Vienna

MEi:CogSci

Middle European interdisciplinary master's programme in Cognitive Science

Detecting Suicide-Related Content from Bereaved Individuals on Twitter – A Machine Learning Approach

Introduction

- **Research showed that suicide rates are associated with** what and how individuals, news agencies and non-governmental organizations write about suicide on **social media**
- Appropriate tone and content may positively influence how people cope with their suicidal thoughts
- To investigate this association it is necessary to **understand what type of social media content relates to an increased or decreased suicide rate**
- Need for large scale studies and automatic categorization of social media content
- The Computational Social Science Lab of Austria established a **novel approach employing a machine learning algorithm** that classifies Tweets into categories such as 'personal reporting', 'celebrity suicide', 'news articles', and 'prevention tweets', among others
- The **existing algorithm can already distinguish six suicide related tweet categories**

Goals

- For this project, **two previously defined categories were selected and a new algorithm that classifies posts into these categories was trained**
- The new categories are supposed to distinguish tweets from **bereaved individuals (individuals who lost a loved one to suicide)** that are either written in a positive tone indicating successful coping, or a negative tone indicating suffering, into the 'positive' or 'negative' category.

Natural Language Processing

Suicide Prevention

Machine Learning

Twitter

TF-IDF

SVM

Methods

First step: in tweets with the word "suicide": keyword search (e.g., "my best friend") in the terminal to identify potential tweets from bereaved individuals

2,6GB CSV

grep "my best friend" all_tweets.csv

Second step: Dataset containing all tweets that include the keyword

3,1MB CSV

Third step: Preprocessing of tweets by lowercasing, cleaning up links, cleaning up @mentions

Fourth step: Adding category labels manually

*@User **i am hurting too...** i just lost my best friend to suicide* 😞💔🙏🏼

*@User i want to rewatch it. my best friend took her life 5 years ago. **#sohelpful***

*my best friend killed himself two years ago after he turned 22.. **it was so hard**. i miss you nathan. </3 https*

*@User: obssesd with @User they really **helped me** with my best friends suicide #awesomedancers #pelevate #fun" love my job*

Fifth step: TF-IDF (Text Frequency - Inverted Document Frequency) translates words into importancy metrics by category

my	best	took	class
0.020	0.001	0.300	**positive**
0.400	0.003	0.100	**negative**

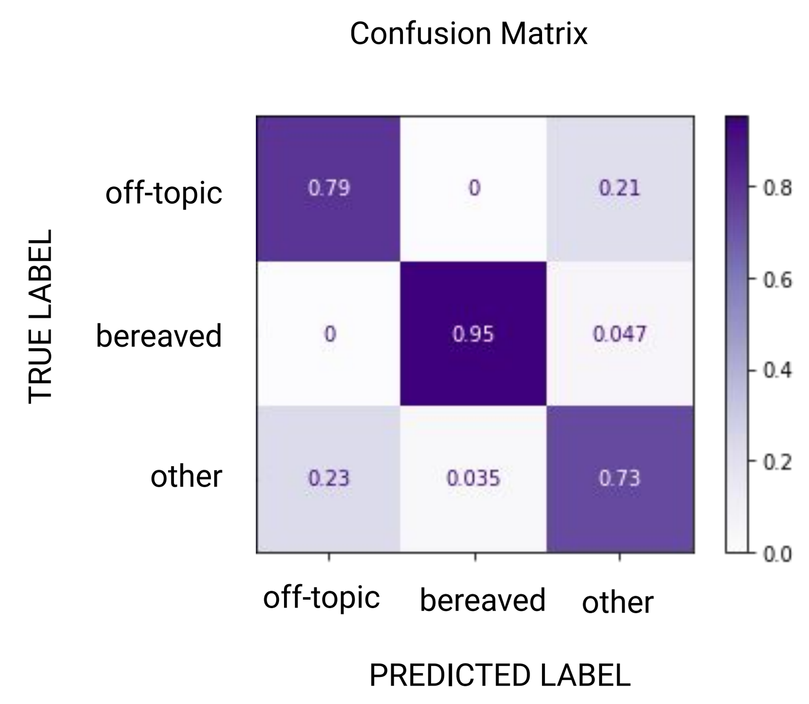
Sixth step: The numeric output from TF-IDF was used to train a support vector machine (SVM) which can then predict categories of new input tweets; GridSearch was used to find best Hyperparameters of the model

```
model = svm.SVC
GS.fit(X_train, y_train)
```

Overall model precision (detected true examples / all true examples) of 60%

Outlook

- Because of the rather poor performance of the classifier, the Computational Social Science Lab of Austria decided that for now an **algorithm which only distinguishes bereaved tweets (label: bereaved) from other suicide related tweets (label: other) and off-topic tweets (label: off-topic)** would be sufficient
- This model achieved a higher **overall precision of 86%**
- The classifier achieved an **even higher detail precision of 93% for the bereaved category**
- The prediction frequencies per category are shown in the confusion matrix:



- To continue this project and improve the classifier, it was suggested to use the deep learning BERT classifier from Google instead of the currently used SVM classifier, because it would not only learn word frequencies, but also consider the context of each word and the syntax of sentences, and can therefore capture more subtle meanings

Conclusion

1. After the first investigation into the data we found that many of the tweets cannot be categorized as positive or negative bereaved stories. We therefore decided to **introduce a third "neutral" category** in the training data
2. The previously defined categories of bereaved negative and bereaved positive tweets are hard to distinguish for an algorithm as well as for a human classifier
3. Due to the third introduced category of neutral bereaved tweets, the training **data sample was not big enough to train the model sufficiently**
4. The **algorithm is likely to be sensitive to the origin of the data** as it was trained with only US tweets which showed some specific country related content ("guns", "veterans", "Nam")
5. Many **bereaved tweets start with direct @mentions**
6. The trained **SVM classifier has a overall precision of 60%**, distinguishing neutral, negative, positive and off-topic tweets from each other. The detailed precision of this model per category for predicting negative and positive tweets was higher with approx. 80%

Contact Details

Anja Huber
a12118890@unet.univie.ac.at
University of Vienna

Supervision
Hannah Metzler
Computational Social Science Lab Austria

GitHub Repository