EXTRACTING USER BEHAVIOURAL CONTROL STYLES BASED ON PROCESS MINING

SHIELADAVI2, DIVYANG SARANGI2, AGNIVA CHAKRABORTY3, C. AKASH MAHADEVAN4
1Assistant Professor, 2,3, 4UG Scholars, Department of CSE, SRMIST Ramapuram, Chennai-89, India
shielad@srmist.edu.in, divyangs2000@gmail.com, codeagni99@gmail.com, akash200400@gmail.com

ABSTRACT

With the increase in advancement in Artificial Intelligence and Data Mining, sentiment analysis is a hot topic. Also, social media is getting more coverage. Public and private opinions on a wide range of topics are constantly shared and distributed through a variety of social media platforms. Twitter is a popular social media platform. Twitter provides businesses with a quick and easy way to assess their consumers' views on issues that are vital to their success in the marketplace. It solves the underlying problem of sentiments that is generally lost in between tweets and retweets. This would improve understanding of the underlying context and the sentiment behind the particular tweet and gain clarity on the subject. We achieve this strategy with techniques such as Data Mining and Natural Language Processing for extractCreating a sentiment analysis software is a method for measuring consumers’ expectations computationally. This paper describes the development of a sentiment analysis that extracts a large number of tweets. Thus, we arrive at a conclusion where we expect to achieve an attractive accuracy rate using this technique.

Keywords: Data Mining, sentiment analysis, Natural Language Processing, tweets, social media platforms.

I. INTRODUCTION

Twitter is one of the largest social media platforms, where people publish short status updates called ‘tweets’. Other users interact with tweets by commenting or reposting the tweets to their followers. The large amount of publicly available data has made twitter sentiment analysis a huge research topic due to its wide applications. Analyzing tweets, helps companies understand public opinion towards their products, helps political parties and candidates develop strategies for campaigning, etc. The objective is the classification of sentiment polarity in a tweet as positive, neutral or negative, using sentiment analysis. Traditional text analysis methods are not efficient because of the causality of tweets, with slangs, acronyms, misspelled words, etc. To tackle these issues, many twitter sentiment analysis methods have been developed, mainly classified in two categories: Fully and Distantly supervised methods. Sentiment classifiers are learnt using sentiment lexicons and data that is manually labeled in a Fully supervised method. It requires more resources and consequently most methods often use very little sentiment lexicons and labelled data to guarantee good performance. In Distantly supervised methods learn sentiment classifiers from data with noisy labels such as emoticons, like, “:)” and “:(” can be used as labels to analyze tweets to be very likely positive or negative, respectively. Although, this method requires less resources, the performance may not be adequate due to the noisy labels.

II. RELATED WORK

The existing system of Sentiment analysis in Twitter has become an interesting topic of research in recent years. Several current Twitter sentiment analysis tools only take into account textual details from tweets, and fail to perform well when dealing with brief and indeterminate tweets. Recent studies indicate that, sentiment diffusion trends on Twitter are closely linked with polarities in sentiment of tweets. As a result, the focus of this paper is on combining text-based information from tweets with sentiment diffusion trends for enhancing sentiment analysis on Twitter data. The existing system, uses Object centric behavioral constraint algorithm. Event logs are used as input for determining insights in Process mining. Cyberbullying events are kept as record and organized using the event logs. A case notion to associate events is often assumed in the current standard. It assumes complications in
organizing the events from social media, since there is difficulty in identifying a case notion due to one-to-many and many-to-many relations that occur in social media platforms.

III. PROPOSED SYSTEM

The proposed system is an interactive automatic system which is called “Content Derivative Algorithm”. It predicts cyberbullying in review or tweets of people posting in social media platforms. Cyberbullying analysis and classification, feature-based classification, negation handling, and Opinion Summarization are all performed by the algorithm. For predicting opinion polarity, a precise method is used, which aids in the improvement of marketing strategies. The primary focus of this proposed system is to indicate cyberbullying analysis in real time in the tweets being acquired from twitter and provide insights into time-based analytics to the end-user. The proposed cyberbullying system performs the following tasks: Categorization and automatic classification of cyberbullying tweets; Provide insights into cyberbullying events and cyberbullies.

The Categorization and Automatic Classification of Tweets deals with the categorization of sentiment driven tweets and classify them based on their polarities as positive, negative and neutral. This would give a clear picture about the mood of the person while writing the post. After analyzing several tweets, we might be provided with insights about the cyberbullying contexts and the deviation from a particular topic.

IV. SYSTEM ARCHITECTURE

The overall working of the system architecture is represented in a diagrammatic form in Fig.1.1

4.1 DataLoading

Text mining is the process of extracting useful information from text data using natural language processing methods and analysis techniques. Twitter is a rich source of data and knowledge that can be utilized for the collection of data on any subject-matter imaginable. This information can be used for a variety of purposes, including identifying patterns relevant to certain keywords, perception of brand sentiment, and acquiring reaction on new products and services. Twitter is a gold mine for information. Unlike various social media platforms, almost the tweets of all users are publicly available and can be searched. If you're trying to collect a large volume of data for analytics, this is a big plus. The information on Twitter is also very unique. The Twitter API helps you to conduct complex queries such as retrieving all tweets about a particular subject as recent as 20 minutes, also retrieving a specific user's tweets which are not re-tweeted. Data encryption and Decryption.

4.2 Data Preprocessing

Preprocessing the data is an important phase in creating a Schizophrenic Discourse model. Data preprocessing includes a number of essential steps, which include cleaning, transformation, and feature selection of data. Schizophrenic Discourse Data cleaning and transformation are techniques used to eliminate deviated data and standardize it so that they can be shaped and conveniently used for construction of a model. The data set of a Schizophrenic Discourse may contain several descriptors; nonetheless, many of these variables might comprise of redundant data. In order to reduce the model's depth, only variables comprising of distinct and relevant information should be considered. Variables which cannot relate to the model of Schizophrenic Discourse may be extracted using data mining procedures. Before we extract and classify features, we go through a series of preprocessing measures. The library is used to perform named entity (NE) identification, reference resolution, and dependency parsing.

4.3 Model Building

A recurrent neural network may be a sort of neural network. A type of neural network that tries to model time or actions relying on sequence like stock market costs, language and electricity need. This is demonstrated by using the output of a neural network layer at a time ‘x’ as input to the exact same network layer at time ‘x + 1’. A network can be interpreted a recurring neural network that utilizes cell blocks instead of traditional layers of a neural network. The input, forget and output gate are all segments of these cells. This memory enables the network to learn long-term dependencies sequentially, allowing it to take in consideration the entire context while making a prediction. Whether it's the next word in a sentence, a sentiment description, or a temperature calculation, the next step is always the same. A network is designed to replicate how humans interpret sequences: instead of responding to individual words, we consider the entire sentence when forming a response.
V. MODULE DESCRIPTION

5.1 Feature Extraction

Feature extraction is a method that dimensionally decreases and transforms a wide set of raw data into smaller sets for processing. The huge number of variables in these large data sets necessitates many computational resources to process. Feature extraction is a name for methods that combine and select variables into features, effectively decreasing the quantity of the knowledge which has got to be processed, while staying accurate and complete when describing the original data set. When you need to decrease the amount of resources necessary for processing without forfeiting valuable information, feature extraction is very useful. Feature extraction also minimizes redundant data in a report. In addition, the reduction of data and the machine's efforts in constructing variable combinations speed up the learning and generalization steps.

5.2 Conformance Checking

A family of process mining techniques called business process conformance testing compares a process model to an event log of a similar process. It's used to see if the model matches the actual execution of a business process as documented in the event log, and vice versa. For example, a process model can indicate that purchasing orders exceeding one 1,000,000 euros require 2 checks. The event log will indicate whether or not this rule is observed. Conformance testing techniques take a process model and an event log as input and return a list of inconsistencies between the event log action and process model. These distinctions can be interpreted as a list of natural language statements, either visually or textually. Some techniques can also generate normalized measurements (between 0 and 1) that show how well the process model and the event log fit. Discrepancies between the model and the log suggest that the model needs to be changed to better capture reality if it is supposed to be descriptive. If the model is normative, such differences may be interpreted in either of two ways: they can disclose undesirable deviations or they can uncover desirable deviations.

5.3 Data Transformation

Data transformation is a method of remodeling data from one format to a different format, generally from a source system's format to a destination system's required format. Many data management and integration functions, like data warehousing and data wrangling, require data transformation. Data transformation is a step in the ELT/ETL process that can be classified as either "simple or complex," depending on the types of changes that must be made to the data before it is sent to its final destination. The data transformation process can be done automatically, manually, or a combination of both. The reality of big data today means that data transformation is more important than ever for companies. Huge amounts of data are constantly generated by an ever-increasing number of programmers, software, and computers. Data compatibility is still in jeopardy with so much disparate data coming in from a number of sources. That's where the data transformation process comes in: it empowers organizations and associations to divert information from any source into a configuration that can be effectively integrated, stored, analyzed, and mined for significant business knowledge.
5.4 Data Analysis

The process of reviewing, cleansing, transforming, and modelling of the data with the aim of finding useful knowledge, conclusions, and assisting decision making is known as data analysis. Data analysis has many aspects and methods, enveloping a wide scope of methods under different names and being utilized in various enterprises, examination, and sociology spaces. Data analysis is essential in the present business world since it assists organizations with settling on more logical choices and run all the more productively. Data mining is a type of data analysis that focuses on mathematical modelling and knowledge exploration for predictive instead of strictly descriptive purposes, while business intelligence is a type of data analysis that mainly relies on aggregation and is primarily concerned with business data. It can be divided into descriptive statistics, EDA and CDA. Confirmatory data analysis in statistical applications focuses on confirming or falsifying existing theories, while exploratory data analysis focuses on finding new features in the data. Text analytics uses mathematical, linguistic, and structural techniques to extract and classify information from textual sources, a type of unstructured data. Predictive analytics focuses on the use of statistical models for prescient anticipating or classifying, while prescient investigation focuses around the use of factual models for prescient estimating or classification. All of the above are examples of data processing. It is a prerequisite for data processing, which is in turn related to data visualization and distribution.

VI. WORKING AND ANALYSIS

The implementation starts with importing all dependencies. We import libraries such as pandas, numpy etc. Initially we extract the raw tweets from a general dataset and then we normalize it, as we need to filter out the various emoticons, bullying tweets, informal tweets from the dataset. We needed to keep every possibilities in a tweet such as hashtags, emoticons, retweets, hyperlinks, mentions etc. Now we classified out tweet texts into single words which occur at a higher frequency and check their frequency count and analyze which word generally has a more frequency count in the dataset. We essentially took a sample size of the total population of words from our dataset and computed the frequency using matplotlib library. We get the resulting graph as shown below.

![Fig: 2 Words with maximum frequency in the dataset](image)

Now we decided to blend multiple words so that we can obtain more clarity on our dataset. Blending the two or more words together would help in understanding the informal accents in natural language contexts. For example in a phrase –“I don’t like Pizza” clearly defines the negative emotion expressed by a person for Pizza. Thus multiple blending of words are performed in order to achieve a more clear understanding and better representation of the system.
After we extract the words from the dataset, we use classification method and represent it as dense feature vector. After representation of the vectors, we infer that the dataset follows Zipf’s law. It is a formulation of quantitative linguistics which results in analyzing the natural flow of a language in a sample space and based on the wherein the word with maximum frequency would occur approximately two times, like the next word with maximum frequency is thrice the maximum frequency word etc. The representation of Zipf’s law is given below. Thus this is the result we obtained from our implementation. This will give us insights in identifying cyberbullying tweets with most common word occurring and in turn analyzing the sentiment of the user behind that tweet.

VII. CONCLUSION AND FUTURE SCOPE

We introduced the idea of cyberbullying Process Mining in this paper and created an application to incorporate it. The aim of process mining is to remove the execution traces of processes in order to find, observe, and develop them. Systems that allow users to be more versatile in their behavior (systems without a fixed structure or unregulated rules for user behavior) hold log files, from which a large number of sequences, often very similar, can be extracted. In such systems, we propose a new approach that helps us to clearly visualize the latent relations between users (or groups of users) with similar actions. Due to the broad dimension of the base user profile vector, traditional methods for constructing a simple user network are not possible in these situations. It was one of the causes that contributed to the discovery of cyberbullying user activity trends. The second explanation was to find representative forms of user actions, or behavioral patterns, in order to better explain the system from a global perspective. In future works, the improvements can be made in preventing the cyber-attacks and securing...
the standard electric values using advanced security algorithms. The electric power system plays a significant role on economics and social benefits. There are several risk factors in building an efficient electric power system right from the production, transportation, storage till the distribution to the end users. So, each and every risk factor should be identified and the right actions should be taken to control those risk factors which would accelerate the effectiveness of the system entirely. The transition cost should be effective. The system should be updated according to the current technologies which eases the way of protecting the data systematically. Every need of the consumer should be analyzed and the model should be redesigned if it is necessary. Thus, building an effective electric power system with essential preventive measures plays a vital role all over the world.

REFERENCES

1. Lei Wang, Jianwei Niu, and Shui Yu, “SentiDiff: Combining Textual Information and Sentiment Diffusion Patterns for Twitter Sentiment Analysis”
2. S. Tsugawa and H. Ohsaki, “Negative messages spread rapidly and widely on social media”
4. E. Cambria, “Affective computing and sentiment analysis,”