Refining Suicide Predictions Based on the BERT Model & Related Statistical Data

Final Report

Jie Wang
CSCI 4502/5502 Affiliate
University of Colorado Boulder
Boulder, CO, USA
jiwa0171@colorado.edu

Tanya Leung
CSCI 4502
University of Colorado Boulder
Boulder, CO, USA
tale9912@colorado.edu

Changbing Yang
CSCI 5502
University of Colorado Boulder
Boulder, CO, USA
chya2547@colorado.edu

ABSTRACT

Suicide is slowly climbing in the modern world. To address this growing problem, we will expand on previous work that predicts suicidal risk based on user's Weibo (Chinese Twitter) post lexicon. In our research, we use and refine a BERT model to further current suicide predictions models identified by Xiaolei Huang. We created a data set from the original Weibo posts with and without emojis tags and compared the BERT model accuracies. We discovered that the data with the emoji tags has a slightly higher accuracy than data without the tags. Additionally, we created some basic data visualizations from the data given and noticed a disproportionately high number of female users in the suicide category. We also noticed a high amount of suicides were attributed to romantic issues and depression and were concentrated in urban centers of China.

KEYWORDS

BERT Model, Suicidal Ideation, Natural Language Processing, Machine Learning, Weibo, Data Visualization, Weibo

INTRODUCTION

Suicide is one of the leading causes of death, but also one of the most preventable. More and more people nowadays like to express their mood on social media, so analyzing some linguistics features from the text could help us estimate the probability of suicide. By identifying linguistic patterns that indicate risky behavior and individual or group suicide behaviour through media propagation, we can intervene and aid those with high probabilities of suicide.

Our studies focus on how emojis can impact analyzing textual data. With the rise of mobile use, there is a heavy emphasis on using non-standard written language to convey more in less words. With that, emoji usage has risen to add more "flavor" or context to messages/tweets with less words. However, being pictures, it can be difficult to use proper language processing as many emojis can have many meanings. Just for example, the fire emoji often does not literally mean fire or hot but is used to convey that something is attractive or amazing. For our studies, we analyzed how emojis can affect the model's ability to classify a tweet and suicidal or not.

In addition to our emojis analysis, we also took a look at what features of a person makes them more likely to express suicidal thoughts on Weibo (location, gender, reason, etc).

PRIOR WORK

We will be working off of Xiaolei Huang's previous studies on identifying suicidal behavior based on online behavior and their lexicon. The main study we are furthering is Huang's "Using Linguistic Features to Estimate Suicide Probability of Chinese Microblog Users" [1] paper, where he used Linguistic Inquiry and Word Count (LIWC) and Latent Dirichlet (LDA) to extract linguistic features from data. A close tie-in article, also by Huang, "Detecting Suicidal Ideation in Chinese

Microblogs with Psychological Lexicons" [2] will also be used to guide our research. In it, Huang and his colleagues train a post detection model with a psychological lexicon dictionary to analyze linguistic features Our data is the same set that Huang used for his research.

DATA

The data set we will be using was acquired from Xiaolei Huang, who previously used it to make initial predictions. This data includes posts from Sina Weibo, a Chinese Twitter. The data is all from confirmed suicide cases and includes categories like tweet text, location, gender, suicide reasons, likes, reposts, comments, repost thread, devices, and follower count, all formatted in JSON. There are about 2,000,000 total unlabeled data points. Data has been separated according to users; there is one text file for one user for a total of 131 users. An example of one user can be seen in **Figure 1**.

TOOLS

To clean our data and make it suitable for the BERT model and our data visualizations, we used Python's statistical libraries Pandas and Numpy.

We refined Huang's predictions by using the Bidirectional Encoding Representation Transformer (BERT) model to train new model based on our data. Pytorch, a machine learning library, will be used in parallel to implement BERT's packages. Data will be fed into an unaltered BERT model to get initial analyses for later comparisons. Training and development data was used to refine a new BERT model. We will analyze its tokenization, encoding, and modeling part and then find a way to improve its performance.

The BERT model uses General Language Understanding Evaluation (GLUE) as part of its processing. It is responsible for the training and evaluation—it will be altered to match our purposes. Data visualization will be implemented using TensorFlow programs, mostly to visualize accuracy, loss, and learning rate of the BERT model.
For our secondary data visualizations, we used Python's matplotlib library to create our bar graph and circle chart. To create our heatmap of China (based on where suicides were occurring), we used Daniel Piner's implementation of Google's Geocharts API [3] that allows the map to be appropriately shaded according to frequency and province.

DATA PROCESSING

The given data set has a significant amount of extra information that is not relevant to what were are analyzing. We parsed the data in order to match our requirements–the labeled data set for the BERT model focuses mostly on just the context of the tweet and if the creator of the tweet committed suicide or not. For any users where the required data was missing, the data point was dropped from the set.

Many tweets include emojis, which cannot be analyzed in its natural state. When the tweet was first textualized, the emojis were indicated by double spaces and the emoji's alt text (e.g. the candle emoji was replaced with " candle "). We removed and replaced the original emoji text with different values and tested how it affected accuracy.

After isolating and reformatting emojis, the data was randomly grouped into training, development, and testing groups. Training contained 80% of the data, while development and testing each had 10%.

DATA ANALYSIS

One of our ultimate goals is to optimize the estimating model to improve the accuracy of suicide detection. We also plan to form a social network graph and analyze the users' interactions, work which has not been done by Xiaolei Huang yet. We plan to use latent social network relationships to identify suicide individuals and groups. Detecting suicide posts may help us further find some efficient features to improve the performance of suicide prediction.

In one of our tasks regarding to improve the performance of estimating model, the data will be split into three parts: train, dev, and test set. We will use 80% of labeled data to train the suicide-estimating model, 10% labeled data as dev set, and 10% labelled and unlabeled data as test set. As to the social network graph, we will use unlabeled data to analyze.

```
"gender": "女",
"good_at": "美食 旅游 ",
"location": -3181371695984046582,
"num_fans": 998,
"num_follow": 174,
"num_weibo": 44,
"relationship_status": "",
"sexual_orientation": "",
"suicide_age": 0,
"suicide_date": "2014年4月8日",
"suicide_reason": "",
"tags": [
    "大猫座",
    "街拍症候群",
    "美食"
],
"verify_info": "",
"verify_type": "",
"weibo": [
    {
        "category": "retweet",
        "meta": {
            "comment_num": 37,
            "date": "2014-04-08 08:45:02 CST+0800",
            "device": "iPhone客户端",
            "retweet_num": 4,
            "social_network": {
                "comment": {
                    "list": [
                        {
                            "content": "婆娘想你了哈 ....8号才来看你了给你写了封信 你姑姑放在盒子里陪你一起下葬了 你看到了吗［爱你］以前我们最爱互相写信了
                            "date": "2015-03-12 22:45:09 CST+0800",
                            "device": "iPhone 6",
                            "u_name": 2761249386939701771,
                            "uid": -3036833537640179501
                        },
                        {
                            "content": "[蜡烛][蜡烛][蜡烛][蜡烛][蜡烛]",
                            "date": "2014-11-25 22:00:09 CST+0800",
                            "device": "iPhone客户端",
                            "u_name": -6958279530904628443,
                            "uid": 7952046759636134715
                        },
                        {
                            "content": "[泪][泪][泪]今天早上知道的，虽然没有那么多接触，但是知道你是个善良的女孩。想到你坎坷的经历，你的好强，你怎么要这样，遇到
                            "date": "2014-10-31 11:23:03 CST+0800",
                            "device": "iPhone 5s",
                            "u_name": -3803161749601886381,
                            "uid": 5345512903402463736
                        },
```


Figure 1: An example of the data format (only an extract, not complete) from one user

RESULTS

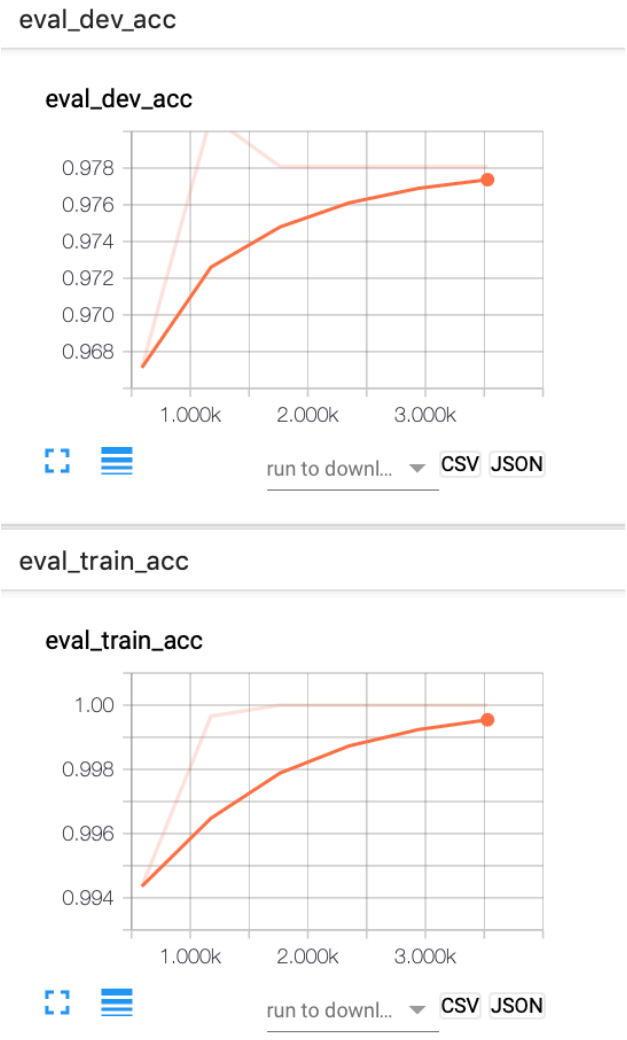

Figure 2.1: BERT accuracy model after development and training without emojis tagged

Wall time	Step	Value
1575500069	588	0.96712327
1575500240	1176	0.9808219075
1575500412	1764	0.97808218
1575500585	2352	0.97808218
1575500757	2940	0.97808218
1575500930	3528	0.97808218

Figure 2.2: Table of the accuracies of untagged data after an *x* amount of iterations

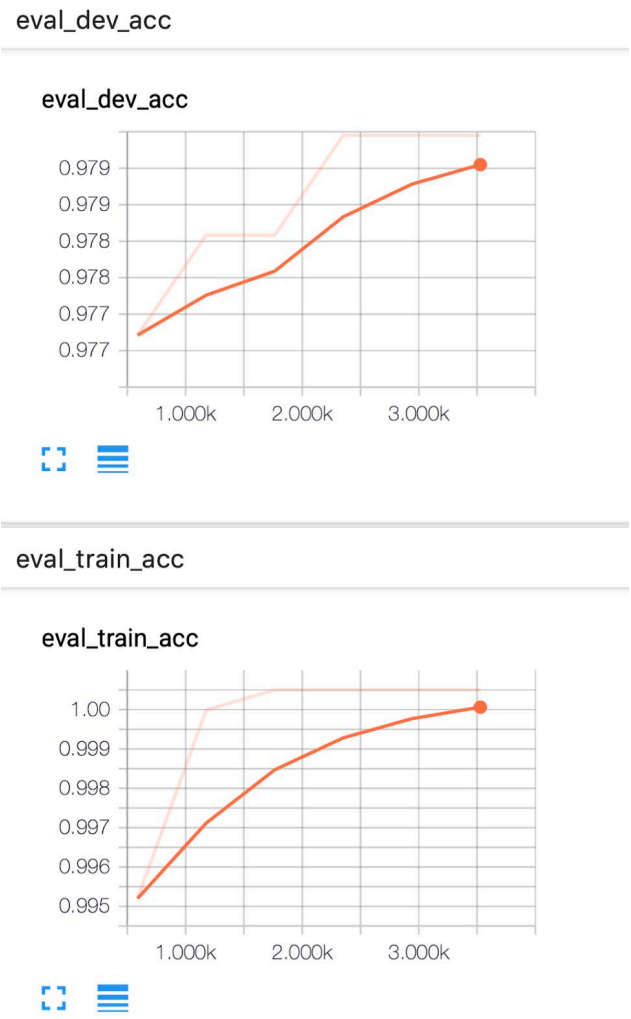

Figure 3.1: BERT accuracy model after development and training with emojis tagged

Wall time	Step	Value
1575509195	588	0.9767123461
1575509368	1176	0.97808218
1575509541	1764	0.97808218
1575509714	2352	0.9794520736
1575509886	2940	0.9794520736
1575510057	3528	0.9794520736

Figure 3.2: Table of the accuracies of tagged data after an *x* amount of iterations

IMPLEMENTATION

To improve the performance of the BERT model, the data will be split into three parts: train, dev, and test set. We will use 80% of data to train the suicide-estimating model, 10% data as development set, and 10% of data as the test set. Different inputs were tested to see how it influences the final accuracy.

Any words that are identified as non-real words (slang, text lingo, etc.) were tokenized by the model. The word is broken up into several substrings and analyzed.

The pure BERT model was compared against our modified BERT model. Based on our newly generated model, we analyzed the interconnectivity of suicidal user interactions. K-fold cross validation was used to assist evaluation.

BERT MODEL DISCUSSION

The BERT model results show a good performance in predicting the users' suicide possibility. In emoji-untagged text, it reached the accuracy at 97.9%, as seen in **Figure 2.1** and **Figure 2.2**. The result with tagged emojis is 98.1% (**Figure 3.1, 3.2**). The features is

learned by the model itself, so it lacks interpretability.

When we compared the results that had emojis tags versus results without emoji tags, we observed the results with emoji tags showed slightly more strength in suicide prediction. The main reason for this difference is some emojis do not express emotions literally, like “tree.” When we replaced the emoji with tags, its influence was limited in the mask layer of the BERT model leading to a more accurate prediction. The subtle improvements may also be due to limited usage of emojis compared to Chinese characters or words.

DATA VISUALIZATION DISCUSSION

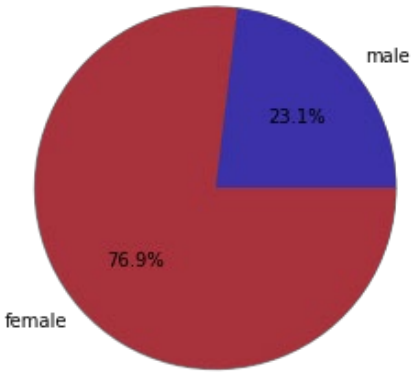

Figure 4: The gender distribution of deceased users

The gender disparity between suicidal users was highlighted in **Figure 4**, where most users were women at 76.4%. According to the World Health Organization (WHO), globally men died by suicide 1.8 times more often in 2017. In the western world, the disparity is even higher: men were more likely to die by suicide three to four times more than females. It is interesting to see a user pool where these statistics are reversed: in our data set, females were 3 times more frequent than males.

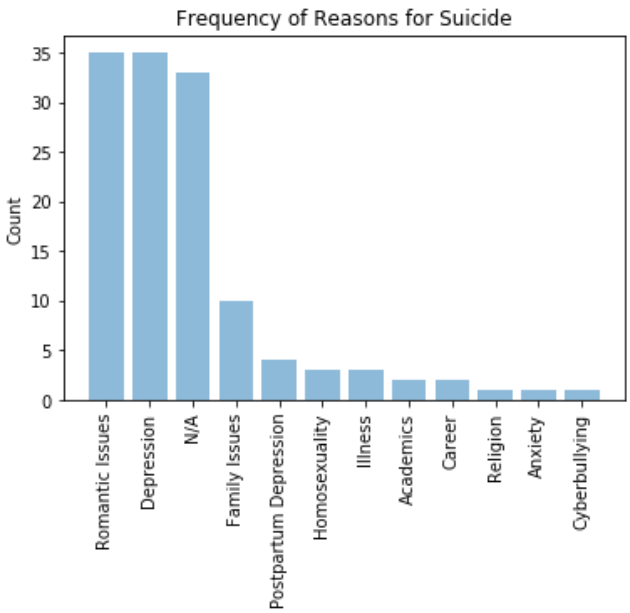

Figure 5: Frequency of identified reasons why users committed suicide

It ties in well with the most frequent reason for suicides in the data set (**Figure 5**): romantic issues and depression, two things that are stereotypically attributed to females. The uneven spread of gender could be attributed to the fact that women usually express their suicidal thoughts more so than men; studies have shown that females show higher rates of suicidal thoughts than men (although men are reported to have a higher rate of successful suicides). As a result, most users that would convey suicidal thoughts are likely to be female. This is support for a troubling issue across the world, where males are unable to freely share their emotions and reach out for help as much as women.

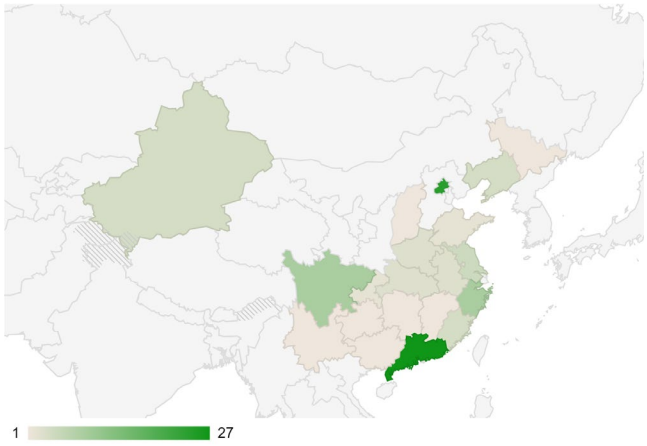

Figure 6: Frequency of users that committed suicide according to the province on their Weibo account

The four provinces with the highest frequencies of suicides are Guangdong (27), Beijing (24), Chongqing (22), Zhejiang (9), and Sichuan (9), which can be seen in **Figure 6**. These results are not too surprising as they mostly correlate with urban centers areas with high populations in China. There are notable provinces with high populations that are missing, such as Shanghai, Henan, and Jiangsu. It could be related to the presence of Western influence that allows users to use apps outside of Weibo to tweet their thoughts; for example, it is no surprise that Hong Kong and Taiwan do not have a heavy presence in the map because, as a result of not being part of China's "Great Firewall," residents are more likely to use Twitter over Weibo.

LIMITATIONS & CHALLENGES

When we ran the data, we found some difficulties. First, when we tokenized non-real words, slang, or unidentified words, these words could potentially interfere with the model. When words are re-tokenized and their influence is reduced, model understands the data set better.

We identified more potential errors.

Firstly, we may have received high results because of limited character usage. The same Chinese characters may have different meanings according to different sentence or word organizations. Moreover, frequently used Chinese characters are limited, so the model did not recognize the difference once their morphological structure was transformed.

Secondly, the small data set may have overfitting problem because the BERT model may have learned too many features.

Lastly, there was duplicate information that may have influenced the results. We have found many users tend to retweet some of the same information, such as advertisements and quotations from others. If this repeat data is split into train and test respectively, it can result in the high accuracy in predictions.

FUTURE WORK

The next step in our work would be to use unique tags for different emojis. Our research showed that including the presence of emojis from the original tweets does increase the accuracy, but we could likely make it more accurate if the emojis' meaning was included. The major obstacle would be correctly describing the emoji—there are already almost ten variants to just the smiley face emoji. It may also be difficult to discern how exactly the emoji is being used, like if it is used literally or sarcastically.

It would also be interesting to take a closer look between the relationship of women and social media. Since there was a disproportionately high amount of female deaths, a closer analysis into how social media might factor into their final decision could be helpful in identifying when help is needed.

Further investigation into the uneven frequencies across urban centers in China could be productive to find out why some cities seem to have more suicidal users than others. It could expose some things about the environments of different cities or weaknesses in data.

CONCLUSION

The BERT model showed that tagged emoji data increased the accuracy of the model, which demonstrates the effect emojis can have on the context of tweets or otherwise short messages. With the spread

of text lingo, emojis are becoming more and more prevalent in messages; natural language processing must begin processing non-language text.

Overall, the research into the data set has shown that an urban female suffering from either romantic issues or depression is most likely to share their suicidal thoughts online. Additionally, our studies are showing a trend that males are less likely to express their suicidal thoughts online, meaning there are less opportunities for people to reach out and perhaps intervene. It's imperative that we reduce the social stigma males have that prevent them from being able to share their emotions.

Suicide is a global problem; it is the 10th leading cause of death worldwide. As such, we hope to continue finding trends that expose risky behavior to identify suicidal individuals before it is too late.

APPENDIX

HONOR CODE

On my honor, as a University of Colorado Boulder student, I have neither given nor received unauthorized assistance. The University of Colorado Honor Code works by receiving the support and participation of all members of the university community. Such an organization is intended to promote a campus culture that consciously upholds the tenets of academic integrity, and moral and ethical conduct. As an international student, it is VERY important to understand CU Boulder's Standards of Academic Integrity.

ACKNOWLEDGMENTS

We are thankful to Xiaolei Huang for providing his data and guidance for our project. We would also like to thank professor Qin Lv, Yichen Wang, and Siddartha Shankar for assisting us in refining this project.

WORK DISTRIBUTION

Tanya implemented the data cleaning and data visualizations in Python.

Changbing processed the cleaned data, parsing through the Weibo post text to identify where emojis occurred and replacing them appropriately to make it fit for the BERT model.

Changbing and Tanya ran the BERT model with the assistance of Jie.

Changbing and Tanya co-wrote the research paper.

REFERENCES

[1] Zhang L., Huang X., Liu T., Li A., Chen Z., Zhu T. (2015) Using Linguistic Features to Estimate Suicide Probability of Chinese Microblog Users. In: Zu Q., Hu B., Gu N., Seng S. (eds) Human Centered Computing. HCC 2014. Lecture Notes in Computer Science, vol 8944. Springer, Cham

[2] X. Huang, L. Zhang, D. Chiu, T. Liu, X. Li and T. Zhu, "Detecting Suicidal Ideation in Chinese Microblogs with Psychological Lexicons," 2014 IEEE 11th Intl Conf on Ubiquitous Intelligence and Computing and 2014 IEEE 11th Intl Conf on Autonomic and Trusted Computing and 2014 IEEE 14th Intl Conf on Scalable Computing and Communications and Its Associated Workshops, Bali, 2014, pp. 844-849.

[3] Pinero, D. (2017, July 12). How to Create a Heat Map of China in Google Geocharts. Retrieved December 1, 2019, from http://www.danielpinero.com/how-to-create-heat-map-china.