Analysis and Derivation of Optimum Domain Specific Semantic Model for Detecting Depression Text From Twitter Stream

[1]C. Ashwini, [2]Aman Ehtesam, [3] Bhaskar Bhakat, [4] Piyush Kumar

[1]Supervisor, Computer Science & Engineering, SRM Institute of Science & Technology, Ramapuram Campus, India

[2 3 4]B.Tech Student, Computer Science & Engineering, SRM Institute of Science & Technology, Ramapuram Campus, India

Abstract- Depression is a serious challenge in personal and public health. Every year one in 15 adults are affected by depression. According to WHO, more than 300 million people globally suffer from depression. An individual with depression is 20 times more likely to die from suicide than someone without depression. Depression specially, prevalent in young individuals from 15 to 29 years old and this age group also has suicide as second leading cause of death. We consider English content gathered from web-based media such as Facebook, Twitter, Reddit, etc., which can give data having utility in an assortment of ways, particularly assessment mining. Web media such as Twitter, Reddit and Facebook are brimming with feelings, sentiments, and assessments of individuals everywhere in the world. Nonetheless, examining and characterizing text based on feelings is a major test and can be considered a high-level Sentiment Analysis. This paper aims to identify 6 different Emotions i.e. Happy, Sad, Fear, Anger, Surprise and Disgust. Our approach is based on machine learning classification algorithms and Natural Language Processing which uses textual features like degree of words and negations, Parts of Speech, grammatical analysis and emoticons. However, on a huge dataset, emotional words are mapped with emotional intensities. On testing, we show our model gives huge exactness in arranging tweets taken from Twitter. At last, through our project, we can successfully identify particular text behaviour and pass judgment on the depression tendency.

Keywords:Depression, Sentimental Analysis, Social Media, Emotion Classification, SVM, Natural Language Processing

I. INTRODUCTION

Sentiments are depicted as genuine slants that are focused on someone or something considering inside or outside events having explicit significance for the individual. Also, the web, today, has got a vital medium through which individuals express their feelings, sentiments, and assessments. Each occasion, news, or movement all throughout the planet, is shared, examined, posted, and remarked via web-based media, by a great many individuals.

Depression causes negative reasoning, less focus in work, decreased efficiency. It additionally influences the propagation arrangement of the person. Depression may cause mental confusion moreover. Early recognition of mental health is easily treated at an early stage. Mental wellbeing handles pressure and is an action for making choices in each stride throughout everyday life. Mental Health plays an important factor in each stage of life whether it be adolescence or a grown-up. Mental wellbeing of the pressure at home and work environment. It will build the profitability of individuals.

In this paper, we propose a strategy to arrange and measure tweets as per six standard feelings that are well-known to everybody. Here, we base our examination on tweets posted on Twitter, yet it tends to be handily stretched out to any sort of text whether it is one-lined features, messages, and posts via online media or bigger lumps of compositions, due to programmed improvement of our preparation set.

While the exactness is higher than the review for the best tweet models and the review is higher than the accuracy for the best content models, the best models for the two modalities have equivalent exactness. Regardless of better execution, the best content models require less highlights than the best tweet models based on highlights from a similar transient amount of information (except for 56 days of information).

We have developed a structure that could score and name any piece of text, especially tweets and posts by means of online media as appeared by six Emotion-Categories: Happiness, Sadness, Fear, Surprise, Anger and Disgust close by their force scores, utilizing its scholarly features, a grouping of NLP gadgets and standard Machine Learning classifiers. Another colossal responsibility is that we have successfully arranged a system that could normally (with no manual effort) collect a gainful planning set for our ML Classifiers, involving a gigantic enough course of action of stamped tweets from all Emotion-Categories. We have made a tremendous bunch of words in English, that contains words conveying a particular inclination close by the force of that inclination. We have achieved an accuracy of about 94%.

II.LITERATURE SURVEY

The Literature survey examines the different models completed by different creators in such a way that aides in the identification of Depression. The strategy uses various strategies identified with AI computations unmistakably and achieves high effectiveness in their specific fields. The total of the writing audits that we have gone through is examined in a word underneath:

Guozheng Rao [1]2020 proposed a model in which they were identifying Depressed Individuals in Online Forums.Their key aim was to perceive despairing using the best immense neural arranging from two of the most standard basic learning approaches in normal language dealing with(CNNs) and (RNNs).

Mark E. Larsen[2] in 2015 conducted a study named as Mapping Emotion on Twitter. His study rehearses and astoundingly mixes clear audit relations with semantic and social signs amassed from Google Play application information (87K applications, 2.9M examinations, and 2.4M intelligent people, assembled over a goliath piece of a year), to see sketchy applications. It accomplishes over 95% exactness referring to the best quality level datasets of malware, deceiving, and genuine applications.

Marcel Trotzek[3] in 2018 conducted a study mainly known for Detection of Depression Indications in Text Sequences that evaluate using AI computations.These are customary in regular day-to-day existence where choices are taken by interlinked various rules. In the proposed structure, AI masterminds a tendency from different viewpoints and evaluation words. For example, in a bistro audit, the master cherishes the food yet disdains the help. The computation depicts the audit by thought words or verbalizations about viewpoints.

BudhadityaSaha, Thin Nguyen, Dinh Phung, and Svetha Venkatesh[4] in 2016conducted a study that shows 75% of the conspicuous malware applications share searching for rank mutilation. Fairplay finds diverse lie applications that advantage right currently keep away from Google Bouncer's receptiveness improvement. Fairplay also helped the openness of more than 1,000 outlines, verifiable for 193 applications, that uncover another sort of "coercive" design crusade: clients are meddling with into framing positive assessments and present and study various applications.

Michael M. Tadesse[5] in 2019 made a comparative study. In this paper, they have done a near examination among five methodologies say TF-IDF, Naive bayes, LSTM, Logistic Regression, Linear help vector.Among every one of the 5 strategies we have tracked down that Long ShortTerm Memory(LSTM)- RNN has the most noteworthy exactness to recognize the burdensome tweets from twitter.

ML Tlachac[6] tried todemonstrate that reflectively collected content messages have an incredible potential when evaluating for wretchedness, more so than freely posted tweets. Utilizing only the earlier fourteen days of instant messages, our strategic relapse models foresee a twofold PHQ-9 score at the moderate gloom cutoff of 15 with a normal F1 score of 0.806 and AUC of 0.832.

Les Servi and Sara Beth Elson [7] consolidating a robotized text examination program with another numerical way to deal with distinguish shifts in Twitter clients' passionate articulations, the work introduced in this paper offers a chance to utilize the information got from web-based media to acquire equitably determined experiences into the elements of Twitter clients.

FirojFattulalShahare[8] in 2017 concluded an expansive viewpoint on organizing compulsion and propose an orchestrating bowing perceiving check structure for versatile Apps. In particular, we from the start proposal to absolutely find the planning craftiness by mining the surprising time frames, to be express driving get-togethers, of flexible Apps. We can use such driving social affairs for seeing the nearby inconsistency instead of a general attribute of App rankings. [9]They have research three sorts of requests, i.e., coordinating based certifications, rating based confirmations, and studybased authentications, by showing Apps' engineering, rating, and study practices through quantifiable hypotheses tests. Proposed an advancement gathering approach to work with everyone in assertion for guile attestation. At last, they assess the proposed structure with veritable App information accumulated from the iOS App Store for quite a while period. In starters, we embrace the savvy instinct of the proposed plan and show the versatility of the disclosure computationfurthermore as some retinene of planning cheating works out.

Norah Saleh Alghamdi [10] inspected Depression Symptoms in an Arabic Psychological Forum through Machine Learning model which uses Naïve Bayes algorithm. The examination shows anprecision of 89%; recall0.99; F1-score 0.946; AUC 90%.

III.PROPOSED ALGORITHM

Our model involves two uncommon, yet related, philosophies. The chief methodology utilizes Natural Language Processing, Emotion-Words Set and a couple of abstract features. It attempts to gathering and score text according to the sentiments present in it. The ensuing strategy utilizes standard classifiers like SMO and J48 to portray tweets. Finally, we join both these approaches manage to propose a Hybrid method to manage separate sentiments in the substance even more effectively. Note that, despite the fact that we are comprehensively utilizing the tweets model all through this paper, this estimation is ordinary and can recognize and assess sentiments in anypiece of text.

We portrayed honest Bayes classifiers as a plan of portrayal algorithms relying upon Support Vector Machine. This single machine learning algorithm, undoubtedly a get-altogether of algorithms in which each one of them distribute a normal in vogue, we liberate as an example each pair of features being amassed from each other.

At that point term recurrence and reverse record recurrence are picked on the highlights to giveaway the feeling highlights a superior score, from that point forward, some unique managed calculation for feeling classification are used by us. Support Vector Machine for feeling classification have been used in this paper, these are automated AI algorithms. We have attempted to explore ourselves by recognizing the feeling from a book archive. Beneath we are introducing our proposed system in figure1.The accompanying segments depict each interaction in subtleties.

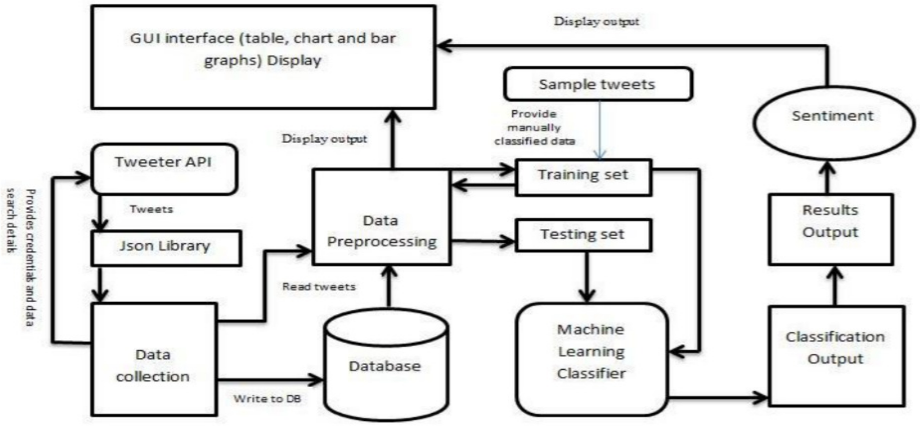


Figure 1: Architecture Diagram of Proposed System

4.1. Dataset Exploration –

In Sentiment Analysis, Exploratory statistics examination is a sample of setting apart via facts trying to find fascinating facts or fashions. Analysts' gift units for learning facts consolidate informational index corporation structures, quantifiable assessment packs, facts mining instruments, discernment devices, and file mills. Since the examination cooperation searches for the unanticipated in a information pushed way, it's miles critical that those units are flawlessly joined so retailers can deftly choose and make mechanical assemblies to apply at every duration of evaluation. Hardly any systems have joined all of those limits either basically or on the UI stage. Look's facts driven approach presents coordination amongst unique software UIs. It makes use of a plan that displays the arranging of visual things to facts in shared informational indexes for Sentiment Analysis.

We use Tweepy to gather tweets, which is a Python library for getting to the Twitter API. It takes as information various limits, similar to headings, range, etc, and after the removal of duplicates, joins, hashtags, and words in various lingos (other than English) from these tweets, stores the tweet-ids, text, and space of the most recent ones in the informational collection. This gave us a rundown of tweets from different areas the nation over. Eg: The Tweets Set, we made for Delhi has around 16,000 sections. Another way we utilized Tweepy is by taking care of it a twitter-username (of a client) as a contribution to store every one of the tweets of that client (till date), in our data set.

4.2. Preprocessing –

In Sentiment Analysis , we play out a progression of preprocessing ventures before highlighting extraction and classification is finished. Named substance (NE) acknowledgment, coreference goal, and reliance parsing are performed using the Stanford CoreNLP library. All references to casualties, including names or family names went before by greetings, specifies, etc, are supplanted with a uniform casualty marker after the reliance parsing step. We likewise eliminate client specifies, retweet marker, hashtags, and URLs from the tweet text after reliance parsing and before grammatical features (POS) labeling with Stanford CoreNLP for Sentiment Analysis. On the off chance that the occasion considered is a previous occasion, momentum news source or internet searcher results would not be acceptable pointers of a referenced element's extremity around there. For those, a rundown is developed dependent on verifiable news identified with the referenced substances.

Preprocessing on twitter text (tweets):

- Select English tweets.
- Transform the text into lowercase.
- Delete URL present in a twitter text.
- Erasure of notification, retweet sees, and pointless numbers.
- hastags can be taken out for better outcome.
- Adding a full design to short construction terms. For example, btw stands for coincidentally
- Changing the emoticons with their importance. For instance (":D")stands for laugh/happiness with respect to table I.
- Striping highlights ['"?!,.():;] from the tokenized words.
- words should not be removed from to create confusion.
- Stemming and lemmatizing the tokenized words.
- Making syntactic highlights tags for those tokenized words in the end.

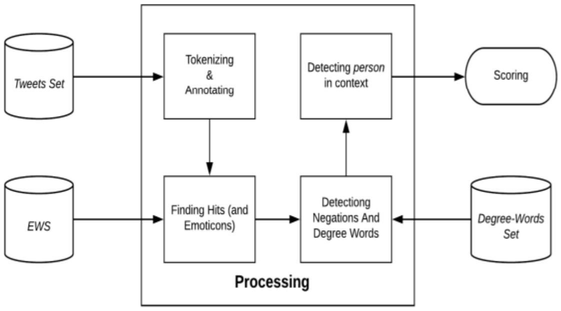
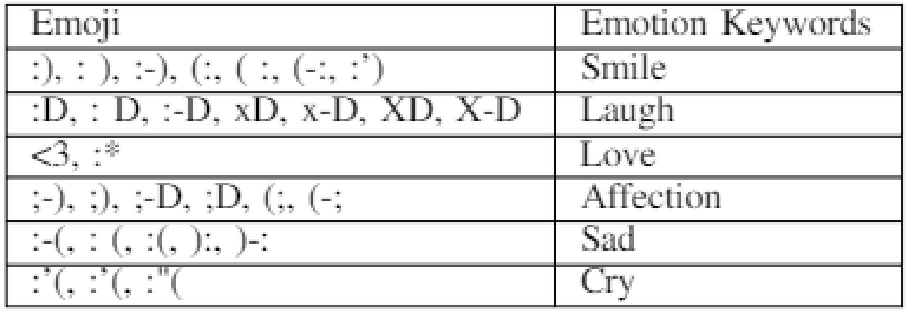

Tweets Set
Tokenizing & Annotating
Detecting person in context
Scoring
EWS
Finding Hits (and Emoticons)
Detectiong Negations And Degree Words
Degree-Words Set
Processing

Figure 2: Preprocessing of Dataset

In our model we have taken 6 types of Emotion Keywords.



Emoji	Emotion Keywords
:), :), :-), (:, (:, (-:, :')	Smile
:D, : D, :-D, xD, x-D, XD, X-D	Laugh
<3, :*	Love
;-), ;), ;-D, ;D, (;, (-;	Affection
:-(, : (, :(,):,)-:	Sad
:'(, :'(, :"(	Cry

Table 1: Emoji to Emotion Keywords

4.3. Feature Selection –

Feature in language handling enlargement insinuates the numeric vectors changed over from text based total facts. Since the published facts here are of various languages, constructing up a component choice shape with no making plans appeared to the most accommodating path of action. To do all things taken into consideration, Term Frequency-Inverse Document Frequency (TF-IDF) technique is picked for it being the most first-rate method. TF-IDF has treated the most normal articulations of being lousy regarding a set of rules execution by consigning less weight.

In Sentiment Analysis, we can portray feature extraction as a sample of keeping apart a lot of the latest capabilities from the capabilities set that is made in containing decision stage.

Term frequency (TF) is the manner as often as possible a phrase appears in a record, removed through the amount of words there are.

$$\mathrm{T\,F-IDF\,(T,d,D)=T\,F(T,d)\times I\,DF(T,D)}$$

Document frequency is the number of documents for your corpus a time seems in (and in opposite file frequency is the multiplicative inverse of this number). Together, those two sums may an extent of ways big a time is to a particular document.

In Sentiment Analysis, include dedication methods can be perceived into 3 classifications: channels, covers, and embedded/pass range machine. Covers strategies carry out well than channel methods because we progress characteristic selection connection for the classifier for use. In Sentiment Analysis, channel methodologies have low computational cost and quicker yet with inefficient constancy in type when stood out from protecting techniques and better practices for high dimensional instructive assortments. Creamer/it virtually made introduced strategies that useadvantages of the two channels and covers strategies. A cream approach makes use of each a loose look at and execution appraisal restriction of the aspect subset. We can also orchestrate channel methodologies into social occasions, explicitly incorporate weighting algorithms and subset seek algorithms as exhibited in Figure. Feature weighting algorithms choose burdens to features freely and rank them reliant upon their importance to the aim thought.

4.4. Classification Techniques –

We have utilized Support Vector Machine in our model. We have picked a lot of seed words, including ordinarily used Emotion-words and emoticons from the EWS, consistently appropriated over all of the Emotion-Categories. We by then inquiry Tweepy using these seed words to develop an enormous data base of around 16,000 tweets. The seed words are used to ensure that we get tweets that express at any rate one of the six sentiments. The retweets are removed out to keep from any emphasis of tweets. The even allocation of seed words ensures that we get even dispersal of tweets over all of the Emotion-Categories, so our classifier isn't uneven. Regardless, we have analyzed that there is an enormous degree of tweets on Twitter imparting bliss, so we have kept the quantity of seed words having the Emotion-Category HAPPINESS, on the higher side.

SVMs accomplish superior text arrangement, however they acknowledge high dimensional element spaces and scanty component vectors. Additionally, text identification uses SVMs which is strong to exceptions and doesn't need any boundary tuning. It finds a most extreme edge isolating hyperplane among2 classes of data. For different class SVM enhance the edge for one versus all classes of data., Linear SVM of the SK-learn pack are utilized. We introduced that straight pieces-based SVM plays out significantly better result.

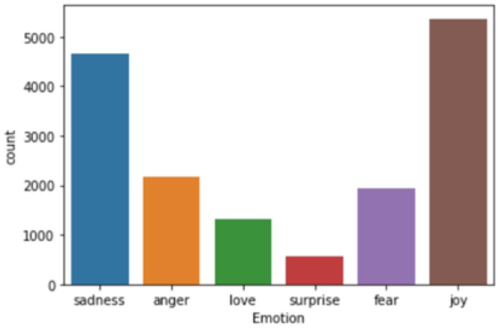
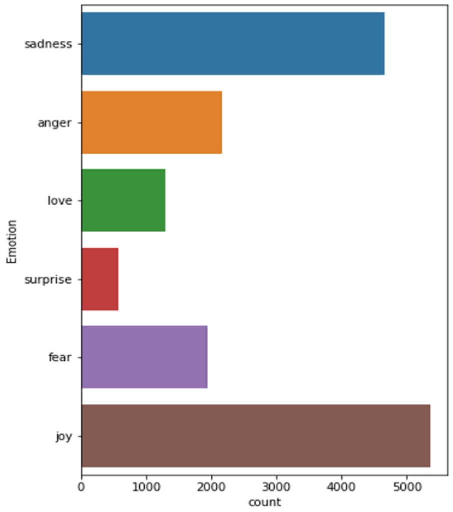


Figure 3: Emotion VS Count

IV.RESULT

In this portion, we will discuss our results of our proposed model in the wake of investigating various roads. We have chosen 16000 tweets from each and every class from the dataset, 16000 models altogether. Various information decreases an extraordinary arrangement bye liminating uproarious tweets during pre-dealing with and arranging WordNet was used to find them passionate words and EmoSenticNet words. We have acknowledged that one tweet has recently a solitary inclination class and just one inclination word in the tweet address that feeling.



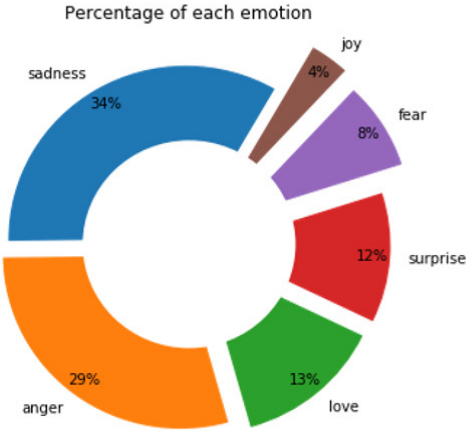


Figure 4: Count of Emotion **Figure 5:**Percentage of each Emotion

	Prec	recall	f1-sc	support
0	0.95	0.96	0.95	695
1	0.95	0.95	0.95	275
2	0.88	0.79	0.83	159
3	0.96	0.98	0.97	581
4	0.89	0.89	0.89	224
5	0.80	0.74	0.77	66

Table 2: Performance of Proposed Algorithm

```
              accuracy                              0.94       2000
                 macro avg0.91      0.89       0.90       2000
weighted avg0.94      0.94       0.94       2000
```

After pre-dealing with and removing unwanted tweets for feeling word, we scrap and plan to test the dataset to get ready and play it on ordinary ML Classifiers. Backing Vector Machine (SVM) was applied on our model. We got a precision about 94%.

V.CONCLUSION

Feeling identification is perhaps the hardest issue to settle. Recognizing feeling from text is testing work and the majority of the exploration works have some caring constraints in particular, language equivocalness, various feeling bearing content, text which doesn't contain any passionate words, and so forth However we have attempted a few ways to deal with recognize feeling from Twitter. In future, a system could be set up for thus invigorating the pack of-words which we made, in view of new tweets and data separated. Using our strategy, numerous captivating applications can be made, for instance, an extra to a long reach casual correspondence site showing the new perspective of all of your mates. Likewise, our investigation of Twitter can be reached out to the advancement of a constant framework, dissecting emotional episodes and feelings on Twitter.

References

[1] Guozheng Rao, Yue Zhang, Li Zhang, Qing Cong and Zhiyong Feng, “MGL-CNN: A Hierarchical Posts Representations Model for Identifying Depressed Individuals in Online Forums”, IEEE Access(Volume: 8),February 2020, DOI: 10.1109/ACCESS.2020.2973737.

[2] Mark E. Larsen, Tjeerd W. Boonstra; Philip J. Batterham; Bridianne O’Dea, Cecile Paris, and Helen Christensen, “We Feel: Mapping Emotion on Twitter,” in IEEE Journal of Biomedical and Health Informatics, 2015, DOI 10.1109/JBHI.2015.2403839.

[3] Marcel Trotzek, Sven Koitka, and Christoph M. Friedrich, “Utilizing Neural Networks and Linguistic Metadata for Early Detection of Depression Indications in Text Sequences,” in IEEE Transactions on Knowledge and Data Engineering (Vol: 32), 2018,DOI: 10.1109/TKDE.2018.2885515.

[4] BudhadityaSaha, Thin Nguyen, Dinh Phung, and Svetha Venkatesh, “A Framework for Classifying Online Mental Health-Related Communities With an Interest in Depression,” in IEEE Journal of Biomedical and Health Informatics, 2016, vol. 20, DOI: 10.1109/JBHI.2016.2543741.

[5] Michael M. Tadesse, Hongfei Lin, and Liang Yang, “Detection of Depression-Related Posts in Reddit Social Media Forum”,in IEEE Access, 2019, DOI: 10.1109/ACCESS.2019.2909180.

[6] ML Tlachac and Elke Rundensteiner, “Screening For Depression With Retrospectively Harvested Private Versus Public Text,” in IEEE Journal of Biomedical and Health Informatics, 2020, DOI: 10.1109/JBHI.2020.2983035.

[7] Les Servi and Sara Beth Elson, “A Mathematical Approach to Gauging Influence by Identifying Shifts in the Emotions of Social Media Users” in IEEE Transactions on Computational Social Systems DOI 10.1109/TCSS.2014.2384216.

[8] FirojFattulalShahare, “Feeling discovery through online media” in Proceedings of the 2017 conference on (ICICCS),DOI 10.1109/ICCONS.2017.8250692.

[9] AltugAkay, Andrei Dragomir,and Björn -Erik Erlandsson, “Assessing Antidepressants Using Intelligent Data Monitoring and Mining of Online Fora”, IEEE Journal of Biomedical and Health Informatics (Volume: 20), 2016, DOI: 10.1109/JBHI.2016.2539972.

[10] Norah Saleh Alghamdi, Hanan A. Hosni Mahmoud, Ajith Abraham, Samar Awadh Alanazi, and Laura García-Hernán, “Predicting Depression Symptoms in an Arabic Psychological Forum”,IEEE Access(Volume:8), 2020, DOI: 10.1109/ACCESS.2020.2981834

[11] Mansur Alp Tocoglu, OkanOzturkmenoglu and Adil Alpkocak, “Emotion Analysis From Turkish Tweets Using Deep Neural Networks,” in Proceedings of the 2019IEEE Access (Volume: 7). ACM, 2017, DOI 10.1109/ACCESS.2019.2960113.