Deep Learning Based Approach For Detecting Suicidal Ideation in Code-Mixed Hindi-English: Baseline and Corpus

Kaustubh Agarwal
Computer Science and Engineering
Netaji Subhas University of Technology
New Delhi, India
kaustubh.co19@nsut.ac.in

Bhavya Dhingra
Electrical Engineering
Netaji Subhas University of Technology
New Delhi, India
bhavya.ee19@nsut.ac.in

Abstract

Suicide rates are rising among the youth, and the high association with suicidal ideation expression on social media necessitates further research into models for detecting suicidal ideation in text, such as tweets, to enable mitigation. Existing research has proven the feasibility of detecting suicidal ideation on social media in a particular language. However, studies have shown that bilingual and multilingual speakers tend to use code-mixed text on social media making the detection of suicidal ideation on code-mixed data crucial, even more so with the increasing number of bilingual and multilingual speakers. In this study we create a code-mixed Hindi-English (Hinglish) dataset for detection of suicidal ideation and evaluate the performance of traditional classifiers, deep learning architectures, and transformers on it. Among the tested classifier architectures, Indic BERT gave the best results with an accuracy of 98.54%.

1 Introduction

A study by the World Health Organization (WHO), has found that nearly 700,000 people die of suicide each year (WHO). Suicidal ideation, the act of thinking about, considering or planning suicide, can be attributed to multiple reasons including mental illness, traumatic stress, loss or fear of loss, social isolation, biological factors, environmental factors, genetic factors and situational factors (Wasserman et al., 2004). In the wake of COronaVIrus Disease-2019 (COVID-19), an increasing number of individuals across the world have become victims to one or more than one of these factors, that has led to increased rates of suicide worldwide (Fortgang et al., 2021).

It has been found that increase in consumption and posting on social media has a direct correlation to the tendency of expressing desires, thoughts, and intentions on pro-suicide platforms before attempting suicide (Gea and Sánchez, 2012). With COVID-19 driving social media consumption up by 72% and posting by 43% such incidents recorded an all-time high (Wold |, 2020). Danet and Herring (2007) mentioned that more than half of the people on social media platforms are not native English speakers and (Hong et al., 2011) confirmed that about 50% of the posts on Twitter are in languages other than English. These studies substantiate the need of a much broader scope for detection of suicidal ideation on social media than just the English language.

(Gupta et al., 2016) found that over 26% of the Indian population speaks more than one language, often in the form of code-switching or code-mixing. Code-switching occurs when an individual alternates between multiple languages in the context of a single conversation or situation while code-mixing is the use of two or more languages by an individual below clause level in a single social context. However, working with code-mixed data presents it's own set of challenges, including the creation of a large number of new constructions for understanding the syntax and semantics of the two or more combined languages, the availability of very small amounts of annotated data, and the use of drastically different approaches when compared to monolingual data (Çetinoğlu et al., 2016).

In this paper, we aim to detect suicidal ideation in code-mixed Hinglish. Although significant work is available for suicidal ideation detection in English (Castillo-Sánchez et al. (2020), Coppersmith et al. (2018), Mbarek et al. (2019), Ophir et al. (2020), Ramírez-Cifuentes et al. (2020), Sawhney et al. (2020), Shaoxiong Ji (2020), Tadesse et al. (2019), Vioules et al. (2018)), detection of suicidal ideation in code-mixed languages is relatively unexplored. To the best of our knowledge, we are the first to identify suicidal ideation in code-mixed

Proceedings of the 18th International Conference on Natural Language Processing, pages 100–105
Silchar, India. December 16 - 19, 2021. ©2021 NLP Association of India (NLPAI)

Table 1: Count and distribution of dataset

Category	Sample Count	Percentage
Suicidal Ideation (Positive Samples)	3098	47.41%
Non Suicidal (Negative Samples)	3435	52.59%

Hindi-English.

The contributions of our work include:

1. There is a significant lack of data in code-mixed suicidal ideation. We attempt to overcome this drawback by creating a dataset for suicidal ideation in code-mixed Hindi-English.

2. We propose the use of various existing models to create a baseline for future work in the field.

2 Methodology

The proposed methodology consists of three major parts, each consisting of a major contribution of our work.

2.1 Dataset Creation and Analysis

Even with the huge surge in suicidal ideation cases in code-mixed Hindi-English on social media platforms, there exists no dataset for suicidal ideation posts in it. Most existing research uses data from special Reddit channels like "Suicide Watch" (Ji et al. (2018), Tadesse et al. (2019)) or Twitter (Mbarek et al., 2019). However, since all of these datasets are in English, they fail to capture a large section of suicidal ideation texts that are unaccounted for due to medium of communication in specialized channels of social media (like subreddit "Suicide Watch"), frequent lack of hashtags, and deletion of such texts by social media companies due to the impact it can have on other users of their social media platform. To overcome a lack of a code-mixed dataset in this domain we scraped data from the subreddits such as Aww, Jokes, History, Discussion, Stories and Entertainment as negative samples and selected text from the "Suicide Watch" subreddit as positive samples. On the 6533 scraped samples thus obtained, we used the approach proposed by Gupta et al. (2020) to obtain code-mixed Hindi-English text from English text. The generated dataset consists of 6533 code-mixed text samples, 3098 of which are labelled as having suicidal ideation and 3435 labelled as having no suicidal ideation.

From Table-1 it can be observed that the data is fairly balanced. It is essential to ensure a fair representation of class labels in this context to eradicate unfounded bias due to training data.

Examples of the annotated data are:

Sample: Main literally aur figuratively khud ko maarnaa caahtaa huun
Translation: I want to kill myself, both literally and figuratively.
Label: Suicidal Ideation

Sample: Tumhara novels ya books mein favorite twist kaunsa hai?
Translation: What is your favorite twist in books or novels?
Label: Non Suicidal

The dataset is available on GitHub .

2.2 Creation of Hindi-English Bi-lingual Word Embeddings

Word embeddings are dense vectors that give semantic and syntactic information of words in a context (Mandelbaum and Shalev, 2016) and are a critical part of text classification tasks. However, creating embeddings requires a large amount of textual data. For this purpose, we have used 412k Hinglish tweets and 320k English tweets from Twitter for code-mixed Hindi-English data and preprocessed them by removing rare words, hashtags, mentions and Uniform Resource Locators (URLs). We experimented with two different experimental settings to form embeddings using two different techniques. For the first setting, we have used only the Hinglish tweets corpus to create embeddings and for the second one, a corpus of both English and Hinglish tweets combined. On each of these experimental settings, we tried the following two embeddings:

1. Word2Vec: This technique was introduced in 2013 (Mikolov et al., 2013) and is widely regarded as a pivotal method for creating dense word embeddings. Since a pre-trained corpus for English embeddings already exists, we trained our Hinglish corpus to create the required embeddings.

2. FastText: FastText was introduced by Facebook (Bojanowski et al., 2017a) as an extension of Word2Vec embeddings (Joulin et al., 2017). Instead of learning weights for words directly, FastText breaks words into multiple sub-words (Bojanowski et al., 2017b). This will be particularly helpful in representing rare words as embeddings since it is highly likely that their n-grams will be a subpart of other words. For example, “aww”, “awww” and it’s variations, which are very common on social media platforms, can be trained appropriately.

2.3 Deep Learning Models

Four traditional classifiers, five deep learning classifiers and two transformers have been used to create a baseline. These models include:

1. Naive Bayes (Yu et al., 2015)
2. Random Forest (Breiman, 2001)
3. Linear SVM (Ladicky and Torr, 2011)
4. RBF Kernel SVM (Daqi and Tao, 2007)
5. Series CNN (Tang et al., 2021)
6. Parallel CNN (Yao et al., 2019)
7. LSTM (Hochreiter and Schmidhuber, 1997)
8. Bi-Directional LSTM (Schuster and Paliwal, 1997)
9. Attention Based Bi-Directional LSTM (Wang et al., 2016)
10. mBERT (Devlin et al., 2019)
11. Indic BERT (Kakwani et al., 2020)

All these architectures were presented with the task of binary classification where each text was predicted as a sample of suicidal ideation or a sample having no suicidal ideation.
While some of these models have been able to detect sarcasm, irony, and other factors that may affect the classification of a suicidal ideation text, its explicit learning and merging could be an avenue for future research.

3 Experimental Settings / Modeling

For training our deep learning models, we made a fifteen percent validation split for a total of 20 epochs while the transformers are trained for 5 epochs on the same split. The model checkpoints are saved at each epoch and the model with highest validation accuracy and lowest difference from training accuracy is saved as the final model to ensure prevention of overfitting.

Word embeddings are trained using negative sampling polarity, an embedding size of 300, a window length of 10. The Adam optimizer is employed in all of the models, coupled with the binary cross entropy loss function. With the exception of the output layer, which has sigmoid activation, all layers have relu activation. We tested CNN models with various kernel sizes, number of kernels, dropouts, and strides to see how well they performed. With the following parameters, the best results are obtained: stride = 1, number of kernels = 200, dropout = 0.5

For all RNNs the hyperparameters used are dropout for recurrent state = 0.25, dropout for input state = 0.25, and number of LSTM units = 400.

4 Results

We have tested our dataset on traditional machine learning classifiers, deep learning models and transformers. The results of well known traditional classifiers have been listed in Table 2. RBF Kernel SVM gives the best results among the traditional classifiers with an accuracy of 60.8% on the given corpus.

Table 2: Accuracy from traditional classifiers

Classifier	Accuracy
Naive Bayes	51.2%
Random Forest	55.7%
Linear SVM	59.2%
RBF Kernel SVM	60.8%

Deep Learning models are tested using Word2Vec and FastText embeddings on Hindi-English (Hinglish) data only, and on Hinglish and English data combined. Table 3 shows the results obtained by training deep learning models on this corpus. The Attention Bi-LSTM trained on a Word2Vec embedding of Hinglish and English data corpus gives the best result of 90.66%. It is observed that deep learning model architectures perform better with embeddings of Hinglish

Table 3: Accuracy from Deep Learning Models

	Hinglish Data		Hinglish + English Data	
Model	Word2Vec	FastText	Word2Vec	FastText
Series CNN	71.26%	71.24%	73.60%	73.12%
Parallel CNN	73.86%	73.86%	74.36%	74.16%
LSTM	79.64%	78.52%	81.72%	80.66%
Bi-LSTM	83.42%	82.64%	85.44%	84.74%
Attention Bi-LSTM	89.66%	87.42%	90.66%	89.82%

and English data combined instead of using just Hinglish data for creating embeddings which may be a result of better semantic and correlation coverage between embeddings on English data and Hinglish data. It's also worth noting that Word2Vec produces slightly better results than the FastText embeddings. This observation could be due to the fact that code-mixed data prevents n-grams from being used as the major classification criterion. Furthermore, the better performance of deep learning models over traditional classifiers can be attributed to the fact that they can learn more about human tendencies like sarcasm and irony (Sentamilselvan et al. (2021), Potamias et al. (2020)), thus reducing incorrect predictions on them.

Given the code-mixed nature of the corpus, the transformers used for classification are mBERT and IndicBERT. Table 4 shows the results of training our corpus on these transformers. IndicBERT slightly outperforms mBERT with an accuracy of 98.54%.

Table 4: Accuracy from Transformers

Classifier	Accuracy
mBERT	96.63%
Indic BERT	98.54%

The method of generation of the dataset could have influenced the results for mBERT's classification performance, however, it is highly unlikely as separate instances of mBERT have been used for each task and the tasks performed by them are highly specialized in the given scenario. The performance of IndicBERT on the same task proves the lack of apparent correlation between the semi-supervised technique used in the creation of the dataset (Gupta et al., 2020) and the classification accuracy of mBERT.

The problem of detecting suicidal thoughts on code-mixed Hindi-English data is compounded by a lack of clean data and linguistic complications connected with code-mixed data. More data, as well as well labelled classes, are necessary to allow the model to accept noise in textual input, spelling errors, diverse contexts, and stemming words.

5 Conclusions and Future Work

The current research is the first attempt to investigate multilingual text classification to predict suicidal ideation in code-mixed Hindi-English texts, and proposes a baseline for further work along with a corpus for validation. For the objective of detecting suicidal text on the created corpus, several deep learning based models are used, including CNN, LSTM, Bi-Directional LSTM, Attention Based Bi-Directional LSTM, mBERT and Indic BERT.Since texts containing suicidal ideation in Hinglish are not available directly, a dataset is created by using a semi-supervised approach to generate code-mixed Hinglish text using pre-trained encoders and transfer learning from anonymized data in English from Reddit.

A comparison of the various models indicated that both BERT-based models mBERT and Indic BERT give exceptional results and have accomplished the task with over 96% accuracy.

Multilingual text classification is still a developing field, and future advancements could lead to better outcomes. Comparing vectors aligned with multilingual word embeddings generated using MUSE to FastText pre-aligned word embeddings may generate better results. Working on factors such as sarcasm, humor and irony that affect the classification of suicidal ideation explicitly, along with their inclusion in the creation of the model could be another potential avenue for future research.

6 CRediT author statement

Kaustubh Agarwal: Conceptualization, Software, Formal Analysis, Investigation, Data Curation, Writing- Original Draft, Review & Editing.
Bhavya Dhingra: Project administration

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017a. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017b. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Leo Breiman. 2001. Random forests. 45(1):5–32.

Gema Castillo-Sánchez, Gonçalo Marques, Enrique Dorronzoro, Octavio Rivera-Romero, Manuel Franco-Martín, and Isabel De la Torre-Díez. 2020. Suicide risk assessment using machine learning and social networks: a scoping review. 44(12).

Özlem Çetinoğlu, Sarah Schulz, and Ngoc Thang Vu. 2016. Challenges of computational processing of code-switching. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 1–11, Austin, Texas. Association for Computational Linguistics.

Glen Coppersmith, Ryan Leary, Patrick Crutchley, and Alex Fine. 2018. Natural language processing of social media as screening for suicide risk. 10:117822261879286.

B. Danet and S.C. Herring. 2007. *The Multilingual Internet: Language, Culture, and Communication Online*. Oxford University Press.

Gao Daqi and Zhang Tao. 2007. Support vector machine classifiers using rbf kernels with clustering-based centers and widths. In *2007 International Joint Conference on Neural Networks*, pages 2971–2976.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Rebecca G. Fortgang, Shirley B. Wang, Alexander J. Millner, Azure Reid-Russell, Anna L. Beukenhorst, Evan M. Kleiman, Kate H. Bentley, Kelly L. Zuromski, Maha Al-Suwaidi, Suzanne A. Bird, Ralph Buonopane, Dylan DeMarco, Adam Haim, Victoria W. Joyce, Erik K. Kastman, Erin Kilbury, Hye-In S. Lee, Patrick Mair, Carol C. Nash, Jukka-Pekka Onnela, Jordan W. Smoller, and Matthew K. Nock. 2021. Increase in suicidal thinking during COVID-19. 9(3):482–488.

PM Gea and CB Sánchez. 2012. Psiquiatria.comSuicidio e Internet. Medidas preventivas y de actuación.

Deepak Gupta, Asif Ekbal, and Pushpak Bhattacharyya. 2020. A semi-supervised approach to generate the code-mixed text using pre-trained encoder and transfer learning. Association for Computational Linguistics.

Sakshi Gupta, Piyush Bansal, and Radhika Mamidi. 2016. Resource creation for hindi-english code mixed social media text.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. 9(8):1735–1780.

Lichan Hong, Gregorio Convertino, and Ed H. Chi. 2011. Language matters in twitter: A large scale study. In *Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21, 2011*. The AAAI Press.

Shaoxiong Ji, Celina Ping Yu, Sai fu Fung, Shirui Pan, and Guodong Long. 2018. Supervised learning for suicidal ideation detection in online user content. 2018:1–10.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In *Findings of EMNLP*.

Lubor Ladicky and Philip Torr. 2011. Linear support vector machines. pages 985–992.

Amit Mandelbaum and Adi Shalev. 2016. Word embeddings and their use in sentence classification tasks.

Atika Mbarek, Salma Jamoussi, Anis Charfi, and Abdelmajid Ben Hamadou. 2019. Suicidal profiles detection in twitter. SCITEPRESS - Science and Technology Publications.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111–3119, Red Hook, NY, USA. Curran Associates Inc.

Yaakov Ophir, Refael Tikochinski, Christa S. C. Asterhan, Itay Sisso, and Roi Reichart. 2020. Deep neural networks detect suicide risk from textual facebook posts. 10(1).

Rolandos Alexandros Potamias, Georgios Siolas, and Andreas Georgios Stafylopatis. 2020. A transformer-based approach to irony and sarcasm detection. *Neural Computing and Applications*, 32(23):17309–17320.

Diana Ramírez-Cifuentes, Ana Freire, Ricardo Baeza-Yates, Joaquim Puntí, Pilar Medina-Bravo, Diego Alejandro Velazquez, Josep Maria Gonfaus, and Jordi Gonzàlez. 2020. Detection of suicidal ideation on social media: Multimodal, relational, and behavioral analysis. 22(7):e17758.

Ramit Sawhney, Harshit Joshi, Saumya Gandhi, and Rajiv Ratn Shah. 2020. A time-aware transformer based model for suicide ideation detection on social media. Association for Computational Linguistics.

M. Schuster and K.K. Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

K. Sentamilselvan, P. Suresh, G K Kamalam, S. Mahendran, and D. Aneri. 2021. Detection on sarcasm using machine learning classifiers and rule based approach. *IOP Conference Series: Materials Science and Engineering*, 1055(1):012105.

Shaoxiong Ji. 2020. Suicidal ideation detection in online social content.

Michael Mesfin Tadesse, Hongfei Lin, Bo Xu, and Liang Yang. 2019. Detection of suicide ideation in social media forums using deep learning. 13(1):7.

Wensi Tang, Guodong Long, Lu Liu, Tianyi Zhou, Jing Jiang, and Michael Blumenstein. 2021. Rethinking 1d-cnn for time series classification: A stronger baseline.

M. J. Vioules, B. Moulahi, J. Aze, and S. Bringay. 2018. Detection of suicide-related posts in twitter data streams. 62(1):7:1–7:12.

Yequan Wang, Minlie Huang, xiaoyan zhu, and Li Zhao. 2016. Attention-based LSTM for aspect-level sentiment classification. Association for Computational Linguistics.

Gail A. Wasserman, Susan J. Ko, and Larkin Mcreynolds. 2004. Assessing the mental health status of youth in juvenile justice settings. juvenile justice bulletin.

WHO. Suicide statistics.

Suzin Wold |. 2020. COVID-19 is changing how, why and how much we're using social media.

Hongdou Yao, Xuejie Zhang, Xiaobing Zhou, and Shengyan Liu. 2019. Parallel structure deep neural network using cnn and rnn with an attention mechanism for breast cancer histology image classification. *Cancers*, 11:1901.

Zhou Yu, Vikram Ramanarayanan, David Suendermann-Oeft, Xinhao Wang, Klaus Zechner, Lei Chen, Jidong Tao, Aliaksei Ivanou, and Yao Qian. 2015. Using bidirectional lstm recurrent neural networks to learn high-level abstractions of sequential features for automated scoring of non-native spontaneous speech. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 338–345.