International Journal of Computer Science and Informatics

Volume 4 | Issue 3 Article 5

February 2022

An Optimized Machine Learing Framework For Extracting Suicide Factors Using K-Means++ Clustering

Naren S R Mr.
Kumaraguru College of Technology, Coimbatore, India, naren.19it@kct.ac.in

Thirumal P C Dr.
Kumaraguru College Of Technology, Coimbatore, India, thirumal.pc.it@kct.ac.in

Sudharson D Dr.
Kumaraguru College of Technology, Coimbatore, India, sudharsondorai.ads@kct.ac.in

Recommended Citation
S R, Naren Mr.; P C, Thirumal Dr.; and D, Sudharson Dr. (2022) "An Optimized Machine Learing Framework For Extracting Suicide Factors Using K-Means++ Clustering," *International Journal of Computer Science and Informatics*: Vol. 4 : Iss. 3 , Article 5.
DOI: 10.47893/IJCSI.2022.1197
Available at: https://www.interscience.in/ijcsi/vol4/iss3/5

An Optimized Machine Learning Framework For Extracting Suicide Factors Using K-Means++ Clustering

Mr. Naren S R[1], Dr. Thirumal P C[2], Dr. Sudharson D[3]
[1,2,3] Kumaraguru College of Technology,
Coimbatore, India

***Abstract*—** Suicide has emerged as one of the serious problems which should be eradicated from the society. People with suicidal thoughts restrict themselves by not expressing thoughts to the people around them. Studies have shown that people show more interest in expressing their thoughts over social media platforms. So, research has been conducted to identify people with suicidal ideation by analyzing the posts which they posted in social media platforms. Certain studies mined out new factors which influenced people to commit suicide, but those factors had certain drawbacks in it. This paper mainly focuses on overcoming those drawbacks in the factors. A new modified approach for extracting those risk factors is introduced as it can be used for future works related to suicidal ideation detection tasks. Statistical methods were imposed on the data to mine out the underlying characteristics of the features. K-Means++ clustering algorithm was implemented to extract the modified features. The modified features were given as an input for a testing classifier, and it attained an accuracy of 75.13%.

Keywords—suicidal factors, suicide ideation detection, k-means clustering, feature extraction, first person singular pronoun, reddit.

I. Introduction

According to World Health Organization, a rough estimate of more than 703,000 people are committing suicide every year [1] which makes it as an important problem in our society. A feel of hopelessness and depression are certain factors by which people develop suicidal thoughts [2]. People who failed in previous suicide attempt are more vulnerable for committing suicide [1]. Suicides can be prevented by diagnosing people with suicidal thought at the right time. In many cases, consulting a trained psychiatrist and unveiling the suicidal thoughts to them will be the best solution to prevent suicides. But the hindrance here is, most of the people with suicidal thoughts restrict themselves by not expressing their intent to the people around them and moreover they won't seek for counselling and get prescribed medications from the physicians. This paper focuses on extracting new features which can provide better accountability for classification of suicidal texts. The text data was extracted from the social media platform, and it was subjected to preprocessing techniques. Statistical analyses were conducted on the text data and patterns were observed. Clustering algorithm was implemented to mine out the features which significantly affected the classification process in a better way.

II. LITERATURE REVIEW

Studies show that people tend to express their thoughts and emotions over social media platforms like twitter, reddit etc. [3]. So, various social media platforms can be used by suicide prevention organizations for diagnosing people with suicidal ideation [3]. Suicidal ideation or suicidal intent is evolved in a person's mind generally because of severe stress, trauma, fear etc. If not treated properly at early stage, this leads people to develop suicidal thoughts which lead to suicide. Several researches related to analyze the depression level of individuals over texts provided accountability for further exploration in identifying suicidal thoughts in people. People with depression use First person singular pronouns than the people who never got depressed [4] and depressed people show a peculiar language pattern in their texts [5,6]. Certain works focused on finding a relation between the word usage of depressed people in various stages of depression like mild, moderate and severe [7,8].

Clustering algorithms were imposed on general texts [9] and the clusters have been implemented based on the use of hashtags in the texts [10,11] which were posted by people over social media platform which helps other people to skim over the genre of the post and to find new patterns in them. Text data were extracted based on the hashtags which people used in their posts and classified them whether individuals have depression or not [12]. To identify posts with suicidal ideation, supervised machine learning classification was implemented to classify whether that specific text had suicidal thought or not [13,14].

To have a better vision on the genre of each text given by people having suicidal thought, several topic modelling techniques like negative matrix factorization [15,16] and Latent Dirichlet Allocation (LDA) [9] were imposed on suicidal texts which showed the topics covered by texts posted by people having suicidal thoughts.

III. METHODOLOGY ADOPTED

This paper focuses on identifying people with suicidal ideation over social platforms in order to recognize their thoughts so that suicides can be prevented. The methods employed in this work are as follows, i) Data collection ii) Data Preprocessing iii) Feature Extraction iv) Unsupervised learning v) Result evaluation. Unsupervised learning algorithm was implemented on the text data to mine out new information to introduce a new feature which enhances the classification process.

IV. DATA COLLECTION AND ANNOTATION

A. *Collection of data from Reddit*

Twitter data has been used in many research works over various languages for identifying people having suicidal ideation [17, 18]. Reddit is an online discussion platform where people express their thoughts within a specific subreddit amongst other people. The main reason for choosing Reddit over Twitter is that Reddit has no word limit barrier for the users and so they can express their thoughts elaborately so that this work can be implemented in any piece of text irrespective of its length. For texts with suicidal ideation, the data was gathered from Suicide Watch subreddit. For texts with no suicidal ideation, the data was extracted from the Teenagers subreddit as it had texts with randomized topics involved in it which can resemble a population with normal thoughts.

B. *Preprocessing*

Preprocessing techniques employed over here includes lowering the characters, removing the punctuations and non-alphabetic characters which included numbers and unwanted whitespaces. To access each word in a post for normalizing it, the sentences were tokenized. Normalizing the data included stop words removal and lemmatization along with part of speech tag. Stop words are words which don’t contribute much meaning to a sentence and so it can be removed. Lemmatization technique was imposed on every sentence as it increases the efficiency by casting every word to its root form.

C. *Identifying suicidal texts based on presence of keywords*

The texts gathered from Suicide Watch subreddit had posts posted by people having suicidal thoughts. On the other hand, it also had posts posted by people who encourage other peers to overcome their thoughts. The texts were labelled as ‘suicidal texts’ only if the text had strong emotion of killing themselves [20,19] So, the posts containing most common key phrases pertaining to suicidal thoughts like ‘kill myself’, ‘end my life’, ‘hang myself’, ‘want to die’, ‘don’t want to live’, ‘my suicide attempt’ were extracted and labelled as suicidal texts. This criterion was brought under consideration as a previous suicide attempt is the strongest factor for people committing suicide [1].

D. *Topic modelling*

Latent Dirichlet Allocation technique was employed to extract the topics of each corpus. It was observed that the normal people without suicidal ideation had a vast extent of topic diversity in their corpus but on the other hand, texts having suicidal ideation had more topics closely related suicidal thoughts as shown in Table *1*. So, the texts which were segregated as suicidal texts based on presence of keywords clearly had suicidal intent in it.

Sl No	**Topics used by people having suicidal ideation**	**Topics used by people not having suicidal ideation**
1	*I, year, my, life, job, work, don't, time, want, a*	*Play, game, you, song, post, bore, video, dm, comment, music*
2	*School, I, my, college, life, class, parent, want, get, a*	*Na, school, gon, wan, class, teacher, online, work, test, high*
3	*I, life, people, a, like, be, make, don't, the, my*	*Filler, text, award, karma, trump, plus, optional, free, post*
4	*Amp, carbon, painless, monoxide, hang, rope, helium, tie, garage, unconscious*	*People, the, say, world, woman, cling, be, u, gay, a*
5	*I, want, don't, feel, know, like, life, kill, think, die*	*Eat, birthday, a, big, step, God, food, cuddle*
6	*I, tell, my, say, be, time, the, know, want, think*	*I, like, know, don't, make, want, feel, think, people, think*
7	*I, my, na, want, mom, like, kill, life*	*Cheese, filler, cat, nnn, chicken, eat, blah, n*
8	*Ugly, look, fat, kill, girl, I, like, hair, face, get*	*Gtpoplt, join, drink, water, cum, wish, luck, Christmas, pm, curious*
9	*Gt, ampbsnp, d, neck, rope, suspension, noose, necktie, attempting*	*Like, I, friend, say, girl, ask, guy, talk, want, tell*
10	*I, feel, like, friend, think, know, really, time, want, to, be*	*Day, my, time, the, today, mom, get, go, I, watch*

Table 1. Topics used by the people of both the categories.

V. FEATURE EXTRACTION

By analyzing various classification techniques employed before**,** a new feature which tries to differentiate whether the text has suicidal ideation or not can be extracted as it will engender a better solution for this problem. Certain works focused on this area and classified texts with new features developed by them like using POS tagging, identifying topic concepts, and calculating sentiment score for each word varying in a scale of 0-1 to add it as extra features for classification [22,23,24]. Number of tokens was also taken as a feature for classification purpose [16]. Language usage of people having suicidal ideation will vary a lot form the normal people without suicidal intent and this can be modified into a feature for classification process.

Depressed people often use **F**irst **P**erson **S**ingular **P**ronoun (**FPSP**) in higher frequency than non-depressed people [4] as depressed people had higher mean value of **12.17** than non-depressed people **10.76**. Many works have been conducted with this statistical approach to analyze texts in different genres and languages [6,7]. FPSP frequency in each post was calculated and it was finally divided by the length of the post to normalize it [25,26]. For better picture of this feature, the frequency of First-Person Singular Pronouns (FPSP) like 'I', 'Me', 'Myself' in suicidal texts were compared with texts having no suicidal ideation. The frequency of FPSPs in suicidal texts was quite higher than the non-suicidal texts as shown in the *Table 2*.

Class of the text	***Corresponding frequency of FPSP words in it***
Suicidal ideation	10,72,786
No suicidal ideation	1,43,224

Table 2. Showing the usage of FPSPs by both the classes in their texts.

The median and Interquartile Range (IQR) for the frequency of FPSP in both the categories were also calculated and both the values were much higher for suicidal texts than non-suicidal texts as shown in *Table 3*.

Metrics	***Frequency of FPSP in suicidal texts***	***Frequency of FPSP in non-suicidal texts***
Median	*20*	*2*
Inter Quartile Range	*26*	*3*

Table 3. Median and IQR for frequency of FPSP in both texts

There was a significant difference in FPSP usage between suicidal texts and non-suicidal texts as suicidal texts had FPSP in higher frequency. And thus, frequency of FPSP was considered as one factor for classification purpose. The frequency of FPSP and length of the post posted by each person was analyzed and visualized. In both the histograms (Fig 1, Fig 2), it was observed that non-suicidal texts were concentrated near the origin as they had very low range of frequency of FPSP and length of the post. On the contrary, suicidal texts skewed on right side showing higher range of frequency and length.

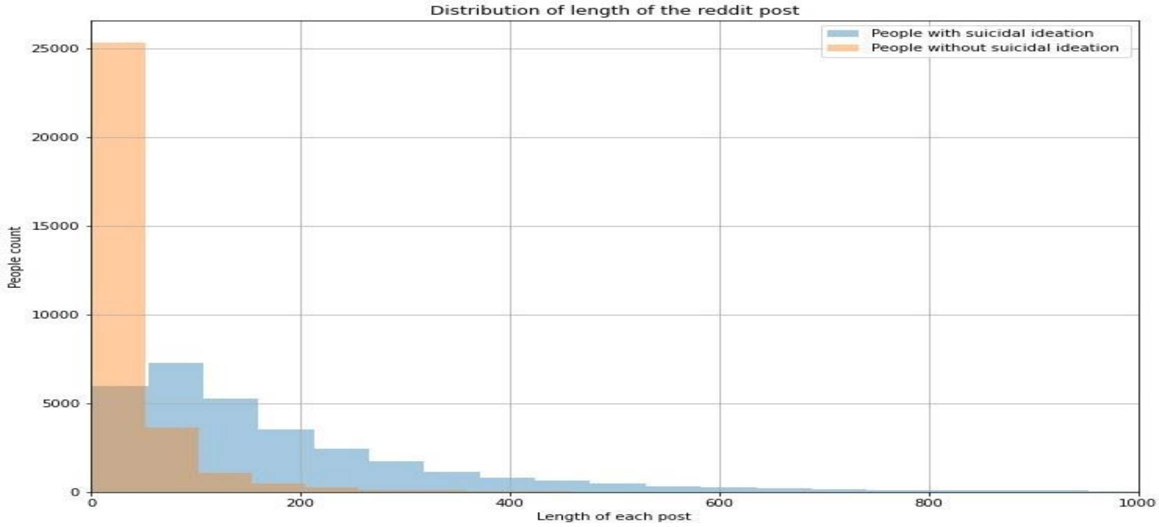

Fig.1. Histogram showing the distribution of Length of the reddit post.

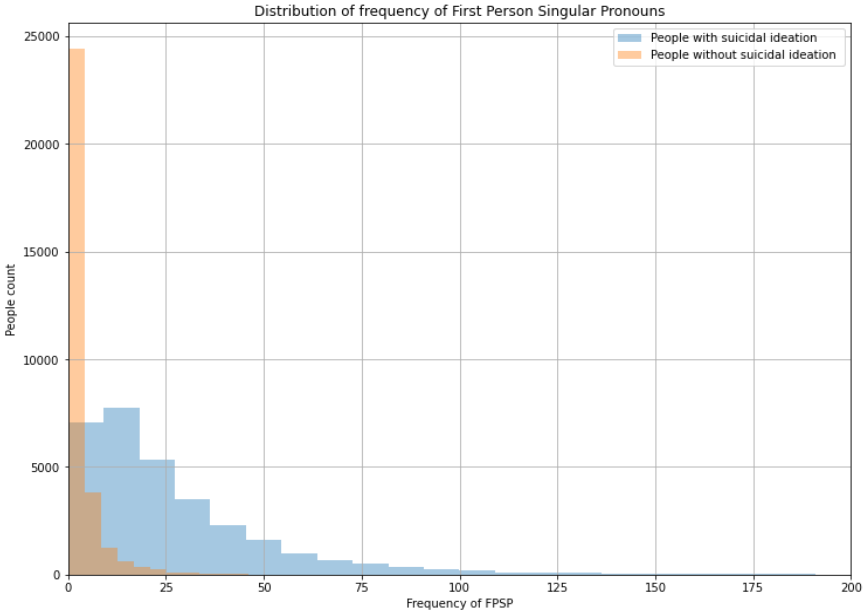

Fig.2. Histogram showing the distribution of Frequency of FPSPs in the reddit posts.

VI. UNSUPERVISED LEARNING

After interpreting the distribution of the frequency of FPSP, it was clear that frequency of FPSP in each post contributes more individually than using it along with length of the post. K-means++ clustering algorithm was imposed on the text data based on the number of words in the reddit post and frequency of FPSPs in it. Clustering is a technique which can find patterns in an unlabeled dataset by forming clusters. Clusters were made to get divided based on the length of the post and corresponding frequency of FPSPs.

The length of the post was taken in X axis and the frequency of FPSP was taken at y axis. These 2 factors were chosen in order to find any patterns in which the algorithm can cluster people. The number of clusters to be initiated by the algorithm is determined by elbow method Fig.3. Elbow method is the process of identifying ideal number of clusters to be given by plotting the k value in X axis and the corresponding inertia value in Y axis. A point should be chosen such that both K value and inertia should be minimum. As the graph starts to decrease in a linear fashion from the k value 4, k=4 was taken as the ideal number of clusters to be initiated for the clustering process (*Fig 3*).

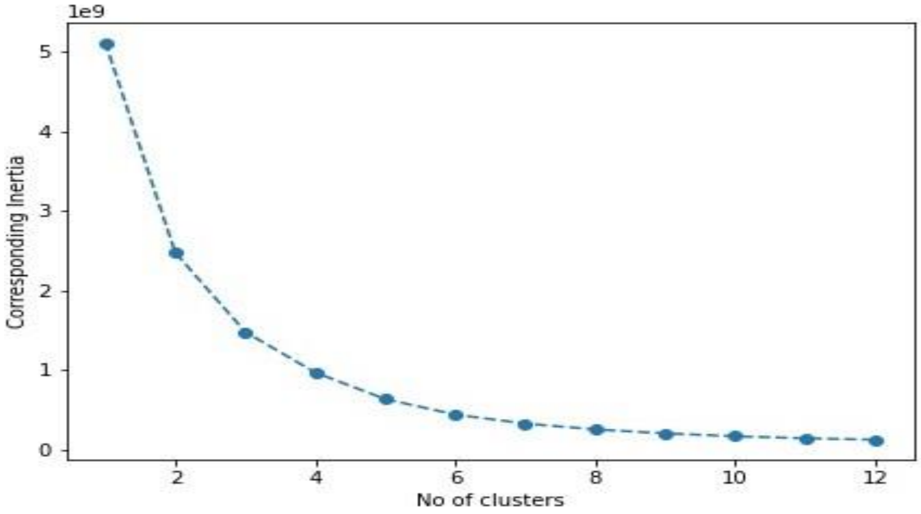

Fig.3. Elbow method to determine the ideal number of clusters

VII. RESULT AND EVALUATION

After clustering (*Fig 4*, *Fig 5*), it was observed that many people who had no suicidal ideation fell within the first few clusters near the origin which had small length and lower frequency of FPSPs (*table.4)*. The clusters had higher level of distinction with respect to the length of the post. Then the K value was increased gradually to observe certain key points like, *i)* How the clusters get separated, either significantly or by getting overlapped when the k value is increased.

ii) How the algorithm separates texts with suicidal ideation from non-suicidal texts.

iii) Is there any possibility that it separates suicidal texts from non-suicidal texts distinctively within certain clusters?

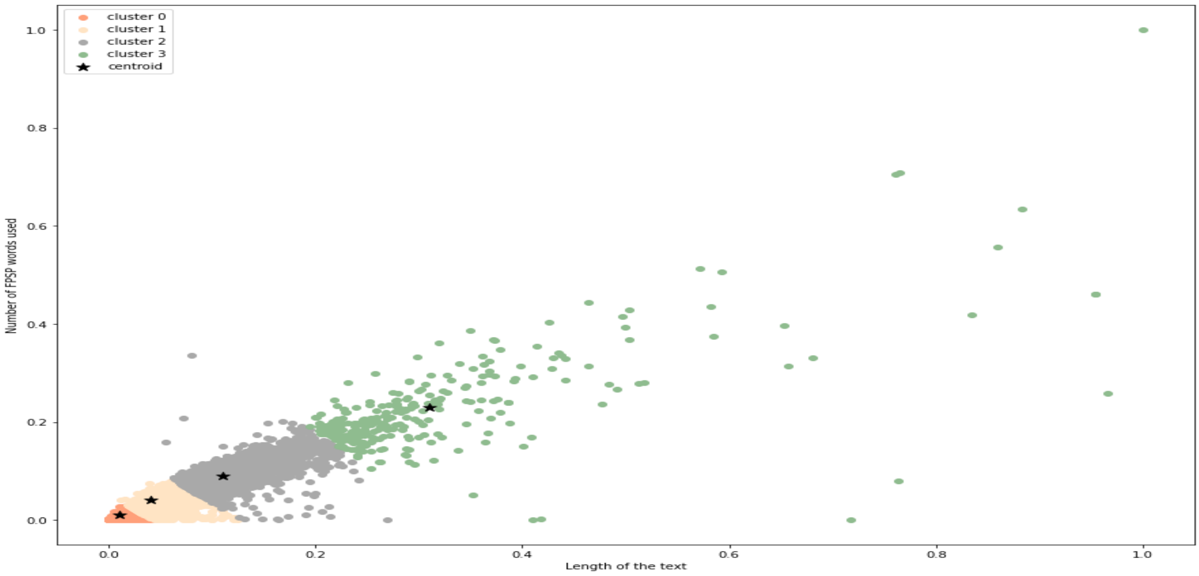

Fig.4 Formed clusters with k value as 4 using elbow method

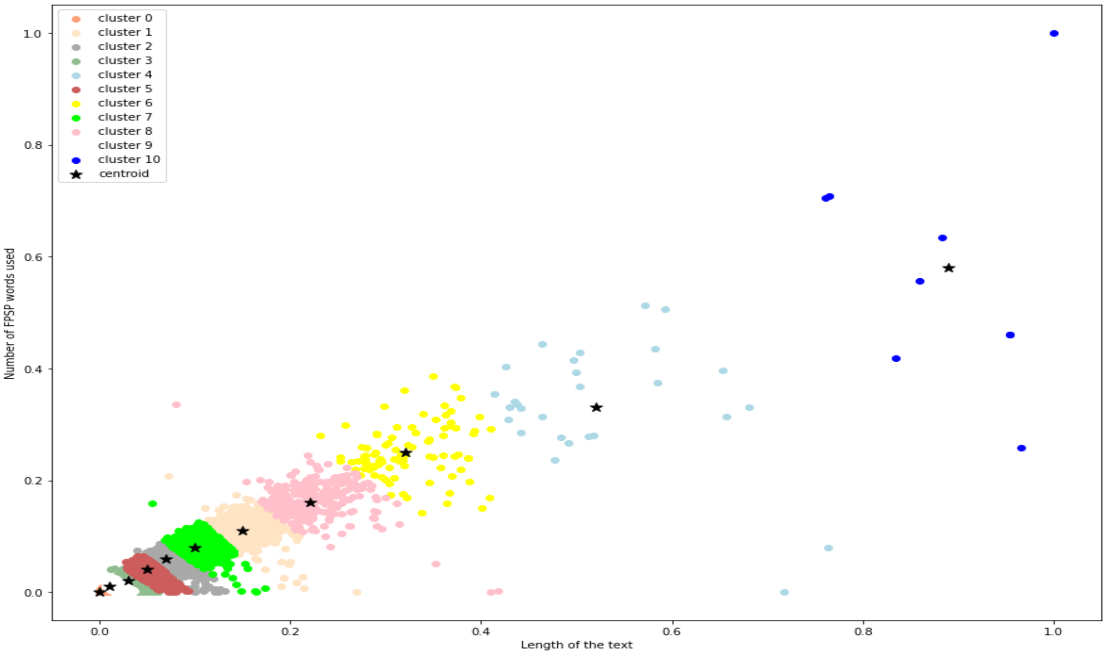

Fig 5. Clusters formed when k value as 11 for analysis

Cluster number	***Range of length of the reddit post***	***Range of frequency of FPSPs in the post***	***Count of non-suicidal texts***	***Count of suicidal texts***
1	0.000 – 0.016	0.000 – 0.012	25892	4614
2	0.005 – 0.038	0.000 – 0.031	3835	9357
3	0.016 - 0.064	0.000 - 0.044	978	7556
4	0.029 – 0.100	0.000 – 0.074	289	4837
5	0.048 – 0.143	0.001 – 0.098	108	2589
6	0.055 – 0.198	0.000 – 0.160	34	1313
7	0.073 – 0.291	0.003 – 0.208	13	569
8	0.080 – 0.393	0.001 – 0.337	12	250
9	0.258 – 0.518	0.150 – 0.443	2	74
10	0.631 – 0.958	0.000 -0.331	7	0
11	0.571 – 1.000	0.315 - 1	1	12

Table 4. Overview of clusters which got separated when k = 11.

People without suicidal intent mostly got stacked within the initial cluster itself. Even if the K value got increased, the pattern of cluster division remained the same as it kept on dividing at specific length and frequency of FPSP. The margin between the suicidal texts count and non-suicidal texts count gradually increased when the length of the post increases. One of the clusters had only non-suicidal texts in it (*Cluster 10 in Table.4*). When examined, it had higher range of length but comparatively lower FPSP frequency range. This shows that both the parameters together influence the texts in certain specific manner.

For inspecting the goodness of the clusters formed at k=4 and k=11, silhouette coefficient was calculated for both. It is calculated by using the mean intra-cluster distance and mean inter-cluster distance. If the value is negative, the data samples are assigned to wrong cluster. If 0, the clusters get overlapped. The clusters are considered to be good if the value is greater than 0. The silhouette score remained positive (*table.5*) even if the k value was increased up to 11 showing that there was no overlapping and no wrong cluster in it. As all the clusters got divided at a certain range of FPSP frequency and length, these 2 metrics can be individually considered for classification purpose, and it can provide better accountability for the problem.

SL NO	***K-VALUE***	***SILHOUETTE SCORE***
1	4	0.6664
2	11	0.5429

Table.5 Silhouette scores for both the K-values are shown.

As research were conducted for extracting new suicide factor for classification, many considered FPSP usage as a feature, but they divided it by the length of the post to normalize it [22]. The problem with this approach is that a person with less FPSPs and less length of the post will be treated equally to a person with higher frequency of FPSP and higher length as the average becomes same for both [27,28]. For this reason, the frequency of FPSPs and length of the post were considered as 2 individual factors for classification as the clusters got divided such that a certain value of FPSP frequency and length fell under a unique cluster. Stochastic Gradient Descent Classifier with hinge loss was imposed as a testing classifier for classifying the texts based on these two factors and it achieved better accuracy with those 2 factors itself.

Two weights, 0.05 & 0.95 was initiated respectively for length and FPSP frequency and for each specific weight combination the accuracy was calculated. It was performed to check whether any individual factor among the 2 factors contributed more. The accuracy score deviated only for a small extent from 75.13% – 71.37%. SGDClassifier was used just as a testing classifier to check the accuracy of the model using the extracted factors.

VIII. CONCLUSION

Current study confirms that frequency of first-person singular pronouns and the corresponding length of the text can be directly considered as a factor for classifying texts with suicidal ideation without normalizing the FPSP frequency by dividing it with length of the post. Further work can be progressed in this by analyzing other factors which influences suicidal intent. Also with small enhancements, Classification process can be done using powerful classification algorithms in future to identify people with suicidal thoughts over social media posts. For further research to solve this problem, we hope to enhance the classification process by implementing ensemble techniques by combining outputs of various algorithms using these modified factors to yield a better classification accuracy.

IX. REFERENCES

[1] World Health Organization, 2014. *Preventing suicide: A global imperative*. World Health Organization.

[2] Furr, S.R., Westefeld, J.S., McConnell, G.N. and Jenkins, J.M., 2001. Suicide and depression among college students: A decade later. *Professional Psychology: Research and Practice*, *32*(1), p.97.

[3] Luxton, D.D., June, J.D. and Fairall, J.M., 2012. Social media and suicide: a public health perspective. *American journal of public health*, *102*(S2), pp.S195-S200.

[4] Rude, S., Gortner, E.M. and Pennebaker, J., 2004. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, *18*(8), pp.1121-1133.

[5] Losada, D.E. and Crestani, F., 2016, September. A test collection for research on depression and language use. In *International Conference of the Cross-Language Evaluation Forum for European Languages* (pp. 28-39). Springer, Cham.

[6] Ramirez-Esparza, N., Chung, C.K., Kacewicz, E. and Pennebaker, J.W., 2008, March. The Psychology of Word Use in Depression Forums in English and in Spanish: Texting Two Text Analytic Approaches. In *ICWSM*.

[7] Smirnova, D., Cumming, P., Sloeva, E., Kuvshinova, N., Romanov, D. and Nosachev, G., 2018. Language patterns discriminate mild depression from normal sadness and euthymic state. *Frontiers in Psychiatry*, *9*, p.105.

[8] Şimşek, Ö.F., 2013. The relationship between language use and depression: Illuminating the importance of self-reflection, self-rumination, and the need for absolute truth. *The Journal of general psychology*, *140*(1), pp.29-44.

[9] Lossio-Ventura, J.A., Morzan, J., Alatrista-Salas, H., Hernandez-Boussard, T. and Bian, J., 2019, November. Clustering and topic modeling over tweets: A comparison over a health dataset. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 1544-1547). IEEE.

[10] Antenucci, D., Handy, G., Modi, A. and Tinkerhess, M., 2011. Classification of tweets via clustering of hashtags. *EECS*, *545*, pp.1-11.

[11] Sudharson, D Prabha, D. A novel machine learning approach for software reliability growth modelling with pareto distribution function. Soft Comput 23, 8379–8387 (2019). https://doi.org/10.1007/s00500-019-04047-7.

[12] Cavazos-Rehg, P.A., Krauss, M.J., Sowles, S., Connolly, S., Rosas, C., Bharadwaj, M. and Bierut, L.J., 2016. A content analysis of depression-related tweets. *Computers in human behavior*, *54*, pp.351-357.

[13] Chiroma, F., Liu, H. and Cocea, M., 2018, July. Text classification for suicide related tweets. In *2018 International Conference on Machine Learning and Cybernetics (ICMLC)* (Vol. 2, pp. 587-592). IEEE.

[14] Shahreen, N., Subhani, M. and Rahman, M.M., 2018, September. Suicidal trend analysis of twitter using machine learning and neural network. In *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)* (pp. 1-5). IEEE.

[15] Luo, J., Du, J., Tao, C., Xu, H. and Zhang, Y., 2018, June. Exploring temporal patterns of suicidal behavior on Twitter. In *2018 IEEE International Conference on Healthcare Informatics Workshop (ICHI-W)* (pp. 55-56). IEEE.

[16]Sudharson, D Prabha, D "Improved EM algorithm in software reliability growth models" International Journal of Powertrains, Vol.9 Issue.3, pp.186-199, 2020,Inderscience Publishers (IEL).

[17]Abboute, A., Boudjeriou, Y., Entringer, G., Azé, J., Bringay, S. and Poncelet, P., 2014, June. Mining twitter for suicide prevention. In *International Conference on Applications of Natural Language to Data Bases/Information Systems* (pp. 250-253). Springer, Cham.

[18]Lee, S.Y. and Kwon, Y., 2018. Twitter as a place where people meet to make suicide pacts. *Public health*, *159*, pp.21-26.

[19]Sudharson D, Ratheeshkumar A.M, P.Divya, Dr.Sakthi Govindaraju, A Novel AI and RF Tutored Student Locating System Via Unsupervised Dataset, Turkish Journal of Physiotherapy and Rehabilitation, vol.32, no.2, pp no. 882-887, 2021

[20]Sawhney, R., Manchanda, P., Singh, R. and Aggarwal, S., 2018, July. A computational approach to feature extraction for identification of suicidal ideation in tweets. In *Proceedings of ACL 2018, Student Research Workshop* (pp. 91-98).

[21]O'dea, B., Wan, S., Batterham, P.J., Calear, A.L., Paris, C. and Christensen, H., 2015. Detecting suicidality on Twitter. *Internet Interventions*, *2*(2), pp.183-188.

[22] Burnap, P., Colombo, G., Amery, R., Hodorog, A. and Scourfield, J., 2017. Multi-class machine classification of suicide-related communication on Twitter. *Online social networks and media*, *2*, pp.32-44.

[23]Sudharson, D., Prabha, D. "Hybrid software reliability model with Pareto distribution and ant colony optimization (PD–ACO)" International Journal of Intelligent Unmanned Systems, 2049-6427, 2020Emerald Publishing Limited

[24]Burnap, P., Colombo, W. and Scourfield, J., 2015, August. Machine classification and analysis of suicide-related communication on twitter. In *Proceedings of the 26th ACM conference on hypertext & social media* (pp. 75-84).

[25]Sudharson D, Dr Prabha D. A Review On The Prominent Factors Of Software Testing For Optimized Output Through Data Analytics, International Journal of Pure and Applied Mathematics, vol. 115, issue. 6, pp no. 95-106,2017.

[26]Vioules, M.J., Moulahi, B., Azé, J. and Bringay, S., 2018. Detection of suicide-related posts in Twitter data streams. *IBM Journal of Research and Development*, *62*(1), pp.7-1.

[27]Sudharson D, Dr Prabha D. A contemporary survey on the aspects of software reliability growth models via size of the software, International Journal of Pure and Applied Mathematics, 119, Special Issue, 1253-1271, 2018.

[28]Dr.B.Arunkumar, D.Sudharson, "A Novel Approach for Boundary Line Detection using IOT During Tennis Matches", Advancement of Electrical, Information and Communication Technologies for Life Application, Volume.13, Issue.4, pp.243-246 2020.