Identifying suicide intention using natural language processing
A social media approach

Pablo González
pablogv@gatech.edu

Abstract

Suicidal ideation detection is a relevant and challenging research topic often characterized by a lack of public information regarding the victims, as well as perceived surprise among their close contacts or lack of these. In light of this, following the literature, we consider the benefits that social networking services offer and analyze how we can generate a machine learning classifier capable of accurately identifying posts that show a suicide intention. For this purpose, using social media posts from Reddit and natural language processing (NLP), the following report explores the use of different supervised algorithms to determine key characteristics that can help solve this classification problem. Our results show how transformer deep neural networks outperform other types of models due to their ability to capture contextual relationships between words in a sentence, as well as that distilled versions of these models provide similar results while also ensuring a more cost-effective, actionable alternative when scalability is required.

Problem Statement:

With the explosion of digital media and advances in the natural language processing (NLP) field, social media has become increasingly important to retrieve and analyze unstructured heterogeneous data about human behavior and, thereby, help make an impact where it was not previously possible. This is particularly the case in areas where publicly available data sets are scarce, nonexistent, or difficult to access due to data privacy concerns, but similar types of posts can be found on social networking services (SNS) such as Twitter, Facebook, and Reddit.

Suicide intention identification and prediction is an area in which the previous remarks are especially true due to the vast amount of social media data on the matter and the lack of publicly available suicide notes. Being the fourth leading cause of death among 15- to 19-year-olds as indicated by the WHO, suicide prevention is certainly a research area where social data science can have a significant impact. According to Won et al. (2013), there is a strong link between suicide-related entries and suicide frequency on a national level. Additionally, both Levis et al. (2021) and Pestian et al. (2008) find value in the use of NLP when studying suicide notes, and, in fact, deep neural networks (which allow for these studies) have already been widely used under similar experimental conditions to classify sentiments on social media (Ain et al., 2020).

There is certainly still much to know about how to prevent suicide attempts. In 2014, Facebook's experiments on their users' emotions showed the impact that social media can have on their consumers' mental health, which we believe could be advantageous to reducing the problem if we first manage to correctly identify suicide intentionality. Thus, in this project, we intend to respond to the following question: How can we generate machine learning classifiers capable of accurately identifying posts that show a suicide intention? For this purpose, we will compare multiple models, selected based on their characteristics, and, having aspects such as context in mind, compare the network's performances and identify characteristics useful to identify suicide intention.

Data Source and exploratory data analys:

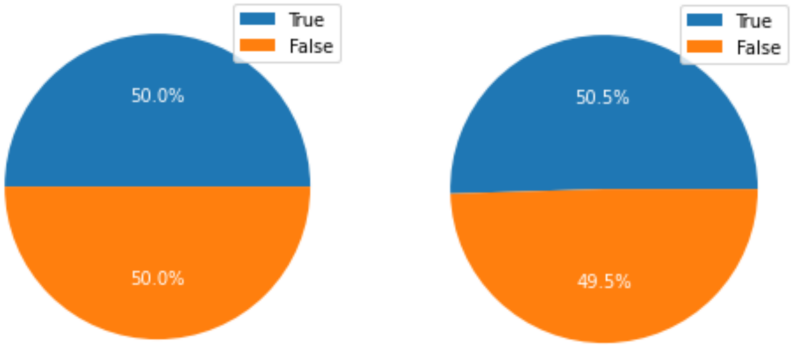
Due to Reddit's characteristic of allowing users to label posts in an easier manner, we chose to use Kaggle's "Suicide and Depression Detection" data set (Komati, 2021), which contains approximately 348 thousand observations, extracted from two different sub-forums and date ranges:

Table 1: Data Source date ranges per sub-forum

Sub-forum	Initial Date	Final Date
r/Suicide	Dec 16, 2008	Jan 2, 2021
r/Depression	Jan 1, 2009	Jan 2, 2021

When first analyzing our data set, we could notice an equal number of suicide and non-suicide (depression) cases. This particularity of the employed data was lost after concluding with the preprocessing procedures. While unexpected, due to the nature of the studied phenomena, as figure 1 illustrates, any potential problem arising from imbalanced data was solved during the web-scrapping stage, and, thus, no additional procedure such as re-sampling methods or special loss functions (focal loss) will be required on this matter.

Figure 1: Proportion of posts with suicide ideation



Original Data After pre-processing

Additionally, upon research, we could notice that there was no-missing observation, but some posts contained irrelevant information to this analysis, such as the first 40 thousand digits of pi, and, additionally, a small amount of observations contained a filler word ("filler"), which would allow any classifier to almost immediately label the post as a non-suicide post.

Table 2: Number of observations containing a filler word among classes

	Suicide	Non-suicide
Contains filler word	31	2554

Methodology

1. Data pre-processing

When analyzing the steps to take at the pre-processing stage, the employed criterion is that the steps followed must be robust, and the language used in the posts must be both understandable and contain valuable information. In light of this, special characters that are not commonly used in the English language, as well as emojis, are deleted. Noise, considered as unusually large words (longer than 100 characters) or word repetitions, is also deleted, as, for example, there was a non-negligible number of observations containing a few words being repeated over a hundred times with no context to be found, which would negatively affect our context-based classifiers and increase the training time due to the unusual length of these posts.

Additionally, since most social networking services (different from Reddit) do not allow for long posts, the length of the inputted texts must also be limited for our approach to be more generalizable to other social media platforms. In this regard, it is also important to consider that for social networking services, both scalability and efficiency would be key when implementing a potential solution, and, therefore, we should also refrain from using any type of lazy learner when selecting potential models.

Furthermore, it is important to note that the pre-processing steps depend on the type of analysis to be conducted and type of embedding to be employed. Thus, while we will not conduct any additional pre-processing steps on the selected transformer models, additional processes such as (1) lower case conversion, (2) stemming , (3) lemmatization, and (4) stop-word removal will be conducted on other types of models that do not benefit from contextual information.

A noteworthy aspect of the NLP field is that while we will be working with text data, regardless of the employed machine learning classifier, text is encoded using embedding models, thus making the inputs to our classifiers numeric vectors, simplifying our, in this case, binary classification problem, and allowing the use of in-class learned machine learning algorithms. In light of this, regarding the employed embeddings, we will employ Word2Vec embeddings, following a skipgram approach, in non-transformer models, a byte-level BPE in the case of "RoBERTa"-based models, and a Wordpiece tokenizer otherwise.

2. Analysis

As previously indicated, in this project we intend to analyze textual data on suicide intention with the goal of generating models capable of identifying whether a certain post should be flagged as a potential suicide risk or is just a regular post, so that, in the future, preventive actions can be followed accordingly. For this purpose, we employed a set of embedded machine learning classifiers and compared their performance to determine whether particular model characteristics, such as the ability to learn from context, for example via an attention mechanism (masked multi-head attention), can help achieve better performance in terms of both accuracy and sensitivity rates.

The employed models are as follows:

- Two baseline traditional machine learning classifiers: (1) A Logistic Regression and (2) a Random Forest classifier using n = 200 estimators, determined via 10-fold cross-validation. These models were selected as base-line models as they do not manage to learn contextual information.

- A Bidirectional LSTM classifier using a hidden layer size of 128 and 64 for the Bi-LSTM and final fully connected component layers, respectively. This network architecture was chosen because it learns from some contextual information, such as the immediate words following or preceding a certain word.

- A set of deep neural networks, using attention mechanism, that correspond to the Bert model's family (transformer models), including: (1) "DistilBERT", (2) "BERT", (3) "DistilRoBERTa" and (4) "RoBERTa", which were chosen to compare whether a higher number of layers and complexity

result in better performance. These models were selected because their use of context vectors allows them to better take advantage of contextual information.

An important clarification regarding the employed transformer models is that the ones using a "Distil" prefix (first and third models), were generated by a process called knowledge distillation, which is the process of transferring knowledge from large models (in this case Bert and Roberta) to smaller ones.

Furthermore, regarding the selected measures of performance, it is important to note that the accuracy and sensitivity rates measure the percentage of well-classified cases and the percentage of correctly classified posts indicating suicide intentionally, respectively. Thus, both measures are relevant to our analysis since our generated models should both perform predictions correctly and only generate warnings when a potential suicide risk is present.

Additionally, the following tables indicate the general architecture of the employed neural networks. It is important to note that for the pre-trained models, n is equal to 12 and 6 for the transformer models and their distilled versions, respectively:

Table 3: Networks' General Architecture

Layer Number	**Description**
Layer 1	Embedding layer using pretrained weights from Word2Vec
Layers 2-3	2 Stacked Bidirectional LSTM layers
Layer 4	Fully connected layer with leaky ReLU activation function ($\alpha = 0.2$)
Layer 5	Fully connected layer with softmax activation function

Layer Number	**Description**
—	Attention Mask
Layers 1 - n	Selected Pre-trained transfomer model
Layer n+1	Fully connected layer with leaky ReLU activation function ($\alpha = 0.2$)
Layer n+2	Fully connected layer with softmax activation function

Finally, it is important to note that most of the analyzed models will be imported from the Hugging-Face repository. The use of pre-trained models is due to how computationally expensive it might become to attempt training a network from scratch when our purpose is related to sentiment intensity and use of vocabulary when using the English language, and, therefore, using models that have already "experimented" this language would be more efficient than re-training every layer of the employed deep neural network.

Conclusion

Results should be evaluated in terms of both accuracy and the model's sensitivity. In other words, we would like to focus on whether the generated model (1) can correctly predict a significant percentage of the test set and (2) manages to identify posts showing potential suicide risk. The following table shows the obtained results per classifier:

Table 4: Models' Performance

Employed Classifier	**Accuracy**	**Sensitivity**
Logistic Regression	51.10%	69.75%
Random Forest	69.94%	68.91%
Bi-LSTM	82.18%	84.54%
DistilBERT	91.06%	93.49%
BERT	90.99%	92.78%
DistilRoBERTa	92.99%	93.16%
RoBERTa	93.37%	95.44%

As can be observed, all of the employed neural networks managed to outperform the employed base-line models. Additionally, transformer models significantly outperformed their counterparts, implying that the use of contextual and syntactic information via their multi-head attention mechanisms is relevant when detecting suicide intention in social media.

Finally, our results show that suicide intention can be correctly detected in most cases, even when the original post has been shortened, and that distilled transformers manage to obtain a similar performance than their original versions, thus implying that additional complexity in terms of layers and parameters might not be necessary for this classification problem.

Future work

Regarding future research, we intend to extend the current analysis by:

1. Generating a new set of rules to better separate those posts that "only" come from the r/Suicide subreddit from those that show a higher risk. This would also transform the current problem into a multi-class classification problem where unbalanced data is likely to be present but might provide some relevant additional information.
2. Studying whether posts showing suicide intention can be predicted and these users detected before they have written these posts. For this purpose, we intend to generate a web scraper that can obtain past posts from the authors of the cases labeled as showing suicide intention/ideation.

Finally, we would also recommend comparing "real" suicide notes with such posts coming from social media and studying their similarities and differences to further improve this analysis.

References:

- Ain, Q. T., Ali, M., Riaz, A., Noureen, A., Kamran, M., Hayat, B., Rehman, A. (2017). Sentiment analysis using deep learning techniques: a review. Int J Adv Comput Sci Appl, 8(6), 424.
  Available at https://www.researchgate.net/profile/Amna-Noureen/publication/318096420

- Anna Rogers, Olga Kovaleva, Anna Rumshisky; A Primer in BERTology: What We Know About How BERT Works. Transactions of the Association for Computational Linguistics 2020; 8 842–866. doi: https://doi.org/10.1162/tacl_a_00349

- Levis, M., Westgate, C. L., Gui, J., Watts, B. V., Shiner, B. (2021). Natural language processing of clinical mental health notes may add predictive value to existing suicide risk models. Psychological medicine, 51(8), 1382-1391.
  Available at https://www.cambridge.org/core/journals/psychological-medicine/article/abs/natural-language-processing-of-clinical-mental-health-notes-may-add-predictive-value-to-existing-suicide-risk-models/B9C3395DB61CAD0F79CBDDC93A35E790

- Panger, G. (2016). Reassessing the Facebook experiment: critical thinking about the validity of Big Data research. Information, Communication Society, 19(8), 1108-1126.
  Avaiable at https://www.tandfonline.com/doi/abs/10.1080/1369118X.2015.1093525

- Pestian, J., Matykiewicz, P., Grupp-Phelan, J., Lavanier, S. A., Combs, J., Kowatch, R. (2008, June). Using natural language processing to classify suicide notes. In Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing (pp. 96-97).
  Available at https://aclanthology.org/W08-0616.pdf

- Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

- Suicide and Depression Detection
  https://www.kaggle.com/nikhileswarkomati/suicide-watch

- World Health Organization. Suicide: Key facts
  Available at https://www.who.int/news-room/fact-sheets/detail/suicide

- Won, H. H., Myung, W., Song, G. Y., Lee, W. H., Kim, J. W., Carroll, B. J., & Kim, D. K. (2013). Predicting national suicide numbers with social media data.
  Available at https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0061809