Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 1618–1626
Marseille, 11–16 May 2020

CEASE, a Corpus of Emotion Annotated Suicide notes in English

Soumitra Ghosh, Asif Ekbal, Pushpak Bhattacharyya
Department of Computer Science and Engineering
Indian Institute of Technology Patna, Patna, India
{1821cs05, asif, pb}@iitp.ac.in

Abstract

A suicide note is usually written shortly before the suicide, and it provides a chance to comprehend the self-destructive state of mind of the deceased. From a psychological point of view, suicide notes have been utilized for recognizing the motive behind the suicide. To the best of our knowledge, there are no openly accessible suicide note corpus at present, making it challenging for the researchers and developers to deep dive into the area of mental health assessment and suicide prevention. In this paper, we create a fine-grained emotion annotated corpus *(CEASE)* of suicide notes in English, and develop various deep learning models to perform emotion detection on the curated dataset. The corpus consists of 2393 sentences from around 205 suicide notes collected from various sources. Each sentence is annotated with a particular emotion class from a set of 15 fine-grained emotion labels, namely *(forgiveness, happiness_peacefulness, love, pride, hopefulness, thankfulness, blame, anger, fear, abuse, sorrow, hopelessness, guilt, information, instructions)*. For the evaluation, we develop an ensemble architecture, where the base models correspond to three supervised deep learning models, namely Convolutional Neural Network (CNN), Gated Recurrent Unit (GRU) and Long Short Term Memory (LSTM).We obtain the highest test accuracy of 60.17%, and cross-validation accuracy of 60.32%.

Keywords: Emotion Analysis, Suicide notes, Deep learning

1. Introduction

Suicide, a serious public well-being concern and a major cause of death worldwide, can be marked as the death caused by self-directed harmful behaviour with any intent to end one's life. Moments before committing this horrendous act, usually an individual's mind is flooded with a range of unpleasant emotions. (Pestian et al., 2008; Pestian et al., 2010; Duch et al., 2008). A suicide note is a significant asset when attempting to evaluate a patient's risk of attempting suicide repeatedly. The suicide note furnishes us with the firsthand information about that individual's specific personality status and mind rationale (Pestian and Matykiewicz, 2008). Suicide notes may serve some informative purpose and may have a remedial role in helping the surviving relatives to understand the reason behind suicide. An understanding of the messages contained inside suicide notes could be helpful for suicide aversion programs. Analyzing the inward feelings uncovered in the suicide note may assist us with identifying individuals who conceivably have suicide ideation (Xu et al., 2012), and thus prevent the misery from happening. Modelling the emotions present in such notes may help health experts in surveying suicide hazard, by contrasting the model with writings written by at-risk subjects, such as mental patients or online content producers. The success of suicide prevention, a significant general well-being concern worldwide, relies on satisfactory suicide hazard evaluation.

Real-world suicide notes are very scarce because of the sensitive nature of the document. The i2b2 Shared Task 2011 (Pestian et al., 2012) introduced a fine-grained (15 emotion labels) emotion annotated corpus of suicide notes in English with the objective to categorize suicide notes into various emotion labels. But as of now, this corpus is not available anymore publicly. So we take this opportunity to create a similar corpus and make it available in the public domain to facilitate research. We create a corpus from the suicide notes in English comprising of 2393 fine-grained emotion annotated sentences from around 205 suicide notes, collected primarily from two sources (web and book). Annotation is done at the sentence level within complete documents (notes), and each sentence can be labelled with only one emotion (called as the primary emotion). Where one would expect most of a suicide notes to be rich with emotional-content, around 58% of the training sentences are annotated with neutral sentiment tags like instructions or information. This leaves very little data for the thirteen remaining emotion labels, which include more conventional emotions such as sorrow, anger, fear and love. We experiment with three different deep learning models, each makes use of a similar feature set, and is trained to classify among the 15 labels.

The main contributions of our current work are two-fold, *viz.* (i). providing a benchmark setup for emotion detection in suicidal notes in English; and (ii). creating a fine-grained emotion labelled corpus in suicidal notes. We build ensemble framework where deep learning models are used as the base models. We obtain an accuracy of 60.17% for the test set and 60.32% for cross-validation. We also compare with several feature-based machine learning models to show the efficacy of our proposed approach.

The rest of the paper is organized as follows. In Section 2, we present a survey of the existing literature. In Section 3, various aspects of resource creation and challenges are discussed. Observations from our created resource are mentioned in Section 4. Next, we discuss the methodologies which we have implemented in our work in Section 5. Experiments and necessary analysis are presented in Section 6. Finally, we conclude in Section 7.

2. Related Work

For quite some time, there has been enthusiasm and a keen interest in analysing actual suicide notes to perceive the state of mind of the deceased while committing the horrendous act (Shneidman, 1973). There are subtle gender-based differences in the behaviour of a suicidal person. It has been observed by (Canetto and Lester, 1995) that the number of suicidal attempts by women is higher than that of men but the completion rate of suicide for men is higher than that of women. (Ho et al., 1998) examined 224 notes written by 154 subjects and made some valuable observations in their works. Most of the deceased who have left a suicide note were found to be young married (non-widowed), mentally stable (with no mental illness) females who also happened to be first-time attempters. When comparing notes of young deceased with that of elders, it was observed that notes written by young people were emotionally richer and longer than the ones written by elderly deceased. While notes of elderly people contained specific instructions, young deceased asked for forgiveness in their writings. (Joiner et al., 2002) in their study observed that sense of burden toward loved ones would characterize completers more so than attempters, whereas, on the other dimensions, completers and attempters would not differ. They could relate perceived burdensomeness to the lethality of means of completed suicide. Recent works revolve around classifying texts (at document and sentence level both) into various emotion categories such as anger, joy, sadness, fear, surprise, and disgust (Pestian et al., 2012; Aman and Szpakowicz, 2007; Neviarouskaya et al., 2011). Unavailability of digitized data related to such domain (suicide, depression) has been a major reason for not being able to apply computational methods in the analysis of suicide notes. In an attempt to find differences between real and simulated suicide notes, (Pestian et al., 2008) observed that the performance of human annotators was outperformed by machine learning tools in terms of accuracy.

Distinguishing suicide notes from newsgroup conversations using unsupervised machine learning methods were done by (Pestian and Matykiewicz, 2008). To our knowledge, the first work on automatic suicide note classification was performed by (Pestian et al., 2010) using machine learning algorithms. First of its kind, (Pestian et al., 2012) introduced a shared task (Task 2 of the 2011 i2b2 NLP Challenge) on automatic identification of emotions (at the sentence level) in suicidal notes. This research was aimed to assess the risk and facilitate prevention strategies with respect to suicides. The authors rightfully introduced an extremely fine-grained set of emotion classes to address the complex task of emotion analysis in the suicidal domain. Among the 15 classes (or, tags), 6 of them belongs to the positive sentiment (forgiveness, happiness_peacefulness, hopefulness, love, pride, thankfulness), 7 belongs to negative sentiment (guilt, hopelessness, blame, sorrow, anger, fear, abuse) and 2 to neutral (objective) classes (information and instructions). All the total 900 multi-labelled (sentence may have multiple emotion tags) annotated notes (at the sentence level) were provided for the task (600 for training and 300 for testing).

Because of the high emotive content in suicide notes and varying types of emotions that cannot be restricted to Ekman's (Ekman, 1992) or Plutchik's (Plutchik, 2001) basic emotions, we adopt the 15 fine-grained emotion tags as introduced by (Pestian et al., 2012) in the i2b2 2011 shared task challenge where the emotions were detected at the sentence level. To achieve this, we train three deep learning classifiers, namely Convolutional Neural Network (CNN), Gated Recurrent Unit (GRU) and Long Short Term Memory (LSTM), and then build two ensembles by combining these base models.

3. Corpus Development

The *CEASE*[1] corpus introduced in this paper is motivated by a similar corpus used for the i2b2 2011 shared task (Pestian et al., 2012) that contains the notes written by 1319 people before they committed suicides. Initially, this corpus was made available for research purpose, but later on discontinued the distribution. To the best of our knowledge, there is no available suicide corpus available in the public domain as of now. We take this opportunity to introduce a corpus, consisting of 2393 English sentences from 205 real-life suicide notes, to facilitate further research and development.

While preparing this corpus, we carry out the following steps.

1. **Data Collection:** document retrieval from various sources
2. **Data Cleaning:** basic pre-processing of raw notes
3. **Data Annotation:** sentence-level emotion labelling

3.1. Data Collection

We collect the data in two different phases:
In the first phase, we perform manual scavenging of actual suicide notes from the internet. Initially, the idea was to crawl suicide notes from google images by doing a google search and get them translated to text through any OCR tool. But the big question was the genuinity of the notes collected. So to avoid the collection of any doctored notes or augmented notes or fake notes, we try to collect the notes published by any popular online news websites, e-newspaper, blogs, etc. (like hufftingtonpost, dailynewsuk, tumblr).[2] Among the collected notes, we label 31 notes as

[1] Resource available at: https://www.iitp.ac.in/~ai-nlp-ml/resources.html#CEASE

[2] List of sources:
https://www.dailymail.co.uk/news/article-2303946/Angelina-Greens-mother-shares-heart-wrenching-suicide-note-seeks-anti-bullying-passed.html
https://www.telegraph.co.uk/news/uknews/law-and-order/10403944/Poem-of-a-schoolgirl-who-killed-herself-after-being-targeted-by-trolls.html
https://www.ocregister.com/2018/03/19/this-16-year-olds-suicide-letters-are-a-cry-for-help-and-a-national-call-for-change/
http://www.academia.edu/12268498/AN_ANALYSIS_OF_ADJECTIVES_USED_IN_SUICIDE_NOTES_IN_TERMS_OF_POSITIVITY_NEGATIVITY_NEUTRALITY
https://www.dailymail.co.uk/news/article-

'anonymous' (where we could not find the deceased person's names) and the rest as 'named' (where the deceased person's name was available). Among the named notes, some were from famous personalities from various fields (like Adolf Hitler, Virginia Woolf, Marilyn Monroe, etc). The second phase of our data collection mainly comprises of notes from a book titled *...Or Not to Be: A Collection of Suicide Notes* by Marc Etkind (Etkind, 1997). This book is a collection of suicide notes accompanied by a brief description of who the individual was, what their note meant and the circumstances which led up to their suicide. In total we found 138 notes, among which 38 are anonymous and the rest are from the known identities. Some notes were from persons whose note we had already collected in the first phase. We did not consider them in our final dataset. A handful of excerpts were not suicide notes, but were written by the deceased a few days or months ago before their suicide. We included such notes as well in our dataset since they carried emotionally-charged contents which were indicative of depression or some uneasy state-of-mind they were in then. We converted the scanned images of notes to the corresponding plain text using a free third-party online OCR software [3] that gave us the considerably good quality output.

3.2. Data Cleaning

OCR errors were present in the data that needed to be cleaned up manually. All the collected notes from both the phases were combined into a single text file and then sentence splitting was performed wherever a sentence ended with a sentence-ending character (., !, or ?) but not grouped with other characters into a token (such as for an abbreviation or number or other punctuation marks). Basic pre-processing like removal of punctuation marks, unnecessary blank spaces removal, conversion of sentences to lowercase, contractions ("i'm" for 'I am', "shan't" for 'shall not', etc) were taken care of. To maintain anonymity of any person mentioned in any note, we replace the name with the NAME tag. Similarly, we replace any address with the ADDRESS tag and any institution/community/organization name with the ORGANIZATION tag.

Abuse it is like they beat the hell out of me with their stupid words
Anger i feel disgusted when i see guys who make a fuss about the entrance examination all the time
Blame they led me to destroy the proof of their misdeeds and still they behave as if they have nothing to do with me
Fear i am afraid somebody will come
Forgiveness dear god please have mercy on my soul please forgive me i cannot stand the pain anymore
Guilt i cannot go on spoiling your life any longer
Happiness_Peacefulness now that it is all said i feel at peace
Hopefulness i am wondering if i will find anything in death
Hopelessness my mood is one of profound discouragement and my personal future appears bleak
Information every one who knows me and hears of it will have a different hypothesis to offer to explain why i did it
Instruction please do not think about causes of my suicide and do not inquire about what i have done before death
Love i embrace you both from the very bottom of my heart
Pride i do not think two people could have been happier than we have been
Sorrow i have been so down so god damn down i cannot get up
Thankfulness i am deeply grateful for all your kindness

Table 1: Annotation Samples from each emotion class

3.3. Data Annotation

The entire data annotation of the corpus was done by three annotators who were asked to review a note and perform sentence-wise emotion labelling with each sentence carrying at most one emotion from the 15-emotion tagset (Pestian et al., 2012) (abuse, anger, blame, fear, forgiveness, guilt, happiness_peacefulness, hopefulness, hopelessness, love, pride, sorrow, and thankfulness, information, and instructions). Some sample instances outside the corpus was used to provide training to the annotators regarding the process of annotation.

Annotators were chosen based on some certain crite-

2040694/Jamey-Rodemyer-suicide-3-bullies-face-hate-crimes-charges-death-gay-boy.html
https://www.huffingtonpost.in/entry/cora-delille-suicide_n_5366546
https://edition.cnn.com/2018/08/20/us/aaron-hernandez-suicide-note-baez-book/index.html
https://www.jconline.com/story/news/local/2013/03/18/a-letter-found-in-angel-greens-room/28936361/
https://www.thebetterindia.com/40976/vishal-pawar-save-farmer-families/
https://www.ranker.com/list/last-words-written-by-famous-people-in-their-suicide-notes/notable-quotables
https://journals.openedition.org/samaj/4481
https://www.phrases.org.uk/quotes/last-words/suicide-notes.html
https://www.storypick.com/suicide-notes/
http://lastwordslastmoments.tumblr.com/
http://suicide--notes.tumblr.com/

[3] We use https://ocr.space/ for the OCR task

ria:

- They are of at least 25 years of age and proficient in reading, writing and speaking in English.
- Mentally healthy, having no emotional attachment or history with any of the deceased whose notes we have collected.
- Willing to read, analyze and annotate notes belonging to such a sensitive domain.

In any annotation task involving multiple raters, finding out the agreement among the various annotators is an essential task to produce a reliable annotated dataset. One such measure is Cohen's Kappa coefficient (Cohen, 1960) that is considered a reliable measure for analyzing the inter-annotator agreement. It is defined as

$$K = \frac{Pr(a)-Pr(e)}{1-Pr(e)}$$

where $Pr(a)$ & $Pr(e)$ are the observed and by chance agreement among raters. The average agreement obtained on our dataset is 71.23% which shows that the annotated dataset is of acceptable quality. The final annotations for the dataset were done via majority voting on the individual annotations of the three annotators.
Some annotated sample instances have been shown in Table 1.

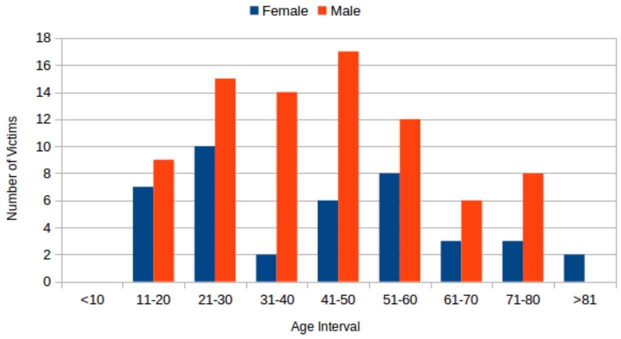

Figure 1: Gender distribution among various age groups of the notes collected

4. Corpus Analysis

Figure 1 gives an insight into the gender distribution of the deceased whose notes we have collected. Among the 120 notes collected, over 65% of them belong to males and the rest to females. Notes of the deceased belonging to the age group, 11 and 20, whom we term as Adolescents, have an almost equal gender distribution. In the age interval between 31 and 40, it seems males are highly vulnerable to commit suicide compared to females. These handful number of notes may not be sufficient to derive such conclusions but they can surely give a rough idea of the possible trend.
Along with gender, length of notes also varies across the different age groups. Figure 2 depicts the relation among the size of notes, the number of notes and age interval of the deceased. Most lengthy notes belong to the young deceased whose age falls between 11 and 30. Also, among the collected notes, over 40% of them belonged to this age interval. The number of notes, as well as the average length of the notes, show a quantifiable fall with an increase in age. Another interesting observation that can be made from the red line in Figure 2 is that the number of notes for an age interval, or in other words, the number of suicides recorded in our collected data for an age group is directly proportional to the average length of the notes for that particular age interval.

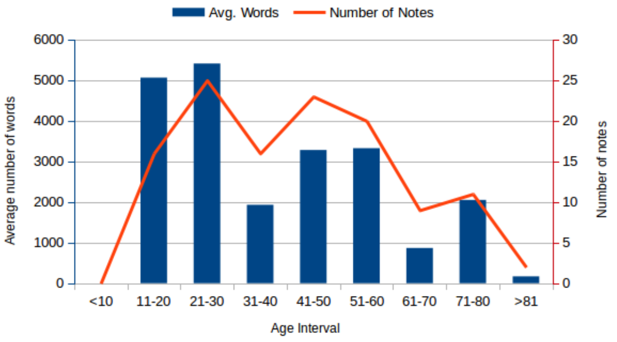

Figure 2: Representation of collected notes and their average note size across various age groups

Sentiment	**Instances**
Positive	370 (15.46%)
Negative	631 (26.37%)
Neutral	1392 (58.17%)
Total	2393

Table 2: Emotion instance distribution across the individual sentiment polarities

The major content of most of the collected notes is neutral in nature, primarily of informative nature. Next major sentiment is that of negative, comprising almost one-fourth of our corpus followed by the positive sentiment content. The distribution of instances over various sentiment polarities is shown in Table 2

5. Methodology

We develop and train three basic deep learning-based models, *viz.* Convolution Neural Network (CNN) (Kim, 2014), Long Short Term Memory (LSTM) network (Hochreiter and Schmidhuber, 1997) and Gated Recurrent Unit (GRU) network (Cho et al., 2014) on top of pre-trained word embeddings. Since GloVe [4] (Pennington et al., 2014) embedding model captures syntactic and semantic relations among words, we utilize it to learn our emotional word embeddings. The representation from the CNN/GRU/LSTM network is passed through an attention layer (Yang et al., 2016) that maps the important and relevant words from the input sentence and assign higher weights to these words, enhancing the accuracy of the output prediction. Finally,

[4] `http://nlp.stanford.edu/data/wordvecs/glove.840B.300d.zip`

we combine predictions of these models using the following two methods: using majority voting (MV), and using a MultiLayer Perceptron (MLP) network. For comparison purpose, we train four popularly used classical supervised models, namely Multinomial Naive Bayes, Support Vector Machine, Random Forest and Logistic Regression on our curated dataset, and compare with our deep learning based implementation.

5.1. Convolution Neural Network (CNN)

CNN is well known for producing good results in image classification tasks, but it is being used extensively for text classification tasks as well for its ability to produce competing results as of other state-of-the-art classifiers. Some of the prior works that make use of CNN for sentiment analysis include ((Kim, 2014); (Akhtar et al., 2016); (Singhal and Bhattacharyya, 2016)). Our CNN based classification system employs 3 convolutional layers in parallel with 100 filters of sizes 2,3 and 4 respectively. The output of the layers is added (merged) to produce a single output of the same shape as the individual layer's output. Max pooling operation (pool size = 2) is performed on the convoluted output which is further passed through an attention layer (Yang et al., 2016) to get an aggregated representation (document vector) of the informative words in a document (tweet). Lastly, the attended output is passed through 2 fully connected layers (with 100 neurons in each layer) with ReLU activation (Glorot et al., 2011) and an output layer (with 15 neurons, one for each class) with softmax activation.

5.2. Long Short Term Memory Network (LSTM)

LSTM (Hochreiter and Schmidhuber, 1997) have been one of the very useful techniques in learning long-range dependencies in text, thus eliminating the vanishing gradient problem. LSTM employs 3 gates (forget, input and output) for regulating the amount of information it wants to retain in its cell state (memory). We use a Bidirectional LSTM layer followed by an LSTM layer having 128 neurons in each layer. Word attention (Yang et al., 2016) is applied on the encoded output (from the LSTM layer) which is further passed through 2 fully connected layers (with 100 neurons in each layer) with ReLU activation (Glorot et al., 2011) and an output layer (with 15 neurons, one for each class) with softmax activation.

5.3. Gated Recurrent Unit (GRU)

Similar to LSTMs, GRUs (Cho et al., 2014) have been effective in mitigating the vanishing gradient problem while efficiently learning long-term dependencies in textual data. Unlike LSTMs where 3 gates are involved, GRUs has 2 gates (update and reset gate) to control the amount of information it wants to retain, making it simpler and faster internally than LSTMs. We use a Bidirectional GRU layer followed by a GRU layer having 128 neurons in each layer. Word attention (Yang et al., 2016) is applied on the encoded output (from the GRU layer) which is further passed through 2 fully connected layers (with 100 neurons in each layer) with ReLU activation (Glorot et al., 2011) and an output layer (with 15 neurons, one for each class) with softmax activation.

We use categorical cross-entropy as the loss function and Adam optimizer (Kingma and Ba, 2014) to train our models through backpropagation. To tackle the problem of overfitting, we employ 25% dropout (Srivastava et al., 2014) in the fully-connected layers.

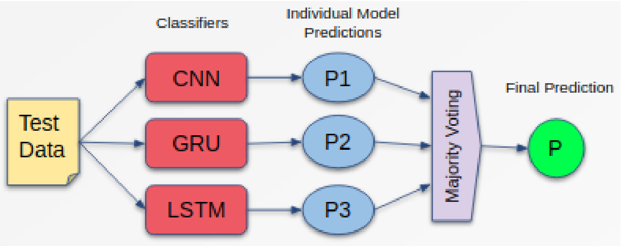

Figure 3: Majority Voting based ensemble

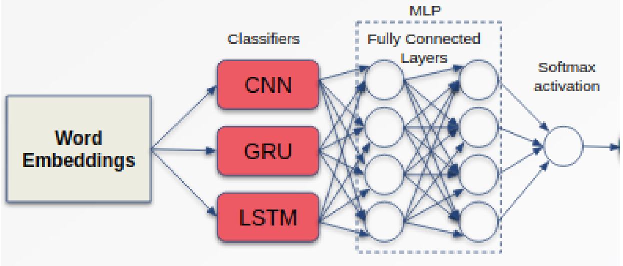

Figure 4: MLP based ensemble architecture

Emotion	Train	Test
Positive Emotions		
Forgiveness	19	5
Pride	13	3
Happiness_Peacefulness	30	8
Hopefulness	135	34
Thankfulness	30	8
Love	68	17
Negative Emotions		
Abuse	10	2
Fear	23	6
Sorrow	244	61
Anger	63	16
Blame	38	9
Guilt	59	15
Hopelessness	68	17
Neutral Emotions		
Information	936	234
Instructions	178	44

Table 3: Emotion instances distribution in individual emotion classes.

5.4. Ensemble Approaches

The ensemble of models brings an improvement in the predictive accuracy by exploiting the benefits of the individual models. It exploits the strengths of all the participating models. In our majority voting approach, for every instance, we assign the particular class which is most frequent among all the three predictions. In case all the three models produce separate predictions, we consider the output of the model having the highest F-score. The second ensemble technique, a MLP (Akhtar et al., 2017) is employed to learn over the predictions of the participating models. We use a small MLP network consisting of 2 hidden layers (with ReLU activation) of 50 and 40 neurons, respectively, followed by an output layer with Softmax activation. To reduce overfitting, 25% dropout is used in the intermediate layers and Adam optimizer to update the network weights through backpropagation. The output of the ensemble network gives us the final classification prediction. An overview of the proposed methods is depicted in Figure 3 and Figure 4.

5.5. Traditional Supervised Classifiers

To get a comparative idea of the performance of our deep-learning models, we use some popular machine-learning based classifiers that have time and again served as strong baselines in many classification tasks. We implement several features (word TF-IDF, Vader-sentiment, lexicon features, etc.) to train our machine learning classifiers for the emotion detection task.

Naive Bayes is a probabilistic classifier propelled by the Bayes theorem under a basic supposition that the traits are conditionally autonomous. The linear time complexity of naive Bayes algorithm makes it easily scalable for larger datasets. *Support Vector Machine (SVM)* is a supervised machine learning algorithm that can be used to solve a regression problem as well as a classification problem. The algorithm works to find an optimal hyperplane that is well able to differentiate among the classes. *Random forest classifier* works similar to the idea of that of divide-and-conquer by building several decision trees on various randomly selected subsets of the training data and combining the votes from the various decision trees to make the final prediction on the test data. This approach proves beneficial in improving the overall prediction accuracy as well as control over-fitting. Another very popular machine learning-based classifier is *Logistic Regression* whose multinomial version has been used for our multi-class classification task. Unlike the traditional logistic regression classifier where sigmoid activation is used in the output layer, softmax activation with cross-entropy loss function is used in the multinomial form of logistic regression.

All the models discussed in Section 5.1 to 5.5 are trained and tuned individually on our prepared dataset. The evaluation shows that results of the deep learning models are better than the classical supervised machine learning approaches by a considerable margin. On the effective combination of deep learning models as in the ensembles, the scores got further increased than the individual model's performance.

6. Experiments

In this section, we describe the experimental setup, report the results and provide necessary analysis.

6.1. Experiment Settings

For our emotion classification experiments we split our prepared dataset into train and test sets in an 80:20 ratio. Mapping the various emotions to the sentiment polarities, we see that there are six emotions of positive sentiment, seven emotions of negative sentiment and two emotions of neutral sentiment. The distribution is depicted in Table 3. We also perform 10-fold cross-validation on the entire dataset, the results of which are discussed in Section 6.2.

We have used the deep-learning models from Python's Keras library.Pre-trained Glove (Pennington et al., 2014) embedding (300 Dimension) has been used to initialize the embedding matrix (lookup table). We have implemented three deep learning models (CNN, GRU and LSTM) and two ensembles (majority voting and MLP) as discussed in Section 5.1 to 5.4. In the i2b2 shared task (Pestian et al., 2012) for Sentiment Analysis on Suicide Notes, the top 2 systems have used supervised classical machine learning approaches (Naive Bayes and SVM) on the feature set, comprising primarily of Lexicon Features, word n-grams, PoS tags. It will be unfair to compare the performance measure of our models with that of the shared task(Pestian et al., 2012), since the datasets are different. Nonetheless, for experimental purpose we trained four machine learning-based classifiers (Multinomial Naive Bayes (MNB), Support Vector Machine (SVC) with linear kernel, Random Forest (RF) and Logistic Regression (LR)) on some hand-crafted features (NLTK POS Tags[5], MPQA Subjectivity Lexicons[6], Opinion Lexicons[7], NRC Emotion Lexicons and NRC Hashtag Emotion Lexicons and [8], Vader[9](Hutto and Gilbert, 2014), Afinn[10](Nielsen, 2011), Word n-grams (n = 1 to 4), Character n-grams (n = 1 to 4)) and evaluated on our test dataset. We depict the results in Table 5.

We calculate macro-averaged scores of precision, recall and F1-score on the test dataset. We adopt these metrics as our dataset has a very high skewed distribution of the various classes, with several relevant but under-represented classes.

6.2. Results and discussion

Table 4 and Table 6 show the evaluation results of our 3 deep learning models and their ensembles. We calculate the per-class precision, recall and F1-scores for all the models which are shown in Table 4 (for all the deep learning-based models) and Table 5 (for best-performing machine learning classifier). We also obtain the classification accuracy and the macro-averaged precision, recall and F1-

[5] https://www.nltk.org/book/ch05.html

[6] http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/

[7] https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

[8] http://sentiment.nrc.ca/lexicons-for-research/

[9] https://github.com/cjhutto/vaderSentiment

[10] https://github.com/fnielsen/afinn

	CNN			GRU			LSTM			Average Voting			MLP		
Emotion	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
Forgiveness	1.00	0.20	0.33	0.67	0.40	0.50	0.80	0.80	0.80	0.75	0.60	0.67	0.60	0.60	0.60
Happiness	0.00	0.00	0.00	0.50	0.25	0.33	0.33	0.12	0.18	0.50	0.25	0.33	0.67	0.25	0.36
Hopefulness	0.71	0.35	0.47	0.63	0.50	0.56	0.52	0.47	0.49	0.62	0.44	0.52	0.59	0.47	0.52
Love	0.55	0.94	0.70	0.76	0.94	0.84	0.71	0.88	0.79	0.76	0.94	0.84	0.70	0.94	0.80
Pride	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Thankfulness	1.00	0.25	0.40	0.89	1.00	0.94	0.89	1.00	0.94	0.89	1.00	0.94	0.89	1.00	0.94
Abuse	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Anger	0.86	0.38	0.52	0.43	0.38	0.40	0.47	0.56	0.51	0.50	0.50	0.50	0.50	0.44	0.47
Blame	0.00	0.00	0.00	0.43	0.33	0.38	0.13	0.22	0.17	0.25	0.11	0.15	0.17	0.11	0.13
Fear	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Guilt	0.67	0.27	0.38	0.54	0.47	0.50	0.60	0.40	0.48	0.58	0.47	0.52	0.60	0.40	0.48
Hopelessness	0.00	0.00	0.00	0.41	0.41	0.41	0.44	0.41	0.42	0.46	0.35	0.40	0.54	0.41	0.47
Sorrow	0.30	0.11	0.17	0.26	0.25	0.25	0.32	0.38	0.34	0.27	0.21	0.24	0.26	0.31	0.28
Information	0.59	0.94	0.73	0.68	0.78	0.72	0.70	0.71	0.71	0.65	0.82	0.73	0.70	0.76	0.73
Instruction	0.68	0.34	0.45	0.62	0.45	0.53	0.64	0.57	0.60	0.71	0.45	0.56	0.68	0.57	0.62
avg / total	0.54	0.59	0.52	0.58	0.60	0.58	**0.59**	0.59	**0.59**	0.58	**0.61**	0.58	**0.59**	0.60	**0.59**

Table 4: Performance of 3 deep learning based models and 2 ensemble approaches on the test dataset. Here, the values in bold signify the maximum attained score per-metric among all the deep learning models.

	Logistic Regression		
Emotion	**P**	**R**	**F**
Forgiveness	0.50	0.20	0.29
Happiness	0.67	0.25	0.36
Hopefulness	0.50	0.21	0.29
Love	0.77	0.59	0.67
Pride	1.00	0.33	0.50
Thankfulness	0.80	0.50	0.62
Abuse	0.00	0.00	0.00
Anger	0.25	0.06	0.10
Blame	0.50	0.22	0.31
Fear	0.00	0.00	0.00
Guilt	0.50	0.27	0.35
Hopelessness	0.45	0.29	0.36
Sorrow	0.25	0.21	0.23
Information	0.58	0.86	0.69
Instruction	0.36	0.11	0.17
avg / total	0.50	0.53	0.48

Table 5: Per-class Precision, Recall and F1-scores of Logistic Regression which is the best performing baseline model.

Models	**P**	**R**	**F1**	**Acc. (%)**	**10F-CV_Acc. (%)**
CNN	0.44	0.36	0.38	59.54	59.43
GRU	0.44	0.40	0.41	58.70	58.54
LSTM	0.43	**0.43**	**0.42**	58.08	57.96
MV	**0.50**	0.41	**0.42**	**60.17**	**60.32**
MLP	0.43	0.40	0.41	59.96	60.07
MNB	0.27	0.22	0.23	43.42	43.21
SVC	0.26	0.22	0.24	46.34	47.35
RF	0.31	0.19	0.22	50.52	52.20
LR	0.47	0.27	0.33	53.44	55.98

Table 6: Test accuracy (Acc.(%)) and macro-average Precision (P), Recall (R) and F1-score (F1) values along with the average 10-fold cross-validation accuracy (10F-CV_Acc.(%)) for all the models

scores on the test set and are shown in Table 6. It is clear that the performance is greatly influenced by the number of instances available for a particular class. It can be observed that the predicted output of different models (i.e. CNN, LSTM, GRU) often varies on the same test instances. To exploit the effectiveness of the individual models, 2 popular ensemble approaches (majority voting and MLP) have been put to task to combine predictions from the individual models and give a final prediction. Both the ensembles improve upon the performance of the individual models with majority voting being the best of the lot. Clearly, all the four classical machine learning approaches fell considerably short in terms of the evaluation scores with respect to the deep learning approaches. LSTM and majority voting outperformed the other models in terms of F1-score parameter (42%) and the former scoring the highest in the recall parameter (43%) as well. Majority voting based ensemble yields the best performance with respect to the classification accuracy on the test set (**60.17%**) and macro-average precision (50%), F1-score (43%). We also perform ***10-fold cross-validation*** on the entire dataset to assess the effectiveness of the models, particularly to mitigate the risk of overfitting. The average accuracies over the 10-folds are shown in Table 6. The best attained average accuracy from 10-fold cross-validation is **60.32%** by the majority voting based ensemble technique. The proposed best performing system (MV) is observed to be statistically significant [11] over the next best performing model (MLP) and also with the best

[11] We perfom *Student's t-test* for assessing the statistical significance

Sentence	Actual	CNN	GRU	LSTM
'This will solve all our problems.'	Hopefulness	Information	Information	Information
'Love repent and see me yourself in everyone because that is what the truth is!'	Instruction	Love	Love	Hopefulness
'You never appreciated my love, kicked me in the face.'	Sorrow	Love	Blame	Blame
'I do not deserve life anymore.'	Hopelessness	Sorrow	Sorrow	Sorrow
'I was scared of getting pregnant but i gave myself completely the pain you have caused me everyday has destroyed every bit of me destroyed my soul.'	Fear	Information	Abuse	Sorrow
'You are the best you always have been.'	Pride	Information	Information	Love
'My first memories as a child are of being raped repeatedly.'	Abuse	Information	Sorrow	Sorrow
'I feel that I will not improve and cannot keep on causing you and the children so much misery.'	Guilt	Sorrow	Sorrow	Sorrow
'I am angry only because he has suspected me without any basis.'	Anger	Information	Blame	Information
'Ganesh, today you have suspected me without any reason as no act of mine is wrong.'	Blame	Information	Information	Information

Table 7: Sample instances that were misclassified by all systems.

performing machine learning based baseline model (LR) when tested against null hypothesis with p-values 0.029 and 0.036 respectively.

6.3. Error Analysis

Class imbalance problem appears to be the primary concern in achieving satisfactory class-wise accuracies or F1-scores since most of the emotion-rich classes are under-represented. Classes like Pride, Abuse and Fear failed to get any correct classification by any of the models. This can be attributed to the non-availability of instances of these classes. An interesting observation is that some other classes like *Forgiveness*, *Love* and *Thankfulness* were also under-represented, but still managed to get well-classified by the three models because of the existence of explicit emotion instances in these classes. On the contrary, classes like *Instruction* and *Sorrow* fail to achieve satisfactory scores despite having a considerable amount of instances. This is because of the high amount of implicit instances in the aforesaid classes. Table 7 shows some sample instances from the test set which were misclassified by all the 3 models, CNN, GRU and LSTM.

7. Conclusion

Analyzing suicide notes can aid in effective suicide prevention as well as rehabilitation of at-risk individuals but unavailability of such data has been the major setback. To bridge this knowledge gap to some extent, we introduce a fine-grained emotion annotated corpus on real-life suicide data containing 2393 sentences from 205 suicide notes. This fine-grained emotion annotated corpus yields several important insights on experiences felt by a person moments before committing suicide. Analysing the distribution of emotions over various subjective categories may help us to understand a pattern which usually a person with suicidal intent may exhibit before committing suicide. Early identification of at-risk individuals may avert this horrendous act from happening and facilitate timely intervention from health experts to take curative measures. The annotated corpus can aid future research to classify critical information from any emotionally charged document, specifically depressive in nature or where sentiment is involved, as well as many other emotion classification tasks such as depression analysis in suicide notes or finding similarity in suicide note content and posts in social media. We have demonstrated the use of this data by training five deep learning prediction models and four machine learning-based classifiers, with best attained accuracy of 60.17% on the test set and 10-fold cross-validation accuracy of 60.32%. The various experiments performed can serve as strong baselines for future works in this direction.

Our future research aims at extending our current work by adding intensity scores corresponding to each emotion labels (per sentence) to support multitask learning of *emotion* classes as well as *arousal* level for that emotion.

8. Acknowledgements

Authors gratefully acknowledge the support from the project titled 'Development of C-DAC Digital Forensic Centre with AI based Knowledge Support Tools', supported by MeitY, Govt. of India and Govt. of Bihar. The authors would also like to thank the linguists: Akash Bhagat, Suman Shekhar (IIT Patna) and Danish Armaan (IIEST Shibpur) for their valuable efforts in labelling the tweets.

9. Bibliographical References

Akhtar, M. S., Kumar, A., Ekbal, A., and Bhattacharyya, P. (2016). A hybrid deep learning architecture for sentiment analysis. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 482–493.

Akhtar, M. S., Gupta, D., Ekbal, A., and Bhattacharyya, P. (2017). Feature selection and ensemble construction: A two-step method for aspect based sentiment analysis. *Knowledge-Based Systems*, 125:116–135.

Aman, S. and Szpakowicz, S. (2007). Identifying expressions of emotion in text. In *International Conference on Text, Speech and Dialogue*, pages 196–205. Springer.

Canetto, S. S. and Lester, D. (1995). *The epidemiology of women's suicidal behavior.* Springer Publishing Co.

Cho, K., Van Merriënboer, B., Bahdanau, D., and Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Duch, W., Matykiewicz, P., and Pestian, J. (2008). Neurolinguistic approach to natural language processing with applications to medical text analysis. *Neural Networks*, 21(10):1500–1510.

Ekman, P. (1992). An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Etkind, M. (1997). *–or Not to be: A Collection of Suicide Notes*. Riverhead Books.

Glorot, X., Bordes, A., and Bengio, Y. (2011). Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323.

Ho, T., Yip, P. S., Chiu, C., and Halliday, P. (1998). Suicide notes: what do they tell us? *Acta Psychiatrica Scandinavica*, 98(6):467–473.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Hutto, C. J. and Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*.

Joiner, T. E., Pettit, J. W., Walker, R. L., Voelz, Z. R., Cruz, J., Rudd, M. D., and Lester, D. (2002). Perceived burdensomeness and suicidality: Two studies on the suicide notes of those attempting and those completing suicide. *Journal of Social and Clinical Psychology*, 21(5):531–545.

Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Neviarouskaya, A., Prendinger, H., and Ishizuka, M. (2011). Affect analysis model: novel rule-based approach to affect sensing from text. *Natural Language Engineering*, 17(1):95–135.

Nielsen, F. Å. (2011). A new anew: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Pestian, J. and Matykiewicz, P. (2008). Classification of suicide notes using natural language processing. *Proceedings of ACL Bio NLP*, 967.

Pestian, J. P., Matykiewicz, P., and Grupp-Phelan, J. (2008). Using natural language processing to classify suicide notes. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, pages 96–97. Association for Computational Linguistics.

Pestian, J., Nasrallah, H., Matykiewicz, P., Bennett, A., and Leenaars, A. (2010). Suicide note classification using natural language processing: A content analysis. *Biomedical informatics insights*, 3:BII–S4706.

Pestian, J. P., Matykiewicz, P., Linn-Gust, M., South, B., Uzuner, O., Wiebe, J., Cohen, K. B., Hurdle, J., and Brew, C. (2012). Sentiment analysis of suicide notes: A shared task. *Biomedical informatics insights*, 5:BII–S9042.

Plutchik, R. (2001). The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, 89(4):344–350.

Shneidman, E. S. (1973). Suicide notes reconsidered. *Psychiatry*, 36(4):379–394.

Singhal, P. and Bhattacharyya, P. (2016). Borrow a little from your rich cousin: Using embeddings and polarities of english words for multilingual sentiment classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3053–3062.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Xu, Y., Wang, Y., Liu, J., Tu, Z., Sun, J.-T., Tsujii, J., and Chang, E. (2012). Suicide note sentiment classification: a supervised approach augmented by web data. *Biomedical informatics insights*, 5:BII–S8956.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., and Hovy, E. (2016). Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489.