GALGOTIAS UNIVERSITY

(Established under Galgotias University Uttar Pradesh Act No. 14 of 2011)

# SENTIMENTAL ANALYSIS OF SOCIAL MEDIA

A Report for the Evaluation 3 of Project 2

Submitted by

**RAJAT BISHT**

**(1613105082 / 16SCSE105074)**

in partial fulfilment for the award of the degree of

**Bachelor of Technology**

**IN**

**Computer Science and Engineering with Specialization of**

**Cloud Computing and Virtualization**

**SCHOOL OF COMPUTING SCIENCE AND ENGINEERING**

**Under the Supervision of**

**Mr C. VAIRAVEL, Assistant Professor**

**APRIL / MAY- 2020**

GALGOTIAS UNIVERSITY

(Established under Galgotias University Uttar Pradesh Act No. 14 of 2011)

# SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

## BONAFIDE CERTIFICATE

Certified that this project report **"SENTIMENTAL ANALYSIS OF SOCIAL MEDIA"** is bonafide work of **"RAJAT BISHT"** who carried out the project under my supervision.

**SIGNATURE OF HEAD**

Dr. Munish Shabarwal,

PhD (Management), Phd (CS)

**PROFESSOR & DEAN,**

**School of Computing Science & Engineering**

**SIGNATURE OF SUPERVISOR**

Mr C.Vairavel

**ASSISTANT PROFESSOR & SUPERVISIOR, SCHOOL OF Computer Science & Engineering**

**TABLE OF CONTENTS**

**ABSTRACT**

On social networking has taken over huge number of users. With over 3.2 billion users in 2019 alone. With these 3.2 billion users, a lot of data is generated. From usual text to images to emoticons to videos, a lot is generated on this basis. For this there are no such analysis going on what type of users are categorized, uploads an image or video, when they post, what do they post or for whom they post. For this, there are many analysis techniques that can be used here. In this paper we are going to use sentimental analysis. Sentiment analysis (also known as opinion mining or emotion AI) refers to the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information.

Both Twitter and Instagram are the top most used social media networking application. Both produce content at a very large.

Twitter offers a fast and effective way to write the content. In first segment, we shall report the design on sentimental analysis of huge number of tweets. The result will signify the tweet which are positive, neutral or negative in a pie chart.

Instagram is another social media platform where user upload photos, updating their lifestyle. Using different kinds of hashtag and different kind of filters on photo. Since it has been launched, it has seen a large jump in its popularity. It was launch on October 2010. In spite taking the place of then largest and most popular social media application, still the research based on this platform is very low. Research community is yet to discover the potential of this platform. So, for this the second social media analysis is for the Instagram, we did the sentimental analysis of the platform with analysis as qualitative and quantitative. We have used some computer technology works as a vision to examine photo content. We have

identified and categorizing what kind of users are active and use the platform. Our study gave us the insight –

1. Photo Categories.
2. User types.
3. Relation of users and their followers.

This research is based on user generated content.

## INTRODUCION

Internet has grown the most out of any sector. People are using it to communicating, sharing thoughts then they are the best ways to do it. Internet has become ways to connect, for online meetings, interviews, academics and much more. Majority of the public likes to listen other opinions before taking any final decision.

Microblogging is blogging about real time blogging, what we do on social media, generating content on the related topics which are discussed on these platforms, these discussing can conclude from current issues, complains, politics, reviews about a product for others to use and write back as well. So, microblogging websites generates loads of day as per day basis. For many companies' social media has become a place to advertise their products and get review about them. These micro blogs act as a user interface which gives them reviews about their products and to see what kind of service users want more, what will make the company's product better. So, the main challenge is to make a summarize form for user's reaction and content.

In this paper, we take a look at two social media giant micro blogs Twitter and Instagram.

a. For Twitter, we have built model which can classify "tweets" into three types positive content, neutral content (no side taken) or negative content sentiments. So, we have built three classes respectively. Advantage of Twitter is that that data collected are in a streaming way, which represent actual meaning of the content.

b. For Instagram, as a life updating platform, Instagram has come a long way to capture the life of one to represent as a great one. It provides easy way to capture and use different filters to make it look better. Instagram has reached its 1 billion active users mark. With unimaginable photos been uploaded, it has become a resourceful place too.

a. For Twitter,

Here are some of the methods used for sentimental analysis:

a.1 Opinion Mining

By using text mining and opinion present in the text or photo, is read through deep analysis and computational study is called as Opinion Mining. This is used in many areas marketing politics etc. by the help of this mining, organization gets the knack what the product is lacking and what the consumers want more, services or the quality, this gives direct relation with the consumers interaction with company.

a.2 Lexicon Based Approach

Each word represents a specific sentiment, by this method we can identify what the word actually means according to the text. It might be possible that the word meaning is different but used as different in content.

Steps used in Lexicon methods are:

1. Pre-processing of tweets by removing punctuation marks.
2. Initialize the polarity score(s) to 0.
3. If token is present or not.

   If present, s is positive

   If absent, s is negative

4. Compare polarity score from post

   If s>threshold, tweets are +ve

   If s<threshold, tweets is -ve.

a.3 ML Approach

1. Supervised machine learning which uses structured data or human annotations from which machine learns in order to make future inferences.

2. Semi supervised machine learning approaches helps to interpret unstructured data without guidelines.

3. Deep learning algorithms are advanced level algorithms which uses algorithms like SVM and ANN to gain higher accuracy in results.

a.4 Application Programming Interface

Python Twitter API performs better than any other is terms of quality and execution time. Twitter API helps us in collecting tweets from twitter database in real time and automatically preprocess the data internally and calculates the frequency of the words than use those values for calculating the sentiments.

b. For Instagram,

To start the analysis, we have taken data from Instagram API, used for categorization of photos and user's characteristic. The ranges from profile information, captions, tags used and other social network that are connected. The first dataset is provided and developed a scheme for categorizing the content.

b.1 Data Methodology

To obtain more range of user's data, asked some people whose photos are appearing on the public timeline. The timeline displays which photos or videos are popular at that particular moment. By this first experiment, we found these IDs were of popular people.

Then by crawling and merging into a single list of users, who are followers and friends of these popular people. Next choosing IDs which are regularly posting on Instagram.

By this experiment we found, different types of users of which some are 1) spammer or brands and 2) at least 30 friends and same followers with 60 posts. Also the count was of 14.6% who satisfied criteria, out of which randomly selected some and used there 20 recent photos and their social connect.

Since the manual coding was required, we only chose a smaller number of 50 users.

## PROPOSED MODEL

a. For Twitter,

The proposed technique is designed for query-based processing with open API on twitter. This API collects tweets in real time on the basis of user's keyword. Below diagram represents system architecture and the work flow of the model.

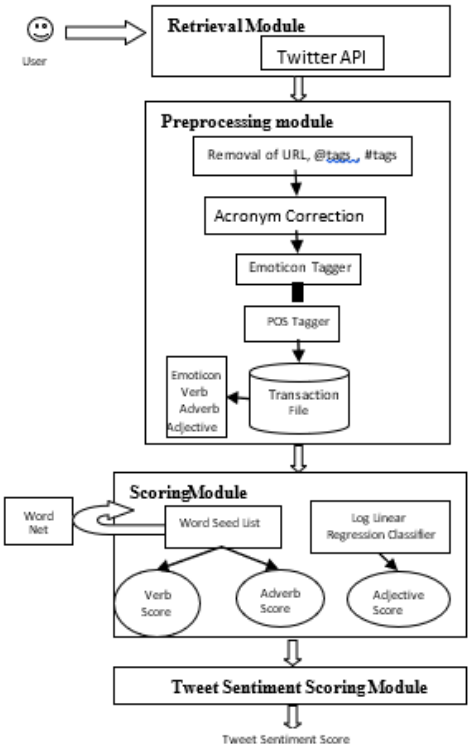
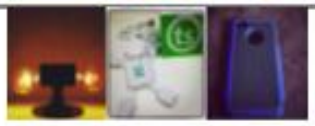



Fig 1: System overview

b. For Instagram

For to know the types of photo posted, using an approach and code a sample 200 photos from 1000 we obtained.

For photos, as photo has richer features than text, so it might be challenging to make the content categories to make them meaningful. So, for this, taking help from computer vision techniques to get an overview of efficient manner. The first technique we used of Lowe (1990) called Scale Invariant Feature Transform algorithm (SIFT), to detect and extract local discriminative features from photos. By using this algorithm, we are using feature vectors, which gave us the 128 dimensions of photos.

| Category | Exemplary Photos |
|---|---|
| Friends (users posing with others friends; At least two human faces are in the photo) |  |
| Food (food, recipes, cakes, drinks, etc.) |  |
| Gadget (electronic goods, tools, motorbikes, cars, etc.) |  |
| Captioned Photo (pictures with embed text, memes, and so on) |  |
| Pet (animals like cats and dogs which are the main objects in the picture) |  |
| Activity (both outdoor & indoor activities, places where activities happen, e.g., concert, landmarks) |  |
| Selfie (self-portraits; only one human face is present in the photo) |  |
| Fashion (shoes, costumes, makeup, personal belongings, etc.) |  |

For Instagram we are going to make categories with the help of above figure which will give us a better view of the content shared on the platform.

## IMPLEMENT

a. For Twitter,

with the help of the above table in proposed model section, I will begin the implementation on twitter content.

Extracted tweets are than used for the purpose of preprocessing. Emoticons are categorized in many types from neutral, negative, extremely negative, positive and extremely positive.

| Emoticons | Polarity |
|---|---|
| :-) :) :o) :] :3 :c) | Positive |
| :D C: | Extremely Positive |
| :-( :( :c :[ | Negative |
| D8 D; D= DXv.v | Extremely negative |
| :\| | Neutral |

Table 1: Emoticons dictionary

Sometimes people also use acronyms in their sentences and that becomes a challenge to overcome this problem we try to change the acronyms used to their real meanings and use them for getting sentiments. For example: “bff” is referred as best friends forever, “btw” is referred as by the way etc. Here is a table which shows some of the acronyms and their expansion.

| Acronym | English expansion |
|---|---|
| Gr8 | Great |
| Tbc | To be continued |
| Not, no never, n’t | Not |
| Lovv, luv, love | Love |

Table 2: acronym and their English expansion

After this set, we give the semantic score from ranging -1 to 1 to the verbs and adverbs which defines the strength. We have chosen the most frequently used verbs in below table.

| *Verb* | *Strength* | *Adverb* | *Strength* |
|---|---|---|---|
| Love | 1 | complete | +1 |
| adore like | 0.9 | most | 0.9 |
| enjoy | 0.8 | totally | 0.8 |
| smile | 0.7 | extremely | 0.7 |
| impress | 0.6 | too very | 0.6 |
| attract | 0.5 | pretty | 0.4 |
| excite | 0.4 | more | 0.3 |
| relax | 0.3 | much | 0.2 |
| reject | 0.2 | any | 0.1 |
| disgust | -0.2 | quite | -0.2 |
| suffer | -0.3 | little | -0.3 |
| dislike | -0.4 | less | -0.4 |
| detest | -0.7 | not | -0.6 |
| suck hate | -0.8 | never | -0.8 |
| | -0.9 | hardly | -0.9 |
| | -1 | | -1 |

Table 3: verbs and adverbs strength

Strength calculation explanation will be given in this paper, where detailed study of the strength calculation is done. After this tweet scoring sentiment is done. Of opinion indicators like verb, adverb and adjective clusters, capitalization, using the average strength of all opinion will be calculated of the tweet. Strength calculation is shown below:

$$S(T) = \frac{(1 + (P_c + \log(Ns) + \log(N_x))/3)}{|OI(R)|} * \sum_{i=1}^{|OI(R)|} S(AG_i) + S(VG_i) + N_{ei} * S(E_i)$$

Where,

|OI(R)| denotes the scale of the set of opinion and emoticons extracted from the tweets,

Pc denotes fraction of tweet in caps,

Ns denotes the count of continual letters,

Nx denotes the count of exclamation marks,

S(AGi) denotes score of the ith adjective cluster,

S(VGi) denotes the score of the ith verb cluster,

S(Ei) denotes the score of the ith facial expression

Nei denotes the count of the ith facial expression.

b. For Instagram,

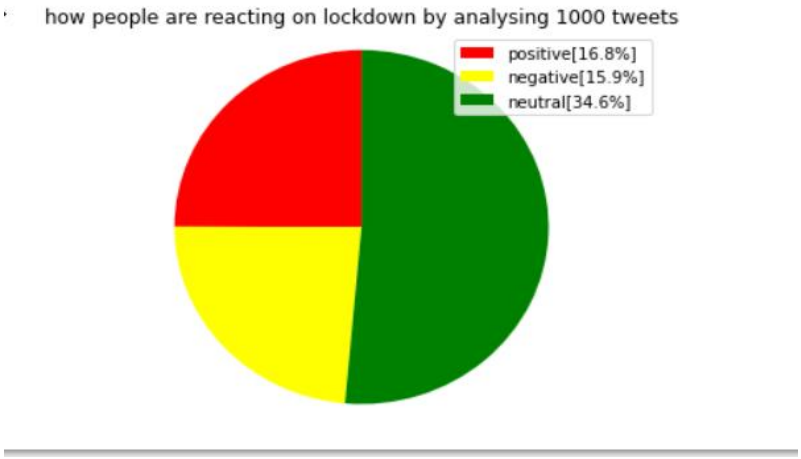
again, carrying on from proposed model section. Following the standard approach called the image vector quantization approach, we obtained 3 codebook vectors of each photo. And finally using ML approach of k means clustering, to obtain 15 clusters of photos and by using Euclidean distance between the codebook vectors are recalculated between two photos to increase the accuracy. These will give the reading regarding which photos are going to be categorize in accordance with their vectors, where each photo will be in one category.

To further test the quality and accuracy of this automated categorizing, we also took help from other two coders who can manual do this, to examine photos in each one of the 15 categories. They analyzed the affinity of the themes within the category and across categories, and manually adjusted categories if necessary (i.e., move photos to a more appropriate category or merge two categories if their themes are overlapped). Finally, through the coder's perspective, they decided with 8 different categories with scheme of photos (see Table 1) where both coders agreed on, i.e., the Fleiss' kappa is $\kappa = 1$. The goal for our manual coding was to provide a different perspective and not to hypothesize on the motivation of the user who is posting photos. Based to 8- category scheme, the two coders manually did the differentiation of the photos according to the categories. (e.g., if a photo has a girl with her dog, and the description of this photo is "look at my cute dog", then this photo is categorized into "Pet" category). The initial Fleiss' kappa is $\kappa = 0.75$. To resolve discrepancies between coders, we asked a third-party judge to view the unresolved photos and assign them to the most appropriate categories.

## RESULT

a. For Twitter

We were able to successfully implement the above discussed idea and have a working model of Twitter Sentimental Analysis. In the developed model it extracts tweets in real time from the twitter and successfully classifies the sentiments into positive, negative and neutral categories. In the previous models there were issues due to the use of emoticons, acronyms and informal languages so we have solved both the issues in this model and due to which more accurate results are produced. We can improve the accuracy by using larger data sets and by targeting tweets which are related to the politics, products and opinions. Sentiments having value greater than zero are considered as positive sentiment, sentiments having values less than zero are considered as negative sentiments and those which have value equal to zero are considered as null sentiment. The result of this analysis is displayed in the pie chart which is as follows:




As shown in Fig 2, Pie chart is representing how many peoples are taking lockdown occurred due to coronavirus in positive way, negative way. Results are presented in different colours with percentage values.

b. For Instagram

So, we performed an analysis of photos and users on Instagram platform, which has been the fastest growing social media platform. In this sentimental analysis how data on the image was handled and analysed to answer the three research questions on Instagram. Our analysis

concluded that there are 8 different types of photo categories on Instagram, based on the user generated content, derived from 5 different types of users. Also, another conclusion is that there is no direct relation between number of followers and type of users characterized in terms of content that is shared by the user.