Deep Learning Approach for Multiclass Sentiment Analysis of Customer Reviews

Gauri Chaudhari Sakshi Rathi Shreya Gajbhiye Shruti Houji
Luddy School of Informatics, Computing and Engineering
Indiana University Bloomington
{gchaudh,sakrathi,sygajbhi,sghouji}@iu.edu

Abstract

This study aims to gain an in-depth understanding of customer perspectives by analyzing their product reviews.To achieve this objective, the research utilizes digital platforms like Amazon.The paper addresses several research questions, including the effective integration of textual and visual information for sentiment analysis and tackling label imbalance in multiclass sentiment analysis. The paper involves several data preprocessing techniques like correcting spelling errors and processing emojis to extract valuable insights. By analyzing more than 82,816 Amazon reviews across three categories, the study demonstrates the efficacy of the methodology, with the Bi-LSTM deep learning model achieving an 88% accuracy rate, followed by Bi-RNN with 82%. Overall, this study contributes to the e-commerce industry by providing insights into customer sentiment.

1 Introduction

The e-commerce industry has grown significantly over the years with the increase in the number of online shoppers. For the businesses to grow, analyzing the customers' reviews and reactions is important for any business to grow.

Our paper, considers the mobile reviews real world data scraped from Amazon where a subsample of a huge dataset UCSD that is available on Kaggle was used.

The paper focuses on the challenges of preprocessing and representing multimodal data for sentiment analysis. We focus on handling the visual cues of reviews like emojis.Pre-processing techniques like emoji to text conversion and emoji mapping using dictionary. To prepare our data to train our models, the textual and visual data are processed and tokenized. The paper also discusses appropriate strategies for dealing with spelling mistakes, commonly used slangs, abbreviations, repeating characters, uppercase words, etc.

Sentiment analysis for customer reviews using neural network models like Recurrent Neural network and Long Short-Term Memory networks(LSTM) is implemented in this paper instead of using traditional machine learning models like Naive Bayes, KNN, Decision Tree, Random Forest, etc.

In this research, upsampling technique is used to balance the distribution of the three sentiment classes - positive, negative, and neutral. This results in a total of 42,867 samples across all sentiment classes, and the model's performance is evaluated based on unbalanced dataset and compared with the balanced dataset.

Overall, this paper covers the research questions like

- How can textual and visual information be integrated effectively for sentiment analysis?
- What is the most effective way to pre-process and represent multimodal data for sentiment analysis? What are the best methods for dealing with ambiguity and noise in visual cues such as emojis?
- How can label imbalance be addressed in multiclass sentiment analysis?

2 Literature Review

(Fang and Zhan, 2015) proposed the scope of the text have three levels of sentiment polarity categorization and entity aspect level then targets on what exactly people like or dislike from their opinions.(Shivaprasad and Shetty, 2017) provided a thorough literature review of sentiment analysis techniques for product reviews, it covers three levels of sentiment analysis, polarity-based classification, and various classification methods.(AlQahtani, 2021) analyzed the importance of data preprocessing to remove noise, such as stop words and

punctuation marks, and to perform feature selection to reduce the dimensionality of the data.

Supervised and unsupervised methods were analyzed by (Nguyen et al., 2018) using two approaches: machine learning and lexicon based with data pre-processed using NLP techniques such as sentence extraction, normalization,tokenization, and stop-word removal. Hence, to further improve the accuracy of the supervised models (Alrehili and Albalawi, 2019) proposed voting ensemble technique like combining Bagging, Naive Bayes, SVMs, Random Forest, and Boosting classifiers. But after a certain limit, traditional machine learning models doesn't solve the complex problems like sarcasm reviews, fake reviews, context based emoji processing. Hence, deep learning models are preferred over ML models.

(Abah, 2021) used deep learning models, CNN, LSTM, to classify sentiment polarity in Amazon Electronic review dataset. GloVe Embeddings are used to process raw text into a word vector representation.

3 Sentiment Analysis on product reviews

Sentiment analysis of product reviews is an important task in natural language processing and has gained significant attention in recent years. The goal of this task is to automatically identify the sentiment expressed in a product review and classify it into one of several categories, such as positive, negative, or neutral. Deep learning approaches have been widely used for sentiment analysis of product reviews due to their ability to automatically extract meaningful features from the input text. These approaches typically involve training a deep neural network on a large dataset of labeled reviews to learn a mapping between the review text and its associated sentiment category. The resulting model can then be used to predict the sentiment of new, unseen reviews. By applying deep learning to sentiment analysis of product reviews, we can improve the accuracy and efficiency of this important task, enabling businesses to gain valuable insights into customer opinions and preferences.

4 Methodology

In this section, we present an outline of our proposed approach for conducting sentiment analysis on mobile phone reviews from Amazon. The various stages of our methodology are illustrated in Figure.1, which includes the process of collecting data and culminates in the evaluation of each classification model.

4.1 Data Selection

The data used for the research experiments is sourced from Amazon obtained from Kaggle.[1]. The dataset consists of 82,816 reviews of unlocked mobile phones and 8 features over a period of 2003-2020. The features and its discerptions are shown in table 1. Following are the features of the dataset:

Features	Description
ASIN	Product unique ID
Reviewer Name	Name of the reviewer
Rating	Reviewer rating [scale of 1-5]
Date	review date
Verified	valid customer
Title	review title
Body	review content
HelpfulVotes	helpful feedback count

Table 1: Features of dataset

4.2 Data Pre-Processing

In the dataset pre-processing, we analyzed all attributes founded in data set in order to see if these attributes useful for our goal or not. Hence, we omitted 5 attributes which are ASIN, Reviewer Name, Date, Verified, HelpfulVotes.

There are five distinct class of Ratings in the dataset ranging from 1-5. We have considered that reviews with ratings from 1-2 would have an negative sentiment, 3 would be neutral and 4-5 would be a positive sentiment. Therefore, a new attribute is introduced called sentiment based on this rating distribution.

The dataset is further scrutinized. Null Ratings are removed, this is because null ratings do not have any valuable information and can not be transformed into the sentiment label. Any blank fields in title, body are replaced with an empty string.

4.2.1 Removing Spanish Reviews

The dataset contains 3 reviews in English and Spanish. Sometimes the title is in English but the reviews are in Spanish and vice-versa. We removed all rows that both contain title and reviews in Spanish since code-switching is not in the scope.

[1]dataset:https://www.kaggle.com/datasets/grikomsn/amazon-cell-phones-reviews/versions/1?select=20190928-reviews.csv

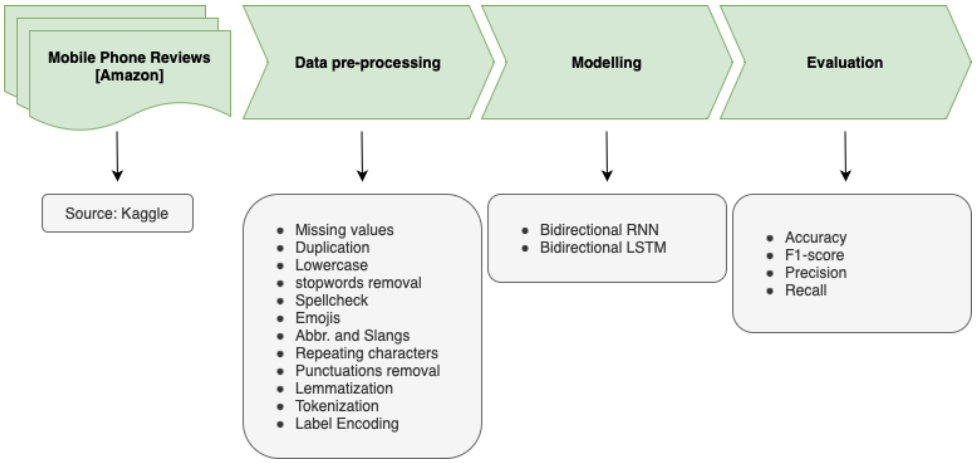

Figure 1: Overall methodology of Sentiment Analysis for Amazon mobile reviews

4.2.2 Exploratory Data Analysis of Reviews

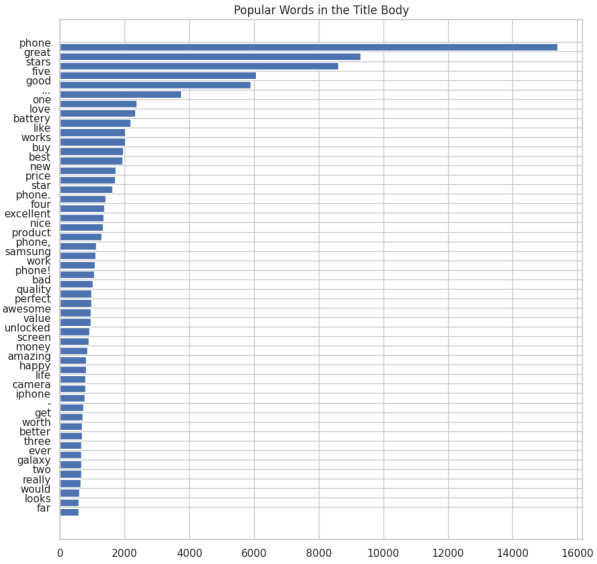

Figure 2: Most popular words in title

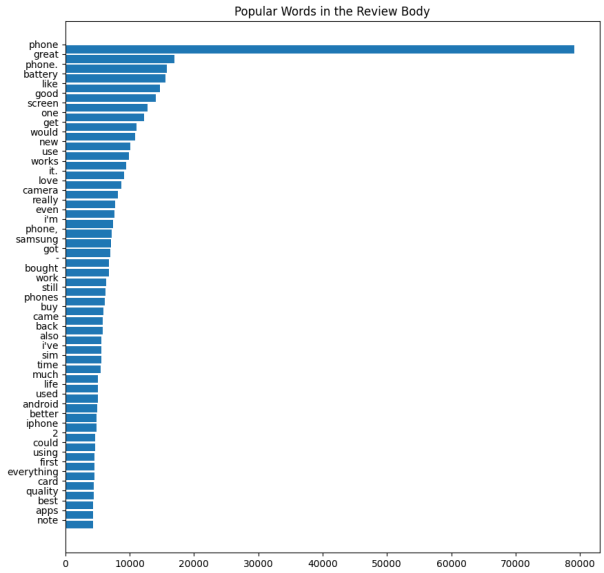

Figure 3: Most popular words in review content

From figure 2, we can see that before preprocessing title, there are stop words and punctuation that also form a major chunk of the reviews.

From figure 3, we can see that before preprocessing the content, there are stop words,punctuation and digit that also form a major chunk of the reviews.

4.2.3 Reviews Cleaning

This part of model focuses on text cleaning. At this stage, all the information that is not required is removed or replaced or transformed from the data.

- **Transform to lowercase:** Text standardization is a common practice in natural language processing that reduces the complexity of datasets by converting all text to lowercase. However, there are situations where capitalization plays an important role in conveying sentiment or other aspects of the text. For example, fully capitalized words may carry emotional significance in a review. To preserve this important information, it may be necessary to retain the original capitalization of certain words. By doing so, we can ensure that the sentiment analysis accurately reflects the original meaning of the text. Therefore, while we have transformed all text to lowercase, we have also made sure to keep words that are completely capitalized.

- **Delete repetitive characters:** We have transformed words that have repetitive characters

like 'Amaaaazing' by only keeping one copy of the repeated character using regex.

- **Deleting digits:** We excluded digits from the text because they did not provide any key information for the task of sentiment analysis.
- **Deleting Punctuations:** Punctuation contains symbols including full stops, commas, question marks, exclamation marks, semicolons, colons, ellipses, and brackets. Using string.punctuation, we eliminated punctuations from the text.
- **Abbreviations and Slangs** This phase consists of correcting any internet-related terminology or acronyms. We use preset dictionaries and incorporate them to translate slang or abbreviations to their real versions. For example,'OMG' is for Öh my goodnessör Öh my God.
- **Removing Stopwords** The words that occur in English language most commonly such as the, a, an, and in. As these words are not going to provide useful information for sentiment analysis therefore, we have removed them.
- **Spelling mistakes** Dealing with spelling mistakes can be quite beneficial. Because users often make spelling errors, it might result in many word attributes belonging to the same root form. For example, various users may misspell the term excellent in different ways, resulting in separate word attributes that must be evaluated, using extra time.
- **Text Normalization** The technique of reducing a token to its basic shape is referred to as lemmatization. WordNetLemmatizer() was used to do lemmatization. We picked lemmatization because it produces better results than stemming but takes much longer. We had to choose between quality and time, and we picked quality by utilizing lemmatization. Even though we are trying to reduce the time complexity for sentiment analysis, the impact of using lemmatization is worth the extra time for our case.

4.2.4 Handling Emojis

There were three different types of emojis in the dataset.

- Text based emojis like :) and :(
- Modern emojis like 👍 👎 ☺ 🙄
- Monochrome emojis like ☺ ☹

For experiments that tested with preprocessing emojis, they were replaced with a description of the visual cue of the emoji using the emoji package. For Text based emojis, a dictionary was created and description was assigned to the most common forms of it.

4.3 Class Imbalance

After transforming the ratings into its respective sentiments, we found that the classified reviews have a class imbalance with more positive reviews and very less neutral and negative reviews. This can cause a potential bias while training the model towards positive reviews.

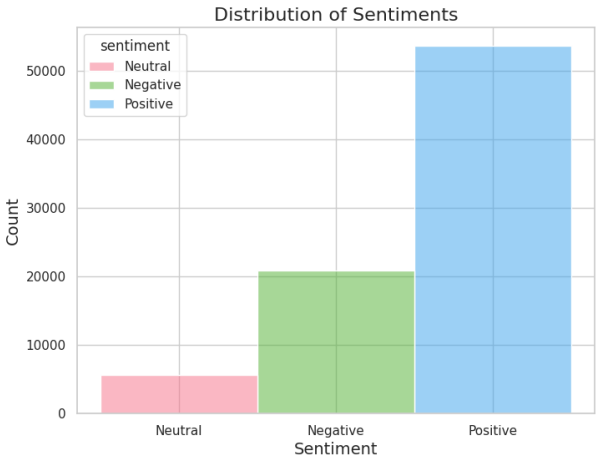

Figure 4: Sentiment Class Imbalance in Amazon Mobile reviews

4.4 Data Models

Two Deep Learning algorithms are applied on the reviews. The architecture of the algorithms, batch size, number of epochs remain the same across all four experiments to allow comparison of the two models. The model architecture are described below.

4.4.1 Bidirectional Simple RNN

It was found that the best Bidirectional Simple RNN model had the following features

- The model is a sequential model with four layers.
- The first layer is an embedding layer that takes the input sequence of length 100 and converts each word into a dense vector of fixed size.

- The second layer is a bidirectional simple RNN layer with 128 units. This layer processes the input sequence in both forward and backward directions and outputs a single vector of size 128 for each input sequence.
- The third layer is a dropout layer that randomly sets a fraction of the input units (0.5) to zero during training, which helps prevent overfitting.
- The fourth layer is a dense layer with three units, which applies the softmax activation function to the input and produces a probability distribution over the three classes.
- The total number of trainable parameters in this model is 1,021,507.

4.4.2 Bidirectional LSTM

It was found that the best Bidirectional LSTM model had the following features

- It is a sequential model with each layer added sequentially.
- The first layer is an embedding layer that converts each word in the input sequence of length 100 into a dense vector of fixed size.
- The second layer is a bidirectional LSTM layer with 128 units that takes the output of the embedding layer as input. This layer has 128 units, which means that the output of each forward and backward LSTM layer is of size 64, and the final output of this layer is a concatenation of the forward and backward LSTM outputs.
- The third layer is a dropout layer that randomly drops out 0.5 of the units in the bidirectional LSTM layer to prevent overfitting.
- The final layer is a dense layer with 3 units, corresponding to the number of output classes in the sentiment analysis task.
- The activation function used in the final layer is softmax, which outputs a probability distribution over the three classes.
- The total number of trainable parameters in this model is 1,084,867.
- The model takes an input sequence of length 100 and outputs a probability distribution over three classes.

4.5 Experiments and Results

Four kind of experiments were conducted with both the models:

1. Sentiment Analysis on only title

Models	Accuracy	
	Without emoji	With emoji
Bi-RNN	81.56	82.04
Bi-LSTM	83.11	84.95

Table 2: Accuracy of sentiment analysis on only title of reviews

2. Sentiment Analysis on only review content

Models	Accuracy	
	Without emoji	With emoji
Bi-RNN	81.01	82.88
Bi-LSTM	83.44	86.94

Table 3: Accuracy of sentiment analysis on only content of reviews

As we can see from Table 2 and Table 3. Bi-LSTM seems to perform slightly better than Bi-RNN overall for both on title and on actual reviews. However as we can see from Figure 4 that there is a class imbalance in data. We can see the impact of the same on the F1-score of the Bi-LSTM model that was computed with emoji.

	Precision	Recall	F1-score
Negative	0.7	0.75	0.72
Neutral	0.25	0.2	0.22
Positive	0.92	0.92	0.92
Accuracy			0.8694

Table 4: Metrics of Bi-LSTM model [With Emoji]

3. Impact of upsampling on the overall performance of Bi-LSTM

After upsampling the data, we obsserved a slight improvement in the accuracy and F1-scores also seem to improve. This can be furthur improved by fine tuning the hyper-parameters of the data model to be more effective.

4. Impact of spelling correction data preprocessing step on Sentiment Analysis

	Precision	Recall	F1-score
Negative	0.89	0.89	0.89
Neutral	0.60	0.60	0.60
Positive	0.97	0.96	0.96
accuracy			0.8853

Table 5: Upsampling of Bi-LSTM model

After balancing the data, we were curious about how effective the step for spelling correction is and if the model can build through the incorrect spellings or not. We ran the above upsampled model again but this time did not pass it through the spelling correction function. After conducting this experiment, we found that performing this step did slightly worsen the accuracy of the model which meant that incorrect spellings do have a little impact on the performance of the overall model.

With Spelling Accuracy	Without Spelling Accuracy
88.53	85.76

Table 6: Comparison of upsampled Bi-LSTM with and without spelling correction

5 Discussions

Data pre-processing play an important role as textual information is made more meaningful by removing duplicates, stopwords, repeated characters, punctuations, digits, and replacing abbreviations and slangs into its meaningful words.

Emojis can be handled using emojis library and hence the results, it is observed that for any model Bidirectional RNN or LSTM, the accuracy without processing emojis is less. After emojis are converted into their corresponding text, the accuracy of the model increases by about 1 to 2 percent. To further improve the accuracy, autocorrect spelling checker library in python is used to correct the mispelled words.

Hyperparameter tuning using grid search plays a crutial role to increase the efficiency of the algorithm. Bidirectional LSTM gives the best accuracy for our data with the best parameters as max_len100, max_words = 10000, embedding_dim = 100, lstm_units = 64, dropout_rate = 0.5, batch_size= 32, epochs = 10 with accuracy= 88.5%. We have calculated F1-score as 96% for positive sentiments, 60% for neutral sentiments and 89% for negative sentiments, as our dataset is imbalanced.

Label imbalance can be addressed in multiclass sentiment analysis by upsampling the dataset and hence better accuracy can be obtained.

6 Limitations

For the title column, pre-processing is done without removing the abbreviations and slangs because for most of the rows, NULL values were obtained. We have not implemented code switching where the reviews consist of a mixture of two languages, English and Spanish. Also, due to the upsampling of the dataset, a high false negative rate (type 2 error) is observed. A smaller number of resources are available to increase the number of epochs above 10 to increase the accuracy of the model. Manual dictionary mapping for emojis is one of the greatest limitations because it is difficult to manage manual intervention for a large dataset. Context-based emoji pre-processing, sarcastic emotions using emojis and text, and detecting fake reviews are other limitations that need to be considered for training. Our model does not work for multi-domain reviews as it is trained for one specific domain.

7 Conclusion and Future Scope

We have applied sentiment classification on the Amazon reviews for electronic mobiles dataset. Several pre-processing techniques like spelling correction, emoji to text conversion, along with the usual techniques like, lower case, removing punctuations, digits, etc, were applied to the text. Out of all the experimentations, that were carried out, the best model was given when emoji to text and spelling corrections, were applied along with the regular pre-processing techniques, and trained using the Bi-LSTM model. The test accuracy achieved was 88% and F1 score was lowest for neutral sentiment , since our data is imbalanced.

The dataset has two text columns body and title, currently we use the text of the body. We can further, use a weighted sentiment approach. By combining the sentiment of the body and the text. This could can lead to a better prediction model. Also, having a negative and a positive connotation in the same sentence, can lead to some sort of ambiguity while prediction. Though, we have dealt with spelling errors, using the NLTK spelling correction function, there are a lot of typos that havent been

dealt with yet. Dealing with variations of typos and some unexplored slang is something that can be explored in the future. Also, converting the text and the emojis both to the vectorized form and finding the accuracy may be worked on in the future. A better sampling technique should be used instead of up-sampling to avoid high type II error. Lastly, since we are dealing with only English language here, code switching would be a part of future exploration.

References

Jemimah Ojima Abah. 2021. *Sentiment Analysis of Amazon Electronic Product Reviews using Deep Learning*. Ph.D. thesis, Dublin Business School.

Arwa SM AlQahtani. 2021. Product sentiment analysis for amazon reviews. *International Journal of Computer Science & Information Technology (IJCSIT) Vol*, 13.

Ahlam Alrehili and Kholood Albalawi. 2019. Sentiment analysis of customer reviews using ensemble method. In *2019 International conference on computer and information sciences (ICCIS)*, pages 1–6. IEEE.

Xing Fang and Justin Zhan. 2015. Sentiment analysis using product review data. *Journal of Big Data*, 2(1):1–14.

Heidi Nguyen, Aravind Veluchamy, Mamadou Diop, and Rashed Iqbal. 2018. Comparative study of sentiment analysis with product reviews using machine learning and lexicon-based approaches. *SMU Data Science Review*, 1(4):7.

TK Shivaprasad and Jyothi Shetty. 2017. Sentiment analysis of product reviews: a review. In *2017 International conference on inventive communication and computational technologies (ICICCT)*, pages 298–301. IEEE.