Research Article

iMedPub Journals
www.imedpub.com

American Journal of Computer Science and Engineering Survey

2021
Vol.9 No.5:8161

Effect of Inter-State Policy using Crime Data Mining

Pooja Mithoo[1*], **Manoj Kumar**

[1]Delhi Technological University, New Delhi, India

[*]**Corresponding author:** Pooja Mithoo, Ibn Zohr Delhi Technological University, New Delhi, India, Tel: 8279839257; Email: dspooja17@gmail.com

Received date: April 14, 2020; **Accepted date:** September 02, 2021**; Published date:** September 10, 2021

Citation: Mithoo P (2021) Effect of Inter-State Policy using Crime Data Mining. Am J Comput Sci Eng Surv Vol: 9 No:4.

Abstract

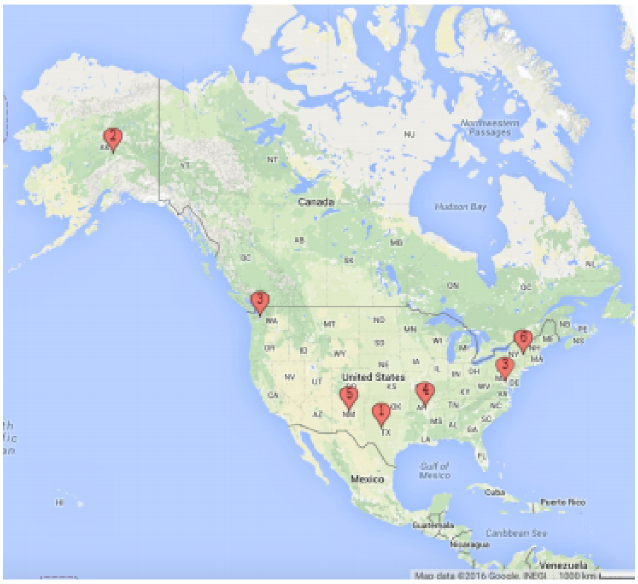
In every nation, there are many states which adopt or influence neighboring or far off states with their policy and criminal activities. Many famous news magazines such as ”Washington Monthly”, ”Washington Post”, ”The New York Times ” and ”The Seattle Times” are filled with the activities related to policies drafted by the department of justice, violent crimes, and property crimes. The question is, ”Can we create a model which can deduce rules stating which two or more states influence each other based on the enormous corpus of information?”. The main hurdle in this process, how to assess the impact of a positive entity ”Policy” over the negative entity ”Crime”. We provide a careful procedure of data transformation followed by rule-based deduction and applied FP-Growth for generating the itemsets. We appreciate our itemsets based on the clustering done on our data set using the category of crimes and states as the centroids. We were able to prove that Washington’s policy affects Alaska 89% of the time and crime in Alaska affect Washington 34% of the time.

Keywords: Data Transformation, Policy, Violent crimes, SVM, Logistic Regression, K-Means and FP-Growth.

Introduction

Over the years, many continents across the globe have implemented policies related to crime, infrastructures, education, democracy, budget, environment, and health. Data Science has been seen as a drift towards crime science since the major terrorist attacks like [1]. These attacks have modified the conscience of the authorities in different nations. These authorities meticulously draft policies and implement them into laws for maintaining peace within the nation. Policies implemented within the state have a sure probability of its effect but are we considering the nation in that policy? This question has been raised in [2] and in this project, we put our thought process towards answering this question. Considering these avenues where policy has played some crucial roles, we see crime is the one, the only sector where a positive element like policy impacts a negative element crime. When we consider the policy of a state, we have an intra-state policy and inter-state policy. Policies are implemented with confidence 1.0 in intra-state but we are left with one open end question ”what is the confidence of influence of inter-state policy”. In this project, we will talk about some of the linguistic paradigms along with data mining paradigm. This project works on an interdisciplinary approach between computer science and policy, more specifically inter-state policy relating to crime. There are two approaches for linking the effect of one state policy on another i.e qualitative[3][4] and quantitative[5]. In this project, we have used a quantitative analysis approach where we have used real data to show the impact of policy drafted for one state on another. Twitter is a popular microblogging service that has many users, which generate many data every moment, through their posts, which express their thoughts, opinions, and preferences. These posts are called tweets. The tweets can contain precious data for analytics because the tweets have hidden within them information like the user’s opinions, preferences (of products, brands, political parties, and others), the user behavior like their day-to-day activities and location.

Figure 1: Geolocations using google maps.



(Figure 1) depicts the geolocation of the states which were considered for crimepolicy relationships. To extract meaningful information from this source of data, we need to apply data mining techniques to reveal patterns to be analyzed in an easy way. We will discuss about some of the crime words used in National Institute of Justice. Following our discussion about the crime related words, we will be focusing about some terms used along with policy in some of the affluent magazines such as ”The Washington Post” and ”The Seattle Times”. We have considered illegal, drugs, murder, kill, rape, crime and assault as the primary labels for our clustering process. These labels were extracted from a corpus of tweets related to crime from The Washington Monthly, The Atlantic, The Texas Monthly. Our primary goal in this project is to device and train a model based on the quantified tweets obtained using some known twitter API's via REST and Streaming processes. We can compare and contrast different states with respect crime and policy. The figure 1 shows the number of states we have considered for our project. We begin our study of the project using some vague notion: there exist a problem space for the problem and later we tend to conclude ourselves using the unsupervised notion of learning to solve the problem. The cluster developed based on the category of crimes and the states supported in deriving rules between the crime tweets and policy tweets from the random sample of the tweets. We provide a disclaimer at the very beginning of the project that Washington and Washington D.C are taken together as we were unable to distinguish them semantically.

Related Work

In this section, we will briefly review some of the past works done about policy, crime, and crime data mining. These reviews form the base for further improvement in crime-related data exploration. We have been seen that policy and crime are usually referred together, but the order can be different. It has been defined in [6] that compute based mapping techniques together with appropriate storage management and tactical analysis provide early warning of any negative happening or provides an intuition of a positive scene. To make crime prediction and analysis real-time, we see that there has been an approach to it in [7], which specifically mentions intelligence-based information systems related to the crime that uses inferences from witnesses and police personnel. Related to crime analysis we have been seeing many advances in forensic intelligence [8] but one constraint in this field is, how to associate crime with a policy so that we can simultaneously address the growth of the state. All the advancements concerning the crime in the past have been through the use of Spatio-temporal information [9] about the area. Such information does not address the policy-related information, hence what we obtain is simply crime as a single entity and not in comparison. In [10], there has been advancement in crime data mining. Through the different stages of development in the nation, there has been development in crime in coherence with policy. This structure in [10] talks about criminal inferences from the crime database. One drawback of such an approach is they don't capture recent or trending news. In a book by [11], it is mentioned that through an appropriate channel of model data design, model design, and transition we can formulate any problem in data mining. Analyzing the large databases of Spatio-temporal information have been assistive in providing some useful predictions concerning the crime, but the twitter information is equally important for validating the crime [12]and also provide information related to policy. So, we build an idea for designing a model for policy and crime relationships and consider the involvement of Spatio-temporal information in the second phase of this project. So, we have seen inference based prediction in crime data mining, but one drawback of rule inference is the size issues. If we have many rules, it would be difficult to comprehend and define fruitful predictions. So in [13], we see an FP-Tree based approach to criminal investigation and pattern identification. This approach is useful as it reduces the amount of traversing the database and hence provide compact rules in terms of an itemset. Since, as stated in [1], that we have large information related to crime, and if policy-related information is added it will double up the time involved. So, the multi-layer neural network can be a solution to such a prediction problem but a naive K-Means clustering together with the rule-based tree is fast and accurate as it is difficult to model vague behavior of crime and policy. We have observed in [14] that, violent crime in Texas (23%), Arkansas (32.6%), Washington is (63%), Albany is (82.4%) and Alaska (83%). A simple inference based on crime will result in Alaska and Albany as the potential pairs whereas if we consider policy, we create links between the states not in terms of crime but also the proximity of their location (figure 1).

ETL Phase

TL is Extraction, Transformation, and Loading. In this section, we will describe in brief essential phases in data collection, processing, and transformation before we initiate the model selection process. Data collection, processing, and transformation is essential and must do concerning the problem we are trying to solve.

Extraction

The data has been collected from the Twitter REST and Streaming APIs. The data obtained using twitter is in a raw format containing many links, multiple spaces, and annotations in the form of hashtags, ”@” and username. If we begin with the process of model selection and preparation on raw data then we will be encircled with textual data containing extraneous symbols and will hence divert our rules deduction. This data extraction process is one of the challenges in the field of data mining as the identification and filtering of outliers and unwanted symbols are essential to formulate a precise data structure needed for the classifiers [15]. Twitter raw tweets have nearly 16 attributes of which the main attribute in this project is the text. For instance: ”text”: ”Mexican rival demands vote recount: The leader of Mexico's leftist Party of the Democratic Revolution, Andres Ma... http://t.co/cHqH0dhP”, in this text we see hyperlinks Http, which are the source of this tweet that has been generated using the UR shortened routine. Considering a bunch of raw tweets similar to this and make a word cloud, we will see that huge words which correspond to high frequency are covered by these hyperlinks. As a result, the main text is shown in small and is negative towards deriving necessary conclusions. So, to improve the data and remove extraneous pieces of information we incline ourselves to the data transformation process.

Transformation

Transformation is defined as the process of converting the raw data into a format that can be used as an input to the classifier so that some useful inferences can be derived. As stated in [16], the transformation process develops formal relationships from the raw data, which are used later to develop conceptual structures using unsupervised [17] and supervised learning [18]. Regular Expressions: Within the perspective of this project, we constrict ourselves to the application of regular expressions (RE) [19]. We have developed some AWK scripts which helped in removing extra spaces, dummy symbols, and hyperlinks. As we will describe natural language toolkit (NLTK) [20] in section 5, which mandates that our textual input should have words from some specific language and should be free from extraneous literals. We will briefly define in section 5, our process of creating the vocabulary necessary for converting our textual data to some real number data, that will serve as an input to the model. Outlier Detection: Within the periphery of data mining, there has been the development of many algorithms relevant to outlier detection and removal. Such as energy of the graph [21] and statistical technique [22]. We are more inclined towards simple central tendency based outlier detection and removal. Our outlier detection process uses standard deviation and means for identifying divergent tuples in the transformed text. Mean is the central measure of the tendency which involves the summation of sentiment vector of the tweets over the total number of tweets. The mean value of the tweets helps in identifying the standard deviation of each tweet. This statistical measure can be seen as a process of selecting those tweets which are above 10%percentile and below 90%percentile which is termed as the Interquartile Range (IQR).

$$T_{vector} = S_{score}(positive_{sentimental}, negative_{sentimental}) \qquad (1)$$

We formulate the twitter data as a 1-D vector (Tvector) of positive and negative sentiments. Now the question is, why we haven't considered neutral sentiments? The answer to this question is described in section 5. Each transformed tweet from regular expression is quantified using equation 1.

$$\mu = \frac{1}{N}\sum_{i=1}^{N}(T_i - \bar{T})^2 \qquad (2)$$

$$\sigma = \sqrt{\mu} \qquad (3)$$

The process of outlier detection involves identifying the central value of the data and check variation in the sample from 1.....N. The variation procedure is similar to the hidden Markov model (HMM) method which uses covariance [23]. So we define our method of detection of outliers using equations 2 and 3. These equations provide the range which is in term of medians, call IQR.

$$Q1_{outlier} < \frac{1}{N}\sum_{i=1}^{N} T_i - \sigma \qquad (4)$$

$$Q3_{outlier} > \frac{1}{N}\sum_{i=1}^{N} T_i + \sigma \qquad (5)$$

We have provided an analogy, equation 4 and 5 works similar to 1.5 rule [24] of medians for outlier detection where Q1 is the lower quartile (10%percentile) and Q3 is the upper quartile (90%percentile). As per the above equation, we can create a box-like structure in which all the tweets are transformed and filtered tweets that are used in section 6. So we have filtered the tweets using regular expression and transformation, we removed those tweets which will result in poor training of the classifiers if we are performing supervised learning or poor and inaccurate clusters if we confide to clustering.

Loading

After we have performed the required work in the above two sections, we move towards the step of loading or storing of the tweets. This step is essential for getting accurate and interpretable results in this paper. There are various formats for storing the data such as comma-separated values (CSV), tab-separated values (TSV), excel spreadsheets (XLS), JavaScript object notation (JSON), and ASCII text. Of these formats, CSV/TSV and JSON are some of the recommended formats in natural language processing (NLP) society [25].

Data Mining Techniques

This section will describe in brief some of the prominent data mining techniques used in this project with a specific focus on the crime mining procedure. Mining of crime-related data requires models that can group the data based on some gaussian surface which can embed data in a vectorized form consisting of entities that are related to one another.

$$Ntweets = \phi_i * \epsilon_o \qquad i_{\epsilon} 1 to \epsilon_o \qquad (6)$$

In equation 6 we have shown that the complete crime mining process comprising of clustering or classification together with rule-based deduction can be mapped to a gaussian surface with ϕi amount of charge 0per clusters. Such a formulation provides an intuition for solving the problem. We will first be considering the classification technique followed by clustering and rule-based technique.

Classification : Logistic Regression

Logistic Regression is a regression-based classifier using a sigmoid function for its classification process. We know from [26] that a classification is a process of developing subset from the superset using some activation function which will provide an output equivalent to the labels in the problem. Logistic regression is an improvement over the linear regression [27] with the involvement of a radial basis kernel.

$$f(Tvector) = \frac{1}{1 + e^{-T_{vector}}} \qquad (7)$$

$$Mclassifier = \beta M_{classifier} + \alpha\frac{1}{1 + e^{-T_{vector}}} \qquad (8)$$

The function represented by equation 7 is an activation function in logistic regression which is continuous and differentiable. Since the activation function in logistic regression is differentiable, there is a certain amount of assurance that the classifier if trained with the correct amount of iterations and minimal errors will not fall into saddle points in the function space of the problem. Learning momentum (Mclassif ier)is defined as the velocity of the learning of a classifier. Higher the velocity signifies faster learning but with a note of caution: the classifier should not be stuck in the saddle point in the graph. Mclassif ier can have a negative value as well as a positive value depending upon gradient descent or gradient ascent. We have considered the parameters of learning rate α and learning momentum rate as β in equation 8 for the classifier but we haven't seen a significant amount of improvement in the classification. This was the reason we started with classification.

Cross Validation: Cross-validation (CV) is an essential procedure incorporated in before the training of the classifier. It is a technique to evaluate the learner (model) concerning the prediction it provides. Since CV segments the complete sample into training and testing states, we can judge the learning ability of the classifier by its predict accuracy from testing sets [28]. So, CV is a method of creating subsets from the set, so that classifier is trained using one of the subsets and is tested another subset. This is a reason to have a large number of the samples in the set as it will create a reasonably good classifier. Another parameter is the number of iterations of the classifier. In equation 8, i in βi depicts the iterations of the classifier. Another note of caution is A very low (< 30%) amount iterations is random learning and very high (> 90%) iterations are overfitting (these values are relative to the sample).

Clustering: K-Means

Clustering is similar to grouping things according to some similarity values [29]. The range of a set of values is used to find the median of the set also termed as centroid, which is the representative element of the set [19]. Now the question is, how to define K in K-Means? For the sake of brevity, we provide a short description of the methods used in the project for identifying K. Apart from the mentioned methods, other methods for identifying K are in [30]. Elbow Method: Elbow method is one such metric for identifying K for the K-Means algorithm. Initially, the method finds the intra-cluster distances for different values of K. For each K values, it normalizes the sum of all intracluster distance. Elbow method finds the k K values at which the elbow is observed. This k is the K for K-Means.

$$ElbowNC_k = \sum_{k=1}^{K} \frac{1}{2 * C_k} \sum_{j=1}^{C_k} \sum_{i=1}^{C_k} \parallel T_{vector_i - T_{vector_j}} \parallel \tag{9}$$

$ElbowNC_k$ is the Normalized Cluster value for a k ϵ K.

$$ElbowNC_{k_{optional}} = \underset{i \epsilon C_{k-1}}{arg} \ maxf(\Theta) = tan(\Theta_i) \tag{10}$$

In the figure 2, we show that K in the range (7,10) Z is the appropriate value for K-Means algorithm. From the figure 2 it is evident that to find ElbowN Ckoptimal , we need to use equation 10 over the curve in the figure 2.

Silhouette metric: Another metric that can be used for identifying the k in K-Means is the silhouette metric. The silhouette metric is the opposite of Elbow method.In this metric we replace, arg maxick−1 with arg minick−1 .The inversion of figure IV-B is the the silhouette graph. In the silhouette metric we look for the lowest point in the graph. In the figure 3, it is evident that K in the range (7,10) Z is the appropriate value for K-Means algorithm and is coherent with figure 2. We define the silhouette score as :

$$SNC_k = \sum_{k=1}^{K} \frac{1}{2 * C_k} \sum_{j=1}^{C_k} \sum_{i=1}^{C_k} \frac{T_{vector_i - T_{vector_j}}}{max(T_{vector_i - T_{vector_j}})} \tag{11}$$

SNC_k is the Normalized Cluster value for a k ϵK using silhouette metric.

$$SNC_{k_{optional}} = \underset{i \epsilon C_{k-1}}{arg} \ minf(\Theta) = tan(\Theta_i) \tag{12}$$

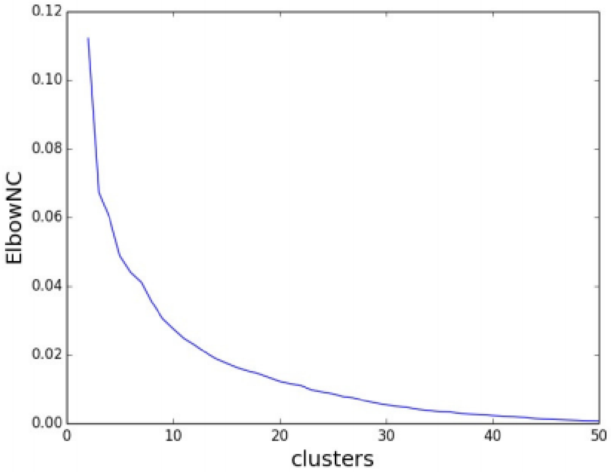


Figure 2: Elbow curve with 7<K.

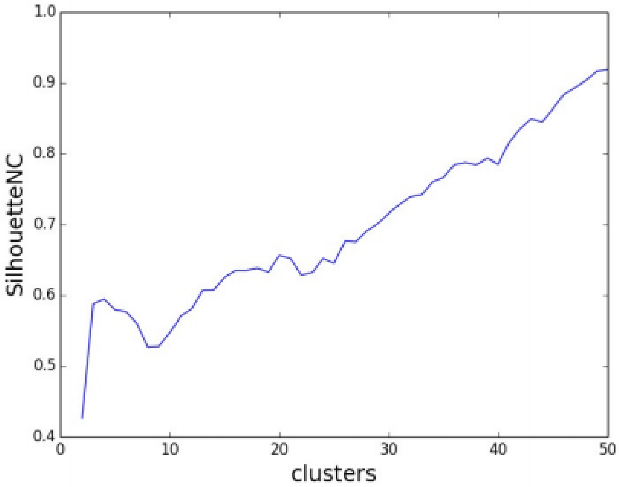


Figure 3: Sihouette score graph with 7<K.

Comparing the equations 9 and 11, we observe that, max Tvectori , Tvectorj is the factor due to which the silhouette is inverted as compared to elbow method. The mathematics behind formulation is extensive as stated in [20]. This factor is replaced by

$$\sqrt{T^2_{vector_i} - T^2_{vector_j}}$$

Frequent Pattern Growth

Frequent Pattern (FP) growth is an algorithm in data mining which creates a tree based on association rules developed from association rule mining (ARM). FP-growth uses the frequency parameter of the item sets for creating the tree. In the procedure of creating the FP tree, the algorithm embeds essential modules from association rule mining. In order to create the FP tree, the algorithm first generates rules based on the clusters. We threshold the rules based on the confidence values and hence, those rules which are alive form the FP tree [31]. The procedure for association rule mining is stated in the following equation :

$$support \geq \underset{support}{min} \tag{13}$$

$$conf_{p->s} = \frac{support_{(p \sqcup s)}}{support_s} \geq \underset{support}{min} \tag{14}$$

FP-Growth algorithm has the potential to compress a large database into a compact Frequent-Pattern tree (FP- tree) structure which is highly condensed, but complete for frequent pattern mining. This algorithm also avoids costly database scans, which matters when we have huge clusters and over a billion samples. From equations 13 and 14, it is derived that the FP tree method with ARM guarantees healthy prediction for any complex problems. We are now in a position to validate this section IV with results and appropriate discussion.

Results and Discussion

In this section, we will be providing a discussion on the results obtained after the application of algorithms on the structured sample obtained. In the figure 4, we observe that there is a good amount of clustering being shown using the positive sentiment value of the Tweets. The policy-related tweets influence positive sentiments and the crime related tweets influence negative sentiments. Since we are focusing on policies that affect the crime, we are more inclined towards those tweets which form clusters in 6. The cluster of tweets formed in figure 5 is crime-related tweets. This cluster provides the very foundation of the rules that will link a state to another state using policy and crime keywords. Observing figure 6, we can nearly say that there are states which influence each other in terms of policy and crime but it can also be possible that we are observing the cluster of state and not states concerning policy and crime.

Figure 4: Cluster of tweets with positive sentiments.

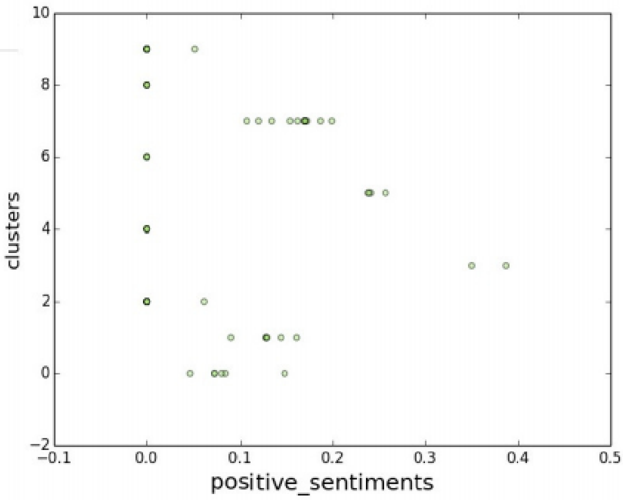


Figure 5: Cluster of tweets with negative sentiments.

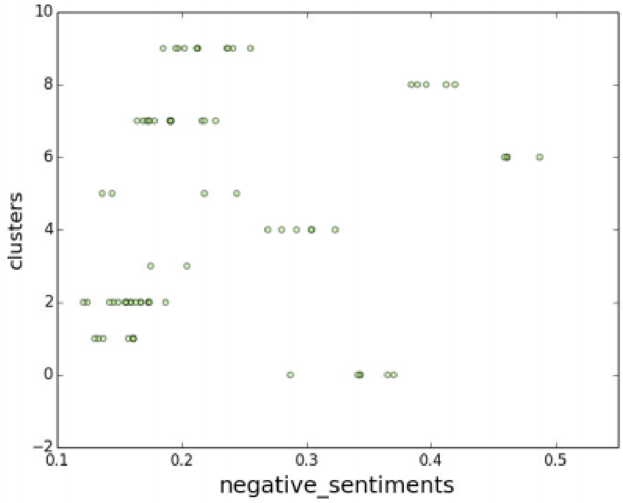


Figure 6: Cluster of tweets with neutral sentiments.

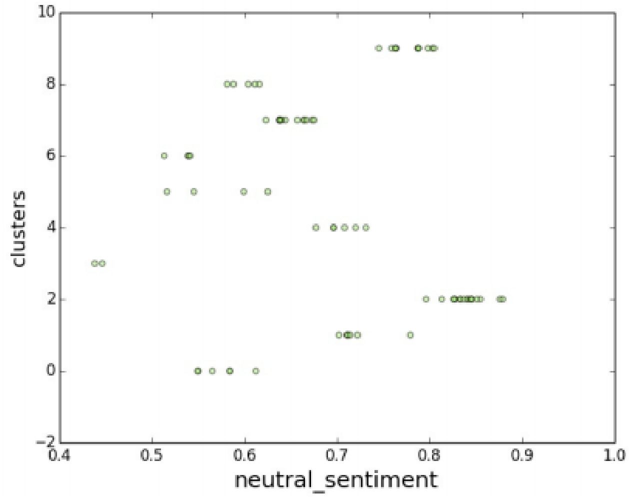
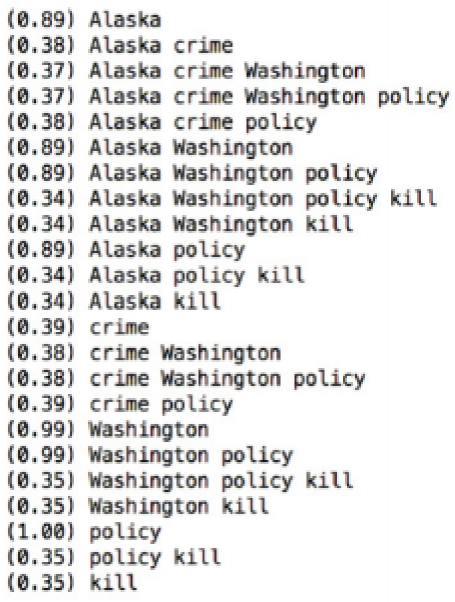


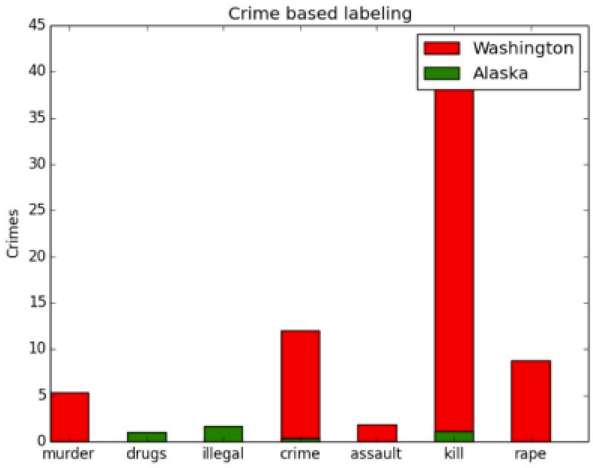
Figure 7: A structure of FP-tree based on neutral sentiments with confidence values.



```
(0.89) Alaska
(0.38) Alaska crime
(0.37) Alaska crime Washington
(0.37) Alaska crime Washington policy
(0.38) Alaska crime policy
(0.89) Alaska Washington
(0.89) Alaska Washington policy
(0.34) Alaska Washington policy kill
(0.34) Alaska Washington kill
(0.89) Alaska policy
(0.34) Alaska policy kill
(0.34) Alaska kill
(0.39) crime
(0.38) crime Washington
(0.38) crime Washington policy
(0.39) crime policy
(0.99) Washington
(0.99) Washington policy
(0.35) Washington policy kill
(0.35) Washington kill
(1.00) policy
(0.35) policy kill
(0.35) kill
```

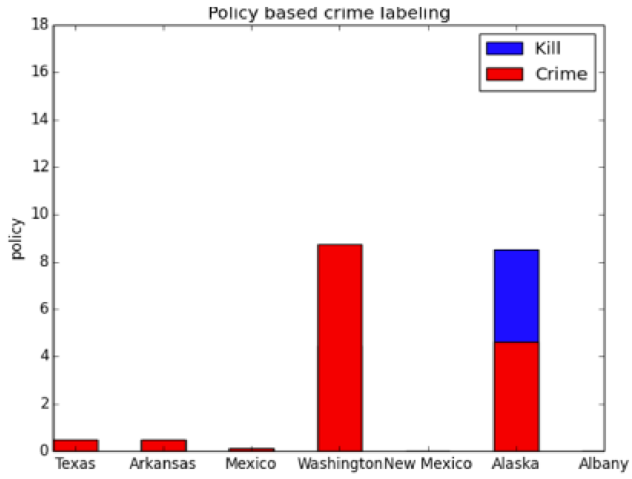
From the figure 7, we say Alaska, Washington, policy, crime arrives 34% of the time which is quite sufficient with to answer the question asked in section I. Furthermore, we are able to show that Alaska, Washington, policy arrives 89% of the time which also support our answer to the question in section I. Lastly, we can able to prove that Alaska, Washington, crime arrives 34% of the time. So we can say with high assurance that Washington and Alaska (W&A) influence greatly. In order to magnify our observation, we used the FP-Tree technique to illustrate those clusters which include intra-state and not inter-state. Since we are focusing on those clusters which are present in neutral sentiments of the tweets, we formulate an FP-Tree based on support and confidence of the rules generated using ARM. We place a threshold of 30% over the support and 30% over the confidence for generating the rules. Observing the FP-Tree, we achieve a strong predisposition towards 2-states, Washington, and Alaska in terms of policy and crime confluence.

Figure 8: Bar plot of Crime based labelling in Washington and Alaska after FP-Tree.



As we have proved that W&A influence each other w.r.t crime and policy, it is evident from figures 8 and 9 that the crime-related tweets specifically property crime and killings related tweets are high in Washington and whereas drugs and illegal jobs related tweets are predominant in Alaska. It can also be observed that Alaska shows a significant number of tweets (negative clusters in figure 5) with respect to crime and killings. Furthermore, we analyze the bar plot in figure 9 which is more influenced by (figure 4).

Figure 9: Bar plot of Policy based labelling in Washington and Alaska after FP-Tree



In (figure 9), we observe that W&A has a significant contribution to the implementation of policies against crime. In respect to section III, we made sure that when we talk about policy, these are not environmental, health, or transportrelated policies. From figure 9, we also observe that Texas, Arkansas, Mexico, New Mexico, and Albany lack tweets related to policy and crime. Though these states have tweets related to crime but when ”policy” word is augmented the confidence values steps down, hence we don’t see results from these states in (figure 7).

Conclusion

lies in proving this linkage between policy and crime using some of the common data mining techniques. The project develops a strong understanding of the need for ETL, model selection, and rule-based inferences for solving any complex and diverse problem. We aimed at providing linkage between states in the USA because, we have observed over the time through different monthly and quarterly magazines, that various state and federal organizations are involved in maintaining peace within the state and neighboring states. Our rule-based deductions sufficiently prove that Twitter tweets play an eminent role in providing useful inferences about the working of the state and its environment. This project is a preliminary approach to policy research but provides an appropriate foundation for further study. We have used limited classification and clustering techniques in their basic form and didn’t include hybrid algorithms. Furthermore, we have restricted ourselves to simple and structural learning without incorporating randomness while making the classifier learn. Lastly, the problem can be formulated in computational intelligence and it can be assumed that much better patterns in terms of rules can be created using Heuristic, Meta-heuristic and Fuzzy methodologies. Also, we haven’t dealt deeply into the computational linguistic perspective of the problem, so this solution though worthy of giving the answer to question in section 1, is still silent from linguistic creativity.

References

1. Josef, [Online], http://list25.com/25-worst-acts-terrorism-committed/(2014).
2. University of Missouri-Columbia, project.asp?projectID=387, (2016).
3. H. L. Larsen, J. M. Blanco, R. P. Pastor, and R. R. Yager, Using Open Data to Detect Organized Crime Threats: Factors Driving Future Crime. Springer, 2017.
4. Lorenc, T., Petticrew, M., Whitehead, M. et al. Fear of crime and the environment: systematic review of UK qualitative evidence. BMC Public Health 13, 496 (2013).
5. Bennett, T., & Brookman, F. (2009). The Role of Violence in Street Crime: A Qualitative Study of Violent Offenders. International Journal of Offender Therapy and Comparative Criminology, 53(6), 617–633.
6. C. R Block, M. Dabdoub, S. Fregly, ”Crime Analysis Through Computer Mapping”, National Criminal Justice Reference Service, (1995).
7. O. Ribaux, P. Margot, ”Inference structures for crime analysis and intelligence: the example of burglary us-ing forensic science data”, Forensic Science International,(1999). 8. Del´emont O., Lock E., Ribaux O.,” Forensic Science and Criminal Investigation”, In: Bruinsma G., Weisburd D. (eds) Encyclopedia of Criminology and Criminal Justice. Springer, New York, NY,(2014).
8. H. Chen, W. Chung, J.J. Xu, ”Crime data mining : a general framework and some examples”, IEEE Computer, (2004).
9. Vladimir Estivill-Castro, Ickjai Lee, ”Data Mining tech-niques for Autor=nomous Exploration of Large Volumes of Geo-referenced Crime Data”, Proc. of the 6th Interna-tional Conference on Geocomputation, (2001).