Information Sciences 439–440 (2018) 61–78

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Online suicide prevention through optimised text classification

Bart Desmet*, Véronique Hoste

Language and Translation Technology Team (LT3), Department of Translation, Interpreting and Communication, Ghent University, Belgium

ARTICLE INFO

Article history:
Received 28 October 2015
Revised 11 December 2017
Accepted 8 February 2018
Available online 8 February 2018

Keywords:
Suicide prevention
Social media
Text classification
Machine learning
Feature selection
Optimisation

ABSTRACT

Online communication platforms are increasingly used to express suicidal thoughts. There is considerable interest in monitoring such messages, both for population-wide and individual prevention purposes, and to inform suicide research and policy. Online information overload prohibits manual detection, which is why keyword search methods are typically used. However, these are imprecise and unable to handle implicit references or linguistic noise. As an alternative, this study investigates supervised text classification to model and detect suicidality in Dutch-language forum posts. Genetic algorithms were used to optimise models through feature selection and hyperparameter optimisation. A variety of features was found to be informative, including token and character ngram bags-of-words, presence of salient suicide-related terms and features based on LSA topic models and polarity lexicons. The results indicate that text classification is a viable and promising strategy for detecting suicide-related and alarming messages, with F-scores comparable to human annotators (93% for relevant messages, 70% for severe messages). Both types of messages can be detected with high precision and minimal noise, even on large high-skew corpora. This suggests that they would be fit for use in a real-world prevention setting.

1. Introduction

Suicidal behaviour is an important public health concern. Globally, an estimated one million people die by suicide each year [42], making it the sixth leading cause of death for adults aged 20–59 years, and the primary cause of death among teenagers [45]. Apart from successful suicides, there are ten to twenty times as many non-fatal attempts, which also have disruptive emotional and economic consequences. Suicide ideation has an even higher incidence: in a Belgian survey, suicidal thoughts were found to have affected 10% of the male and 15% of the female population between 15 and 24 years old [10].

In spite of these alarming numbers, suicide is generally considered a preventable death: regardless of a victim's stage in the suicidal process (i.e. the progressive stadia of suicidal thoughts, attempt(s) and actual suicide), there often remains ambivalence between life and death. It is a common adage in prevention discourse that suicide is a permanent solution to a temporary problem. Prevention is typically aimed at either the general population, by reducing risk factors and removing barriers to mental health access, or at people who are known or expected to have suicidal tendencies, with adequate risk assessment, medication, therapy and acute crisis support (e.g. suicide hotlines). However, these two prevention types fail to

[*] Corresponding author.
E-mail addresses: bart.desmet@ugent.be (B. Desmet), veronique.hoste@ugent.be (V. Hoste).

https://doi.org/10.1016/j.ins.2018.02.014
0020-0255/© 2018 Elsevier Inc. All rights reserved.

adequately reach the blind spot in between: at-risk individuals who have not yet exhibited suicidal behaviour or found their way to secondary prevention. Efforts to bridge that gap may benefit significantly from suicidality detection on social media.

The rise of the ‘social’ Web 2.0 has had far-reaching implications for human communication. It opened up the possibility to interact and form communities online. Inevitably, these developments have also had an impact on how people communicate about suicidal behaviour. Peter et al. [32] found evidence of reduced inhibition and more self-disclosure in online communication, since it can offer anonymity and a sense of control. Social media have indeed become an outlet for people contemplating suicide to share their thoughts and feelings. Such suicidal expressions can be recognised and responded to by peers, although this may happen in an inappropriate or untimely fashion, if at all. It is therefore preferable to also have trained website administrators or suicide prevention workers monitor user-generated content, if this is not in conflict with users’ preferences, safety and privacy concerns.

Given the massive volume of online content that is continually produced, manual monitoring is practically infeasible. Automatic approaches are therefore required. A search-based approach that uses keywords to locate relevant content would reduce the volume, but still presents a number of problems:

- Specific search queries may only cover a limited range of explicit suicidal expressions (e.g. *suicide* or *kill myself*). Search terms are inadequate for detecting implicit mentions, such as *Wouldn’t it be better if I went now?* or *I would like to end the pain forever*.
- The number of possible (explicit) expressions is too large to capture effectively with keywords. Adding multiple or broader search terms inevitably increases the amount of false positives, adding to the burden for prevention workers who monitor the results. Even highly topical search terms yield false positives, e.g. *political suicide*.
- User-generated content tends to deviate from the linguistic norm. Typical problems include misspellings, the use of abbreviations, phonetic text and colloquial or ungrammatical language use. This may hinder keyword retrieval considerably (e.g. *siucide*).

In this paper, we present the first approach based on text classification to automatically detect suicide-related online content. The focus is on forum and blog messages in Dutch. Text classification of suicidal posts is a high-skew classification problem. To address the skew and data sparsity inherent to the problem, we investigate a wide range of potential features to model suicidality in text, and perform model optimisation through feature selection, hyperparameter tuning, and joint optimisation. The usability of the resulting system is evaluated on large datasets with realistic proportions of suicidal content.

2. Related research

Research conducted on the topic of ‘suicidal text’ has revolved primarily around suicide notes, arguably the most prototypical (albeit rare) textual expression of the suicide victim. For this reason, the genre has long been studied from psychological and psychiatric perspectives [25,37,38]. The field recently saw the introduction of machine learning techniques: in [27], unsupervised clustering techniques are used to separate suicide notes from online newsgroup postings, and Pestian et al. [31] applied supervised classification to distinguish genuine from fake notes. A corpus of 900 genuine suicide notes, annotated with fine-grained emotions, was released in the framework of the 2011 i2b2 NLP Challenge on emotion classification [30], allowing research on which emotions might be indicative of suicidal behaviour, and how they can be found automatically.

Machine learning techniques have been applied in other areas of suicide research as well. Tran et al. [41] built a predictive model to identify patients at high risk from suicidal behaviour, using the information contained in electronic health records (EHR), such as administrative and demographic data, information on prior self-harm episodes and mental and physical health diagnoses. In addition to the clinical codes and numerical data, EHRs also contain free text (e.g. admission notes and discharge summaries), a source of unstructured information that is harder to take advantage of in data mining applications. Haerian et al. [14] explored the use of NLP techniques to extract structured output from EHR notes, and used it in combination with clinical codes to detect potential relationships between drugs (e.g. antidepressants) or psychosocial stressors (e.g. depression, eating disorders, domestic abuse) to the incidence of suicidality. Models that incorporated information from free text were found to have much higher predictive value than those that only included clinical codes.

Work on the automatic detection of suicidal content in online media is scarce. Huang et al. [18] explored the possibility to identify bloggers at risk of suicide, by weighing profiles based on the occurrence of suicide-related keywords. The setup suffered from low precision (35% on the 20 highest-ranking profiles), and did not allow to measure recall, i.e. the number of actually suicidal bloggers that are missing from the results.

A study by Jashinsky et al. [19] also takes a keyword-based approach to detect at-risk content, on Twitter. Keywords were manually selected, along with exclusion terms (e.g. *cutting myself* and *shaving, accidentally* and *slack*). The approach was validated by collecting geolocated tweets that matched the terms, comparing them to tweets from random users from the same US state, and calculating the proportion of at-risk users versus background users. Proportions that departed from the expected (nation-wide) proportion were found to be strongly correlated to the actual age-adjusted state suicide rates, indicating that Twitter may be viable for large-scale monitoring of suicide risk factors. A limitation of the study is that it may not be reliable on an atomic level, i.e. for specific Twitter users.