Assessing the Impact of Mining Techniques on Criminal Data Quality

Lucas Zanco Ladeira[1], **Matheus Ferraroni Sanches**[1], **Cassio Viana**[2], **Leonardo Castro Botega**[2,3]

[1] Computer Networks Lab
State University of Campinas (UNICAMP)
Cidade Universitária "Zeferino Vaz", Campinas - SP

[2]Human-Computer Interaction Group
University Centre Eurípides of Marília (UNIVEM)
529 Avenida Hygino Muzzi Filho, Marília - SP

[3]Information Science Department
State University of São Paulo (UNESP)
737 Avenida Hygino Muzzi Filho, Marília - SP

lucaszl@lrc.ic.unicamp.br

Abstract. *Crime data refers to crime events reported in natural language to the emergency response center of the police forces. Furthermore, it comprehends the textual description of the event and may be utilized to understand what characterizes crime situations, considering weapon of crime, objects stolen, criminal activity and more. In this work it's applied data pre-processing, transformation and mining techniques to discover hidden crime details in the dataset relating similar records. Consequently, the crime records are classified into 3 groups considering the sophistication of the criminal action, being: A (low sophistication), B (medium sophistication), or C (high sophistication). To find out the impact of absence and usage of pre-processing techniques and which data mining technique achieves the best results, two experiments were performed and had their mean accuracy compared. The usage of pre-processing and Random Forest algorithm achieved better results and also the capability of understanding a high dimensional and dynamic data. Consequently, the joint of these techniques can provide better information to police forces.*

1. Introduction

Textual data is being generated all the time by social network users, news websites, e-mails, crime records, and more. According to Noyes [Noyes 2018] every 60 seconds 510,000 comments are posted and 293,000 statuses are updated. This creates a huge amount of data to be mined, serving as an input to computational systems that analyze and present visualizations to users. For instance, a system that analyzes crime data searching for patterns, and provides a time-line visualization with the growth or decrease of crimes considering different regions, therefore it is usually used at police stations. Many works may be found in the literature that proposes the prediction [Azeez and Aravindhar 2015, Sivaranjani et al. 2016, Bogomolov et al. 2014,

Babakura et al. 2014, Tayebi et al. 2015, Aghababaei and Makrehchi 2016] or classification [Kumar and Gopal 2015, Jung and Yoon 2015, Aljrees et al. 2016] of crime data. In addition, it's important to consider the visualization of the crime incidence [Ballesteros et al. 2012, Schünke et al. 2014].

Considering crime records, it's necessary to point out some characteristics of the textual data comprehending that people with different levels of education describe the same situation differently, maybe at the stress level of a robbery, and words usage changes according to region. In addition, concerning the data quality it's a HUMINT data what implies that data may be incomplete, imprecise, spurious, and more. To mine this information it's necessary to apply pre-processing, transformation, and data mining techniques. Pre-processing improves the data quality by removing words that doesn't contribute to the characterization of the scenario. Transformation quantifies each word utilized as a number so that data mining techniques may understand. Finally, data mining techniques discovers hidden information and patterns to provide to an user or a system [Silva 2016]. A illustration of each step described may be observed at Figure 1

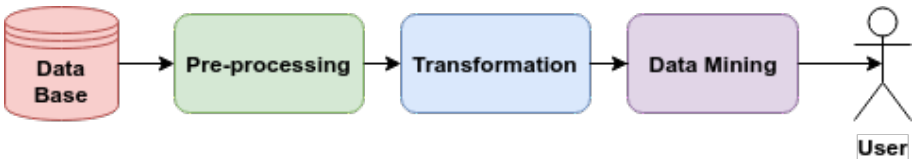

Figure 1. Pre-processing, transformation and data mining flow

The data mining may provide a typification of similar crimes implying that crimes of the same type will have a similar way to contain. In this work the typification of crime records is utilized to characterize similar events and discover, by experiments, which technique works better with textual data. The existence or absence of pre-processing techniques are compared, and utilized the accuracy of data mining techniques to identify the best strategy. Furthermore, the data mining step uses machine learning classification algorithms to classify records into each crime event type, and a confusion matrix is generated to evaluate their understanding of the data distribution of each type.

This work is organized as follows: section 2 shows articles that utilize textual data and reference the techniques used. Section 3 presents the strategies utilized to pre-process textual data, and the implication of each technique into the data quality. At section 4 the experiments are presented, and their results described and visualized. Finally, at section 5 the results of the techniques are discussed, and proposed new ways to pre-process the data.

2. Related Work

The characteristics of textual HUMINT data were already presented, so it's necessary to focus on the techniques of pre-processing, transformation, and data mining. Where these techniques are applied with distinct combinations and achieve different results according to the input data distribution.

George [George et al. 2012] tries to solve the problem of categorizing web surveys. The dataset of surveys were built with the SurveyMonkey tool that provides specific

models comprehending various categories. To pre-process data were utilized Natural Language Processing (NLP), TinySegmenter, and MeCab, comprehending each technique for a group of distinct languages.

To classify each survey the questions were considered by the usage of a variation of the BOW (Bag of Words) transformation technique called TFIDF (Term Frequency Inverse Document Frequency), which only considers the occurrence of the words ignoring the appearance order. To divide the dataset into classes was utilized a clustering algorithm based on Fuzzy Clustering represented into a Latent Semantic Indexing (LSI) space. The languages French got the best result at almost all the survey types, while portuguese had the worst, however it had the least number of surveys considered.

In the work described by Vyas [Vyas et al. 2015] the importance of emails on internet is highlighted, along with the problems caused by spam. Spam emails are common and can cause damage to hardware or even be used to steal sensitive data from victims. The work described how some actual mechanisms work to filter spam emails. Some of the techniques utilized to perform this filtering are based on machine learning algorithms. In this work tests of efficiency, performance, and time were realized with 6 techniques of machine learning using WEKA. The following techniques were evaluated: K-means, Decision Trees (DT) ID3, Decision (DT) Trees J48, Naive Bayes (NB), Support Vector Machine (SVM) and Multilayer Perceptron (MLP). Also to pre-process data the techniques removal of stop words and stemming were applied. In addition, the transformation technique called TFIDF (Term Frequency Inverse Document Frequency) were used.

The dataset utilized during the tests is composed by 1020 emails, 44% are spam. The objectives were performed the correct identification of spam and decrease the rate of false positives and false negatives. According to the results presented in this paper it is possible to realize that NB can achieve good results over other techniques (except SVM and DT ID3). The techniques SVM and DT ID3 were more precise during the results, however they take a longer time to be trained.

Anzi [Al-Anzi and AbuZeina 2015] performed a study of stemming impact on Arabic text categorization performance with the objective to decide if there is an agreement upon the stemming performance on Arabic text categorization. The stemming process can be summarized as the process to find the lexical root or stem. According to the author the use of stemming is not the best choice for feature reduction, the author also says that stemming can enhance the results in some cases and be worse for others. It was suggested during the conclusions that the artificial intelligence used can try to detect the words contexts before categorization process.

The work developed by Aghababaei [Aghababaei and Makrehchi 2016] studied how information from Twitter can provide social-behavioral signals to predict crime rate. Using the content inside the posts, not just the crime cause may be identified but signals to predict future incidents. The example tweets used during the training step were labeled with crime trends information. The dataset was constructed of data from 4 cities from United States, Chicago, Philadelphia, San Francisco and Houston. A total of 101 million tweets were collected to this experiment.

As pre-processing were applied stemming, removal of stop words, and low frequency removal at the data, consequently the time-series were smoothed to increase

predictability and reduce noise. To quantify data the techniques BOW binarization and TFIDF were used. The machine learning algorithm Support Vector Machines (SVM) was applied to classify data into groups. Finally, this work presented a new model of prediction that was able to identify, using the content of tweets, if the rate of crimes in a certain region is increasing or decreasing.

The works presented use pre-processing techniques comprehending the removal of stop words, stemming, feature selection, low frequency removal, and more. These techniques try to identify which words are important to the scenario, by analyzing their frequency, previous knowledge about the words, and analytical strategies. Machine learning algorithms found comprehend Support Vector Machines, Multilayer Perceptron, Decision Trees, Naive Bayes, and more.

3. Data Quality on Feature Selection

To begin with, it is necessary to explain how textual data is represented as numbers so machine learning algorithms may comprehend. Following the Knowledge Discovery [Piatetsky-Shapiro 1991] steps it refers to the step of Transformation, which applies techniques to represent the data differently. In this step the data is already divide into two sets being training set and testing set, the set used to train the technique and the set used to verify the accuracy, respectively.

The most common Transformation techniques are: Bag of Words (BOW), Bag of Words Binarization (BOWB), Term Frequency Inverse Document Frequency (TFIDF). The technique BOW may be formally described as: Let $\tau = \{\eta_1, \eta_2, ..., \eta_n\}$ be the set of all the words found at the training set. Then $\upsilon = \{\nu_1, \nu_2, ..., \nu_n\}$ the set of sums of all the words frequency found at the training set, where ν_1 refers to the word η_1, ν_2 refers to η_2, and so on. The quantification of the test data record comprehends a set $\theta = \{\nu_1 * \mu_1, \nu_2 * \mu_2, ..., \nu_n * \mu_n\}$ where μ_i is the count of the word η_i in the specific record.

Consequently, for each word found at the training set there is an index on each set. At data mining, usually are used machine learning algorithms that consider these indexes as features or dimensions. There are techniques that analyze the whole dataset and understand which features doesn't help to characterize the records and remove them (feature extraction), in addition other techniques analyze the data and maps it to a smaller dimension, for instance, the algorithm called Principle Component Analysis (PCA). PCA uses linear transformation to map the original feature vector into a smaller dimension vector, where the biggest variance is mapped to the first feature, the second biggest is mapped to the second feature, and so on. Some algorithms have better results than others with high dimensional data, also considering their representation strategy, and feature selection technique.

Considering data quality it's compared techniques of preprocessing that improves data quality and have impact in the number of features. The first technique is called removal of stop words, where stop words represent words that doesn't contribute to the characterization of the record. For instance, words as: of, the, at, in, on. These words may be removed from the dataset, disregarding one feature for each word removed. There is a technique called stemming, the aim of this technique is to represent similar words as the same getting the root of each word. For example: play, played, player, plays, playing will

all be considered the same feature as play.

Furthermore, it's possible to correct misspelling words by replacing them for the correct ones. This technique requires a dictionary with all the words and verify if every dataset word is correct or not. It has a big impact considering that textual crime information may have many misspelling words as people may be at the stress level of a robbery. In the Figure 2 it's exemplify mistakes of misspelling as they are considered different features in the dataset, even though the stemming technique corrects some mistakes some of them are only corrected by the usage of a dictionary. More problems found at textual data are: depending on the region of the person the words usage changes, it's very dynamic, have missing fragments that correspond to situation explanation, and is not structured.

Figure 2. Misspelling words example

Inspired by the removal of words that doesn't contribute to the characterization of the record, it's possible to aim at the identification of important words. There are different techniques that tries to identify which words are necessary to this task, for instance to apply filter utilizing the word frequency and a threshold, words that have frequency lower than the threshold are not considered as a feature. This technique assumes that important words would have a minimum frequency, however words common to a language are also considered. In the crime scenario it's possible to cite the words "victim" and "criminal", as to analyze crime data is expected to find a victim and a criminal at each record. Furthermore, it's possible to apply the threshold on the opposite direction, removing words with more frequency than a threshold, so words too common wouldn't be considered by the algorithm.

Finally, it's possible to predict missing fragments of data using machine learning algorithms of regression. However, analyzing the scenario it's possible to assert that if a crime record doesn't have a fire weapon and the prediction algorithm understands that it has it may chance the context of the crime. The existence of a fire weapon refers to a criminal with more sophistication comparing it with a criminal with only a knife, or a criminal without any weapon. In addition, characteristics of criminals may be predicted and filled into the crime record leading to wrong criminal description.

The aim of this work is to compare three different aspects of data quality and the capability of data mining techniques to work with high dimensional data. The first one, it's compared data with different quality to find out the impact at data mining techniques considering accuracy, recall and precision. The data quality is improved by the usage

of preprocessing techniques of stemming and removal of stop words, which is expected the decreasing of feature count. Will also be compared the impact of the usage of PCA technique. Furthermore, will be compared the machine learning algorithms against each other using the BOW transformation technique.

4. Crime Classification

The crime records used in this article come from the state of Rondônia and comprehend 1000 crimes of steal and robbery. According to the characteristics of textual data presented at section 3, it's possible to cite that there is a possibility that some records are duplicated as many people may call the police to describe the crime event. So, crime event characterizes the crime as it's the description of the crime by a victim or a individual present at the scenario. Also, there are distinct ways to describe the same situation, considering that people may have access to less fragments of data about the event. The data refers to the crime event considering the data about quantity of suspects, criminal characteristics, object stolen, crime weapon, scape route, and more.

Crimes are classified according to the sophistication of criminals into 3 types: A, is a crime with low sophistication, with no fire weapons, 1 criminal involved, low value objects stolen (wallet, cellphone). B, is a crime with medium sophistication, with fire weapons, 1 or 2 criminals involved, medium value objects stolen (notebook, vehicle). C, is a crime with high level sophistication, with special weapons (explosives), with 3 or more criminals involved, high value objects stolen (ATM machine, artwork, jewelry). Therefore, it has 400 records of type A, 300 of type B, and 300 of type C, what counts 1000 crime records.

Comparing the proposed classification and the one already utilized by Brazilian police forces, it's possible to address the usage of a more complex characterization by the police forces. The result of a criminal action, the will of the criminal, and more are considered to classify the crime event. This implies that a crime analyst has to identify more scenarios to make a decision. The proposed classification simplifies the crime characterization facilitating the crime analyst to identify similar crime events and make a decision.

To evaluate classification machine learning algorithms there are some metrics that may be utilized, for instance, accuracy, recall, precision, true negative rate (TNR), precision, negative predictive value (NPV) [Buczak and Guven 2016, Hailong et al. 2014]. In this work it's utilized accuracy that is based on the results presented at a confusion matrix. Considering the notation TP (true positive), TN (true negative), FP (false positive), and FN (false negative) it's referred to the equation 1 of the metric. The parametrization of the algorithms considered executions varying each parameter to verify the best configuration for each algorithm. The number of executions comprehend the number of parameters for each algorithm considering 1000 records and BOW transformation technique.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Hence, two experiments were performed, the first one compares the usage and the absence of pre-processing techniques, where these techniques comprehend the usage of

stemming, and removal of stop words. In addition, the PCA technique is applied to data and compared with the usage of preprocessing to understand the need to represent the information into a small dimensional space. The second experiment compares the usage of different data mining techniques to discover which one behaves better considering accuracy, precision, and recall. This experiment presents the capability to work with high dimensional data, with such complexity as textual data. On both experiments the techniques utilized were: Naive Bayes (NB), Random Forest (RF), Decision Tree (DT), Support Vector Machines (SVM), Multilayer Perceptron (MLP).

4.1. Pre-processing Experiment

At first, an experiment was conducted to test the accuracy of algorithms considering the usage or absence of pre-processing techniques. The results may be observed at table 1, which shows that pre-processing improved the accuracy of algorithms varying from 10.2% to 5.28%. At the feature count the variation found starts at 490.25 and goes to 1991.75. It shows that as the dataset size increases also the quantity of features used grows together as more words may appear at the training data. Was only considered the features of the training data as BOW creates the features vector based on it. The feature count has distinct effects considering each type of strategy utilized to understand the data distribution. Consequently, algorithms accuracy may decreased after a certain input size.

Table 1. Absence and Usage of Pre-processing Techniques Comparison

	Accuracy Mean (%)		Feature Count Mean (%)	
Techniques	100 (A)	1000 (B)	100 (A)	1000 (B)
No Pre-processing	61.8	70.5	1945.75	6741
Pre-processing	72	75.78	1455.5	4749.25

To verify the execution considering each algorithm the Figure 3 presents the results of the absence of pre-processing techniques. The algorithm MLP got the worst accuracy results, even though it presented the biggest growth. The other algorithms presented similar results at input size B, however at input size A Decision Tree and Naive Bayes tied at first place. Observing the Figure 4 the results of each algorithm with pre-processing techniques are presented. At the input size A MLP got the same accuracy as Decision tree, but after the increase of input size it decreased the accuracy. SVM, NB, and RF also achieved a similar accuracy of input size A, while SVM got the first place at input size B.

It is necessary to verify the impact of dimensions comparing the pre-processing techniques and the PCA technique along with the pre-processing. Consequently, was executed another experiment disregarding NB as PCA generated negative values. The results are shown at table 2. The usage of only pre-processing achieve better results even though the quantity of dimensions decreased. To analyze the accuracy it is necessary to consider that PCA only maps the data into a lower dimension hyperplane. It's a generalization of features where the feature that has more incidence is placed at the beginning of the feature vector. It doesn't helps to identify which feature is important following a predefined rule, as the appearance of a fire weapon only once may change the classification of the algorithm. Consequently, for this scenario the PCA is not recommended, considering the complexity and dynamism of textual data.

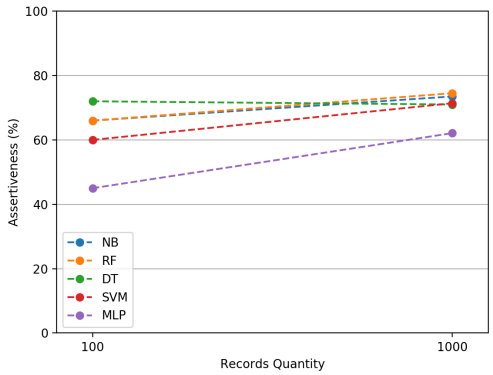

Figure 3. No pre-processing algorithms results

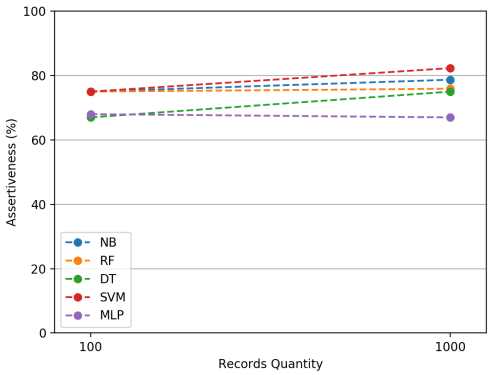

Figure 4. Pre-processing algorithms results

Table 2. Absence and Usage of PCA Technique Comparison

	Accuracy Mean (%)		Feature Count Mean (%)	
Techniques	100 (A)	1000 (B)	100 (A)	1000 (B)
Pre-process	64.25	76.62	1458	4749.75
Pre-process + PCA	48	62.12	100	500

To verify the accuracy of each algorithm the Figures 5, 6. At both Figures MLP got the worst results and maintained after the input size increased. The best algorithms were RF at the input size A and SVM at the input size B. That shows that RF doesn't need so much data to understands the scenario. The algorithm presented consistent results considering the growth of data compared with the adversaries.

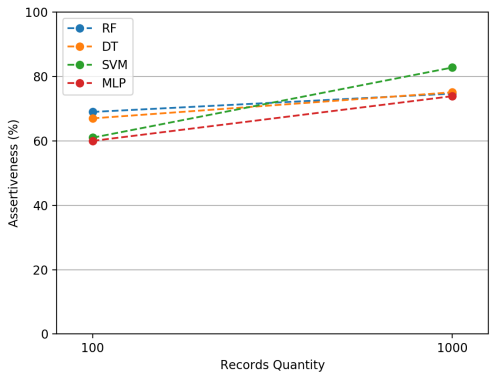

Figure 5. Pre-processing algorithms results

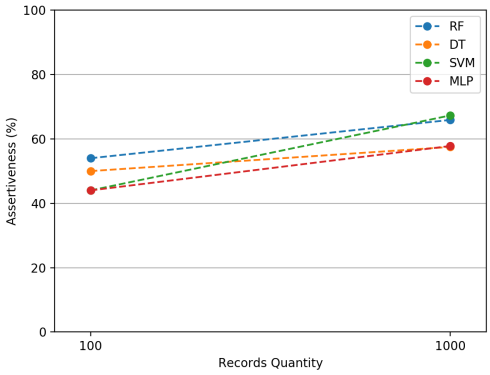

Figure 6. Preprocessing with PCA technique algorithms results

4.2. Data Mining Experiment

The second experiment executed shows the impact of dimensions by the visualization of confusion matrices. Only the usage of pre-processing techniques were considered, as it showed the best results at section 4.1. In addition, were executed only a single kfold step with 1000 input data, 750 to train and 250 to test. At Figures 7, 8, 9, 10 the results

may be observed as the accuracy of each algorithm were: RF (80.8%), DT (75.2%), SVM (79.2%), MLP (61.2%), NB (76.8%).

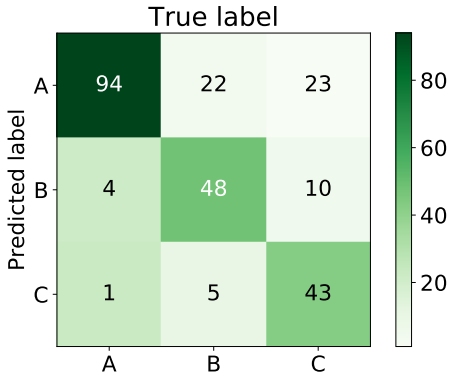

Figure 7. Random Forest algorithm's confusion matrix

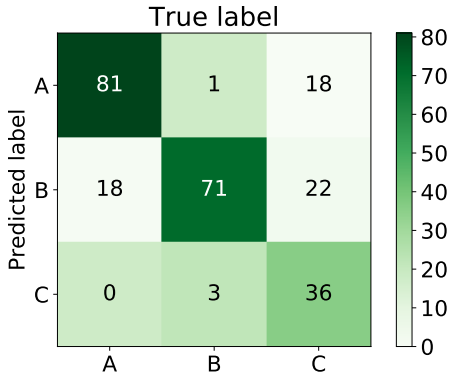

Figure 8. Decision Trees algorithm's confusion matrix

At Figure 7 it's possible to observe the misclassifications of the Random Forest algorithm, where the highest was found at C type with 33 records. The second were type B with 27 records. The highest accuracy were found at type A misclassifying only 5 records. At Figure 8 the results of Decision Trees algorithm are presented, comprehending the misclassification of 40 records at type C. The second were type A, with 18 records, and only 4 misclassified at type B.

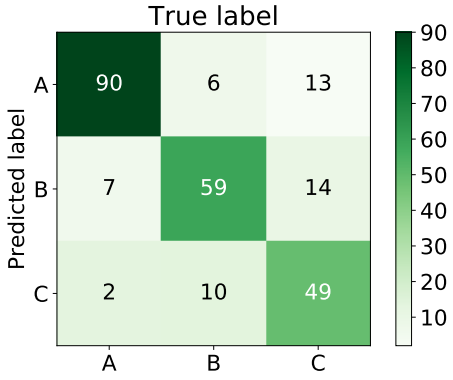

Figure 9. Support Vector Machines algorithm's confusion matrix

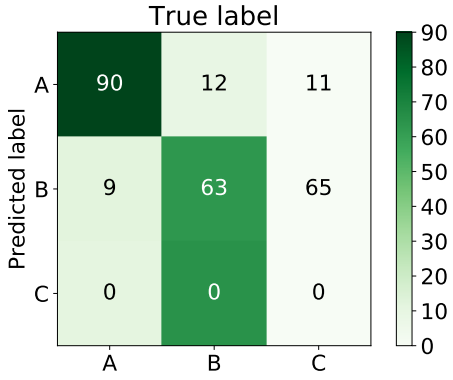

Figure 10. Multilayer Perceptron algorithm's confusion matrix

At Figure 9 the results of Support Vector Machines may be observed as the classification type C got 27 misclassified records, with 16 at type B, and only 9 at type A. At Figure 10 it's shown that the Multilayer Perceptron algorithm couldn't classify type C records correctly with 76 records. In addition, the algorithm didn't classify any record as type C, implying that it couldn't understand how was the data distribution, necessitating a lower dimension input data. At Figure 11 the Naive Bayes algorithm results are shown. It missed 24 records at type C, 20 at type B, and 14 at type A.

Analyzing the results obtained, it's easy to asses that the type C is the hardest class to classify correctly. Depending on the algorithm, type C records are classified mostly as

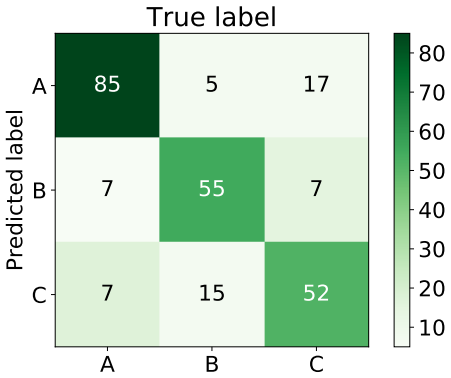

Figure 11. Naive Bayes algorithm's confusion matrix

type A or B, even though the classifier has been trained with a maximum of 300 records of type C. The interesting result is shown by the MLP algorithm where it could not classify correctly any type C. A record is classified as type C if there is a high value product, more than two criminals, or there is fire weapons or explosives, what requires the incidence of the words that describe the situation in the training data. The top 2 best algorithms were Random Forest and SVM considering the accuracy and capability to understand the data distribution at all the crime types.

The experiments showed that the improvement of data quality has a direct impact into the accuracy results of the algorithms. While this results imply that the information is reliable to be used by a crime analyst that needs to analyze the information presented and make a decision according to it's knowledge of the scenario, also visualizations may be developed based on the typification to present an overview of the global scenario. Preprocessing provides better data to be processed by data mining techniques, further it may be transfered to a database to be utilized by different systems.

5. Conclusion

In this work was presented the rehearsals were performed to discover the impact of preprocessing techniques at machine learning algorithms accuracy considering the number of features utilized, and consequently the data quality. In addition, the experiments presented which classification algorithm works better with textual information, requiring the understanding of the distribution of data into a high dimensional space. The usage of preprocessing techniques diminish greatly the mean quantity of dimensions and presented a better accuracy result. The best overall data mining technique found was the machine learning algorithm Random Forest, which achieved the accuracy result of 80.8% at the second experiment.

To improve the results, the identification of important words should be verified as even by removing stop words there are words that doesn't contribute to the scenario. For instance, the word "hat" doesn't help to characterize the scenario and classify the record into a specific class. What may be applied is the utilization of a feature selection method ignoring words with the frequency lower than a threshold. However, it has to be utilized carefully, therefore it may decrease the accuracy of type C records because it's not usual to find records with explosives on it.

Another alternative is to utilize a dictionary of words, that lists words important for the scenario, where the existence may change the type of the record. Even though, the dictionary would have many words, it probably wouldn't have them all requiring a guidance of a scenario specialist to identify them. Furthermore, robbery changes according to regions, as distinct objects are specific of regions, requiring a iterative task of updating the dictionary and adding new words.

Considering the scenario of smart cities, the amount of information grows according to the interconnectivity of systems, requiring an automation to classify incoming crimes and provide it to a visualization systems. This ecosystem helps crime analysts to make decisions to combat crimes to decrease the crime incidence. The typification based on the criminal sophistication improves the capability of an user to manage personal, as crimes with the same typification require similar ways to contain. The capability of the data mining techniques to classify into types facilitate the automation requiring crime records with objective information about crime weapon, objects stolen, quantity of criminals, and more, as input data.

References

Aghababaei, S. and Makrehchi, M. (2016). Mining social media content for crime prediction. In *2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 526–531.

Al-Anzi, F. S. and AbuZeina, D. (2015). Stemming impact on arabic text categorization performance: A survey. In *2015 5th International Conference on Information Communication Technology and Accessibility (ICTA)*, pages 1–7.

Aljrees, T., Shi, D., Windridge, D., and Wong, W. (2016). Criminal pattern identification based on modified k-means clustering. In *2016 International Conference on Machine Learning and Cybernetics (ICMLC)*, volume 2, pages 799–806.

Azeez, J. and Aravindhar, D. J. (2015). Hybrid approach to crime prediction using deep learning. In *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 1701–1710.

Babakura, A., Sulaiman, M. N., and Yusuf, M. A. (2014). Improved method of classification algorithms for crime prediction. In *2014 International Symposium on Biometrics and Security Technologies (ISBAST)*, pages 250–255.

Ballesteros, J., Rahman, M., Carbunar, B., and Rishe, N. (2012). Safe cities. a participatory sensing approach. In *37th Annual IEEE Conference on Local Computer Networks*, pages 626–634.

Bogomolov, A., Lepri, B., Staiano, J., Oliver, N., Pianesi, F., and Pentland, A. (2014). Once upon a crime: Towards crime prediction from demographics and mobile data. In *Proceedings of the 16th International Conference on Multimodal Interaction*, ICMI '14, pages 427–434, New York, NY, USA. ACM.

Buczak, A. L. and Guven, E. (2016). A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications Surveys Tutorials*, 18(2):1153–1176.

George, C. P., Wang, D. Z., Wilson, J. N., Epstein, L. M., Garland, P., and Suh, A. (2012). A machine learning based topic exploration and categorization on surveys. In *2012 11th International Conference on Machine Learning and Applications*, volume 2, pages 7–12.

Hailong, Z., Wenyan, G., and Bo, J. (2014). Machine learning and lexicon based methods for sentiment classification: A survey. In *2014 11th Web Information System and Application Conference*, pages 262–265.

Jung, Y. and Yoon, Y. (2015). Behavior tracking model in dynamic situation using the risk ratio em. In *2015 International Conference on Information Networking (ICOIN)*, pages 444–448.

Kumar, A. S. and Gopal, R. K. (2015). Data mining based crime investigation systems: Taxonomy and relevance. In *2015 Global Conference on Communication Technologies (GCCT)*, pages 850–853.

Noyes, D. (2018). The top 20 valuable facebook statistics.

Piatetsky-Shapiro, G. (1991). Knowledge discovery in real databases: A report on the ijcai-89 workshop. *AI Mag.*, 11(5):68–70.

Schünke, L. C., de Oliveira, L. P. L., and Villamil, M. B. (2014). Visualization and analysis of interacting occurences in a smart city. In *2014 IEEE Symposium on Computers and Communications (ISCC)*, pages 1–7.

Silva, L. A. (2016). *Introdução à Mineração de Dados*. Elsevier.

Sivaranjani, S., Sivakumari, S., and Aasha, M. (2016). Crime prediction and forecasting in tamilnadu using clustering approaches. In *2016 International Conference on Emerging Technological Trends (ICETT)*, pages 1–6.

Tayebi, M. A., Glasser, U., and Brantingham, P. L. (2015). Learning where to inspect: Location learning for crime prediction. In *2015 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 25–30.

Vyas, T., Prajapati, P., and Gadhwal, S. (2015). A survey and evaluation of supervised machine learning techniques for spam e-mail filtering. In *2015 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, pages 1–7.