Received: May 31, 2022. Revised: August 22, 2022.

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Named Entity Emotion Intensity Tagging for Suicidal Ideation Detection From Social Media Texts During MT

K. Soumya[1]* **Vijay Kumar Garg[2]**

[1]*School of Computer Science and Engineering,*
Lovely Professional University - Punjab, and VBIT – Hyderabad, India
[2]*School of Computer Science and Engineering, Lovely Professional University - Punjab, India.*
* Corresponding author's Email: kothapallisowmya@gmail.com

Abstract: Detecting suicide ideation from social media texts is challenging as it necessitates analyzing the content and context of the text utterances. Modeling the temporal trajectory of suicide through various stages like stress, depression, thought and strengthening of the thought becomes difficult with content features alone and this needs context cues too. This work proposes a hybrid feature representation incorporating both content and context features. The richness of features is enhanced using adversarial learning. Neural machine translation is done on hybrid feature representation to provide named entities with emotional intensity tagging. The sequence of emotional intensity tagging is mapped by Bi-directional Long short term memory (LSTM) to suicidal ideation label. The proposed integration of emotion tagging to SI detection with Bi-directional LSTM provided an SI detection accuracy of 95.54% which is atleast 3% higher compared to existing works.

Keywords: Suicidal ideation, Adversarial learning, Named entity tagging, Neural machine translation.

1. Introduction

Suicidal ideation (SI) is the intensive desire for death, thoughts about dying, and conceiving plans about suicide [1]. Every year 800,000 people commit suicide worldwide [2]. The rate is high at adolescence [3]. It is difficult to detect suicidal ideation as the thoughts are not expressed often. With rapid availability of internet to more than half of world population, social media has become a way of life for most adults. Social media is being increasingly used for communications and expression oneself. People express their life events and feelings in social media. There is increasing evidence of people expressing their feelings which are higher correlated to suicidal ideation over various social media like Twitter, Reddit or Instagram [4-6]. Compared to traditional methods like clinical interviews or questionnaire analysis, social media can be used to identify SI in time. Jashinsky et al [7-8] found a strong correlation between suicide rates and keywords/phrases extracted from Twitter about suicide risk factors. Thus social media text analysis looks a promising approach for SI detection. Emotion, thoughts and intentions expressed on social media are valid indicators of SI [9]. Most of the current approaches for SI detection either use human coders or automated systems to analyze the words and phrases which are semantically and lexically consistent with SI. Automated approaches for SI detection can be categorized to: content analysis, feature engineering and deep learning approaches [10]. Content analysis approaches do exploratory data analysis on user generated content like lexicon based filtering, statistical linguistic features and topic modeling. Feature engineering involves extraction of N-gram features, knowledge based features, syntactic and context features etc from social media posts. Using these features, machine learning classifiers like support vector machine (SVM) [11], Artificial neural network (ANN)[12] , conditional random fields [13] etc are trained to classify SI. Deep learning classification is being successfully used in many applications like computer vision, natural language processing and medical diagnosis. Deep learning classifiers can learn text features automatically far better compared to feature engineering approaches. Natural language texts are converted to distributed vector space using word embedding techniques like word2vec [14] and Glove [15]. The sequences of work embedding are classified to SI using deep learning models like Convolutional neural network (CNN), RNN (Recurrent neural network) or LSTM. The current deep learning models have two important issues: lack of sufficient labeled dataset for SI detection and not consideration of context information in classifying SI. Addressing these two issues, this work proposes a hybrid feature representation incorporating content and context features with feature adversarial learning for compensating lack of sufficient dataset for SI. The hybrid feature representation is converted to named entity emotional intensity tagged representation using neural machine translation guided by attention vector. The sequences of emotion labels in emotion intensity tag are classified by a Bi-directional LSTM classifier to detect SI. Following are the novel contributions of this work.

(i) A hybrid feature representation incorporating both content and context features with feature richness enhanced through adversarial learning

(ii) Neural machine translation (NMT) with attention vector for translating hybrid features to named entity emotional intensity tagging.

(ii) Temporal modeling of SI using emotional intensity tagging information with a Bi-directional LSTM

The rest of the paper is organized as below. Section 2 provides the survey on SI detection approaches and their issues. Section 3 presents the proposed solution for SI detection. Section 4 presents the results of the proposed solution and comparison to most recent works. Section 5 presents the conclusion and scope of future research.

2. Related work

H Ji et al [16] extracted statistical, syntactic, linguistic, word embedding, and topic features from twitter dataset and trained supervised machine learning classifier to detect SI. The results are snapshot based and method has higher false positives due to non consideration for context information. Coppersmith et al [17] extracted character n gram features from the tweets of people who attempted suicide and normal. Using character n gram a character language model (CLM) score was calculated. Authors found a significant difference between the CLM scores of suicide attempted people and normal people. However, the method is very coarse and did not consider the context in which suicide related words are uttered. This results in higher false positives. Mulholland et al [18] extracted vocabulary features, syntactic features, semantic class features and N gram features from text and classified using Simplecart algorithm to classify between two classes or suicide or no suicide. Features were snapshot based without consideration for temporal correlation. Huang et al[19] combined machine learning with psychological knowledge to detect suicidal posts in Chinese micro blogs. Linguistic features learnt from psychological lexicon dictionary are extracted and classified using SVM classifier to detect suicidal posts. The work was specific to Chinese language lexicons. Liakata et al [20] extracted sentence based features like n-gram, negative expressions, self directed emotions etc and classified the features using SVM classifier to detect nine emotion categories of anger, sorrow, hopefulness, happiness, peacefulness, fear, pride, abuse and forgiveness spread across suicide notes. From this spread of different emotions, the posts are classified to suicide / non suicide posts. Pestian et al [21] explored the roles of computational algorithm in extraction of suicide related thoughts from suicidal notes. By applying natural language processing author classified between genuine and elicited suicide notes. Exploring different feature extractions methods like bag of words, latent semantic analysis and heterogeneous selection, authors found heterogeneous selection yielding better results. Braithwaite et al [22] extracted three features of global high level properties of text(4 variables), relative use of word categories (71 variables), relative use of language markers(17). The count of each variable is divided by total number of words in the tweet. This feature vector of 92 variables is then classified using decision tree classifier to two target labels of suicidal and non suicidal posts. Authors considered only content features. Nobles et al [23] extracted psycholinguistic features and term frequency-inverse document frequency (TF-IDF) from the short messages and classified using dense neural network to two target class of suicide and non suicide messages. The result is snapshot based and it did not correlate between the messages of same person. Coppersmith et al [24] converted text to Glove embeddings and used a sequence of embeddings as input to LSTM model to classify SI in texts. The attention mechanism is used to apply weights to the timesteps of the sequence such that the most informative subsequences are more strongly considered in the final prediction. But the method to find information subsequences is based only on

Received: May 31, 2022. Revised: August 22, 2022.

Table 1. Survey summary

Author	**Solution**	**Gap**	**How Proposed solution address the problem**
Ji et al [16]	Extracted statistical, syntactic, linguistic, word embedding, and topic features from twitter dataset and trained supervised machine learning classifier to detect SI	Snapshot based with consideration for context	Bi-directional LSTM is used in proposed work to consider context
Coppersmith et al [17]	Character n gram based score	Score is based on individual tweet without accommodating the emotion context.	Emotion context is considered in the proposed work
Mulholland et al [18]	Authors extracted vocabulary features, syntactic features, semantic class features and N gram features from text and classified using Simplecart algorithm	Features were snapshot based without consideration for temporal correlation	Temporal correlation is realized using Bi-directional LSTM
Huang et al[19]	Linguistic features learnt from psychological lexicon dictionary are extracted and classified using SVM classifier to detect suicidal posts	Context spread across multiple posts	Proposed solution use short term context information spread across posts.
Liakata et al [20]	Authors extracted sentence based features like n-gram, negative expressions, self directed emotions etc and classified the features using SVM classifier	Temporal correlation is not considered	Temporal correlation is realized using Bi-directional LSTM
Braithwaite et al [22]	Authors extracted three features of global high level properties of text(4 variables), relative use of word categories (71 variables), relative use of language markers(17).	Based only on statistical distribution of words and does not consider emotional context	Emotional context is considered in proposed solution
Nobles et al [23]	Authors extracted psycholinguistic features and term frequency-inverse document frequency (TF-IDF) from the short messages and classified using dense neural network	Based only on statistical distribution of words and does not consider emotional context	Emotional context is considered in proposed solution
Coppersmith et al [24]	Authors converted text to Glove embeddings and used a sequence of embeddings as input to LSTM model to classify SI in texts.	Based only on content and does not consider context	Proposed solution consider both content and context
Sawhney et al [25]	Authors extracted three features of character n grams, TF-IDF and Bag of words. These features are used in combination with	Does not consider temporal correlation	Bi-directional LSTM is used for temporal correlation in proposed solution

	RNN, LSTM and C-LSTM classifiers to classify SI.		
Tadesse et al [26]	Authors combined LSTM with CNN classifier to classify text to SI	Snapshot based decision without considering temporal context	Temporal context is considered in proposed solution
Ji et al [27]	relational encoding model combining suicidal ideation with sentimental indicators	Sarcastic comments are misinterpreted.	Sarcasm is detected and correct polarity is given in proposed solution
Matero et al [28]	Dual modality approach using both traditional ML models and RNN model	User personality factors input needed for SI detection	The proposed solution does not depend on user personality features
Chen et al [29]	Authors proposed a hybrid approach combining behavioral and language model for detecting SI	Snapshot based with considering context	Context information is considered in proposed solution

content and not on context. Sawhney et al [25] extracted three features of character n grams, TF-IDF and Bag of words. These features are used in combination with RNN, LSTM and C-LSTM classifiers to classify SI. But the approaches could not achieve higher performance gains due to lack of feature selection methods. Tadesse et al [26] combined LSTM with CNN classifier to classify text to SI. Text are converted to word embedding vector and passed to LSTM. The output of LSTM is fed to CNN classifier. Since the CNN decision is based on sequence of sentences, the accuracy is better than snapshot based decision. But the model considers only content features due to which the accuracy is limited. Ji et al [27] proposed a relational encoding model combining suicidal ideation with sentimental indicators and life event related topical indicators and classified SI with relational encoded features. Sarcastic sentiments can cause error propagation and affects the SI detection accuracy. Matero et al [28] proposed a dual modality approach using both traditional ML models and RNN model. Both models were trained using word embeddings. In addition to classification of SI, authors incorporated user personality factors to classify the SI risk levels into low, medium and high. Chen et al [29] proposed a hybrid approach combining behavioral and language model for detecting SI. Behavioral model is built taking sentiment, content and posting behavior as input. SVM classifier is built with behavioral input to detect SI. Language model is built to classify the user posts to risk level. Combining both the results, suicide risk level is detected. Results are snap shot based without temporal consideration. Wicentowski et al [30] proposed a ensemble of supervised maximum entropy classifiers to find emotions expressed in suicide notes. Lexical and syntactic features were extracted from suicide notes. For each emotion (15 emotions) a maximum entropy classifier is trained to classify that particular emotion. The approach identified the emotions spread in the suicide notes but it did not classify the SI based on the spread of emotions. Schoene et al [31] extracted sentiment, linguistics , word frequency features from suicide notes and classified the SI using regression tree classifier. The work classified genuine suicide notes but it could not identify SI in texts. Wang et al [32] explored different features for SI detection and found that features gathered around discussion context were found to be reliable indicator of SI. Though the work found the importance of context, it could be provide method to remove sarcastic suicide notes in contexts. Choudhury et al [33] proposed a statistical approach to derive distinct markers which provide information on shift from mental illness to SI. These markers can be used in combination with other text feature to increase the prediction accuracy of SI. Guan et al [34] used profile and linguistic features to detect SI. Social media profile features and content based linguistic features are extracted and classified using random forest classifier. This method is suitable only for initial screening as accuracy is less than 70%. Also without context features, the false positive is very high in this approach.

The summary of the survey is presented in Table 1. From the survey, most approaches don’t consider the context and mistake sarcastic or funny suicide comments as genuine. This is due more emphasis on content based features. Though many works classified the spread of emotions, they did not analyze the shift of emotions in detecting SI. From the survey, it could be observed that incorporating context features and analyzing the shift of emotions could further enhance the effectiveness of SI detection.

Received: May 31, 2022. Revised: August 22, 2022.

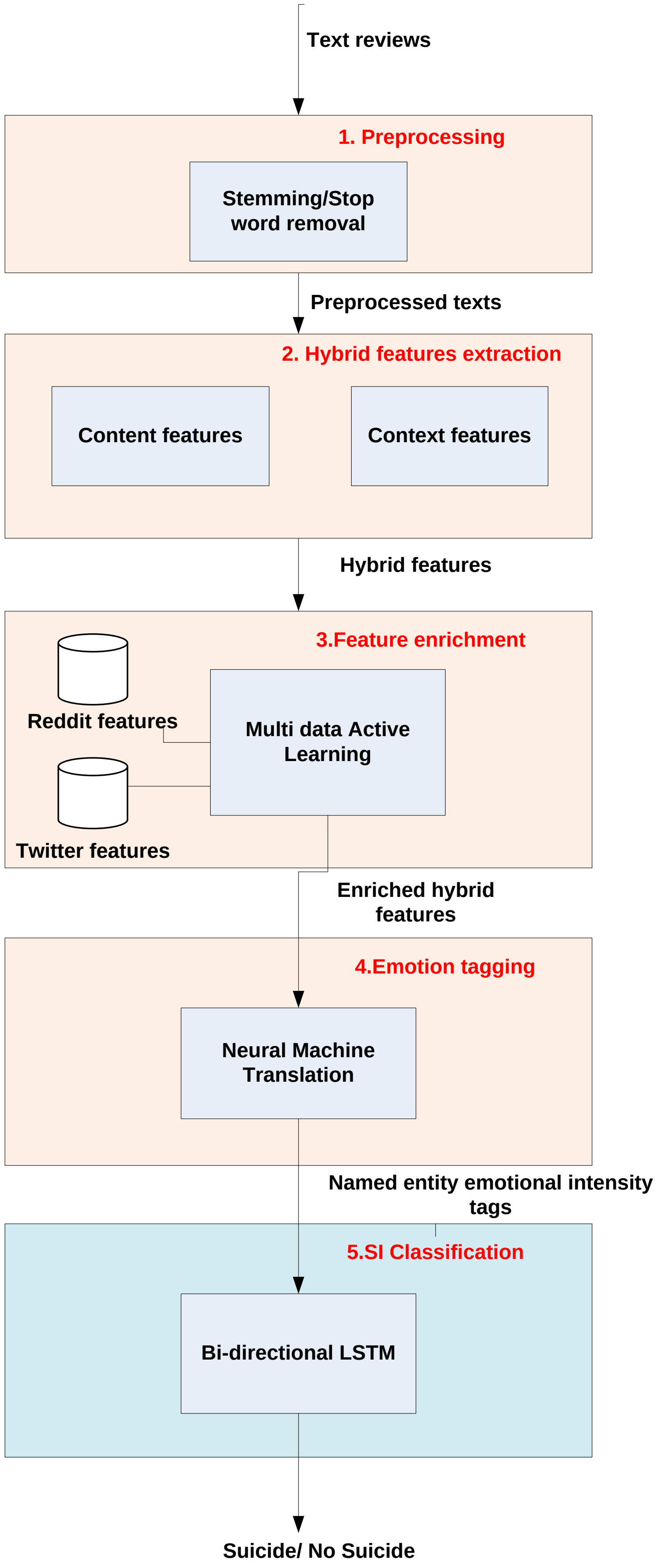

Figure. 1 Architecture of proposed solution

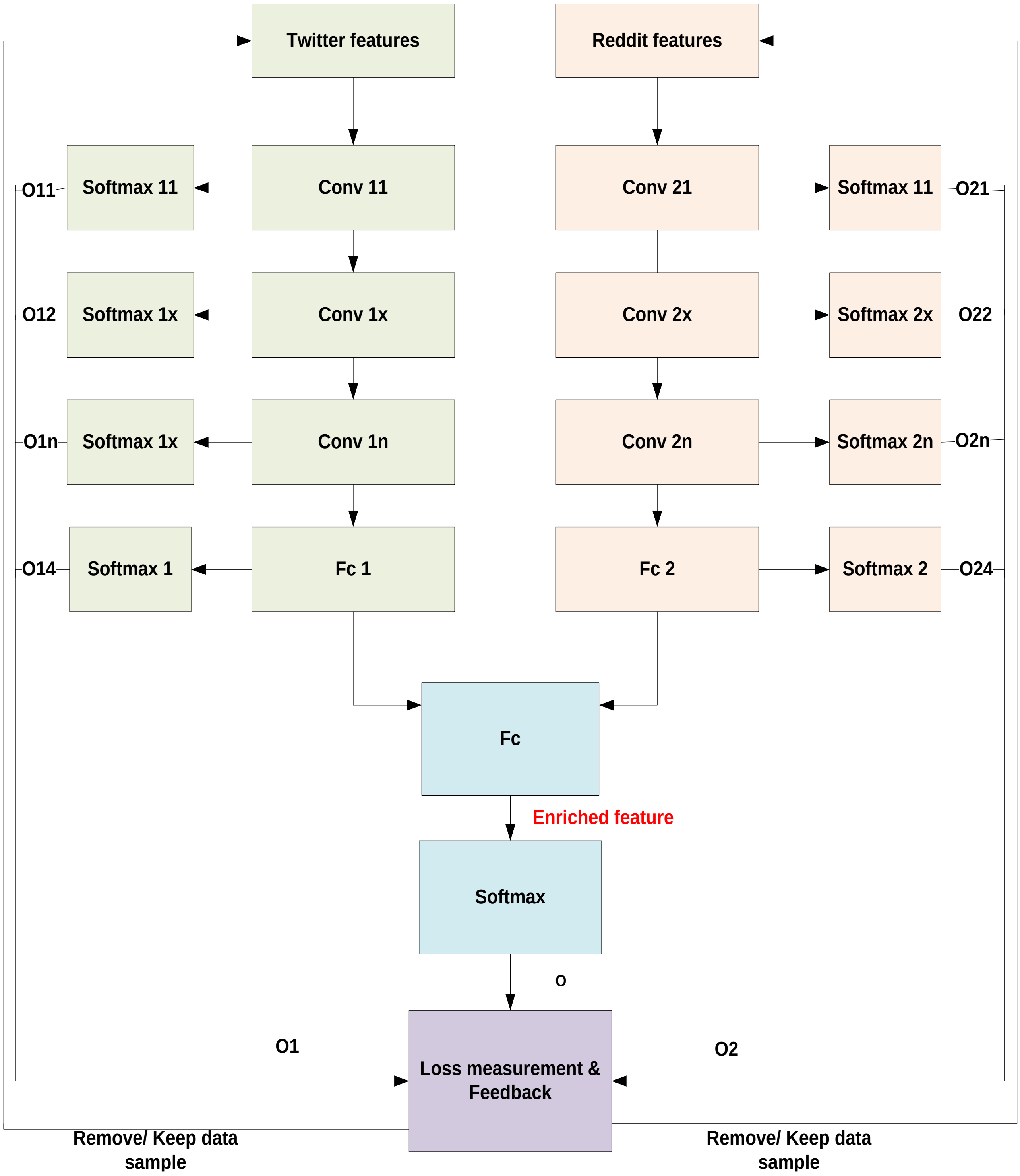

Figure. 2 Multi data active learning

3. Named entity emotional intensity tagging for SI detection

The architecture of the proposed solution is given in Fig. 1. From the tweet text, the content and context features are extracted and a hybrid feature vector is created. To hybrid feature vector representation is enriched with active learning using cross domain datasets of SI. Neural machine translation is invoked to translate the hybrid feature representation to a named entity emotion intensity tags representation. The sequence of emotion intensity tags is passed to Bi-directional LSTM to provide the classifier label: suicide or no suicide. Each of stages of the proposed solution is explained in below subsections.

Table 2. Sentiment emoticon mapping

Sentiment	**Emoticon patterns**
Positive	:-(, :(, :-|, ;-(, ;-<,|- {
Negative	-), :), :o, :-}, , ;-}, :->, ;-)
Sarcastic	(, [:, ;], -?[), p, P]

Table 3. Sentence incongruity rules

Candidate term(Verb positive/negative)	**Positive/Negative patterns**
Verb	Verb followed by Verb
Verb present participle	Verb followed by Adverb
Verb Gerund	Adverb followed by Verb
Verb past participle	Verb followed by a proposition
Verb past form	Verb followed by adjective
Verb present participle third person singular	Verb followed by noun

A. *Preprocessing*

The raw review texts are preprocessed before passing to the feature extraction stage. The review comments with very few words are filtered out. The remaining reviews are tokenized to words and space characters are removed. From the tokenized words, the most common stop words in English are removed. The remaining words are then stemmed to standard form using the Porter stemmer algorithm.

B. *Feature extraction*

Content and context features are extracted from the preprocessed texts.

Content features

Emoticon features and word embedding features are the content features used in this work. Emoticons visually represent the facial expression. Emoticons have high correlation to uttering sarcastic statements [16]. A smiley emoticon with negative situation word is an example of sarcastic utterances.

The emoticons patterns considered in this work are given Table 2.

The reviews are processed to look up for the emoticon patterns and the frequency of the emotion patterns. The feature vector is constructed for emoticon patterns in form of

$$E = < f_p, f_n, f_s > \qquad (1)$$

Where $f_p$the frequency of positive emoticons is, f_p is the frequency of negative emoticons and f_s is the frequency of sarcastic emoticons.

Glove word embedding features [35] are extracted from the preprocessed reviews after removing all the emoticons. Glove is a powerful word embedding algorithm. It learns the words vector representation by doing dimensionality reduction on the co-occurrence matrix. Glove is an unsupervised technique to obtain meaningful vector for each individual word in the corpus. The vector for the word is constructed in such a way that similar words cluster together and different words repel against each other. Glove embedding is selected in this work due to it ability to capture both local statistics and global statistics to obtain word vectors compared to other word embedding models like word2vec.

Context features

Sentiment information and sentiment contradiction are two context features extracted from the texts. Sentiment information is the measure of sentiment expressed in texts. It has two components intrinsic and extrinsic. Intrinsic refers to the person's internal sentimental state. Extrinsic refers to event topic which acts as risk indicator. Intrinsic sentiment information is extracted for text using domain specific sentiment lexicon approach proposed in [36]. The output of intrinsic sentiment information is a work vector with polarity score(+1/-1) for each of domain specific word in the text. Extrinsic sentiment information is extracted using latent dirichlet allocation (LDA) topic modeling approach proposed in [37]. The output of topic modeling is a vector of scores with each element representing the topic score.

Text like "I am so happy today that I don't want to live anymore" is contradictory statement and it should not be falsely judged as SI. These kind of sentiment contradictions can be found by looking for signals of sentence incongruity. Sentence incongruity is the concept of polarity contrast between the positive candidate term and negative phrase or negative candidate term with positive phrase. The order of occurrence is not important. Camp [38] detailed the incongruity patterns in English language sentences and the summary is presented in Table 3

The sentences are POS tagged and count of number of patterns as defined in Table 2 are found and given as sentence incongruity (si) feature.

A hybrid feature representation is formed combining the content and context features as a one dimensional feature vector.

C. *Feature enrichment*

The hybrid features can be enriched by learning

Table 4. Emotion tagging

Text	Emotion tagging
I have to look at life in her perspective, and it would break anyone’s heart	Break anyone’s heart **<sadness, high>**
I hate it when certain people always seem to be better at me in everything they do	Hate it **<disgust, low>**
I felt bored and wanted to leave at intermission, but my wife was really enjoying it, so we stayed	Felt bored **<sadness,low>** enjoying it **<happiness,medium>**

the representation from more datasets. This work proposes multi data active learning to learn a hybrid representation. On each dataset, a convolutional neural network (CNN) is trained and the output of last fully connected layer are merged to get final feature representation. The architecture of multi data learning is given in Fig. 2. It is shown for two datasets but the same concept can be extended for multiple datasets. The output of each convolution layer and FC layer are passed to softmax classifier to get the output. In addition the output fused feature is also passed to softmax classifier to get the output. For a same sample, the outputs of the convolution ($O_{11} \dots . O_{1n}, O_{21}, \dots O_{2n,}, O_1, O_2, O$) must be consistent. Based on this criteria, the samples with minimum loss to consistency of the outputs is selected and samples whose loss is higher are rejected. The loss is measured as

$$loss = \frac{O_{1n}}{O_{1n}+O_{2n}} E_1 + \frac{O_{2n}}{O_{1n}+O_{2n}} E_2 + E \qquad (2)$$

Where

$$E_x = \sum_{j=1}^{n-1} O_{xj} . e_{xj} \qquad (3)$$

e_{xj} is the entropy of output skipping the final output and E is the entropy of the final output.

$$e_{xj} = -\sum_{k=1}^{m} p_i^k \, log \, p_i^k \qquad (4)$$

p is the prediction label out of k candidate out labels with $k = 2$ in this work (suicide or not suicide).The loss is calculated for each sample and average loss is calculated for all samples. The samples whose loss is less then average are skipped from training set and the rest is used to retrain the model to get the enriched features as the output at fully connected (FC) layer fusing both the individual CNN model FC outputs. This enriched hybrid features is used for emotion tagging.

*D. **Emotion tagging***

Neural machine translation is used in this work to translate the sentence to named entity with emotion tag and emotion intensity value. Named entity is the span of sentence marked with one or more words which convey emotional content in the sentence. Emotion tag is one of following emotions of: happiness, sadness, anger, disgust, surprise and fear. Emotion intensity value is one of : low, medium and high. Some of the example of translation is given below

Neural machine translation is trained to take the text as input and provide the named entity, its corresponding emotion tag and emotion intensity as the output.

Self-attentive based neural machine translation is used in the work. It is based on encoder – decoder architecture, where source text $X$is first converted to a feature representation C by the encoder. The decoder takes$C$as input and provides the Y <named entity, emotion, emotion intensity> as the output. Given a parallel dataset of sentence pairs $D = \{(X, Y)\}$ where X is the source sentence and Y is the target sentence, the loss function is defined as

$$L(D; \theta) = \sum_{(X,Y)\epsilon D} \log p(Y|X; \theta) \qquad (5)$$

Where p represents the probability function and θ represents the weight.

For the encoder part, the CNN model encoding text to enriched hybrid features (discussed in subsection C) is used. Attention mechanism is incorporated in decoder part to select the named entity. Attention based LSTM is used in the decoder part. The dimension of features for encoding are reduced significantly due to presence of forget gates in the LSTM. This allows representing long term dependencies between words efficiently. Flexible access to context information both previous and subsequent is facilitated by integration of attention mechanism in LSTM. The highly relevant features are selected from the entire set of features by the attention layer.

NMT translation is trained with GoodNewsEveryone dataset[39]. This dataset has annotations for text in terms of emotional contents in text, emotion and emotional intensity. The trained NMT is used for converting the source texts to named entity, emotion and emotional intensity.

Table 5. Classifier parameters

Parameters	Values
Number of filters	2,4,6,8
Kernel sizes	2,3,4
Padding	'same'
Activation function	'Relu'
Pooling type	Max-pooling
Embedding dimension	300
Batch size	50
Epochs	100
Dropout	0.5
Fully connected layer	Softmax

E. SI Classification

The sequence of emotion tags of form $S = \{< e_1, I_2 >, < e_2, I_2 >, \ldots . < e_n, I_n >\}$ where e_x is the emotion tag and I_x is the intensity of the emotion tag e_x. The emotion tag and intensity is represented in form of binary vector of size $K = m + n$ where m is the number of unique emotions and n is the number of unique intensities. The sequence of emotions tags are converted to binary sequences $BS = \{ b_1, b_2, \ldots . b_n\}$. The binary sequences are provided as input to Bi-LSTM classifier to classify the binary sequences SI label: suicide or not suicide. Bi-LSTM is an extension of traditional LSTM which is an extension of recurrent neural network [22]. Temporal correlation between sequential data is learned effectively using LSTM. LSTM can decide to learn long term or forget long term information with gating mechanisms.

A combination of an input vector x and the previous hidden state is taken as input by an LSTM node.

A new candidate cell activation ˜c is calculated by the LSTM. It is calculated as the weighted sum of the inputs and bias b. The result is then passed to a hyperbolic tangent activation function as given below

$$c_t = \emptyset_t(W_c x_t + U_c h_{t-1} + b_c) \qquad (6)$$

$\mathrm{c_t}$ is the candidate cell activation. $\mathrm{x_t}$ is the input vector. W and U are the weight matrices.$\mathrm{h_{t-1}}$ is the hidden state vector at the previous time step and $\mathrm{b_c}$ is the bias.

The gates control how much of activation must be retained and how much must be forgot. Input gate control how must activation to retain and forget gate decided how much cell activation must be forgot. The final gate is incorporated to calculate the hidden state.

$$f_t = \emptyset_s(W_f x_t + U_f h_{t-1} + b_f) \qquad (7)$$

$$i_t = \emptyset_s(W_i x_t + U_i h_{t-1} + b_i) \qquad (8)$$

$$o_t = \emptyset_s(W_o x_t + U_o h_{t-1} + b_o) \qquad (9)$$

$\mathrm{f_t}$ is the forgot gate vector. $\mathrm{i_t}$ is the input gate vector. $\mathrm{o_t}$ is the output gate vector. It has two operation networks: forward and backward. The clause information between words in forward direction is learnt using forward operation network and it is learnt in backward direction by the backward operation network. The hidden states generated by these two operation network is merged to generate the hidden states of Bi-LSTM. The output vector of Bi-LSTM is calculated as

$$y_t = \sigma(\vec{h}, h^{\leftrightarrow} t) \qquad (10)$$

Where the result of forward operation network and backward operation network are given as input to function involving sum, multiplication and concatenation. In terms of vector representation, the output of Bi-LSTM is given as

$$Y_t = [y_{t-n}, \ldots \ldots y_{t-1}] \qquad (11)$$

Thus, concatenating the Bi-directional layer and LSTM layer constructs Bi-LSTM, and the LSTM results will be automatically concatenated.

The bi-directional LSTM was configured with parameters in Table 5.

4. Results

The effectiveness of proposed solution is tested by implementing the solution in Python. The performance is tested against UMD Reddit suicidality dataset [40] and Reddit SWMH dataset [37]. UMD Reddit suicidality dataset was collected from discussions on Reddit.com. The training set has posts from 620 users and test set has posts from 245 users. The posts are labelled with four labels of : no risk, low, moderate and severe. The labels "no risk" and "low" are grouped as "false" and labels "moderate" and "severe" are grouped as "true". The users are split into 80:20 ratio and posts from users in training set is used for training the classifiers. Posts from users in test set are used for testing the classifiers. Reddit SWMH Dataset has 54,412 posts which are labeled with five labels of : depression, suicide watch, anxiety, offmychest and bipolar. Labels "suicide watch" is changed as "true" and other labels are made as "false". The training set of UMD Reddit suicidality dataset and full

Received: May 31, 2022. Revised: August 22, 2022.

Table 6. Comparison of proposed solution to existing works

Solutions	Precision	Recall	Accuracy	F1-score
Proposed	95.93	95.21	94.76	95.54
Tadesse et al	94.8	90.5	91.7	92.6
Ji et al	83.81	83.85	83.85	83.77
Chen et al	89	81	86	85

Table 7. Comparison to context features

Solutions	Precision	Recall	Accuracy	F1-score
Hybrid features	95.93	95.21	94.76	95.54
Content features	94.1	90.1	91.2	91.7

Table 8. Comparison of feature enrichment

Solutions	Precision	Recall	Accuracy	F1-score
With enrichment	95.93	95.21	94.76	95.54
Without enrichment	94.1	93.1	92.2	94.7

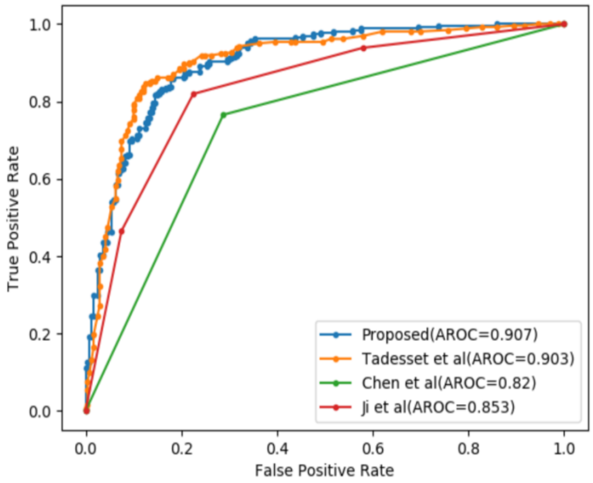

Figure. 2 Comparison of ROC

Reddit SWMH dataset are used to train the Multi data active learning component of proposed solution. The test set from umd reddit suicidality dataset is used for testing the effectiveness of the proposed solution.

The performance is compared against LSTM-CNN algorithm proposed by Tadesse et al [26], attention relation networks solution proposed by Ji et al [27] and hybrid model proposed by Chen et al [29]. The proposed solution, LSTM-CNN by Tadesse et al, relation encoding model by Ji et al, hybrid model by Chen et al are implemented in Python using Keras and tensor flow modules. The performance of all the solutions are tested against UMD Reddit suicidality test dataset. The performance is measured in terms of standard metrics: accuracy, precision, recall and F1 score.

The results of accuracy, precision, recall and F1-score across solutions are given in Table 6.

The accuracy in proposed solution is 94.76% which is 3% higher compared to Tadesse et al, 10.91% higher compared to Ji et al and 8.76% higher compared to Chen et al. The accuracy has improved in proposed solution due to involvement of context features apart from content features, feature enrichment and emotional intensity tagging. Tadesse et al though applied temporal modelling of emotions similar to proposed solution, they used only word embedding feature representation without considering the context. Ji et al used attention relation networks involving topic modelling and sentiment, however they did not consider sentence incongruence is resolving funny suicide comments. Though Chen et al used hybrid model to increasethe confidence in detection, their solution was based only on content features.

The performance of the proposed solution with content feature alone and hybrid features is measured and results are given in Table 7. With content features alone, the accuracy in proposed solution was at 91.2% which is almost similar to that Tadesse et al. Inclusion of context and feature enrichment has increased the accuracy in proposed solution.

The performance of the proposed solution with and without feature enrichment is measured and the result is given in Table 8.

The feature enrichment due to multi data active learning has increased the accuracy by 2.56%.
The ROC is compared across the solutions and the result is shown in Fig. 3.

Higher the ROC area, better is the classifier performance in classifying the reviews. Plotting the true positive rate against various false positive rate, the ROC area in proposed solution is 0.907 which is higher compared to Tadesset et al (0.903), Ji et al (0.853) and Chen et al (0.82).

The accuracy and loss in proposed solution is measured for various epocs and given in Fig. 4 and Fig. 5.

The lowest value for loss is achieved at epochs of 100 and highest value for accuracy is achieved at epochs of 15.

The accuracy of emotion tagging is measured for various emotions in proposed solution for GoodNewsEveryone dataset and the result is given in Table 9.

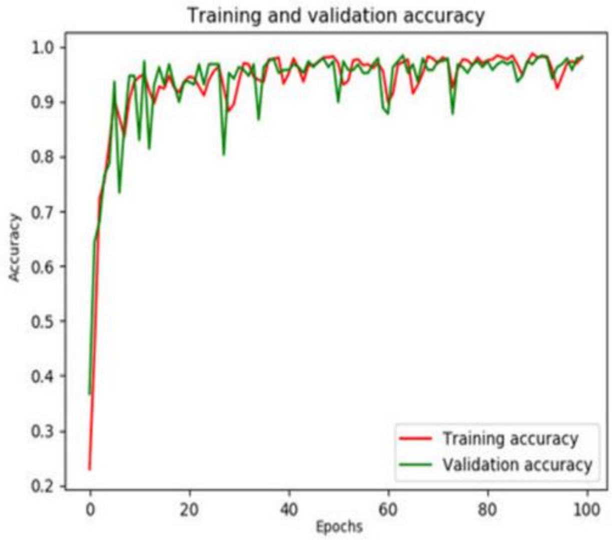

Figure. 4 Accuracy in proposed solution

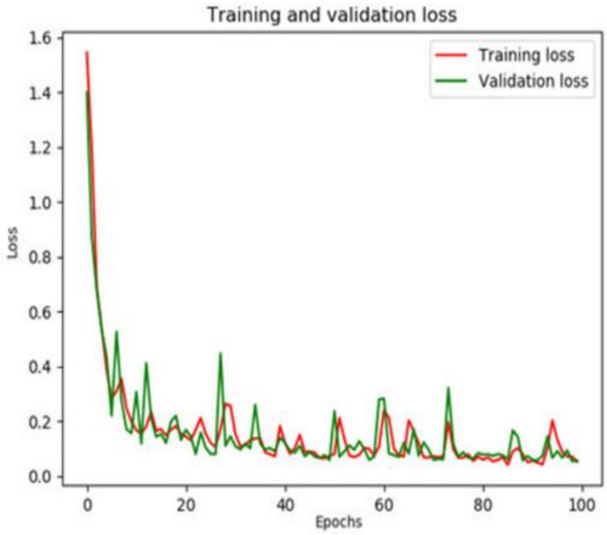

Figure. 5 Loss in proposed solution

Table 9. class wise accuracy

Emotions	**Accuracy**
Happiness	95.56
Sadness	97.67
Anger	96.12
Disgust	94.32
Surprise	91.12
Fear	97.12
Average	**95.31**

The average accuracy of emotion detection is 95.31 in proposed NMT and highest value of accuracy is achieved for Sadness emotion at 97.67.

5. Conclusion

In this work, a named entity emotion intensity tagging for SI detection from social media texts is proposed. As part of work, a hybrid feature representation combining both content and context feature is proposed. The hybrid feature is enriched using a multi data active learning technique. The enriched hybrid features are used for named entity emotion intensity tagging using neural machine translation. SI detection with emotion intensity tagging provided higher accuracy compared to existing works. The proposed solution achieved 95.54% accuracy which is atleast 3% higher compared to existing works. Extending the solution to multi lingual texts and extending the multi data learning for multi lingual datasets is in scope of future work.

Conflicts of interest

Authors declare no conflict of interest.

Author contributions

K. Soumya conceptualized the idea, experimented with the solution and documented the paper.

References

[1] R. Morese and C. Longobardi, “Suicidal ideation in adolescence: a perspective view on the role of the ventromedial prefrontal cortex”, *Front Psychol*, pp. 11-713, 2020.

[2] World Health Organization, *World Health Statistics 2021: Monitoring Health for the SDGs, Sustainable Development Goals. Industry and Higher Education*, Geneva, Switzerland: World Health Organization, 2021.

[3] D. Klonsky, M. May and Y. Saffer, “Suicide, suicide attempts, and suicidal ideation”, *Annu Rev Clin Psychol*,Vol. 12,pp. 307–30, 2016.

[4] M. Kumar, M. Dredze, G. Coppersmith, and M. Choudhury,“Detecting Changes in Suicide Content Manifested in Social Media Following Celebrity Suicides”, In: *Proc. of the 26th ACM Conference*, 2015.

[5] E. Carlyle, D. Guidry, and K. Williams, “Suicide conversations on InstagramTM: contagion or caring?”, *J. Commun. Healthc.*, Vol. 11, No. 1, pp. 12-18, 2018.

[6] J. Jashinsky, H. Burton, L. Hanson, and J. West, “Tracking suicide risk factors through twitter in the US. Crisis”, *Journal of Crisis Intervention and Suicide Prevention*,Vol. 35, No. 1, pp. 1-9, 2014.

[7] J. Wang, M. Plöderl, M. Häusermann, and G. Weiss, “Understanding suicide attempts among gay men from their self-perceived causes”, *J. Nerv. Ment. Dis.*, Vol. 203, No. 7,

Received: May 31, 2022. Revised: August 22, 2022.

pp. 499–506, 2015.

[8] S. Chattopadhyay, "A Study on Suicidal Risk Analysis", In: *Proc. of International Conference on e-Health Networking, Application and Services*, pp. 74-78,2007.

[9] D. Meshi, I. Tamir, and R. Heekeren, "The emerging neuroscience of social media. Trends", *Cogn. Sci.*, Vol. 19, No. 12, pp. 771–782, 2015.

[10] S. Ji, S. Pan, and X. Li, "Suicidal Ideation Detection: A Review of Machine Learning Methods and Applications", *IEEE Transactions on Computational Social Systems*, Vol. 8, No. 1, pp. 214-226, 2021.

[11] W. Wang, L. Chen, M. Tan, and P. Sheth, "Discovering fine-grained sentiment in suicide notes", *Biomedical Informatics Insights*, Vol. 5, No. Suppl 1, p. 137, 2012.

[12] M. Tai and W. Chiu, "Artificial neural network analysis on suicide and self-harm history of taiwanese soldiers", In: *Proc. of Second International Conference on Innovative Computing, Information and Control. IEEE*, pp. 363–363, 2007.

[13] M. Liakata, J. H. Kim, S. Saha, J. Hastings, and D. Rebholzschuhmann, "Three hybrid classifiers for the detection of emotions in suicide notes", *Biomedical Informatics Insights*, Vol. 2012, No. Suppl. 1, pp. 175–184, 2012.

[14] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space", *arXiv*,Vol. 1, 2013.

[15] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation", In: *Proc. of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1532–1543, 2014.

[16] S. Ji, P. Yu, S. Pan, and G. Long, "Supervised learning for suicidal ideation detection in online user content", *Complexity*, Vol. 2018.

[17] G. Coppersmith, R. Leary, E. Whyne, and T. Wood, "Quantifying suicidal ideation via language usage on social media", In *Proc. of Joint Statistics Meetings Proceedings, Statistical Computing Section, JSM*, 2015.

[18] M. Mulholland and J. Quinn, "Suicidal tendencies: The automatic classification of suicidal and non-suicidal lyricists using nlp", In: *Proc. of Sixth International Joint Conference on Natural Language Processing*, pp. 680–684, 2013.

[19] X. Huang, L. Zhang, D. Chiu, T. Liu, X. Li, and T. Zhu, "Detecting suicidal ideation in chinese microblogs with psychological lexicons", In: *Proc. of IEEE 11th International Conference on Ubiquitous Intelligence and Computing and Autonomic and Trusted Computing*, pp. 844–849,2014.

[20] M. Liakata, J. H. Kim, S. Saha, J. Hastings, and D. Rebholzschuhmann, "Three hybrid classifiers for the detection of emotions in suicide notes", *Biomedical Informatics Insights*, Vol. 2012, No. Suppl. 1, pp. 175–184, 2012.

[21] J. Pestian, H. Nasrallah, P. Matykiewicz, A. Bennett, and A. Leenaars, "Suicide note classification using natural language processing: A content analysis", *Biomedical Informatics Insights*, Vol. 2010, No. 3, p. 19, 2010.

[22] R. Braithwaite, C. G. Carrier, J. West, D. Barnes, and L. Hanson, "Validating machine learning algorithms for twitter data against established measures of suicidality", *JMIR Mental Health*, Vol. 3, No. 2, p. 21, 2016.

[23] L. Nobles, J. Glenn, K. Kowsari, and E. Barnes, "Identification of imminent suicide risk among young adults using text messages," In: *Proc. of the CHI Conference on Human Factors in Computing Systems*, pp. 1–11, 2018.

[24] G. Coppersmith, R. Leary, P. Crutchley, and A. Fine, "Natural language processing of social media as screening for suicide risk", *Biomedical Informatics Insights*,Vol. 10, 2018.

[25] R. Sawhney, P. Manchanda, P. Mathur, R. Shah, and R. Singh, "Exploring and learning suicidal ideation connotations on social media with deep learning", In *Proc. of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp. 167–175,2018.

[26] M. Tadesse, H. Lin, B. Xu, and L. Yang, "Detection of suicide ideation in social media forums using deep learning", *Algorithms*, Vol. 13, No. 1, p. 7, 2020.

[27] S. Ji, X. Li, Z. Huang, and E. Cambria, "Suicidal ideation and mental disorder detection with attentive relation networks", *arXiv:2004.07601*, 2020.

[28] M. Matero, A. Idnani, Y. Son, S. Giorgi, and A. Schwartz, "Suicide risk assessment with multi-level dual-context language and bert," In *Proc. of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pp. 39–44, 2019.

[29] L. Chen, A. Aldayel, N. Bogoychev, and T. Gong, "Similar minds post alike: Assessment of suicide risk using a hybrid model", In: *Proc. of the Sixth Workshop on Computational*

Linguistics and Clinical Psychology, pp. 152–157, 2019.

[30] R. Wicentowski and R. Sydes, "Emotion detection in suicide notes using maximum entropy classification", *Biomedical Informatics Insights*, Vol. 5, pp. BII–S8972, 2012.

[31] M. Schoene and N. Dethlefs, "Automatic identification of suicide notes from linguistic and sentiment features", In: *Proc. of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage*, Social Sciences, and Humanities, pp. 128–133, 2016.

[32] Y. Wang, S. Wan, and C. Paris, "The role of features and context on suicide ideation detection", In: *Proc. of the Australasian Language Technology Association Workshop*, pp. 94–102, 2016.

[33] M. Choudhury, E. Kiciman, M. Dredze, G. Coppersmith, and M. Kumar, "Discovering shifts to suicidal ideation from mental health content in social media", In *Proc. of the CHI Conference on Human Factors in Computing Systems*,ACM, pp. 2098–2110, 2016.

[34] L. Guan, B. Hao, Q. Cheng, S. Yip, and T. Zhu, "Identifying chinese microblog users with high suicide probability using internet-based profile and linguistic features: classification model", *JMIR Mental Health*, Vol. 2, No. 2, p. e17, 2015.

[35] A. George, B. Ganesh, A. Kumar, and K. Soman, "Significance of Global Vectors Representation in Protein Sequences Analysis. Computer Aided Intervention and Diagnostics in Clinical and Medical Images", *Springer*, pp. 261–269, 2019.

[36] L. William, K. Clar, and D. Jurafsky, "Inducing domain-specific sentiment lexicons from unlabeled corpora", In: *Proc. of EMNLP*, Vol. 2016, p. 595, 2016.

[37] S. Ji, X. Li, and Z. Huang, "Suicidal ideation and mental disorder detection with attentive relation networks", *Neural Comput & Applic*, 2021.

[38] E. Camp, "Sarcasm, pretense, and the semantics/pragmatics distinction", *Noûs*, Vol. 4, No. 46, pp. 587-634, 2012.

[39] https://www.ims.unistuttgart.de/en/research/resources/corpora/goodnewseveryone/

[40] http://users.umiacs.umd.edu/~resnik/umd_reddit_suicidality_dataset.html