Department of Computer Science and Engineering
University of Barishal

Sentiment Analysis

Submitted By

Group 1
Computer Science and Engineering

Submitted To

Dr. Abdullah Al Masud
Associate professor and Chairman
Department of Management Studies
University of Barisal

January 14, 2024

Table of Contents

Chapter 1

Methodology

Opinions expressed on social networks wield significant influence over public behaviour, spanning domains such as consumer product choices, gauging stock market trends, and impacting political preferences (Bai, 2011; Eirinaki, Pisal, Singh, 2012). The opinions generated on the web, particularly in blogs and social networks, have emerged as valuable data for extracting user sentiments. This information proves instrumental in applications like customer relationship management, monitoring public opinion, and refining text filtering mechanisms (Zhang, Zeng, Li, Wang, Zuo, 2009). Recent analyses of online opinions increasingly employ sentiment analysis (SA) techniques to gain insights into the prevailing sentiments and attitudes within these digital spaces. In fact knowledge obtained such as Apple products through many e-commerce websites have been shown to be extremely valuable to marketing research companies, public opinion organisations and other text mining entities. Apple products enjoy popularity based on a spectrum of customer reviews. Positively, users often highlight the sleek design, premium build quality, and the seamless integration of hardware and software across the Apple ecosystem. The intuitive user interface, coupled with a smooth and responsive performance, receives consistent praise. Additionally, the App Store's extensive collection of high-quality apps and Apple's commitment to regular software updates contribute to positive customer experiences which is found in many websites. On the negative side, some customers criticise the high price tags associated with Apple products, expressing concerns about affordability and value for money. The closed ecosystem and limited customization options are also points of contention for certain users who prefer more flexibility in their devices.Some customers may express frustration with Apple's proprietary connectors and the lack of compatibility with non-Apple accessories. In neutral reviews, users may appreciate the standardisation and ease of use, acknowledging the consistent quality across Apple's product lineup. In essence, the popularity of Apple products stems from a mix of positive design and functionality, negative cost considerations for some, and neutral sentiments about standardisation and reliability.

1.1 Proposed Approach

As reviews play a vital role in marketing of a particular product ,we planned a model through the opinions of customers about Apple products. Here is the graphical representation of the proposed model. In this study, section 1.2 describes the lexicon, 1.3 describes the full data collection system, 1.4 describes the classification techniques, 1.5 section describes the evaluation model parameters, 1.6 describes implementation, and result and discussion are described in section 2 and 3 respectively.

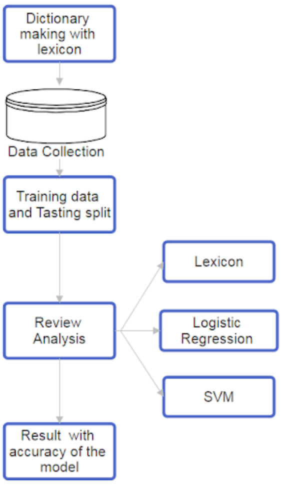

Dictionary making with lexicon
Data Collection
Training data and Tasting split
Review Analysis
Lexicon
Logistic Regression
SVM
Result with accuracy of the model

Fig 1: Proposed approach

1.2 Lexicon

Classifying words for sentiment analysis is one of the major steps of this study. There are two widely used methods for sentiment orientation identification: the lexicon-based approach and the corpus-based method [1]. In this study, we used a lexicon-based method for categorising words. A lexicon is a list of words used in a particular language, or a dictionary. The sentiment is determined by comparing words against the expert-defined dictionary entry making it easy to determine the polarity of a specific sentence; positive, negative, and neutral. In this study, the lexicon contains 8222 entries. The positive words are designated by 1, negative by -1, and neutral by 0 in the lexicon. The lexicon is defined by Stepthom (subjectivity$_{c}lues_{h}ltemnlp05/lexicon_{e}asy.csv$).

1.3 Data Collection

The data collection process for sentiment analysis in our study involves obtaining opinions from various sources, mainly blogs, open sources and socia networks. This information is crucial for understanding user sentiments about Apple products. We have collected

5540 data (Text, Account Id and Publishing date). Blogs and social networks are primary sources. We have collected mixed data from public opinion, most of them under age of 40. Because they are main customer of apple products.

1.4 Classification Techniques

In this paper, the focus revolves around leveraging social networks' text mining techniques to analyze consumer brand sentiments. In this paper uses two prominent machine learning algorithms, namely Logistic Regression and Support Vector Machines (SVM), in the analysis process. The Logistic Regression algorithm is utilized to model the relationship between textual features extracted from social media content and the sentiment expressed towards consumer brands.Using this algorithm from our Dataset we get the Accuracy: 82Additionally, Support Vector Machines are employed to enhance the sentiment analysis process. SVMs excel in handling high-dimensional data and are proficient in capturing complex relationships between features. By leveraging the power of SVMs, the study aims to improve the accuracy and robustness of sentiment classification, providing a more nuanced understanding of consumer perceptions. Using this algorithm from our Dataset we get the Accuracy: 82The integration of Logistic Regression and SVM algorithms in this research is strategic, as it combines the simplicity and interpretability of Logistic Regression with the sophisticated pattern recognition capabilities of SVMs. Through this dual approach, the paper aims to contribute valuable insights into the complex landscape of consumer brand sentiments on social networks. The comparative analysis of these two algorithms sheds light on their respective strengths and limitations in the context of social media text mining, offering a comprehensive view of their performance and applicability in understanding and interpreting consumer sentiments towards brands.

1.5 Evaluation model parameters

The evaluation of the sentiment analysis model proposed in the study involves assessing specific parameters:

- **Accuracy:** Both Logistic Regression and Support Vector Machines (SVM) achieved an accuracy of 82%, indicating their effectiveness in correctly classifying sentiments.

- **Precision:** Both algorithms show a precision of 81%, demonstrating their ability to accurately identify positive sentiments when they occur.

- **Recall:** Both algorithms exhibit a recall of 82%, indicating their capability to capture a substantial portion of positive sentiments in the dataset.

1.5.1 Sentiment Distribution

- **For Lexicon Process:**
  - **Positive: 72.2%**
  - **Negative: 7.0%**
  - **Neutral: 20.8%**
- **For LR and SVM:**
  - **Positive: 73.47%**
  - **Negative: 6.77%**
  - **Neutral: 19.77%**

The dataset has a distribution of sentiments with the majority being positive.

1.5.2 Algorithm Efficiency

SVM is considered more efficient than Logistic Regression, especially in handling high-dimensional data and capturing complex relationships between features. But we can conclude that Lexicon-based-approach is much more efficient than using Machine Learning.

1.5.3 Dual Algorithm Approach

The integration of both Logistic Regression and SVM aims to provide a comprehensive understanding of consumer brand sentiments on social networks, leveraging the strengths of both algorithms.

1.6 Implementation

1.6.1 Data Loading and Preprocessing

The 'apple$_{p}roduct_{r}eview.csv'datasetwasloadedusingthe`csv`module.Textualdata, includingcomments, IDs$
$ASCIIcharacters, andadditionalsteps, suchasconvertingtexttolowercaseandeliminatinghashtagsandre-$
$comments, wereimplemented.$

1.6.2 Sentiment Scoring

A sentiment lexicon from 'lexicon$_easy.csv'wasemployedtoassignsentimentscorestowordswithinthecommen$

1.6.3 Feature Extraction

The TF-IDF vectorizer from the ‘sklearn‘ library was utilized to convert textual data into numerical vectors, facilitating the machine learning models’ analysis. Subsequently, the dataset was split into training and testing sets using the ‘train$_test_split`function.$

1.6.4 Lexicon Process

The sentiment scoring is used here for checking each word in a comment against the lexicon dictionary, and its corresponding sentiment score is accumulated. The comment is then classified as positive, negative, or neutral based on the overall sentiment score. This lexicon-based approach provides a simple method to identify the sentiment in text by leveraging predefined scores for individual words.

1.6.5 Logistic Regression Model

A Logistic Regression model from ‘sklearn‘ was employed for sentiment analysis. The TF-IDF vectors of the training set were used to train the model, followed by predictions on the test set. Performance metrics, including accuracy, precision, and recall, were then evaluated. Visualizations, such as a horizontal histogram and a pie chart using ‘matplotlib‘, were generated to provide a comprehensive overview of the Logistic Regression model’s effectiveness and sentiment distribution.

1.6.6 Support Vector Machine (SVM) Model

Simultaneously, a Support Vector Machine (SVM) model using the ‘SVC‘ class from ‘sklearn.svm‘ was implemented. Similar preprocessing and feature extraction steps were applied to train and evaluate the SVM model. Performance metrics, including accuracy, precision, and recall, were visualized using a horizontal histogram. A pie chart illustrated the distribution of sentiments in the test set as predicted by the SVM model.

1.6.7 Comparative Analysis

A comprehensive comparative analysis was conducted, assessing the strengths and weaknesses of both the Logistic Regression and SVM models. Performance metrics, visualizations, and qualitative insights were collectively utilized to evaluate their effectiveness in sentiment analysis.

1.6.8 Top Reviews Identification

Based on predictions and associated probabilities from both models, the top positive, negative, and neutral reviews were identified. This qualitative assessment contributed to a nuanced understanding of the models' decision-making processes.

1.6.9 Google Colab Integration

The entire implementation was carried out in Google Colab, providing a collaborative and accessible environment for Python coding. Specific configurations or considerations for running the code in this environment were addressed.

Chapter 2

Results

This section presents the testing results of a Lexicon analysis and two machine learning model Regression (LR) and Support Vector Machines (SVM). The results of the sentiment analysis on reviews about Apple products are summarised in table 1 and table 2.

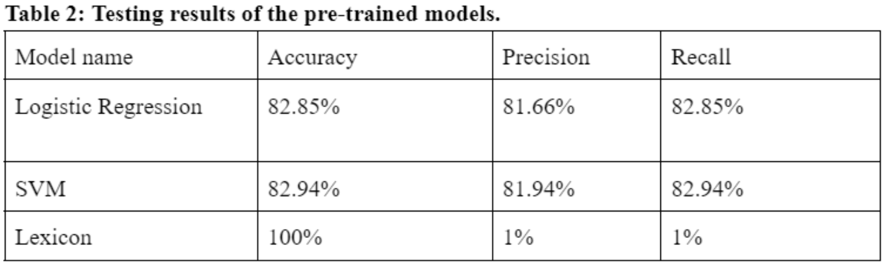


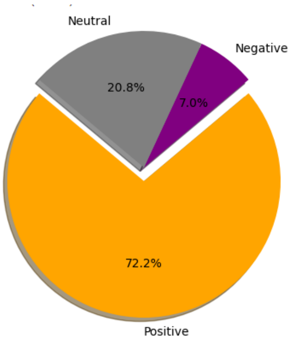
Table 2: Testing results of the pre-trained models.

Model name	Accuracy	Precision	Recall
Logistic Regression	82.85%	81.66%	82.85%
SVM	82.94%	81.94%	82.94%
Lexicon	100%	1%	1%

Fig 2: Table representing positive, negative, and neutral parentage

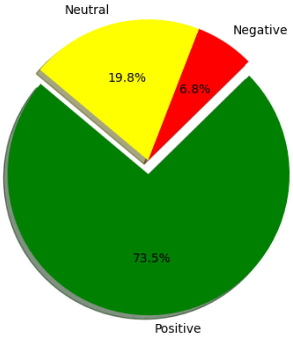
In these experiments, two pretrained machine learning models, namely, LR and SVM were implemented to distinguish sentiments into positive, negative, and neutral categories. The Sentiment Percentages indicate the distribution of sentiments in the dataset. The majority of opinions are positive, with a smaller percentage being negative and a moderate percentage being neutral.

The sentiment percentages are similar for both LR and SVM, suggesting that both models are capturing sentiments in a comparable manner.The lexicon-based method yields slightly different sentiment percentages compared to LR and SVM. It indicates a positive sentiment of 72.2%, a negative sentiment of 7.0%, and a neutral sentiment of 20.8%. Accuracy, precision, and recall values are closely aligned between the two models, indicating analogous sentiment classification capabilities. Both LR and SVM demonstrate high accuracy, precision, and recall, indicating their effectiveness in discerning sentiments in the context of Apple product opinions. The Lexicon-based method achieves a perfect accuracy of 100%. This implies that all predictions made by the lexicon-based approach align with the actual sentiments in the dataset.



.5

Figure 2.1: Fig 3: Laxicon sentiment Percentages chart



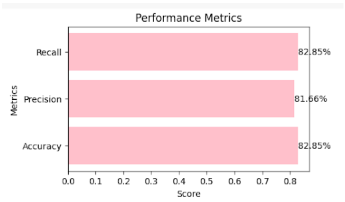
.5

Figure 2.2: Fig 4: LR SVM sentiment Percentages chart

Figure 2.3: Review Sentiment Percentages

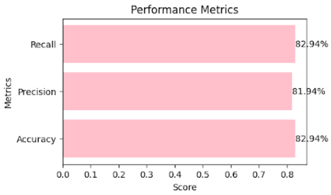
Fig 2: Table representing positive, negative, and neutral parentage

The model appears to effectively capture and distinguish between positive and negative sentiments in the provided reviews, as evidenced by the high probabilities associated with the correct sentiment labels. Table 3 provides sentiment probabilities and associated reviews that showcase the outcomes of a sentiment analysis task. Some reviews all receive very high positive sentiment probabilities, indicating a strong likelihood that they express positive sentiments. The language used in these reviews is generally positive and appreciative. There are instances where the sentiment probabilities for negative reviews may not align with the actual sentiment expressed in the text. Some reviews all receive very high negative sentiment probabilities, indicating a strong likelihood that they express negative sentiments.



.5

Figure 2.4: Fig 3: Performance Metrics for LR



.5

Figure 2.5: Fig 4: Performance Metrics for SVM

Figure 2.6: Performance Metrics

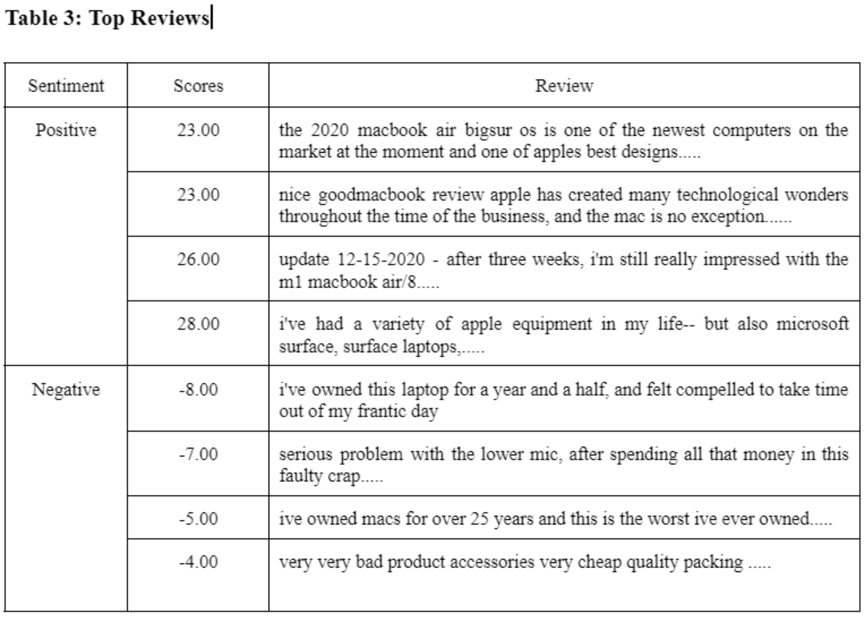


Table 3: Top Reviews

Sentiment	Scores	Review
Positive	23.00	the 2020 macbook air bigsur os is one of the newest computers on the market at the moment and one of apples best designs.....
	23.00	nice goodmacbook review apple has created many technological wonders throughout the time of the business, and the mac is no exception......
	26.00	update 12-15-2020 - after three weeks, i'm still really impressed with the m1 macbook air/8.....
	28.00	i've had a variety of apple equipment in my life-- but also microsoft surface, surface laptops,.....
Negative	-8.00	i've owned this laptop for a year and a half, and felt compelled to take time out of my frantic day
	-7.00	serious problem with the lower mic, after spending all that money in this faulty crap.....
	-5.00	ive owned macs for over 25 years and this is the worst ive ever owned.....
	-4.00	very very bad product accessories very cheap quality packing .....

Chapter 3

Discussion

The paper is about the implementation of sentiment analysis of opinions about Apple products extracted from blogs, social networks, and other online sources. The research aims to discern sentiments into positive, negative, and neutral categories by leveraging machine learning algorithms such as Logistic Regression(LR) and Support Vector Machines (SVM). The proposed approach involves a lexicon-based method for sentiment orientation identification, using a dictionary containing 8222 entries. Data collection encompasses 5540 opinions from various sources, primarily targeting individuals under the age of 40, as the main consumers of Apple products. The comparative analysis between Logistic Regression and SVM reveals similar results in terms of sentiment percentages and model evaluation metrics, demonstrating their efficiency in understanding consumer brand sentiments on social networks. But SVM is more efficient than LR in this case. The implementation involves data loading, preprocessing, sentiment scoring, feature extraction, and the application of machine learning models, with Google Colab serving as the coding environment. Here in this paper we use only two machine learning algorithms (Logistic Regression) and Support Vector Machine(SVM) for the analysis. The accuracy and the efficiency of the analysis can be made better by using deep learning based algorithms and also by another machine learning algorithms sush as LDA, CART etc.

References

[1]. Miao, Q., Li, Q., Zeng, D. (2010). Fine-grained opinion mining by integrating multiple review sources. Journal of the American Society for Information Science and Technology, 61, 2288–2299