Sentiment Analysis using Social and Topic Context for Suicide Prediction

E. Rajesh Kumar[1]*
Department of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation, Vaddeswaram
Guntur 522502, Andhra Pradesh, India

K.V.S.N. Rama Rao[2]
Department of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation
Aziznagar, Moinabad (m), Hyderabad, Telangana

***Abstract*—In many fields, analysing large user-generated microblogs is very crucial and drawing many researchers to study. However, processing such short and noisy microblogs is very difficult and challenging. Most prior studies use only texts to find the polarity of sentiment and presume that microblog site is independent and distributed identically, ignoring networked data from microblogs. Consequently, not satisfied with performance motivated by emotional and sentimental sociological approaches. This paper proposes a new methodology that incorporates social and topic context to analyze sentiment on microblogs by introducing the meaning of structure similarity into social context. Unlike from previous research employing direct relations from user and by suggesting a new method to quantify structure similarity. In addition, to design the microblog semantic relation, topic context is introduced. The Laplacian matrix of these graph produced by these context combines social and topic context and Laplacian regularization is applied to the microblogging sentiment model. The Experimental results on the two datasets show that, the suggested model had reliably and substantially outperformed the baseline methods that is helpful for suicide prediction.**

***Keywords*—Social context; topic context; microblogging; Laplacian matrix; emotional and sentimental**

I. Introduction

Getting real user sentiment from huge collections of social media content created by users (e.g., microblogs) is a great challenge. Often, it is a great benefit and has a broad range of application opportunities for the sentiment of mining customers, such as business intelligence, recommendation system, customer management and relationship [1,2]. The role of automated sentimental study requires the system or machine to understand in deep of natural language [3], which has attained some results in formal analysis of sentiment related to text [4,5]. However, its output and performance are dropped when applied to microblogging sentimental analysis as it may consider text are independent and identically distributed. Microblogs are significantly shorter and have different type of expression compared to 'long formal text', e.g. 'it's so cooooooool!' and 'lol', aggravates the vocabulary sparsity problem. Conversely, social networking offers various kind of metadata, like user relation that can be controlled to boost the accuracy of sentimental analysis.

The study of effect on microblog sentiment analysis of other metadata a head of texts ('called social context') has recently exhibited more interest from researchers, such as applying direct user relationships to sentimental analysis models [6,7]. To support these approaches, two sociological theorems: emotional contagion [8], sentimental consistency [9] are used. As per social context, sentiment consistency is known as user context, it indicates that all post tends to have same sentiment label posted by same individual: Emotional contagion (EC) states that same opinion may appear to have for similar kind of people called friends context. While these studies were already exploited for sentimental analysis by considering the effects of direct relationship from users and ignoring the influence of indirect user relationship [6,7]. But social network connections are heterogeneous [10], so analysing sentiment analysis in microblogs using direct user relationship is not appropriate. For example, in Fig. 1, the blue dialog box signifies positive sentiment of text and red dialog box represents negative sentiment. Text reflected in black dialog box are the one to be categorized. From the user relation between Sam and John no direct relation exists but have common friends Alex and Joe. All users have different opinion about suicide, users may also be in a depressive mindset that may lead to suicide. Sam had posted a tweet related to suicide "I will never be unhappy", this sentence is a positive comment towards suicide. However, for a machine it is very difficult to detect the polarity for any sentence from its literal meaning. Further, by using direct user relationship among users to support sentimental analysis the text classification cannot be classified into classes as John's friend Alex and Joe has no comments on depression that results in a classifier error.

According to recent research, Indirect user relationships have recently been used into recommendation systems [11,34]. The principle of these works is that same preferences or behavioral patterns were found among similar users. However, based on small literature it studies the indirect user relationship in analysing sentiment. In same instance, homophily [12] has gained much more popularity with the growth of sociological theory. The principle is that interaction between similar individuals occurs at a higher rate than between dissimilar individuals [13], which has a significant impact on the creation of friendships. The data that flows through the network such as behaviour and culture appears to be limited. In addition, some indication of both negative and positive sentiment of homophily has found in social network [14].

Based on these research works; a new model is proposed to analyze microblog sentiment using user indirect relations by structure similarity (SS). This approach is by an assumption: similar user's opinions must be similar and tested this

*Corresponding Author

assumption experimentally. First, through mutual friend relationships, related users are found, and similarity matrix is established. Finding similar or related users through mutual friends [15,16] is a regular practice and new connections are generated by similarity [12]. In addition, the same opinion [15] may be shared by two users who may have a new link between them. The principle of this approach is to search for possible or potential relationship that could be friends among users and by considering them into model of sentiment analysis. Second, topic factors are created, and context matrix is of the subject is formed. On the same topic [13] the occurrence of homophily is highly significant and the context of the subject or topic in turn will better exploit the homophily theory. Finally, the context of user structure similarity and topic context are merged into a model of a graph, this graphs Laplacian matrix is used to evaluate microblogging sentiment. From Fig. 1, users Sam and John have common friends Alex and Joe. So, there can be several probabilities of being friends where they can express the same sentiment with specific probability by assumption. Therefore, Sam could also have negative comments on #suicide in response to Johns negative comments on #suicide, so the accuracy by relationship for sentimental analysis can be assured.

The main key contributions in this paper include:

- Proposing a technique for modelling homophily by applying structure similarity in social networks.
- The emergence of structural similarity as a replacement for user-direct relationships in the social context of microblogs.
- To model the semantic relationships between microblogs by introducing topic or subject context.
- Proposing a new sentimental analysis model for microblogs that integrates context of structure similarity, user context, topic, and text information context.
- Extensive evaluation of the proposed method using real-world datasets to recognize the working of proposed method.

II. Related Works

In this section, some related works about microblogging sentiment and sentimental analysis are reviewed.

A. Sentimental Analysis

First, confirm that you have the correct template for your paper size. This template has been tailored for output on the US-letter paper size. If you are using A4-sized paper, please close this file and download the file "MSW_A4_format". Existing model consists of two major categories: machine learning and lexicon-based method. Method based on lexicon [14,15,32] generally uses SenticNet [16], SentiWordNet [17], to tag positive and negative labels for terms occurring in sentence, then by summarizing the complete sentence by tagged words the sentiment of the document can be judged. Methods based on lexicons that do not require polarity label datasets are unsupervised. These approaches however depend too much on lexicons and domain related due to polarity change of words from domain to domain. The methods of machine learning view sentiment analysis as an issue of text classification [18,19]. In these methods, using text features like unigram, bigram and word embedding are extracted and applied to different classification techniques such as NB, SVM and so on. Machine Learning (ML) techniques are supervised with polarity labels and typically requires lot of training data. Hence, Classification accuracy is related to size of the data.

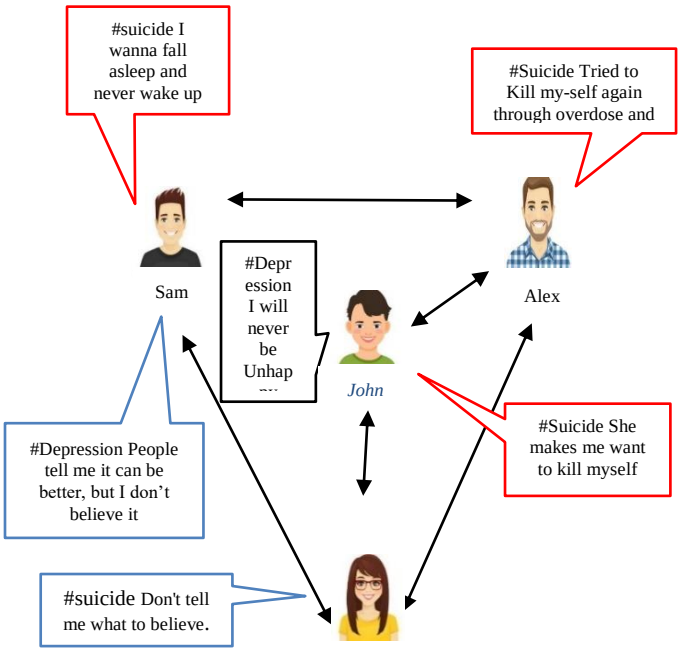

Fig. 1. User Direct Relationship.

B. Microblogging Sentimental Analysis

Over the years, microblog sentimental analysis is noisy, short, and become a hot research subject or topic [6,7,20] due to this problem many techniques are proposed to solve this issue. To analyze the opinion of tweets, [21] used emoticons features. In [22], repeated punctuations, generalized emoticons and words repeated were used to build a label propagation algorithm co-occurrence graph, this graph was used to identify the polarities of tweets sentiment. Using the relationship between emoticons and words, [23] used lexicon for feature sentiment extraction by developing a sentimental lexicon and for microblog analysis. All these above-mentioned strategies use only text information and ignores the additional information provided by microblog.

In recent years, there have been more and more studies on how to use user data to determine sentiment. [7,35] suggested a method to classify the sentiment of users on twitter using user '@' information and follow relationship. [24] brought user sentimental analysis to a specific subject or topic as a collaborative filtering problem, to predict user sentiment relationship between users were applied. Similarly, [25] has also manipulated the graph of user relation. In order to classify sentiment, entropy model with maximum result were used as labels and author applied label propagation approach. These working are method of classification of user topic level or user level sentiment, whereas the proposed model is microblogging level. In [6] author Hu et al., proposed a structure called SANT

('a sociological approach to noisy and short Text handling') that incorporates social context to characterize microblog sentiment. Based on [6], [26] included similarity of contents to the SANT system and suggested a semi-supervised model for tweet sentiment recognition. The author in [27] contended that context proposed by [6] was completely a content-based model, so for prediction level they suggested a framework for structured Microblog sentiment classification (SMSC). There also exists some works that have applied microblogs retrieval to user relationship [28]. All these approaches ignore user similarities and employ user direct relation.

III. Methodology

A. Dataset

In this paper, the experiment is performed on two different twitter datasets. These datasets were used for suicidal detection using different classification methods that included raw data with labelled sentiment.

Dataset1 is collected from microblogging websites consisting of trained dataset and test data set with labels positive and negative. It consists of five topics 'Evidence of suicide attempts', 'Suicide flippant reference', 'support or information', 'campaign or fight', 'suicide reporting', 'Condolence or memorial', 'None of these [33]. Dataset2 is created based on keywords [32]. Finally, users with friends are considered and remaining microblogs with no user friends are deleted.

B. Notations used

Uppercase letters such as M represents matrices, ***m*** indicates vector in bold, *m* represents scalar, M_{*i} is used to denote j^{th} column and M_{i*} denotes i^{th} row of the matrix. The matrix entry for row and column can be M_{ij}. Transpose of a matrix can be calculated by M^T. $||M||_F$ indicates Frobenius norm of matrix *M* and *tra(.)* for matrix traces.

The main objective of this paper is to construct a classifier W € Xm*p using the training matrix L € Xn*m (n- indicates feature and m represents number of microblogs) and labels the matrix Y € Xn*p (p indicates no. of polarities), classifier B € Xm*p is used to predict microblogs y that are unseen. Variable B indicates truth table, Ŷ=LC € Xm*c is used to represent matrix B truth table. Here, only binary classification is considered for sentiment, i.e., p=2. Consequently, the truth table is Bi*= [+1-1] for positive microblog and B*i = [-1+1] for negative microblog sentiment.

Consider an undirected graph G= (V, E) where, V indicates vertices and E represents edge. Mr represents microblog adjacency or relation matrix, Lm=Dm - Mr is Laplacian Matrix of G. Dm represents diagonal matrix and Dii is degree of ith vertex.

In Eq. (1) prediction feature is applied to identify microblog that are unseen. In Table II, variables, type, and their definitions are shown.

$$f(y)=\begin{cases}+1 & if \quad yC_{*1} > yC_{*2} \\ -1 & if \; yC_{*1} < yC_{*2} \\ +1 \quad or \quad -1 & randomly \quad if \; yW_{*1} = \quad yW_{*2}\end{cases} \tag{1}$$

TABLE I. Dataset Statistics

Emoticon	Dataset1	Dataset2
# of Twitter users	4562	5632
# of Twitter Tweets	147318	89141
# of Positive Tweets	53412	32765
# of Negative Tweets	78951	47697
Tweets avg. per user	32.21	15.84
Avg. friends per user	241.5	127.2

C. Microblog Content Modelling

For information related to text, standard least square method is used to satisfy the classification method. It aims to understand c-classifier and optimization problem after execution of (2) in terms of classification task in multiclass.

$$\min_y \frac{1}{2} \| LC - B \|^2_{R_m} \tag{2}$$

Unlike conventional text results, microblog leads to unigram sparse matrix due to noise and short in form. To deal this issue, sparse L1 regularization standard to seek for feature space by sparse reconstruction. To minimize the reconstruction error based on L1 norm, feature selection can be automatically implemented, and a sparse description can be obtained. To achieve a more stable model, L1 norm is implemented in the proposed model as shown in (3).

$$\min_y f(C;L;B) = \min_y \frac{1}{2} \| LC - B \|^2_{R_m} + \beta \| C \|_1 \tag{3}$$

Where, β is regulation weight.

D. Factors beside Text

In this section, various context is incorporated and integrated into a final model.

1) Integration of topic context: Hashtag is a mechanism that microblog services offer, using this service any user can insert information related to topics in microblogs. For instance, in a twitter tweets #symbols is used to tag tweet topic, let tweet "I wish to end my #life" represents tweets about "suicide thoughts". To express any kind of emotions, based on different topic people post in various microblogging site, in connection to same topic there can be same opinion by different users; different opinion for different topic, opinion of same topic with same person usually depend on each other. The importance of topic content is built to check whether more than one microblog text refer to same topic, to design connection with microblogs, it is better to include subject information into microblogging for sentimental analysis rather than text similarity. The microblog similarity value may be less by usage of text similarity leading to failure of sentiment efficiency. Using matrix-Mm in (4) "microblog-microblog matrix" for topic is obtained.

$$M_m = M_x + M_x{}^{M_t} \tag{4}$$

TABLE II. VARIABLE MEANING

Variable	Variable Meaning	Variable Type
M in Uppercase	Matrix representation	-
m in Lowercase	scalar representation	-
m in bold and lowercase	Vector representation	-
M_{*i}	Matrix-M (i^{th} column)	-
M_{i*}	Matrix-M (i^{th} row)	-
L	Feature matrix representation	X^{n*m}
B	Matrix truth table	X^{n*Sc}
$\hat{Y}$	Fitted label matrix	X^{n*Sc}
n	No. of features	*Int*
t	No. of training dataset	*Int*
Sc	Sentiment classification count	*Int*
x	No. of topic content	*Int*
C	Classifier	X^{t*Sc}
Y	Microblog feature vector	X^{t}
U_m	User Matrix	X^{r*n}
r	No. of users	*Int*
Ss	Structure similarity representation	X^{r*r}
M_t	Topic Matrix	X^{n*x}
M_m	Topic Microblog-Microblog Matrix	X^{n*n}
M_r	Relation Microblog-Microblog Matrix	X^{n*n}
D_m	Diagonal Matrix representation	X^{n*n}
L_m	Laplacian Matrix representation	X^{n*n}
R_m	User-user relation direct matrix	X^{r*r}
G	Graph representation	-
E	Edge representation	-
V	Vertices representation	-

Where, $Mt \in Xn*x$ is atopic matrix and $M_{t_{ij}}$ =1 if and only if (iff) i^{th} microblog is about jth topic information, $M_{m_{ij}}$ =1, if p_i and p_j microblog refers to same topic, D_m indicates diagonal matrix all assigned to zero.

2) Integration of user context: It is based on a theory called sentimental consistency. It recommends that tweet sentiment posted by same user in two microblog has greater probability than selection of random microblogs. Here, $M_r \in X^{n*n}$ indicates matrix sentiment consistency ('microblog-microblog').

To calculate $M_{r_{sc}}$ use (5), $Um \in X^{r*n}$ is a microblog matrix of user, where $U_{m_{ij}}$ =1, iff user-i post microblog-j and 'r' indicates no. of users.

$$M_{r_{sc}} = U_m{}^{M_t} + U_m \tag{5}$$

Where, $M_{r_{sc}}$ =1, iff p_i and p_j microblogs are posted by similar users.

3) Structure similarity content: This section is focused on sociological theory; emotional contagion, this indicates that sentiment posted by similar users in two microblogs has greater probability than selecting in microblogs randomly. In existing work, if sentiment posted by two different users in two microblogs with friends/followers related are connected, an approach is constructed to make two microblogs user sentiment to be closer as possible, it is known as friends' context. This is denoted by $M_{r_{ec}} = U_m{}^{M_t} * R_m * U_m$, where $R_m \in X^{r*r}$ indicates user matrix ('user-user') and $R_{m_{ij}}$ =1, iff there is a follower/followee relation among ith user and jth user. But existing work focused on direct user relationship and eliminating friends and follower's relationship.

As discussed in previous section, users can share the sentiment to the user "who is a friend of his/her friend", which is homophily communication. In this context, structure similarity is used to develop the EC sociological theory by considering friend relation. New relationship can be caused by common friends [29], example, if A and B have friend C in common, the friend probability will be increased between the users, this is known as "Triadic closure" [30].

The reality is A and B have friend relation to C that supports with confidence which is lacking with strangers at friendship creation is one of the causes for "Triadic closure". The second explanation is related on the motivation for C: it can minimize C's latent stress in two different relationship about bringing A and C closure.

In twitter with respect to three different users and three cases where users are closure and connected based on following two relations as shown in Fig. 2, 3 and 4. The incoming arrow to the user is pointed as followee and opposite user is follower. In case one as shown in Fig. 2, indicates the flow of communication between users, there can be a flow of opinion between Starc and Tony over Shane. In case two as shown in Fig. 3, two users tony and Starc with common followee Shane determining "friend of friend" relationship, if more followee's exist among two users, the construction of the relation can be made easier.

In case three as shown in Fig. 4, two users Tony and Starc with Shane as common follower. All three cases are indicating user similarity expressions, it implies possibility or relationship can be created between unconnected users. Due to this reason an undirected graph is taken for follower relationship.

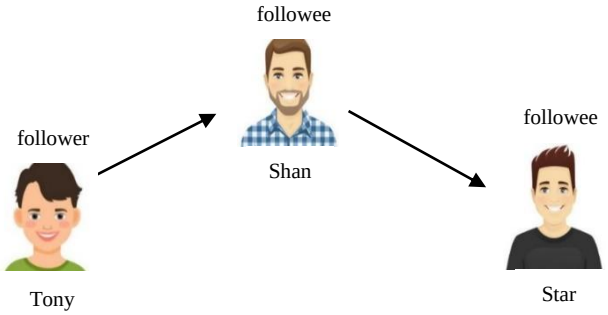

Fig. 2. Relation Type: Case One.

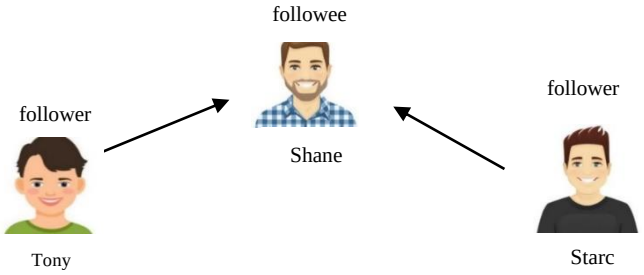

Fig. 3. Relation Type: Case Two.

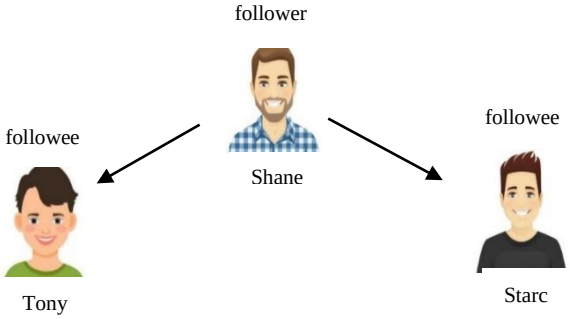

Fig. 4. Relation Type: Case Three.

Consider users v_i and v_j, Structure similarity can be analyzed by (6).

$$S_{s_{ij}} = sim(v_i, v_j) = |W_{ui} \cap W_{uj}| \tag{6}$$

Using similar friends between two user's Structural similarity (SS) can be determined. W_{v_i} indicates user V_i. $|W_{vi} \cap W_{v_j}|$ neighbours representing v_i and v_j number of friends in common. By including the condition as shown in Fig. 5, user Shane and Tony have common friends Joe and Starc. But in Fig. 6, user Tony has many friends, using (6) to calculate to SS for user Shane and user Tony in Fig. 6 will produce same SS value as obtained for Fig. 5. To manage this issue, all friends between two users are included to calculate SS.

$$s_{s_{ij}} = sim(v_i, v_j) = \begin{cases} \dfrac{|W_{v_i} \cap W_{v_j}|}{|W_{v_i} \cup W_{v_j}|} \\ \dfrac{|W_{v_i} \cap W_{v_j}|}{|W_{v_i} \cup W_{v_j}|} + 1 \end{cases} \tag{7}$$

Here, $W_{v_i} \cup W_{v_j}$ implies all set of friends (union) of both v_i and v_j users and $|W_{v_i} \cap W_{v_j}|$ indicates total users in the set. Once SS matrix Ss value is obtained, $M_{r_{ec}} \in$ C matrix can be calculated using (8).

$$M_{r_{ec}} = U_m^{M_t} * S_s * U_m \tag{8}$$

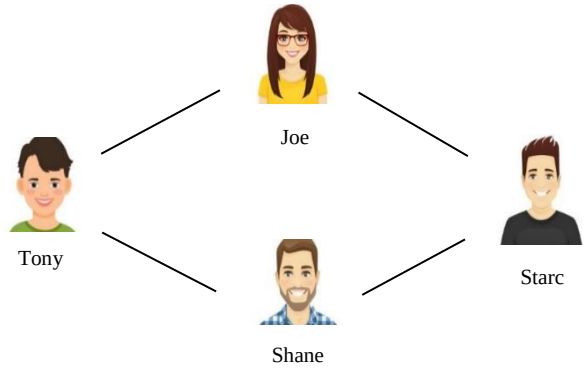

Fig. 5. Similarity Expression1.

4) Proposed model: Three types of context are incorporated in a framework. M_{r1} X^{n*n} represents both combination of SS context and user context that can be computed using (9). $M_{r2} \in X^{n*}$ indicates SS context, user context and subject or topic context, this can be calculated using (10).

$$M_{r_1} = M_{r_{sc}} + \beta * M_{r_{ec}} \tag{9}$$

$$M_{r_2} = (M_{r_{sc}} + \beta * M_{r_{sc}}) \circ M_m \tag{10}$$

Here, β=1, symbol denotes Hadamard product.

The main problem in existing system is microblogging text are noisy and short, there exist two kinds of problems. First problem is due to the large-scale nature of the vocabulary and dataset, this contributes to a high-dimension feature space. Second problem is the noisy and short text of data make data characterization extremely sparse. To avoid this problem sparse reconstruction is done as that change date robust to noise in terms of feature as discussed in Section 3.3.

The basic principle is to make two microblogs as identical as possible if they are posted by the similar users or two users similar to each other or same users based on sentiment consistency and emotional contagion to combine sentiment relation among microblogs in sentimental classification. This condition can be minimized by using (11).

$$= \min_C \sum_{l=1}^{c} \hat{Y}_{*l}^{T} (D^{M_t}{}_m - M_r) \hat{B}_{*l} \tag{11}$$

If only SS and user context is used then, $M_r = M_{r_1}$, $M_r = M_{r_2}$ for subject or topic content. Hence, final model combines social content and text information by using (12).

$$f(C; L; B) = \min_C \frac{1}{2} \| LC - B \|^2_{R_m} + \frac{\delta}{2} tri(C^{M_t} L^{M_t} L_m LC) + \beta \| C \|_1 \tag{12}$$

Where, δ indicates social content weight, β is regularization weight.

It is observed that (12) leads to non-smoothing optimization problem. It is reduced by convex smooth reformulation. To

solve this (12) is reformulated as shown in (13) which is a "constrained convex smooth optimization" problem.

$$\min_{C\in Z} L_m(C;L;B) = \frac{1}{2}\| LC-B \|^2_f + \frac{\delta}{r} tri(C^{M_t} L^{M_t} L_m LC) \quad (13)$$

Where, $z=\{C \| C \|_1 \le z\}$, differentiable part is $L_m(C;L;B)$, z -indicates non-differentiable part. L_{m_1} ball radius is z>=0, and one-to-one correspondence is present among z and β. For linear function $L_m(C;L;B)$, The smooth optimization problem is equivalently reformulated as proximal regularization [31] at C_x defines as $C_{x+1} = \arg\min_C G_{\gamma x C x}(C)$.

$$G_{\gamma x C x}(C) = L_m(C_x;L;B) + < \nabla L_m(C_x;L;B), C-C_x > + \frac{\gamma_x}{2}\| C-C_x \|^2_{R_m} \quad (14)$$

Algorithm1: Proposed Sentiment analysis using SS.

Proposed Sentiment analysis using SS **(PSASS)**

Input: L, B, C, δ, μ

Output: C

1. Randomly initialize C0
2. assign Ω=0,1, C1=C0, x=1
3. while till not convergence do
4. Calculate
5. Calculate
6. While condition is true do
7. Calculate
8. Calculate considering (16)
9. If then
10. Assign
11. Break
12. End if
13. Assign
14. End while
15. If x> max_iteration then
16. return
17. End if
18. Assign
19. Assign x=x+1
20. End while

Where, γx denotes size of step in x iteration. Therefore, the gradient of Laplacian matrix $L_m(C;L;B)$ respect to C can be evaluated using (15).

$$\nabla L_m(C;L;B) = L^x(LC-B) + \delta L^x L_m LC) \quad (15)$$

In considering β, Z constraint from (13) and (x+1)-th C can be calculated using (16).

$$(C_{x+1})_{j^*} = \begin{cases} \left(1 - \frac{\mu}{\lambda_x \| (U_m)_{j^*} \|}\right)(U_m)_{j^*}, if \| (U_m)_{j^*} \| \ge \frac{\mu}{\gamma_x} \\ 0, otherwise \end{cases} \quad (16)$$

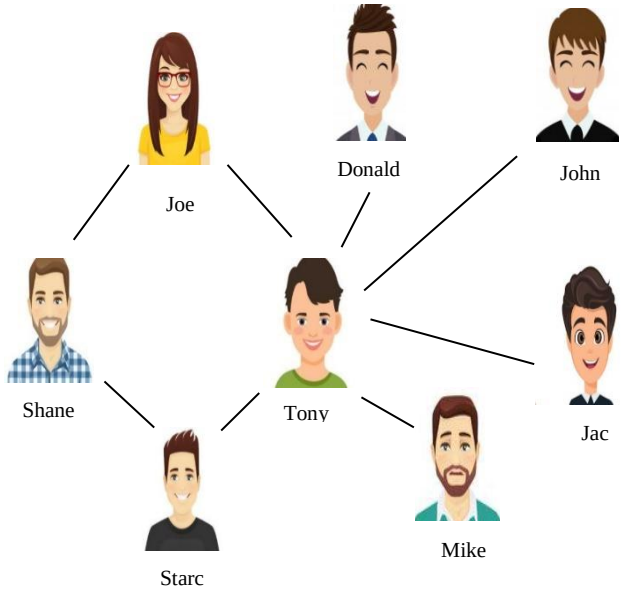

Fig. 6. Similarity Expression2.

Here, $U_{m_t} = C_x - \frac{1}{\gamma_x}\nabla L_m(C_x;L;B)$ For best convergence, optimization problem can be further accelerated and smoothened. Sequence C_x and V_x are utilized in the algorithm, C_x is sequence of estimated solution, V_x is combination affine of C_x and C_{x-1} is search point sequence. The final combination can be calculated by using (17).

$$V_x = C_x + \sigma_x(C_x - C_{x-1}) \quad (17)$$

Where, σ is the grouping coefficient. C_{x-1} suitable solution is calculated as "gradient" of V_x step through $G_{\gamma i V x}$. Finally, the algorithm for optimization is discussed as follows.

IV. SS AND SENTIMENT CORRELATION

The relation between sentiment label and friend's context are verified in [6,7]. A statistical analysis is done by illustrating how Sentiment labels in microblog and SS correlate. Consider G as an undirected graph where G=(V,E) is used to construct relation on two microblogs. Edge consisting of similar sentiment label is calculated by (18).

$$E = \frac{\sum_{i=1}^{x}\sum_{j=1}^{x} 1(B_{i^*} = B_{j^*}, e_{ij} \in E)}{\sum_{i=1}^{x}\sum_{j=1}^{x} 1(e_{ij} \in E)} \quad (18)$$

Where 1 is a function named indicator. The same calculation can be done in a weighted matrix using (19), the weights are regarded based on sentiment label. The same equation can be used for evaluating correlation among sentiment label in microblog and text similarity. In (19) I represent index, K indicates the weight of matrix G.

$$I = \frac{\sum_{i=1}^{x}\sum_{j=1}^{x} 1(B_{i*} = B_{j*}, e_{ij} \in E).K_{ij}}{\sum_{i=1}^{x}\sum_{j=1}^{x} 1(e_{ij} \in E).K_{ij}} \quad (19)$$

SS indicates two microblogging graphs built by SS, SS-topic represents two microblogging graph constructed SS and topic context. It is observed that ratio of SS and SS-Topic is greater between dataset1 and dataset2 as shown in Fig. 7. This possesses a positive relation between sentiment label and SS that helps to cover the study for exploring SS into microblogging sentimental analysis. Ratio of SS-topic method is said to be greater than SS model, it is because of homophily on similar topic and user may tend to have "same opinion on same topic".

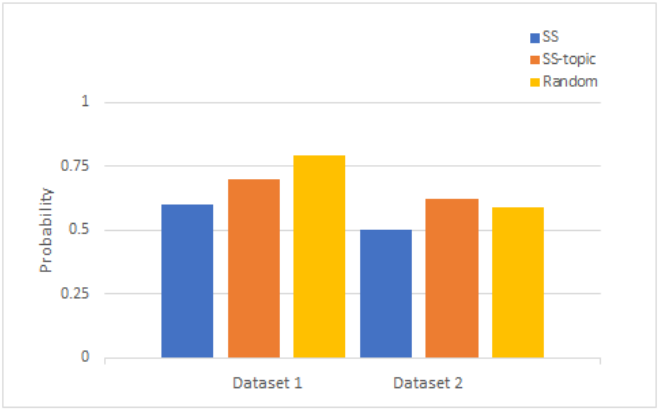

Fig. 7. SS-Conditioned Sentimental Polarity.

Thus, by including topic or subject content can explore heterogeneous relationship among microblogs.

V. Discussions

A. Advantage of Social Context

It is used to check the lifetime of various context whether they can improve in sentimental classification with respect to accuracy. 80% microblogging information are used to train the model. Here, "TC" denotes Text context, "UC" indicates user context and text. Likewise, "SSC" indicates text and SS context, "FC" implies Friends context and text. Accuracy is calculated as shown in Table III for the above-mentioned cases used by a metric, $accuracy = \frac{(TP+TN)}{(n)}$ where, n indicates number of negative and positive samples in the training dataset. TP is true positive, and TN is true negative labeled classes.

TABLE III. Context Performance

	FC	TC	UC	SSC
Dataset1	0.771	0.660	0.768	**0.799**
Dataset2	0.769	0.657	0.778	**0.788**

The following statements are determined from Table III.

- Using the "Social context", sentimental analysis on dataset1 and dataset2 can be improved in performance. Methods tested on Social content has better accuracy compared to text, it validates the utility of the context of the user, the context of friends and the context of structure similarity. The output of the social context illustrates that, in the microblogging network, emotional contagion and sentiment consistency (two theories) hold true.
- The user context performance is less compared to social contexts. This is mostly due to average friend's relationship usage post is more than the single user average post, contributing to more sentiment sparse consistency matrix. Example, according to Table I, every user in dataset1 has 32.21 average tweets with average friends 241.5.
- Approaches using SS context obtains best results in all context related to social text. SS the reason behind its better results than other context, it can get more data from friend's relationship than direct user relationship and weight on user who has more influence.

Structure similarity, which is the reason behind its better results than others, can get more data than direct relationships such as common friends and weights on whose effect is greater on users.

B. Performance Analysis and Comparison

Random sampling approach is used to check the accuracy of various methods with change in training dataset size. Methods are as follows:

1) Lasso: it is "Least absolute shrinkable and selection operator" is one of the regression analyses models which works on regularization and selection to improve accuracy of prediction it produces.

2) Least Square (LS): it is also one kind of approach in regression analysis, it is a statistical process for best fit to some set of information to be noticed. It is also used to predict related variable behaviour.

3) Support Vector Machine (SVM): it is one of a supervised ML model. They can categorize new text from the labelled training dataset for every group.

4) Naive bayes: it is one of a supervised ML model. The classifier of Naive Bayes assumes that the existence of a certain feature in a class is not linked to the presence of any other feature.

5) Logistic Regression (LR): it is also a statistical model used to design certain classes based on probability. It evaluates a dataset in which an outcome is calculated by individual variables.

6) SANT: Sentiment Analysis for Norwegian Text method proposed by [6] that combines two geological theories.

7) SMSC: Structured Microblog Sentiment Classification. All friends and user context are considered with equal priority.

8) SASS: Sentiment analysis based on SS is the proposed model to evaluate sentiment by using user context and SS. In this model, two initial positive parameters δ, μ are used. Assigning δ=0.0005 and μ=1 set by 'cross-validation'. Parameter μ denotes sparse regularization, parameter δ governs the information of social context contribution. For experimentation, from original or actual data testing dataset and training dataset are selected randomly. Percentage of the training dataset set is expressed by percentage and the remaining data for testing purpose. All the model's results are compared, and the observation are as follows:

- Models that use texts alone produced less result than methods using social context. Considered two samples and one tail T experiments are performed, and the results indicates that social context technique can boost the accuracy of sentimental classification with a significant level 0.01. In microblogging site, text data is very noisy, cynicism, and sarcasm are often used to convey user's negative emotions. Techniques like LS, NB, SVM and LR may not manage this scenario, by applying social context this problem can extended to some degree as the techniques take microblogs into account that are linked to perform a better result.
- SASS outperform SMSC and SANT and achieved better result on dataset1 and dataset2 with difference in size of training data significantly and consistently.

Compared with all models SASS has better performance with all different size of dataset with an accuracy of 0.799 for dataset1 and 0.792 for dataset2 as shown in Table IV. The performance is with respect to both dataset1 and dataset2 SANT and SMSC models had used friends and user context. But, in proposed model using SS it can explore relationship between microblogs intensely by using potential relationship between friends; every microblog possesses different impact to the suicidal sentiment compared to other microblogs but, in SMSC and SANT model all sentiment from microblogs has similar contribution to another microblogs.

SASS model is susceptible to change in size of the data for training. This shows that it is significant that "lot of labelling cost can be reduced" in spite labelling all training dataset manually.

9) Advantage of topic context: Topic context is introduced in proposed SASS model and SASS is compared with topic or subject context (SASS-T) varying with training dataset size from 60% to 90%. The result of classification is shown in Table V, from result it is observed that after incorporating topic context there is increase in accuracy of sentimental analysis in microblog compared to SASS model. Finally, SASS-T had produced better results 0.821 for dataset1 and 0.834 for dataset2 compared to all other models. The results signify the positive impact of applying topic or subject context in microblogging sentimental analysis to design the semantic relation among microblogs. The reason behind incorporating topic context is "the views of same person and similar kind of users on same topic usually remains consistent with each other that may help in prediction of suicide".

10)Parameter testing: Impact of two parameters δ, μ are tested for selection. These two parameters play a major role in managing the contribution of proposed model that might gain from SASS-T regularization constraints. The value of δ is assigned in some range {0,0.01,0.1,1,10 and 100} to study the impact on prediction accuracy. When SASS-T achieved better accuracy, the value of μ and δ are not same. So, the value of μ vary from {0, 1e-4, le-3,0.01,0.1,1} to study the impact of μ. When δ=μ=0, it reduces to SANT model, when δ>0, μ=0, it is SMSC model and when δ=0.005, μ>0, it is SASS model. So only suitable value of δ, μ can lead to a major improvement. Thus, proposed model has obtained better accuracy when δ=10, μ=0.1.

TABLE IV. PROPOSED AND BASELINE MODELS COMPARISION

	Training	**LS**	**LASSO**	**NB**	**LR**	**SVM**	**SANT**	**SMSC**	**SASS**
Dataset1 Without Topic content	60%	*0.644*	*0.693*	*0.759*	*0.724*	*0.718*	*0.770*	*0.762*	*0.776*
	70%	*0.622*	*0.665*	*0.768*	*0.731*	*0.725*	*0.763*	*0.771*	*0.772*
	80%	*0.612*	*0.695*	*0.754*	*0.733*	*0.731*	*0.759*	*0.767*	*0.779*
	90%	*0.612*	*0.721*	*0.745*	*0.729*	*0.745*	*0.769*	*0.761*	***0.799***
Dataset2 Without Topic content	60%	*0.653*	*0.677*	*0.702*	*0.718*	*0.709*	*0.701*	*0.723*	*0.733*
	70%	*0.659*	*0.692*	*0.705*	*0.716*	*0.718*	*0.717*	*0.722*	*0.744*
	80%	*0.662*	*0.671*	*0.691*	*0.709*	*0.723*	*0.730*	*0.739*	*0.774*
	90%	*0.645*	*0.740*	*0.681*	*0.722*	*0.731*	*0.749*	*0.755*	***0.792***

TABLE V. CLASSIFICATION-ACCURACY

	Training	**SASS**	**SASS-T**
Dataset1 With Topic content	60%	0.768	0.792
	70%	0.784	0.769
	80%	0.781	0.789
	90%	0.792	**0.821**
Dataset2 With Topic content	60%	0.736	0.754
	70%	0.750	0.756
	80%	0.771	0.772
	90%	0.792	**0.834**

VI. Conclusion

In this paper, a new technique is proposed to identify and facilitate sentiment classification as inspired by emotional contagion and sentiment consistency. Three types of context are considered in the proposed model: structure similarity, user context, and topic or subject context. structure similarity matrix and topic or subject context matrix are constructed, these contexts are added to the model using Laplacian matrix build by the contexts. experimental analysis showed that SS context produced better result compare to direct user relation. Thus, by incorporating topical context in SASS model aided in improving the outcome of sentimental classification compared to all other models with an accuracy 0.821 for dataset1 and 0.834 for dataset2. This result can be useful for suicide prediction among users based on the emotion of tweets posted by users that may help individual from attempting from suicide by informing to any NGO's.

VII. Future Work

The experiment is done on two different datasets with respect to topics and keywords with prediction results, further research is required based on different suicide risk factors that can be used for suicidal prediction and analysing the timeline tweets by examining retweets exhibiting suicidal contents with friends, followers, and tweeters.

Reference

[1] J, Mao H, Zeng X, Twitter mood predicts the stock market, Journal of Computational Science, vol.2, pp. 1-8, 2011.

[2] Cambria E, Mar, Affective computing and sentiment analysis. IEEE Intelligent Systems, vol.2, pp.102-107,2016.

[3] Cambria E, Schuller B, Xia Y, White B, New avenues in knowledge bases for natural language processing. Knowledge Based System, 108, pp.1-4, 2016.

[4] Fuji Ren, Ye Wu, IEEE transaction on affective computing, Predicting user opinions in Twitter with social and Topic content, vol.4, pp.412-424, 2013.

[5] Mei Q, Ling X, Wondra M, Su H, Zhai C, Topic sentiment mixture: modeling facets and opinions in weblogs, In Proceedings of the 16th international conference on World Wide Web-ACM, pp. 171-180, 2007.

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

[7] Tan C, Lee L, Tang J, Jiang L, Zhou M, Li P, User-level sentiment analysis incorporating social networks, In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, pp. 1397-1405, 2011.

[8] Hatfield E, Cacioppo JT, Rapson RL, Emotional contagion. Cambridge university press,1994.

[9] Abelson RP, Whatever became of consistency theory? Personality and Social Psychology Bulletin, 1983.

[10] Tang J, Hu X, Gao H, Liu H, Exploiting local and global social context for recommendation, In International Joint Conference on Artificial Intelligence, pp. 2712-2718, 2013.

[11] Tang J, Wang S, Hu X, Yin D, Bi Y, Chang Y, Liu H, Recommendation with social dimensions, The Thirtieth AAAI Conference on Artificial Intelligence. AAAI Press, pp. 251-257,2016.

[12] Mcpherson Miller and Smithlovin Lynn and Cook James M, BIRDS OF A FEATHER: Homophily in Social Networks. Review of Sociology vol.27(1), pp.415-444, 2001.

[13] Crimaldi Irene and Vicario Michela Del and Morrison Greg and Quattrociocchi Walter and Riccaboni Massimo, Homophily and Triadic Closure in Evolving Social Networks, arXiv: Social and Information Networks, 2015.

[14] Thelwall Mike, Emotion Homophily in Social Network Site Messages. First Monday, vol.15(4), 2010.

[15] Liang Y, Li Q, Incorporating interest preference and social proximity into collaborative filtering for folk recommendation, In Workshop on Social Web Search and Mining, 2011.

[16] Xie Yan Bo and Zhou Tao and Wang Bing Hong, Scale-free networks without growth, Physica A Statistical Mechanics & Its Applications, vol.387(7), pp.1683-1688, 2008.

[17] Liu J, Ji S, Ye J, Multi-task feature learning via efficient l 2, 1 -norm minimization. Conference on Uncertainty in Artificial Intelligence, AUAI Press, pp. 339-348, 2009.

[18] Ortigosa-HernaÂndez J, RodrõÂguez JD, Alzate L, Lucania M, Inza I, Lozano JA, Approaching sentiment analysis by using semi-supervised learning of multi-dimensional classifiers. Neurocomputing 92, pp. 98-115, 2012.

[19] Wang Y, Huang M, Zhu X, Zhao L, Attention-based LSTM for aspect-level sentiment classification. Conference on Empirical Methods in Natural Language Processing. pp. 606-615, 2016.

[20] Pandarachalil R, Sendhilkumar S, Mahalakshmi G, Twitter sentiment analysis for large-scale data: an unsupervised approach, Cognitive Computation, vol. 7(2), pp. 254-262, 2015.

[21] Cheng C-H, Chen H-H, Sentimental text mining based on an additional features method for text classification. PLoS ONE, vol.14(6): e0217591, 2019.

[22] Cui A, Zhang M, Liu Y, Ma S, Emotion tokens: Bridging the gap among multilingual twitter sentiment analysis, In Asia Information Retrieval Symposium. Springer, pp. 238-249, 2011.

[23] Kiritchenko S, Zhu X, Mohammad SM, Sentiment analysis of short informal texts, Journal of Artificial Intelligence Research, vol.50, pp. 723-762, 2014.

[24] Ren F, Wu Y, Predicting user-topic opinions in twitter with social and topical context, IEEE Transactions on Affective Computing, vol.4(4), pp.412-424, 2013.

[25] Speriosu M, Sudan N, Upadhyay S, Baldridge J, Twitter polarity classification with label propagation over lexical links and the follower graph, In Proceedings of the First workshop on Unsupervised Learning in NLP, Association for Computational Linguistics, pp. 53-63, 2011.

[26] Lu T-J, Semi-supervised microblog sentiment analysis using social relation and text similarity, In International Conference on Big Data and Smart Computing, IEEE, pp.194-201, 2015.

[27] Wu F, Huang Y, Song Y, Structured microblog sentiment classification via social context regularization, Neurocomputing 175, pp.599-609, 2016.

[28] Vosecky J, Leung KW, Ng W, Collaborative personalized Twitter search with topic-language models, international ACM SIGIR conference on research and development in information retrieval, pp. 53-62, 2014.

[29] Easley D, Kleinberg J, Networks, crowds, and markets: Reasoning about a highly connected world. Cambridge University Press, 2010.

[30] Jackson MO, Rogers BW, Meeting Strangers and Friends of Friends: How Random Are Social Networks? American Economic Review, vol.97(3), pp.890-915, 2007.

[31] Kernighan BW, Lin S, An efficient heuristic procedure for partitioning graphs, The Bell System Technical Journal, vol.49(2), pp.291-307, 1970.

[32] E. Rajesh Kumar, K.V.S.N. Rama Rao, Soumya Ranjan Nayak & Ramesh Chandra, Suicidal ideation prediction in twitter data using machine learning techniques, Journal of Interdisciplinary Mathematics, vol.23:1, pp.117-125, 2020.

[33] Gualtiero B. Colombo, Pete Burnap, Andrei Hodorog, Jonathan Scourfield, Analysing the connectivity and communication of suicidal users on twitter, Computer Communications, pp.1-10, 2015.