ISSN 2278-3091
Volume 9, No.1, January – February 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse97912020.pdf
https://doi.org/10.30534/ijatcse/2020/97912020

Sentiments Analysis On Public Land Transport Infrastructure in Davao Region using Machine Learning Algorithms

Mark Van M. Buladaco[1], Jumar S. Buladaco[2], Laarni M. Cantero[3]
[1] Davao del Norte State College, Philippines, markvan.buladaco@dnsc.edu.ph
[2] Davao del Norte State College, Philippines, jumaw@dnsc.edu.ph
[3] Davao del Norte State College, Philippines, laarni.cantero@dnsc.edu.ph

ABSTRACT

Land transport infrastructure has been a vital part of a city. People nowadays use social media to post their sentiments towards developments in a city such as land transport. Government agencies have difficulty in identifying issues that arises from the people using social media towards land transport infrastructure. These social media posts in the form of a text can be analyzed using sentiments analysis, which is a significant task of Natural Language Processing (NLP). This research experimented on creating a model of sentiments on land transport infrastructure in Region XI (Davao Region), the Philippines from the social media website, and test a data set on the accuracy of the model. There are a total of 1,200 text data sets, and it's divided into two: test dataset is 25%, and the training dataset is 75%. The machine learning text classifiers used are Support Vector Machines (SVM), Random Forest (RF) and Multinomial Naïve Bayesian (MNB) to process sentiments analysis of the text data sets. The performance of each classification model is estimated by generating confusion metric with the calculation of precision and recall, the f1-score. The accuracy rating was also computed. A comparison was also conducted based on the results of experiments of the three machine learning classifiers. Based on the resulting experiment, SVM has the highest accuracy, with 76.12% and a f1-score of 71.98%. This research will be utilized as support and notes for policy-making and development for land transport infrastructure in the Davao Region.

Key words: Data Mining, Land Transport Infrastructure, Machine Learning, Sentiments Analysis, Philippines.

1. INTRODUCTION

Land transport infrastructure has been a vital part of a city. Development of land transport infrastructure, through its immediate and roundabout impacts, has a course on the manageability of development and in general advancement of a nation. Aside from improving the network, the improvement of streets can open up to this point detached districts to exchange and speculate and venture up access to products, administrations, and work openings. This is fortified by the way that spending on the framework has enormous multiplier impacts. Over the long haul, this is accomplished by getting higher returns from private ventures, however just if this impact is more prominent than the negative effect of the increased tax rates expected to pay for it [1].

These infrastructures have a positive and negative impression on the citizens of a country. Though it is for the general betterment of a nation, it also produces problems. Traffic congestions can have a great effect on peoples' behavior. Social media have become an emotional outlet of people, especially when stuck in traffic. Traffic congestion inevitably causes delays that can poorly impact one's emotional state, as well as physical aspects of their lives, such as business meetings or enjoyment of being with friends and family. Stress and frustration are both common emotions caused by delays on the road. According to a study, 40% of commuters and drivers in the road reported negative emotions of stressed or frustrated to describe their feelings about travel in their city [2]. Increased use of public transport can be a key solution for congestion problems, which, if successful, would make private vehicle users will switch to public transport and will reduce the amount of volume of vehicles on the road, which would reduce congestion. There are researches and technologies that address traffic flow and congestion in the road. One of which is the method developed in controlling the traffic for increasing the likelihood and efficiency of cryptographic parties [3]. It is implemented as support in the Intelligent Transport System (ITS), which basically focuses on resolving different issued of road traffic. Another research investigates and proposed big data processing design for data flow specific to Smart Cities and implemented on ITS [4]. It led to convincing results in terms of performance and speed using different data mining technologies.

The Philippines is a country that has heavy traffic on major public roads and highways. Metropolitan cities in the country always have traffic congestions during peak hours. The Philippines is also a country that is much attuned to social

media and mobile technology. The country is nicknamed the "social media capital of the world" [5]. In the Asia-Pacific region, it is actually the country with the highest social networking penetration [6] and the eight in global usage [5]. Even government agencies in the Philippines have Twitter pages which use it to post updates such as the Metro Manila Development Authority, which post traffic conditions of the Greater Manila Area. The country also started a multi-billion project, namely the "Build, Build, Build." President Duterte initiated the "Build, Build, Build" (BBB) Program, which seeks to accelerate infrastructure spending and develop industries that will yield robust growth, create jobs and improve the lives of Filipinos [7]. In the Philippines, there are 75 flagship projects consist of airports, railways, bus rapid transits, roads and bridges, and seaports that seek to basically mark down the costs of production, improve income in the rural areas, encourage investments in the province, make the movement of goods and people more efficient, and most importantly, create a projected 1.7 million jobs by 2022 [8].

The use of public transport and Twitter have close ties with each other. People usually post tweets of their opinions on the use of these services and during traffic congestion, especially Filipinos. Twitter also has become a major source of news and updates in the country. In fact, studies have been done to analyze whether Twitter is a social network or a news media outlet, and some initial results show that some characteristics of Twitter deviate from other social media and that the majority of the user-generated content on Twitter is news-related [9]. In a study conducted by Effendy et al., the ease of access in the delivery of opinion could be an opportunity to be used as an assessment and evaluation of city public transport services [10]. Twitter could be a major factor in developing policies in the future. To generate information from the existing opinion data, the data is processed with sentiment analysis that will separate opinion in a positive or negative sentiment class, and infer what factors are often discussed in those opinions. Sentiment analysis is a process of the deriving sentiment of a particular statement or sentence [11]. It's a classification technique that derives opinion from the tweets and formulates a sentiment and on the basis of which, sentiment classification is performed.

There are two ways to conduct sentiments analysis in social media platforms: lexicon-based approach and machine learning. Some researchers applied a lexical approach to identify emotions in text. For example, the researchers constructed a large lexicon annotated for six basic emotions: anger, disgust, fear, joy, sadness, and surprise [12]. Machine learning algorithms are one of the methods of determining a summary of sentiments from a social media platform. In machine learning, supervised learning algorithms are utilized. There are also unsupervised learning and hybrid learning approach that is used. A training data set will be produced first as the basis of the classification of the text to be either positive, negative, or neutral.

The objective of this study will determine the polarity classification of Filipino sentiments in Twitter using three machine learning algorithms on classifying: Support Vector Machines, Random Forest, and Naïve Bayesian Classifier. Further, this study will also determine the best classifier algorithm for the sentiments on land transport infrastructure by investigating the accuracy of each classifier. The result of this study will also be given to appropriate agencies as notes to policymaking in relation to land transportation. This study will also identify issues related to the use of public transport in a city. This research can be used as a recommendation to government agencies and offices in the improvement of policies for increasing the performance and efficiency of a city public transport infrastructure.

2. SENTIMENTS ANALYSIS METHODOLOGY

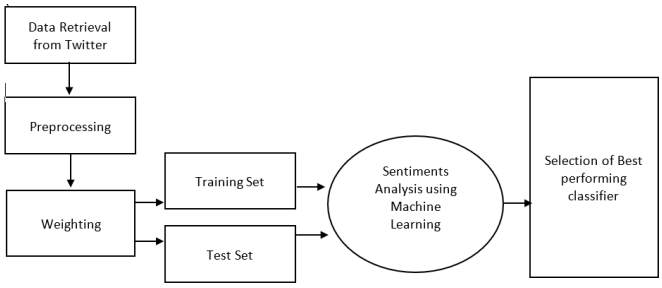

Figure 1: Conceptual Framework of the Sentiments Analysis Method

The method of classifying polarity in text mining is based on the sentiments analysis methodology implemented by Gupta [11]. The data set will be retrieved first from a social media network such as Twitter, and it will undergo pre-processing. After the dataset is pre-processed, it will undergo through weighting where a number of occurrences of a document will be counted and be weighted. After that, a training set and the test set will be developed and extracted from the collected data and divided using the percentile approach. The test set will undergo the classification of polarity using Support Vector Machines, Random Forest, and Naïve Bayesian Classifier. After the experiment, the accuracy and other measurements of each classifier will be measured to determine the best classifier in the given dataset on transportation.

2.1 Dataset Source

Twitter is the source of the dataset of this research. In order to study Filipino sentiments on public transport and traffic congestion, the researchers will extract text data from twitter. In social media Twitter, users tend to use non-formal language, without using proper grammar and appeared many slang words. A developer account was created for a Twitter API, and an application was developed in order to execute scripts for getting datasets. Tweets will be automatically collected using the script. Filipino and Bisaya dialect tweets were also considered, and it will then be later be translated to English using the Google Translate API.

A corpus will be developed once the data from Twitter will be collected. A text corpus is a large and structured set of texts [13]. They are used to do statistical analysis and hypothesis testing, checking occurrences, or validating linguistic rules within a specific language territory. The research focuses on the land transport infrastructure in four large cities in the Davao region, namely: Davao City, Digos City, Panabo City, and Tagum City. These are the cities that experience heavy traffic during peak hours and are the chartered cities in the Davao region. Davao City is the capital of the region and is considered as one of the fast-growing and developing chartered metropolitan cities in the Philippines in terms of population, the geographic and economic condition in the country and in Southeast Asia [14]. The dataset must consist of tweets that contain the text presented in Table 1.

Table 1: Datasets filters

City	Transport Infrastructure	Text Inclusion Criteria
Davao	Roads	(“Davao”) AND (“Dalan” OR “Road” OR “Kalsada”)
Davao	Flyover	“Davao” AND “Flyover”
Davao	Bridge	“Davao” AND (“Tulay” OR “Bridge”)
Tagum	Roads	“Tagum” AND (“Dalan” OR “Road” OR “Kalsada”)
Tagum	Bridge	“Tagum” AND (“Tulay” OR “Bridge”)
Digos	Roads	“Digos” AND (“Dalan” OR “Road” OR “Kalsada”)
Digos	Bridge	“Digos” AND (“Tulay” OR “Bridge”)
Panabo	Roads	“Digos” AND (“Dalan” OR “Road” OR “Kalsada”)
Panabo	Bridge	“Digos” AND (“Tulay” OR “Bridge”)

2.2 Data Cleansing

The process consists of case folding and removes noise. Noise, in this case, is a character other than letters (numbers, symbols, and punctuation). If the text data has non-English words, the sentence will be translated first to English using Google translate API. The data are divided into two parts: 75% is the training data, and 25 % is the test data. The training data set will undergo preprocessing.

2.3 Preprocessing

Data from Twitter is unstructured, which is information from people posted his or her feelings, opinions, attitudes, behavior, emotions, etc. [15]. Tweets will then undergo preprocessing to prepare for the classification using the different classifiers and preprocessing of the tweets is shown in Table 2.

Table 2: Preprocessing Techniques

Technique	Description
Translation	Text data that are posted in Filipino or Bisaya are translated to English using Google Translator.
Tokenization	The process of cutting a row of words in the document into single word piece or unigrams.
Part of Speech (POS) Tagging	POS tagging is the process of tagging the word. In this research, the process of tagging using Hidden Markov Models (HMM) word and Rule-Based POS Tagging.
Stop word Removal	The stop word removal process is removing words that often appear but do not have a specific meaning and is not considered necessary in the opinion classification.

2.3 Weighting

The weighting process is performed based on the number of occurrences of words in a document so that the material can be represented in a vector. Feature weighting used is a unigram, and Term Frequency-Inverse Document Frequency (TF-IDF) weighting method. A document terms matrix function is performed using the TF-IDF algorithm. It is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. The Frequency representation is symbolized as tf, and the output will be a frequency of terms in the dataset where the columns in the matrix represent the terms in the dataset and the row represent the documents. This tf is not directly used to find similarities between texts, but instead, it is used in calculating the weights of every term. The formula (1) is:

$$wtf(t) = tf * idf(t) \text{ for term } t, \quad idf = \log 2(N/n) \tag{1}$$

In the formula (1), where N is the number of all documents and n is the number of documents where that term appears. The wtf calculation has to be done for every document and for every term that appears in it. The return value corresponds to the degree of similarity between the training document and the new document as a parameter.

2.3 Polarity Classification

After datasets underwent preprocessing and weighting process, it will now experience a process of classification of its polarity by utilizing the classifier models of machine learning. The classification models selected for categorization are Naïve Bayesian, Random Forest, and Support Vector Machines.

Support Vector Machines (SVM)

SVM can predict class based on the model or pattern of the results of the training process. Classification is done by finding the hyperplane or boundary line (decision boundary) that separates between classes. SVM search hyperplane value by using a support vector and the cost of margin [16]. Separator function or hyperplane is a linear function in equation.

$$w \times x + b = 0 \quad (2)$$

In equation (2) equation, w is the weight that presents the position of a hyperplane in the normal field, x is the vector of input data, and b is the bias that represents the position of the field relative to the origin. To optimize the hyperplane, the problem essentially transforms to the minimization of $\|W\|$, which is eventually computed as $\sum_{i=0}^{n} a_i y_i x_i$ where ai are numeric parameters, and yi are labels based on support vectors xi. That is if yi = 1, then $\sum_{i=0}^{n} w_i x_i \geq -1$ [16]. It was concluded that the use of SVM provides superior results compared to other methods that the level of accuracy is up to 82.2% [17].

Moreover, the SVM method itself can be used to classify the data based on attribute valuation opinion held to be separated opinions belong to a class of positive or negative [10].

Random Forest (RF)

Random forests are a group learning strategy for order, relapse, and different assignments that works by developing a large number of choice trees at preparing time and yielding the class that is the method of the categories (arrangement) or mean forecast (relapse) of the individual trees [19]. Random forests right for choice trees' propensity for overfitting to their preparation set. These random trees are combined to form the aggregated regression estimate. The formula for the random forest classifier is (3).

$$RFfi_i = \frac{\sum_j normfi_{ij}}{\sum_{j \in all\ features, k \in all\ trees} normfi_{jk}} \quad (3)$$

Naïve Bayesian Algorithm (NB)

The Naive Bayes Classifier technique is based on the so-called Bayesian theorem and is particularly suited when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods. Specifically, in this research, the researcher utilized the Multibinomial Naïve Bayesian classifier.

The NB classifier works as follows: Suppose that there exists a set of training data, D, in which each tuple is represented by an n-dimensional feature vector, X=x1,x2,..,xn, indicating n measurements made on the tuple from n attributes or features [20]. Assume that there are m classes, C1, C2, ..., Cm. Given a tuple X, the classifier will predict that X belongs to Ci if and only if: P (Ci | X) >P (Cj | X), where i, j ∈ [1, m] and i≠ j. P (Ci | X) is computed as (4):

$$P(C|X) = \prod_{k=1}^{n} P(x_k|C_i) \quad (4)$$

3. EXPERIMENT RESULTS

After preprocessing the data, weighting, and feature extraction, training the model was conducted using classifier machine learning algorithms: Support Vector Machines (SVM), Random Forest (RF), and Naïve Bayesian Classifier (NB). The training data set was used for training the models using the three classifiers. The dataset will be divided into two: 75% became training data, and 25% of it is test data. Labeling of the training data is done by undergoing it through a lexicon-based algorithm called SentiWordnet as to whether the tweet is positive, negative, or neutral polarity. The machine learning approach is more efficient and accurate to detect polarity classification when compared and subject to a lexicon-based approach [21]. Scikit-learn library of Python was utilized for the experiment and training the model using the three algorithms.

The performance of each classification model is estimated by generating confusion metric with the calculation of precision and recall. Using precision and recall value, F1-score is calculated (5), and results are compared. The accuracy rating was also computed.

$$F_1 = 2 * \frac{precision * recall}{precision + recall} \quad (5)$$

Results was obtained by applying different machine learning classifying methods using the scikit-learn library in Python. It is presented in Table 3.

Table 3. Results of the experiment

Machine Learning Algorithm (Classifier)	Accuracy	F1-Score
SVM	76.12%	71.98%
Random Forest	68.02%	66.99%
Naïve Bayesian Algorithm	72.22%	70.80%

The following data in table 2 shows the performance of the three algorithms for evaluation. Out of which, SVM turns out to be best among all. This result corresponds with the findings of Pang and Lee [17] that SVM is superior among other classifiers.

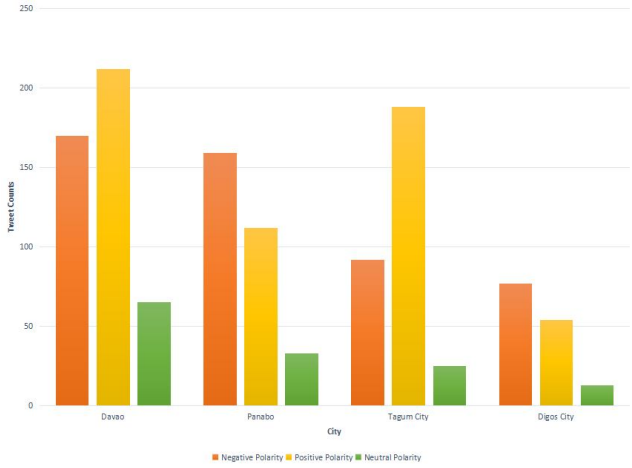

Figure 2: Tweet Counts and its polarity by city

In figure 2 shows the tweet counts of polarity by city, and it shows that Davao city has a higher positive polarity, and Panabo City has a higher negative polarity. This may be related to the construction of the flyover infrastructure as it leads to traffic congestion in the city. There is also the reconstruction of the road, which significantly affects the negative emotions of the citizens in the town. In Davao city, it is evident with the different construction of new land transport infrastructure that will significantly benefit the citizen of the town. These include development and construcion of the different bypass roads, renovation of bridges, and development of the coastal roads with the four-lanes facility and has four segments.

4. CONCLUSION

Based on the results of this research, it can be concluded that sentiment analysis on Twitter data about the use of city public transport can be done machine learning algorithms. A comparison was also conducted based on the results of experiments of the three machine learning classifiers, namely, Support Vector Machines (SVM), Random Forest (RF) and Naïve Bayesian Algorithm (NB). The experiment was conducted and evaluated using k-fold validation up to k=10. Based on the resulting experiment, SVM has the highest accuracy, with 76.12% and a f1-score of 71.98%.

This research will be utilized as support and notes for policy-making and development for land transport infrastructure in the Davao Region. The positives opinions can be utilized to be able to improve or can be applied to other types of transportation, as well as for negative opinions could be used as advice for stakeholders to be able to improve public transport services, especially city public transport. Issues and reasons for the traffic congestion in a city will be part of the output.

REFERENCES

1. Shruti Tripathi and Vikash Gautam. **Road Transport Infrastructure and Economic Growth in India**, *Journal of Infrastructure Development*, vol 2, pp 135-151. https://doi.org/10.1177/097493061100200204
2. G. Cantarella. **The Impact of Traffic Environmental Vision Pressure on Driver Behaviour**, *Journal of Advanced Transportation*, vol. 2018, 2018. https://doi.org/10.1155/2018/4941605
3. A. Ghasempour, Z. Mohd.Hanapi, M. Salehi, and Z. Vahdati. **Using Traffic Control Scheme In Intelligent Transportation System**, *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 8, no.1.4, 2019. https://doi.org/10.30534/ijatcse/2019/2581.42019
4. S. El Mendili, Y. El Bouzekri El Idrissi, N. Hmina.**Big Data Processing Platform on Intelligent Transportation Systems**, *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 8, no. 4, 2019. https://doi.org/10.30534/ijatcse/2019/16842019
5. J. Boaz, M. Ybañez, M. De Leon, and M.R. Estuar. **Understanding the Behavior of Filipino Twitter Users during Disaster**, *GSTF International Journal on Computing (JoC)*, vol. 3, no. 2 July 2013. https://doi.org/10.7603/s40601-013-0007-z
6. Maureen Atienza. **Social Media Marketing(SMM): Measuring Its Effects To Resorts In Batangas Province, Philippines**, *IOER International Multidisciplinary Research Journal*, vol. 1, no. 1, Mar. 2019
7. J.P. Cruz, C.P. Delgra, and J.M. Enriquez. **Governing the 'Golden Age of Infrastructure': Build, Build, Build Through an Accountability Perspective**, *SSRN Electronic Journal*, Jan 2018 https://doi.org/10.2139/ssrn.3300597
8. E. Francia, **Build, Build, Build to highlight infrastructure, transportation, and connectivity.** PhilStar Dec. 2018, available at https://www.philstar.com/business/2018/05/27/1818831/build-build-build-highlight-infrastructure-transportation-and-connectivity
9. H. Kwak, C. Lee, H. Park, and S. Moon. **What Is Twitter, a Social Network or a News Media?**, *Proceedings of the 19th International Conference on World Wide Web*, 2010. https://doi.org/10.1145/1772690.1772751.
10. N. Effendy, Sabariah. **Sentiment Analysis on Twitter about the Use of City Public Transportation Using Support Vector Machine Method**, International Journal on ICT, vol. 2 no. 1 pp. 57-66, 2016. https://doi.org/10.21108/IJOICT.2016.2.1.85
11. B. Gupta. **Study of Twitter Sentiment Analysis using Machine Learning Algorithms on Python**, *International Journal of Computer Applications*, vol. 165, no. 9, 2017

12. C. Strapparava, and R. Mihalcea. **Learning to identify emotions in text**. *Proceedings of the 2008 ACM symposium on Applied computing*, pp.1556-1560, 2008
13. K. Wołk and K. Marasek. **A Sentence Meaning Based Alignment Method for Parallel Text Corpora Preparation**. *Advances in Intelligent Systems and Computing*, vol. 275, pp 107–114 Springer, 2015 ISBN 978-3-319-05950-1. ISSN 2194-5357
14. M.V. Buladaco. **Alternative Mass Transport System for Davao City: A Geographic Information System Approach**, *International Journal of Multidisciplinary and Current Research*, vol. 6, Aug 2018, https://doi.org/10.14741/ijmcr/v.6.5.21
15. S. Wankhede et al. **Data Preprocessing for Efficient Sentimental Analysis**, *Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, 2018, 723-726. https://doi.org/10.1109/ICICCT.2018.8473277.
16. J. Han, J. Pei and M. Kamber. **Data Mining Concept and Technique Third Edition**. Published by Elsevier Inc., Waltham, MA, 2006
17. B. Pang and L. Lee. **Opinion Mining and Sentiment Analysis**, *Foundations and Trends in Information Retrieval*, vol. 2, no. 1- 2, pp. 1-135, Jan. 2008.
18. A. Go, R. Bhayani and L. Huang. **Twitter Sentiment Classification using Distant Supervision**. Project Report, Stanford, 2005
19. T. Hastie, R. Tibshirani and J. Friedman. **The Elements of Statistical Learning (2nd ed.)**. Published by Springer, 2008. ISBN 0-387-95284-5.
20. Alexander Pak and Patrick Paroubek. **Twitter as a Corpus for Sentiment Analysis and Opinion Mining**. *Proceedings of Seventh International Conference on Language Resources and Evaluation*, LREC 2010, 17-23 May 2010, Valletta, Malta
21. M. Syamala and N.J. Nalini. **A Deep Analysis on Aspect based Sentiment Text Classification Approaches**, *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 8, no. 5, 2019. https://doi.org/10.30534/ijatcse/2019/01852019