TEXT CLASSIFICATION FOR SUICIDE RELATED TWEETS

FATIMA CHIROMA[1], HAN LIU[2], MIHAELA COCEA[1]

[1]School of Computing, University of Portsmouth, Portsmouth, United Kingdom
[2]School of Computer Science and Informatics
Cardiff University, Cardiff, United Kingdom
E-MAIL: fatima.chiroma@port.ac.uk, LiuH48@cardiff.ac.uk, mihaela.cocea@port.ac.uk

Abstract:

Online social networks have become a vital medium for communication. With these platforms, users have the freedom to share their opinions as well as receive information from a diverse group of people. Although this could be beneficial, there are some growing concerns regarding its negative impact on the safety of its users such as the spread of suicidal ideation. Therefore, in this study, we aim to determine the performance of machine classifiers in identifying suicide-related text from Twitter (tweets). The experiment for the study was conducted using four popular machine classifiers: Decision Tree, Naive Bayes, Random Forest and Support Vector Machine. The results of the experiment showed an F-measure ranging from 0.346 to 0.778 for suicide-related communication, with the best performance being achieved using the Decision Tree classifier.

Keywords:

Text classification; Machine Classifier; Social media; Suicide

1. Introduction

A social network is a type of social media platform which provides a web-based service that enable users to communicate with each other [1], such as Facebook and Twitter. These platforms have provided some benefit to the society as well as pose some threat to vulnerable web users who are at potential risk of harming themselves due to information they receive, such as the spread of suicidal ideation [2]. Several studies have shown the association between social media and suicidal behaviour [3–5] and according to the World Health Organization, as of 2015 suicide was the second main cause of death in individuals of age 15 – 29 years [6, 7]. They have estimated that about 800,000 people commit suicide worldwide on a yearly basis and a lot more attempt it [7]. Unfortunately, this age group are also the majority of users on social media [8, 9].

As such, this has led to some growing concerns [6, 10, 11] about the impact or influence of social media on these vulnerable users. However, if handled correctly, these platforms have abundant information regarding peoples' daily lives and behaviours, which can be used to study and understand suicide and possibly intervene [3]. According to [2], in order to assist social media users who are suicidal, it is important to understand the communication of suicidal ideation. Studies such as [2, 6, 12] have shown that it is more likely for an individual to look for non-professional support through social media instead of professional support due to concerns regarding social stigma. Therefore, this study is aimed to contribute to the ongoing research on suicide in social media by conducting a baseline experiment to measure the performance of popular machine classifiers in distinguishing between suicide-related and non-suicide related communication. Also, we undertake data manipulation to draw different versions of training data and investigate the impact of data manipulation on the performance.

The rest of this paper is organized as follows: Section 2 describes the related works on text classification and social media suicide. Section 3 explains the experimental approach and Section 4 presents the experimental results. Section 5 concludes the study and the future work is stated in Section 6.

2. Related Work

Text classification or sentiment classification [13] has been applied to different texts including tweets, and overall, it is a well-studied field [13–16]. However, studies on text classification for suicide-related communication is still in its infancy stage; as such, the research in this area is limited. Some other related works have been carried out regarding suicidal communication using other methods than text classification (e.g. statistical analysis). For example, [3–5] found a correlation be-

tween suicidal behaviour and social media, thereby raising concerns regarding human safety and building a platform for further studies to be carried out in order to help vulnerable users.

The authors in [17] created word lists of suicide-related topics and emotions; their studies aimed at classifying tweets into risky and non-risky language using machine learning. They obtained an accuracy of approximately 63%. The work in [3] demonstrated that users at risk of suicide may be detected using social media. They focused on the United States and identified suicide-related risk factors from Twitter conversations; they found a strong correlation between the Twitter data and the geographic suicide rates. The result of their studies showed the proportions of suicide-related tweets per state, with Midwestern and western states having a higher proportion than the other states. Furthermore, [4] examined the potential of social media in predicting suicide at population level. They developed and validated prediction models, and their results suggested including social media data when surveilling trends and strategies for prevention that relates to suicide.

However, to the best of our knowledge, the study that is most related to the aim of our study is [2]. They measured the performance of machine classifiers in classifying suicide-related text from Twitter by extracting text using lexical terms and names of deceased (suicide) as search key-terms. The result of their base-line experiment gave an F-measure of 0.702; this result was further improved by building and applying an ensemble classifier which achieved an improve F-measure of 0.728. In this paper we build on the work in [2] to further investigate the detection of suicide-related communication.

3 Experiments

In this section, we give a description of the experiments, which includes the data collection and annotation, preprocessing, text classification and evaluation.

3.1 Data Collection and Annotation

For the data collection and annotation, we used the data provided by [2]. They collected tweets using lexicon terms from known suicide websites to collect data that contain suicidal ideation, and also names of deceased as search keywords to collect tweets that are connected to suicide. A total of 2,000 tweets were collected and classified into seven categories by expert human annotators.

Table 1 shows the suicide communication classes and their weight in the dataset. Class c1–c6 are suicide related communications, with c1 and c3 indicating potential suicidal intention, whereas c7 contain communications that does not fall into any of the other six classes.

TABLE 1. Labelled suicide related communications and their weight [2]

Class	*Description*	*% of Dataset*
c1	Evidence of possible suicidal intent	13
c2	Campaigning (i.e. petitions etc.)	5
c3	Flippant reference to suicide	30
c4	Information or support	6
c5	Memorial or condolence	5
c6	Reporting of suicide (not bombing)	15
c7	None of the above	26

3.2 Pre-processing

Social media data are noisy [2, 18–20]. Reducing this noise will improve the quality of data and the performance of classifiers [18, 20]. Therefore, there is a need to clean the text (tweets) to prepare it for classification; this process is known as pre-processing [20]. We followed established methods [2, 21] and removed all the tweets that have less than 75% annotator agreement score. A total of 936 instances were removed, leaving 1064 instances. Table 2 shows the total instances per class and each of the classes have been assigned a new shorter description based on the description in Table 1. This descriptions will be used subsequently to identify the classes.

TABLE 2. Total instances per class

Class	*Description*	*Instances*
c1	Suicide	159
c2	Campaign	158
c3	Flippant	133
c4	Support	178
c5	Memorial	142
c6	Reports	165
c7	Other	129

These texts were further transformed using standard preprocessing techniques that are mostly used in text mining [2,20, 22], i.e. the removal of URLs, stop words and non-ASCII characters, case conversion, stemming (to reduce redundancy) and POS (Part of Speech) tagging. Furthermore, the Bag of Words (BOW) and the Inverse Document Frequency (IDF) representations were applied. BOW focus on words rather than context, it treats a text as a collection of words and ignores the syntactic and semantic information [22]. IDF gives more weight to the less frequent or rare terms while the frequent terms are likely to weigh less [22]. Subsequently, a document vector was created for each of the terms in the document using the IDF as a vector value i.e. for each unique document or term in the bag of words,

one input table column was created and each of this term column was assigned a value based on the IDF. This resulted in the extraction of 2393 terms.

3.3 Text Classification and Evaluation

According to [23], classification is among the most popular machine learning tasks in practice and it has been applied to different areas including sentiment analysis. Furthermore, they stated that classification can be specialized into two: the binary classification, which contains two categories (classes) for data instances; and multi-class classification, which contains multiple categories of instances.

Therefore, in our experiment, we organised the data into two datasets for binary and multi-class classification. The binary dataset consists of the Suicide and Flippant classes, whereas the multi-class dataset contains the Suicide, Flippant and Non-Suicide classes. The Non-suicide class comprises all the classes that are not related to suicidal ideation or have any flippant reference to suicide (i.e. Suicide or Flippant). Therefore, the classes campaign, support, memorial, reports and other, all belong to the non-suicide class in this particular case. Table 3 shows the two datasets and the instances distribution, where *Raw* are the raw instances before pre-processing and *Processed* are the instances after pre-processing.

TABLE 3. Datasets and Instances Distribution

Data-sets	Description	Class	*Raw*	*Processed*
Binary	Suicide	c1	159	156
	Flippant	c3	133	133
Multi-class	Suicide	c1	159	156
	Flippant	c3	133	133
	Non-suicide	c2, c4, c5, c6, c7	772	771

The two datasets, binary and multi-class, were then used to measure the performance of four machine classifiers – Decision Tree (DT), Naïve Bayes (NB), Random Forest (RF) and Support Vector Machine (SVM) – in classifying suicide-related text. These classifiers were chosen due to their popularity, as well as their properties: SVM works considerably well with text that are short and informal [2,12,20,24], DT is usually used to detect features to maximize information gain during classification [2], NB makes classification decision based on the likelihood of feature occurrence [2], while RF help decrease the number of false negative results [2].

The experiments were conducted using 10-fold cross-validation to minimize the influence of the training set variability on the results [24].

The next section shows the performance results of the individual classifiers i.e. the standard classification measure scores: Precision (P), Recall (R), F-measure (F) and Accuracy (A), for both the binary and multi-class datasets.

4 Results

The results for this experiment have been split into three sections to better observe and analyse the performance of the algorithms. These sections are the binary dataset, multi-class dataset, and the binary and multi-class datasets comparison.

4.1 Binary Data-set

The results per class for the binary dataset experiment (Table 4) show that RF has the best Precision (0.856 for the flippant class) and F-measure (0.806 for the suicide) scores, while NB has the best Recall (0.942 for the suicide class). As Table 4 shows, two (recall and F-measure) of the three best scores belong to the suicide class.

TABLE 4. Binary Data-set Result

Classifier	Measure	*Suicide*	*Flippant*
DT	Precision	0.803	0.752
	Recall	0.782	0.774
	F-measure	0.792	0.763
NB	Precision	0.54	0.471
	Recall	0.942	0.06
	F-measure	0.687	0.107
RF	Precision	0.719	0.856
	Recall	0.917	0.579
	F-measure	0.806	0.691
SVM	Precision	0.735	0.808
	Recall	0.872	0.632
	F-measure	0.798	0.709

4.2 Multi-class Dataset

Table 5 shows the results per class for the multi-class dataset with SVM having the best score for precision (0.892), while NB has the best recall (0.973) and DT has the best F-measure (0.879). All of these scores belong to the non-suicide class which is the majority class; thus, the good performance for this class is not surprising. The performance for the other two classes however, is much lower; for example, the class with the fewest instances (i.e. flippant) has poor results – the best values for this class are: precision of 0.550, recall of 0.376 and F-measure of 0.446, all obtained by the DT classifier.

TABLE 5. Multi-class Result

Classifier	Measure	*Suicide*	*Flippant*	*Non-suicide*
DT	Precision	0.564	0.550	0.864
	Recall	0.647	0.376	0.890
	F-measure	0.603	0.446	0.879
NB	Precision	0.862	0.214	0.794
	Recall	0.481	0.045	0.973
	F-measure	0.617	0.075	0.874
RF	Precision	0.588	0.429	0.827
	Recall	0.731	0.023	0.922
	F-measure	0.651	0.043	0.872
SVM	Precision	0.528	0.378	0.892
	Recall	0.846	0.278	0.824
	F-measure	0.650	0.320	0.856

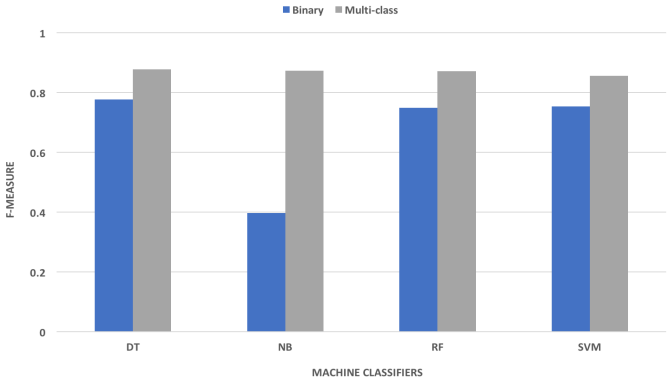

FIGURE 1. Graphical Comparison: Binary and Multi-class Data-sets

4.3 Datasets Comparison: Binary and Multi-class

For the binary and multi-class data-sets comparison, we measure the performance of the classifiers using the F-measure and accuracy. The F-measure is known to be a more appropriate measure for the performance of classifiers than accuracy, as it balances the influence between precision and recall [2, 22, 25].

The results of the comparison show that DT has the best F-measure (0.879) and accuracy (0.790), both belonging to the multi-class data-set. The result of this comparison is shown in Table 6 and Figure 1, respectively.

TABLE 6. Binary and Multi-class Data-set Comparison

Classifier	Measure	*Binary*	*Multi-class*
DT	F-measure	0.778	0.879
	Accuracy	0.779	0.790
NB	F-measure	0.398	0.874
	Accuracy	0.536	0.784
RF	F-measure	0.749	0.872
	Accuracy	0.761	0.781
SVM	F-measure	0.754	0.856
	Accuracy	0.761	0.758

It has been observed that the performance (F-measure) of the suicide and flippant class are, individually, higher for the binary dataset than for the multi-class dataset. While the previous comparison focused on the overall performance, we now focus on the performance on the different datasets for the suicide-related classes, i.e. suicide and flippant.

Thus, we calculate and compare the average scores for these classes from both datasets. This also allows us to observe the impact or influence of the majority class (or class imbalance) in the multi-class dataset on the classification of the classes of interest (suicide and flippant). Table 7 and Figure 2 shows the result of the comparison.

As the table illustrates, the best performance on the two classes of interest is obtained on the binary dataset, while the best F-score performance achieved for the multi-class in this case is 0.525. Thus, the majority non-suicide class in the multi-class dataset has a large negative influence on the prediction on the classes of interest (suicide and flippant).

TABLE 7. Class Comparison: Flippant and Suicide Only

Classifier	Measure	*Binary*	*Multi-class*
DT	Precision	0.778	0.564
	Recall	0.778	0.512
	F-measure	0.778	0.525
NB	Precision	0.54	0.538
	Recall	0.501	0.263
	F-measure	0.398	0.346
RF	Precision	0.788	0.509
	Recall	0.748	0.377
	F-measure	0.749	0.347
SVM	Precision	0.772	0.453
	Recall	0.752	0.562
	F-measure	0.754	0.485

5 Discussion

Tweets, which are short and informal text, are known to be complex because they are noisy (as any social media data) and also limited to 140 characters. As such, the users are restricted to using a language which is short and informal to express themselves. This form of communication differs from the regular means of expressing feelings on other platforms like Facebook or on paper [2]. Despite these challenges, we were able to successfully identify suicide-related communication with an F-measure of 0.778 obtained by the DT classifier on the binary dataset. While the best overall performance is achieved on the multi-class dataset (F-measure of 0.879 with DT), the higher

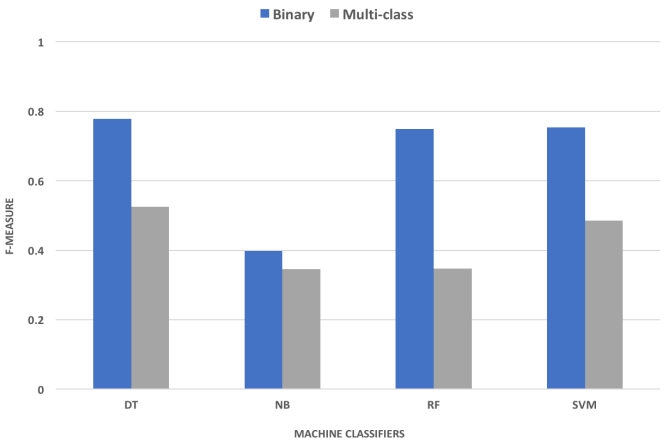

FIGURE 2. F-measure Graphical Comparison: Flippant and Suicide Only

performance is due to the detection of the non-suicide class (as shown in Section 4.3).

Our experiments indicate that focusing on the minority class(es) of interest through data manipulation may lead to better detection for these classes than when using all the data available, as the other classes which are not of interest may introduce more noise.

6 Conclusion & Future Work

In this paper, we measured the performance of four popular machine classifiers, i.e DT, NB, RF and SVM, in classifying suicide-related tweets. The results of the experiments showed that the best performance was an F-measure of 0.778 for the suicide-related communication (suicide and flippant classes).

However, to improve the the performance of machine learning techniques in classifying suicide-related communication, it is required to further examine and compare the performance of other machine learning techniques with the result of this experiment. Therefore, we intend to investigate the performance of ensemble learning approaches, which could potentially be more accurate and robust, as shown by other research [2, 26, 27].

We also plan to investigate different approaches for feature selection, which has been shown to improve classification performance in many applications [20–22, 24].

Acknowledgements

This is independent research commissioned and funded by the Department of Health Policy Research Programme (Understanding the Role of Social Media in the Aftermath of Youth Suicides, Project Number 023/0165). The views expressed in this publication are those of the author(s) and not necessarily those of the Department of Health. 9. The authors would also like to extend their gratitude to the Petroleum Technology Development Fund for their support.

References

[1] J. S. Liu, M. H. C. Ho, and L. Y. Lu, "Recent themes in social networking service research," *PLoS ONE*, vol. 12, no. 1, pp. 1–17, 2017.

[2] P. Burnap, W. Colombo, and J. Scourfield, "Machine Classification and Analysis of Suicide-Related Communication on Twitter," *Proceedings of the 26th ACM Conference on Hypertext & Social Media - HT '15*, pp. 75–84, 2015.

[3] J. Jashinsky, S. H. Burton, C. L. Hanson, J. West, C. Giraud-Carrier, M. D. Barnes, and T. Argyle, "Tracking suicide risk factors through Twitter in the US," *Crisis*, vol. 35, no. 1, pp. 51–59, 2014.

[4] H. H. Won, W. Myung, G. Y. Song, W. H. Lee, J. W. Kim, B. J. Carroll, and D. K. Kim, "Predicting National Suicide Numbers with Social Media Data," *PLoS ONE*, vol. 8, no. 4, 2013.

[5] H. Sueki, "The association of suicide-related Twitter use with suicidal behaviour: A cross-sectional study of young internet users in Japan," *Journal of Affective Disorders*, vol. 170, pp. 155–160, 2015.

[6] P. A. Cavazos-Rehg, M. J. Krauss, S. Sowles, S. Connolly, C. Rosas, M. Bharadwaj, and L. J. Bierut, "A content analysis of depression-related tweets," *Computers in Human Behavior*, vol. 54, pp. 351–357, 2016.

[7] "Suicide," Jan 2018. [Online]. Available: http://www.who.int/mediacentre/factsheets/fs398/en/

[8] M. Duggan and J. Brenner, "The demographics of social media users - 2012," Feb 2013. [Online]. Available: http://www.pewinternet.org/2013/02/14/the-demographics-of-social-media-users-2012/

[9] L. Sloan, J. Morgan, P. Burnap, and M. Williams, "Who tweets? deriving the demographic characteristics of age, occupation and social class from twitter user meta-data," *PLoS ONE*, vol. 10, no. 3, pp. 1–20, 2015.

[10] Q. Cheng, C. L. Kwok, T. Zhu, L. Guan, and P. S. Yip, "Suicide communication on social media and its psychological mechanisms: An examination of chinese microblog users," *International Journal of Environmental Research and Public Health*, vol. 12, no. 9, pp. 11 506–11 527, 2015.

[11] S. Rice, J. Robinson, S. Bendall, S. Hetrick, G. Cox, E. Bailey, J. Gleeson, and M. Alvarez-Jimenez, "Online and social media suicide prevention interventions for young people: A Focus on implementation and moderation," *Journal of the Canadian Academy of Child and Adolescent Psychiatry*, vol. 25, no. 2, pp. 80–86, 2016.

[12] C. M. Homan, R. Johar, T. Liu, M. Lytle, V. Silenzio, and C. O. Alm, "Toward Macro-Insights for Suicide Prevention: Analyzing Fine-Grained Distress at Scale," *Association for Computational Linguistics 2014*, pp. 107–117, 2014.

[13] C. Catal and M. Nangir, "A sentiment classification model based on multiple classifiers," *Applied Soft Computing Journal*, vol. 50, pp. 135–141, 2017.

[14] S. Kiritchenko, X. Zhu, and S. M. Mohammad, "Sentiment analysis of short informal texts," *Journal of Artificial Intelligence Research*, vol. 50, pp. 723–762, 2014.

[15] K. Nigam, A. K. Mccallum, S. Thrun, and T. Mitchell, "Text Classification from Labeled and Unlabeled Documents using EM," *Machine Learning*, vol. 39, pp. 103–134, 2000.

[16] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas, "Short text classification in twitter to improve information filtering," in *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, 2010, pp. 841–842.

[17] A. Abboute, Y. Boudjeriou, G. Entringer, J. Azé, S. Bringay, and P. Poncelet, "Mining twitter for suicide prevention," in *International Conference on Applications of Natural Language to Data Bases/Information Systems*. Springer, 2014, pp. 250–253.

[18] L. Barbosa and J. Feng, "Robust sentiment detection on twitter from biased and noisy data," in *Proceedings of the 23rd international conference on computational linguistics: posters*, 2010, pp. 36–44.

[19] G. B. Colombo, P. Burnap, A. Hodorog, and J. Scourfield, "Analysing the connectivity and communication of suicidal users on twitter," *Computer Communications*, vol. 73, pp. 291–300, 2016.

[20] E. Haddi, X. Liu, and Y. Shi, "The role of text preprocessing in sentiment analysis," *Procedia Computer Science*, vol. 17, pp. 26–32, 2013.

[21] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, "Sentiment strength detection in short informal text," *Journal of the Association for Information Science and Technology*, vol. 61, no. 12, pp. 2544–2558, 2010.

[22] H. Schütze, C. D. Manning, and P. Raghavan, *Introduction to information retrieval*. Cambridge University Press, 2008, vol. 39.

[23] H. Liu, M. Cocea, A. Mohasseb, and M. Bader, "Transformation of discriminative single-task classification into generative multi-task classification in machine learning context," *9th International Conference on Advanced Computational Intelligence*, pp. 66–73, 2017.

[24] Y. Liu, J. W. Bi, and Z. P. Fan, "Multi-class sentiment classification: The experimental comparisons of feature selection and machine learning algorithms," *Expert Systems with Applications*, vol. 80, pp. 323–339, 2017.

[25] Z. Liu, W. Yu, W. Chen, S. Wang, and F. Wu, "Short text feature selection for micro-blog mining," in *Computational Intelligence and Software Engineering (CiSE), 2010 International Conference on*, 2010, pp. 1–4.

[26] M. P. Ponti, "Combining classifiers: From the creation of ensembles to the decision fusion," *Proceedings - 24th SIBGRAPI Conference on Graphics, Patterns, and Images Tutorials*, pp. 1–10, 2011.

[27] C. Poulin, B. Shiner, P. Thompson, L. Vepstas, Y. Young-Xu, B. Goertzel, B. Watts, L. Flashman, and T. McAllister, "Predicting the risk of suicide by analyzing the text of clinical notes," *PLoS ONE*, vol. 9, no. 1, pp. 1–7, 2014.