ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

PCON-2019

National Conference and Seminar on Innovations in Engineering & Technology

Govt Polytechnic College, Palakkad

Vol. 8, Special Issue 1, January 2019

A Survey on Detecting Crisis of Online Social Media Users

Aswathy K S[1], Dr. Rafeeque P C[2]

PG Student, Department of Computer Science and Engineering, Govt. Engineering College, Palakkad, Kerala, India[1]

Head of the Department, Department of Computer Science and Engineering, Govt. Engineering College, Palakkad, Kerala,India[2]

Abstract: A crisis (e.g. suicide, self-harm, abuse, or eating disorders) is any event that is going (or is expected) to lead to an unstable and dangerous situation affecting an individual, group, community, or whole society. It is very difficult to automatically and accurately detect crisis from social media text due to the commmplexity of identifying crisis states. However, detecting a general state of crisis without explaining why has limited applications. An explanation in this context a coherent, concise subset of the text that rationalizes the crisis detection. Explore several methods to detect and explain crisis using a combination of neural and non-neural techniques. Evaluate these techniques on different dataset obtained from different platforms.

Keywords: Natural Language Processing, Support Vector Machine (SVM), Linguistic Inquiry & Word Count (LIWC)

I. INTRODUCTION

People are increasingly relying on social media platforms like messaging applications and voice assistants to disclose emotions and moods as well as share their personal statuses. A crisis (e.g. suicide, self-harm, abuse, or eating disorders) is any event that is going (or is expected) to lead to an unstable and dangerous situation affecting an individual, group, community, or whole society. Someone who is in crisis will be in need of some form of immediate support. Automatically and accurately detecting whether a social media user is under crisis can have profound consequences. However, without explaining why a particular user is under crisis has limited applications. An explanation in this context is a phrase or clause in the post that most strongly identifies that rational behind the crisis label.

Detecting and explaining crisis in social media can be considered as a complex task, mainly due to complicated nature of mental disorders. In recent years, this research area started to evolve with the continuous increase in popularity of social media platforms that became an integral part of peoples life. For evaluating the explanations generated by the model against human reference explanations, it is not practical to collect enough explanations to train the model. Classifying users under crisis is a simple binary decision task, it can be done more cheaply and quickly. But collecting an explanation requires an annotator to highlight text for every case of crisis.

The applications of crisis detection system is spread over variety of ares. The main scope of this system comes under medical health care centers. The informations obtained as a result from the crisis detection system can be given as extra information for doctors. Which will help the doctors to provide a better suitable treatment.

Basic idea of the crisis detection system is shown in the Fig. 1. given below. This survey aims to discuss about the different methodologies for user crisis detection and highlight the advantages and disadvantages of each of them. Such a comparative study is very much relevant because of the wide range of applications that makes use of crisis detection systems. This survey will provide some insights for choosing the right methodology for developing a crisis detection system for a given domain based on its requirements.

This paper is organized as follows: Section II gives a brief idea about different dataset used and discusses about various classification (machine learning) approaches and neural network approaches for user crisis detection and Section III provides a comparison between them. Section IV gives a brief comparison between the results and also discuss about some future works.

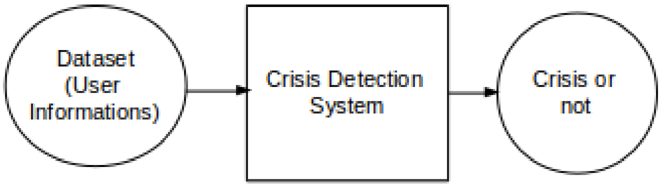

Fig. 1.Basic idea of crisis detection system.

II. CRISIS DETECTION SYSTEM

Based on the methodologies used previous works done for detecting crisis can be classified as two. Those are, machine learning based and neural network based approaches. Most of the papers follow traditional machine learning approaches like Support Vector Machines (SVM), Random Forest (RF), Decision trees etc. Each of them apply their methods on different datasets.

A. *Datasets*

Different datasets used for crisis detection are discussed here, 1) *Koko Dataset:* Koko is a mental health app specially developed for individuals struggling with mood or depressive disorders. Which is available for chatbots on a variety of social media platforms like mobile devices, computers (via Koko's specialized messenger, Facebook Messenger, Kik, Telegram, or Twitter). Koko try to understand the user's specific needs and goals starting a text-like conversation. The dataset mainly contains 106,000 posts labeled either as crisis or not.

2) *Depressive Symptoms and Psychosocial Stressors Associated with Depression (SAD) Dataset:* The SAD dataset was specially developed by Mowery et al. (2016) [1] for automatically classifying depressive symptoms from Twitter data. The SAD dataset contains 9,300 tweets and annotated using three annotators - two psychology undergraduates and a postdoctoral biomedical informatics researcher. In this dataset each tweet was annotated with one or more classes based on DSM-5 (American Psychiatric Association, 2013) and DSM-IV (American Psychiatric Association, 2000).

3) *Twitter Dataset:* Which is a dataset obtained from Twitter, a popular micro blogging service. A sample of public tweets that originated from the New York City (NYC) metropolitan area was extracted using Twitter search API. The collection period started on May 18, 2010 and was one month long. Thus obtained Twitter dataset contains 6,237 tweets of individuals.

It introduces a systematic way of parsing NL text, based on context-free grammar. Evaluation is enhanced in terms of both dataset construction and evaluation mechanisms, that means accuracy, precision, and recall of system is improved. DOL has some nice characteristics that are critical to building a high-precision math problem solving system. Once an answer is provided by this approach, it has a very high probability of being correct.

B. *Different approaches for crisis detection*

Various approaches are put forth for detecting crisis. Some of them are discussed here: 1) *Methodology 1 - Michael Thaul Lehrman et al. (2012):*This approach [2] perform automatic analysis of short written texts based on relevant linguistic text features to identify whether the authors of such texts are experiencing distress. The work is based on Natural Language Processing (NLP) using supervised machine learning.

This work mainly focus on some fundamental supervised classification methods and text-based features to the challenging task of automatically classifying mental affect states in short texts based on just a small dataset. The methodology represents a binary classification problem, were short texts are classified as either distressed or non-distressed. At a more fine-grained level, four classes of text are: high distress, low distress, response, and happy. Any post stating an active intent to harm someone or oneself was classified as high distress, while posts simply discussing bad feelings were usually classified as low distress.

Initially the unlabelled short forum texts are labelled by using an annotator. From the training set several features are extracted and they are used for classification task. The following features were automatically extracted from text, using Python and NLTK with unique unigrams. They are excerpt length in sentences, excerpt and sentence lengths in words, positive vs. negative polarity word list matches, happy, sad, afraid, and angry affect word list matches, first-, second-, and third-person pronouns, and, finally, nouns, verbs, adjectives, adverbs, and pronouns.

Three fundamental supervised classification methods such as Naive Bayes, Maximum Entropy, and Decision Tree are used here for classification task. Figure 2.2 shows the overall system working for this methodology.

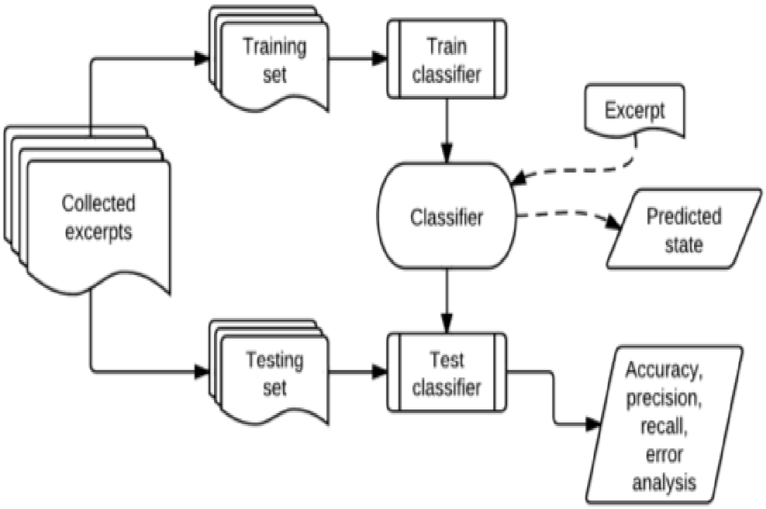

Fig. 2. Computational experimentation process, Lehrman et al. (2012)

2) Methodology 2 - Christopher M. Homan et al. (2014):

This paper [3] takes an initial step toward the automatic detection of suicidal risk factors through social media activity, with no reliance on self-reporting. Mainly focus on a particular aspect of suicidality, namely distress. Use methods that take advantage of lexical analysis to retrieve micro-blog posts (tweets) from Twitter and compare the performance of human annotators - one being an expert and others no.

The methods involve four main phases. They are :

- Filtered a corpus, obtained from Sadilek et al. (2012) [4].
- Annotated each tweets with their level of distress, and also analyzed the annotations in detail.
- Then trained Support Vector Machines (SVM)and topic models.
- Assessed the effectiveness of these methods on validation set.

To facilitate the discovery of distress-related tweets, converted all texts to lower case, stripped out punctuation and special characters, and mapped informal terms (such as abbreviations and net speak) to more standard ones. They used two different methods to filter tweets that are Linguistic inquiry and word count (LIWC) [5] to capture 1,370 tweets by sampling randomly from among the 2,000 tweets with the highest LIWC sad score and next, adopted a collection of inclusive search terms/phrases which was designed specifically for capturing tweets related to suicide risk factors.

For the annotation process divided the resulting set of filtered tweets into two randomized sets of 1,000 tweets each. A novice annotated the first set and a counselling psychologist with experience in suicide related research annotated the second set. Each tweet in each set was rated on a four point scale (H, ND, LD, HD) according to the level of distress.

In the modelling phase each tweet is mapped to a feature spac composed of the unigrams, bigrams, and trigrams in the corpus. Performed topic modelling on dataset. A topic is a set of lexical items that are likely to occur in the same tweet. Topic models are capable of associating words with similar meanings and distinguishing among the different meanings of a single word. Here used latent Dirichlet allocation (LDA) [6] to create these topics. Before doing so, removed stop words and words that occur only once in the dataset. Then applied LDA algorithm on the data to discover three topics.

DA algorithm on the data to discover three topics. The another method used here is Support Vector Machines (SVM). Which is a machine learning method that is used to train a classification model that can assign class labels to previously unseen tweets. SVMs treat each tweet as a point in an extremely high dimensional space (one dimension per uni-, bi-, and tri- gram in the corpus). SVMs are a form of linear separator that can also distinguish between non-linearly separable classes of data by warping the feature space.

3) Methodology 3 - Mowery et al. (2016):

In this work [7], developed classifiers for discerning whether a Twitter tweet represents no evidence of depression or evidence of depression. If there was evidence of depression, then classified whether the tweet contained a depressive symptom and if so, which of three subtypes: depressed mood, disturbed sleep, or fatigue or loss of energy. Different feature groups described in this paper are shown below. Light purple boxes are depressive
symptom subtypes. No evidence of depression and evidence of depression are mutually exclusive classes.

A quantitative study is conducted to train and test a variety of machine learning classifiers. Variety of binary features (present: 1 or absent: 0) are included for study. Some of them are,

- **N-grams** - It may provide meaningful, highly predictive terms indicative of a particular symptom e.g., tired may indicate fatigue or loss of energy. Encoded unigrams using the Twokenizer.
- **Syntax** - Which has been shown to be useful for discerning whether a person is depressed or not e.g., usage of first person vs third person pronouns. Encoded parts of speech using ARK [8].
- **Emoticons** - They are used to demonstrate positive or negative emotion, which could be an indicator of whether an individual is experiencing a depressive mood. Encoded whether the tweet contained emoticons representing four values: happy, sad, both, or neither.
- **Age/ Gender** - Correlated with some depressive symptoms. Because age and gender information is not readily available with tweets, here applied age and gender lexicons to predict the age and gender for each tweet.
- **Sentiment subjectivity terms and polarity terms** - May indicate a persons sentiment and its strength toward people, events, and things. Here leveraged the Multi-Perspective Question Answering lexicons to encode these subjectivity and polarity scales.
- **Personality traits** - Those have been useful predictors of depressive states e.g., depressed individuals exhibit more inward-looking behaviour. Encoded personality traits of openness, conscientiousness, extraversion/introversion, agreeableness/antagonism, neuroticism.
- **Linguistic Inquiry Word Counts** - Words associated with negative emotion, biological state: health and death, cognitive mechanisms including cause and tentativeness have been used to accurately distinguish a depressed individual.

Trained and tested supervised machine learning classifiers for predicting depression-related classes. In this methodology assessed six supervised machine learners. Those include,

- **Decision Tree** -Tree structures for interpretation can be simply represented using decision trees. Learns a prediction model by determining a sequence of the most informative features that maximize the split distinguishing one output class label from another. Tested models produced with an optimized version of the CART algorithm.
- **Random Forests** - Learn many decision trees during its training and classifying a predicted class label based on the mode of the classes or the mean of the prediction of the aggregate individual trees; thus, reducing the likelihood of over fitting by a single decision tree model.
- **Logistic Regression** - Learns a logic regression model in which the dependent variable is the class label. Logistic regression models that leverage regularization avoid over-fitting particularly when the dataset contains only a few number of training examples for a class label, many irrelevant features for classification, and a largenumber of parameters that must be learned. The models are tested with both L 1 and L 2 regularization.
- **Support Vector Machine** - Learns a model that linearly separates two classes in a high dimensional space. Chose to train classifiers using support vector machines because of their ability to tolerate a large number of features while maintaining high performance, to minimize the likelihood of over-fitting by using support vectors for classification, and to withstand sparse data vectors that could be produced by encoding a high number of features. In this approach trained the model using a linear kernel.
- **Linear Perceptron** - Learns a prediction model based on a linear predictor function leveraging a set of weights from a feature vector. Here chose linear perceptron because of their efficiency and ability to be easily trained with large datasets.
- **Naive Bayes** - Learns a prediction model that leverages posterior probabilities of each class and conditional probabilities of the class for each individual feature. Here chose naive Bayes because a naive assumption of independence between features can prove effective for many similar text classification problems.

As a result they analyzed that Support Vector Machine perform as the most accurate classifier with a high F1-score. Also reached in an assumption that Decision trees are the most precise classifiers. This work shows that in most cases the use of machine learning classifiers improve precision in identifying depression symptom and subtype-related tweets compared to the use of keywords alone.

4) Methodology - Robert Morris et al. (2017):

In the current generation many of the users on social media are going through various states of crisis (e.g. suicide, self-harm, abuse, or eating disorders). In this paper [9] they explore several methods to detect and explain crisis using a combination of neural and non-neural techniques. An explanation is a phrase or clause in the post that most strongly identifies the rationale behind the crisis label.

The training set mainly consist of N examples,$[X^i, Y^i]^N{}_{i=1}$. Where the input, X^i is a sequence of tokens $[w_1 , w_2 , ..., w_{\mathbf{T}}$] and the output, Y^i is a binary indicator of crisis. Some annotation is applied to get the explanations for each crisis samples. GloVe, an extension to the word2vec method is used to map each input token to an embedding. Here used the 200 dimensional embeddings for all our experiments, so each word w_t is mapped to $x_{t€}$R 200. The resulting fully embedded sequence is represented as $x_{1:T}$.

A recurrent neural network (RNN) recursively encode a sequence of vectors, $x_{1:T}$. The hidden state of the RNN at t-1 is fed back into the RNN for the next time step.

The model is five layers deep, with a word embedding layer, a two-layer The gated recurrent units (GRU) [4] as encoder and a two-layer long short-memory (LSTM) [5] as decoder. Both the encoder and decoder contain 512 nodes. The reason to use GRU as the encoder instead of LSTM is that the GRU has less parameters and less likely to be over fitted on small dataset.

$$h_t = f\ (x_t\ , h_{t-1}\ ; \Theta)$$

The sequence is then encoded using Gated Reccurant Unit (GRU) [10]. The GRU employs an update gate z_t and reset gate r_t . The update gate helps the model to determine how much of the past information (from previous time steps) needs to be passed along to the future and Reset gate is used from the model to decide how much of the past information to forget. Then use a bidirectional RNN (running one model in each direction) and concatenate the hidden states of each model for each word to obtain a contextual word representation $h^{bi}{}_t$.

$$r_t = \sigma\ (W_r\, x_t + U_r\, h_{t1})$$
$$z_t = \sigma\ (W_z\, x_t + U_z\, h_{t1})$$
$$\tilde{h}_t = \tanh\ (W\ x_t + U\ (r_t\ \ \Theta\ h_{t-1}\ ))$$

Employed attention mechanism to extract words that are important to the meaning of the sentence and aggregate the representation of those informative words to form a sentence vector. With attention, a scoring function scores the relevance of each contextual word representation $h^{bi}{}_t$.

$$u_t = \tanh\ (W_w\ h^{bi}{}_t + b_w\ )$$

The attention mechanism serves two purposes. d acts as a contextual document representation which can be fed into a downstream model component for detection. In addition, the score vector $u_{1:N}$, is utilized to seed the explanation. Optionally for detection, encode the document by using the last hidden state of a single forward GRU. The final document encoding of each sample, d, is fed into a sigmoid layer with one node to detect.

To generate explanations for each input the first step is to build a subset of words which seed the explanation generation function. The seed function is meant to give a set of tokens from the input that most influenced the prediction. For the task of detecting crisis, descriptive content words, such as adjectives, nouns, and verbs, are desirable compared to stop words or punctuation. Three techniques used for seeding words for a given input include:

*1) **Logistic Coefficients:***

- Logistic regression is a linear model that learns a vector of weights for a fixed set of features to detect in binary classification.
- Train a logistic regression model on unigrams to learn a vector of weights for each word in the vocabulary.
- Find the *k* most highly-weighted activated features according to the model.

*2) **Neural Attention:***

- Select seeds by sorting the words by their attention weights u.

*3) **LIME:***

- It is the short for Local Interpretable Model-agnostic Explanations.
- The LIME API contains a num-features parameter in the explain instance function.
- Each explanation will then result in learning an interpretable model, which can be used to then seed the explanation.
- The LIME API is applied to both models, the baseline logistic and the neural model.

The explanation generation algorithm acts on the input text and the *k* explanation seeds. The sentence of importance is identified by taking the sentence with the most seeds. The identified sentence is then parsed with a dependency parser [11] and traversed from the root to find the highest seed in the sentence. If the highest seed token is not a verb and not the head of the entire sentence, then traverse to the seed’s head node. The subtree phrase of the highest seed is used for the explanation.

The main drawback of the system is, in many cases the generated explanation contained more text than is necessary to accurately capture the gold explanation. It is difficult to obtain more precise and accurate explanations.

III. RESULTS AND DISCUSSION

Different methods for detecting crisis are discussed above. In which the methodology suggested by Robert Morris et al. [9] performs the most. The best models presented by them are both effective at detection and produce explanations similar to those produced by human annotators. The Table 1 compares those methodologies based on different facts.

IV. CONCLUSION AND FUTURE WORK

In this paper, present and compare explanation oriented methods for the detection of crisis in social media text. Here introduce a modular approach to generating explanations and make use of neural techniques that significantly outperform the baseline. The best models presented are both effective at detection and produce explanations similar to those produced by human annotators. The researchers find this exciting for two reasons: Within the domain of crisis identification, successes in explanation help to build the trust in trained models that is necessary to deploy them in such a sensitive context.

Few suggestions for building better Detecting Crisis System are :

- In many cases, the generated explanation contained more text than is necessary to accurately capture the gold explanation. This may suggest room for improvement in the explanation generation technique.
- In the future experiments, it is expected to explore human evaluation of the generated explanations as an indicator of trust in the model, to investigate compression based approaches to explanation, and to consider richer architectures for text classification.

Author and Year	Method Used	Dataset Used
Robert Morris et al. (2017)	Recurrent Neural Network, Attention mechanism and Explanation generation algorithm	Data set obtained from Koko
Mowery et al. (2016)	Decision Tree, Random Forests, Logistic Regression, Support Vector Machine (SVM), Linear Perceptron and Naive Bayes	Depressive Symptoms and Psychosocial Stressors Associated with Depression (SAD) dataset
Christopher M. Homan et al. (2014)	Latent Dirichlet Allocation (LDA) and SVM	Twitter dataset
Michael Thaul Lehrman et al. (2012)	Naive Bayes, Maximum Entropy, and Decision Tree	No readily available

REFERENCES

[1]. D. Mowery, H. A. Smith, T. Cheney, C. Bryan, and M. Conway, "Identifying depression-related tweets from twitter for public health monitoring," Online Journal of Public Health Informatics, vol. 8, no. 1, 2016.

[2]. M. T. L. C. O. Alm and R. A. Proaño, "Detecting distressed and non-distressed affect states in short forum texts," NAACL-HLT 2012, p. 9, 2012.

[3]. C. Homan, R. Johar, T. Liu, M. Lytle, V. Silenzio, and C. O. Alm

[4]. "Toward macro-insights for suicide prevention: Analyzing fine grained distress at scale," in Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pp. 107–117, 2014.

[5]. J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic inquiry and word count: Liwc 2001," Mahway: Lawrence Erlbaum Associates, vol. 71, no. 2001, p. 2001, 2001.

[6]. D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," Journal of machine Learning research, vol. 3, no. Jan, pp. 993–1022, 2003.

[7]. D. L. Mowery, A. Park, C. Bryan, and M. Conway, "Towards automatically classifying depressive symptoms from twitter data for population health," in Proceedings of the Workshop on Computational Modelling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES), pp. 182–191, 2016.

[8]. K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N. A. Smith, "Part-of-speech tagging for twitter: Annotation, features, and experiments," tech. Rep., Carnegie-Mellon Univ Pittsburgh Pa School of Computer Science, 2010.

[9]. R. Kshirsagar, R. Morris, and S. Bowman, "Detecting and explaining crisis," arXiv preprint arXiv:1705.09585, 2017.

[10]. J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modelling," arXiv preprint arXiv:1412.3555, 2014.

[11]. M. Honnibal and M. Johnson, "An improved non-monotonic transition system for dependency parsing," in Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 1373–1378, 2015.