emotions through social media interactions in a region. Social media is a large platform where people share their views, opinions and thus expressing emotions through their posts.

Emotion analysis is a way to study these human interactions and classify them into different categories such as Happiness, Sadness, Fear, Anger, Surprise and Disgust.

3.1. Proposed Methodology

Since people usually give their opinions and views on social media platforms such as Twitter, Facebook and so on, the main source of data in the twitter data set is obtained from twitter streaming API. A list of 40 hashtags belonging to each of the six emotion categories are provided as keywords to extract the tweets. The tweets are also labelled at the same time. Data is then pre-processed to remove noise. Noise includes various elements such as URLs, hashtags, user handles and so on.

The classification methods involve two parts: first, use the pre-labelled processed data to train SMO, Naïve Bayes and J48 classifiers using the Weka tool. Second, classify the data using the NLP technique of vectorization. We have also proposed a hybrid way where in the NLP labelled data is used to train the classifiers.

4. DESIGN

4.1. Conceptual Design

Fig. 7. Stages of the design

4.2. Gathering Data

Many users using social network express their emotions and daily feelings on platforms like twitter. We can detect a person's mental and emotional state through his tweets and classify emotions. Therefore Twitter was chosen as the social media to gather data for emotion analysis.