International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 5, Issue 5
May 2018

Detecting Stress based on Social Interactions in Social Networks

Mrs. S.Sudeshna[1],
Assistant Professor
Department of Computer Science & Engineering,
St. Peter's Engineering College, Kompally,
Hyderabad, Telangana

B.Deepika Rani[2],
Department of Computer Science and Engineering,
St.Peter's Engineering College, Kompally,
Hyderabad, Telangana

N.Aakash[3],
Department of Computer Science and Engineering,
St.Peter's Engineering College, Kompally,
Hyderabad, Telangana

V.Keerthi[4],
Department of Computer Science and Engineering,
St. Peter's Engineering College, Kompally,
Hyderabad, Telangana

Abstract- Psychological stress is becoming a threat to people's health nowadays. With the rapid pace of life, more and more people are feeling stressed. According to a worldwide survey reported by new business in 2010, over half of the population has experienced an appreciable rise in stress over the last two years. Though stress itself is non-clinical and common in our life, excessive and chronic stress can be rather harmful to people's physical and mental health. The rise of social media is changing people's life, as well as research in healthcare and wellness. With the development of social networks like Twitter and Facebook, more and more people are willing to share their daily events and moods and interact with friends through the social networks. Psychological stress detection is related to the topics of sentiment analysis and emotion detection. Research on tweet-level emotion detection in social networks is done. Through Computer-aided detection, analysis, we are proposing a hybrid model which combines the factor graph model (FGM) with a convolution neural network Where we can detect the people who are under stress based on their tweets.

keywords-KDD,CART,CHAID

I. INTRODUCTION

Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. It is an essential process where intelligent methods are applied to extract data patterns. It is an interdisciplinary subfield of computer science. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.[1] Aside from the raw analysis step, it involves database and data management aspects, datapreprocessing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.[1] Data mining is the analysis step of the "knowledge discovery in databases" process, or KDD.[5]

In the 1960s, statisticians and economists used terms like *data fishing* or *data dredging* to refer to what they considered the bad practice of analyzing data without an a-priori hypothesis. The term "data mining" was used in a similarly critical way by economist Michael Lovell in an article published in the Review of Economic Studies 1983. Lovell indicates that the practice "masquerades under a variety of aliases, ranging from "experimentation" (positive) to "fishing" or "snooping" (negative).[10]

The term *data mining* appeared around 1990 in the database community, generally with positive connotations. For a short time in 1980s, a phrase "database mining"™, was used, but since it was trademarked by HNC, a San Diego-based company, to pitch their Database Mining Workstation;[11] researchers consequently turned to *data mining*. Other terms used include *data archaeology*, *information harvesting*, *information discovery*, *knowledge extraction*, etc. Gregory Piatetsky-Shapiro coined the term "knowledge discovery in databases" for the first workshop on the same topic (KDD-1989) and this term became more popular in AI and machine learning community. However, the term data mining became more popular in the business and press communities.[12] Currently, the terms data mining and knowledge discovery are used interchangeably.The related terms data dredging, data fishing, and data snooping refer to the use of data mining methods to sample parts of a larger population data set that are (or may be) too small for reliable statistical inferences to be made about the validity of any patterns discovered. These methods can, however, be used in creating new hypotheses to test against the larger data populations.

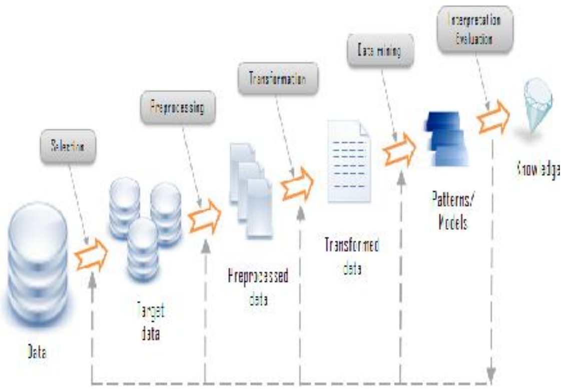


Figure 1: Structure Of Data Mining

The above figure depicts how data is mined through various processes and required useful data is obtained.

II. WORKING OF DATA MINING

While large-scale information technology has been evolving separate transaction and analytical systems, data mining provides the link between the two. Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries. Several types of analytical software are available: statistical, machine learning, and neural networks. Generally, any of four types of relationships are sought:

- *Classes:* Stored data is used to locate data in predetermined groups. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials.
- *Clusters:* Data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities.
- *Associations:* Data can be mined to identify associations. The beer-diaper example is an example of associative mining.
- *Sequential patterns:* Data is mined to anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.

Data mining consists of five major elements:

1.Extract, transform, and load transaction data onto the data warehouse system.

2. Store and manage the data in a multidimensional database system.

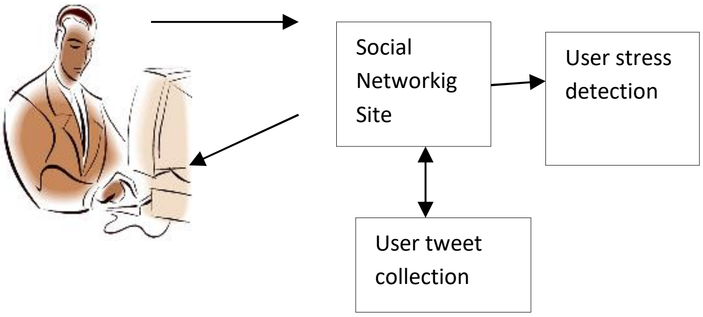
3. Provide data access to business analysts and information technology professionals.

4. Analyze the data by application software.

5. Present the data in a useful format, such as a graph or table.

SYSTEM DESIGN

SYSTEM ARCHITECTURE:



III. DIFFERENT TYPES OF ANALYSIS

- *Artificial neural networks:* Non-linear predictive models that learn through training and resemble biological neural networks in structure.
- *Genetic algorithms:* Optimization techniques that use process such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution.

- *Decision trees:* Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID). CART and CHAID are decision tree techniques used for classification of a dataset. They provide a set of rules that you can apply to a new (unclassified) dataset to predict which records will have a given outcome. CART segments a dataset by creating 2-way splits while CHAID segments using chi square tests to create multi-way splits. CART typically requires less data preparation than CHAID.
- *Nearest neighbor method:* A technique that classifies each record in a dataset based on a combination of the classes of the *k* record(s) most similar to it in a historical dataset (where *k*=1). Sometimes called the *k*-nearest neighbor technique.
- *Rule induction:* The extraction of useful if-then rules from data based on statistical significance.
- *Data visualization:* The visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships.

IV. CHARACTERISTICS OF DATA MINING

Following are the characteristics:

- Large quantities of data: The volume of data so great it has to be analyzed by automated techniques e.g. satellite information, credit card transactions etc.
- Noisy, incomplete data: Imprecise data is the characteristic of all data collection.
- Complex data structure: conventional statistical analysis not possible
- Heterogeneous data stored in legacy systems

V. BENEFITS OF DATA MINING

1. It's one of the most effective services that are available today. With the help of data mining, one can discover precious information about the customers and their behavior for a specific set of products and evaluate and analyze, store, mine and load data related to them

2. An analytical CRM model and strategic business related decisions can be made with the help of data mining as it helps in providing a complete synopsis of customers

3. An endless number of organizations have installed data mining projects and it has helped them see their own companies make an unprecedented improvement in their marketing strategies (Campaigns)

4. Data mining is generally used by organizations with a solid customer focus. For its flexible nature as far as applicability is concerned is being used vehemently in applications to foresee crucial data including industry analysis and consumer buying behaviors

5. Fast paced and prompt access to data along with economic processing techniques have made data mining one of the most suitable services that a company seek

VI. ADVANTAGES OF DATAMINIG

A. *Marketing / Retail:*

Data mining helps marketing companies build models based on historical data to predict who will respond to the new marketing campaigns such as direct mail, online marketing campaign…etc. Through the results, marketers will have appropriate approach to sell profitable products to targeted customers.

Data mining brings a lot of benefits to retail companies in the same way as marketing. Through market basket analysis, a store can have an appropriate production arrangement in a way that customers can buy frequent buying products together with pleasant. In addition, it also helps the retail companies offer certain discounts for particular products that will attract more customers.

B. Finance / Banking

Data mining gives financial institutions information about loan information and credit reporting. By building a model from historical customer's data, the bank and financial institution can determine good and bad loans. In addition, data mining helps banks detect fraudulent credit card transactions to protect credit card's owner.

C. Manufacturing

By applying data mining in operational engineering data, manufacturers can detect faulty equipments and determine optimal control parameters. For example semi-conductor manufacturers has a challenge that even the conditions of manufacturing environments at different wafer production plants are similar, the quality of wafer are lot the same and some for unknown reasons even has defects. Data mining has been applying to determine the ranges of control parameters that lead to the production of golden wafer. Then those optimal control parameters are used to manufacture wafers with desired quality.

D. Governments

Data mining helps government agency by digging and analyzing records of financial transaction to build patterns that can detect money laundering or criminal activities.

E. Law enforcement:

Data mining can aid law enforcers in identifying criminal suspects as well as apprehending these criminals by examining trends in location, crime type, habit, and other patterns of behaviors.

F. Researchers:

Data mining can assist researchers by speeding up their data analyzing process; thus, allowing those more time to work on other projects.

VI. LITERATURE SURVEY

Research has proven that stress reduces quality of life and causes many diseases. For this reason, several researchers devised stress detection systems based on physiological parameters. However, these systems require that obtrusive sensors are continuously carried by the user. In our paper, we propose an alternative approach providing evidence that daily stress can be reliably recognized based on behavioral metrics, derived from the user's mobile phone activity and from additional indicators, such as the weather conditions (data pertaining to transitory properties of the environment) and the personality traits (data concerning permanent dispositions of individuals). Our multifactorial statistical model, which is person-independent, obtains the accuracy score of 72.28% for a 2-class daily stress recognition problem. The model is efficient to implement for most of multimedia applications due to highly reduced low-dimensional feature space (32d). Moreover, we identify and discuss the indicators which have strong predictive power.

This paper examines whether the Cranfield evaluation methodology is robust to gross violations of the completeness assumption (i.e., the assumption that all relevant documents within a test collection have been identified and are present in the collection). We show that current evaluation measures are not robust to substantially incomplete relevance judgments. A new measure is introduced that is both highly correlated with existing measures when complete judgments are available and more robust to incomplete judgment sets. This finding suggests that substantially larger or dynamic test collections built using current pooling practices should be viable laboratory tools, despite the fact that the relevance information will be incomplete and imperfect.

We focus on detecting complex events in unconstrained Internet videos. While most existing works rely on the abundance of labeled training data, we consider a more difficult zero-shot setting where no training data is supplied. We first pre-train a number of concept classifiers using data from other sources. Then we evaluate the semantic correlation of each concept w.r.t. the event of interest. After further refinement to take prediction inaccuracy and discriminative power into account, we apply the discovered concept classifiers on all test videos and obtain multiple score vectors. These distinct score vectors are converted into pairwise comparison matrices and the nuclear norm rank aggregation framework is adopted to seek consensus. To address the challenging optimization formulation, we propose an efficient, highly scalable algorithm that is an order of magnitude faster than existing alternatives. Experiments on recent TRECVID datasets verify the superiority of the proposed approach.

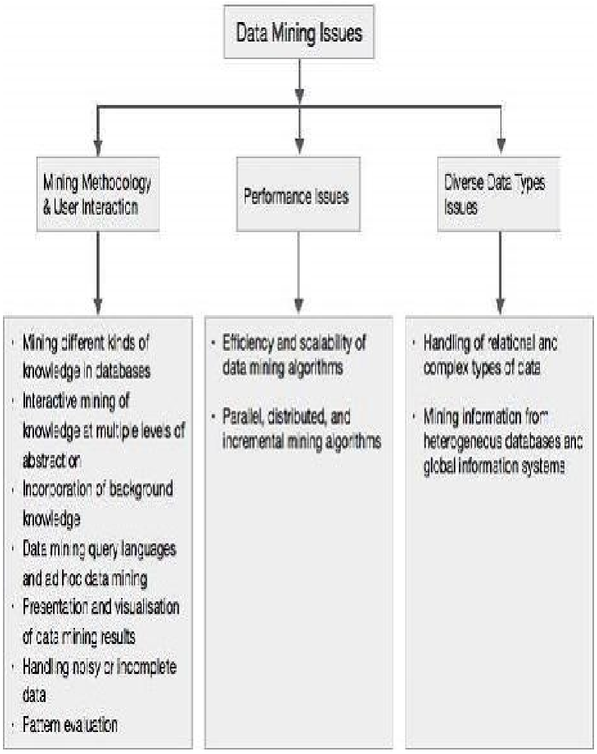
Traditional mental health studies rely on information primarily collected through personal contact with a health care professional. Recent work has shown the utility of social media data for studying depression, but there have been limited evaluations of other mental health conditions. We consider post traumatic stress disorder (PTSD), a serious condition that affects millions worldwide, with especially high rates in military veterans. We also present a novel method to obtain a PTSD classifier for social media using simple searches of available Twitter data, a significant reduction in training data cost compared to previous work. We demonstrate its utility by examining differences in language use between PTSD and random individuals, building classifiers to separate these two groups and by detecting elevated rates of PTSD at and around U.S. military bases using our classifiers.

Online gaming is one of the largest industries on the Internet, generating tens of billions of dollars in revenues annually. One core problem in online game is to find and convert free users into paying customers, which is of great importance for the sustainable development of almost all online games. Although much research has been conducted, there are still several challenges that remain largely unsolved: What are the fundamental factors that trigger the users to pay? How does users? paying behavior influence each other in the game social network? How to design a prediction model to recognize those potential users who are likely to pay? In this paper, employing two large online games as the basis, we study how a user becomes a new paying user in the games. In particular, we examine how users' paying behavior influences each other in the game social network. We study this problem from various sociological perspectives including strong/weak ties, social structural diversity and social influence. Based on the discovered patterns, we propose a learning framework to predict potential new payers. The framework can learn a model using features associated with users and then use the social relationships between users to refine the learned model. We test the proposed framework using nearly 50 billion user activities from two real games. Our experiments show that the proposed framework significantly improves the prediction accuracy by up to 3-11% compared to several alternative methods. The study also unveils several intriguing social phenomena from the data. For example, influence indeed exists among users for the paying behavior. The likelihood of a user becoming a new paying user is 5 times higher than chance when he has 5 paying neighbors of strong tie. We have deployed the proposed algorithm into the game, and the Lift_Ratio has been improved up to 196% compared to the prior strategy.

VII. MAJOR CHALLENGES/ISSUES

Data mining is not an easy task, as the algorithms used can get very complex and data is not always available at one place. It needs to be integrated from various heterogeneous data sources. These factors also create some issues.

The following diagram describes the major issues.


Data Mining Issues
Mining Methodology & User Interaction
Performance Issues
Diverse Data Types Issues
Mining different kinds of knowledge in databases
Interactive mining of knowledge at multiple levels of abstraction
Incorporation of background knowledge
Data mining query languages and ad hoc data mining
Presentation and visualisation of data mining results
Handling noisy or incomplete data
Pattern evaluation
Efficiency and scalability of data mining algorithms
Parallel, distributed, and incremental mining algorithms
Handling of relational and complex types of data
Mining information from heterogeneous databases and global information systems

A. Mining Methodology and User Interaction

It refers to the following kinds of issues −Mining different kinds of knowledge in databases − Different users may be interested in different kinds of knowledge. Therefore it is necessary for data mining to cover a broad range of knowledge discovery task.

Interactive mining of knowledge at multiple levels of abstraction − The data mining process needs to be interactive because it allows users to focus the search for patterns, providing and refining data mining requests based on the returned results.

Incorporation of background knowledge − To guide discovery process and to express the discovered patterns, the background knowledge can be used. Background knowledge may be used to express the discovered patterns not only in concise terms but at multiple levels of abstraction.

Data mining query languages and ad hoc data mining − Data Mining Query language that allows the user to describe ad hoc mining tasks, should be integrated with a data warehouse query language and optimized for efficient and flexible data mining.

Presentation and visualization of data mining results − Once the patterns are discovered it needs to be expressed in high level languages, and visual representations. These representations should be easily understandable.

Handling noisy or incomplete data − The data cleaning methods are required to handle the noise and incomplete objects while mining the data regularities. If the data cleaning methods are not there then the accuracy of the discovered patterns will be poor.

Pattern evaluation − The patterns discovered should be interesting because either they represent common knowledge or lack novelty.

B. Performance Issues

There can be performance-related issues such as follows −Efficiency and scalability of data mining algorithms − In order to effectively extract the information from huge amount of data in databases, data mining algorithm must be efficient and scalable.

Parallel, distributed, and incremental mining algorithms − The factors such as huge size of databases, wide distribution of data, and complexity of data mining methods motivate the development of parallel and distributed data mining algorithms. These algorithms divide the data into partitions which is further processed in a parallel fashion. Then the results from the partitions is merged. The incremental algorithms, update databases without mining the data again from scratch.

C. Diverse Data Types Issues

Handling of relational and complex types of data − The database may contain complex data objects, multimedia data objects, spatial data, temporal data etc. It is not possible for one system to mine all these kind of data.Mining information from heterogeneous databases and global information systems − The data is available at different data sources on LAN or WAN. These data source may be structured, semi structured or unstructured. Therefore mining the knowledge from them adds challenges to data mining.

VIII. IMPLEMENTATION

A. System Framework:

In this framework we propose a novel hybrid model - a factor graph model combined with Convolution Neural Network to leverage tweet content and social interaction information for stress detection. Experimental results show that the proposed model can improve the detection performance by 6-9% in F1-score. By further analyzing the social interaction data, we also discover several intriguing phenomena, i.e. the number of social structures of sparse connections (i.e. with no delta connections) of stressed users is around 14% higher than that of non-stressed users, indicating that the social structure of stressed users' friends tend to be less connected and less complicated than that of non-stressed users.

B. Social Interactions:

We analyze the correlation of users' stress states and their social interactions on the networks, and address the problem from the standpoints of: (1) social interaction content, by investigating the content differences between stressed and non-stressed users' social interactions; and (2) social interaction structure, by investigating the structure differences in terms of structural diversity, social influence, and strong/weak tie. Our investigation unveils some intriguing social phenomena. For example, we find that the number of social structures of sparse connection (i.e. with no delta connections4) of stressed users is around 14% higher than that of non-stressed users, indicating that the social structure of stressed users' friends tend to be less connected and complicated, compared to that of non-stressed users.

C. Attributes categorization

We first define two sets of attributes to measure the differences of the stressed and non-stressed users on social media platforms: 1) tweet-level attributes from a user's single tweet; 2) user level attributes summarized from a user's weekly tweets.

Tweet-level Attributes

Tweet-level attributes describe the linguistic and visual content, as well as social attention factors (being liked, commented, and retweeted) of a single tweet. We can classify words into different categories, e.g. positive/negative emotion words, degree adverbs. Furthermore, we extract linguistic attributes of emoticons, so we can map the keyword in square brackets to find the emoticons. Twitter adopts Unicode as the representation for all emojis, which can be extracted directly.

User-Level Attributes

Compared to tweet-level attributes extracted from a single tweet, user-level attributes are extracted from a list of user's tweets in a specific sampling period. We use one week as the sampling period in this paper. On one hand, psychological stress often results from cumulative events or mental states. On the other hand, users may express their chronic stress in a series of tweets rather than one. Besides, the aforementioned social interaction patterns of users in a period of time also contain useful information for stress detection. Moreover, as aforementioned, the information in tweets is limited and sparse. We need to integrate more complementary information around tweets, e.g., users' social interactions with friends.

There are some techniques to detect the intrusion that occurs in the cloud computing environment. They are:

1. Signature-based Detection: A set of rules that can be used to design given pattern is that of an intruder.
2. Anomaly-based Detection: Identifying events that appear to be anomalous with respect to normal system.
3. Stateful Protocol Analysis: the intrusion detection system could know and trace the protocol states.

IX. SYSTEM ANALYSIS

Existing system:

Many studies on social media based emotion analysis are at the tweet level, using text-based linguistic features and classic classification approaches. A system called *MoodLens* to perform emotion analysis on the Chinese micro-blog platform Weibo, classifying the emotion categories into four types, i.e., angry, disgusting, joyful, and sad.

- A existing system studied the emotion propagation problem in social networks, and found that anger has a stronger correlation among different users than joy, indicating that negative emotions could spread more quickly and broadly in the network. As stress is mostly considered as a negative emotion, this conclusion can help us in combining the social influence of users for stress detection.

Disadvantages of existing system:

- Traditional psychological stress detection is mainly based on face-to face interviews, self-report questionnaires or wearable sensors. However, traditional methods are actually reactive, which are usually labor-consuming, time-costing and hysteretic.
- These works mainly leverage the textual contents in social networks. In reality, data in social networks is usually composed of sequential and inter-connected items from diverse sources and modalities, making it be actually cross-media data.
- Though some user-level emotion detection studies have been done, the role that social relationships plays in one's psychological stress states, and how we can incorporate such information into stress detection have not been examined yet.

*Proposed system**:***

- Inspired by psychological theories, we first define a set of attributes for stress detection from tweet-level and user-level aspects respectively: 1) **tweet-level attributes** from content of user's single tweet, and 2) **user-level attributes** from user's weekly tweets.
- The *tweet-level attributes* are mainly composed of linguistic, visual, and social attention (i.e., being liked, retweeted, or commented) attributes extracted from a single-tweet's text, image, and attention list. The *user-level attributes* however are composed of: (a) *posting behavior attributes* as summarized from a user's weekly tweet postings; and (b) *social interaction attributes* extracted from a user's social interactions with friends.
- In particular, the *social interaction attributes* can further be broken into: (i) *social interaction content attributes* extracted from the content of users' social interactions with friends; and (ii) *social interaction structure attributes* extracted from the structures of users' social interactions with friends.

Advantages of proposed system:

- Experimental results show that by exploiting the users' social interaction attributes, the proposed model can improve the detection performance (F1-score) by 6-9% over that of the state-of-art methods. This indicates that the proposed attributes can serve as good cues in tackling the data sparsity and ambiguity problem. Moreover, the proposed model can also efficiently combine tweet content and social interaction to enhance the stress detection performance.

- Beyond user’s tweeting contents, we analyze the correlation of users’ stress states and their social interactions on the networks, and address the problem from the standpoints of: (1) **social interaction content**, by investigating the content differences between stressed and non-stressed users’ social interactions; and (2) **social interaction structure**, by investigating the structure differences in terms of structural diversity, social influence, and strong/weak tie.
- We build several stressed-twitter-posting datasets by different ground-truth labeling methods from several popular social media platforms and thoroughly evaluate our proposed method on multiple aspects.
- We carry out in-depth studies on a real-world large scale dataset and gain insights on correlations between social interactions and stress, as well as social structures of stressed users.

X.CONCLUSION

. In this paper, we presented a framework for detecting users’ psychological stress states from users’ weekly social media data, leveraging tweets’ content as well as users’ social interactions. Employing real-world social media data as the basis, we studied the correlation between user’ psychological stress states and their social interaction behaviors. To fully leverage both content and social interaction information of users’ tweets, we proposed a hybrid model which combines the factor graph model (FGM) with a convolutional neural network (CNN).

In this work, we also discovered several intriguing phenomena of stress. We found that the number of social structures of sparse connection (i.e. with no delta connections) of stressed users is around 14% higher than that of nonstressed users, indicating that the social structure of stressed users’ friends tend to be less connected and less complicated than that of non-stressed users. These phenomena could be useful references for future related studies.

ACKNOWLEDGEMENT

The authors also express gratitude towards the assistance provided by JP InfoTech Pvt. Ltd. in preparing the manuscript.

XI.REFERENCES

[1] Andrey Bogomolov, Bruno Lepri, Michela Ferron, Fabio Pianesi, and Alex Pentland. Daily stress recognition from mobile phone data, weather conditions and individual traits. In *ACM International Conference on Multimedia*, pages 477–486, 2014.

[2] Chris Buckley and EllenM Voorhees. Retrieval evaluation with incomplete information. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 25–32, 2004.

[3] Xiaojun Chang, Yi Yang, Alexander G Hauptmann, Eric P Xing, and Yao-Liang Yu. Semantic concept discovery for large-scale zero-shot event detection. In *Proceedings of International Joint Conference on Artificial Intelligence*, pages 2234–2240, 2015.

[4] Wanxiang Che, Zhenghua Li, and Ting Liu. Ltp: A chinese language technology platform. In *Proceedings of International Conference on Computational Linguistics*, pages 13–16, 2010.

[5] Chih chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM TRANSACTIONS ON INTELLIGENT SYSTEMS AND TECHNOLOGY*, 2(3):389–396, 2001.

[6] Dan C Ciresan, Ueli Meier, Jonathan Masci, Luca Maria Gambardella, and J ¨ urgen Schmidhuber. Flexible, high performance convolutional neural networks for image classification. In *Proceedings of International Joint Conference on Artificial Intelligence*, pages 1237–1242, 2011.

[7] Sheldon Cohen and Thomas A. W. Stress, social support, and the buffering hypothesis. *Psychological Bulletin*, 98(2):310–357, 1985.

[8] Glen Coppersmith, Craig Harman, and Mark Dredze. Measuring post traumatic stress disorder in twitter. In *Proceedings of the International Conference on Weblogs and Social Media*, pages 579–582, 2014.