APPLICATION OF SENTIMENT ANALYSIS ON SOCIAL MEDIA REVIEWS OF PUBLIC TRANSIT SYSTEMS

| **Lalithanjali Kesava Pillai**  *anjali.kp@rutgers.edu*  Bloustein School of Planning and Public Policy Rutgers University | **Sri Akshita Velpuru**  *akshita.velpuru@rutgers.edu*  Bloustein School of Planning and Public Policy Rutgers University | **Davyonn Jones**  *dayvonn.jones@rutgers.edu*  Bloustein School of Planning and Public Policy Rutgers University |
| --- | --- | --- |

**Abstract**

Sentiment analysis is a widely studied Natural Language Processing technique to systematically determine the opinions and emotions of users towards a product or a service. Today, a wide variety of social media platforms have become a source of information, and social media data has become a significant part of the world of sentiment analysis. Massive amounts of user-generated data from various social media platforms have been used to provide insights to businesses and governments. This paper attempts to determine user sentiment about a public transit system by extracting data in Spanish from social media sources such as Twitter and Tripadvisor. In this paper, we perform sentiment analysis on Spanish text by translating Spanish text to English using translation tools and modeling multiple sentiment analyzers to explore their efficacy in retaining the sentiment of the original Spanish text. This paper gives insights into the results of the sentiment analysis compared to a Gold Standard and the implementation process and also highlights several challenges and gaps in the analysis methods. The paper contributes to the research on sentiment analysis and can help practitioners select a suitable methodology for their applications.

**Keywords:** *Natural Language Processing, Public Transit, Renfe, Sentiment Analysis, Textual Analytics*

**1 Introduction**

Social media has been a helpful tool for extracting insight with commentary and various interactions with other forms of commentary. These interactions are responses that have different names throughout different platforms. For example, in platforms, Tripadvisor and Twitter interactions include and are not limited to likes, mentions, and retweets. Although social media is used mostly to complain, the two platforms provided excellent data to identify commentary on public transportation. The public transportation system The users of public transportation help provide important feedback and information that help improve quality standards for the transportation systems. When a user provides feedback on public transportation, it is usually to commend or condemn in relation to their experience.

This study aims to help language users know if the translated sentiment from artificial intelligence can properly imitate that of a native language speaker. Examining the sentiment of a comment or question is difficult because there are various aspects within the sentence. Machine learning models can be binary and unable to recognize sentence constructs such as sarcasm because the programming is still fairly new.

**2 Literature Review**

**2.1 Textual Analytics**

Any language has infinite dimensions, and expressing thought through a language depends on the thinking behind it. This makes it essential to study various aspects of the dynamic thinking process to understand the patterns better. One of the common facets of any language is the bias associated with it based on beliefs subjective to several stereotypes. Malsburg et al., 2020 discuss gender bias in election event prediction through linguistic choices in the US and the UK. For the years 2016 and 2017, the study found that usage of 'she' pronouns had a severe bias in the expectation of the election win. Studies like this emphasize the importance of linguistic preferences in real-world events and that language plays a vital part in the psychological thinking of humans. Artificial Intelligence paves the way for detailed research in analyzing a language using a process called Natural Language Processing (NLP). Sentiment analysis is a widely researched NLP task whose goal is to determine users' opinions, emotions, and evaluations of a product, entity, or service they are reviewing. Sentiment analysis is the automated process of analyzing text to determine if the sentiment expressed is positive, negative, or neutral. It is one of the most popular ways to analyze text, such as survey responses, customer support issues, online reviews, and live chats, to help companies stay on top of customer satisfaction. With the recent advances in deep learning, the ability of algorithms to analyze text has improved considerably. These advancements have led to the application of sentiment analysis in various domains, and this paper explores the usage of sentiment analysis and various techniques used and also identifies the challenges and issues with it.

**2.2 Sentiment Analysis**

(Moreno & Redondo, n.d., #) explores the applications of textual analytics and the emergence of computational linguistics. A typical text analytics application consists of information extraction through named entity recognition; topic tracking through keyword analytics; text summarization with deep and shallow analysis; clustering with defined topics; categorization and classification by theming each document source; concept linking by combining attributes in each document; information visualization by mapping to produce a capable browsing source; question answering with predictive answers to fit the themes and typologies and deep learning techniques involving recurrent neural networks, and convolutional neural networks. Text analytics is constantly evolving as information expressed in the text contains data that can be used with cognitive applications to filter information and provide pathways and clarity to important and relevant information.

Once such research was a computerized textual analytics process done by categorizing words into psychologically meaningful sections, called Linguistic Inquiry and Word Count (LIWC). Comparing words from a specific dataset with dictionary words is the base of the LIWC technique. Each word is compared to the dictionary and categorized into its specific form, such as pronouns, etc. Similarly, the data's emotional states, intentions, and motivations can be analyzed and summarized through LIWC, which establishes the relation between the language and the psychological process (Tausczik and Pennebaker., 2010 #).

By investigating the dominance of group communication with chat transcripts through manual and automatic coding, the efficiency of textual analytics can be studied (Samuel et al., 2014, #). The results from both types of coding were compared, and the relevance was determined to provide insights into costs and other resources associated with it. A custom-developed case was given to seven groups with six members in each, and a decision-making task with clues given to each group was observed. Through the communication within the groups, the dominant factor that led to the decision-making was studied. Texts from the group chats were coded with a custom parsing program and manual coding using Bales' interaction process. Expression of Dominance(ED) was calculated at the member level in all the groups, three logit regression models were formed to study the relationship between ED and chat comment features using the PROC LOGISTIC function in SAS. Refraining from classifying positive or negative connotations, the study only focused on the detection of dominance, and it can be automated. Results demonstrated that automatic coding is as efficient as the manual coding of texts.

(Gefen et al., 2017, #) developed a guide that introduced the concept of Latent Semantic Analysis in analyzing text specifically for behavioral scientists. The methods outlined by the authors include pre-processing textual data to build the Term Document Matrix followed by a Latent Semantic Analysis. LSA is a powerful tool, and applying it correctly requires both understanding what it does by understanding the tool’s strengths and weaknesses and how to run it. Drieger (2013) explored if semantic models can be used to support knowledge-building, analytical reasoning, and explorative analysis. The model was explored with network structure by statistical methods. Edges that represent the relationship between two models using statistical quantities of adjacent nodes. A path is a set of connected edges which describes the semantic relations, nodes, hubs, subgraphs, and clusters. The takeaway was those general methods for semantics review quantitative and qualitative perspectives from the analysis and the interpretation of network structures.

Samuel (2020) questioned How can we meaningfully and effectively exploit big data so as to extract insights in ways that are friendly to human sensory abilities and cognitive capabilities? The methodology was Visual analytics by analytical problems and general application areas. Principal Component Analysis to represent data in a condensed dimensional structure and Multi-Dimensional Scaling to map higher dimensional data to lower dimensional data. Modeling was shown by line and bar graphs, and word frequency visualizations. The overall idea was that there is a growing availability and notoriety of artificial intelligence that will help people perform better data visualization.

**2.2.1 Sentiment Analysis on Covid-related events**

The Coronavirus pandemic led to a rich source of data that can be used for sentiment analysis. (SAMUEL et al., 2020) identifies public sentiment in the context of the lockdown's problematic socioeconomic consequences and investigate four potential public sentiment-related scenarios. This study examines public sentiment using Twitter data, which is time-aligned to the COVID-19 reopening debate, to identify dominant sentiment trends associated with the push to reopen the economy, including statistical validation. The results show that tweets data from American Twitter users show more positive sentiment support than negative sentiment support for reopening the US economy. This study creates a novel sentiment polarity-based public sentiment scenarios (PSS) framework that can be useful for future crisis analysis long after COVID-19. This research stream, with additional validation, could present valuable time-sensitive opportunities for state governments, the federal government, corporations, and societal leaders to guide local, regional, and national communities into a successful new normal future.

(Nawaz Ali et al., 2020) A study was carried out to gain insights through the temporal dynamics of the sentiments associated with COVID-19 vaccination in the United States for facilitating necessary policy implementations. The framework of this research was to illustrate common sentiment scenarios and associate the polarity of the Public Sentiment Scores (PSS) with the vaccine administration. The findings are then compared with CDC data for state-wise vaccination rates mapping and with HPS data to study the sentiments. Technologies used were the rtweet and SentimentR package to download tweets with filters such as the "vaccine" keyword, cleaning and processing a million tweets, and a custom algorithm to replace abusive words and assign scores, in the range greater than 0.1 for positive statements, less than -0.1 for negative, and between this range for neutral statements. To further reinstate these sentiment scores, Python-based lexicon libraries, Textblob and VADER were used to identify the sentiments. With the scores obtained from all three methods close to each other, the next step is to identify the word frequency to generate patterns and themes of the text corpus.

(Nawaz Ali et al., 2020) It was found that in the United States, there was an increase in positive sentiments in the first month and a declining trend in the second. Negative sentiments seemed to increase over the period; however, neutral remained the same. State-level scores led to inferring that fear was the most associated sentiment in most states. Overall, trust sentiment seemed to have a moderate to low score and had seen a decrease in all states. Vaccination propensities and administration rates were observed. The shortcomings of the study were the overall representativeness of data since most disadvantaged people do not have access to Twitter.

To dive deep into the fear sentiment associated with the COVID pandemic broadly, another study was conducted (Samuel et al., 2020). Data from Twitter was used to perform a descriptive analysis of textual data in addition to the visualization and comparing the classification of sentiment analysis methods used. Generating explanatory textual data with the frequency of occurrence of the words using Word Cloud packages in R and Python was done to summarize the textual corpus. Data obtained from Twitter was cleaned for stop words, and abusive words and converted into tokens, and further analysis of textual elements such as stemming or lemmatization was carried out to simplify the text. The N-grams technique was used to identify the frequency of words and also word pairs and chains. Hashtags from different types of devices were also analyzed as endogenous analytics research. Machine learning methods such as Naive Bayes and logistic regression were used to classify the tweets statistically. Mapping sentiments against time is critical during crises like the Coronavirus pandemic. This research demonstrated textual and non-textual features of a text corpus as visualizations. However, the data chosen was based only on Twitter data, so the sentiments associated with only that platform were analyzed and not varied by gaining data from other platforms.

**2.2.2 Sentiment Analysis in other domains**

Finance and Political Science domains have also shown the usage of sentiment analysis. (Xing et al., 2018, #) investigates if Natural Language Based Financial Forecasting (NLFF) can yield accurate results in stock prediction. In this study, financial data was sourced from corporate disclosures, financial reports, professional periodicals, aggregated news, message boards, and social media. Text data were preprocessed as a well-formatted input to feed into the algorithms implemented in the predictive model. The predictive model employed various linear regression models and NLP techniques to yield the results. The results show that there is a direct correlation between public sentiment and market trends, the illusion of growth, the positive effect of predictions, and poor market trends.

(Anson, 2020, #) reviewed textual data in the political science sphere to understand the present and future of writing analytics. The author explored a burgeoning area of research and quantitative tool for the analysis of the political text. Political data was sourced from judicial records because of its descriptive nature through speeches, press releases, political advertisements, lobbying, and public opinion through polls and survey data. The results show that text as data is more than writing analytics. Writing analytics is more interdisciplinary and multidisciplinary than political text-as-data, and learning new techniques would allow more substantive research results.

**2.2.3 Challenges in Sentiment Analysis**

While many studies have shown that machine learning techniques have improved the efficiency of sentiment analysis, it is important to understand the challenges that scholars have encountered. (Xu et al., 2022, #) explores the challenges and potential problems in studying sentiment analysis in social media. In studies on sentiment analysis, data volume is one of the first challenges in addition to the imbalance of data, its lengths, bias associated with annotated comments, and the prevalence of figurative language. Social media data is also unstructured and contains noisy, inefficient, and irrelevant information that generally does not provide textual emotional content leading to data processing issues. In terms of the method of evaluation, a major challenge is the versatility of methods employed (i.e., lexicon-based and machine learning based). The lexicon-based approach depends on the presence of predefined words in dictionaries and the absence of a comprehensive dictionary makes it difficult to realize the versatility of lexicon-based methods. On the other hand, a machine-learning-based approach trains sentiment classifiers using labeled data samples from a specific domain. Although these methods achieve great application in the studied domain, they are not great in other domains. Another challenge is the efficiency of the methods employed and the evaluation metrics used to determine the model’s efficiency.

**2.3 Multilingual Textual Analytics**

The authors (Serna A et al., 2021, #) studied the performance of Sentiment Analysis systems by comparing the performance of a classic, unsupervised knowledge-based approach based on sentiment lexicon, SentiWordNet, with respect to a deep-learning classifier obtained by fine-tuning a large pre-trained language model based on the Transformer architecture, namely, XLM-RoBERTa. Their results revealed that the Gold Standard Corpora showed that the supervised approach outperforms the lexicon-based one by a large margin. These results show the huge benefits that can be obtained by using large language models when the amount of training data is very small.

(Can et al., 2018, #) build a reusable sentiment analysis model that does not utilize any lexicons and train a Recurrent Neural Network (RNN) model to predict the polarity of textual data (i.e., reviews from different domains like movie reviews, product reviews etc.,). To evaluate the reusability of the sentiment analysis model, non-English datasets were tested. The non-English test sets were translated to English and used the pre-trained model to score polarity in the translated text. In this way, the proposed approach eliminates the need to train language-dependent models and uses sentiment lexicons and word embeddings for each language. The proposed approach for multilingual sentiment analysis was evaluated using two sets of corpora: training sets and test sets. Training sets included English reviews which were used to build the generalized sentiment analysis model and test sets were used for the evaluation of multilingual approaches for four different languages (Spanish, Turkish, Dutch, Russian). Experimental results showed that the proposed multilingual approach outperforms both the majority baseline and the lexicon-based approaches. It was found that for the Spanish dataset, the results of the lexicon-based approach were below the majority baseline approach, indicating that solely translating data and using lexicons is not sufficient to achieve good results in multilingual sentiment analysis. To see the differences between baseline and RNN the results of each method were grouped and compared to the means. Post hoc comparisons Post hoc comparisons using the Tukey HSD test indicated that the mean accuracies for baselines (majority and lexicon-based) are significantly different than RNN accuracies – meaning each test is significant.

Pham, H., Liang, P. P., Manzini, T., Morency, L.-P., & Póczos, B. (2019) explore learning joint multimodal representations through modality translations. The methodology was using several models to test the design decision, which helped investigate the impact of cyclic translations, modality ordering, and hierarchical structure. The conclusion was that cyclic translations are useful for learning joint representations.

**2.4 Sentiment Analysis using social media data**

Social media platforms such as Facebook, Instagram, and Twitter have drastically altered the way information is generated and disseminated. The profusion of data generated through social media has proved to have the potential to improve the efficiency of existing traffic management systems and transportation analytics. The study by (Haghigh et al., 2018, #) proposes a framework to evaluate transit riders’ opinions about the quality of transit service using Twitter data. Twitter data was used to evaluate people’s opinions about the quality of transit service in Salt Lake City, Utah, and developed a framework to effectively extract tweets relevant to public transit service performance, using a machine learning technique—Latent Dirichlet allocation (LDA) model. Sentiment analysis was then performed to evaluate transit users’ feedback on the quality of transit service and explore the underlying reasons causing users’ dissatisfaction. The findings verify the potential of social media data in analyzing the quality of transit services but various sources of sampling biases persist, which need to be addressed. For instance, in this study, only tweets in English were extracted, which might have excluded members of other communities. Also, low-income people and senior residents are less likely to use smartphones, which represent a significant portion of the missing inputs in our analysis (Haghigh et al., 2018, #).

A. Valdivia, M. V. Luzón and F. Herrera (2017) measured the helpfulness of a review. The research question was the measure of reviews on the three tourist spaces in Spain. The methodology was to extract the polarity of the reviews as found on Tripadvisor and review the differences in the polarity for the negative, neutral, and positive sentiment reviews. The conclusions were Aspect Based Sentiment Analysis (ABSA) is necessary to understand what specific information is being given for a specific entity. Business managers find reviews more helpful where and when a specific opinion is given about a review. Spam, irony, and sarcasm are discovered when sifting through the results, and make the task with automation more difficult to discern the specific helpfulness in a review. Troiano, L., Cambria, E., & Herrera, F. (2019) examined polarities within a review. The research question explored if SAM and ABSA be used to examine user polarity and the polarity within the reviews. The methodology showed that the Sentiment Analysis Methods (SAM) are lexicon dictionary-based methods, machine learning-based models, and deep learning-based methods. The Aspect Based Sentiment Analysis (ABSA) used one input layer, two convolution layers, two max pool layers, and a fully connected layer with a softmax input. The data is from web scraping within Tripadvisor. The conclusion was users tend to rate the overall experience but within the review, there are different polarities, and the study was limited. The behavior model was studied using only one parameter of the overall sentiment.

To understand the relationship between textual analytics and transportation modes, this study analyzed lyrics of popular songs from 60 years to identify the change in social attitudes toward automobiles. A list of songs was prepared from Billboard's year-end hot 100 singles and their genre was identified from iTunes. A total of 2400 songs were chosen and tokens related to automobile mobility were found. Eight different categories were determined to classify the identified lyrics and analyze the most associated sentiment. Two open source algorithms – IBM Alchemy language to give a continuous score from negative to positive and IBM Watson Tone Analyzer to find the emotion of the song were used for further analysis. A bivariate correlation was done to establish the significance between year and sentiment, which turned out to be significant. However, when the regression was done between all four classifications – 2 manual and 2 algorithms, all 4 models were found to be significant but the goodness of fit was low for algorithms. These models identified the sentiment as negative but the human scoring showed it is mostly positive. This contradiction made it clear that a clear interpretation was nearly impossible to determine. Hence, this research shows us the importance of training a model.

Another study on analyzing the traffic safety of roads was conducted by examining sentiments of reviews on the Web. Autostrada was used as the data source for topic-related text processing and mining on up-to-date road conditions. Text cleaning was done using CountVectorizer and TFIDFVectorizer for vectorization and lexical analysis to process the downloaded data by removing noise such as punctuation and normalizing the data. To classify the sentiments, MultinomialNB and SGDClassifier were used by inputting 255 tagged tweets. Cross-validation between the various combinations was done and the linear model with Ngram was found to be efficient. The study identified that positive reviews were 75% with a classification accuracy of 71.94% for the overall data. The next step for this study would be to classify reviews based on topic groups.

**2.5 Sentiment analysis on public transportation**

G. T. Giancristofaro and A. Panangadan (2016) questioned if feelings towards public transportation can be predicted using social media. 1,200 posts from IG using the IG API that was filtered for only *caltrans* using the hashtag #caltrans. Images and texts were stored in a database and then cleaned to the final 1,010 posts and annotated by two people using a java based web application called PNClassifier that used MySQL. Values were created and data sets we pre-processed and trained. It was understood that visual features with social media posts are not more informative than textual posts. Visual and textual posts together present more information and increase the accuracy of the sentiment prediction.

Paget-Seekins, L., & Tironi, M. (2016) questioned the sincere view of how public transportation operates. It was a review of several transit options that were located in various Latin American cities Chile, Colombia, Ecuador, and Mexico. Public transportation is not treated as a shared space with the users and the creators. The commonality in a common space can create a collective experience or promote disharmony. The politicization of the planning portion creates a problem in which people who are unprofessional planners have opinions about the creation. The financing is often obscure because state ownership has to allocate funds from different organizations and that leads to fundamental differences of opinion. Public transport is always seen as a backup option that creates participation with spatial justice. There are several options for relationships between public usage, government agencies, and private operators in which the monopoly and participation can be public or private.

To study the sentiments related to transit usage, we found this research closely associated with improving the performance of the commuter electric line, Trans Jakarta. User reviews were analyzed for positive and negative sentiments by gathering the data through Web crawling using the request, ‘commuter line’ in the Indonesian language for a one-year period. Manually checked for positive, negative, and neutral tweets, and data was cleaned following the processes of stemming, normalization, and stop word removal before feeding into 3 algorithms – Naïve Bayes, Random Forest, and Support Vector Machine. Data was fed into the algorithms before and after cleaning to make sure the three models were efficient to an extent. Out of these, the Support Vector Machine algorithm was found to perform better and analyzed the data to be most negatively sentimentalized.

Using data mining to evaluate public transportation can be extremely beneficial to transportation planners and policymakers. Analyzing the service for spatial and temporal dynamics can result in streamlined research findings that will help improve the services. A study on metro services in Shenzhen, China (Shuli Luo and Sylvia Y He ., 2021 #) was done based on user-generated text from Weibo microblogs. The focus of this research was to establish Service Quality Attributes based on user perceptions and not a planner's or decision-makers assumption. The transportation network is growing at a faster pace and the data associated with that is dynamic. Hence, analyzing it based on various parameters will give focused findings. For this study, web crawling was used to download data with various advanced search criteria to filter out duplicate or marketing accounts, using the keyword "metro". Data cleaning was done by filtering irrelevant blogs associated with the keyword, manually coding marketing accounts, removing symbols, and converting traditional to simplified Chinese.

Jieba segmentation model was used to convert the text into bag-of-words and a custom dictionary was used with transit-related words to develop probability language modeling. The next step is to form keyword-in-context lists and match them with appropriate service categories from the six categories that were formed. Sentiment analysis was done by counting the number of positive and negative words in a sentence and a score was assigned to it. Temporal information along with geo-tagged location was fed into ArcGIS to establish clusters based on each SQA. Results successfully established five SQAs- crowdedness, reliability, waiting conditions, personal behavior, and safety which can be used to analyze any transit service to make improvements. The shortcomings of this study were that it included only people who used microblogs to express their opinions, which left most of the population over 40 years of age. This age group tends to comment most on the safety of the network, also usage of the keyword metro might have brought much-unstructured text that could have led to missing any quality SQA. However, analyzing public sentiments toward a transit network is critical to make improvements.

**3 Literature Synthesis**

The literature review showed ways that the sentient analysis can be displayed through numerous methods, and how it could be helpful in determining if the analyses through different sentiment lexicons were more accurate than a native speaker’s knowledge.

According to Serna A et al., 2021 using large language models for Spanish is good when the amount of training data is very small. Studies on Sentiment Analysis with social media have been completed and were used as guidance to help understand the various approaches that can be used within this project. The commentary on transit service in Salt Lake City, Utah was analyzed to show the relationship between transit service and rider experience. The analysis utilized a Latent Dirichlet Allocation model that used the rider’s and user’s experience and disappointment to highlight reasons the transit service needs to be improved. This study identified social media data can be utilized by a service provider to explore a bias in user experience. An additional note in this study was that the user bias is representative of a small sample because there can be focus groups that do not share opinions. Every review has a polarity and each review has a polarity calculated to show a score for negative, neutral, and positive portions in a comment or question. In this study, an ASBA was necessary to retrieve the specific scores to identify the rating for a specific entity. This would help the SAM to rate the user experience of the overall feeling. Studied the safety of roads by examining Twitter data to validate the accuracy and efficiency of the road. It was discovered that positive reviews were more accurate than negative or neutral reviews and kept better track of updating road conditions.

In the study where user reviews of Trans Jakarta were collected and analyzed to improve the experience and service for their commuter service lines. For one year reviews from Twitter were verified for categorical significance negative, neutral, and positive. The reviews were fed into several algorithms in an effort to find a pattern. Unsurprisingly, most of the user reviews were negative, and this was determined by the most accurate and therefore better performing algorithm *Support Vector Machine*. From all these studies, we have understood that social media data can be used to identify sentiments towards a service or a product using artificial intelligence technologies. For languages other than English, models need to be trained before performing the sentiment analysis.

**4 Methods**

This section explains the methodology of deriving the sentiment of the users of the Renfe transit system. The data was sourced from *Twitter* and *Trip Advisor*. The analysis was completed in a *Google Colaboratory* to execute arbitrary python code in a browser.

**4.1 Data Collection**

The data was collected by sourcing *Twitter* data with the *Twitter API* and *Rtweet* package in *R Studio* using the word Renfe in the search query. In addition, data was collected from public user responses on *TripAdvisor*. The data from Twitter was chosen within a 10-day time frame in the month of November 2022 and the TripAdvisor data was chosen from 2017 to 2022.

**4.2 Data Cleaning and Processing**

The data was cleaned by removing special characters, stopwords (and, are, I, is, the, my, Renfe, train), interactions (which include embedded media, hashtags, likes, links, replies, retweets, and usernames) and cleaning grammatical errors with the help of a native Spanish speaker before processing. The data had to be simplified into three categories based on their length to verify the efficiency of the sentiment The categories were divided into short with 1 to 6 words, medium with 7 to 14 words, and long with 15 to 25 words. The maximum length of the data did not exceed 25 words because there were processing time issues, and the translators were not responding as a result of runtime errors.

**4.3 Data Transformation**

The text was then translated from Spanish to English using *Google* and *Opus MT* machine translation engines to generate a gold-standard corpus. In this study, the gold standard corpus was manually labeled negative, neutral, and positive sentiment information of the original text. The gold standard corpus was then used to compare the sentiment scores derived from the sentiment analyzers.

**4.4 Choice of Data Mining Task and Algorithms**

We applied three sentiment analyzers *NLTK*, *Pattern*, and *Textblob* for text classification at the sentence level. The resulting sentiment scores were then compared with the gold standard.

**4.5 Evaluation and Interpretation**

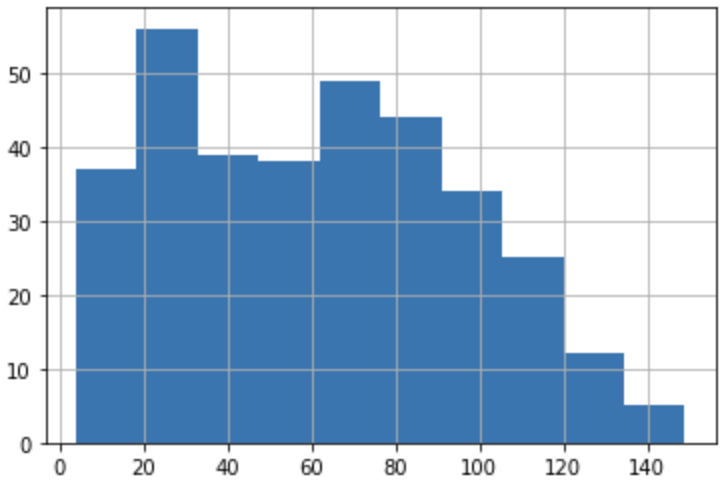
The data was evaluated and interpreted using *gensim*, *pyldavis*, *seaborn*, *sklearn*, *spacy*, *textstat*, and *word cloud* packages. Error analyses were performed to identify the gaps between the gold standard corpus and the translated text data to determine the efficacy of the machine translation engines.

**5 Discussion**

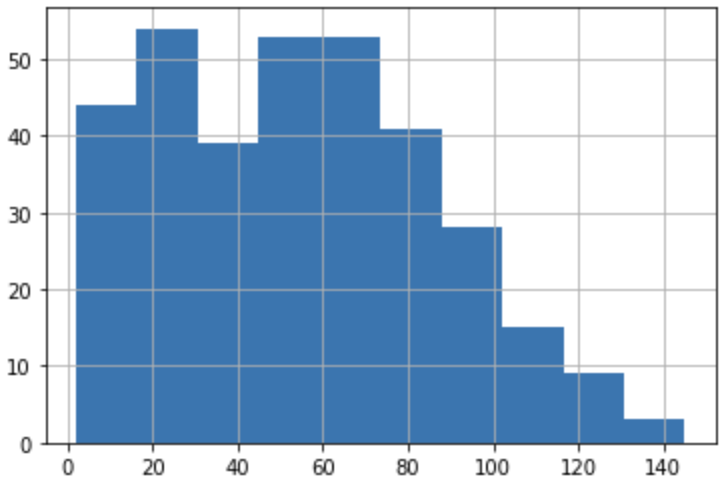
This section discusses the features of the dataset and the polarity scores obtained from NLTK, Textblob and Pattern sentiment analyzers while also highlighting the errors and gaps with the tools applied in determining the sentiment analysis. We noticed that while cleaning the tweets, some words had a special character, and when the punctuation was being cleaned the letter was removed and the resulting data was broken and incomplete. The language expert reviewed the broken and incomplete words to repair a few comments, so they could be utilized by machine learning techniques.

**5.1. Text Statistics**

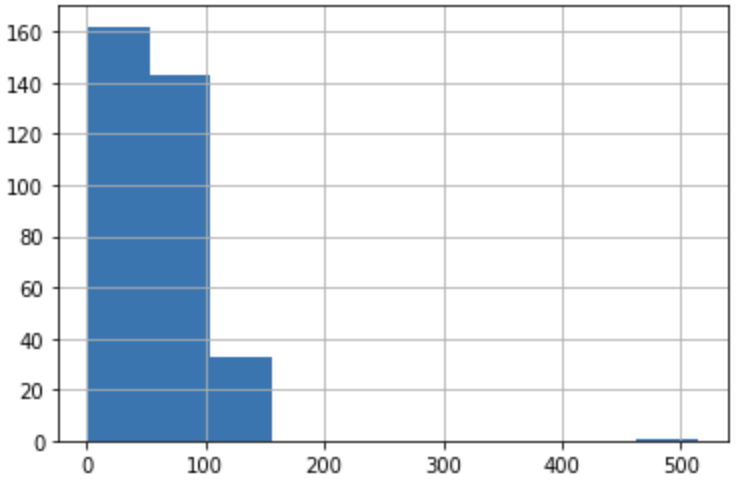
Text statistic visualizations have been used to explore the fundamental characteristics of textual data. The histograms in figures 1-3 below show the number of characteristics present in each sentence of the original Spanish text, google translated text, and opus-mt translated text. It can be said that the selected textual data ranges from 5 to 150 characteristics. The x-axis in the figures below represents the character count and the y-axis represents the number of sentences.



*Figure 1: Character Count of Original Spanish Text*

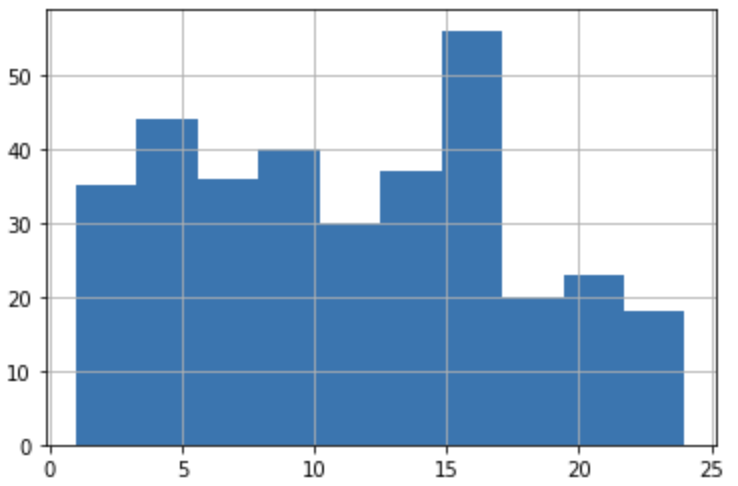


*Figure 2: Character count of Google Translate Text*

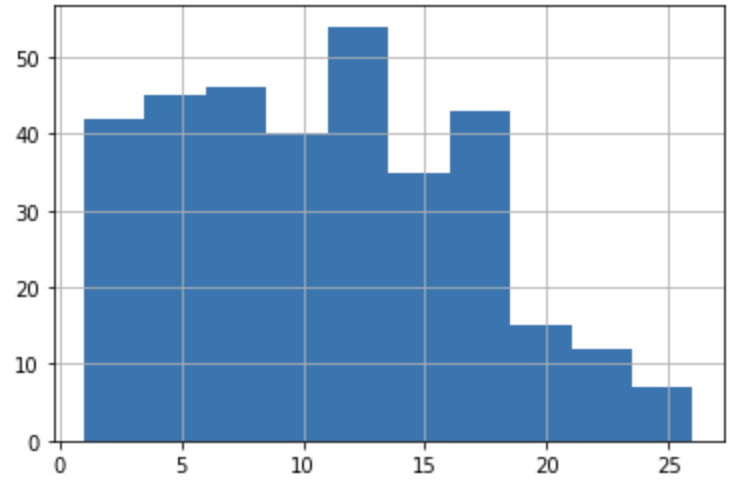


*Figure 3: Character count of Opus-mt Translate Text*

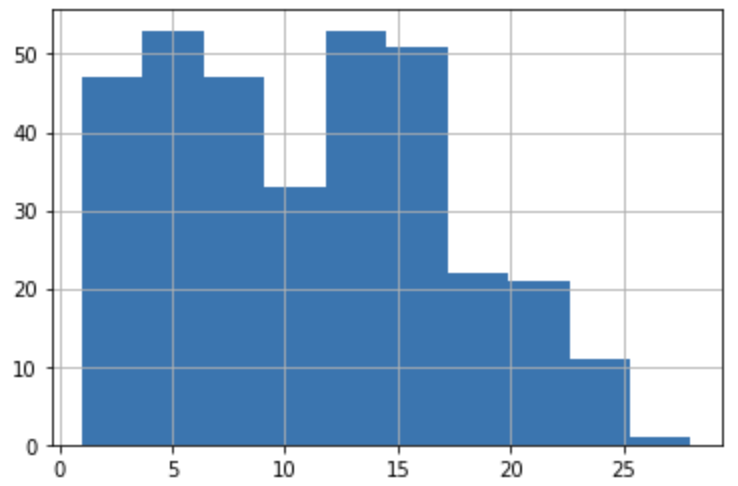
The histograms in figures 4-6 below show the word counts of the textual data for all the three categories (i.e., Original Spanish Text, Google translate and Opus-mt translated text). It can be seen that the word count ranges from 1 to 27 words for all categories of textual data



*Figure 4: Word count of Original Spanish Text*

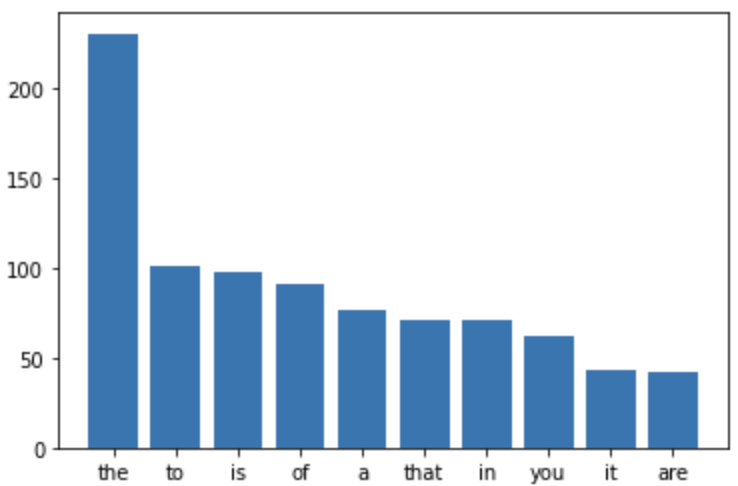
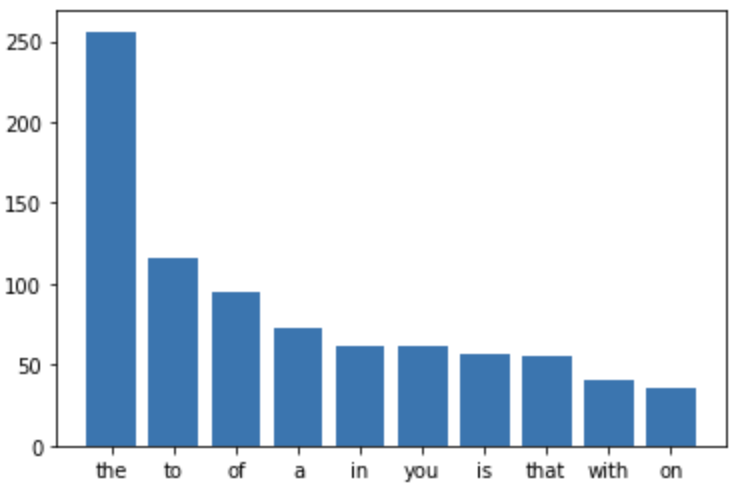


*Figure 5: Word count of Google Translate Text*



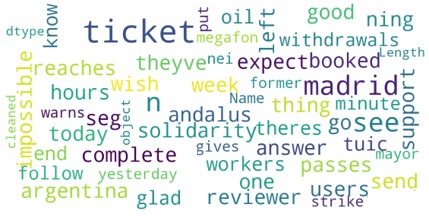
*Figure 6: Word count of Opus-mt Translate Text*

We also observed the frequency of stop words in the data set before cleaning the translated Spanish text to avoid discrepancies in the analysis. It can be seen that the top 10 stop words in Google translated text and Opus-mt translated text are different from each other. .

*Figure 7: Stop words of Google Translate Text Figure 8: Stop words of Opus-mt Translate Text*

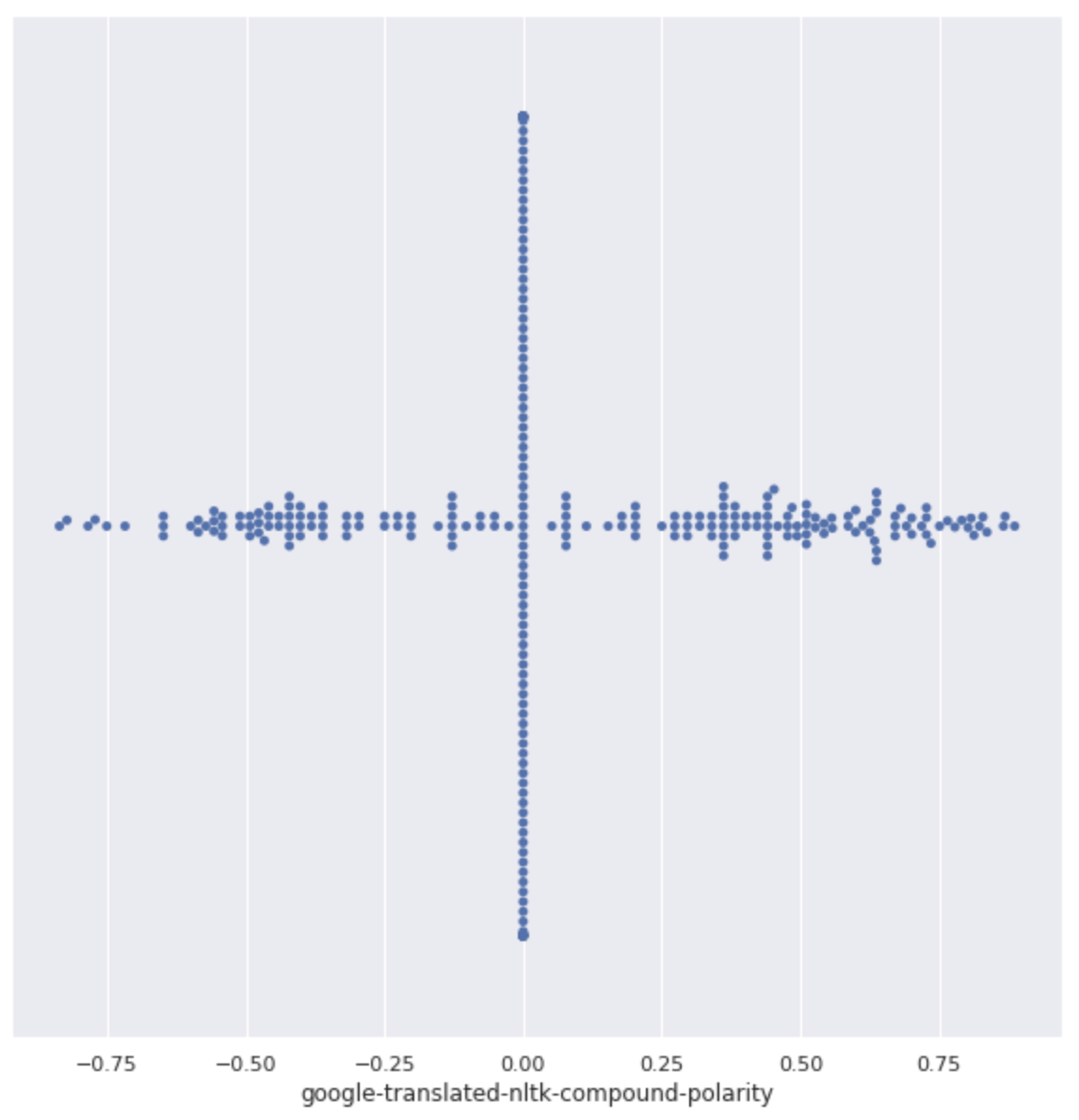
Additionally, we looked at word clouds to represent text data. The size and color of each word that appears in the word cloud indicate its frequency and importance. The following word clouds indicate the frequency of words in google translated text and opus-mt translated text.

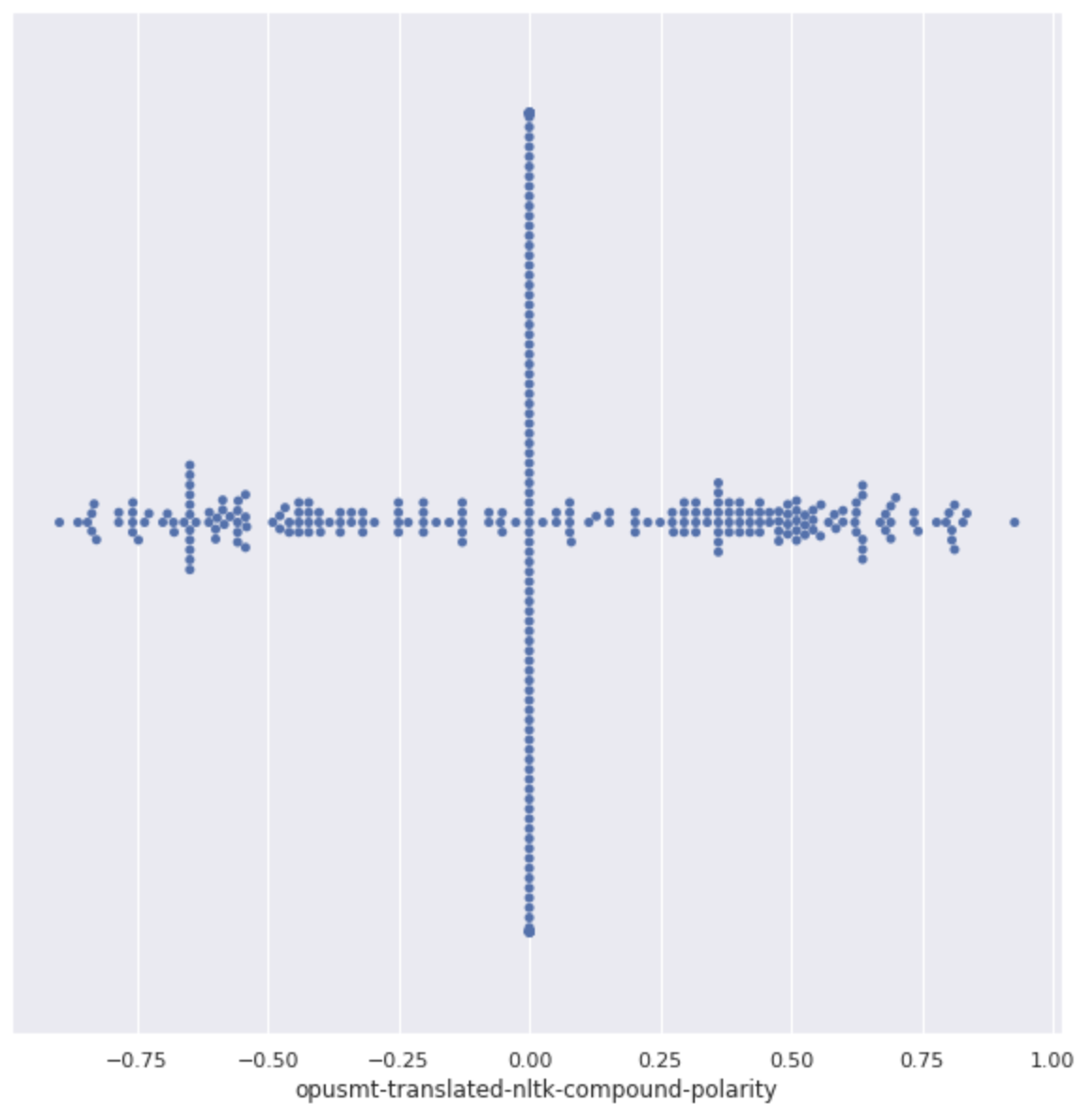
*Figure 9: Word cloud of Google Translate Text Figure 10: Word cloud of Opus-mt Translate Text*

**5.2 Sentiment Analysis Statistics**

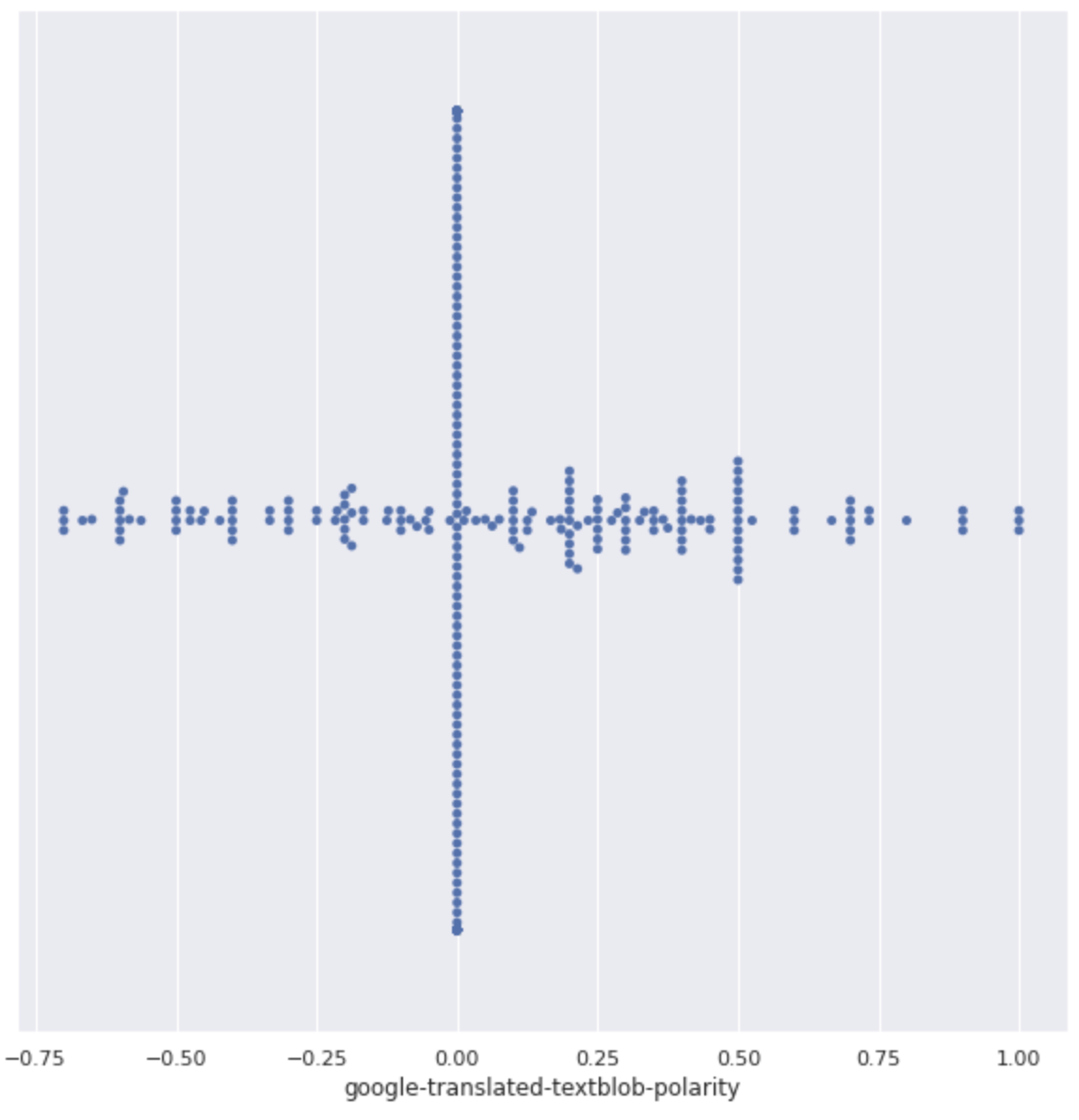
To understand the distribution of polarity scores for both the translated texts and for all three sentiment analyzers, we used Swarm Plots to show the observed scores of an underlying distribution. The figures below indicate that NLTK polarity for both translations ranges from -0.75 to 0.9 while Textblob polarity ranges from -0.75 to 1 and Pattern polarity ranges from -1 to 1.



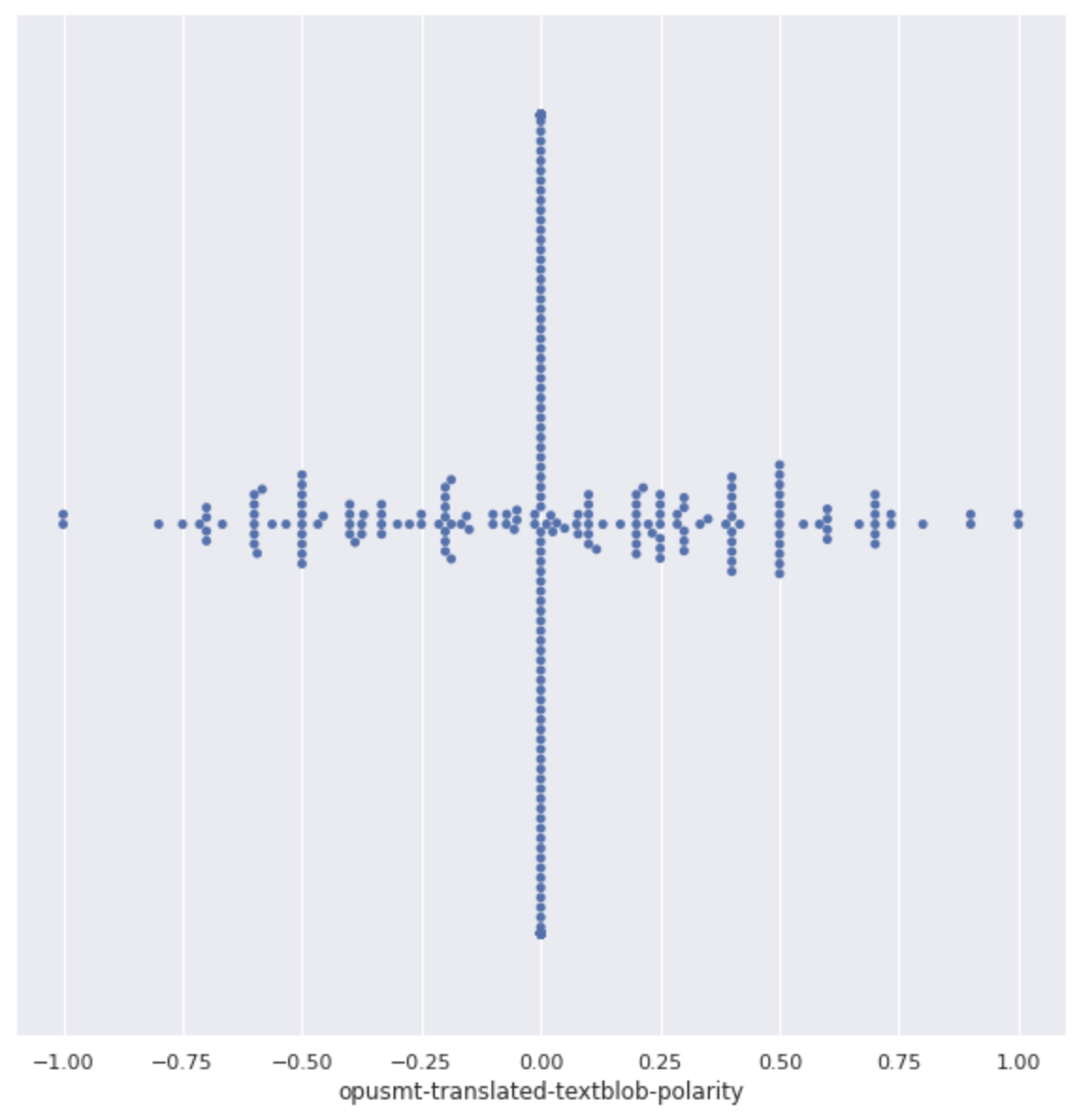
*Figure11: Swarm Plot of NLK Polarity for Google translated text*



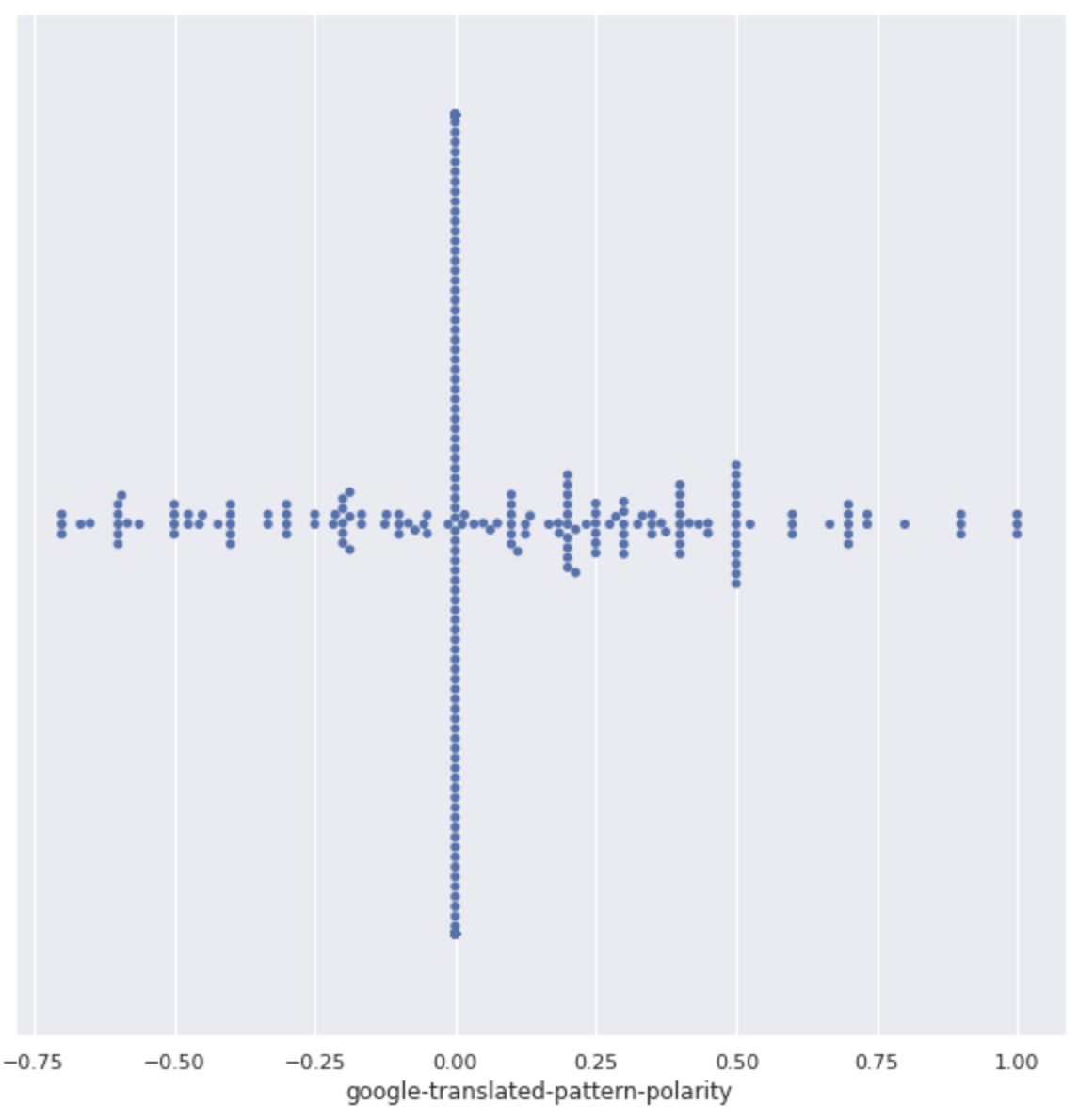
*Figure 12: Swarm Plot of NLK Polarity for Opus-mt translated text*



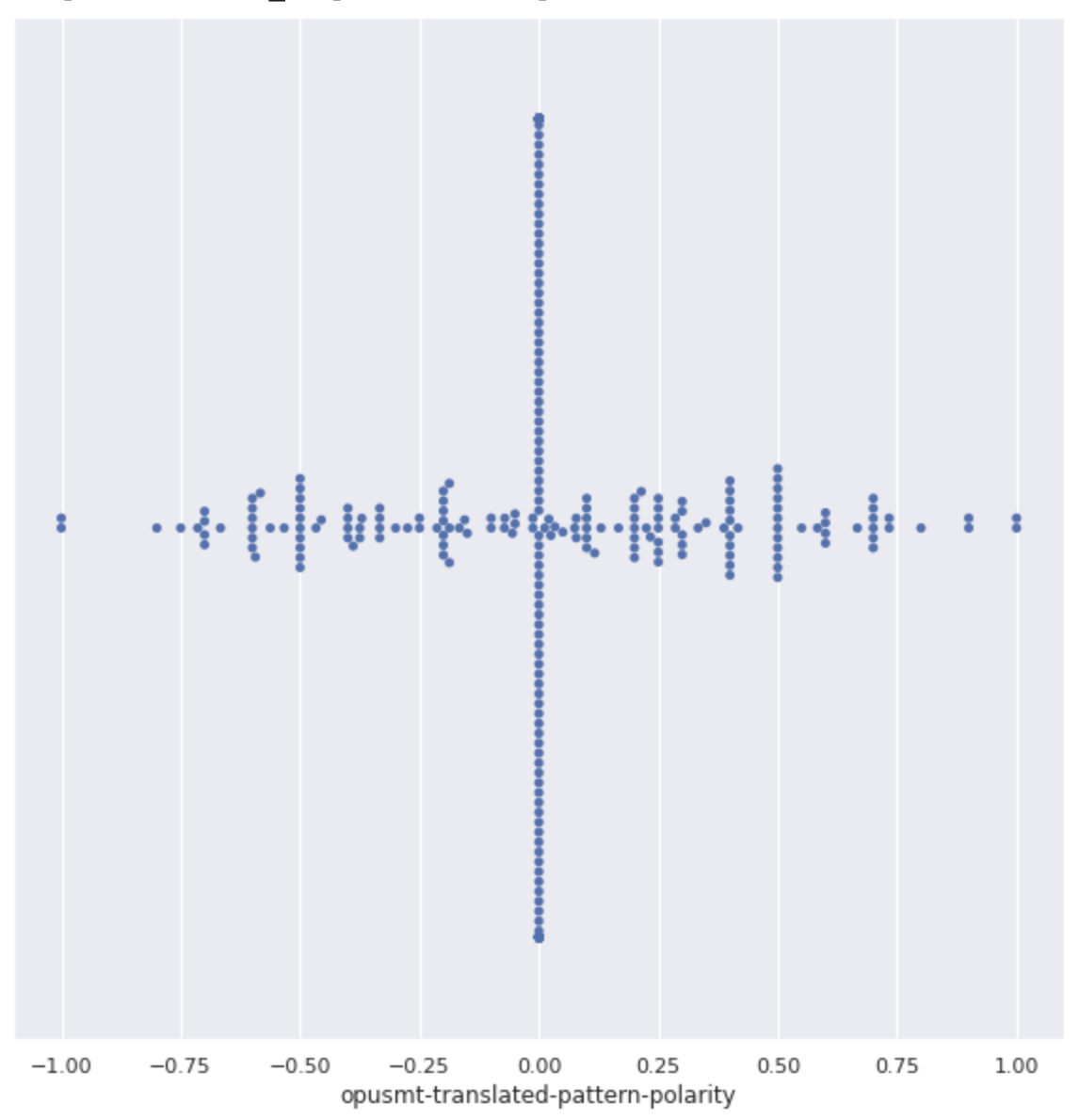
*Figure 13: Swarm Plot of Textblob Polarity for Google translated text*



*Figure 14: Swarm Plot of Textblob Polarity for Opus-mt translated text*

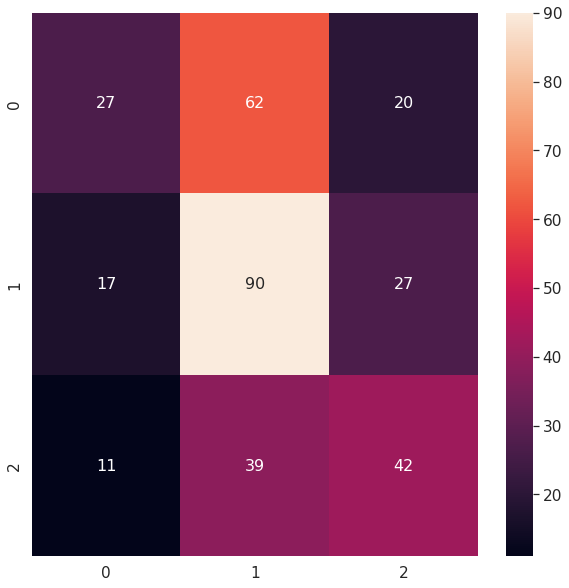
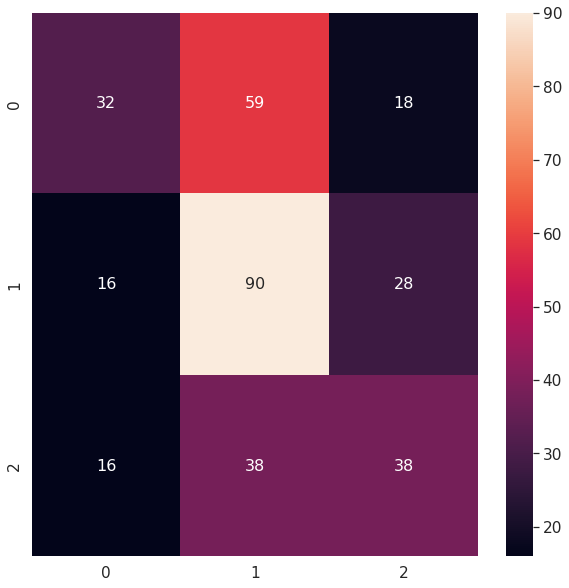


*Figure 15: Swarm Plot of Pattern Polarity for Google translated text*



*Figure 16: Swarm Plot of Pattern Polarity for Opus-mt translated text*

To summarize the performance of the models, we use confusion matrices that provide a class-wise distribution of the sentiment analyzers. The three parameters accuracy, recall rate, and specificity rate were derived to measure the efficiency. Accuracy is the percentage of data that was predicted correctly compared to the total data, Recall rate is the number of positive sentences that were predicted correctly and Specificity is the number of negative sentences that were predicted correctly. All these parameters were computed using the gold standard as the base of the matrices.

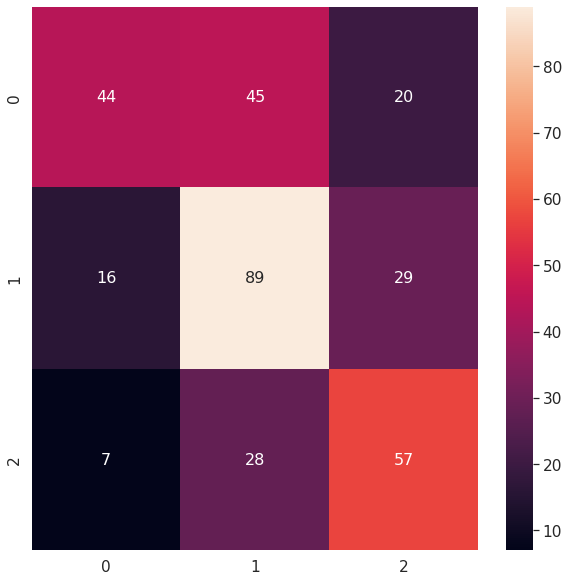
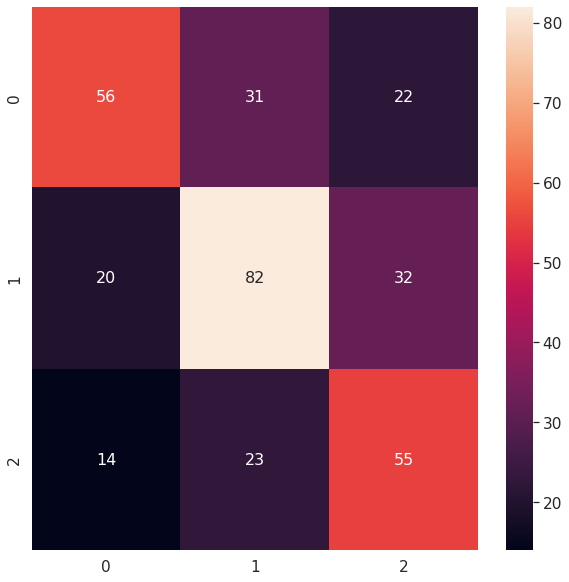
** **

*Fig 17:Confusion matrix of Textblob for Google translate Fig 18:Confusion matrix of Textblob for Opus translate*

**Accuracy rate = 47.46% Accuracy rate = 47.76%**

**Recall rate = 0.5744 Recall rate = 0.64**

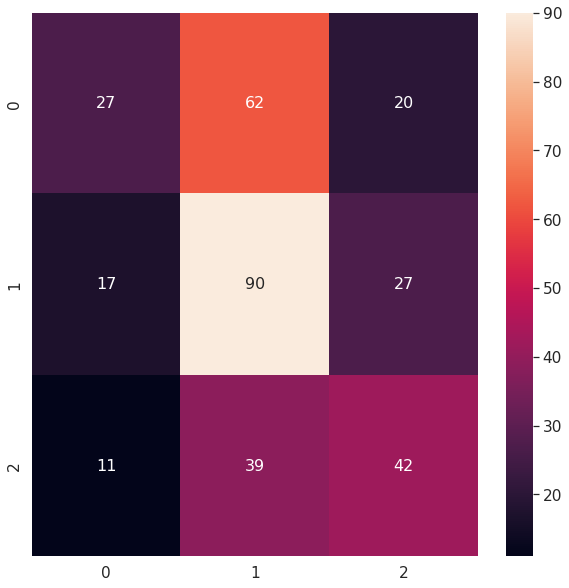
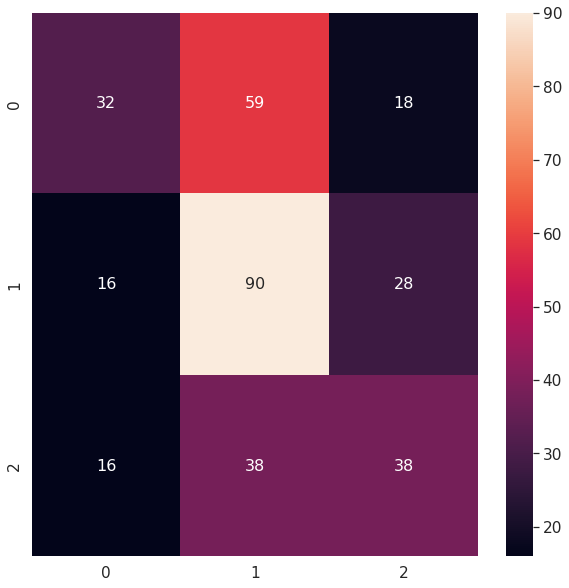
**Specificity rate = 0.7924 Specificity rate = 0.7037**

** ***Fig 19:Confusion matrix of NLTK for Google translate Fig 20:Confusion matrix of NLTK for Opus translate*

**Accuracy rate = 56.72% Accuracy rate = 57.6%**

**Recall rate = 0.6875 Recall rate = 0.7179**

**Specificity rate = 0.8906 Specificity rate = 0.7971**

** **

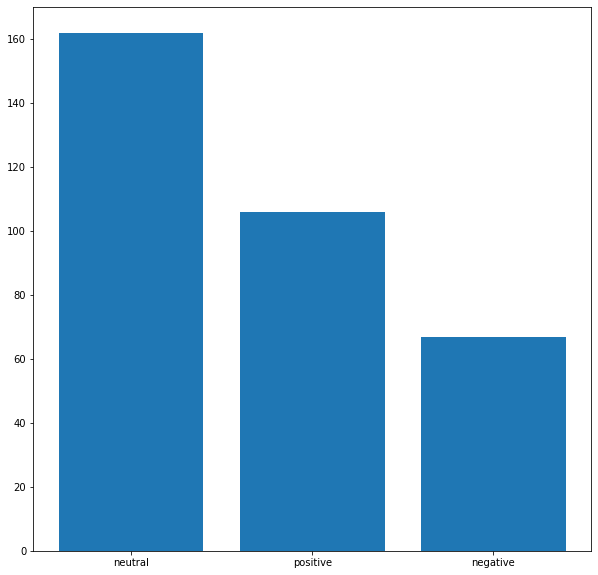
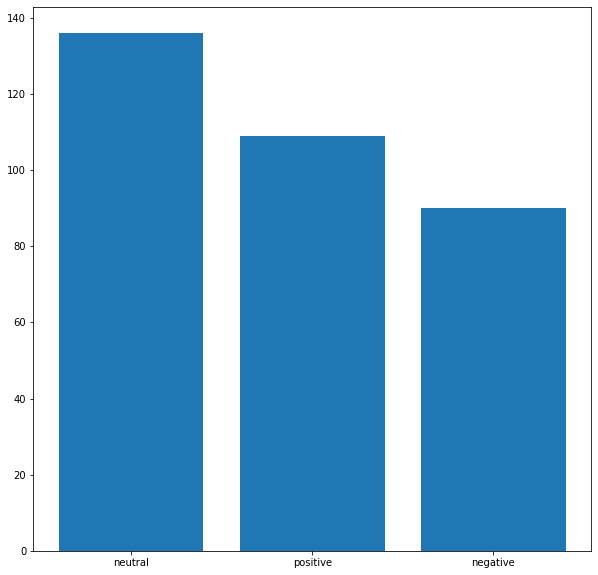
*Fig 21:Confusion matrix of Pattern for Google translate Fig 22:Confusion matrix of Pattern for Opus translate*

**Accuracy rate = 47.46% Accuracy rate = 47.76%**

**Recall rate = 0.5744 Recall rate = 0.64**

**Specificity rate = 0.7924 Specificity rate = 0.7037**

Based on the accuracy and other parameters from the confusion matrices, we have found that among the three analyzers used, NLTK had higher efficiency. So we decided to proceed with NLTK’s polarity scores to determine the final results of our research study. Figures 23 and 24 show the composition of positive, negative, and neutral sentences in the whole dataset with both Google and Opus MT translators and sentiment analyzer as NLTK. In both graphs, it can be observed that the sentiment in a majority of sentences is neutral. This could be due to the limited capability of translators to convert a few Spanish words or identify sarcasm, which is one of the shortcomings of these models.

** **

*Fig 23:Distribution of sentiments for Google translate Fig 24:Distribution of sentiments for Opus MT translate*

**5.3 Error Analysis**

The sentiment analyzers identified the sentiments for all the sentences that was translated with Google translate and Marian MT translator. When compared with the gold standard provided by the language expert, there were patterns of errors from the translations which could have potentially altered the sentiments. Table1 shows some of the identified errors in the translations of short sentences. One of the words that stood out was laundering and washing money, as seen in the table. Similarly, tables 2 and 3 show errors in the translation of medium and short sentences.

*Table 1: Error analysis of short sentences*

| **Statement** | **Expert Translation** | **Gold Standard** | **Google Translate** | **Opus-mt Translate** | **NLTK** | **Textblob** | **Pattern** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| gracias un por joderle los planes alguien | thank you for playing with somones plans | Positive | thanks f\*\*\*ing plans someone | thank screwing plans someone | 0.4404 | -0.2 | -0.2 |
| habra alguien que vayas bien | there will be someone doing well | Positive | do well | theres someone whos going well | 0.2732 | 0 | 0 |
| estan lavando dinero | They are laundering money | Negative | washing money | theyre washing money | 0 | 0 | 0 |
| adjunto foto la cola ya estaba peque | i attached the photo the queue was already small | Neutral | attached photo tail already small | attached photo queue already small | 0 | -0.25 | -0.25 |
| pero sigue siendo obligatorio en renfe metro | but it is still mandatory in the renfe metro | Neutral | still mandatory renfe metro | still mandatory renfe metro | 0.0772 | 0 | 0 |

*Table 2: Error analysis of medium sentences*

| **Statement** | **Expert Translation** | **Gold Standard** | **Google Translate** | **Opus-mt Translate** | **NLTK** | **Textblob** | **Pattern** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| sabe gestionar esto hay que ser tiles e ineficaces | he knows how to manage this you have to be useful and effective | Positive | knows manage tiles ineffective | know manage subtle ineffective | -0.128 | 0 | 0 |
| se quien eres persona random de la renfe pero felicidades por subir de nivel tu pou | i know who you are random person from renfe but congratulations for leveling up | Positive | know random person renfe congratulations leveling pou | know random person renfe congratulations leveling pou | 0.5994 | -0.5 | -0.5 |
| lo le respondes este los dem que te han dejado en ridiculo | You have responded to others that have made a fool of you | Negative | answer dem left rid | answer one dem left ass | 0 | 0 | 0 |
| critica las salas club de renfe | critique of the renfe club rooms | Negative | renfe club salas | crtica club halls renfe | 0 | 0 | 0 |
| se van pareciendo las consultas de la atencion primaria en madrid esta barato todo | Primary care consultations in Madrid are looking cheap | Neutral | Primary care consultations in Madrid are looking like everything is cheap | Primary care consultations in Madrid are looking like everything is cheap | 0.6908 | 0.4 | 0.4 |
| sale cuando quiero comprar ida vuelta siempre los tengo que comprar por separado | exits to when i want to buy roundtrip always have to buy separate | Neutral | leaves want buy round round always buy separately | comes want buy back always buy separately | 0.0772 | -0.2 | -0.2 |

Table 2 has error analysis of medium sentences from the two translators. In the first sentence, the word useful was translated into tiles in google translation and subtle in opus translation. The real meaning of the word was not captured in either of the translators and has shown a neutral score in Textblob and Pattern while the real sentiment is close to positive.

*Table 3: Error analysis of long sentences*

| **Statement** | **Expert Translation** | **Gold Standard** | **Google Translate** | **Opus-mt Translate** | **NLTK** | **Textblob** | **Pattern** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| parada en la renfe entre plaza cada santos rodeada de franceses dios da sus peores batallas sus mejores guerreros yo | he stops at the renfe between plaza surrounded by every saint god gives his worst battles to his best warriors me | Positive | stop renfe plaza cata sants surrounded french god gives worst battles best warriors yo | stop renfe square cata sants surrounded french god gives worst battles best warriors | -0.3818 | 0 | 0 |
| uno puede quejarse si le da la gana nos tienes que dar permiso es alucinante | you can complain if you feel like but you have to gives us permission it is amazing | Negative | one may complain feel like give us permission amazing | complain feel like give us permission amazing | 0.59 | 0.6 | 0.6 |
| pero como lo hacen en argentina ocurre lo mismo es imposible conseguir pasajes de tren | They do the same thing in Argentina it is difficult to get tickets | Negative | thing happens argentina impossible get train tickets | argentina thing impossible get train tickets | 0 | -0.67 | -0.67 |
| trenes comprados de segunda mano renfe espa en extra condiciones queda la mitad de la flota funcionando | second hand bought trains with space in extra conditions half of the fleet remains working | Neutral | second hand buy tranes renfe espo extra conditions half fleet working | trains bought second hand renfe spa extra conditions half fleet running | 0.49 | -0.06 | -0.06 |
| una multa los usuarios que hayan reservado vayan en el tren ya ver como espabilan | The fine users who have booked the train lets see how they wake up | Neutral | fine users reserved go train see spab | fine users booked go train see expect | 0.2 | 0.42 | 0.42 |

**6 Future Research Possibilities**

In the future, the data could be retrieved with enhanced search criteria by omitting advertisements tweets and by adding location and topic-specific jargon. When we reviewed stations, there was a varied amount, and during the analysis, it was determined that a specific route name or station identification could retrieve a better sentiment. This would include improving our search criteria by adding abbreviations. Additionally, there were a lot of Twitter interactions in the data set, and it was hard to sort because it was already present and counted as a response. Therefore, most retweets did not provide true sentiment from a user experience. The study could have been improved by exploring other machine-based sentiment analyzers for our data analysis. Furthermore, a public transit system's performance can be improved by considering temporal information within a specific time frame to understand the user experience.

**7 Conclusion**

Our study demonstrates the usage of various translators and sentiment analyzers for textual analytics with multilingual data. Based on our results, it can be derived that the existing models have a low accuracy rate when using languages other than English. By using finetuned datasets and training the models, there is potential to increase the efficiency of these analyzers to detect accurate sentiments. Various machine learning models, which are usually more accurate than dictionary-based models, can be used to compare and improve the accuracy of sentiments. Future studies should be done to minimize the shortcomings of this study to establish a strong sentiment analysis of the public transit system for policy recommendations and improving the services. To conclude, textual analytics is of paramount importance in the coming years in natural language processing as it helps improve a myriad of daily events.

**8 References**

A. Valdivia, M. V. Luzón and F. Herrera, "Sentiment Analysis in TripAdvisor," in IEEE Intelligent Systems, vol. 32, no. 4, pp. 72-77, 2017, DOI: 10.1109/MIS.2017.3121555.

Ali, G. G. Md. N., Rahman, Md. M., Hossain, Md. A., Rahman, Md. S., Paul, K. C., Thill, J.-C., & Samuel, J. (2021). Public Perceptions of COVID-19 Vaccines: Policy Implications from US Spatiotemporal Sentiment Analytics. *Healthcare*, *9*(9), 1110.<https://doi.org/10.3390/healthcare9091110>

Can, E. F., Ezen-Can, A., & Can, F. (2018, July 12). Multilingual Sentiment Analysis: An RNN-Based Framework for Limited Data. Ann Arbor, Michigan, USA: ACM. . DOI.

discovery of dominance in synchronous computer-mediated communication. In 2014 47th Hawaii International Conference on System Sciences (pp. 1804-1812). *IEEE Access*.

Drieger, P. (2013). Semantic Network Analysis as a Method for Visual Text Analytics. *Procedia - Social and Behavioral Sciences*, *79*, 4–17.<https://doi.org/10.1016/j.sbspro.2013.05.053>

G. T. Giancristofaro and A. Panangadan, "Predicting sentiment toward transportation in social media using visual and textual features," 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC), 2016, pp. 2113-2118, DOI: 10.1109/ITSC.2016.7795898.

Gefen, D., Endicott, J. E., Fresneda, J. E., Miller, J., & Larsen, K. R. (2017). A Guide to Text Analysis with Latent Semantic Analysis in R with Annotated Code Studying Online Reviews and the Stack Exchange Community. Communications of the *Association for Information Systems*, *41*, 450–496.<https://doi.org/10.17705/1cais.04121>

Lang, S. (2020). Writing Analytics: Expanding the Body of Knowledge. The *Journal of Writing Analytics*, *4*(1), i–iv.<https://doi.org/10.37514/jwa-j.2020.4.1.01>

Luo, S., & He, S. Y. (2021). Using data mining to explore the spatial and temporal dynamics of perceptions of metro services in China: The case of Shenzhen. Environment and Planning B: Urban Analytics and City Science, 48(3), 449–466.<https://doi.org/10.1177/2399808320974693>

Malsburg, T., Poppels, T., & Levy, R. P. (2020). Implicit Gender Bias in Linguistic Descriptions for Expected Events: The Cases of the 2016 United States and 2017 United Kingdom Elections. *Psychological Science*, *31*(2), 115–128.<https://doi.org/10.1177/0956797619890619>

Paget-Seekins, L., & Tironi, M. (2016). The publicness of public transport: The changing nature of public transport in Latin American cities. In Transport Policy (Vol. 49, pp. 176–183). Elsevier BV. <https://doi.org/10.1016/j.tranpol.2016.05.003>

Pham, H., Liang, P. P., Manzini, T., Morency, L.-P., & Póczos, B. (2019). Found in Translation: Learning Robust Joint Representations by Cyclic Translations between Modalities. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 6892-6899.<https://doi.org/10.1609/aaai.v33i01.33016892>

Pratama, M. O. (n.d.). The sentiment analysis of Indonesia commuter line using machine learning based on Twitter data. IOPscience.

Qianwen Ariel Xu, V. C. (2022). A Systematic Review of Social Media-Based Sentiment Analysis: Emerging Trends and Challenges. *Decision Analytics Journal 3*

Redondo, T., & Sandoval, A. M. (2016). Text Analytics: the convergence of Big Data and Artificial Intelligence. *International Journal of Interactive Multimedia and Artificial Intelligence*, *3*(Special Issue on Big Data and AI).<https://www.ijimai.org/journal/bibcite/reference/2540>

Samuel, J. (2020, July 1). *Conceptual Frameworks for Big-Data Visualization: Discussion on Models, Methods and Artificial Intelligence for Graphical Representations of Data - Call for Extension*. Papers.ssrn.com.<https://ssrn.com/abstract=3640058>

Samuel, J., Ali, G. G. Md. N., Rahman, Md. M., Esawi, E., & Samuel, Y. (2020). COVID-19 Public Sentiment Insights and Machine Learning for Tweets Classification. *Information*, *11*(6), 314.<https://doi.org/10.3390/info11060314>

Samuel, J., Holowczak, R., Benbunan-Fich, R., & Levine, I. (2014, January). Automating

Samuel, J., Rahman, Md. M., Ali, G. G. Md. N., Samuel, Y., Pelaez, A., Chong, P. H. J., & Yakubov, M. (2020). Feeling Positive About Reopening? New Normal Scenarios From COVID-19 US Reopen Sentiment Analytics. *IEEE Access*, 8, 142173–142190.<https://doi.org/10.1109/access.2020.3013933>

Serna, A., Soroa, A., & Agerri, R. (2021). Applying Deep Learning Techniques for Sentiment Analysis to Assess Sustainable Transport. *Sustainability*, 1-19.

Shao, N. N. (2018, August 02). Using Twitter data for transit performance assessment: a framework for evaluating transit riders’ opinions about quality of service. *Springer*, pp. 363–377.

Tausczik, Y. R., & Pennebaker, J. W. (2009). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. <https://doi.org/10.1177/0261927x09351676>

Troiano, L., Cambria, E., & Herrera, F. (2019). Inconsistencies on TripAdvisor reviews: A unified index between users and Sentiment Analysis Methods. In Neurocomputing (Vol. 353, pp. 3–16). Elsevier BV.<https://doi.org/10.1016/j.neucom.2018.09.096>

Wu, C., Le Vine, S., Bengel, E., Czerwinski, J., & Polak, J. (2022). Sentiment analysis of popular-music references to automobiles, 1950s to 2010s. *Transportation*, *49*(2), 641–678.

Xing, F. Z., Cambria, E., & Welsch, R. E. (2017). Natural language based financial forecasting: a survey. *Artificial Intelligence Review*, *50*(1), 49–73.<https://doi.org/10.1007/s10462-017-9588-9>

Yaroslav Seliverstov, Svyatoslav Seliverstov, Igor Malygin, Oleg Korolev, Traffic safety evaluation in Northwestern Federal District using sentiment analysis of Internet users’ reviews, *Transportation Research Procedia, Volume 50,*2020.