Emotional Influence in Social Networks

Siri Chandana S
SU ID-829955672
MS., CS First Year
CIS700-Social Media Mining
Syracuse University

Sachin Ramesh
SU ID-384471510
MS., CS First Year
CIS700-Social Media Mining
Syracuse University

Introduction

Social Media emerged in the late 90's, it was during this time that it took various forms over the internet. The various social media platforms have gained popularity over the years and attracted many people towards them. Even while most of them vary in functionality, the primary cause for their existence remains the same. The main reason why these platforms have attracted many humans is because they help in portraying the social characteristic that exists within us. It is in the nature of human beings to communicate or express ones thoughts. While humans express different thoughts through the social media platform, we believe that there are lot of emotions flowing within them. It is true that not all forms of expressions hold emotions, but majority of them certainly do. It can be seen that throughout all forms of social media that people post or comment with emotions like happiness, sadness, angry, fear. Our project is based on this underlying assumption that emotions that people hold drive social networks.

Facebook and twitter are considered to be the most popular social networking sites where any piece of information can be tweeted or liked. Twitter is also the most researched domain of social media. Another such social networking site that we have chosen for our project is Quora. Quora is a platform where humans tend to express emotions indirectly. Due to the question and answer method of communication that Quora uses, there is a lot of open space for constructive discussions. We believe that it is a platform where one shares his opinions in an honest way or expresses ones ideas related to a certain topic. These opinions that people hold can be based on facts and we consider this factor during our project. Even though facts do play a role in opinions there are some topics where emotions seem to play the important role. That is one of the reasons why we have picked the topic - 'Kashmir Conflict'. This topic deals with the territorial dispute of the Kashmir region between the Indian and Pakistan government. It dates back to the days of India and Pakistan separation. While there may be some facts related to governments involved in this, we feel that emotions of individuals play a vital role in this discussion forum. In this project, we wish to explore the depth of emotions held within individuals in this forum.

As emotional analysis is the most researched topic in recent times, there are many algorithms that have been developed to evaluate the emotions behind writings of humans. Among the many that exist, we will be using the VADER Sentiment analysis algorithm to scrutinize the responses given by individuals on this topic. Also, as our project focuses on the depth of emotions or a variety of emotions, this factor will be evaluated by clustering the users. By keeping our assumption in mind, we will be analyzing our results with known factors. In Quora, every topic has a section called the 'most viewed writers'. The results that we obtain in our project will be compared with this factor and this will help us evaluate the emotions that users hold within this forum. As we believe that emotions drive social networks, our project will be giving these individuals an emotional metric. This metric will be based on the emotions their writings hold and it will be estimated based on the language that they use, the emphasis of punctuations and considering all the factors related to writings that showcase emotions. As the topic that we have chosen has a wide range of emotions like support, anger, fear within it, we are confident that the dataset that we will extract from this topic will provide us interesting information.

Prior Work

Sentiment analysis is extensively performed on social media websites to understand the opinions of users under various contexts. Even though most of the analysis is subjective, logical objectives have been deduced from these analyses. Some of the related works in this field are mentioned below-

1. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text (C.J. Hutto Eric Gilbert)- Valence Aware Dictionary and sEntiment Reasoner(VADER), analyses emotions, opinions and sentiments based on the computational treatment of the subjectivity in the text. VADER method has been implemented to analyze ground truth in multiple domain context.
    - Social Media Context: 4000 tweets were pulled from twitter's public time line 200 contrived tweets that specifically test syntactical and grammatical conventions of conveying differences in sentiment intensity. A 0.881 correlation to ground truth has been observed
    - Movie reviews: includes 10,605 sentence-level snippets from rottentomatoes.com. The snippets were derived from an original set of 2000 movie reviews (1000 positive and 1000 negative) in Pang & Lee (2004); they used the NLTK tokenizer to segment the reviews into sentence phrases, and added sentiment intensity ratings. 0.451 co-relation to ground truth.

- Technical product reviews: includes 3,708 sentence level snippets from 309 customer reviews on 5 assorted products. 0.565 co-relation to ground truth.
- Opinion news articles: includes 5,190 sentence-level snippets from 500 New York Times opinion editorials. 0.492 co-relation to ground truth.

2. EmpaTweet: Annotating and Detecting Emotions on Twitter (Kirk Roberts, Michael A. Roach, Joseph Johnson, Josh Guthrie, Sanda M. Harabagiu)- This research emphasizes on Micro-blogging services such as Twitter. Seven emotions are being assigned to the corpus collected from twitter. Authors also analyze how emotions are distributed in the data that is being annotated and compare it to other corpora. It idealizes that an emotion can be divided into three main categories namely negative, positive and neutral. These categories can be further divided in to emotions such as anger, disgust, sadness, fear as subset of negative emotion; joy, love as subset of positive emotion and fear and surprise as the subset of neutral emotions. These emotions ontology is created to aid the annotators to understand how emotions relate. The data is downloaded for most emotions driven topics from Twitter API using hashtag queries, enforcing maximum overlap of 0.8 hashtags, punctuations and the urls were removed from raw data. Annotation was divided into initial teaching phase, independent annotation phase and bulk annotation phase to maximize number of annotations. Linguistic analysis is done by grouping different words according to their psychological meaning to different emotional group. In our project we have grouped words according to three basic groups namely negative, positive and neutral. We use the values assigned to these words to calculate the total value of the sentence and hence the whole answer provided for certain questions.

Algorithms

Over the wide variety of algorithms that are available to us, it is important to chose the right one which gives us appropriate results with effective computation mechanisms. For our project we have used three main algorithms which are explained below-

1. **Crawler-** The dataset that is extracted consists of questions and the answers for each question. CSV files are used to store this data. Also, each answer of the question is mapped to the users who has answered this question. For crawling the contents on Quora we have used the following list of libraries in python-
   - **Beautiful Soup**- It is available in bs4 package and is an efficient library to pull out data from HTML and XML files. It uses HTML Parser to extract the data from the pages. The beautiful soup object represents the document from which the data will be extracted and using this object various html tags like paragraph tag, a tag can be accessed.

- **Selenium-** This is another package that was used for crawling data. It provides us with the facility to access the chrome driver. So, the Selenium Server is necessary for running the Remote Web driver.
- **Urllib**- This package consists of many useful methods like the urlopen() and read() which have been used to read the data. These methods are very helpful in accessing the data present in the world wide web. It also contains the request package, which is used to verify the SSL certificate for certain web pages.
- **HTMLParser**- This is a parser that parses through the entires content on HTML file. With the help of this parser the links and answers content data can be extracted and stored in our CSV files.

2. **Sentiment Analysis Algorithm-** The sentiment analysis algorithm that suits our project requirements is VADER. It was developed by Hutto, C.J. & Gilbert, E.E. (2014) especially for the social media text. It a rule based algorithm developed in python. It measures the given text between a range of -1 to 1 where 1 is positive and -1 is negative. Also, for each text it gives four measures which are positive, negative, neutral and compound. The compound measure is given by considering the positive, negative and neutral values(overall metric). For our project we will be using the compound measure to understand the distribution. This algorithm works on lexicons and has large range of lexicons that it takes into consideration. The algorithm can analyze slang words, emoticons, different languages and emphasis of punctuations. It also works to analyze certain english idioms. Along with this we will also be using the NLTK library provided by python to tokenize our answers into sentences. In this algorithm each of the lexicon is given a valence score and the compound score is calculated based on this lexicon. The score is normalized so that it ranges from -1 to 1. Some examples of certain sentences and their compound scores are given below-

- Showing emphasis on the punctuations and the words used, the following are the compound scores of the slightly modified sentences.

  VADER is very smart, handsome, and funny. {"compound': 0.8545}
  VADER is VERY SMART, handsome, and FUNNY. {'compound': 0.9227}
  VADER is VERY SMART, handsome, and FUNNY!!!{'compound': 0.9342}

- For some tricky variations of a sentence, the VADER gives compound scores as shown in the below sentences.

  Sentiment analysis has never been good. {'compound': -0.3412}

Sentiment analysis has never been this good!{'compound': 0.5672}
Most automated sentiment analysis tools are shit.{"compound': -0.5574}

- Below is an example of how VADER deals with paragraphs, if we consider the paragraph as- “It was one of the worst movies I've seen, despite good reviews. Unbelievably bad acting!! Poor direction. VERY poor production. The movie was bad. Very bad movie. VERY BAD movie!”. Then the VADER will analyze the paragraphs sentences as follows-

  It was one of the worst movies I've seen, despite good reviews.------ -0.7584
  Unbelievably bad acting!!---------- -0.6572
  Poor direction.---------- -0.4767
  VERY poor production.---------- -0.6281
  The movie was bad.---------- -0.5423
  Very bad movie.---------- -0.5849
  VERY BAD movie!---------- -0.7616
  AVERAGE SENTIMENT FOR PARAGRAPH: -0.6299

3. **Clustering Algorithm-** Clustering is the process of grouping the objects based on some similarity measure. Cluster is a collection of similar objects and these objects are dissimilar to objects of other groups. There are many clustering algorithms like centroid based clustering, density based clustering and so on. For our project, we will be using the k-means clustering algorithm. In the k-means clustering algorithm, there will be k clusters in which the users of the topic will be grouped into. The advantage of using k-means clustering algorithm is that we can see the variations of emotions in the clusters by varying the size of the clusters. In the clustering algorithm that we have written we use the Euclidean distance to find the clusters. The most important aspect of our algorithm is the plotting of the centroids for each of the clusters. This centroid shows where the clusters’ points are actually located and when we compare one cluster centroid with another cluster centroid then we will understand the variations in the clusters.

$$\mathrm{d}(\mathbf{p},\mathbf{q}) = \mathrm{d}(\mathbf{q},\mathbf{p}) = \sqrt{(q_1-p_1)^2 + (q_2-p_2)^2 + \cdots + (q_n-p_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n}(q_i-p_i)^2}.$$

The formula given above is the Euclidean Distance Formula that our project uses to find the different clusters. The clusters that the users are being divided into are based on the three values(negative, neutral and positive) which the VADER algorithm gives.

Expected Results

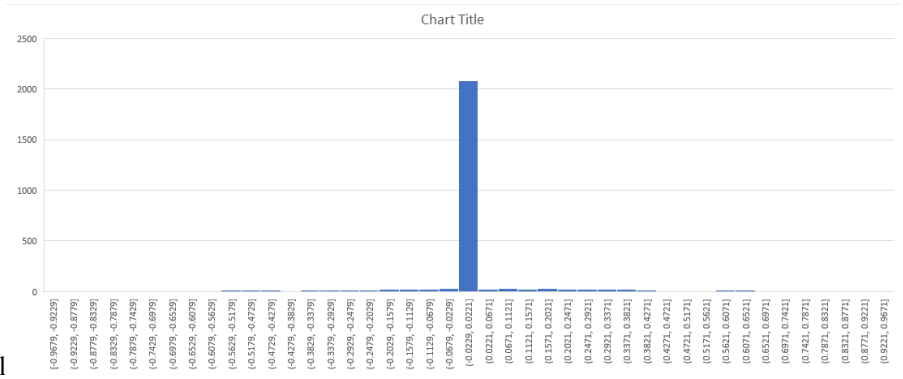
As our project focus on the emotions of the users, given below is a list of some of the findings that our project seeks to make-

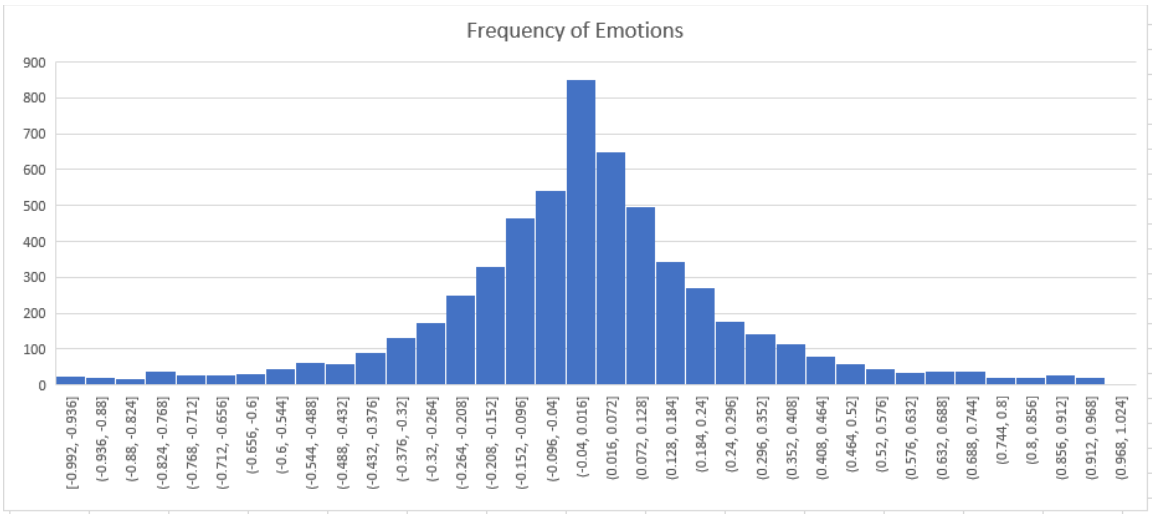
- When we measure the polarity of the users, the polarity values for the users usually will lie in 3 ranges positive, neutral and negative. There will be more overlapping of the clusters that we create and neutral cluster will be more dense when compared to other cluster
- When dividing these users into two clusters, we expect our project to yield overlapping clusters and the centroid values to be close. This is because even though we understand that users of different emotions are present in the forum, the variation of emotions is not completely distinct or extreme.
- When including the most viewed writers in our clusters, we expect them to belong to a single cluster and wish to identify which emotional cluster they belong to.

Observed Results

Crawler extracted the profiles of the users and the answers that they have given for each question. The crawler that we have used accessed the most active users and most recent questions from the URL link(https://www.quora.com/topic/Kashmir-Conflict). It has extracted the profiles of 5096 users. The total number of answers that were available on the topic for various questions are 10512 unique answers. The questions that the crawler has extracted are the ones posted in the past 2 years and some of these questions were posted long time ago but they were a topic of discussion in the past 2 years. The total number of questions that were analyzed are 2616 out of 6000. In these 6000 questions there are around 1700 unanswered questions.

Sentiment Analysis algorithm ‘VADER’ was used to analyze the sentiments of questions and the answers. Based on the emotional quotient which is the four values that VADER gives as output, our project assigned an average emotional quotient to the users. Based on the results obtained 2 different bar plots can be obtained(for these bar plots the compound sentiment metric is considered)-





The fig 1 bar plot is emotional quotient vs the total number of questions, while the fig 2 bar plot is emotional quotient vs total number of users. According to fig 1, most of the questions lie in the [-0.22, 0.22] interval, while most of the answers lie in the [-0.04, 0.016] interval.

The K-means clustering algorithm uses the three value- negative, positive and neutral and based on this emotional quotients, the algorithm has created clusters. While clustering each of the cluster has computed a centroid(shown in red). The following is the number of clusters that were created for the users-

1. Cluster Size 2- In this case , the users were mainly grouped as positive and negative. The dot points belong to cluster 0 and the square points belong to cluster 1.The cluster 0 which is represented in blue color is the negative cluster and the cluster 1 which is represented in the green color is the positive cluster. The plot is given below-

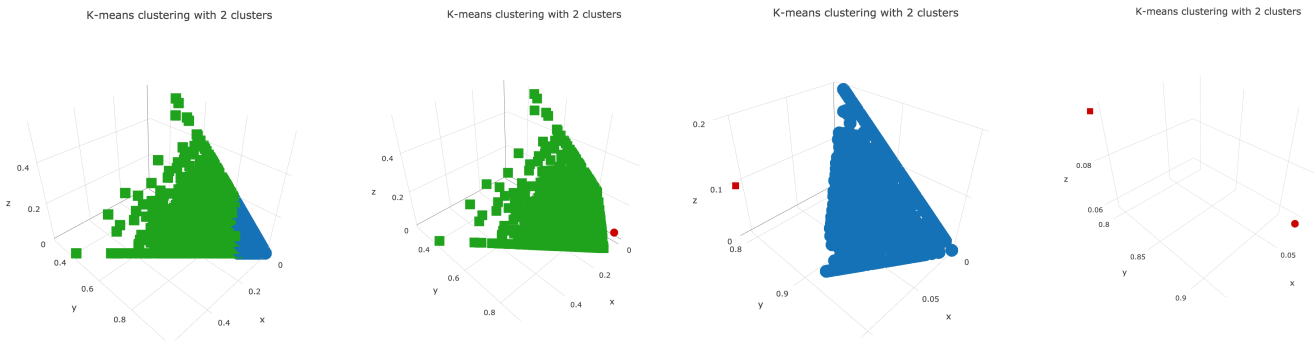


Fig 3- Grouping the profiles into cluster size =2

2. Cluster Size 3- For this the users were grouped as negative, neutral and positive. The dots belong to cluster 0, the rectangles belong to cluster 1 and the rhombus belongs to cluster 2. The

blue color and purple color belong to negative and positive respectively . While, the green belongs to neutral. The following is the plot-

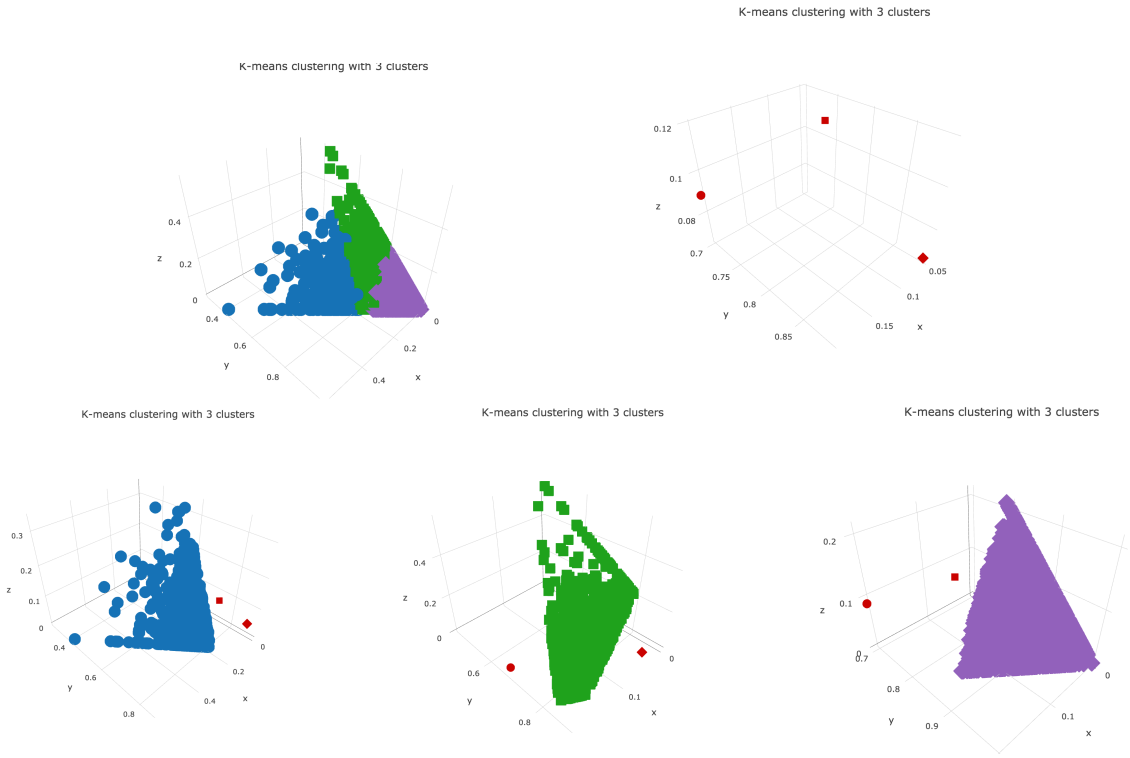


Fig 4- Clustering users with cluster size 3

3. Cluster size 5- In this case, the users were divided into groups as most negative(blue color, filled circles), least negative(green, filled rectangles), neutral(purple, rhombus), least positive(pink, empty circles) and most positive(yellow, empty rectangles). The plot is given below-

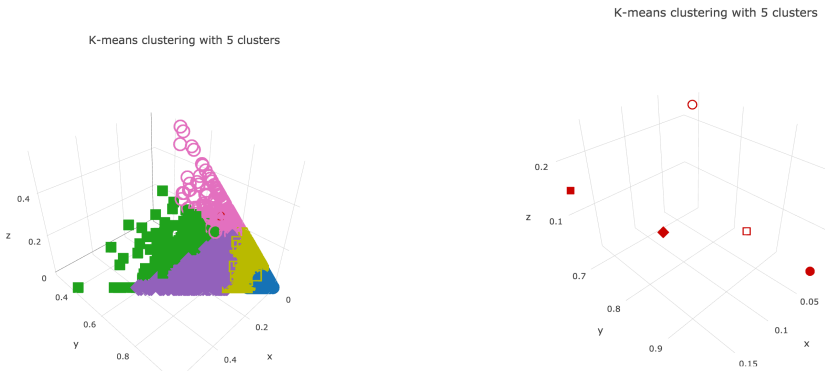
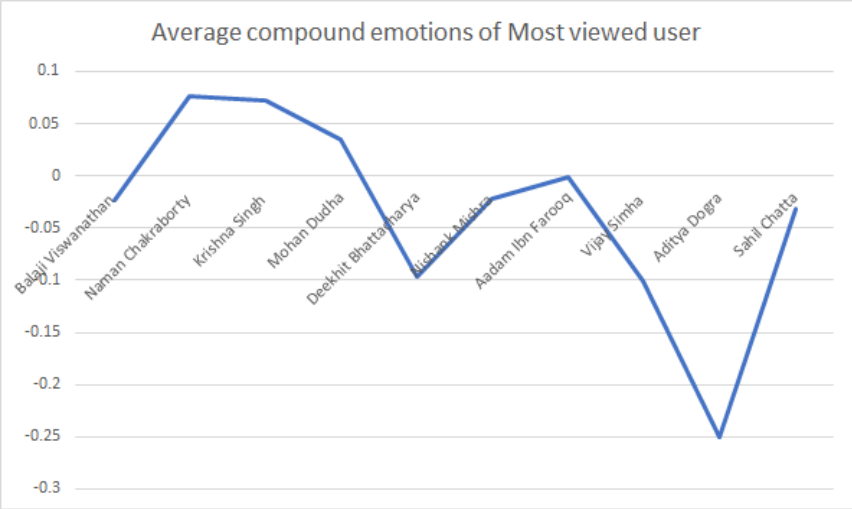


Fig 5- Clustering of users with cluster size 5

These are the results that we have obtained on our Quora topic- Kashmir Conflict from the various algorithms that we have used. Further more analysis has been performed and the results are given in the form of graphs which can be found in the folder(CIS700_SMM_Project II\Results And Plots).

Most Viewed Writers

Quora decides on the top 10 most viewed writers depending on the answers that they have given for the topic. This is given after analyzing the data for 30 days. So, every 30 days the list of writers will be updated. The most viewed writers as on April 25th, 2017 were taken for our observations. The clusters in which these writers lie were analyzed. The expected result was that they would belong to one cluster, however the outcome of this was not according to our expectation. They belong to two different clusters and also even with variations of the cluster size, they remained to be divided into two clusters. The main assumption for our project is that emotions drive social networks and if emotions were not playing any role then the most viewed writers for this topic should belong to the neutral cluster. However, after analyzing the emotions of the most viewed writers, the observations made is that they belong to the negative cluster.



Future

Our project is based on a single topic in Quora due to time constraints. Our results could be more generalized and accurate if the same analysis is performed on the various other topics in Qoura. Also, our project uses only the K-means clustering algorithm and there are various other clustering algorithms like DBSCAN which would give us more precise results on our current dataset. The results obtained for these variety of algorithms would help us in drawing better

conclusions related to emotions of people. Based on the results that our project has obtained the following is the fields where it can be beneficial-

- Predicting of the emotions of group of people depending on which clusters they would belong to and this would help analyze the emotional region in which there answers would belong.
- Understanding the perception of the topics and questions being answered related to those topics, based on emotional quotient of people.
- Deriving the polarity (Negative / Positive) of the entity based the queries/answers of the people under study.
- Analyzing how metrics that are being studied balance each other for the entity under discussion.
- The method can be extended to analyze reviews on products, stocks etc helping the business holders to make important business decisions.

Conclusion

After performing various experiments and analyzing our dataset from various aspects, the following conclusions are drawn-

- The centroid values of the clusters when the cluster size is 2 portrays that the emotions are extreme. As, the size of the clusters is increased these centroid values seem to come closer for the negative and positive cluster. The reason behind this is that there is a lot of variation of emotions held in the topic.
- When we compared the density of the emotions in each cluster, it was observed that each cluster was equally dense when compared to other cluster. There was no cluster of emotion which was extremely dense.
- Overlapping of clusters emotions is observed by increasing the size of the clusters. Initially for cluster size 2, there was high overlap of clusters. However, as the cluster size increased, the overlap of the clusters also decreased. This also shows the existence of variation of emotions.
- Most viewed writers didn't belong to a single cluster, instead it could be seen that they belonged to two different clusters. However, the majority of them belonged to negative cluster and it can be said that emotion they hold in their writings is more closer to the negative side.

Therefore, based on all the conclusions drawn, it can be asserted that emotions do play an essential role in social media and our belief that emotions drive social networks is true.

References:

1.VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text, Clayton.J. Hutto, Eric Gilbert ICWSM,2014

2. EmpaTweet: Annotating and Detecting Emotions on Twitter , Kirk Roberts, Michael A. Roach, Joseph Johnson, Josh Guthrie, Sanda M. Harabagiu Human Language Technology,Research Institute, University of Texas at Dallas, Richardson TX 75080

3. Sentiment Analysis of Twitter Data, Apoorv Agarwal Boyi Xie Ilia Vovsha Owen Rambow Rebecca Passonneau Department of Computer Science Columbia University New York, NY 10027 USA

4. Natural Language Processing with Python, Authors- Edward Loper and Steven Bird

5. NLTK: The Natural Language Toolkit Steven Bird Department of Computer Science and Software Engineering University of Melbourne, Victoria 3010, AUSTRALIA Linguistic Data Consortium, University of Pennsylvania, Philadelphia PA 19104-2653, USA

6. https://www.utdallas.edu/~kzhang/Publications/THMS13.pdf