Reducing suicide contagion effect by detecting sentences from media reports with explicit methods of suicide

Shima Gerani[1*], **Raphael Tissot**[1], **Annie Ying**[1], **Jennifer Redmon**[2], **Artemio Rimando** and **Riley Hun**[1]
[1]Cisco Vancouver AI Lab
[2]Cisco Research Triangle Park
{sgerani, rtissot, anying,jeredmon, rhun }@cisco.com

Abstract

Research has shown that suicide rate can increase by 13% when suicide is not reported responsibly. For example, irresponsible reporting includes specific details that depict suicide methods. To promote more responsible journalism to save lives, we propose a novel problem called "suicide method detection", which determines if a sentence in a news article contains a description of a suicide method. Our results show two promising approaches: a rule-based approach using category pattern matching and a BERT model with data augmentation, both of which reach over 0.9 in F-measure.

1 Introduction

According to the World Health Organization, close to 800,000 people die by suicide every year. The way an individual's suicide is reported has shown to be a contributing factor in whether the death will be followed by another suicide. Research in this suicide contagion effect has shown that there is a 13% increase in suicide rates in the US following highly reported suicides [7].

To reduce this suicide contagion effect, mental health experts have created best practice guidelines for media reporting.[1] One of the guidelines that research has shown to be the most harmful is explicitly depicting the suicide method in a media report, especially affecting individuals at-risk of suicide.[2] This specific guideline can be particularly unintuitive to many reporters as it conflicts with a universal journalistic standard of reporting accurate details of an event.

In this paper, we propose a novel classification problem called "suicide method detection" which can help journalists and media organizations automatically pre-screen articles before publication. This classification problem determines if a sentence in a textual news article contains a description of a suicide method that can have a harmful suicide contagion effect. To tackle this problem, we created a labeled data set of such harmful versus not harmful sentences and then proposed two approaches and evaluated them against multiple baselines. Currently, we have the baseline implementation deployed.[3]

Detecting harmful sentences including detecting suicide method in text is useful in a number of use cases:

- Media organizations and journalists can use this approach to automatically pre-screen and correct articles before publication.
- Social media organizations can potentially use this tool to filter out harmful sentences especially for at-risk individuals.
- For increasing awareness of how words can be harmful to individuals at risk, our approach can be embedded any time content is generated on the web or being communicated.

The contributions of the paper are as follows:

- We propose the problem of detecting harmful language from text that contributes to suicide contagion and the sub-problem of the suicide method classification.
- We automate the suicide method classification task using two approaches and compare them against several baseline systems.

2 Dataset and Experimental Setup

To the best of our knowledge there is no publicly available dataset of labeled documents with respect to ethical reporting of suicide. Therefore, we created our own data set. We pulled articles from the NewsAPI - there were 300,000 articles originally and we filtered that down to a set of 1625 articles that are in English and contained a report on the suicide of an individual in a textual format. A group of volunteers then manually labeled this dataset and found 281 articles that report the method of suicide.

We further labeled these 281 news articles at the sentence level, accounting to a total of 7350 labeled sentences, of which 397 were harmful. For our experiments, we randomly split the 281 positive labeled news articles into train, development and test sets and then considered the sentences in the associated split for the experiments. By doing the split at

*Contact Author
[1]https://reportingonsuicide.org
[2]https://www.poynter.org/archive/2003/reporting-on-suicide/
[3]https://reportingonsuicide.cisco.com/

the article level we decrease the chance of data leakage due to the presence of similar sentences about the suicide of the same person in train and test sets.

Since the number of positive news sentences was too small to build any classification model, we used an external data-source to augment the training data. We crawled from Wikipedia a list of famous people who died by suicide.[4] In each the people's page, we located the sentences referencing the method of suicide. As a result, we obtained 900 documents, containing both valid and harmful sentences. Table 1 shows the statistics on the number of sentences in each set.

Table 1: Statistics on the number of harmful sentences in our training, dev and test set.

split	harmful sentences	valid sentences
train	1274	14797
dev	70	1079
test	102	1705

3 Method

In this section, we explain the approaches for classifying sentences as valid or harmful. We started with a simple model based on a dictionary of suicide method terms. We then improved the dictionary-based approach by targeting each category of suicide method separately. Finally, we investigated the impact of fine-tuning BERT, a state of the art sentence embedding model. We found that the lack of labeled training data as a challenge in training a strong model. As such, we propose augmentation methods to generate high quality labeled training sentences and show the effectiveness of such data in training a strong BERT-based classifier.

3.1 Baseline: Dictionary based approach

Our baseline approach leveraged a dictionary of terms consisting of suicide action terms such as (e.g. *hang, shot, jump, etc*), suicide object terms such (e.g. *bridge, cliff, poison, etc)* as well as suicide indicator terms (e.g. *suicide, killed himself*). We manually extracted these terms from the *suicide methods* article on Wikipedia[5]. Table 2 shows the category of suicide methods that we considered as well as our dictionary terms.

We investigated two variations of the dictionary-based approach: V1) we consider a sentence as harmful if it contains at least one action (e.g jumped) or object term (e.g. gun). V2) we add a constraint that a sentence is harmful only if it contains at least one action or object term along with a suicide indicating term in that sentence. The purpose behind the added constraint is to decrease false positives by increasing the probability that the action/object terms are used in the context of explaining the suicide method.

The first two rows in Table 4 show the performance of these two dictionary-based methods on our test data set. It was observed that both variations had low *F1* while V1 has a higher recall and lower precision. Adding the constraint of referring to suicide in the same sentence in V2 has clearly helped in increasing the precision but lowers the recall by introducing more false negatives.

3.2 CPM: Category-based pattern matching

The *suicide methods* article on Wikipedia provides a categorization over different methods of suicide. We devised a representative pattern for each category and marked a sentence as harmful if it matched at least one of the patterns. Similar to the dictionary-based approach, CPM applies pattern matching. However, as opposed to exact matching of words, patterns in CPM take the POS tags and dependency structure of the sentences into account whenever necessary. Algorithm 1 shows an example of the pattern for the *"jumping from height"* category of suicide methods. Error analysis of the dictionary-based approach over training data revealed that false positives in this category were on sentences reporting a person jumping from a height to escape and not for the purpose of suicide. Having category specific patterns helps us be more specific and enables us to rule out the potential false positives of each category of suicide methods.

In terms of the results (3^{rd} row in Table 4), CPM hugely increased the precision (0.97) while having high recall (0.87) as well. There is still room to improve the recall: just like any rule-based approach, CPM is sensitive to the words and patterns and might not be general enough to capture all possible variations of writing a sentence with a specific meaning.

Algorithm 1 Jumping from heights

$text \leftarrow input$
$verb_{jump} \leftarrow [jump, leap]$
$prep_{jump} \leftarrow [from, out, off]$
$obj_{jump} \leftarrow [window, cliff, apartment, building, balcony, roof]$

$pattern_{jump} \leftarrow verb_{jump} prep_{jump} obj_{jump}$
return text matches $pattern_{jump}$ and *escape* $\notin$ *text*

3.3 BERT

BERT (**B**idirectional **E**ncoder **R**epresentations from **T**ransformers) [2] is a state of the art language representation model based on Transformers. It is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both the left and the right context in all layers. Training in BERT is based on a novel technique called Masked Language Model (MLM) which allows bidirectional training as well as Next Sentence Prediction. Experimental results show that a language model which is bidirectionally trained can have a deeper sense of language context than single-direction language models [2].

The input to BERT is a *sequence* which can be a single sentence or a pair of sentences packed together. The first token of every sequence should always be a special token ([CLS]). The final hidden state corresponding to this token denoted as $C \in R^H$ is used as the aggregate sequence representation for classification tasks.

[4]https://en.wikipedia.org/wiki/List_of_suicides

[5]https://en.wikipedia.org/wiki/Suicide_methods

Suffocation

See also: Suicide bag and Inert gas asphyxiation

Suffocation as a classification of suicide method includes strangulation and hanging.[59][60] Suicide by suffocation is the act of restricting breathing or the amount of oxygen taken in, causing asphyxia and eventually hypoxia. This may involve the use of a plastic suicide bag.[61]

It is not possible to die simply by holding the breath since a reflex causes the respiratory muscles to contract, forcing an in-breath, and the re-establishment of a normal breathing rhythm.[62] Therefore inhaling an inert gas such as helium, nitrogen, and argon, or a toxic gas such as carbon monoxide is a common method used to bring about rapid unconsciousness.[63][64]

Figure 1: Example of wikipedia article on suffocation as a suicide method.

Table 2: Suicide Method Dictionary Terms

Suicide Method Category	Dictionary terms
Firearms	gun, firearm, gunshot, bullet, self-inflicted, pistol, rifle, handgun, revolver
Suffocation	strangulation , suffocation, asphyxiation, hanging, noose, neck, ligature, plastic bag
Poisoning	carbon monoxide, pesticide, cyanide, poison, exhaust, arsenic, fume, gas, toxin, helium
Drug overdose	barbiturate, pill, heroin, cocaine, opioid, drug overdose
Fire	fire, ablaze, burn, self-immolation
Jumping from height	jump, leap, window, bridge, root, balcony, building, apartment, floor of
Wounding	wrist, throat, vein, artery, razor, sword,
Drowning	drowning, river, sea, bridge, pool, gulf, dam
Transportation	train, rail, car, bus, metro, truck, subway, vehicle
Ritual Suicide	fasting, seppuku
Other Methods	volcano jumping, lava, aircraft, gas oven
Suicide Terms	suicide, killed ...self, took ... own life, end ... life

BERT is a big neural network architecture with parameters ranging from 100 million to over 300 million. Thus, training a BERT model from scratch on a small dataset would result in overfitting. The common approach is to use a pre-trained BERT model trained on a huge dataset as a starting point and then fine-tune the model on a relatively smaller dataset related to a specific task. The pre-trained BERT model provides a powerful context-dependent sentence representation and can be used for various target tasks such as sentiment classification, question answering, etc.

In this paper, we used the "bert-base-uncased" model (12 layers, 12 self-attention heads, and with a hidden size 768, 110M parameters). We fine-tuned the model by inputting the final hidden state of the special token ([CLS]) to a dropout layer and finally the output layer.

In terms of the results (the fifth row in Table 4), fine-tuning BERT on our train set for our classification task achieves a very high recall (0.98) but not a good precision (0.49). While precision and recall are both much higher than the simple dictionary-based model, precision is not at an acceptable level and is much lower than CPM approach. We attribute the low precision to the lack of sufficient training data with only 1400 harmful sentences. In the next section, we propose an improvement by augmenting the training data with different techniques.

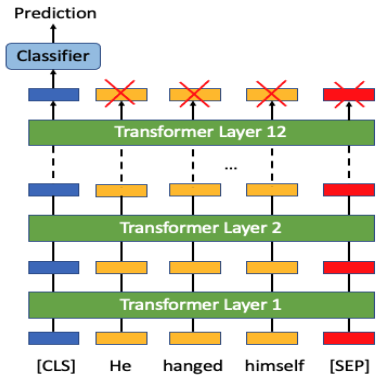

Figure 2: Fine-tuning BERT for harmful sentence classification. [6]

3.4 Text Data Augmentation

Automatic text data augmentation is a complex task; i.e., one cannot simply replace a word in a sentence with another word as it might violate grammar or even change the meaning of the sentence. Several word replacement based methods have been proposed for text data augmentation [1, 9]: *Thesaurus-based methods* [10] leverage an existing thesaurus such as WordNet to replace words or phrases with their synonyms. Further, [8] suggested the use of pre-trained embeddings such as Word2Vec for finding replacement candidates. These static word embedding based approaches have the limitation of not taking context into account when replacing a word Recent

[6] Image adapted from https://mccormickml.com/2019/07/22/BERT-fine-tuning

studies [3, 4] take contextual information into account while replacing a word for data augmentation.

Contextualized word embedding

In this paper, we used the simple implementation of *contextualized word embeddings* by the nlpaug library [6] which substitutes words by contextual word embeddings (BERT, DistilBERT, RoBERTA or XLNet). We chose the "bert-base-uncased" model for the augmentation and generated 10 samples from each positive example in the training set. The method (named BERT-context-aug in Table 4) shows the performance of the BERT model with augmented training data. While a boost in precision by BERT using augmented data can be observed, the increase is not as high as we expected. Looking at the data, we noticed that there are lots of low quality sentences generated using the contextual embedding with a bigger problem of the model sometimes replacing a word with one that completely reverses the label of the sentence. Note that we considered the label of the original sentence as the label for the augmented ones. Table 3 shows three augmented sentences generated from the harmful sentence *"He committed suicide by jumping off the bridge"*. This sentence is labeled as harmful because it mentions jumping as the method of suicide. However, not all generated sentences from this sentence are about suicide or mentioning the method of suicide. For instance, in the second sentence, the word *"suicide"* is replaced by *"robbery"* and this completely changes the label of the sentence as it does not report a method of suicide anymore. Using such augmented data with frequent wrong labels for training would negatively impact the performance of the resulting model. Some previous works have proposed conditioning on the label of the sentence as well as the context while predicting best word to replace a target word to prevent the generated words from reversing the information related to the labels of the sentences [4, 5]. We left implementing the proposed approaches and testing them for future work.

Table 3: Example outputs from contextual embedding model

sentence	true label
He committed suicide by jumping off the wall	pos
He committed robbery by jumping on the bridge	neg
He committed suicide by jumping off the roof	pos

Removing bad quality augmented sentences

To improve the quality of augmented data, we performed post processing by filtering the potentially wrong labeled sentences generated by contextual word embedding model. We classified the generated sentences using the CPM approach. We then removed augmented sentences that were labeled as valid by CPM. While we might lose a number of valid sentences due to false negative error from CPM, we increase the performance by only adding high quality positive sentences to the training set. This filtering left us with around 6700 augmented positive sentences. BERT-context-aug-filtered in Table 4 shows the result of training BERT on filtered augmented sentences from contextual embedding model. Observations showed that there is a great improvement on precision when training BERT on filtered high quality training data.

Rule-based Augmentation

As another approach we built a domain specific augmentation model. We considered the most popular synonyms and alternatives for the words related to suicide method. We crafted rules and sets of alternative words and generated new sentences from the original harmful sentences of the training set by substituting words or phrases with the alternatives in their associated set. For example, we allowed the phrase *"committed suicide"* to be replaced by one of [*"killed himself", "died by suicide", "took his own life", "ended his life"*] because they belong to the same set of suicide indicating phrases. We also had sets of alternative action/objects terms and rules to control their substitution. The resulting sentences (approximately 6300 positive sentences) are of high quality and preserve the labels. BERT-rule-aug in Table 4 shows the great boost in precision as a result of fine-tuning BERT on this rule-based augmented dataset. Finally, we combined the two sets of augmented sentences with the original data to have the final data set with around 14K positive and 14K negative samples. The last row in Table 4 shows the high performance of the BERT model using the properly augmented data.

Table 4: Classification Model Results

Method	Precision	Recall	F1
Dictionary-v1	0.28	0.83	0.41
Dictionary-v2	0.79	0.15	0.25
CPM	**0.97**	0.87	0.92
BERT	0.49	0.98	0.65
BERT-context-aug	0.66	0.94	0.78
BERT-context-aug-filtered	0.87	0.91	0.89
BERT-rule-aug	0.85	0.90	0.88
BERT-context-rule-aug-filtered	**0.90**	**0.92**	0.91

4 Discussion and Conclusion

In this paper we proposed "suicide method detection" as a novel classification problem and our approach in obtaining labeled data for developing and evaluating classification models. We proposed two promising approaches: category pattern matching and a BERT model with data augmentation, both of which reaching over 0.9 in F-measure. To better understand the strength and weaknesses of these two models, we performed an error analysis: out of a sample of 500 sentences, we examined the false positives of the 2 approaches. For the false positives, augmented BERT had 17, most of which were about when and where the suicide occured but without depicting the method; CPM had 2 false positives. This error analysis led us to believe that augmented BERT has potential room for improvement via additional data augmentation using sentences about suicide but without the method as negative instances. Another area of improvement for BERT-based models is stacking a bidirectional LSTM or a CNN on top of BERT to better capture the sequences depicting suicide method.

References

[1] Zeshan Hussain Jared Dunnmon Alexander J Ratner, Henry Ehrenberg and Christopher Ré. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. In *Advances in Neural Information Processing Systems*, pages 3236–3246, 2017.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[3] Marzieh Fadaee, Arianna Bisazza, and Christof Monz. Data augmentation for low-resource neural machine translation. *CoRR*, abs/1705.00440, 2017.

[4] Sosuke Kobayashi. Contextual augmentation: Data augmentation by words with paradigmatic relation. In *Proceedings of the NAACL-HLT*, 2018.

[5] Varun Kumar, Ashutosh Choudhary, and Eunah Cho. Data augmentation using pre-trained transformer models, 2020.

[6] Edward Ma. Nlp augmentation. https://github.com/makcedward/nlpaug, 2019.

[7] T Niederkrotenthaler, M Braun, J Pirkis, B Till, S Stack, M Sinyor, US Tran, M Voracek, Q Cheng, F Arendt, et al. Association between suicide reporting in the media and suicide: systematic review and meta-analysis. *BMJ (Clinical Research ed.)*, 368:m575–m575, 2020.

[8] William Yang Wang and Diyi Yang. That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using #petpeeve tweets. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2557–2563, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[9] Jason Wei and Kai Zhou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

[10] Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. Character-level convolutional networks for text classification. *CoRR*, abs/1509.01626, 2015.