Chapter 10
Behaviour Anomaly Detection With Similarity-Based Sampling for Imbalanced Data

Isha Y. Agarwal
Sardar Vallabhbhai National Institute of Technology, Surat, India

Dipti P. Rana
https://orcid.org/0000-0002-5058-1355
Sardar Vallabhbhai National Institute of Technology, Surat, India

Kshitij R. Suri
Sardar Vallabhbhai National Institute of Technology, Surat, India

Punitkumar Jain
Sardar Vallabhbhai National Institute of Technology, Surat, India

Saumya Awasthi
Sardar Vallabhbhai National Institute of Technology, Surat, India

Krittika Roy
Sardar Vallabhbhai National Institute of Technology, Surat, India

ABSTRACT

Mental health is a major issue in our society, and people treat this issue as a subject that should not be spoken about. So, many such individuals utilize social media as a platform to share their thoughts and fears. This emphasizes the researchers to identify sufferers who require treatment. Many approaches have been devised to detect early markers of mental health illness, some of which include learning algorithms based on the heuristic of equally distributed balanced data. However, they yield biased results towards the majority data (i.e., normal behaviour). Thus, new perception is needed to explore the available data. This research deals with the first identification of such users from weblog data, and the similarity-based sampled data is then given to the classifier. The experiment analysis shows the effectiveness of this work and will provide the user's mental state information early to take timely necessary steps.

DOI: 10.4018/978-1-7998-7371-6.ch010

INTRODUCTION

According to The World Health Organization, around 800,000 people die due to suicide every year ("Suicide data", 2020). That means every 40 seconds one person commits suicide. A first-year Master's student in IIT Madras committed suicide due to academic reasons (Lobo, 2020). A renowned singer committed suicide after suffering from depression for a long time. Unfortunately, several such cases have been reported in the past few years. It is a serious and troubling issue, for the person who commits it, their family, and a loss for the entire Nation. The World Health Organization also states that a person who has committed suicide is likely to have attempted up to 20 times before the actual incident. According to the American Foundation for Suicide Prevention, 5075 percent of the individuals who attempt suicide talk about or express their intentions (Pappas, 2017). Early detection of such behaviour is crucial. Social Media has become one of the most widely used platforms for people to share and express their thoughts and feelings. In the proposed model, with the help of appropriate data mining techniques, the user's data can be analyzed to determine the complex behavioural patterns exhibited by them. The company or institute access logs are used to monitor the frequency of accessing social media websites and the content present in them is analyzed so that users can be classified based on these parameters. With the help of such a model if abnormal behaviour can be tracked down, then these suicides can be prevented and people's lives can be saved.

BACKGROUND

Unstructured text from social media is a common aspect of both clustering and information retrieval-based algorithms. It is imperative to first structure the given data into meaningful categories and then performs algorithm-based analysis. In this section, the different approaches have been discussed that are implemented by prior works to analyze textual data by categorizing its contents.

Content Classification

In order to structure user data by classifying it into meaningful categories, prior works have followed three major approaches, as described below.

Topic-Based Classification

Sergei Koltcov et al. (Koltcov, Koltsova & Nikolenko, 2014) showed that one should be cautious while using the Latent Dirichlet Allocation (LDA) model (Armstrong, 2015), especially for Topic Modelling because the algorithm contains an inherent uncertainty. The authors introduced a new metric of the document and word ratios for evaluating the different aspects of the topic modelling. The use of this new metric decreased the size of vocabulary used in the topic similarity metric based on Kullback–Leibler divergence.

Bo Dao et al. (Dao, Nguyen, Venkatesh & Phung, 2017) presented an approach for analyzing general and mental health-based online communities on social media platforms. These online communities were analyzed based on different categorization methods such as mood tags, text analysis, writing style analysis, and generic words. The data obtained through these categorizations was given to Hierarchical