Master thesis on Intelligent Interactive Systems

Universitat Pompeu Fabra

Suicide on Twitter: banal use and social support

Lara Solà Gallego

Supervisor: Ana Freire

Co-Supervisor: Diana Ramírez-Cifuentes

July 2019

Master thesis on Intelligent Interactive Systems

Universitat Pompeu Fabra

Suicide on Twitter: banal use and social support

Lara Solà Gallego

Supervisor: Ana Freire

Co-Supervisor: Diana Ramírez-Cifuentes

July 2019

Contents

Dedication

I would like to dedicate this work to my grandfather, Enric Solà, because you really were my father. Although you will not be able to read this, I am very grateful for your love and support along the years and above all, for believing in me.

Acknowledgement

I would like to express my sincere gratitude to Ana Freire for the given opportunity to collaborate with her, for her supervision and guidance.

To Diana Ramírez-Cifuentes for helping me in the beginning of this thesis and for always being at my disposal to solve my doubts.

And to my family and boyfriend for all the support and patience.

Abstract

In the past years, social networks like Twitter have had a huge increase in popularity. Users with suicidal ideation often turn to these social networks to express their feelings, seek for help, ask for advice or leave suicide notes. At the same time, other users employ suicide terms as a way to manifest distress for an isolated event, leading to its trivialisation. This thesis has two objectives: (a) build a classifier that distinguishes between users with suicidal ideation and those that banalise about suicide on Twitter; and (b) measure the social support received by users with suicidal ideation. To solve (a) we apply machine learning techniques to the processed texts written by Twitter users to extract a set of features based on topic modelling and sentiment analysis. Moreover, other features from the tweet's structure (e.g: number of followers, retweet count, etc.) are also exploited. Finally, we build a 83% accuracy Random Forest-based classifier which serves as a preliminary approach for further studies. For (b) we propose an engagement measure that takes into account the user's tweet polarity along with the number of favourites and retweets. After applying the measure to a suicide ideation sample and a control group, we find that the suicide ideation users receive a higher amount of favourites, less retweets and an overall higher reaction for negative tweets than the control group.

Keywords: suicide, social media, machine learning.

Chapter 1

Introduction

Suicidal behaviour is a worldwide cause of death and disability. On average more than 800,000 people die by suicide each year, representing an annual global age-standardised suicide rate of 11.4 per 100,000 population (15.0 for males and 8.0 for females). Most notably is that among young people (15 to 29 years of age), suicide accounts for 8.5% of all deaths making it the second leading cause of death (after traffic accidents). Remarkably, in high income countries and in low- and middle-income countries of the South-East Asia Regions, suicide in this age range accounts for 17.6% and 16.6% respectively, representing the leading cause of death for both males and females. Among adults aged 30 to 49 years it accounts for 4.1% of all deaths and is ranked the fifth leading cause of death [1].

Although there is no agreement on the definition of the suicidal stages yet [2], there are three commonly used characterisations: within the domain of suicidal self-directed violence, suicide is defined as death caused by self-directed injurious behaviour with an intent to die as a result of the behaviour; suicide attempt is defined as a non-fatal, self-directed, potentially injurious behaviour with an intent to die as a result of the behaviour even if the behaviour does not result in injury; and suicidal ideation, also known as suicidal thoughts [3], is defined as thinking about, considering, or planning suicide [4].

These definitions lead to a division in two main classes: ideators and attempters.

Studies show that for individuals with suicidal ideation, mental disorders became very weak predictors of suicide attempts [4]. Finding other indicators would be essential to detect and prevent attempts of suicide.

1.1 Motivation and Objectives

In the past years, social networks like Twitter have had a huge increase in popularity. Users often turn to these social networks to express their feelings, seek for help or ask for advice. Twitter currently has 326 million active users and around 500 million messages, or tweets, are sent every day. Since the vast majority of these messages are publicly available, and with the advent of Artificial Intelligence and Big Data, there has been a surge in the interest from companies and scientists to analyse the contents of these messages.

Tweet analysis has been used for a wide variety of topics like stock market movement prediction [5], sentiment analysis [6], disease outbreak prediction [7] or sarcasm detection [8].

Within the context of a suicide prevention project, our interest is to use Twitter to detect cases of users at risk of dying by suicide. Having a lack of strong suicide predictors, Twitter appears as a very promising tool, since the demographics of Twitter users are similar to the aforementioned ones for suicide.

In particular, we want to use Twitter information to learn to differentiate between cases of suicide trivialisation and cases where there is suicidal ideation or a real danger of suicide.

Over the years, suicide threats and attempts have been labelled attention-seeking or manipulative [9][10], contributing to the guilt and shame felt around suicide [11][12][13]. Therefore, the recovery of these patients can be greatly affected by what other people perceive from their experiences and feelings[14][15], which are often expressed through social media content.

Hence, albeit language that is used in an hyperbolic manner to express emotions,

such as in "If I have to listen to another minute of this talk I will slit my wrists", may appear ridiculous outside its context, when uttered under experiences of trauma, psychache or chronic mental illness, the implied connotations are totally different. It is of the utmost importance to be able to tell the difference between flippant language and intentional words [16].

In this sense, two issues are identified within the social media environment: the influence that other people may have on the worsening or recovery of users that manifest suicidal thoughts, and the trivialisation of the usage of terms that refer to suicide. Both of these topics, combined with the great potential of Twitter for screening purposes, have led clinical psychologists to define two main questions:

- Is it possible to develop a tool to effectively distinguish users that trivialise about suicide from those who have suicidal ideation?
- Do users with suicidal ideation receive social media support?

In this thesis we aim to find an answer to these questions by applying Artificial Intelligence algorithms to extract information from the messages that users share on Twitter. Despite the fact that this work uses only information shared on Twitter, the proposed method can be easily extrapolated to other social networks.

1.2 Outline

In Chapter 2, we present the research questions that this thesis aims to answer and the data collection methodology used in order to build our dataset.

Chapter 3 centres on answering whether we can develop a tool to distinguish users that trivialise about suicide from those that have suicidal ideation and is divided in three main parts. It starts with the methodology which includes a preliminary analysis of the data from a topic analysis perspective, the description of the extracted features and the used algorithms along with some performance evaluation guidelines. Followed by the presentation of the results and discussion.

Chapter 4 focuses on the level of social media engagement received by the users with suicidal ideation. First, it introduces the features that are going to be taken into account and continues by defining the engagement measures proposed. The results and discussion are introduced at the end.

Finally, Chapters 5 and 6 comprise the conclusions and future work respectively.

Chapter 2

Proposal

2.1 Research questions

This thesis aims to answer two questions brought up by clinical psychologists during consultations. The novelty of our approach is the use of social media in order to find the answer to these questions:

- **Q1: Are we able to automatically and effectively distinguish between users that trivialise about suicide from those who have suicidal ideation?** We will study how Twitter users make use of suicide related sentences, in order to build a classification system able to distinguish between users with suicidal behaviour and users that trivialise about suicide. For instance, we should differentiate a user who writes content related to their continuous suicidal thoughts from a user who just uses the term suicide in a trivial way (e.g.: *Tomorrow I start my final exams, I'm gonna kill myself*). We will also compare both types of users to the average one, to see how different or close are both profiles in contrast with our random sample of users.

- **Q2: Do users with suicidal ideation have social support?** By considering some social media features, such as the number of likes or retweets received, we will estimate the social support that a user has.

2.2 Methodology for data collection

In this section we will describe the data collection used in order to answer both of our research questions.

The first required step corresponds to obtaining the data from the messages that the users are posting on Twitter. To do so, we will perform two crawlers. The first one aims to find examples of trivialisation (Q1) and the second one of suicidal ideation (Q1 and Q2). We also perform a random collection of user timelines to have a baseline against which to compare the statistical analysis (Q1 and Q2).

We use Twitter's standard Application Programming Interface (API) to retrieve publicly available tweets/timelines by performing the data collection proposed in section 2.2. Notice that this API has certain limitations:

- For every query only the tweets posted in the last 7 days are returned.
- For every user's timeline a maximum of 3,200 most recent tweets are returned.

The crawling process for the cases of trivialisation and suicidal ideation is the following (see Figure 1):

1. Collect tweets by filtering the pertinent keywords.
2. For each tweet collected retrieve its author timeline.
3. Manually annotate the user timeline upon its analysis.

   A user is labelled with suicidal ideation if its timeline clearly displays suicidal thoughts [4], labelled as suicide trivialisation if it shows signs of banalisation, or discarded otherwise.

Finally, the resulting dataset is divided into three categories of users:

Users that trivialise about suicide: in order to gather these user profiles, we will filter tweets using a set of keywords and sentences related to suicide themes

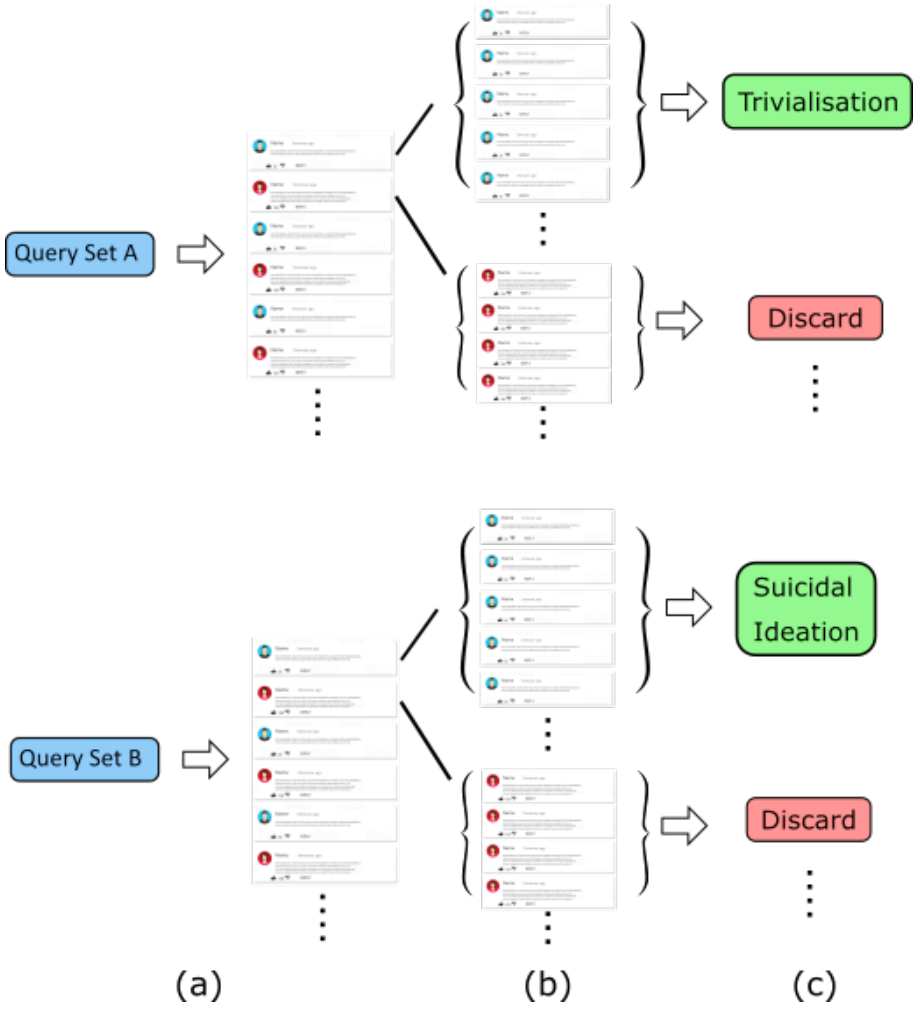


Figure 1: Data collection. (a) Searches according to several queries are made. (b) For each tweet obtained we retrieve the user's timeline. (c) After considering the tweet and the timeline, the user is classified as Trivialisation, Suicidal Ideation, or discarded.

that have been observed on messages where users refer to suicide in a trivial way (e.g.: *kill myself*, *end my life*, etc. the full list can be found on Table 5, Appendix A).

We obtain a total of 4,383 tweets from which we take a random sample of 200 tweets. We extract the timeline for each user and upon inspection we manually label 52 users as suicide trivialisation and 10 users for suicide ideation, 138 are discarded.

Users with suicidal ideation: in this case, the set of keywords has been obtained by mining posts for common terms or phrases where users make reference

to their own suicidal thoughts from a site that has been used by previous studies [17][18] as data source for suicide risk assessment tasks: the subreddit SuicideWatch[1] . Afterwards a clinical psychologist has analysed them to select the most representative ones and also added new ones from his own experience (the full list can be found on Table 6, Appendix A).

We obtain a total of 7,110 tweets from which we take a random sample of 500 tweets, 300 more than in the trivialisation set because the rate of finding a positive of suicidal ideation in this set is lower than finding a positive of banalisation in the trivialisation set. We extract the timeline for each user and upon inspection we manually label 25 users as suicidal ideation, resulting in 475 users discarded.

Random users: for this group we will just use the Twitter[2] API which has a method that returns a small random sample of all public status. Rapid visual inspection completes the process in order to avoid the selection of profiles where users either trivialise or have suicidal ideation.

We obtain a total of 57 users from which we extract their timeline.

The resulting dataset is composed of $N = 52$ users that trivialise about suicide ($\mathcal{T}$), $M = 35$ users with suicidal ideation ($\mathcal{S}$) and $T = 57$ random ($\mathcal{R}$) users. A total of 613 users are discarded because they are news profile accounts, suicide aware institutions, etc.

From the dataset statistics in Table 1 we can observe that users in the Suicidal Ideation group tend to write longer tweets than the rest (average tweet length of 90.80, in contrast with 71.42 and 60.51 for the Trivialisation and Random groups respectively). While there is a clear difference of more than 1,000 tweets between the average number of tweets per user between the Trivialisation (1,3593.12) and Suicide Ideation (1,073.38) groups versus the Random (236.59) group, the Trivialisation group is the one where users tend to post the most.

[1] `https://www.reddit.com/r/SuicideWatch/`
[2] `https://developer.twitter.com/en/docs`

Statistics	Trivialisation (T)	Suicide Ideation (S)	Random (R)
Num users	52	35	57
Num tweets	70,674	377,568	13,486
Avg. num tweets	1,359.12	1,073.38	236.59
Avg. length tweet	71.42	90.80	60.51

Table 1: Main statistics of the dataset

This difference in the average number of tweets per class could indicate that the Random group does not make use of Twitter as a social platform to share their daily life. Also, from the higher average length of the tweets in the suicidal ideation class we can extrapolate that they express more complex ideas or emotions and therefore need more text space.

Chapter 3

Suicide trivialisation

In this chapter we describe our approach for answering the question of whether we are we able to automatically and effectively distinguish between users that trivialise about suicide from those who have suicidal ideation (Q1). To this end we propose a methodology, show the results and expose our discussion.

3.1 Methodology

In this section we want to portray the methodology used in order to answer (Q1). To that end, we want to distinguish between users that trivialise about suicide from users that have suicidal ideation from the extracted data (section 2.2).

To do so, we first analyse the three groups of users obtained from the data collection (i.e.: suicidal ideation, suicide trivialisation and random) to see if there is a real difference in their timelines. Afterwards we extract a series of features that are used as a mathematical representation of the user. Then, we use these features as input of machine learning algorithms to distinguish between both cases and set guidelines to evaluate the performance.

3.1.1 Topic analysis

Before starting with the feature extraction to later classify between users with suicidal ideation and users that banalise about suicide, we analyse the timelines of both groups and the random one to see if we can observe three clear users' profiles.

In order to accomplish this task, we will perform a lexical analysis over each users' timeline's category and view which are the most common topics and emotions displayed on each group.

Empath [19], an open source library whose categories are highly correlated ($r = 0.906$) with similar categories in the Linguistic Inquiry and Word Count (LIWC)[20], is the proposed tool for the above task. It uses the following process to generate word categories:

1. Draw connotations between words and phrases by deep learning neural embedding on around 2 billion words of modern fiction.
2. From these connotations and given a small set of seed words that characterise a category it discovers new related terms.
3. The category is validated by a crowd-powered filter.

The library counts category terms in a text document and has 194 built-in, prevalidated categories generated from common topics like *neglect*, *social media*, *violence*, etc. (the full list of categories can be found in Table 7, Appendix B)

We expect to find distinct presence of the categories showing that the suicidal ideation, suicide trivialisation and random groups display different interests and emotions. In particular, we would like to see a similarity between the suicidal ideation and trivialisation profiles in contrast with the random control group.

3.1.2 Feature Extraction

In order to be used as input to the learning algorithm, we construct a set of features for each user. The details for each feature can be found bellow and they are

summarised in Table 2.

Feature Type	Details and resources	Number of features
Lexical features	Empath	194
Sentiment analysis	VADER	1
Followers	Number of followers	1
Favourite count	Average favourite count	1
Retweet count	Average retweet count	1
Tweet length	Average tweet length	1

Table 2: Extracted features

For this study we will only consider the Suicidal Ideation group and the Suicide Trivialisation group, therefore the number of user timelines being: $M + N = S$.

Lexical features: To characterise each user's timeline, words are classified according to several topics and emotions that represent different semantic fields.

For this task we propose using Empath, the tool introduced in section 3.1.1. We will aggregate Empath's results in the following manner:

Given user u we calculate the word count for each one of Empath's 194 built-in categories as

$$\vec{E}_{u,tweets} = (e_1, e_2, \ldots, e_{193}, e_{194}) \tag{3.1}$$

where e_i is the number of words from category i, resulting in a vector for each user. Since the user timelines vary in terms of number of tweets, a normalised vector is created defined for a user u as

$$\vec{EN}_{u,tweets} = \frac{\vec{E}_{u,tweets}}{\sum_{i=1}^{194} e_i} = \left(\frac{e_1}{\sum e_i}, \frac{e_2}{\sum e_i}, \ldots, \frac{e_{193}}{\sum e_i}, \frac{e_{194}}{\sum e_i} \right) \tag{3.2}$$

which contains the normalised fraction of words presented in each Empath

category. Finally we obtain the lexical feature matrix for all users:

$$C = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_S \end{pmatrix} \tag{3.3}$$

where c_i is the calculated normalised fraction of words $\vec{EN}_{u_i,tweets}$ for each user u_i.

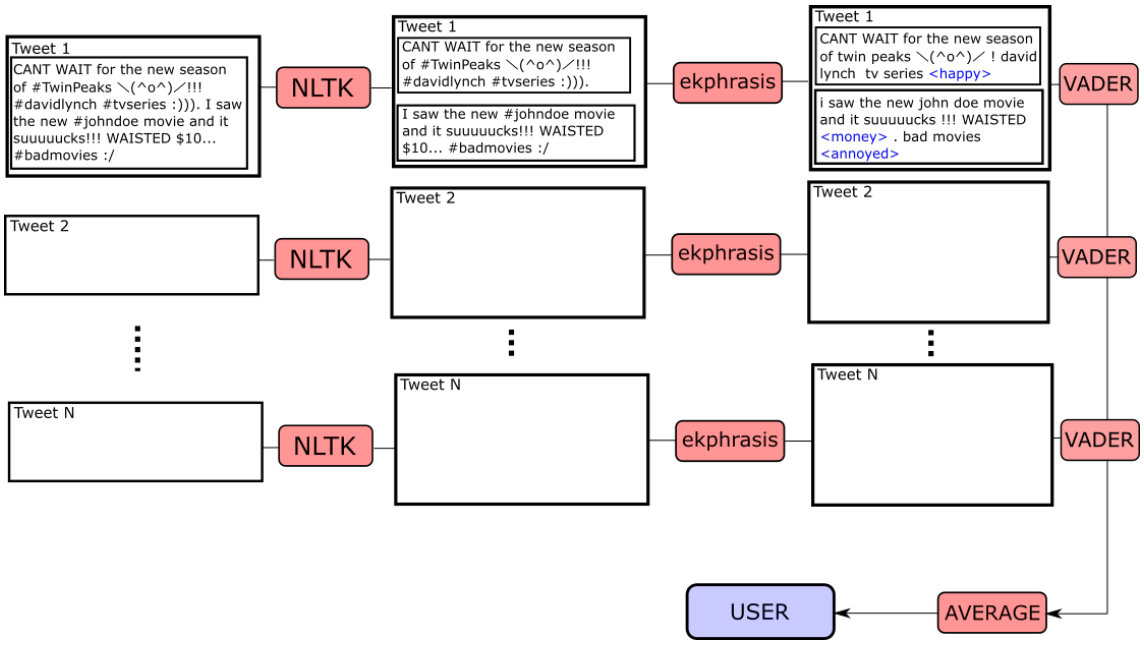


Figure 2: Polarity score feature extraction for a user given its timeline.

Sentiment analysis: Sentiment analysis is the process of computationally identifying and categorising opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral (polarity) [21].

To assign the polarity to each user's tweet, the Valence Aware Dictionary for sEntiment Reasoning (VADER) open source library, a sentiment lexicon designed for social media, is chosen because of its outstanding performance in that domain [22].

The library produces four features corresponding to the negative, positive,

neutral and compound percentages associated to the sentiment polarity of a text.

Since the compound score is a normalised $[-1, +1]$, weighted composite score of the negative, positive and neutral valence scores of each word in the lexicon, this feature is chosen as the sentiment for a given sentence.

As stated in the VADER API demo[1], although VADER can analyse long texts it performs better at the sentence level, because of that, we propose the following process to extract the polarity score of a user's timeline (Figure 2):

1. Use the Natural Language Toolkit (NLTK)[2] package to separate each tweet into sentences.
2. Remove usernames, numbers, urls, etc. and break hashtags into their constituent words (e.g.: #tvseries into *tv series*) by using the Ekphrasis [23] API, a tool geared towards social network messages, which performs tokenization, word normalisation, segmentation and spell correction.
3. Give a tweet's polarity score by averaging the polarity of the tweet sentences.
4. Give a user's polarity score by averaging the polarity of all tweets from its timeline.

The above process produces a polarity vector $\vec{p} = (p_{u_1}, p_{u_2}, \ldots, p_{u_S})$, where p_{u_i} corresponds to the polarity assigned to the user's u_i timeline.

Number of followers: $\vec{w} = (w_{u_1}, w_{u_2}, \ldots, w_{u_S})$, where w_{u_i} corresponds to the number of users following the user u_i, i.e. subscribed to the messages of that user.

Average favourite count: $\vec{f} = (f_{u_1}, f_{u_2}, \ldots, f_{u_S})$, where f_{u_i} is the average of the number of times that each tweet receives a favourite status by another user, i.e. another user indicates that they like that tweet, for the user u_i.

[1] `https://github.com/cjhutto/vaderSentiment`
[2] `https://www.nltk.org`

Average retweet count: $\vec{r} = (r_{u_1}, r_{u_2}, \ldots, r_{u_S})$, where r_{u_i} is the average number of times that each tweet is retweeted by other users for the user u_i. When another user retweets a tweet, he is forwarding it to their own followers.

Average tweet length: $\vec{l} = (l_{u_1}, l_{u_2}, \ldots, l_{u_S})$, where l_{u_i} is the average number of characters per tweet for the user u_i.

Traditionally, the maximum number of characters per tweet was limited to 140. However on the year 2019, this limit was doubled to 280. It is worth mentioning though, that the average message length in the platform is 33 characters and it did not change after the limit extension[3].

3.1.3 Learning Algorithms

In this section we will present two supervised learning algorithms Support Vector Machines (SVM) and Random Forest (RF).

Support Vector Machines: The SVM algorithm is a non-probabilistic binary linear classifier [24]. It aims to map the labelled training data into a higher dimensional input space and construct an optimal separating hyperplane in that space. New examples are mapped into that same space and predicted to belong to a category based on which side of separating hyper-plane they fall (see Figure 3).

Thanks to the kernel trick [25] SVMs can also perform non-linear classification by implicitly mapping the inputs to high-dimensional feature spaces.

[3]Techcrunch article.

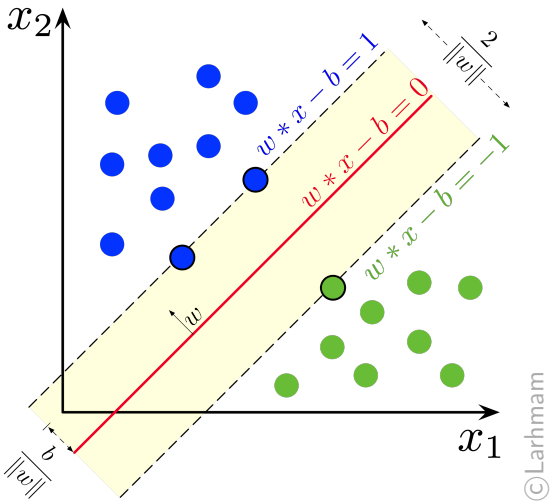


Figure 3: Maximum-margin hyperplane and margins for an SVM trained with samples from two classes. Samples on the margin are called the support vectors.

Random Forest: Random Forests are an ensemble learning [26] method for classification and regression that generates a collection of decision trees at training time [27].

A decision tree has a tree-like graph structure where the leaves represent the class labels and the branches represent the decisions that lead to those classes [28].

Because decision trees tend to overfit [27], Random Forests are used to improve the trade-off between the variance and bias by outputting the class that is the mode of the classes (classification, see Figure 4) or mean prediction (regression) of the individual trees [29][30].

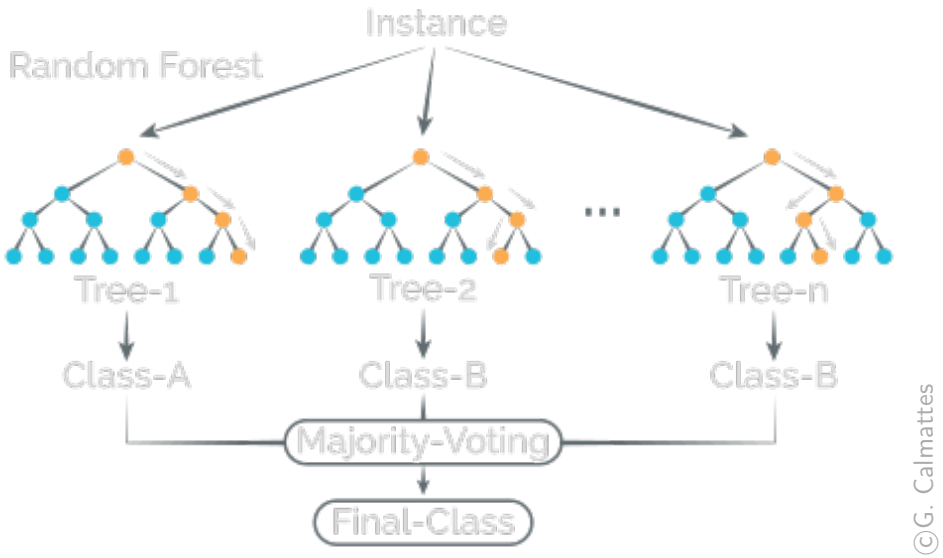


Figure 4: Random Forests classification system simplified.

3.1.4 Evaluation

In order to evaluate the recognition algorithms we will explain some terms to better understand the results:

True Positives (TP): Equivalent to hits, the algorithm predicted correctly a class.

True negatives (TN): Equivalent to rejection, the algorithm correctly predicted that a class X was not a class Y.

False positives (FP): Equivalent to false alarm, the algorithm predicted that a class was X when it actually was Y.

False negatives (FN): Equivalent to miss, the algorithm predicted that a class was not X when it actually was X.

Accuracy: How often the algorithm predicts correctly.

$$Accuracy = \frac{TP + TN}{Total} \tag{3.4}$$

Recall: Also known as sensitivity or true positive rate (TPR), it measures the rate

of hits.

$$Recall = \frac{TP}{TP + FN} \tag{3.5}$$

Precision: Also known as positive predictive value (PPV), it measures how often the prediction is correct.

$$Precision = \frac{TP}{TP + FP} \tag{3.6}$$

F_1-measure: A measure that combines precision and recall is the harmonic mean of precision and recall:

$$F_1\text{-measure} = 2 \cdot \frac{PPV \cdot TPR}{PPV + TPR} \tag{3.7}$$

Confusion matrix: Table that allows visualisation of the recognition algorithm. Each column of the matrix represents the instances in a predicted class while each row represents the instances in an actual class. It helps in seeing if the system is mislabelling two classes.

3.2 Results

The experiments are conducted over the dataset obtained in section 2.2.

In order to answer (Q1) we follow the steps described below:

1. We perform a topic analysis over the whole dataset (Suicide Ideation, Banalisation and Random users).

2. With just the Suicide Ideation and Banalisation users from the dataset, we perform the following actions:

   (a) We inspect the features importance by taking advantage of the calculations performed by decision trees ensemble methods to determine the split that will help the most at distinguishing between classes. To do so, we use a method from the *scikit-learn*[31] Python library that returns an array of each feature's importance in determining the splits.

   (b) We classify both groups using Support Vector Machines (SVM) and Random Forest (RF) from the *scikit-learn* Python library.

   Since we have a reduced number of samples, 87 in total between both classes, we use a 10-fold cross validation approach in the evaluation of our classification experiments. This approach trains iteratively the classifier on 90% of the training data and tests on the remaining 10%. After 10 iterations, the results are calculated by taking the mean accuracy across all models. We also perform a grid search to optimise the parameters for each classification model.

3.2.1 Topics analysis

First, we compare individually the Suicidal Ideation ($\mathcal{S}$) and Banalisation/Trivialisation ($\mathcal{T}$) classes against the Random ($\mathcal{R}$) class. Second, we talk about the main differences between $\mathcal{S}$ and $\mathcal{T}$ classes and $\mathcal{R}$ class. Finally, we study the similarities between $\mathcal{S}$ and $\mathcal{T}$ classes.

Figure 5 shows the summary of the top fifteen categories for each class ($\mathcal{S}$, $\mathcal{T}$ and $\mathcal{R}$) aggregated, resulting in a final total number of 26 categories given that for some classes the topics do not overlap (e.g: death only appears for the suicidal ideation's top fifteen).

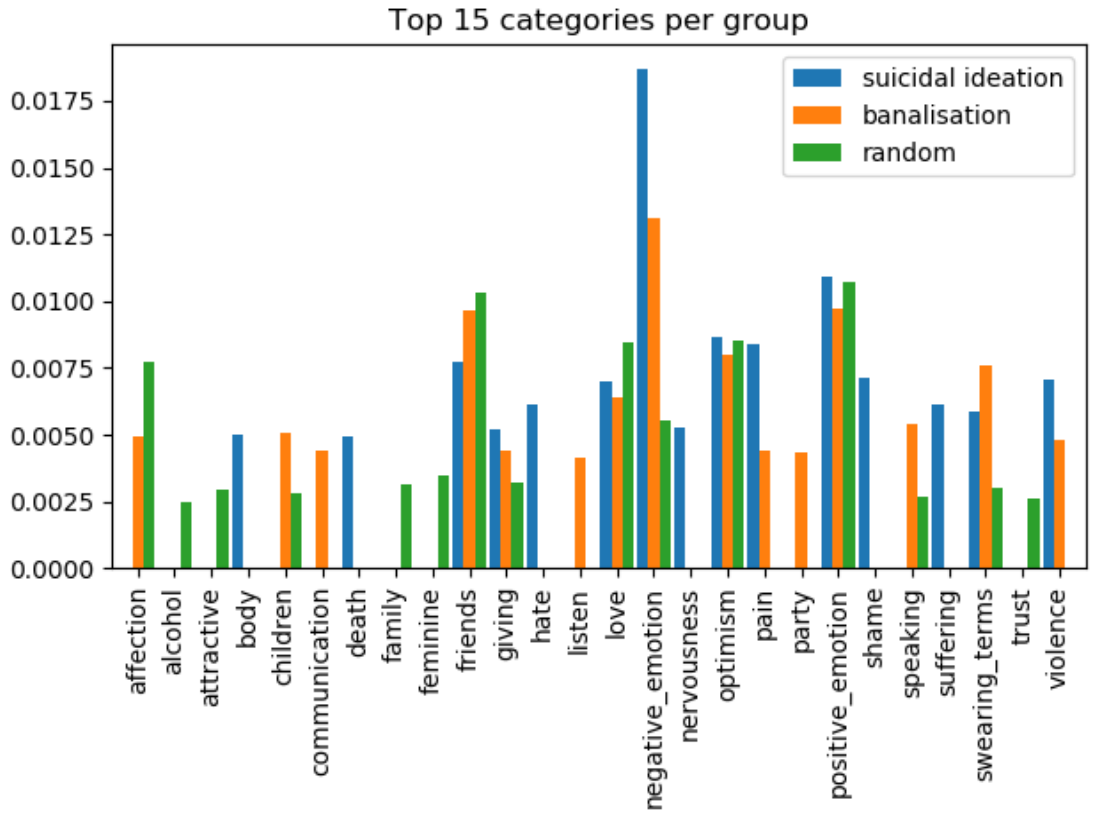


Figure 5: Depicts the 15 categories that appear the most in each class (suicidal ideation ($\mathcal{S}$), banalisation ($\mathcal{T}$) and random ($\mathcal{R}$)), resulting in 26 categories, given that some classes do not share categories in their top fifteen (e.g: family only appears for the random class).

Suicide Banalisation versus Random: The terms related to speaking and children appear more in the Banalisation class than the Random class. While affection has a higher occurrence on the Random class. Also the terms communication, listen and party do not appear on the top fifteen of the Random class.

Suicidal Ideation versus Random: Suicidal Ideation looks like the class that deviates the most from the Random class, with 6 categories related to negative

emotions and terms: hate, death, shame, nervousness, suffering and body not present in the Random class' top fifteen.

It is clearly visible how Suicidal Ideation and Suicide Trivialisation users display a greater use of negative emotion and swearing terms than the Random class. Pain and violence are part of $\mathcal{S}$ and $\mathcal{T}$ top fifteen while they do not appear on the Random class. On the other hand, the topics family, feminine, alcohol, attractive and trust which do pertain to the Random's top do not appear on the $\mathcal{S}$ and $\mathcal{T}$'s.

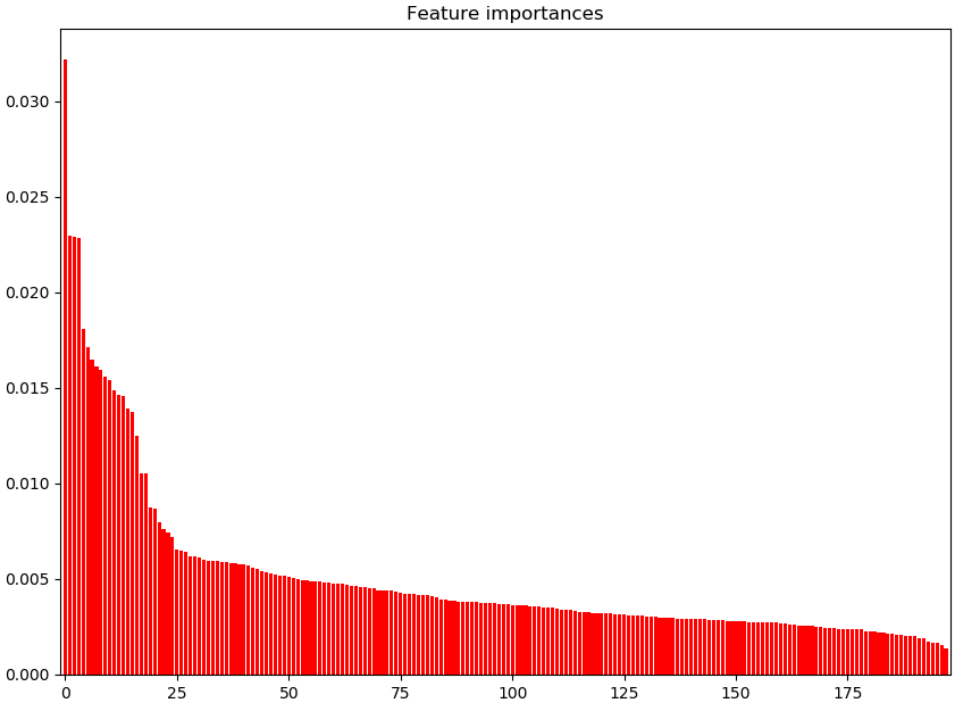
The Suicide Ideation and Suicide Trivialisation classes look very similar, of particular interest is the overall negativity of the $\mathcal{S}$ class:

- it has higher negative emotion, pain and violence terms than the $\mathcal{T}$ class.
- the terms death, hate, nervousness, shame, suffering do not appear on the top fifteen of the $\mathcal{T}$ class.

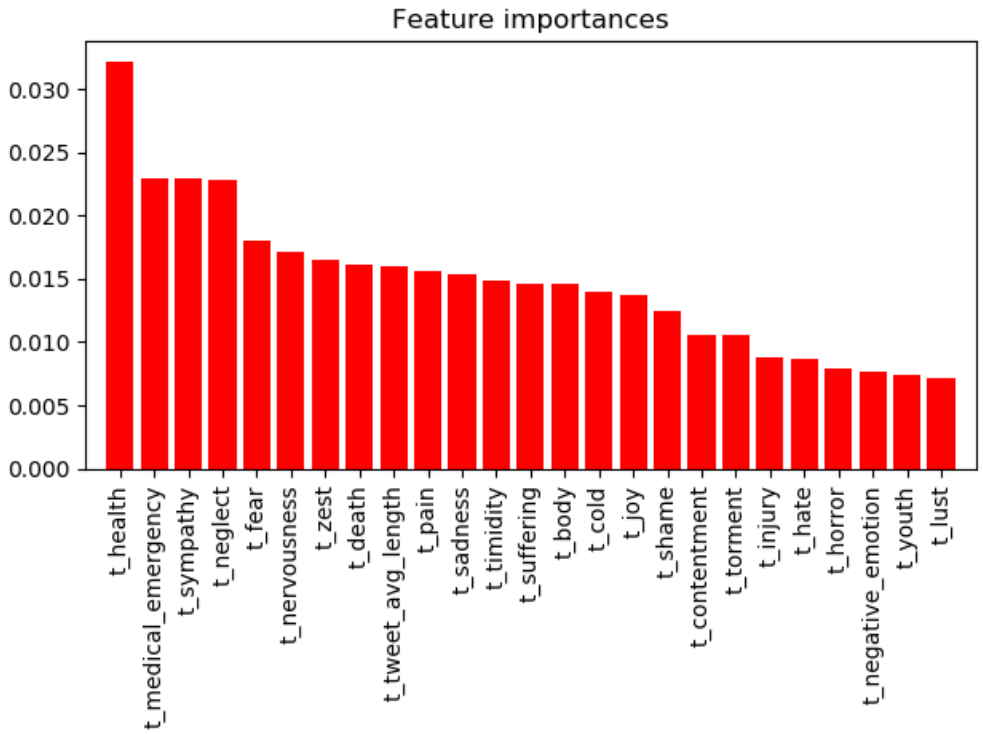
Body-related terms also do not appear on the $\mathcal{T}$ class, but although we could infer that they have a negative connotation in relation with the rest of the terms associated to the class, we can not be sure just by the current data. The complementary terms from the $\mathcal{T}$ class do not have a negative connotation, they include: affection, children, communication, listen, party and speaking.

3.2.2 Feature importance

From Figure 6a it can be seen that the most important features comprise just a 13% of the total, taking the first 25 as the most relevant ones. Figure 6b focuses on these 25 first features: the average length of a tweet and some lexical categories obtained from Empath. The topics with higher importance are health, medical emergency, sympathy, neglect, fear and nervousness, all of them representing emotions or categories that are usually present in subject's with suicidal ideation.



(a) All feature importances: x axis only represents the feature number, not its index.



(b) Top 25 feature importances. All of them are categories' features obtained from Empath but one, the average length of a tweet.

Figure 6: Feature importances.

3.2.3 Classification models

After performing 10-fold cross validation and grid search over the dataset composed only by the Banalisation and Suicide Ideation classes over the set of features extracted from 3.1.2 (i.e.: lexical features, sentiment analysis, followers, favourite count, retweet count, average tweet length) we obtain the following parameters for our models:

- **Support Vector Machines:** linear kernel and $C = 1000$ (penalty parameter of the error term).
- **Random Forest:**

  bootstrapping: bootstrap samples are used to build the trees.

  entropy: the function to measure the quality of a split.

  no max depth: nodes are expanded until all leaves are pure or until all leaves contain less than 10 samples.

  maximum features: 10 features will be considered for the best split.

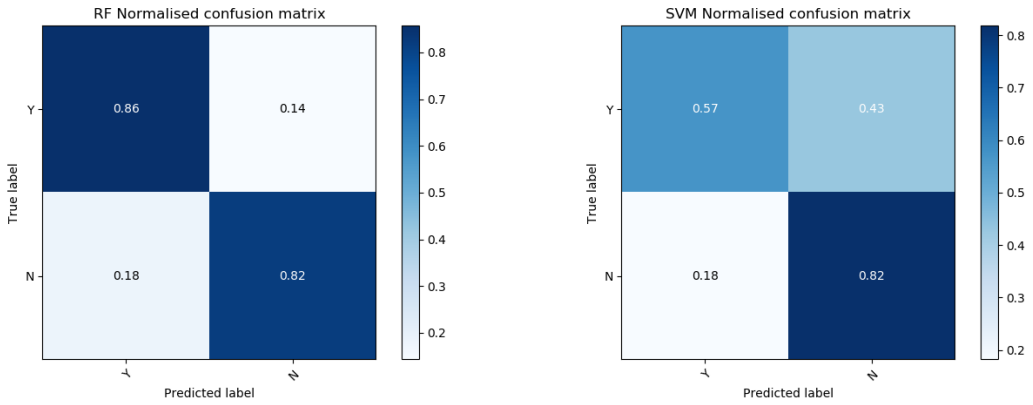
  minimum split samples: 10 minimum samples required to split an internal node.

We evaluate both SVM and RF algorithms having 80% of the data as training and the remaining 20% as test. We obtain 72% and 83% accuracy for the SVM and RF models respectively.

In Table 3 we have a summary of the measures for each class and classifier method, in particular, we can observe recall being 17% higher in the Suicidal Ideation class for RF than SVM. Overall the RF model does a better job at correctly classifying each class. Figure 7 shows the classification performed by both models.

Classifier	Class	Precision	Recall	F_1-measure
SVM	Suicidal Ideation	0.71	0.69	0.70
	Banalisation	0.75	0.82	0.78
RF	Suicidal Ideation	0.75	0.86	0.80
	Banalisation	0.90	0.82	0.86

Table 3: Classifiers' results by class



(a) Random Forest (b) Support Vector Machines

Figure 7: Depicts the confusion matrices for RF and SVM. Labels: {Y = Suicide Ideation, N = Banalisation}. As can be seen on (a) RF outperforms SVM on the task of correctly classifying the Suicide Ideation class.

3.3 Discussion

In this section we will start by discussing the results from the tests performed in order to answer (Q1)

From the topic analysis we observe that suicidal ideation and banalisation users have similar profiles, than the random group, probably coming from their higher engagement on Twitter. They also share a higher display of negative emotions and swearing terms which would be expected from both groups, from the suicidal ideation class, because of their suffering, and from the banalisation class because they probably make a lot of use of irony and flippant vocabulary. The suicidal ideation group displays, though, more emotions related to pain and violence, probably related to the extensive communication of their psychache.

On the feature importance analysis we observe that the average tweet length paired with the features related to the categories that have a correlation with suicidal ideation, either because of the emotions displayed or the consequences of such a condition, are the ones with the higher importance: health, medical emergency, emergency sympathy, neglect, feat, etc. Indicating that the rest of features like polarity, number of followers, categories non related to suicidal ideation, etc. are not clear discriminators between Suicidal Ideation and Suicide Banalisation users. Regardless, given the small size of the sample used in the study we keep all the features for the classification evaluation.

From the models obtained in section 3.2.3 we see that the classification of both classes, suicidal ideation and banalisation, with the proposed features is possible and that RF would be a better choice than SVM, given its higher overall performance and the better ability to discriminate suicidal ideation profiles (higher recall for this class).

Chapter 4

Social support

In this chapter we describe our approach for answering the question of whether users with suicidal ideation receive social support (Q2). To this end we propose a methodology, show the results and expose our discussion.

4.1 Methodology

In this section we will describe the measures that will help us to answer (Q2). First, we start by portraying the features that will be later used to define the scores to analyse the user's social support. Second, we provide a list of the definitions and scores.

4.1.1 Feature Extraction

For this study we will only consider the Suicidal Ideation group and the Random group, therefore the number of user timelines being: $M + T = F$.

In order to perform a statistical analysis we extract a set of features for each user. The details for each feature can be found below.

We select the last tweets (in descending chronological order) posted on a user's timeline, obtaining a tweet sample $s = (t_1, t_2, \ldots, t_Z)$, where Z is the sample size, for each user.

1. Given a tweet t_i we extract the following features:

   Sentiment analysis (p_i): Polarity obtained by applying steps 1) to 3) from the sentiment analysis feature extraction (section 3.1.2) to t_i.

   Number of favourites (f_i): Number of times that t_i receives a favourite by another user.

   Number of retweets (r_i): Number of times that t_i is retweeted by other users.

2. From the above features we obtain three vectors per user:

   Tweets polarity: $p = (p_1, p_2, \ldots, p_Z)$, where $p_i \in [-1, 1]$

   Tweets favourites: $f = (f_1, f_2, \ldots, f_Z)$

   Tweets retweets: $r = (r_1, r_2, \ldots, r_Z)$

3. We also normalise the favourites (f) and retweets (r) vectors:

$$f_{norm} = \frac{f - \min f}{\max f - \min f} \tag{4.1}$$

$$r_{norm} = \frac{r - \min r}{\max r - \min r} \tag{4.2}$$

4.1.2 Scores and definitions

In order to analyse the social support a user receives we use the following definitions and measures:

Total number of tweets (Z): Z = sample size of last tweets retrieved from the user's timeline.

Total number of Favourites (F):

$$\mathrm{F} = \sum_{i=1}^{Z} f_i \tag{4.3}$$

Total number of ReTweets (RT):

$$\mathrm{RT} = \sum_{i=1}^{Z} r_i \tag{4.4}$$

Favourites Ratio (FR):

$$\mathrm{FR} = \frac{\mathrm{F}}{Z} \tag{4.5}$$

Retweets Ratio (RR):

$$\mathrm{RR} = \frac{\mathrm{RT}}{Z} \tag{4.6}$$

Positive tweet: We consider a tweet t_i positive when its polarity is: $p_i \geq 0$.

Negative tweet: We consider a tweet t_i negative when its polarity is: $p_i < 0$.

Total number of Positive Tweets (PT): Number of tweets with $p_i \geq 0$

Total number of Negative Tweets (NT): Number of tweets with $p_i < 0$

Positive Tweets Ratio (PTR):

$$\mathrm{PTR} = \frac{\mathrm{PT}}{Z} \tag{4.7}$$

Negative Tweets Ratio (NTR):

$$\mathrm{NTR} = \frac{\mathrm{NT}}{Z} \tag{4.8}$$

Positive Favourites Engagement (PFE):

$$\mathrm{PFE} = \frac{\sum_{i=1}^{Z} p_i f_{norm_i} [p_i \geq 0]}{Z} \tag{4.9}$$

Negative Favourites Engagement (NFE):

$$\mathrm{NFE} = \frac{\sum_{i=1}^{Z} p_i f_{norm_i} [p_i < 0]}{Z} \tag{4.10}$$

Favourites Engagement (FE):

$$\mathrm{FE} = \frac{\sum_{i=1}^{Z} p_i f_{norm_i}}{Z} \tag{4.11}$$

Positive Retweets Engagement (PRE):

$$\mathrm{PRE} = \frac{\sum_{i=1}^{Z} p_i r_{norm_i} [p_i \geq 0]}{Z} \tag{4.12}$$

Negative Retweets Engagement (NRE):

$$\mathrm{NRE} = \frac{\sum_{i=1}^{Z} p_i r_{norm_i} [p_i < 0]}{Z} \tag{4.13}$$

Retweets Engagement (RE):

$$\mathrm{RE} = \frac{\sum_{i=1}^{Z} p_i r_{norm_i}}{Z} \tag{4.14}$$

Positive Engagement (PE):

$$\mathrm{PE} = \frac{\mathrm{PFE} + \mathrm{PRE}}{2} \tag{4.15}$$

Negative Engagement (NE):

$$\mathrm{NE} = \frac{\mathrm{NFE} + \mathrm{NRE}}{2} \tag{4.16}$$

Engagement Score (ES):

$$\mathrm{ES} = \frac{\mathrm{FE} + \mathrm{RE}}{2} \tag{4.17}$$

4.2 Results

The experiments are conducted over the dataset obtained in section 2.2.

For this study we calculate an average for the classes Suicidal Ideation and Random of the engagement scores defined in section 4.1.2 (see Table 4).

Scores	Suicidal Ideation	Random
Favourites Ratio (FR)	2.2447	1.1138
Retweets Ratio (RR)	0.2893	0.5965
Positive Tweets Ratio (PTR)	0.6502	0.8619
Negative Tweets Ratio (NTR)	0.3498	0.1381
Positive Favourites Engagement (PFE)	1.6815	0.6961
Negative Favourites Engagement (NFE)	-1.5848	-0.3338
Favourites Engagement (FE)	0.0967	0.3623
Positive Retweets Engagement (PRE)	0.4706	0.3570
Negative Retweets Engagement (NRE)	-0.4218	-0.1324
Retweets Engagement (RE)	0.0488	0.2246
Positive Engagement (PE)	1.0760	0.5265
Negative Engagement (NE)	-1.0033	-0.2331
Engagement Score (ES)	0.0728	0.2934

Table 4: Engagement scores

Suicidal ideation: Since the ES is close to 0 (0.0728) users receive a similar engagement for positive and negative tweets. The FR is eleven times the RR and the PTR is higher than the NTR.

Random: Here the ES of 0.2934 indicates a higher engagement for positive tweets than negative tweets. The FR doubles the RR and the PTR is higher than the NTR, this last being very low in comparison with the suicidal as well.

Overall, all positive engagement measures for the Random group are comparatively higher in respect to their negative counterparts than in the Suicidal ideation sample. Also, Suicidal ideation users have a higher FR than the Random group, while the RR is higher in the Random group.

4.3 Discussion

In this section we will discuss the results from the tests performed in order to answer (Q2).

From Table 4 we can observe that Suicidal Ideation users receive more likes per tweet than the Random group while on the former higher retweets rates are displayed. This could be explained by the fact that the tweets of the suicide ideation group are personal or suicide related, so people can favourite them but they do not feel like there is a reason to retweet them.

Also, as expected, Suicidal ideation users have more negative tweets than the random group, but receive the same overall amount of likes/retweets for both negative and positive posts, unlike the random group, which produces less engagement when tweeting negative posts. This reinforces the idea that Suicidal ideation accounts might receive positive engagement from the twitter community, in the form of likes and retweets, for their negative tweets in order to give moral support to these users in distress.

Chapter 5

Conclusions

5.1 Q1: Suicide trivialisation

In order to answer the question of being able to automatically and effectively distinguish between users that trivialise about suicide and those that have suicidal ideation, we have started by performing a topic analysis over three user groups: Suicidal Ideation, Banalisation and Random (our control group), from this study we have determined that while the Suicide Ideation and Banalisation profiles are similar, there are definitely distinguishable traits indicating that classification is possible.

Secondly, we have performed a feature importance analysis to see which of the proposed features for classification are the most relevant; from this inspection we have concluded that features related to categories associated to emotions or consequences to suicidal ideation and the average length of tweets are the most significant.

Lastly, we have build two classification models, Random Forest (RF) and Support Vector Machines (SVM), from which we have ruled out the use of SVM to solve this problem because of its difficulty in classifying the Suicidal Ideation class. We resolve that RF performs relatively well given the small size of the dataset (F_1-measure = 0.83).

We conclude that it is effectively possible to build a classification system able to distinguish between Suicidal Ideation and Suicide Banalisation users.

5.2 Q2: Social support

To solve the question about whether users with suicidal ideation receive social media support, we have proposed different measures of engagement based on the number of favourites or retweets received depending on the positivity or negativity of the tweet. From the results of this measures applied to both the Suicide Ideation and Random groups we have concluded that although the number of retweets in the Suicide Ideation class is less than in the Random group, the Suicidal Ideation class receives substantially more tweets or favourites for negative tweets than the random class, indicating that for those negative tweets, users with suicidal ideation do receive, at least, a higher response from the Twitter community.

Chapter 6

Future work

6.1 Q1: Suicide trivialisation

While we have obtained some interesting and promising results, future research is needed in order to improve them, since they have been conducted on an limited size set of annotated Twitter user's timelines. In fact, even if we started from a relatively large dataset, the user's timelines classified as containing suicidal ideation would not appear to be included in large percentages because of the inherent characteristics of this type of users and content. Therefore, the first limitation has been the magnitude of our dataset which could be improved by using the classification tool we have developed to obtain a larger dataset, followed by performing a manual inspection afterwards.

Other lines of work would include the use of the classification algorithm as a second filter for suicide ideation detection, i.e.: using a more general classifier to distinguish between suicide ideation users and the rest, and then applying our model to the users that have been classified as positive for suicide ideation, in order to reduce False Positives. To that end, new features and learning algorithms could be tested.

6.2 Q2: Social support

Since we have seen that users with Suicidal Ideation receive more engagement than the average user for their negative tweets, future work would include the analysis of each reply, especially those on negative tweets to be able to discern whether the social media support is positive or negative.

Other improvements would include an analysis of the users' followers and people they follow to see if there is a community with similar ideation, and analyse the type of interactions they have.

List of Figures

List of Tables

Frequently Used Acronyms

NLTK Natural Language Toolkit

VADER Valence Aware Dictionary for sEntiment Reasoning

API Application Programming Interface

SVM Support Vector Machines

RF Random Forest

TP True Positives

F Total number of Favourites

RT Total number of ReTweets

FR Favourites Ratio

RR Retweets Ratio

PT Total number of Positive Tweets

NT Total number of Negative Tweets

PTR Positive Tweets Ratio

NTR Negative Tweets Ratio

PFE Positive Favourites Engagement

NFE Negative Favourites Engagement

FE Favourites Engagement

PRE Positive Retweets Engagement

NRE Negative Retweets Engagement

RE Retweets Engagement

PE Positive Engagement

NE Negative Engagement

ES Engagement Score

Bibliography

[1] Organization, W. H. *et al. Preventing suicide: A global imperative* (World Health Organization, 2014).

[2] Goodfellow, B., Kõlves, K. & de Leo, D. Contemporary Definitions of Suicidal Behavior: A Systematic Literature Review. *Suicide and Life-Threatening Behavior* **49**, 488–504 (2019).

[3] Gliatto, M. F. & Rai, A. K. Evaluation and treatment of patients with suicidal ideation. *American family physician* **59**, 1500–1506 (1999).

[4] Klonsky, E. D., May, A. M. & Saffer, B. Y. Suicide, Suicide Attempts, and Suicidal Ideation. *Annual Review of Clinical Psychology* **12**, 307–330 (2016).

[5] Bollen, J., Mao, H. & Zeng, X. Twitter mood predicts the stock market. *Journal of Computational Science* **2**, 1–8 (2011).

[6] Go, A., Bhayani, R. & Huang, L. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford* **1**, 2009 (2009).

[7] St Louis, C. & Zorlu, G. Can Twitter predict disease outbreaks? *BMJ : British Medical Journal* **344**, e2353 (2012).

[8] González-Ibáñez, R., Muresan, S. & Wacholder, N. Identifying sarcasm in twitter: A closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, 581–586 (Association for Computational Linguistics, Stroudsburg, PA, USA, 2011).

[9] Canetto, S. S. Meanings of gender and suicidal behavior during adolescence. *Suicide and Life-Threatening Behavior* **27**, 339–351 (1997).

[10] Canetto, S. S. & Lester, D. Gender, culture, and suicidal behavior. *Transcultural Psychiatry* **35**, 163–190 (1998).

[11] Cvinar, J. G. Do suicide survivors suffer social stigma: a review of the literature. *Perspectives in psychiatric care* **41**, 14–21 (2005).

[12] Minois, G. *History of suicide: Voluntary death in western culture.* (American Psychological Association, 2001).

[13] Sudak, H., Maxim, K. & Carpenter, M. Suicide and stigma: a review of the literature and personal reflections. *Academic Psychiatry* **32**, 136–142 (2008).

[14] Maple, M., Edwards, H., Plummer, D. & Minichiello, V. Silenced voices: hearing the stories of parents bereaved through the suicide death of a young adult child. *Health & social care in the community* **18**, 241–248 (2010).

[15] Kreitman, N., Smith, P. & Tan, E.-S. Attempted suicide as language: An empirical study. *The British Journal of Psychiatry* **116**, 465–473 (1970).

[16] McKay, K. *et al.* Sticks and stones: How words and language impact upon social inclusion (2015).

[17] Shing, H.-C. *et al.* Expert, crowdsourced, and machine assessment of suicide risk via online postings. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, 25–36 (2018).

[18] Zirikly, A., Resnik, P., Uzuner, O. & Hollingshead, K. Clpsych 2019 shared task: Predicting the degree of suicide risk in reddit posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, 24–33 (2019).

[19] Fast, E., Chen, B. & Bernstein, M. S. Empath: Understanding Topic Signals in Large-Scale Text. In *Proceedings of the 2016 CHI Conference on Human*

Factors in Computing Systems, CHI '16, 4647–4657 (ACM, New York, NY, USA, 2016).

[20] Pennebaker, J. W., Boyd, R. L., Jordan, K. & Blackburn, K. The development and psychometric properties of liwc2015. Tech. Rep. (2015).

[21] Aggarwal, C. C. & Zhai, C. A Survey of Text Clustering Algorithms BT - Mining Text Data. 77–128 (Springer US, Boston, MA, 2012).

[22] Hutto, C. J. & Gilbert, E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media* (2014).

[23] Baziotis, C., Pelekis, N. & Doulkeridis, C. Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 747–754 (2017).

[24] Cortes, C. & Vapnik, V. Support-vector networks. *Machine learning* **20**, 273–297 (1995).

[25] Boser, B. E., Guyon, I. M. & Vapnik, V. N. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, 144–152 (ACM, 1992).

[26] Opitz, D. & Maclin, R. Popular ensemble methods: An empirical study. *Journal of artificial intelligence research* **11**, 169–198 (1999).

[27] Breiman, L. Random forests. *Machine learning* **45**, 5–32 (2001).

[28] Quinlan, J. R. Induction of decision trees. *Machine learning* **1**, 81–106 (1986).

[29] Ho, T. K. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, vol. 1, 278–282 (IEEE, 1995).

[30] Tin Kam Ho. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20**, 832–844 (1998).

[31] Pedregosa, F. *et al.* Scikit-learn: Machine learning in python. *Journal of machine learning research* **12**, 2825–2830 (2011).

Appendix A

Keywords Lists

Table 5: List of suicide trivialisation keywords

suicide
suicidal
kill oneself
kill myself
end my life
jump out of the window
throw myself out of the window
throw myself in front of a car
blow my head
drink myself to death
slit my wrists
i want to die
i hope i choke in my own vomit
harakiri
toaster bath

Table 6: List of suicide ideation keywords

about to commit suicide	the only thing stopping me is my family
nobody cares about me	I need someone to listen to me
nobody cares how I'm doing	I can't kill myself
about to end my life	I can't overcome this anxiety
about to hang myself	don't send me flowers
about to kill myself	I will never be happy
self harm	I can't stop crying
farewell letter	I can't live with this
how to kill oneself	I don't want to be in this world
how to commit suicide	I don't want anyone to have to put up with me
slit my wrists	I don't feel like living
should I kill myself	don't come to my funeral
hoping to die	I have never felt so bad
say goodbye to all	suicidal thoughts
diagnosed with depression	Thinking of ending my life
I disgust people	thinking about killing myself
chronic mental illness	thinking of suicide
tired of existing	Why should I stay alive
strength to kill myself	Why should i go on with this life
strength to commit suicide	Why stay alive
I make my family suffer	I will probably be dead
i make everyone suffer	I will probably commit suicide
planned my death	I want to go with him to heaven
planned my funeral	I want to go with her to heaven
tried to suffocate myself	reason to continue living
I will end my life	reasons to kill myself
suicidal impulse	reasons to die
suicidal impulses	reasons to commit suicide
tried to drown myself	no desire to live
suicide attempt	I just want to die
suicide attempts	I'm useless to society
ready to end my life	suicidal
I cry every night	easy suicide
planned death	quick suicide
I hate myself	commit suicide without pain
I feel worse and worse	suicide is the best option
better for everyone if I was dead	inevitable suicide
way to commit suicide	I have anxiety
ways to commit suicide	I have plans to kill myself
die from overdose	therapy doesn't work
die on my birthday	hrow myself to the train
death is the best option	

Appendix B

List of Empath Categories

Table 7: List of Empath built-in categories

achievement	breaking	dance	fabric
affection	business	death	family
aggression	car	deception	farming
air_travel	celebration	disappointment	fashion
alcohol	cheerfulness	disgust	fear
ancient	childish	dispute	feminine
anger	children	divine	fight
animal	cleaning	domestic_work	fire
anonymity	clothing	dominant_heirarchical	friends
anticipation	cold	dominant_personality	fun
appearance	college	driving	furniture
art	communication	eating	gain
attractive	competing	economics	giving
banking	computer	emotional	government
beach	confusion	envy	hate
beauty	contentment	exasperation	healing
blue_collar_job	cooking	exercise	health
body	crime	exotic	hearing

help
heroic
hiking
hipster
home
horror
hygiene
independence
injury
internet
irritability
journalism
joy
kill
law
leader
legend
leisure
liquid
listen
love
lust
magic
masculine
medical_emergency
medieval
meeting
messaging
military
money
monster
morning
movement
music
musical
negative_emotion
neglect
negotiate
nervousness
night
noise
occupation
ocean
office
optimism
order
pain
party
payment
pet
philosophy
phone
plant
play
politeness
politics
poor
positive_emotion
power
pride
prison
programming
rage
reading
real_estate
religion
restaurant
ridicule
royalty
rural
sadness
sailing
school
science
sexual
shame
shape_and_size
ship
shopping
sleep
smell
social_media
sound
speaking
sports
stealing
strength
suffering
superhero
surprise
swearing_terms
swimming
sympathy
technology
terrorism
timidity
tool
torment
tourism
toy
traveling
trust
ugliness
urban
vacation
valuable
vehicle
violence
war
warmth
water
weakness
wealthy
weapon
weather
wedding
white_collar_job
work
worship
writing

youth	zest