International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-4, November 2019

Algorithm for Emotion Prediction using Twitter Dataset

K. Arulmozhi, R. Ponnusamy

Abstract: ***In today's internet world almost each and everyone uses Smartphone and they are all also active in various social media. In general social media contains a huge amount of data that can be extracted and utilized to find various data insights including polarity emotion etc... This research paper mainly investigates in emotion predection using a machine learning approach . Here a novel algorithm was introduced to predict the emotion of tweets . The algorithm mainly deals with emotion Prediction by utilizing various parameters like unigram , bigram , edge weight matrix , frequency matrix and so on . Finally , the result was predicted with the emotions of the tweets . While testing with various search terms this algorithm performs well in Predicting the emotion like anger, happiness and so on .***

Keywords: Machine Learning, Emotion Prediction, Tweets Extraction.

I. INTRODUCTION

In this modern word there are many social network tools such us twitter, facebooketc..ehich allows to express our thoughts or opinion in short text or images. The tweets in the twitter are usually in text format which contains various emotion related to happiness,angry, sadness etc..At the same time the emotion may be about a single persot or thing or it can be a group emotion like protest or gatherings. Twitter is the largest database which contains various tweets related to different type of emotions or sentiments.

For example the tweet " Great Christmas spent with my amazing family " expresses a happy mood and the tweet " Feelings Hurt Tonight !" expresses sadness . In general many professionals belive that human has aa very small set of basic emotion which are discrete.Various researchers had been working in this issue to figure out the basic emotion of the people by using heris tweets or comments.these emotions are common amoung all people throughout the globe but it may differ in various forms. The basic emotion cab be classified into various types like tension, depression , anger , vigour , fatigue , confusion from Twitter .

For instance , it is unclear if "surprise " should be considered an emotion since it can assume negative , neutral or positive valence . Tweets or status updates of twitter users. : By using the information which contains in the tweets or microblogs we need a system to answer the basic questions automatically like :wheather the author is feeling happy , sad from his tweets itself

Given some text , emotion recognition algorithms detect which emotions the writer wanted to express when composing it . To treat this problem as a special case of text classification , we need to define a set of basic emotions . Although emotions have long been studied by psychologists, there is no single , standard set of basic emotions .

II. BACKGROUND

Sentiment analysis or Opinion mining is the term which is used to derive the group of text or microblogs or tweets to find out the motion or polarity of the text. Sentiment analysis mainly deals with the consumer cooments or reviews or shortext summary to find the expression or mood of the consumner in which he was while typing that.A basic thought of the sentiment analysis was to classify the tweets ot text into 3 major groups first they are "positive","negative" and "neutral". They are also reffered as polarity . Based on the polarity these can be further calssified into various emotions like "sad","angry","happy", etc..

Sentiment analysis will provide a clear approach of the given text. It can be used in Customer relationship managemet mainly.

III. RELATED WORK

CMiner: Opinion Extraction and Summarization for Chinese Microblogs

Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao 2016.

Microblog sentifent classification bas been mainly used in polarity classification to find it as positive or negative.But at the same time time classification of the polarity from the sentances is often a difficult task to identify the emotion of a particular topic related to any situation.In this the author had designed apowerfull method for opinion ming of microblags and namedit as CMiner. The Cminer mainly focus on Chinese microblogs to classify the sentiments. –the CMiner provides different opinion for different groups based on Chinese microblogs and also summarize the opinion regarding the topic.The tradition of hashtags is quite comman in twitter to analyse the emotion in group for example #love, #angry etc.. These hashtags can be used to classify the emotion of the tweets or microblogs.

Manuscript published on November 30, 2019.
* Correspondence Author
K. Arulmozhi*, Research Scholar in Computer Science, Mother Teresa Women's University, Kodaikanal, Tamilnadu, India.
Dr. R. Ponnusamy, Professor, Department of Computer Science and Engineering, CVR College of Engineering, Ibrahimpatan, Hyderabad, Telangana, India.

TASC : Topic Adaptive Sentiment Classification on Dynamic Tweets

Shenghua Liu , Xueqi Cheng , Fuxin Li , and Fangtao Li 2015

The topics which is discussed in twitter is mostly unpredictable and diverse.But Sentiment classification is basically domain oriented or domain specific. For example if a classifier was trained in a particular topic and was used to test another data which belongs to different topic it performs poor. The primary reason behind this is the words which we used in a particular topic will differ from anther topic to express our emotions. If there is a comment like "Read a Book" in a microblog it can be positive in a amazon book review dataset since it defines it ,but at the same time it may be wrong if it was in some other domain like movies .the same comment will be nagatine in movies and positive in books. In real world scenario each user may have different opinion on different topics. Since we need a frame work to identify the different domain and cpable of classifying the texts coorectly.

IV. DATASET DESCRIPTION

In this paper we are using live dataset instead of using predownloaded dataset . i . e . we are extracting the data or tweets from the twitter in run time . and the tweets can also be filtered based on language and geo code also . so that you can even run our algorithm to find the sentiment of a particular search term in various location . since the live dataset makes our work unique . becasue all the paper which we referred in literature survey is using only the pre downloasded dataset not the live one . but this is not he only difference in our work there are various steps involved in the emotion Prediction which are explained below.

V. PSUEDOCODE

I ←Input tweets;

pI ← Remove stopwords i → j & symbols i →j;

Token t ← split (pI, " ");

Unigram U←Distinct(t);

Bigram B←pI;

TT=0,TU=0,AOF=0
TT=Count(t);
TU=Count(U);
AOF=TU/TT

F←Frequence matrix Ui→j

F← F>AOF

E ←Edge weight of t

D←Equidian distance[E]

C←From E[Centrality]

CD←Cache matrix

R←Predictemotion(CD,C,D,E,B)

VI. ALGORITHM FOR EMOTION PREDICTION

A novel algorithm for emotion Prediction is designed and developed by using R tool .It's core is designed to extract the tweets , preprocess it and predict the emotion of the tweets . It includes various steps which is explained below .

A. Extract Tweets

In the stage, the desired amount of tweets (I) will be extracted based on the search term by using the Twitter API credentials .

B. Preprocessing

By using the extracted tweets which we got from the previous stage , stop words like " a ", " is ", " was ", " then ", " they", " it ", " so", " we " etc … will be removed and also symbols like "!","@","$" etc … will also be removed and all the tweet sentence will be converted into lower case on the whole (pI).

C. Tokenization

The preprocessed tweets from the previous stage is taken into account. It is then further divided into words by using regular expression methods to form tokens that is each and every splitted words is hereafter called as tokens (t).

D. Preliminary Unigram Matrix

The Duplicate tokens are identified from the token matrix and the remaining token matrix is considered as preliminary unigram matrix(U).

E. Bigram Matrix

The bigram matrix is also formed from the token matrix to give a clear idea about the tweets consecutive two words are kept in a matrix formation to form a bigram matrix (B).

F. Total Number Of Token

The total count of the token matrix is taken into account to calculate the total number of the token (TT).

G.Total Number of Unigram

Like the previous step total number of unigram matrix is considered as total unigram here (TU).

H.Preliminary Frequency Matrix

Preliminary frequency matrix is nothing but the number of time the particular term or a token appears in the token matrix and the result it is considered as primary frequency matrix (F).

I. Average of Frequency

The average of frequency is calculated by dividing the total number of unigrams by a total number of token and stores in a variable AOF .

J. Final Frequency Matrix

The final frequency matrix is obtained by using the AOF value and F . By using the average frequency value the frequency matrix F is filtered . The term matrix value which is less than the average value is eliminated to form the final frequency matrix .

K. Final Unigram Matrix

Based on the final frequency matrix the unigram matrix (U) will be redefined and re ordered to form the final unigram matrix (U).

L. Edge Weight Matrix

Edge Weight matrix is calculated by using the tweets . Each and every word in the final unitary matrix is considered into account to calculate the Edge weight matrix . For example , a token from document A is Taken as input and will be associated with all documents and find the link between various documents . and the final result will be called an edge weight matrix (E).

M. Euclidean Distance

The Euclidean distance (D) is calculated for the edge weight matrix . this step will help us fiding the similar words in the extracted tweets for example " good and "goodddddd " both sounds the same but it will be considered as different tokens . But from now based on this it will get a closer value to its predecessor . By using that we can find closer words also to predict the emotion.

N. Centrality

The Central node (C) is identified from the edge weight matrix which will act as center point for all extracted tweets.

O.Cache Data

The cache data (CD) Matrix is used to exclude a particular word or a list of words from further proceedings. If a list of words wants to be excluded in the result then this cache data matrix will be used else it can be left empty depends upon the situation.

P. Sentiment Prediction

By using the Cachedata matrix ,Centrality , Distance Matrix, Edge Weight Matrix and Bigram as parameters and also a sntiment dataset is taken as input and the tweets are classified and ranked based the input parametes . The final sentiment of the tweets is predicted and given as output for the tweets.

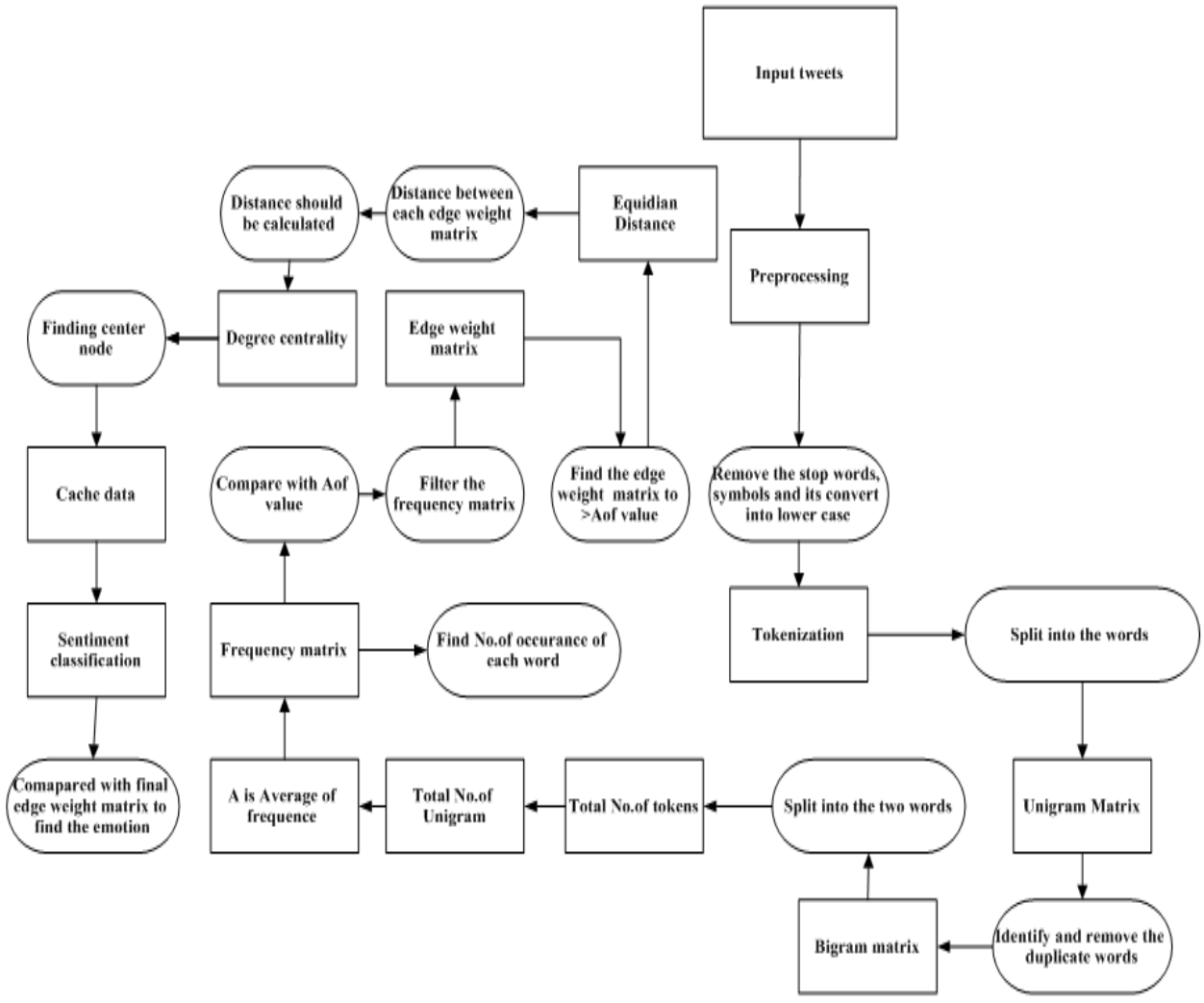
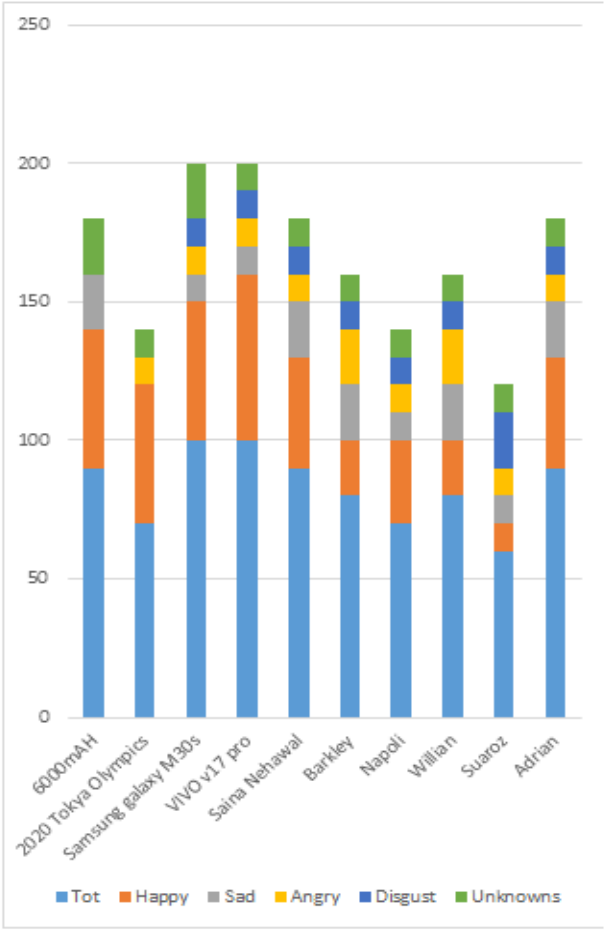


Fig: Data Flow of EmotionPrediction Algorithm

VII. EXPERIMENTAL RESULTS

SEARCH TERMS	TOTAL	HAPPY	SAD	ANGRY	DISGUST	UNKNOWN
6000mAH	90	50	20	0	0	20
2020 Tokya Olympics	70	50	0	10	0	10
Samsung galaxy M30s	100	50	10	10	10	20
VIVO v17 pro	100	60	10	10	10	10
SainaNehawal	90	40	20	10	10	10
Barkley	80	20	20	20	10	10
Napoli	70	30	10	10	10	10
Willian	80	20	20	20	10	10
Suaroz	60	10	10	10	20	10
Adrian	90	40	20	10	10	10



In recent years ,the trending twitter hash tags are mentioned in the chart by using this chart we analyze the persons emotions classification with the help of hashtag details and the emotions are classified as happy, sad , angry , disgust and unknowns .

VIII. LIMITATION

The algorithm is designed only for the text classification . but now a days many tweets contains smilies which is also a best way to express the emotion . this algorithm will not consider the smilies as input all those symbols will be neglected in the preprocessing stage itself . this is one of the limitation in this algorithm .

IX. CONCLUSION

Text Document is one of the thoughest job now a days. Since it involves many human behavious its hard to predict the emotion in which the author is while posting a tweet . here we have to analyse the tweets only with the text not the face reading etc …. so just by keeping the text as input we have identified the emotion of the author by using our algorithm but at the same time there also may be some limitation . we are also working in it and the future updates will be reveled soon . except that our algorithm works fine in various search terms and in various locations also.

REFERENCES

1. TASC:Topic-Adaptive Sentiment Classification on Dynamic Tweets Shenghua Liu, Xueqi Cheng, Fuxin Li, and Fangtao Li-2015.
2. Studying the Scope of Negation for Spanish Sentiment Analysis on Twitter SaludMarıa Jimenez-Zafra, M. Teresa Martın-Valdivia, Eugenio Martınez-Camara and L. Alfonso Urena-Lopez-2019.
3. CMiner: Opinion Extraction and Summarization for Chinese Microblogs Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao-2016.
4. Sentiment analysis in a cross-media analysis framework Yonas Woldemariam-2016.
5. Entity-Level Sentiment Analysis of Issue Comments Jin Ding ; Hailong Sun ; Xu Wang ; Xudong Liu-2018
6. Analyzing Sentiments Expressed on Twitter by UK Energy Company Consumers Victoria Ikoro ; Maria Sharmina ; Khaleel Malik ; Riza Batista-Navarro-2018.
7. A Sentiment Analysis Method of Short Texts in Microblog Jie Li ; Lirong Qiu-2017.
8. Twitter based model for emotional state classification Ravinder Ahuja ; Rohan Gupta ; Saurabh Sharma ; AyushGovil ; Karthik Venkataraman-2017.
9. Fuzzy classification-based emotional context recognition from online social networks messages Imen Ben Sassi ; Sadok Ben Yahia ; Sehl Mellouli-2017.
10. State-of-the-art review on Twitter Sentiment Analysis Norah Fahad Alshammari ; Amal Abdullah AlMansour-2019.
11. Opinion mining and sentiment analysis on a Twitter data stream Balakrishnan Gokulakrishnan ; PavalanathanPriyanthan ; ThiruchittampalamRagavan ; Nadarajah Prasath ; AShehan Perera-2011.