Detecting Psychological Stress using Machine Learning over Social Media Interaction

ABSTRACT

Many of the population now face stress leading to psychological issues. Therefore, stress factors must be identified before a big health problem is involved. BP, heart failure, or death are usually caused by excessive stress. The social media data like posting on Facebook profiles, twits on twitter, etc are used to recognize human stress, as people communicate their social media feelings, promote the acquirement of social data, and detect stress based on their behavior. As traditional solutions are highly time-consuming and expensive. Twitter data collection and user tweets are also posted on the site. User stress states are called anxious or unstressed users using the algorithm.

Index Terms: Stress detection, social media, NLP OR Naive Bayes , healthcare, data mining, social interaction.

INTRODUCTION

Social networks to share photographs, videos, and reviews with friends have been created. The consumer will communicate anywhere. Social networking admission principles are assessed in customer behavior evaluation. During the stress recovery procedure, the relation between the user interface and stress is dealt with. For stress recognition processes, Textual, visual and social attributes are measured accordingly. A new hybrid model has been suggested which includes factor graphics and NLP OR NAIVE BAYES for the pull of tweet contents and social-interaction stress sensing. Psychological stress today is continually portrayed as a big health concern. Stress is not therapeutic and natural in any lifespan, unnecessary stress can be harmful and excessive stress is a significant cause of self-morality. Stress should be detected before it is a big health problem. The welfare network consists of building social networks as well as health research, such as Twitter that post everyday events and satire and communicate with them through social networking websites. Social media, digital media. digital media. As such social media information timely represents users' daily life and feelings, it presents new opportunities to represent, quantify, model, and consumer behavior patterns across the wide-ranging social networks and social information in psychology studies may identify their theoretical basis. Some opinion analysis uses emoticons as labels to reduce dependence on emotion classification in computer teaching techniques. Hashtags and smileys on Twitter to improve emotional awareness. Reinforced Twitter opinion or goal grouping by taking into account specific features and relevant

Edit with WPS Office

tweets. Use machine learning approaches, such as NLP OR NAIVE BAYES , to classify social data for stress identification. A NLP OR NAIVE BAYES classifier is a neural network algorithm for classifying objects. NLP OR NAIVE BAYES classification will presume the attributes in the data dots are strongly independent. NLP OR NAIVE BAYES classification is commonly used for machine learning because they are easy to implement. Section I deals with the application of the classification method for human stress recognition, such as NLP OR NAIVE BAYES . In Part II, the literature review of existing systems is discussed, while in Part III the special details of the proposed system execution are provided in Section IV, experimental analysis, observations, and discussion of the system proposed. Section V ends the introduced plan. While reference paper on the final table is given.

HARDWARE REQUIREMENTS

- System: Pentium IV 2.4 GHz.
- Hard Disk: 500 GB.
- Ram: 4 GB
- Any desktop / Laptop system with above configuration or higher level

SOFTWARE REQUIREMENTS

- Operating system: Windows XP / 7
- Coding Language: Python 2.7 and above
- Scripting tool: Jupyter notebook
- Libraries: Pandas, scipy, numpy, matplotlib

LITERATURE SURVEY

1. Statistical Features-Based Real-Time Detection of Drifted Twitter Spam

Twitter spam has become a critical problem nowadays. Recent works focus on applying machine learning techniques for Twitter spam detection, which make use of the statistical features of tweets. In our labeled tweets data set, however, we observe that the statistical properties of spam tweets vary over time, and thus, the

Edit with WPS Office

performance of existing machine learning-based classifiers decreases. This issue is referred to as “Twitter Spam Drift”. In order to tackle this problem, we first carry out a deep analysis on the statistical features of one million spam tweets and one million non-spam tweets, and then propose a novel Lfun scheme. The proposed scheme can discover “changed” spam tweets from unlabeled tweets and incorporate them into classifier’s training process. A number of experiments are performed to evaluate the proposed scheme. The results show that our proposed Lfun scheme can significantly improve the spam detection accuracy in real-world scenarios.

2. Effect of Spam on Hash tag Recommendation for Tweets

Presence of spam tweets in a dataset may affect the choices of feature selection, algorithm formulation, and system evaluation for many applications. However, most existing studies have not considered the impact of spam tweets. In this paper, we study the impact of spam tweets on hashtag recommendation for hyperlinked tweets (i.e., tweets containing URLs) in HSpam14 dataset. HSpam14 is a collection of 14 million tweets with annotations of being spam and ham (i.e., non-spam). In our experiments, we observe that it is much easier to recommend “correct” hashtags for spam tweets than ham tweets, because of the near duplicates in spam tweets. Simple approaches like recommending most popular hashtags achieves very good accuracy on spam tweets. On the other hand, features that are highly effective on ham tweets may not be effective on spam tweets. Our findings suggest that without removing spam tweets from the data collection (as in most studies), the results obtained could be misleading for hashtag recommendation tasks.

3. HSpam14: A Collection of 14 Million Tweets for Hashtag-Oriented Spam Research

Hashtag facilitates information diffusion in Twitter by creating dynamic and virtual communities for information aggregation from all Twitter users. Because hashtags

Edit with WPS Office

serve as additional channels for one's tweets to be potentially accessed by other users than her own followers, hashtags are targeted for spamming purposes (e.g., hashtag hijacking), particularly the popular and trending hashtags. Although much effort has been devoted to fighting against email/web spam, limited studies are on hashtag-oriented spam in tweets. In this paper, we collected 14 million tweets that matched some trending hashtags in two months' time and then conducted systematic annotation of the tweets being spam and ham (i.e., non-spam). We name the annotated dataset HSpam14. Our annotation process includes four major steps: (i) heuristic-based selection to search for tweets that are more likely to be spam, (ii) near-duplicate cluster based annotation to firstly group similar tweets into clusters and then label the clusters, (iii) reliable ham tweets detection to label tweets that are non-spam, and (iv) Expectation-Maximization (EM)-based label prediction to predict the labels of remaining unlabeled tweets. One major contribution of this work is the creation of HSpam14 dataset, which can be used for hashtag-oriented spam research in tweets. Another contribution is the observations made from the preliminary analysis of the HSpam14 dataset.

4. An Analysis of 14 Million Tweets on Hashtag-Oriented Spamming

Over the years, Twitter has become a popular platform for information dissemination and information gathering. However, the popularity of Twitter has attracted not only legitimate users but also spammers who exploit social graphs, popular keywords, and hashtags for malicious purposes. In this paper, we present a detailed analysis of the HSpam14 dataset, which contains 14 million tweets with spam and ham (i.e., non-spam) labels, to understand spamming activities on Twitter. The primary focus of this paper is to analyze various aspects of spam on Twitter based on hashtags, tweet contents, and user profiles, which are useful for both tweet-level and user-level spam detection. First, we compare the usage of hashtags in spam and ham tweets based on frequency, position, orthography, and co-occurrence. Second, for content-based analysis, we analyze the variations in word usage, metadata, and near-duplicate tweets. Third, for user-based analysis, we investigate

Edit with WPS Office

user profile information. In our study, we validate that spammers use popular hashtags to promote their tweets. We also observe differences in the usage of words in spam and ham tweets. Spam tweets are more likely to be emphasized using exclamation points and capitalized words. Further, we observe that spammers use multiple accounts to post near-duplicate tweets to promote their services and products. Unlike spammers, legitimate users are likely to provide more information such as their locations and personal descriptions in their profiles. In summary, this study presents a comprehensive analysis of hashtags, tweet contents, and user profiles in Twitter spamming.

5. Semi-Supervised Spam Detection in Twitter Stream

Most existing techniques for spam detection on Twitter aim to identify and block users who post spam tweets. In this paper, we propose a Semi-Supervised Spam Detection (S3D) framework for spam detection at tweet-level. The proposed framework consists of two main modules: spam detection module operating in real-time mode, and model update module operating in batch mode. The spam detection module consists of four light-weight detectors: (i) blacklisted domain detector to label tweets containing blacklisted URLs, (ii) near-duplicate detector to label tweets that are nearduplicates of confidently pre-labeled tweets, (iii) reliable ham detector to label tweets that are posted by trusted users and that do not contain spammy words, and (iv) multi-classifier based detector labels the remaining tweets. The information required by the detection module are updated in batch mode based on the tweets that are labeled in the previous time window. Experiments on a large scale dataset show that the framework adaptively learns patterns of new spam activities and maintain good accuracy for spam detection in a tweet stream.

SYSTEM REQUIREMENT SPECIFICATION

A Software Requirement Specification (SRS) is basically an organization's understanding of a customer or potential client's system requirements and dependencies at a particular point prior to any actual design or development work.

Edit with WPS Office

The information gathered during the analysis is translated into a document that defines a sets of requirements. It gives the brief description of the services that the system should provide and also the constraints under which, the system should operate. Generally, the SRS is a document that completely describes what the proposed software should do without describing how the software will do it. It's a two-way insurance policy that assures that both the client and the organization understand the other's requirements from that perspective at a given point in time.
The SRS document itself states in precise and explicit language those functions and capabilities a software system must provide, as well as states any required constraints by which the system must abide. The SRS also functions as a blueprint for completing a project with as little cost growth as possible. The SRS is often referred to as the "parent" document because all subsequent project management documents, such as design specifications, statements of work, software architecture specifications, testing and validation plans, and documentation plans, are related to it. Requirement is a condition or capability to which the system must conform. Requirement Management is a systematic approach towards eliciting, organizing and documenting the requirements of the system clearly along with the applicable attributes. The elusive difficulties of Requirements are not always obvious and can come from any number of sources.

FUNCTIONAL REQIREMENTS

A function of software system is defined in functional requirement and the behavior of the system is evaluated when presented with specific inputs or conditions which may include calculations, data manipulation and processing and other specific functionality. The functional requirements of the project are one of the most important aspects in terms of entire mechanism of modules.

NON-FUNCTIONAL REQUIREMENTS

Usability

Simple is the key here. The system must be simple that people like to use it, but not so complex that people avoid using it. The user must be familiar with the user interfaces and should not have problems in migrating to a new system with a new

Edit with WPS Office

environment. The menus, buttons and dialog boxes should be named in a manner that they provide clear understanding of the functionality. Several users are going to use the system simultaneously, so the usability of the system should not get affected with respect to individual users.

Reliability

The system should be trustworthy and reliable in providing the functionalities. Once a user has made some changes, the changes must be made visible by the system. The changes made by the Programmer should be visible both to the Project leader as well as the Test engineer.

CONCLUSION

Human stress has risen for several days, leading to significant health problems. Throughout the management of stress, higher stress is therefore necessary. A model has been suggested to describe the burden of social information as people seek, through social media tweets ou mails, to express their feelings. The Twitter dataset is used for user behavior and NLP OR NAIVE BAYES for training content classification. NLP OR NAIVE BAYES is used for classification. Since it is very difficult to identify big data files with a conventional approach, social media data is used to describe the training file. Class 0, class1, and class2 of the tests of the tweet uses. Class 0 suggests a positive stress level, Class 1 displays negative stress levels and Class 2 corresponds to a neutral stress level.

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] Huijie Lin, Jia Jia, Jiezhon Qiu, Yongfeng Zhang, Lexing Xie, Jie Tang, Ling Feng, and Tat-Seng Chua, "Detecting Stress Based on Social Interactions in Social Networks", IEEE Transactions on Knowledge and Data Engineering, 2017.

[3] Xiaojun Chang, Yi Yang, Alexander G Hauptmann, Eric P Xing, and YaoLiang Yu. Semantic concept discovery for large-scale zero-shot event detection. In

Edit with WPS Office

Proceedings of International Joint Conference on Artificial Intelligence, pages 2234–2240, 2015.

[4] Li-fang Zhang, -Occupational stress teaching approach among Chinese academics, Educational Psychology, Volume 29, Issue 2, March 2009, pages 203 – 219.

[5] Q. Li, Y. Xue, J. Jia, and L. Feng, "Helping teenagers relieve psychological pressures: A microblog based system," in EDBT, 2014.

[6] Aditya Mogadala, "Twitter User Behavior Understanding with Mood Transition Prediction" 20

[7] Yuanyuan Xue, Qi Li , Li Jin, Ling Feng, David A. Clifton,Gari D. Clifford ,"Detecting Adolescent Psychological Pressures from Micro-Blog", 2013.

[8] 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.

Edit with WPS Office