Sentiment Analysis on Text Messages

Andrea Loizidou

Abstract

This paper delves into the application of Natural Language techniques in the realm of text messaging. It examines the emotional expressions conveyed through SMS across various countries by using sentiment analysis. By integrating Data Preprocessing and exploratory data analysis, this project aims to identify patterns and insights regarding how sentiments are expressed in text messages. The project also aims to categorize messages into positive, negative and neutral by using Naïve Bayes Classifier.

1 Introduction

In today's digital world, text messaging is the most widely used communication method, breaking through geographical and cultural barriers. Short Messages (SMS) contain valuable information and insights into emotional states and sentiments. This project aims to apply data preprocessing and exploratory data analysis techniques to explore various natural language processing techniques and uncover underlying patterns and information on emotional expression across multiple countries worldwide. More specifically, sentiment analysis will be used to analyze the sentiment of messages. The paper is divided into several sections. The Related Work section reviews existing literature on SMS text analysis and sentiment analysis. The objective section discusses the main goal of the projection. The data section where the data preprocessing steps, and description of the data are discussed. The methodology section where exploratory data analysis is discussed and the main NLP techniques are discussed, subsequent sections present the project approach the Naïve Bayes Classifier approach and conclude with the conclusion section.

2 Related Work

According to Qurat Tul Ain (Ain et al., 2017), sentiment analysis is a process that involves recognizing and extracting subjective information from textual data. It includes analyzing opinions, attitudes, emotions, and feelings expressed in written content and classifying them as positive, negative, or neutral sentences. I will use a general sentiment analysis framework for my project, represented by Figure 1. As shown in the figure by Karthick Prasad Gunasekaranx (Gunasekaran, 2023), the first step in sentiment analysis is to input the text data. The next step is preprocessing, which involves several steps such as tokenization, stop word filtering, and stemming. During text preprocessing, noise will be removed from the text by eliminating empty messages, punctuation marks, stop words, URLs, and HTML tags. Text normalization will then take place by converting it to lowercase. In recent years, a great deal of focus has been on exploring various aspects of sentiment analysis. This includes detecting subjective or objective sentences, classifying sentences as positive, negative, neutral, joyful, fearful, or angry, and applying sentiment analysis in industries such as commerce, disaster management, and health. Sentiment analysis has been used for customer reviews and headlines (Bellegarda, 2010), novels (Boucouvalas, 2002), emails and text messages (Liu, Lieberman, Selker, 2003; Mohammad Yang, 2011), and most notably, the analysis of tweets (Mohammad, 2012).

3 Objective

The project's goal is to perform a thorough analysis using natural language processing (NLP) techniques discussed in class. Once the data is preprocessed, the focus will be on sentiment analysis of the message variable in the dataset. With the use of sentiment Analysis, we will understand the emotions and opinions in text messages (SMS) and with the use of statistical analysis we will also compare the sentiment across multiple countries found in our dataset.

4 Data Preprocessing

Data Preprocessing is an important step in training machine learning models to extract meaningful insights from the dataset. A large amount of redundant, irrelevant, or noisy data present can pose a significant challenge in discovering valuable information for our project. In general, redundant, or irrelevant data can lead to overfitting, which reduces the model's generalization performance, and noisy data can introduce errors in the training process. Therefore, data preprocessing is necessary to clean and transform the dataset before feeding it into the ML model.

4.1 Data preprocessing steps

Below are some of the commonly used techniques in text preprocessing used in the analysis of text for this specific project:

4.1.1 Stop Word Removal

Stop words are common words in a language that do not provide any meaning to the text. For example, "the," "a," "an," "and," "or,".

4.1.2 Stemming and Lemmatization

Stemming and lemmatization are used to reduce inflection in words and reduce them to their root form. Stemming refers to removing the word's suffixes, while lemmatization refers to reducing the word to its base form. For example, the word "preprocessed" can be stemmed to "preprocess" and lemmatized to "process."

4.1.3 Tokenization

Tokenization refers to the brake down of text to smaller parts. For instance, breaking down a paragraph into sentences, phrases, or words. By tokenizing the text, we can analyze each unit of the text independently.

4.1.4 Erase Punctuations

Punctuation marks (periods, commas, exclamation marks, question marks), do not contribute to the sentiment analysis of the text. Therefore, removing them improves the accuracy of the sentiment analysis.

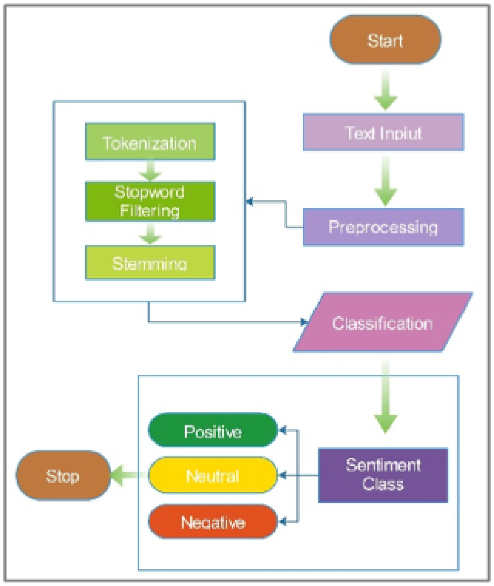

Figure 1: General Framework of SA process

4.2 Data Description

The data used is a corpus of SMS messages collected for research at the Department of Computer Science at the National University of Singapore. Collecting these messages was to analyze the language used in SMS messages and understand how people communicate. The corpus includes 67,093 SMS messages collected on March 9, 2015. The corpus is essential in analyzing patterns of SMS usage in Singapore and understanding how people use SMS to communicate. article booktabs

Table 1: Summary of Message Data Variables

Variable	Description
id	Unique identifier for each message.
Message	The message contents.
length	Total number of characters in the message.
country	Country the sender is from.
Date	Month and Year a message was sent.

5 Methodology

5.1 Exploratory Data Analysis

Exploratory data analysis of the prepared SMS can be performed either at the concept or category level, according to Chee Kian Leong (Leong et al., 2011). Exploratory data analysis refers to exploring data to find insight and identify patterns and trends. At the concept level, extracting the relevant statistics such as frequency and percentage of occurrence is important. Similarly, at the category level refers to identifying the multiple categories and the corresponding number of concepts. This is important because it understand the dataset better and identify which categories are most important. For example, in the case of the dataset used for this project, some data analysis at the category level includes figuring out the unique countries. This helps to identify which countries are most frequently mentioned in the dataset and can provide insights into the geographical distribution of the data. At the concept level, it is essential to determine the number of messages sent by each country, which helps identify the most active countries in the dataset.

6 NLP Techniques for Sentiment Analysis

6.1 Lexicon-based techniques

Lexicon-based sentiment analysis refers to using a pre-defined list of terms with corresponding sentiment scores to evaluate the sentiment of a text (Aung Myo, 2017). Sentiment score denotes the degree of emotion, such as a positive score for a positive statement and a negative score for a negative statement. Usually, the lexicon is created by selecting words that indicate either positive or negative sentiment and assigning scores to each word based on the intensity of the sentiment (Almatarneh Gamallo, 2018). All the scores are added together to calculate the sentiment score for a text (Ma, Cheng, Hsiao, 2018). While lexicon-based techniques are more straightforward to implement and require less training data than machine learning methods, they may not be as precise because they do not consider the context of the words or the relationships between words in a phrase.

6.2 Machine learning-based techniques:

Machine learning (ML) techniques have the ability to learn from large amounts of data without explicit training (Rathi et al., 2018). ML approaches are split into supervised learning, unsupervised learning, and deep learning. All three are discussed below.

6.2.1 Supervised Learning

Supervised learning is a type of machine learning where a model is trained using labeled data with a clear association between the input text and its corresponding sentiment label (Baid, Gupta, Chaplot, 2017). This model is trained by presenting the model with examples of text, each paired with a pre-assigned label. The model then learns to recognize the textual patterns and can later assign a sentiment label.

6.2.2 Unsupervised Learning

Unsupervised learning refers to training a model without labeled input. Unsupervised learning is helpful in cases where there is no pre-defined label for the input text, and the model must identify the underlying structure in the data by itself.

6.2.3 Deep Learning

Deep learning uses multi-layered neural networks to learn and extract complex patterns and correlations in data. This approach can learn more complex patterns and correlations in the data, leading to better performance than traditional machine learning approaches.

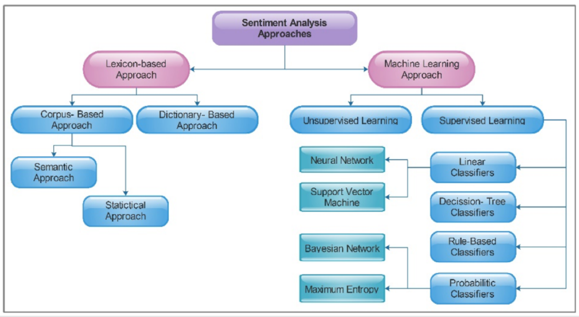

Figure 2: Sentiment Analysis Approaches

6.3 Rule - Based Technique:

The rule-based algorithm works by processing two sets of words: a positive and a negative set of words. These sets are predefined and contain words that are associated with positive and negative sentiments. The way this algorithm words is it scans through the text and looks for words defined in either set. The algorithm keeps track of the number of positive or negative words, and then if more positive words are relevant, then the sentiment is positive; if more negative words are prevalent, then the sentiment is negative. This approach is useful when labeled data is scarce or where existing pre-trained models are not as effective. One drawback that the rule-based approach has is that the results tend to be inadequate since they are not very flexible or precise. However, the rule-based approach is useful in certain contexts because it can identify the overall tone of the message. For instance, it can help customer support teams respond appropriately to feedback and complaints.

7 Comparison Of The Three NLP Techniques Discussed

According to Devika M D (Devika et al.), the Machine Learning Approach, which can be either supervised or unsupervised, has the advantage of not requiring a dictionary and is known for its high accuracy in classification. However, its disadvantage is that a classifier trained on texts from one domain often fails to work with other domains. The Rule Based Approach also employs both supervised and unsupervised learning. It boasts a high-performance accuracy, 91 per cent at the review level and 86 per cent at the sentence level, and performs better in sentence-level sentiment classification than at the word level. Its downside is that its efficiency and accuracy are contingent on the rules that are defined. Lastly, the Lexicon-based Approach utilizes unsupervised learning and has the benefit of not needing labeled data or a specific learning procedure. However, it relies on powerful linguistic resources, which may not always be available

8 Project Approach

After going through numerous NLP techniques that can be used to solve this problem, I decided to include the following NLP tasks: Firstly, I will preprocess the text using regex and NLTK. I will continue by doing exploratory data analysis with pandas and seaborns. Finally, I will use a bag of words and a Naïve Bayes Classifier to continue my sentiment analysis. I want to distinguish whether the data is positive or negative. Moreover, I am interested in the countries with the most positive/negative text messages.

9 Naïve Bayes Classifier:

The Naive Bayes classifier is a common probabilistic model used in NLP tasks. This model is used to categorize text into positive and negative sentiments. The classifier is based on the Bayes rule and assumes that words are conditionally independent. This assumption enables fast classification algorithms for the problem without sacrificing accuracy. For example, in the case of sentiment analysis in text messages, each word is viewed as independent. For instance words such as "happy," "joy," and "awesome" are typically associated with positive sentiments, while "hate," "madness," and "disappointing" are related to negative sentiments. The Naïve Bayes Classifier estimates the probability of the message belonging to each sentiment category based on the occurrence of these words in a text message and then assigns it to the category with the highest probability. Even though this approach is simple it can accurately distinguish between positive and negative sentiments in text messages.

10 Conclusion

To conclude, this project aims to demonstrate the efficacy of employing the Naïve Bayes Classifier for Sentiment Analysis in text messages. This project aims to contribute to the broad fields of natural language processing by providing insights into sentiment trends in text messages across the countries of the data set.

References

- Ain, Q. T., Ali, M., Riaz, A., Noureen, A., Kamran, M., Hayat, B., Rehman, A. (2017). Sentiment Analysis Using Deep Learning Techniques: A Review. International Journal of Advanced Computer Science and Applications (IJACSA), 8(6). https://dx.doi.org/10.14569/IJACSA.2017.080657

- Almatarneh, S. Gamallo, P.J.P. (2018). A lexicon-based method to search for extreme opinions. PLOS ONE, 13(5), e0197816.

- Aung, K.Z., Myo, N.N. (2017). Sentiment Analysis of Students' Comments Using a Lexicon-Based Approach. In Proceedings of the 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS). IEEE.

- Baid, P., Gupta, A., Chaplot, N.J.I.J.o.C.A. (2017). Sentiment analysis of movie reviews using machine learning techniques. International Journal of Computer Applications, 179(7), 45-49.

- Gunasekaran, K. P. (2023). Exploring Sentiment Analysis Techniques in Natural Language Processing: A Comprehensive Review. arXiv preprint arXiv:2305.14842. DOI: 10.48550/arXiv.2305.14842.

- Leong, C. K., Lee, Y. H., Mak, W. K. (2011). Mining sentiments in SMS texts for teaching evaluation. Expert Systems with Applications. https://doi.org/10.1016/j.eswa.2011.08.113.

- Ma, E., Cheng, M., Hsiao, A.J.I.J.o.C.H.M. (2018). Sentiment analysis–a review and agenda for future research in hospitality contexts. International Journal of Contemporary Hospitality Management, 30(11), 3287-3308.
- Qurat Tul Ain, Mubashir Ali, Amna Riaz, Amna Noureen, Muhammad Kamran, Babar Hayat, and A. Rehman. 2017. Sentiment Analysis Using Deep Learning Techniques: A Review. In International Journal of Advanced Computer Science and Applications (IJACSA), 8(6). DOI: 10.14569/IJACSA.2017.080657.
- Rathi, M., et al. (2018). Sentiment analysis of tweets using a machine learning approach. In Proceedings of the 2018 Eleventh International Conference on Contemporary Computing (IC3). IEEE.
- Wongkar, M., Angdresey, A. (2019). Sentiment Analysis Using Naive Bayes Algorithm Of The Data Crawler: Twitter. In Proceedings of the Fourth International Conference on Informatics and Computing (ICIC), Semarang, Indonesia (pp. 1-5).
- Devika, M. D., Sunitha, C., Ganesh, A. (Year). Sentiment Analysis: A Comparative Study on Different Approaches. In Proceedings of the Fourth International Conference on Recent Trends in Computer Science Engineering. Chennai, Tamil Nadu, India. https://doi.org/10.1016/j.procs.2016.05.124