Amazon Cell Phone Reviews Sentiment Analysis & Topic Extraction

Neil Sharma
Ismael Josue Tapia Tamayo
Preetham Kamath

I. Abstract

Semantic analysis is a type of natural language processing to extract and quantify valuable information from texts. In this study, a sentiment analysis is developed to assess customer reviews of cell phone products. While most approaches focus on building sentiment lexicon and training models, this paper adds an analysis of topic extraction, using Latent Dirichlet Allocation (LDA). The sentiment lexicon is built based on word2vec algorithms to train binarized models, which include K Nearest Neighbor, BERT, and Neural Networks. The results show that the BERT algorithm provides the highest accuracy for this case study.

II. Introduction

The advent of website reviews has allowed a large number of customers to express their experience or feelings towards the product they purchased. Traditionally, companies invested efforts in getting customer feedback for a specific product through surveys. The customer product feedback is valuable for improving product quality and for identification of service weaknesses [1]. Nowadays, e-commerce has enabled more and more customers to write reviews of products through their websites. These vast number of reviews have led to the raise of automatic harvesting processes such as sentiment analysis.

Sentiment analysis is a natural language processing mechanism to automatically extract opinions or feelings from texts. Usually, sentiment analysis focuses on three key areas: sentiment opinion, product-feature analysis, and comparative opinion mining [2]. Wei et al [3] develops an analysis using word vectors and Naïve Bayes to classify positive and negative opinions for an app review. Although Wei et al shows that word vectors are effective for this analysis, future work recommends exploring other algorithms such as Latent Dirichlet Allocation (LDA) for topic extraction. Additionally, Lui et al [4] proposes a product-feature based analysis, which essentially associates opinion words with a product feature category. This association can identify which product-feature the review is about and the polarity towards that feature.

The aim of this paper is to develop a sentiment analysis using customer reviews for a cell phone product. The structure of this paper will start with the methodology used for our study along with a short review on various current approaches from literature. Then, the paper shows the results obtained with the proposed model and includes a discussion about potential issues with the proposed model. Lastly, conclusions are presented.

III. Methodology

The methodologies for sentiment analysis are large. Perhaps, availability of different algorithms and evaluation methods enables several approaches for sentiment analysis. Among the most used classification algorithms in current research are the Naïve Bayes [5], Support Vector Machine [6], Logistic Regression and so forth. The evaluation methods used to assess different classification algorithms are diverse and include MAcroP, recall, precision, and F-measure methods. Additionally, there

are different methods proposed for semantic lexicon building. One is a thesaurus-based method, which consists of building a dictionary of words to compare with the text or reviews. Although this approach is simple, it is not accurate enough since the method disregards context of the text. The corpus-based method is another method that offers a more automated process to build semantics based on the search for "seed words" [7]. Lately, a fully automated lexicon building process has been implemented, using recurrent convolutional neural networks for text classification [8].

A. System Architecture

This section aims to provide insights on the architecture taken for this sentiment analysis study. Figure 1. illustrates the process flow of this architecture.

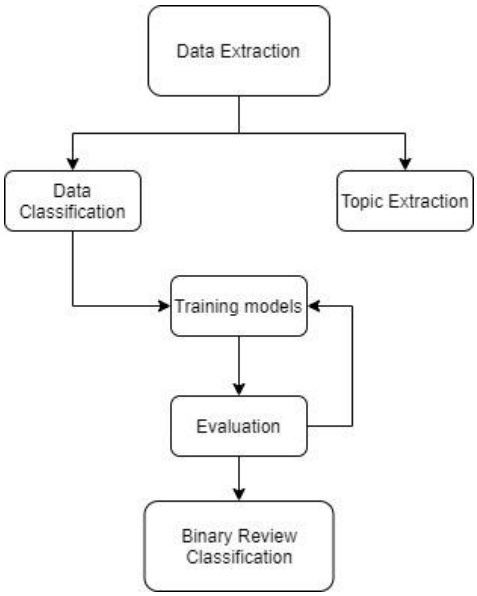

Figure 1. System Architecture

Data Extraction and Cleaning

The data used for this analysis was scrapped from Amazon Cell Phone Reviews posted on kaggle[1]. The challenge with mining reviews is the nature of the text that is posted by customers. The language tends to be grammatically incorrect and uses short forms of various words and phrases. Additionally, the structure of language is not followed correctly, and the grammar tends to be more spoken in nature than the written language.

The approach to clean data includes several steps. First, data is collected and removes irrelevant words, using a stopword file. Second, an identification process to spot bi-grams is developed to ensure the accuracy of meaning. Lastly, words are lemmatized since previous analysis showed there are several words that are verbs and are repeated in a gerund form, past or present tense. For instance, the word "Care" is repeated multiple times as "Caring", "Cares", or "Cared."

Word Cloud

[1] https://www.kaggle.com/grikomsn/amazon-cell-phones-reviews?select=20191226-reviews.csv

Figure 2. Positive Word Cloud

Figure 3. Negative Word Cloud

The dataset contains various reviews that are tightly related to each other and words that have similar usage, but there are also opposite polarities in the reviews that must be considered while making the sentiment classification. To understand the polarity of the words and how closely they are related to each other, the word cloud approach is an important aspect of data exploration in NLP.

Figure 4. Important Words on Amazon Logo

From Figure 1 and 2, it can be inferred that the word “phone” is the most important word. There are also other frequent words in the review text such as “app”, “work”, “screen”, “battery life”, “excellent”, “used”, etc**.** This makes sense since the reviews are about phones, and there are about 67% four and five star reviews so excellent word is also important. We can generate the word cloud of important words out of positive and negative reviews on a specific image.

Topic Extraction

The topic extraction in this study is done using LDA. LDA is a topic modeling algorithm which considers that a document is created from a cluster of topics. Each topic then generates a set of words based on the probability distribution. The LDA converts a document into a set of matrices. The algorithm then continuously associates the word within the document with a topic within the matrix and the probability of the document and the topic and the word and the topic to eventually determine the topic for which the distribution is the most stable across.

The key to any successful topic extraction is the cleanup of the data and conversion to the appropriate form of text to properly identify the Nouns and associated adjectives. In this implementation,

the data includes the topics extracted specifically from the lowest ratings of one and the highest rating of five.

Extracting the topics specifically from this subset of data allows us to identify the key areas that consumers are looking for when buying a mobile. We can expand the analysis further by subsequently including the rating two and four to provide more insight. The data set considered for this report has a majority of reviews under rating one and rating five and hence it would be sufficient for our study to include only this subset.

Data Classification

Word2vec: Since the release of Word2Vec in 2013 by Thomas Mikolov et al. [9], there is almost a decade of algorithmic improvement in NLP because of the competition. POS tagging and TF-IDF is used for feature engineering, but they are not able to capture the semantic meaning of the sentences, in Word2Vec model the word embeddings are trained on a large corpus of Google news data and was very efficient in capturing the semantic meaning. There are two general types of Word2vec models based on Auto-Encoders architecture.

- CBOW(Continuous bag of words) - in this approach the focus word is predicted based on the words around the focused words, where each word is represented as one-hot encoded representation. Linear activation is used, output is measured by the softmax function
- Skip gram

Training Models

Once the semantic lexicon is developed, the next step is to train models to classify reviews as either positive or negative. The training and testing data have been split into 80% for training and 20% for testing. Neural Networks and K-Nearest Neighbors algorithms have been implemented in this study.

Artificial Neural Networks primarily consists of three layers: Input Layer, Hidden Layers and Output Layer. In this work, it is assumed that one Input layer with a total number of Neurons is equal to features, hidden layers, with a number of neurons equal to half of the feature, and the final single neuron is equal to the output layer. A SoftMax layer is used at the end of the network to interpret the class labels based on probability. The following formula is used to calculate the number of neurons in the hidden layers:

$$Nh = Ns / (\alpha*(Ni+No))$$

Ni = Number of input neurons

No= Number of output neurons

Ns = Number of samples in training data set

Using the above formula, we can calculate that there will be approximately 100 neurons in the hidden layers, and we have also used a dropout value of 0.3 after every layer. Total layers in the model are fifteen.

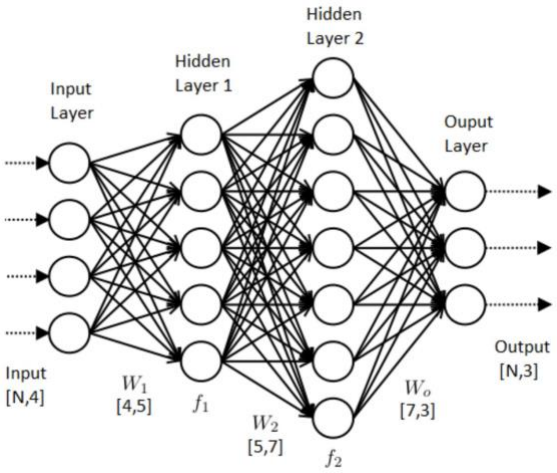

Figure 5. Two-layers shallow neural network

Additionally, a weight optimizing algorithm together with back-propagation is employed to tune the random weights to calculate errors in our prediction. Backpropagation compares the output against the random weights to the desired labels for the specific data row. Generally, a cost function J is used to find the error rate before updating the weights, the error is calculated in each of the hidden layers in the back-propagation step.

The K Nearest Neighbor is a non-parametric classifier considered in this study. There are five neighbors assigned to this model in addition to four different distance metrics such as "euclidean", "manhattan", "chebyshev", and "cosine" metrics.

Bert is a transformer-based model, it consists of twelve stacks of encoder decoder models, every encoder model consists of Attention model and feed-forward neural network. It comprises a total of 112 million parameters. To train the model, we used ktrain to load the BERT model and train it over the text reviews and score of each review, the output on new text can be generated as **score** being positive and **not_score** being negative.

Evaluation

The evaluation process in this model is done by implementing the F1 score and confusion matrix evaluation process. The F1 score evaluator is an “average” type of measure that considers the precision and recall. The F1 score is obtained by dividing the number of correctly predicted positive results by the number of positive results. Additionally, the second evaluator is a confusion matrix, which essentially measures the number of false or true positives or negatives. For an imbalance class distribution, accuracy metric is generally misleading, so we have used both accuracy and F1-Score, which is a harmonic measure of recall and precision.

Polarity Review Classification

The reviews are classified at five levels based on the ratings. However, we would need to classify the ratings as overall positive or negative. We would club the ratings for four and five as positive and rating 1 and 2 as negative and use this data to train the models. The polarity would allow the model to accurately classify the review as positive or negative. In this analysis, we are only classifying the reviews as positive or negative. As the model grows to be more accurate the scale of the classification can be increased.

The determined polarity can then be specifically calculated against each model to identify the overall sentiment or acceptance of a specific model or brand. The ratings are normally skewed towards a specific model, but sentiment analysis and classification of the reviews can identify the brand recognition, its acceptance or its dislike.

IV. Results

Topic Extraction
The following key topics were extracted from reviews, which allow for up to five ratings:

Topic 0: The topic identified is covering the best phones in the review. Automatically the Samsung Note and iPhone come up as topics. The main feature that is seen here is the fast response time, quality of the phone and phone's screen.
[('great', 14336.76), ('screen', 6917.20), ('fast', 4364.56), ('samsung', 3884.92), ('best', 3691.23), ('quality', 3617.71579045398), ('iphone', 3138.10), ('nice', 2929.87), ('note', 2907.96), ('card', 2846.89)]

Topic 1: The topic identified here is the Camera quality compared with the pictures captured by it and consists of the adjectives associated with the camera of different phones.
[('good', 9708.72), ('camera', 6950.04), ('price', 5820.17), ('issue', 2549.07), ('perfect', 2452.92), ('device', 2248.5617691317784), ('picture', 2064), ('awesome', 1936.47), ('day', 1688.36), ('mobile', 1675.60)]

Topic 2: The topic extracted here evidently points towards the battery life of the different phones
[('battery', 8064.17), ('love', 7342.11), ('life', 4124.79), ('time', 3621.78), ('android', 3124.69), ('problem', 2320.21), ('brand', 1941.54), ('review', 1899.33), ('charger', 1707.53), ('verizon', 1462.56)]

Topic 3: This topic adds to various additional features found in different phones such as looks, fingerprint sensor, hand gestures and display types
[('phone', 48381.13), ('feature', 3143.55), ('look', 2729.10), ('galaxy', 1987.81), ('fingerprint', 1772.88), ('google', 1722.16), ('feel', 1691.09), ('want', 1613.71), ('display', 1333.31), ('hand', 1195.91)]

Topic 4: The last topic covers abstract topics such as apps that are available on phones and recommendations of phones to other users.
[('work', 8141.84), ('need', 2606.31), ('happy', 2598.82), ('product', 2413.27), ('apps', 2353.39), ('easy', 2347.8806855445696), ('year', 1930.56), ('recommend', 1927.63), ('little', 1738.91), ('purchase', 1736.37)]

One scored rating reviews topics:

Topic 0: The topic identified talks about various features that can be problematic in a phone such as the screen, battery life, and customer support. It can be observed that problems faced by customers immediately lead to disappointment as expectations are not met by the product.
[('screen', 3076.48), ('problem', 1467.74), ('product', 1458.10), ('charge', 1141.70), ('doesn', 1055.09), ('service', 982.68), ('week', 900.78), ('year', 836.5), ('thing', 807.09), ('disappointed', 689.40)]

Topic 1: The topic covers mostly how the products were sent or received by the customer in a bad condition. The samsung products get highlighted here.
[('work', 3123.81), ('month', 2002.63), ('samsung', 1734.02), ('day', 928.22), ('turn', 801.23), ('charger', 780.20), ('apps', 726.74), ('bad', 701.88), ('send', 654.82), ('sent', 652.76)]

Topic 2: The Google pixel is directly compared with the iphone. Since these reviews do not include the iphone reviews it can be easily concluded that the pixel does not meet the expectations set. Warranty is another aspect that is identified.
[('amazon', 1357.56), ('card', 1208.75), ('warranty', 1065.10), ('seller', 1035.03), ('hour', 855.64), ('customer', 667.15), ('need', 640.04), ('iphone', 608.42), ('pixel', 546.28), ('text', 538.29)]

Topic 3: The iphone continues to set a benchmark in service. The phones are compared, and this topic mainly covers situations where the phones were returned to the seller and refunds were issued to the customers. Sprint and Verizon networks show up in this case and would need to be reviewed.
[('phone', 21680.20), ('issue', 1759.14), ('verizon', 1449.08), ('good', 1275.40), ('return', 1256.14), ('support', 843.84), ('review', 663.27), ('life', 650.74), ('refund', 593.75), ('sprint', 453.03)]

Topic 4: The battery life, update cycle, cost of the phone, camera are the main topics which are identified.
[('time', 2359.05), ('battery', 2209.39), ('bought', 1553.02), ('unlocked', 1404.02), ('device', 1100.38), ('money', 1092.01), ('camera', 1082.67), ('update', 1044.86), ('great', 976.76), ('google', 925.92)]

Sentiment Analysis

Given a sentiment lexicon and training and testing data, this study performs a sentiment analysis and determines the accuracy and evaluation score for each algorithm. The results are shown in Table 1. In

Model	Accuracy	F1-Score
KNN	78.18	86.12
Deep Neural Network	86.16	86.21
BERT	95.57	96.34

Table 1. Performance of three classifiers

KNN Confusion Matrix-

- TN stands for True Negative = Observation is negative,and predicted to be negative
- TP stands for True Positive = Observation is positive,and predicted to be positive
- FP stands for False Positive = Observation is negative,but predicted as positive
- FN stands for False Negative = Observation is positive,but predicted as negative

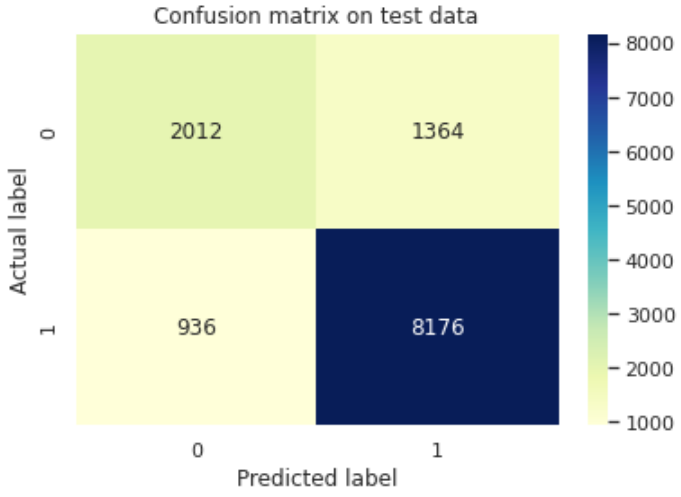

Figure. 6. Confusion Matrix using KNN

Deep Neural Network -

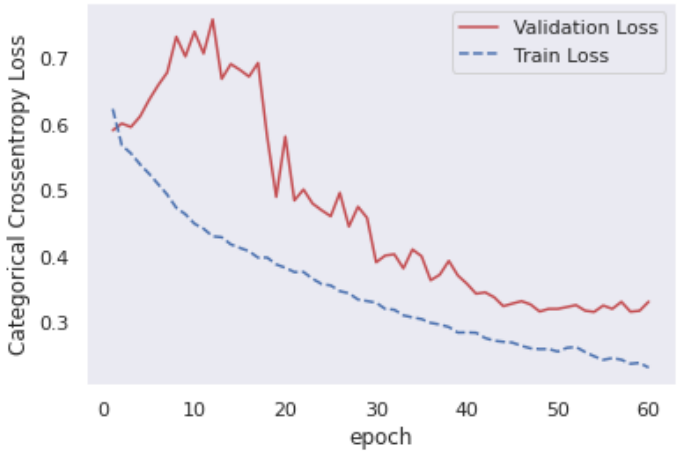

Figure 7. NN Train Loss Curve

V. Discussion

There are some issues related to semantic review that the models may not be able to pinpoint. Sarcasm could impact the classification of polarity. For example, a customer could comment "The product is advertised as a great product, but it is not" or questions like "I do not see what is great about this product." Furthermore, negation of reviews may hinder the true meaning of the review. For instance, a review that states "I wasn't fascinated with the service" could be classified as a positive comment since the word "fascinated" is in the comment. Lastly, irony may be another important factor that could lead to wrong polarity classification. Customers could write a review "I am enjoying the fast speed of this phone" while showing a picture of a phone loading for minutes.

Moreover, the training data is a static snapshot of opinions, which can be considered as a bias in the training data. However, these opinions are subjected to variation over the time. Customer perception of the product changes according to the modification or adjustments that are implemented over the product. After modifying and adjusting a feature that triggered a negative customer opinion, the customers will evaluate the new feature differently with regards to the new feature. This new evaluation will generate a new set of opinions for text mining purposes.

The LDA model has its own limitations such as the number of topics should be pre-determined and are not correlated. The data included in these reviews is a limited subset and includes additional languages and emoticons that the LDA model cannot deduce and the sentence structure is not considered.

The KNN model has some disadvantages especially because it is sensitive to noisy data which in case of reviews always tends to be noisy in nature as it consists of various inconsistencies such as languages, emoticons, images etc. Additionally, we would need to perform scaling of the data by normalizing and standardizing the data in order to achieve higher accuracy.

Similarly, artificial neural networks are also susceptible to noisy data. The major disadvantage of deep neural networks is that it can only improve the accuracy by adding the hidden layers upto a certain point, after that the saturation is reached and we have the problem of vanishing gradient and overfitting. The model becomes lazy and performs poorly on the test dataset (new), so we used dropout to prevent overfitting in the model, with the dropout value of 0.3 we reduced the overfitting significantly.

In addition, it has been observed that the highest F1-Score of 96.34 was achieved using transformer-based model - BERT, but it is very expensive computationally to train a single epoch of BERT, it will take 16 hours to train 2 epochs of the model using GPU and 20 minutes using TPU. We trained the transformer model using Google Colab.

There are two problems in BERT, which can be resolved in XL net transformer models [10]. Some of the transformer models like GPT-3 [11] are currently state of the art and perform better than BERT, so they can also be implemented on the review text data and a variety of Natural Language Understanding tasks. We can generate an end-to-end application with a frontend implementing state of the art models to

analyze the sentiment of new reviews on a mobile phone, it will be beneficial for the cell phone manufacturing companies to use the system to automate the feedback gathering task and improve on their performance.

VI. Conclusion

In this paper, we developed a sentiment lexicon based on word2vec algorithm. Based on the obtained sentiment lexicon, three algorithms were trained and tested. The results show that the algorithm that offers the best accuracy is BERT although this algorithm may demand relatively higher computational resources in comparison with the remaining algorithms.

Topic detection algorithms are critical in identifying the main data points for analysis. Word Clouds serve a limited purpose in visualizing the key words involved in an analysis. However, topic detection allows us to identify hidden concepts within the text and creates a logical outcome. In our analysis, it was easily observed that the phones were consistently compared to the iphone though the reviews themselves did not include any specific feature for the iphone. A cell phone company such as Samsung or Motorola would need to ensure they take a note of various features and services that Apple offers so that leads to users comparing their phones.

Mining reviews of various products is a critical area for many companies. It allows them to identify strengths and weaknesses of their own products compared to their competitors without manually going through the reviews and drive concrete conclusions. If implemented correctly, organizations can also predict how a particular product would be received in the market based on the features which are introduced. The prediction would allow companies to perform market sensitivity analysis and adjust to its demand or make improvements or changes to overcome a perceived reception of the product.

VII. References

[1] M. Zairi, "The art of benchmarking: using customer feedback to establish a performance gap," Total Quality Management, vol. 3, p. 2, 1992.

[2] B. Liu, Web Data Mining: Exploring Hyperlinks, Contents and Usage Data, Springer, 2007.

[3] Wei et al, "Apply Word Vectors for Sentiment Analysis of APP Reviews," in The 3rd International Conference on Systems and Informatics, 2016.

[4] Y. Zhu, B. Swen and S. Yu, "Mining Feature-based Opinion Expressions by Mutual Information Approach," International Journal of Computer Processing of Oriental Languages, vol. 20, pp. 133-150, 2007.

[5] H. K. M. Azharul, S. Mir Shahriar and A. Zakia, "Opinion Mining using Naïve Bayes," in IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE), Dhaka, Bangladesh, 2015.

[6] F. Bodendorf and C. Kaiser, "Mining Customer Opinions on the Internet A Case Study in the Automotive Industry," in Third International Conference on Knowledge Discovery and Data Mining, 2010.

[7] P. Chathuranga, S. Lorensuhewa and M. Kalyani, "Sinhala Sentiment Analysis using Corpus based Sentiment Lexicon," in International Conference on Advances in ICT for Emerging Regions, 2019.

[8] S. Lai, L. Xu, K. Liu and J. Zhao, "Recurrent Convolutional Neural Networks for Text Classification," in Association for the Advancement of Artificial, 2015.

[9] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in Advances in neural information processing systems, 2013, pp. 3111-3119.
[10] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, Xlnet, "Generalized autoregressive pretraining for language understanding," in Advances in neural information processing systems, 2019, pp. 5753-5763.
[11] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell et al., "Language models are few-shot learners" arXiv preprint arXiv:2005.14165, 2020.