Twitter Based Crime Rate Analysis of U.S cities

Shaunak Sangdod
Department of Computer Science
University At Albany, SUNY
Albany, USA
ssangdod@albany.edu

Nayanika Bhargava
Department of Computer Science
University At Albany, SUNY
Albany, USA
nbhargava@albany.edu

Dr. Feng Chen
Department of Computer Science
University At Albany, SUNY
Albany, USA
fchen5@albany.edu

Chunpai Wang
Department of Computer Science.
University At Albany, SUNY
Albany, USA
cwang25@allbany.edu

***Abstract*—Crimes are a social nuisance and cost our society in several ways. They create a havoc and unrest among the common people. Any research that may help in solving crimes will pay for itself. Data Mining is one most important techniques that will help us in analyzing the data related to the crime and model the crime rate analysis problems. Our approach addresses the crime rate in several cities of the United States of America based on the twitter data and classify the cities as safe or unsafe. We have used the Support Vector Machine for supervised learning to classify our data after data collection and preprocessing. We have also compared our Model with Logistic regression and provide visualizations in the form of bar graphs, spatial map and word clouds. We got the model accuracy of approximately 94 percent and finally classified the cities as safe or unsafe.**

Keywords—geo-tagging, Support Vector Machine (SVM), crime detection, crime rate analysis, Data mining techniques, Logistic regression, Data Visualization.

I. INTRODUCTION

Crime, in all aspects, is a common social problem affecting the quality of life and economic development of the society. It is considered one of the most important factors for people to move to a new city and places to avoid for other purposes [1]. With the increasing crime rate, the law enforcement agencies continue to demand better systems and improvised data mining techniques to improve crime analytics and protect the society. According to the sources [2], Violent crimes are on a constant in many of the largest cities of United States. United states saw an increase of 11 percent in crimes from the previous year. It has also been noticed that such crimes are often concentrated in a handful of neighborhoods. Our analysis aims to predict several cities of the United States of America as safe or unsafe based on the social media content of the publicly available forum Twitter [3] and list of the cities of U.S from Wikipedia [4] by utilizing the data mining techniques such as data collection and preprocessing, Support vector machine(SVM) and logistic regression for classification and visualization. The crimes taken in consideration have been categorized into two broad categories of violent and property crimes. We have collected our data from twitter in the form of geotagged tweets and considered the cities and crimes from Wikipedia. After collection of the data for the training set and labelling of 2000 and more tweets as related or not, we classified them using SVM and predicted another set of tweets for enhancing out training set. After generating the training and testing set and their respective feature vectors and taking our class label as Safe (1) or unsafe (0), we classified our model using SVM and logistic regression. This helped us in analyzing the cities as safe or safe based on the crime rate and also generate visualizations in the form of bar graphs, word cloud and spatial maps. Our implementation is elaborated in detail in section IV. The model proposed achieved an accuracy of 94.11 percent. Therefor our model can also help in granting police resources efficiently based on the intensity of the crime rate of cities and make our communities a better place for living.

Our paper is divided into eight sections including- Introduction, Related Work, Proposed Approaches, System Design and Implementation, Results and analysis, Limitations and Challenges and finally Conclusion and References.

II. RELATED WORK

There has been countless of work done in the field of crime detection and analysis using data mining techniques. Huge amount of data sets has been reviewed and collected based on locations and crimes to help the law enforcement agencies and the society. As shown by Nath [5], k-means clustering is used to aid the process of identification of crimes and semi supervised learning technique for knowledge discovery. Instead of using k-means clustering, we have collected our data in the form of geo-tagged tweets from various cities in the United states and manually labelled them for classification. There has been another approach for crime prediction based on crime types and criminal hotspots for two cities of United states by Almanie, Mirza and Lor [6]. Their approach helps in identifying the future crimes for the cities. We decided to categorize our crimes in property and violent based using Wikipedia and increase our scope for several cities. Another approach by Kiani, Mahdavi and Keshavarzi [7] takes into account the optimization of outlier detection and also provide weights to the features to improve accuracy of their classification model. We initially planned to follow the

approach of assigning weights, instead we used two support vector machines to improve our model accuracy. Malathi, Baboo [8] have also based their approach on crime pattern detection and prediction. We have also focused on classifying the cities in consideration as safe or unsafe in interest of the public and society. Agarwal, Nagpal and Sehgal [9] have also performed crime analysis using k-means clustering and rapid miner tool.

III. Proposed Approach

In this section, we describe the approaches we have used to achieve our goal. Below is the list of approaches used for designing our model and analysis the crime to predict the city as safe or unsafe. The detailed description of our approach is given in section IV.

- We have used Twitter API for collecting our data as geo tagged-tweets and preprocess it for fetching the text related to crime and location as the city of United States with reference to Wikipedia
- After collecting the data, we have labelled 2000 and more tweets as positive or negative related to the crime and referred as the Training set
- We trained an SVM model based on the above training set to predict labels for more tweets and add them to the Training set for further enhancement.
- After generating the Training set, we collected some more tweets for the other set of cities as the Testing set. We trained another SVM and a Logistic regression model on our Training and Testing set.
- After data collection and classification, we labelled the city as Safe (1) or Unsafe (0) and visualized bar graphs, word clouds and spatial maps for our analysis.
- The analysis from our graphs word clouds and spatial maps explained in Section V

IV. System Design and Implementation

In this section, we describe each and every component of our model for the crime rate analysis. We begin by describing the proposed architecture of our system, description of the datasets used, and the tasks performed in our model as the implementation. The section is divided into three sub-sections. Below is the detailed description of each sub section:

A. Proposed Architecture

In this section, we propose the architecture of our system in the form of the flowchart "Fig. 1". As mentioned above, our analysis aims to predict the cities of United States as safe or unsafe based on the twitter data. Our system is based on Python 2.7 and makes use of the libraries such as twitter, sys, json, datetime, geopy, sklearn for SVM and Logistic regression, GridSearch and cross validation. For visualizations and other purposes, we have used NumPy, Matplotlib, word cloud, plotly and pylab [13-17]

The data is collected using the Twitter REST API [3] and python. The set of the cities have been from the Wikipedia [4]. After collecting the data based on the categories of crime such as violent and property crimes, we label the tweets as positive or negative making the Training set. We train an SVM model on these tweets to generate more labeled tweets for the Training set.

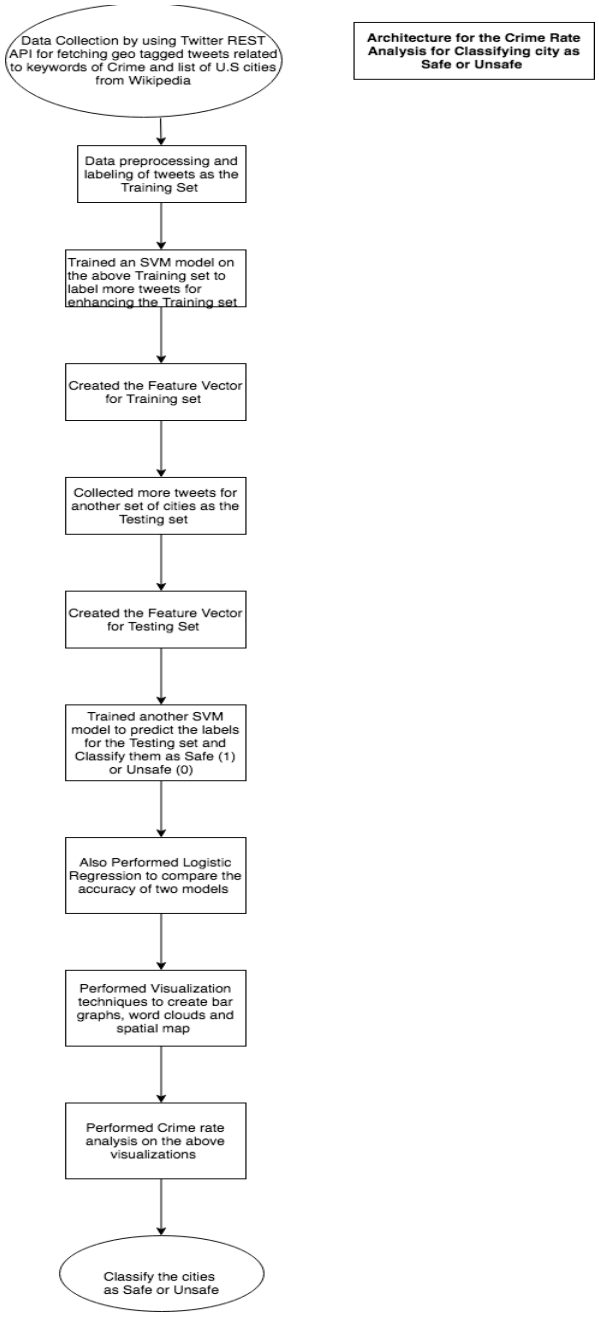

Figure 1 Architecture of Crime Rate Analysis [11]

In our experience and based on literature survey, the model gives a better accuracy based on the amount of Training set. Following, we create the feature vector where each row is the city and each column is the crime keyword.

Our class label is Safe (1) or Unsafe (0) with crime keywords being the other labels. We proceed to fetch more tweets for another set of cities as the Testing set and create the similar feature vector. We finally train our SVM model based on the Training and Testing set and predict the labels for the Testing set. Based on our judgment and the crime rates already given on Wikipedia [4], we classify the testing cities as Safe (1) or Unsafe (0). In order to make our analysis better, we have also performed data visualization to create bar graphs, word clouds and spatial maps. The analysis from these visualizations is explained in Section V. We now explain our datasets, query used, and the list of crime related key words.

B. *Characteristics of DataSet*

The first data set used in our system is the data from Twitter [3]. We have fetched the twitter data using the python Twitter API [3] in the form of .txt file. In order to use the REST API, we were required to register our twitter accounts and generate the API tokens and keys. The API has a query limit of 180 per 15 minutes, so we collected our data over a period of time. The Search query requires the query format for fetching content from twitter based on location, radius and keywords. Our tweets our geo-tagged and containing crime related keywords. We have used the locations as list of cities categorized as training regions and testing regions and used the geolocator to fetch the longitude and latitude of each location. The format of our Search query is "Fig 2".

Search Query Format:
Centroid (Latitude, Longitude) + Radius (e.g., 50 miles) + keywords

Figure 2 Search Query Format

Our crime related keywords are "Fig 3"

```
crimes_list = ['kill','manslaughter','homicide','murder','assassination',
               'rape', 'molestation',
               'robbery', 'larceny', 'steal', 'stole', 'mugging', 'mugged',
               'hit', 'assault', 'violence', 'punch', 'attack',
               'burglary', 'house break', 'loot',
               'theft', 'shoplift',
               'forgery', 'fraud', '"money laundering"', 'bribe']
```

Figure 3 Crime related keywords

The locations as list of cities for Training and testing set are taken from the Wikipedia [4].

We have also added an attributed in the format of the tweet as for location to store the location of the tweet and it can be modified in future. We have labelled around 2000 and more tweets as the Training set and added more tweets to the training set by training an SVM on it and predicting the labels. We have classified our model by finally training another SVM on the Training and testing set. We have also preprocessed the data to fetch the required text and locations from the tweets and saved them in separate file.

C. *Implementation*

After describing our architecture and dataset, we finally describe each and every component of our implementation. After performing the Data collection and data preprocessing as mentioned above, We first train an SVM on the Training set itself to label more tweets for our originial Training set. This saves a lot of time and manual efforts. After predicting the labels and adding more tweets to our Training set, we created the Feature vector "Table 1". We fetched more tweets for another set of cities as Testing set and preprocessed them. Similarly, a feature vector was created for the Testing set "Table 2" . Here, The training set consists of 51 cities and the testing set consists of 16 cities. Our class label for the SVM is Safe (1) or Unsafe (0) and the other labels are the crime related keywords "Fig 3" . We finally trained our classification model using SVM on these Training and Testing sets and classified the cities as Safe (1) or Unsafe (0) as seen in "Table 2".

TABLE I. THE FEATURE VECTOR FOR TRAINING SET

Training Set	**Crime Related Key Words**			**Safe(1)/ Unsafe (0)**
	Crime 1	***Crime 2***	***Crime m***	
City 1	Frequency of Crime			1/0
City 2				1/0
.. ..				
City n				1/0

[a] (*The class label as Safe (1) or Unsafe (0) and other labels*)

TABLE II. THE FEATURE VECTOR FOR TESTING SET

Testing Set	**Crime Related Key Words**			**Safe(1)/ Unsafe (0)**
	Crime 1	***Crime 2***	***Crime m***	
City 1	Frequency of Crime			0
City 2				1
.. ..				
City n				1

[b] (*The class label as Safe (1) or Unsafe (0) and other labels*)

In order to support our analysis We have also performed Data visualization to generate bar graphs whose X axis denotes the cities and Y axis denotes the frequency of crimes. The Word clouds depict the most frequent crime and the spatial maps also depict the crime rate in various cities of the United States. Our implementation also includes another classification model

known as Logistic regression to compare the accuracy of two classification models SVM and Logistic Regression.With the Classification model SVM, we got the model accuracy of 94.11 percent and with the Classification model of Logistic Regression, we got the model accuracy of approximately 82 percent. So, we chose to proceed further with the SVM Classification Model. We have described our visualizations and analysis in the next section.

V. Results and Analysis

In this section, we describe our visualizations and explain the analysis. We begin by first describing the first form of Visualizations as the Bar graphs. "Fig 4" depicts the Frequency of crime per city as the Training Set and "Fig 5" depicts the Frequency of crime per city as the Testing set. The X axis of the graphs have the set of cities as training and testing sets respectively and the Y axis of the graphs denote the frequency of crimes. The training data has all range of values including high crime rates, no crime rates and no values in between so that the SVM could be trained efficiently. Looking at the dataset, we inferred a threshold of 30 being the frequency of crime rate, in order to classify the set of cities and testing set as Safe (1) or Unsafe (0).

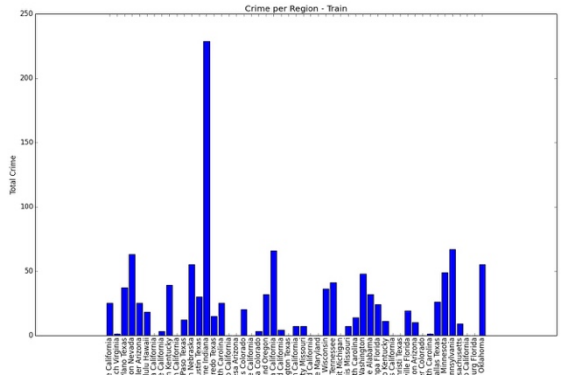

Figure 4 Bar Graph for Training Data

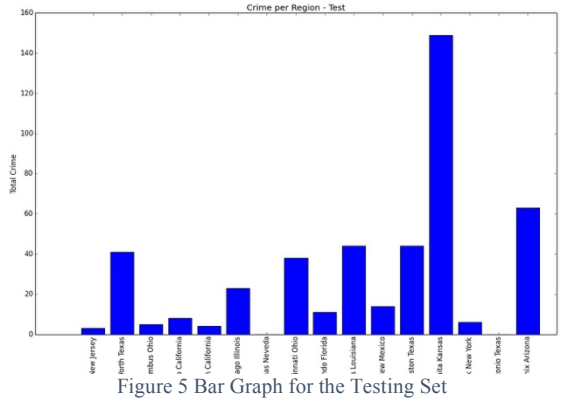

Figure 5 Bar Graph for the Testing Set

We tried to decide the threshold by going through Wikipedia [4] but there is a vast difference between the time periods where the data is collected on the Wikipedia and the tweets of the users, for example, according to Wikipedia [4] some city would be high rated in crime. Whereas, people might not be active on twitter or reporting it on twitter for the time period in which the data was collected. This creates a huge disconnect between the results and produces inconsistent predictions. A fair way to go forward in this scenario was to set a threshold ourselves after carefully observing the data. Furthermore, looking at the crime rates 30 seemed to be a good threshold to classify the cities as safe (1) or Unsafe (0). The next visualization we have generated are the word clouds which give an idea of the most occurring crime in a region, which will help in managing the resources efficiently. Some of the word clouds generated for the Testing set are as follows "Fig 6-8".

Figure 6 Word Cloud for New Orleans Louisiana

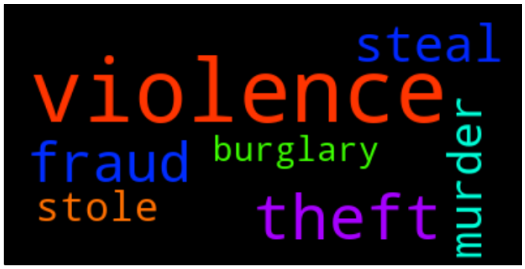

Figure 7 Word Cloud for Phoenix Arizona

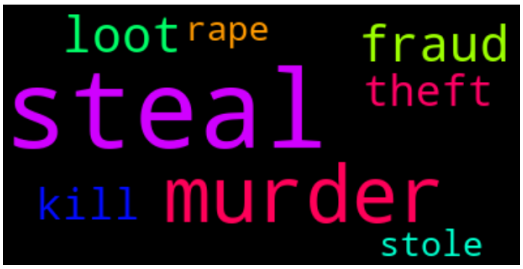

Figure 8 Word Cloud for Wichita Kansas

The word clouds above depict the most frequent crime in each of the cities of United states. Dues to the volume of the graphs, we have included only three of all the graphs in our paper. The first word cloud depicts the crimes in New Orleans, Lousiana. It can be seen from the word cloud that theft is the most common crime in New Orleans. Similarly, violence is quite frequent in Phoenix, Arizona and steps can be taken to curb the crime. For the city of Wichita, Kansas, stealing and murder are the frequent issues.

Our final visualization is the spatial map, which depicts the crime rate of the regions in Training set and Testing set on the map of United States of America. The map for the training set is "Fig 9". The range of 0-30 with the green dots are the regions of lowest crime rate and the range of 51-60 with the purple dots

are the regions with highest crime rates. We have generated these spatial plots using the plot.ly library.

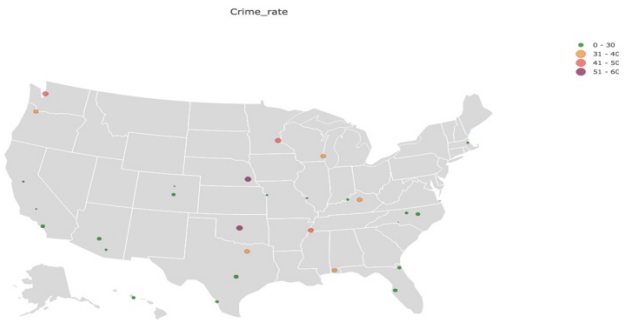

Figure 9 Spatial Map for Training Set

Another spatial plot for the Testing set "Fig 10" depicts the frequency of crime rate for the regions of testing set.

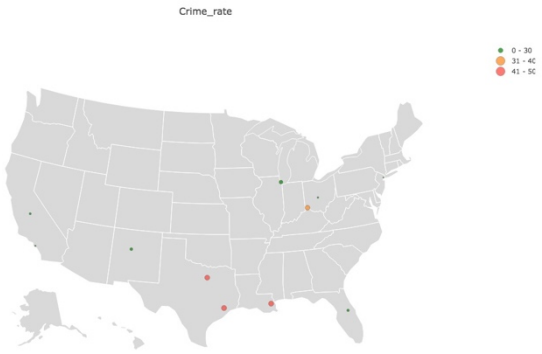

Figure 10 Spatial plot for the Testing Set

In the spatial plot above, the range of 0-30 with the green dots depict the lowest crime rate and can be classified as safe (1) whereas the range of 41-50 depict with red dots depict the highest crime rate and can be classified as unsafe (0).

VI. Limitations and Challenges

One of the main challenges we faced were during our initial approach towards the problem where we tried classifying 5 regions of New York City which made the training data less efficient for classification. Our initial solution to it was to fetch more tweets over a period of time but at the later stages we decided to go with the more efficient approach of including 51 cities of Unites States instead, which gave us a high volume of data. While doing so, we wanted to fetch significantly large amount of dataset to have an accurate model. Our solution to the problem was to build an SVM over our initial 2000 and more labelled tweets and train additional tweets to fetch their labels and include them in the Training set. This helped us get a high accuracy of 94.11 percent while classifying our model on both the training and testing sets.

VII. Conclusion

In this paper, we have analyzed the cities of United states of America as safe or unsafe based on the geo-tagged twitter data and the sources of Wikipedia [4]. We have trained our SVM model on 51 cities and tested it on another 16 cities. We have also compared our model accuracy with logistic regression and decided to proceed with SVM as we got a significantly high accuracy. Our model will help classify the cities as safe or unsafe which may further help the law enforcement agencies to improve and efficiently plan their resources and help the society have a better environment in terms of safety and living. We have supported our analysis with data visualization techniques such as data exploration for bar graphs, word clouds and spatial maps. Therefore, our model can help the law enforcement agencies and the public decide the safety of the cities for their respective purposes.

Acknowledgment

This paper has been written in order to complete the requirements for the course CSI 531 Data Mining. Our work is supported by the guidance of Dr. Feng Chen and his teaching methods which greatly helped us in understanding the nuances of data mining and build an intuitive thinking. Our work is also supported by the Teaching Assistant of the course Chunpai Wang who helped us in deciding the course of this project and give his valuable critiques. We would like to thank them for their support and guidance at every step of the way.

VIII. References

[1] T. Almanie, R. Mirza and E. Lor, "Crime Prediction Based on Crime Types and Using Spatial and Temporal Criminal Hotspots", International Journal of Data Mining & Knowledge Management Process, vol. 5, no. 4, pp. 01-19, 2015.

[2]"Installation & Testing — python-twitter 3.4.1 documentation", *Python-twitter.readthedocs.io*, 2018. [Online]. Available: http://python-twitter.readthedocs.io/en/latest/installation.html. [Accessed: 17- May- 2018].

[3] "Violent Crime Is on the Rise in U.S. Cities", Time, 2018. [Online]. Available: http://time.com/4651122/homicides-increase-cities-2016/. [Accessed: 17- May- 2018].

[4]"List of United States cities by crime rate", En.wikipedia.org, 2018. [Online]. Available: https://en.wikipedia.org/wiki/List_of_United_States_cities_by_crime_rate. [Accessed: 17- May- 2018].

[5] C. facts, "Topic: Crime in the United States", www.statista.com, 2018. [Online]. Available: https://www.statista.com/topics/2153/crime-in-the-united-states/. [Accessed: 17- May- 2018].

[6] S. V. Nath, "Crime Pattern Detection Using Data Mining," 2006 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops, Hong Kong, 2006, pp. 41-44. doi: 10.1109/WI-IATW.2006.55

[7] R. Kiani, S. Mahdavi and A. Keshavarzi, "Analysis and Prediction of Crimes by Clustering and

Classification", International Journal of Advanced Research in Artificial Intelligence, vol. 4, no. 8, 2015.
[8] M. A and S. Santhosh Baboo, "An Enhanced Algorithm to Predict a Future Crime using Data Mining", International Journal of Computer Applications, vol. 21, no. 1, pp. 1-6, 2011.
[9] J. Agarwal, R. Nagpal and R. Sehgal, "Crime Analysis using K-Means Clustering", International Journal of Computer Applications, vol. 83, no. 4, pp. 1-4, 2013.
[10] H. Schulzrinne, "Writing Systems and Networking Articles", Europa.nvc.cs.vt.edu, 2018. [Online]. Available: http://europa.nvc.cs.vt.edu/~ctlu/Course/2018S/CS5604/File/Writing-styles.htm. [Accessed: 17- May- 2018].
[11] "Flowchart Maker & Online Diagram Software", Draw.io, 2018. [Online]. Available: https://www.draw.io/. [Accessed: 17- May- 2018].
[12] "Bubble Maps", Plot.ly, 2018. [Online]. Available: https://plot.ly/python/bubble-maps/. [Accessed: 17- May- 2018]
[13]"1.4. Support Vector Machines — scikit-learn 0.19.1 documentation", Scikit-learn.org, 2018. [Online]. Available: http://scikit-learn.org/stable/modules/svm.html. [Accessed: 17- May- 2018].
[14] "Matplotlib: Python plotting — Matplotlib 2.2.2 documentation", Matplotlib.org, 2018. [Online]. Available: https://matplotlib.org/. [Accessed: 17- May- 2018].
[15]"NumPy — NumPy", Numpy.org, 2018. [Online]. Available: http://www.numpy.org/. [Accessed: 17- May- 2018].
[16] "Python Data Analysis Library — pandas: Python Data Analysis Library", Pandas.pydata.org, 2018. [Online]. Available: https://pandas.pydata.org/. [Accessed: 17- May- 2018].
[17] D. Baidari and S. Sajjan, "Location Based Crime Detection Using Data Mining", Bonfring International Journal of Software Engineering and Soft Computing, vol. 6, no., pp. 208-212, 2016.