SENTIMENT ANALYSIS OF AMAZON REVIEWS USING MACHINE LEARNING MODEL

OMAR MOHAMMED ALI ALBAAGARI

UNIVERSITI TEKNOLOGI MALAYSIA

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

Using natural language processing, this chapter provides an illustration of the general framework that is used in the process of doing research for the purpose of performing sentiment analysis on evaluations of Amazon office products. Everything from the preliminary investigation of the sentiment analysis to the assessment and comparison of the models is included in the research process. In this chapter, the data that were employed and the types of models that were applied will be identified and illustrated.

3.2 Research Framework

In order to accomplished the sentiment analysis in its entirety, the research was carried out in 4 distinct phases. The completion of each phase brought about the achievement of a crucial milestone. The following processes are provided in the order that they are shown: data collection, data preparation and exploratory data analysis, VADER sentiment analysis and Reberta analysis, and model assessment and comparison correspondingly.

There is an illustration of the process of the work in figure 3.1. There will be a discussion of each step in the sub section that follows.

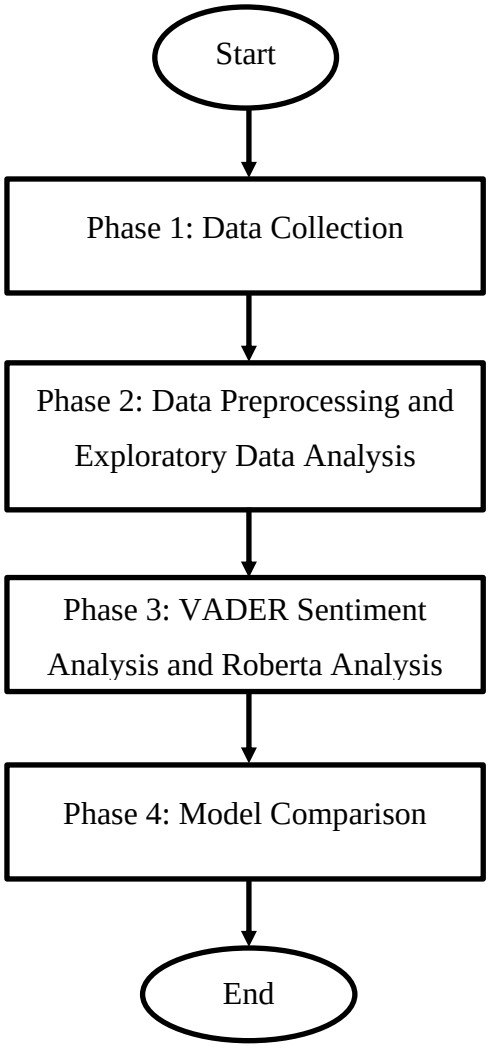

Figure 1 Overall research methodology

3.2.1 Data Collection

A dataset consisting of reviews of office products was gathered from the Amazon Reviews Repository collection. There are a total of 500,000 rows of data that represent each unique review, and there are eight columns that include the following information: parent asin, user id, helpful vote, asin, review, timestamp, verified buy, and rating. It offered a valuable perspective on the total pleasure of purchasers.

	parent_asin	user_id	helpful_vote	asin	text	timestamp	images	verified_purchase	title	rating
0	B01MZ3SD2X	AFKZENTNBQ7A7V7UXW5JJI6UGRYQ	0	B01AHHL4X2	Lovely ink. Writes well. The right amount of w...	1677939345945	[]	True	Pretty & I love it!	5.0
1	B08L6H23JZ	AFKZENTNBQ7A7V7UXW5JJI6UGRYQ	0	B08L6H23JZ	Overall I'm pretty happy with this purchase bc...	1677939160682	[]	True	2 excellent 1 extremely dry (blue)	4.0
2	B07JDZ5J46	AFKZENTNBQ7A7V7UXW5JJI6UGRYQ	2	B07JDZ5J46	[[VIDEOID:63276c19932aa4f3687042b8b9f8613c]] U...	1660188831933	[]	True	I don't get the reviews. Mine are garbage.	1.0
3	B07BR2PBJN	AFKZENTNBQ7A7V7UXW5JJI6UGRYQ	0	B004MNX7EW	It's a beautiful color, but even though it had...	1659806066713	[{'small_image_url': 'https://m.media-amazon.c...	True	Ordering Ink online: never a good idea I guess.	4.0
4	B097SFY5ZS	AFKZENTNBQ7A7V7UXW5JJI6UGRYQ	0	B019YLRFFS	Idk if I just got a bad batch which is possibl...	1659799390978	[]	True	Mine are iffy at best.	3.0

Figure 2 Buyer Review Dataset

3.2.2 Data Preprocessing and Exploratory data analysis

Due to the fact that that particular user-provided product review data is absent from the dataset that we are using, it is not possible to conduct an assessment using this sample. As a consequence of this, it is imperative that they be removed. Additional, in the event that the ratings have values that are missing. In light of this, it is able to either substitute NaN with the average of the other samples or get rid of those samples altogether. On account of the large number of samples that we have, it is possible to remove them.

When doing any kind of data analysis that is associated with text, the first step is to clean the text data. This is done in order to establish some straightforward methods that can be used to clean and prepare text data for modelling and machine learning.

- lowercase
- links starting with “http” or “https”, or “www” are replaced by “ ”.
- remove whitespaces
- remove HTML tags
- replace digit with spaces
- replace punctuations with spaces
- remove extra spaces and tabs

Additionally, First, the text should be tokenized into words or sub-word units, and then stemming should be used in order to standardize word forms. Next, after cleaning.

3.2.3 VADER Sentiment Analysis and Roberta Analysis

To begin, make use of VADER, which stands for Valence Aware Dictionary and sentiment Reasoner. This tool offers a speedy dictionary-based method to determine the polarity of sentiment and provide compound scores that range from -1 (the most negative) to +1 (the most positive). This technique is especially useful for analyzing shorter text samples, and it provides an easy threshold-based classification (for example, positive, negative, and neutral). However, in order to capture language and context with a greater degree of detail, I additionally fine-tune a Roberta model. Roberta, which is an improved transformer-based framework, not only has a deeper understanding of the links between words, but she also has a more precise ability to manage small alterations in mood. The combination of these two approaches provides me with a time-efficient, rule-based sentiment analysis alternative known as VADER, as well as a strong context-aware model known as Roberta, which has the potential to greatly increase classification results.

3.2.4 Model Evaluation and Comparison

After the implementation of natural language processing, the accuracy, precision, and recall of the model may be evaluated. The F1-score is a metric that compares the model's capacity to recognize all positive data to the harmonic mean of precision and recall. This is done to determine how effectively the model is able to categorize reviews as either positive, neutral, or negative. And last, choose the model that is the most correct.

3.3 Chapter Summary

The methodology of the research is broken out in great detail in this chapter, beginning with the gathering of data and continuing with the evaluation of the categorization model and comparison. This procedure guarantees that the procedure of doing sentiment analysis of the office product reviews is carried out in a methodical and data-driven manner.