Mining Twitter for Suicide Prevention

Amayas ABBOUTE[1], Yasser BOUDJERIOU[1], Gilles ENTRINGER[1], Jérôme AZÉ[1], Sandra BRINGAY[1,2], and Pascal PONCELET[1]

1-LIRMM UMR 5506, CNRS, University of Montpellier 2
bringay,aze,poncelet@lirmm.fr,
2-AMIS, University of Montpellier 3

Abstract. Automatically detect suicidal people in social networks is a real social issue. In France, suicide attempt is an economic burden with strong socio-economic consequences. In this paper, we describe a complete process to automatically collect suspect tweets according to a vocabulary of topics suicidal persons are used to talk. We automatically capture tweets indicating suicidal risky behaviour based on simple classification methods. An interface for psychiatrists has been implemented to enable them to consult suspect tweets and profiles associated with these tweets. The method has been validated on real datasets. The early feedback of psychiatrists is encouraging and allow to consider a personalised response according to the estimated level of risk.

Keywords: Classification, Suicide, Tweets.

1 Introduction et motivations

According to the French website Sante.gouv.fr[1], nearly 10,500 people die each year in France by suicide (3 times more than traffic accidents). Approximately 220,000 suicide attempts are supported by Emergency department. The economic burden of suicide is estimated at 5 billion euros for 2009 in France. Suicide is a major public health issue with strong socio-economic consequences. The main objective of this study is to detect, as early as possible, people with suicidal risky behaviour. To do this, we focus on recent information retrieval techniques to identify relevant information in texts from the Twitter social network. These messages are used to learn a predictive model of suicide risk.

Societal benefits associated with such a tool are numerous. The semi-automatic detection model of suicidal profiles can be used by social web services providers. For example, moderators can use such a model to prevent suicide attempts: by communicating directly with the concerned person, by contacting relatives when possible or by displaying targeted advertisements such as *SOS Amitié* (translation: SOS Friendship) advertisement which appears when users enter special terms in google search. A detailed analysis of identified messages can also help psychiatrists to identify emerging causal chains between socio-economic inequalities and different suicidal practices.

[1] http://www.sante.gouv.fr/

We will address three major challenges: 1) Building vocabulary to collect messages from Twitter social network and dealing with various topics related to suicide (e.g. depression, anorexia); 2) Mining messages which are extremely variable from the point of view inter and intra individual in order to propose a classification model to effectively trigger alerts and thus identify people with a high risky behavior; and 3) Presentation of suspect messages in a web interface for health professionals.

The challenges associated with this study are numerous because text analysis is difficult. Most of the NLP methods used in health domain have been applied to publications and hospitalization reports. Their transposition to tweets is far from trivial (limited to 140 characters texts with nonconforming grammatical structures, misspelling, abbreviations, slang). The originality of our solution is to be language-independent and to cover the entire knowledge extraction process: research, acquisition, storage, data mining, classification, visualisation of suspect profiles. To demonstrate the technological feasibility, a prototype is available online[2]. This type of approach can be generalized to widely varying textual data (e.g. blogs, forums, chat, email) and other areas (e.g. cyberbulling, natural disasters detection) for which it is important to identify behaviors known as abnormal based on lexicons.

2 Methodology

The method is divided into 4 steps:

1. **Vocabulary definition.** Thanks to [1][2], we identified 9 topics suicidal people generally talk about: *Sadness/psychological injury, Mental State, Depression, Fear, Loneliness, Description of the suicidal attempt*[3], *Insults, Cyberbullying, Anorexia*. We have defined manually a set of keywords related to these topics on specialized sites and obtained a vocabulary of 583 inputs. We have collected whole sentences that have been used in proven cases of suicide such as "I want to die" or "You would be better off without me".
2. **Suspects and proven messages.** We automatically collected a corpus of tweets containing the words of the defined vocabulary through the API Twitter (about 6000 suspicious messages). We also collected messages from accounts identified as those of persons having committed suicide (proved cases identified thanks to newspapers) (about 30 proven messages).
3. **Manual annotation of messages into two categories: risky tweets and non risky tweets.** Three computers scientists manually classify messages into the two categories (according to the information collected on proven cases). About 150 messages have then been used for the learning phase of the classification process.
4. **Automatic classification.** We automatically classified suspects tweets into risky and non risky tweets. Using WEKA[4], the performances of six classifiers

[2] `http://info-demo.lirmm.fr:8080/suicide2/`
[3] 70% of suicidal people describe concretely how they will realise the suicidal attempt
[4] `http://www.cs.waikato.ac.nz/ml/weka/`

(JRIP, IBK, IB1, J48, Naive Bayes, SMO) were compared using a Leave One Out validation (LOO) and also with a 10-fold Cross-Validation (10-CV) and the results have been averaged under 10 iterations. We first apply no filter to attributes and then remove *Depression* that is very frequent in the "no risky tweet" and rare in the "risky tweets". Removing *Depression* has a significant impact for many classifiers but Naive Bayes remains the best one.

5. **Presentation of results via a web interface.** The objective is to enable psychiatrists to consult tweets indicating suicidal risky behavior, find the latest tweets of the profile and edit statistics.

3 Preliminary results

Table 1 shows the results obtained with the different classifiers tested via WEKA. The best classifier is Naive Bayes both in LOO and 10-CV validation, with an accuracy of 63.15% in LOO and 63.27% in 10-CV. As the two classes have the same number of tweets, the accuracy's baseline is equal to 0% for LOO and to 47.33% for the 10-CV, which is significantly lower than 63.27% for Naive Bayes. Figure 1 shows the distribution of the different categories of the vocabulary in the 6,000 tweets collected and indicating suicidal risky behavior. The most represented categories are *Insults* and *Hurt* for risky tweets.

Dataset	baseline	JRip	IB1	IB3	J48	NB	RF	SMO
10 CV	47.33	55.37	60.16	55.24	57.14	**63.27**	61.03	60.56
LOO	0.00	47.04	60.53	48.68	44.74	**63.16**	59.14	60.72
Without Depression Attribute								
10 CV	47.33	61.23	62.38	61.03	58.65	**63.54**	62.34	60.66
LOO	0.00	61.78	60.00	61.18	57.24	63.16	**63.42**	59.34

Table 1. Accuracy for different classification algorithms. Best results are in bold.

An online prototype was developed to start discussions with psychiatrists. The initial results were encouraging. They confirm that such a tool can be used to provide practical and efficient solutions for suicide prevention. They plan to base on care algorithms which could be personalised to take into account suicidal risk level (e.g. redirection to prevention websites, addressing to the nearest suicidal crisis unit, contacting relatives, etc.). These care algorithms must also integrate well-established risk factors (e.g. age, gender).

4 Conclusions and prospects

In this article, we presented a complete process implementing automatic language processing and learning methods to identify Twitter messages indicating suicidal risky behavior. Initial feedbacks from psychiatrists are encouraging

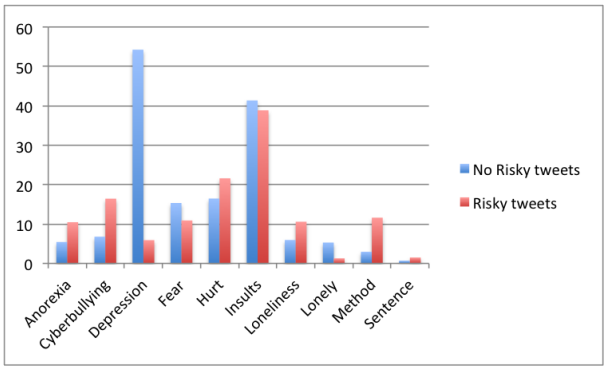

Fig. 1. Risky Tweets: Percentage by subcategories of the vocabulary

and allow them to define prevention methods customised by the level of risk. Prospects are numerous. We plan to collect more information about tweets to improve the learning phase of the automatic classification. First, we will extend the vocabularies with synonyms, antonyms to enlarge the scope of the suspects messages. Web statistic measures will be used to limit noise [3]. We will also use a general vocabulary of emotions to capture more special mental states [4]. Non-textual information such as the increase of the frequency of the tweets posting will also be taken into account. We will also improve the classification phase by using majority vote of multiple classifiers. We will obtain a list of tweets classified according to the level of risk. As our method is language independent, we will reproduce the study for French and Spanish to show its generality. For a medical point of view, we will conduct analyzes contrasting age, gender, location or any other information identified via user profiles Twitter.

5 Acknowledgement

We thank Ph. Courtet and S. Guillaume, Professors for their medical expertise.

References

1. Gunn, J., Lester, D.: Twitter postings and suicide: An analysis of the postings of a fatal suicide in the 24 hours prior to death. Suicidologi **17**(3) (2012) 28–30
2. Luyckx, K., Vaassen, F., Peersman, C., Daelemans, W.: Fine-grained emotion detection in suicide notes: A thresholding approach to multi-label classification. Biomed Inform Insights **5**(1) (2012) 61–69
3. Roche, M., Garbasevschi, O.M.: WeMiT: Web-Mining for Translation. In: Conference on Prestigious Applications of Intelligent Systems, Montpellier, France (August 2012) 993–994
4. Mohammad, S.M., Turney, P.D.: Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon. In: Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, Stroudsburg, PA, USA, ACL (2010) 26–34