ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 8, Issue 5, May 2021

Human Behavior Analysis: A Comprehensive Review

Suraj Raina

M.Tech Student, SMVDU, Katra
Email: 19mms014@smvdu.ac.in

***Abstract*----Social networks are rising too quickly now. Every single person is on social media and talking to each other. All this data could be used to evaluate and forecast the very actions of a person so that human psychology could be encapsulated in a machine. What a person thinks or what is in its mind can be now predicted by using sentiment analysis techniques. Various researchers have worked on sentiment analysis techniques and predicted the output parameters like emotions, stress, etc. The objective of this paper is to analyze the human behavior in terms of his anxiety, mood, mental issues hence, state of mind. In this paper we have done the comprehensive analysis of various such techniques and the databases used by them.**

***Keywords*--- Sentiment analysis, social network**

I. INTRODUCTION

We humans still want to be linked to similar minds and communities where we might share our views on almost everything. In such an atmosphere, people are trying to communicate their feelings and these thoughts may be grouped together on the basis of common interests. Social media plays a crucial role in gathering all these knowledge and getting people together and with the rise of investing more and more time on social networks. Forecasting has become a central issue of actions of a human being. Every work today is related to social media, social networks and online communities. As has already been said, social networks deal with people who like to communicate and expand their contact list whereas online groups are small and they only like to be in groups where they could find people with similar type of mindset.

Analysis[1] is the techniques for arranging the perspective passed on in relation to the solitary article. With the preparation of various technical instruments, it has become a significant measure to be conscious of the mass cognizance o industry, in merchandise or in issues of general similarity and dismay. Following the assumptions underlying web-based media messages will assist with associating the setting in which the shopper reacts and progresses.

Sentiment analysis refers to mining of text which recognizes and separates abstract data in source material, assisting business with understanding the social opinion of their image, item or administration while observing on the web discussions. It is the most widely recognized content arrangement apparatus that investigates a coming message and tells whether the hidden emotion is positive, negative our non-partisan.

The study consists of brief steps in the analysis and categorization of the clustered data on the basis of various parameters. The review paper is presented by different researchers and different models are defined and a general consistency is presented on, which provides a reliable and unified performance of what to achieve.

The journals have been checked and all work carried out has been briefed. Although separate but powerful methods are offered for the work undertaken, a more cohesive solution could be made that could outperform current ones in terms of pre-processing and convergence of parameters.

II. RELATED WORK

In Chakraborty, K., et al [1] A synopsis of data acquisition, pre-processing, Sentiment analysis and opinion mining has been outlined.

Data is generally readily available on Internet and pre-processing of data is done usually clustering similar data together, this is done using k-means, apriori and eclat algorithms A person could be influenced by various posts, news surrounding him and selecting the most influential node is necessary in terms of detecting the source of a news, SenderRank is a model made for such findings.

Positive and negative links of a person could be find out using sentiment and opinion mining methods.

Sentiment analysis is done using lexicon based or classification based approaches.In lexicon based, lexicals are processed. An example of such is term frequency and term scoring approach. In classification based, classifiers like support vector machine, Naive bayes, logistic,etc are

used, With all such algorithms it is found that Naive bayes and Logistic regression performs well than other classifiers with an accuracy of 75% on IMDB dataset of movie reviews.

Bashir, S. (2016).et al SentiMI [2] provides opinion mining and sentiment analysis. It uses an already made model named SentiWordNet with an addition of sentiment dictionary based on mutual information of terms. SentiWordNet is an unsupervised classification model which provides sentiment scores for English synsets with respect to the part of speech tag and their rank. The rank demonstrates the sense wherein a word is most normally utilized.

The mutual information process matching is similar to K-NN classifier. It matches the terms of testing data to training data and find the one which are more correlated to each other .The more correlated one's are ranked and the terms in training data is classified as per most matching like nearest k- nearest neighbors. This mutual information is calculated after classifying the terms into synsets. The synsets could be positive,negative or subjective. If the synset score is >0, it is termed as positive, if it's <0, it is termed as negative otherwise if it's 0, it's subjective. Together they make the SentiMI sentiment dictionary .The model gives an accuracy of >75% upto 80% on the dataset which is better than the previous one.

Gilbert, E.,et al (2014, May). [3] proposed a model named VADER(Valence Aware dictionary for Sentiment Reasoning)

The model uses a gold standard sentiment lexicon dictionary comprising of all micro-blog texts.

Models such as LIWC(linguistic Inquiry and Word count) and GI(General Inquirer) which categories sentiment of a word on the basis of sentiment score suffers from a problem of sentiment intensity and processing of micro-blog texts. To eliminate such issues, a dictionary incorporated with Effective norms for English Words(ANEW) was developed to attach a sentiment valence for intensity consideration. The model outperforms on twitter tweets with an accuracy of above 90% where slangs and punctuation were used while on the IMDB dataset it had an accuracy of above 75%.

Brooke, J.,et al [4] .It is again a tool to extract sentiments from text and they are particularly defined as positive and negative. The model works same like SentiWordNet with an improvement in taking word comparisons into account. Rather than ranking the words according to their sentiment score, it compares alike words and then put a SO score to it. When tested on Movie reviews it showed an accuracy of 75-80% which is somewhat higher than previous basic model.

Thelwall, M. (2017). TensiStrength [5], it is a framework to distinguish the strength of stress and relaxation in online media instant messages.A lexical approach and a set of rules are used to detect direct and indirect expressions of stress or relaxation. The model focuses more on putting more intensity where repeated punctuation's are used and are marked in the range of -1 to -5(stress) and +1 to +5(relaxation).

Exclamation marks boosts/rise the power of stress or relaxation within a sentence by 1.

Repeated punctuation with one/more exclamation marks rise the power of stress or relaxation within a sentence by 1.

Guo, Q.,et al [6], presents a deep neural model to detect psychological stress. Unlike previous Machine learning models, CNN provides a high classification rate and less preprocessing .The model uses LIWC dictionary to measure frequent stress and non-stressed text.

The model is then demonstrated to make client scope content attributes from low-level content attributes utilizing cross auto encoders in CNN. It performed well than other classification model like SVM, Naive Bayes with an accuracy of upto 80% on twitter data.

From all such models, it can be observed that though such classification techniques such as Logistic regression, SVM give a good score in determining the current mood of a person but when the data is in abundance and a lot of words containing data is inputted, the accuracy of these models fluctuate a lot.

Thus, it can be concurred that CNN in performing to such performs well and gives a good and stable accuracies.

III. DISCUSSIONS

There are various feature extraction and classification methods available. Deep learning techniques have shown promising results in overall processing and results. Combining various such models together could result in reducing a lot of overhead and achieving higher accuracy results. Table 1 illustrates the analysis of different models used for sentiment analysis and different parameters related to it.

Feature extraction and selection is the most important part of any text classification. Before feature selection , preprocessing of the data must be done. Reducing Pre-processing overheads have always been a topic to work on and future scopes are there related to it. Also, with analysing all such data together and predicting of human behavior, an important application of reducing suicide ideations could be made. Suicide ideation can be viewed as

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 8, Issue 5, May 2021

an task of binary or multi-class classification.

Table 2 shows the overall accuracies analysis of different algorithms when performed on the database of imdb movie reviews and twitter texts. We can clearly see that while VADER gives more accurate results at some areas like miceoblogs but CNN provides a steady results as compared to rest of the algorithms in terms of a variety of texts.

Figure 1 gives a two-dimentional perception of Spearman correlation of words with stress on X-axis, and with web-based media stage on Y-axis. The words towards the X-axis are exceptionally associated with stress, yet less particular along singular web-based media stages. The words at Y-axis are exceptionally related with platform utilization yet less prescient of stress. With the end goal of this plot, we have disposed of words with non-huge relationships just as $\rho <= 0.05$. In this manner, the words towards the diagonals are the ones which best recognize Twitter use from Facebook use to communicate stress.

The absence of words in right base quadrant recommends that language on Twitter of high stress is a subset of high stress language on Facebook. These discoveries spur our proposition to adjust stress models prepared on Facebook language, a more comprehensive corpus, to Twitter language, with the end goal of region level stress forecast.

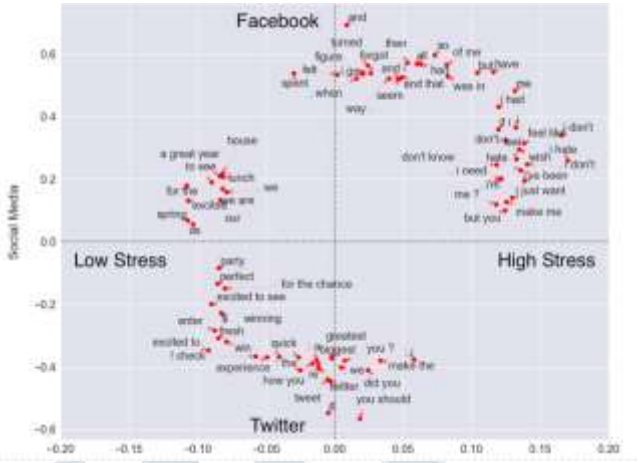

Fig 1: Bonferroni corrected at p < 0.01. Correlation of words with Social Media platform (Facebook on the positive y-axis and Twitter on the negative y-axis) and Stress Scale. Correlations are remarkable [7]

Table 1 Literature review

Reference and Year	Objective and Paper	Dataset	Access	Feature Extarction methods	Classification method	Implementation results
Koyel Chakraborty et al., 2019 [1]	Description of the data collection, pre-processing, Emotional analysis on social media data	IMDB movie review dataset	https://www.kaggle.com/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews	Lexicon based	Naive Bayesian, Logistic regression	Accuracy of 75% overall
F.H.Khan et al., 2016[2]	Senti MI: Introducing point-wise mutual information with SentiWordNet to improve sentiment polarity detection	IMDB movie review dataset	--do--	SentiWordNet	K-NN	Accuracy of upto 80%
Hutto, C., & Gilbert, E. (2014, May)[3]	VADER: a parsimonious model of social media sentiment analysis	IMDB Movie review dataset and Twitter microblogs	--do--	LIWC	-	Accuracy of upto 90% on microblog texts and 75% on normal

Brooke, J., (2011)[4]	Lexicon based Sentiment analysis methods SO-CAL	IMDB movie review	--do--	LIWC	-	Accuracy of upto 75-80%
Feng, L. (2014, November)[6]	Deep neural model for identification of psycological tension	Twitter texts	NA	LIWC	Cross-encoding	Accuracy of upto 80%
Buffone, A., et al., (2019, July) [7]	Understanding and measuring psychological stress using social media.	Facebook and Twitter Texts	NA	LIWC	Correlation between TensiStrength score and Topics	NA

Table 2 Accuracy analysis of all such algorithms on the above database.

Technique	Accuracy
SVM	51%
Logistic Regression	75%
Naive Bayes	75%
SentiMI	>75%
VADER	75-90%
Lexicon	75%
TensiStrength	-
CNN	80%

IV. CONCLUSION

The conclusion can be summarized in terms of parametric outcome. Some of the models illustrated could be really good in getting a good score of sentiment in normal texts but others do exceptionally good in terms of micro-blogs. A unified approach of such could be a future scope of it with minimization of pre-processing time. Modifying the feature exaction (pre-processing phase) could really boost the outcome of any classification model.

Due to widespread of internet in coming era a lot of possibilities could be attained in determining the human behavior with respect to whether it is mental illness or whether a random post is influencing the behaviour of a person or not. With this certain precautionary steps to self and publicly could be find out.

REFERENCES

[1] Chakraborty, K., Bhattacharyya, S., & Bag, R. (2020). A survey of sentiment analysis from social media data. *IEEE Transactions on Computational Social Systems*, *7*(2), 450-464.

[2] Khan, F. H., Qamar, U., & Bashir, S. (2016). SentiMI:Introducing point-wise mutual information with SentiWordNet to improve sentiment polarity detection. *Applied Soft Computing*, *39*, 140-153

[3] Hutto, C., & Gilbert, E. (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 8, No. 1).

[4] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics*, *37*(2), 267-307

[5] Thelwall, M. (2017). TensiStrength: Stress and relaxation magnitude detection for social media texts. *Information Processing & Management*, *53*(1), 106-121

[6] Lin, H., Jia, J., Guo, Q., Xue, Y., Li, Q., Huang, J., ... & Feng, L. (2014, November). User-level psychological stress detection from social media using deep neural network. In *Proceedings of the22nd ACM international conference on Multimedia* (pp. 507-516).

[7] Guntuku, S. C., Buffone, A., Jaidka, K., Eichstaedt, J. C., & Ungar, L. H. (2019, July). Understanding and measuring psychological stress using social media. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 13, pp. 214-225).

[8] Hu, X., Tang, L., Tang, J., & Liu, H. (2013, February). Exploiting social relations for sentiment analysis in microblogging. In *Proceedings of the sixth ACM international conference on Web search and data mining* (pp. 537-546).

[9] Gopalakrishna Pillai, R., Thelwall, M., & Orasan, C. (2018, April). Detection of stress and relaxation

magnitudes for tweets. In *Companion Proceedings of the The Web Conference 2018* (pp. 1677-1684).

[10] Hasan, A., Moin, S., Karim, A., & Shamshirband, S. (2018). Machine learning-based sentiment analysis for twitter accounts. *Mathematical and Computational Applications*, *23*(1), 11.

[11] Yuan, J., You, Q., & Luo, J. (2015). Sentiment analysis using social multimedia. In *Multimedia data mining and analytics* (pp. 31-59). Springer, Cham

[12] Ortega, R., Fonseca, A., & Montoyo, A. (2013, June). SSA- UO: unsupervised Twitter sentiment analysis. In *Second joint conference on lexical and computational semantics (* SEM)* (Vol. 2, pp. 501-507)