Feasibility of Emotions as Features for Suicide Ideation Detection in Social Media

Ratchakrit Arreerard & Scott Piao

School of Computing and Communications, Lancaster University, UK

Abstract

Suicide-related social media message detection is an important issue. Such messages can reveal a warning sign of suicidal behaviour. This paper examines the efficacy of using emotions as sole features to detect suicide-related messages. We investigated two methods which use a single emotion and a set of seven emotions as features respectively. For emotion classification, we used a classifier based on BERT named "Emotion English DistilRoBERTa-base". For detecting suicide-related messages, we tested Naive Bayes and Support Vector Machine. As our training/test data for suicide message detection, we used a publicly available dataset collected from Reddit in which each post is labelled as "suicide" or "non-suicide". Our method obtained accuracies of 76.2% and 76.8% for detecting suicide-related messages with Naive Bayes and Support Vector Machine respectively. Our experiment also shows that three emotion categories, "anger", "fear" and "sadness", have a strongest correlation with suicide-related messages.

Introduction

Over the recent years, people have become increasingly aware of mental issues, and people's mental wellbeing has received an increasing attention from society and research communities, including Natural Language Processing area. An important issue in this regard is the prevention of suicide and early detection of warning signs of such behaviour.

The World Health Organization (WHO) reported that suicide is the fourth leading cause of death[1]. There are various causes of suicide, but struggling to cope with hardship is one of the most common causes of suicidal feelings. Such hardship includes mental health problems, being abused, or long-term physical pain or illness [1].

Nowadays, many people who need a help tend to seek help on the Internet, hoping they can find information on how to solve their issues online. Particularly, with the help of social media, people can interact with others virtually regardless of time and living places, allowing people to seek support from their peers. Moreover, people can use social media anonymously. In this way, they can express their feelings and mental issues without much concern. This makes it possible for us to access textual data in social media related to various mental issues, including suicidal feelings.

[1] https://www.who.int/health-topics/mental-health#tab=tab_1

Over the recent years, a number of studies on mental issues have been reported, which were based on social media textual data that contains information about people's opinions and feelings. For example, De Choudhury et al.[2] predicted users who suffer from depression based on users' online behaviour on Twitter. Sawhney et al.[3] detected suicidal ideation from Twitter messages.

In the past, an exploratory analysis by Coppersmith et al.[4] investigated into people's feelings before and after suicide attempts. They found people tend to feel angry and sad prior to the suicide attempt. After the attempt, there is an increase in sadness shown in their online messages. Therefore, we assume emotion is one of the key features that can be used for suicide attempt detection.

Emotion detection from social media messages is a challenging task in Natural Language Processing (NLP). Most emotion detection models are designed to predict a single emotion from a given message, such as [5]. However, a single message may convey multiple emotions depending on the readers' perspective. In this paper, we investigate the feasibility of detecting multiple emotion information contained in individual social media messages and use such information to detect suicide ideation from social media. We first create a pipeline connecting emotion detection model and suicide-related text detection model, then we compare the performance of two methods which use single-emotion and multiple-emotions as features respectively.

Related Work

Social media is a platform that facilitates people in expressing and exchanging opinions, thoughts, and feelings. Moreover, it can help people relieve themselves from negative mental issues. Similarly, it can also be a place where people use to talk about their struggles and hardships, which can potentially lead to suicide attempts. Therefore, social media provides a source for researchers to study and develop methods for the early detection of suicidal ideation.

Emotion might be used as an indicative signal of suicidal behaviour [6]. Furthermore, a later study found a link between negative emotions (anger and sadness) and suicide attempts [4]. However, a recent survey showed a trend of suicidal ideation (SI) detection based on general textual features [7], such as Frequency-Inverse Document (TF-IDF) combined with machine learning algorithm [8]. Some studies compared different textual features, which generally shows embedding models combined with deep learning approaches yield better performance [3, 9].

Another approach relies on lexicons, such as counting words related to sentiment using Linguistic Inquiry and Word Count (LIWC) etc. Shah et al. [10] combined LIWC with TF-IDF and n-gram methods and produced a promising result in detecting suicide-related messages. Nevertheless, a few such attempts, including [11, 12, 13], mainly relied on subjective lexicons to detect sentiment that are used as one of features for suicide ideation detection. Their method mainly resorts to percentage of positive and negative words in the text to help detect suicide ideation.

In this work, we test two methods which use a single emotion category and a set of emotion categories respectively as sole features for detecting suicide ideation. Furthermore, we examine the level of relevance between emotion categories and suicide ideation.

Suicide Ideation Detection

As mentioned previously, people who experience suicidal thoughts tend to feel negative emotions. Moreover, they are more likely to express their feeling through social media messages. In this section, we discuss our method by which we detect suicide warning signal based on emotion information.

Test Data

In our experiment, as our test data, we used a publicly available dataset collected from Reddit platform [14] (dubbed *Suicide Dataset* henceforth). This dataset contains posts from "SuicideWatch" and "teenagers" subreddits, of which posts from "SuicideWatch" are labeled as *suicide* and posts from "teenagers" are labeled as *non-suicide*. This dataset contains 116,037 posts for each label.

With respect to the quality of the data labelling, we manually checked randomly sampled 100 posts from the "SuicideWatch" subreddits, and we found 87% of them are truly posted by people who talk about their issues. Therefore, we assume labeling of this dataset is generally reliable and it provides a reliable test data for our experiments.

Emotion Classification of Posts

In order to use the emotion as features for suicide ideation detection, we needed to detect emotion of the posts of Suicide Dataset. For this purpose, We used an available BERT emotion model [5]. This model classify text into seven emotions, including anger, disgust, fear, joy, neutral, sadness, and surprise. Figure 1 shows the statistics of each emotion category detected in the Suicide Dataset. As shown by the figure, *suicide* posts are significantly correlated to *sadness*, followed by *fear*. On the other hand, the emotion types are roughly evenly distributed across non-suicide posts, except for *disgust*.

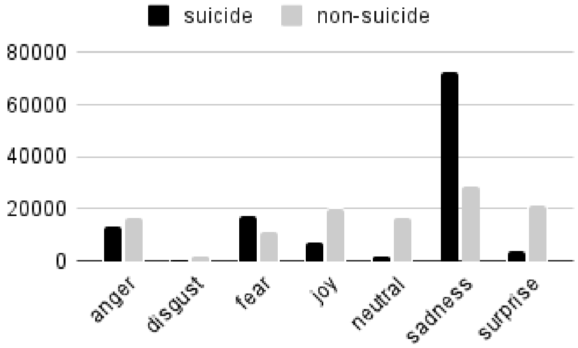

Figure 1: Statistics of emotion categories in Suicide Dataset.

According to our observation on the emotion categories distribution across the posts, we propose a suicide-related text detection model, in which the emotion detection model provides input to the suicide text detection model. Figure 2 shows the outline of this model. As shown in the figure, the first step is to detect emotion of text using an emotion detection model. Let E bea set of seven emotions, containing *anger*, *disgust*,*fear*,*joy*, *neutral*, *sadness*, and *surprise*, and P be a set of a probability distribution produced by the emotion detection model p_i of each emotion e_i, where $e_i \in E$. The emotion type of each post is determined by the maximum probability in P. Then, emotion will be used as a feature in the suicide text detection model.

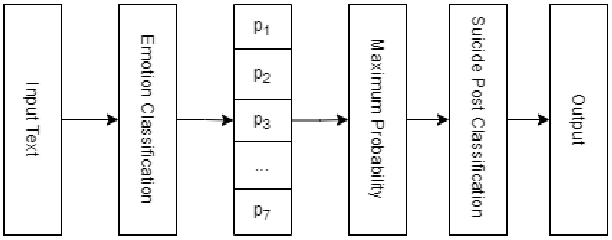

Figure 2: Suicide-related text detection framework using single-emotion feature.

However, as we mentioned previously, a single post may contain more than one emotion type. To capture the multiple emotions of a single post, we exploit the characteristic of the emotion model. The emotion model predicts the emotion of an input text by probability, which means the input text will be labeled with the emotion type that yields the maximum probability. For example, emotion of the post "Oh wow. I didn't know that." is classified as a *surprise* ($p_{surprise} = 0.975$), as other emotions have lower probabilities ranging between 0.006 and 0.003.

Usually, we neglect those low probabilities and focus on the maximum probability when we assign a single emotion type to each text. However, we can view these low probabilities as the model's opinion on other emotions. In this way, the previous sample post not only contains *surprise* emotion, but also contains a small likelihood of other emotions. Therefore, we can consider the probability set P as a feature set for suicide-related text detection, to create another framework, as shown in Figure 3.

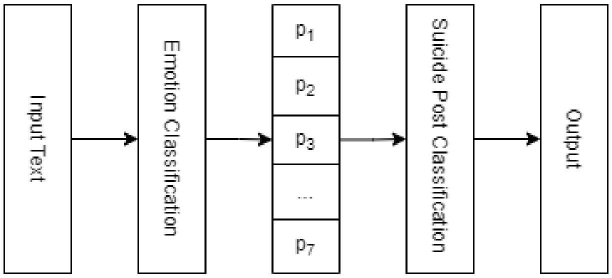

Figure 3: Suicide-related text detection framework using multiple-emotion features.

Regarding suicide-related text detection, we selected Support Vector Machine (SVM) and Naive Bayes (NB) as classifiers. In our experiment, we used SVM and NB functions of the Scikit-learn [15], a Python library.

Evaluation

In the experiment, we evaluated the two suicide-related text detection frameworks using the Suicide Dataset as test data. These methods carry out binary classification with *suicide* or *non-suicide* categories. For evaluation metrics, we employed precision, recall, and F-measure. They are calculated by the following formulae.

$$Precision = \frac{TP}{(TP + FP)} \tag{1}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{2}$$

$$F\_Measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{3}$$

where

TP = A predicted result that correctly indicates *suicide*.
TN = A predicted result that correctly indicates *non-suicide*.
FP = A predicted result that wrongly indicates *suicide*.
FN = A predicted result that wrongly indicates *non-suicide*.

In the evaluation, we adopted the 10-fold cross-validation method over the Suicide Dataset. In this way, the performance of our frameworks is measured by the average of the three evaluation metrics obtained in the ten training and testing sessions. Table 1 shows the evaluation results.

	Single-Emo		Multi-Emo	
	NB	SVM	NB	SVM
Pre.	0.248	0.490	0.762	**0.768**
Rec.	0.498	0.492	**0.760**	0.754
F.	0.331	0.491	**0.761**	**0.761**

Table 1: The performance of suicide-related post detection.

Table 1 shows the results. In the table, *Single-Emo* and *Multi-Emo* columns show the results obtained using the single and multiple-emotion features respectively.

Regarding the single-emotion feature, SVM produced the best performance, with 0.491 F-measure. On the other hand, the multiple-emotion features approach improved the performance of both NB and SVM classifiers, increasing F-measures from 0.331 and 0.491 to 0.761 respectively.

The performance of the single-emotion feature method was expected. As discussed previously, more than half of suicide-related posts are labeled as *sadness*. Hence, the baseline performance of the classifiers was expected to be around 50% precision. However, as shown above, additional emotion information improved the performance of the classifiers.

To gain further insight into the reason for the improvement, we examined further by selecting posts that were classified incorrectly by either SVM or NB. Our observation shows all the posts that were classified incorrectly by single emotion feature framework contained *sadness*, *fear*, or *anger* as the main emotion. But each of those posts also contains minor probability of other emotion types. To observe the probability distribution of the emotions in the mis-classified posts, we grouped those posts by their main emotions, one of *sadness*, *fear*, *anger*, obtaining three groups of posts. Then we calculated the average probability of each emotion type for each group of posts. Table 2 shows the average probabilities of the emotions for each group, with three rows representing three groups.

Main Emotion	Emotions						
	anger	disgust	fear	joy	neutral	sadness	surprise
anger	**0.619**	0.003	0.090	0.032	0.004	0.238	0.011
fear	0.062	0.001	**0.792**	0.024	0.003	0.102	0.014
sadness	0.050	0.001	0.034	0.038	0.005	**0.854**	0.015

Table 2: Average probability of seven emotion of suicide-related posts.

As shown in the table, the three emotions are significantly related with each other. For example, when *anger* is the main emotion, *sadness* and *fear* have probabilities of 0.238 and 0.09 respectively, which are significantly higher than others'. In the Multiple Emotion framework, the combination of the strongly correlated emotion categories improves the performance of the classifiers.

With regard to the classification of non-suicide-related posts, although they also contain negative emotions of *anger*,*fear*, or *sadness*, the probabilities of these emotions tend to have a lower weight relative to other emotions, as shown in Table 3.

Main Emotion	Emotions						
	anger	disgust	fear	joy	neutral	sadness	surprise
anger	**0.571**	0.097	0.034	0.037	0.098	0.075	0.085
fear	0.111	0.043	**0.459**	0.067	0.142	0.063	0.112
sadness	0.054	0.021	0.023	0.095	0.154	**0.458**	0.192

Table 3: Average probability of seven emotions in non-suicide-related posts.

As shown in Table 3, the presence of *neutral* and *surprise* emotions significantly help the classifiers to predict correctly, although they have those three negative emotions as the main emotion.

Conclusion

In this paper, we discussed the feasibility of using the emotion information as sole features to detect suicide-related social media posts. Furthermore, we compared the performance of two frameworks using single-emotion and multiple-emotions as features respectively. In the evaluation, our multiple-emotion feature based framework produced a promising result, with 0.761 F-measure.

We are aware that there are other features which can contribute to the detection of suicide-related social media messages. In future, we will explore combining other features with emotion information to develop a better framework for early warning of suicide ideation.

References

[1] Common causes of suicidal feelings; 2020. Accessed: 2023-03-10. https://www.mind.org.uk/information-support/types-of-mental-health-problems/suicidal-feelings/causes-of-suicidal-feelings/.

[2] De Choudhury M, Counts S, Horvitz E. Social Media as a Measurement Tool of Depression in Populations. In: Proceedings of the 5th Annual ACM Web Science Conference. WebSci '13. New York, NY, USA: Association for Computing Machinery; 2013. p. 47–56. Available from: https://doi.org/10.1145/2464464.2464480.

[3] Sawhney R, Manchanda P, Mathur P, Shah R, Singh R. Exploring and Learning Suicidal Ideation Connotations on Social Media with Deep Learning. In: Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. Brussels, Belgium: Association for Computational Linguistics; 2018. p. 167-75. Available from: https://aclanthology.org/W18-6223.

[4] Coppersmith G, Ngo K, Leary R, Wood A. Exploratory Analysis of Social Media Prior to a Suicide Attempt. In: Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology. San Diego, CA, USA: Association for Computational Linguistics; 2016. p. 106-17. Available from: https://aclanthology.org/W16-0311.

[5] Hartmann J. Emotion English DistilRoBERTa-base; 2022. https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/.

[6] Desmet B, Hoste V. Emotion detection in suicide notes. Expert Systems with Applications. 2013;40(16):6351-8. Available from: https://www.sciencedirect.com/science/article/pii/S0957417413003485.

[7] Abdulsalam A, Alhothali A. Suicidal Ideation Detection on Social Media: A Review of Machine Learning Methods. CoRR. 2022;abs/2201.10515. Available from: https://arxiv.org/abs/2201.10515.

[8] O'Dea B, Wan S, Batterham PJ, Calear AL, Paris C, Christensen H. Detecting suicidality on Twitter. Internet Interventions. 2015;2(2):183-8. Available from: https://www.sciencedirect.com/science/article/pii/S2214782915000160.

[9] Tadesse MM, Lin H, Xu B, Yang L. Detection of Suicide Ideation in Social Media Forums Using Deep Learning. Algorithms. 2020;13(1). Available from: https://www.mdpi.com/1999-4893/13/1/7.

[10] Shah FM, Haque F, Un Nur R, Al Jahan S, Mamud Z. A Hybridized Feature Extraction Approach To Suicidal Ideation Detection From Social Media Post. In: 2020 IEEE Region 10 Symposium (TENSYMP); 2020. p. 985-8.

[11] Desmet B, Hoste V. Online suicide prevention through optimised text classification. Information Sciences. 2018;439-440:61-78. Available from: https://www.sciencedirect.com/science/article/pii/S002002551830094X.

[12] Gao Y, Li B, Wang X, Wang J, Zhou Y, Bai S, et al. Detecting suicide ideation from Sina microblog. In: 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC); 2017. p. 182-7.

[13] Pestian J, Nasrallah H, Matykiewicz P, Bennett A, Leenaars A. Suicide note classification using natural language processing: A content analysis. Biomedical informatics insights. 2010;3:BII-S4706.

[14] Nikhileswar K, Vishal D, Sphoorthi L, Fathimabi S. Suicide Ideation Detection in Social Media Forums. In: 2021 2nd International Conference on Smart Electronics and Communication (ICOSEC); 2021. p. 1741-7.

[15] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12:2825-30.