Sentiment Analysis on Large Scale Amazon Product Reviews

Tanjim Ul Haque
Department of
Computer Science & Engineering
Ahsanullah University of Science
& Technology
Dhaka, Bangladesh
tuhaque@gmail.com

Nudrat Nawal Saber
Department of
Computer Science & Engineering
Ahsanullah University of Science
& Technology
Dhaka, Bangladesh
nudratsaber@gmail.com

Faisal Muhammad Shah
Department of
Computer Science & Engineering
Ahsanullah University of Science
& Technology
Dhaka, Bangladesh
faisal505@hotmail.com

***Abstract*—The world we see nowadays is becoming more digitalized. In this digitalized world e-commerce is taking the ascendancy by making products available within the reach of customers where the customer doesn't have to go out of their house. As now a day's people are relying on online products so the importance of a review is going higher. For selecting a product, a customer needs to go through thousands of reviews to understand a product. But in this prospering day of machine learning, going through thousands of reviews would be much easier if a model is used to polarize those reviews and learn from it. We used supervised learning method on a large scale amazon dataset to polarize it and get satisfactory accuracy.**

***Keywords*—Sentiment analysis, pool based active learning, feature extraction, text classification, machine learning.**

I. INTRODUCTION

As the commercial site of the world is almost fully undergone in online platform people is trading products through different e-commerce website. And for that reason reviewing products before buying is also a common scenario. Also now a day, customers are more inclined towards the reviews to buy a product. So analyzing the data from those customer reviews to make the data more dynamic is an essential field nowadays. In this age of increasing machine learning based algorithms reading thousands of reviews to understand a product is rather time consuming where we can polarize a review on particular category to understand its popularity among the buyers all over the world.

The objective of this paper is to categorize the positive and negative feedbacks of the customers over different products and build a supervised learning model to polarize large amount of reviews. A study on amazon last year revealed over 88% of online shoppers trust reviews as much as personal recommendations. Any online item with large amount of positive reviews provides a powerful comment of the legitimacy of the item. Conversely, books, or any other online item, without reviews puts potential prospects in a state of distrust. Quite simply, more reviews look more convincing. People value the consent and experience of others and the review on a material is the only way to understand others impression on the product. Opinions, collected from users' experiences regarding specific products or topics, straightforwardly influence future customer purchase decisions [1]. Similarly, negative reviews often cause sales loss [2]. For those understanding the feedback of customers and polarizing accordingly over a large amount of data is the goal. There are some similar works done over amazon dataset. In [5] did opinion mining over small set of dataset of Amazon product reviews to understand the polarized attitudes towards the products.

In our model, we used both manual and active learning approach to label our datasets. In the active learning process different classifiers are used to provide accuracy until reaching satisfactory level. After getting satisfactory result we took those labeled datasets and processed it. From the processed dataset we extracted features that are then classified by different classifiers. We used combination of two kinds of approaches to extract features: the bag of words approach and tf-idf & Chi square approach for getting higher accuracy.

II. RELATED WORKS

So far, much of the research papers related to product reviews, sentiment analysis or opinion mining has been done recently. In the work [3] Elli, Maria and Yi-Fan extracted sentiment from the reviews and analyze the result to build up a business model. They have claimed that demonstrated tools were robust enough to give them high accuracy. The use of business analytics made their decision more appropriate. They also worked on detecting emotions from review, gender based on the names, also detecting fake reviews. The commonly used programming language was python and R. They mainly used Multinomial Naïve Bayesian (MNB) and support vector machine (SVM) as their main classifiers. In paper [4] the author applied existing supervised learning algorithms to predict a reviews rating on a given numerical scale using only text. They have used hold out cross validation using 70% data as training data and 30% data as testing data. In this paper the author used different classifiers to determine the precision and recall values. The author in Paper [5] applied and extended the current work in the field of natural language processing and sentiment analysis to data from Amazon review datasets. Naïve Bayesian and decision list classifiers were used to tag a given review as positive or negative. They have selected books and kindle section review from amazon. The author in [6] aimed to build a system that visualizes the reviews sentiment in the form of charts. They have used data scraping from amazon url to get the data and

978-1-5386-5283-1/18/$31.00 ©2018 IEEE

978-1-5386-5696-9/18/$31.00 ©2018 IEI

preprocessed it. In this paper they have applied NB, SVM and maximum entropy. AS the paper claims that they summarize the product review to be the main point so there is no accuracy showed. They showed their result in statistical chart. In the paper [7] authors built a model for predicting the product ratings based on rating text using a bag-of-words. These models tested utilized unigrams and bigrams. They used a subset Amazon video game user reviews from UCSD Time-based models didn't work well as the variance in average rating between each year month, or day was relatively small. Between unigrams and bigrams, unigrams produced the most accurate result. And popular unigrams were extremely useful predictor for ratings for their larger variance. Unigram results had a 15.89% better performance than bigrams. In paper [8] various feature extraction or selection techniques for sentiment analysis are performed. They collected Amazon dataset at first and then performed preprocessing for stop words and special characters' removal. They applied phrase level, single word and multiword feature selection or extraction technique. Naive Bayes is used as the classifier. They concluded that Naive Bayes gives better result for phrase level than single word and multiword. The main cons of this paper are, they used only naive Bayes classifier algorithm from which we cannot get a sufficient result. In paper [9] it has used easier algorithms so it is easy to understand. The system gives high accuracy on svm and so it cannot work properly on huge dataset. They used support vector machine (svm), logistic regression, decision trees method. In paper [10] tfidf is used here as an additional experiment. It can predict rating by using bag of words. But Classifiers used here are only few. They used root mean square error, linear regression model. So, those are some related works mentioned above, we tried to make our work more efficient by choosing best ideas from them and applied those together.

In our system, we used large amount of datasets so it gave efficient result and we could take better decision. Moreover, we have used active learning approach to label datasets which can dramatically accelerate many machine learning tasks. Our system also consists of several types of feature extraction methods. To the best of our knowledge, our proposed approach gave higher accuracy than the existing research works.

III. METHODOLOGY

Amazon is one of the largest E-commerce site as for that there are innumerous amount of reviews that can be seen. We used data named Amazon product data which was provided by researchers from [14]. The dataset was unlabeled and to use it in a supervised learning model we had to label the data. We used three JSON files where the structure of the data is as follows:

"reviewerID": ID of the reviewer

"asin": ID of the product

"reviewerName": name of the reviewer

"helpful": helpfulness rating of the review

"reviewText": text of the review

"overall": rating of the product

"summary": summary of the review

"reviewTime": time of the review (raw)

For data we selected three categories from Amazon products Electronics reviews, Cell Phone and Accessories Reviews and Musical Instruments product reviews which consists of approximately **48500** product reviews. Where 21600 reviews are from mobile phones, 24352 are from electronics & 2548 from musical instruments data. From the formats used for analyzing the review polarity we used review Text & Overall from it. We can see an overview of our methodology:

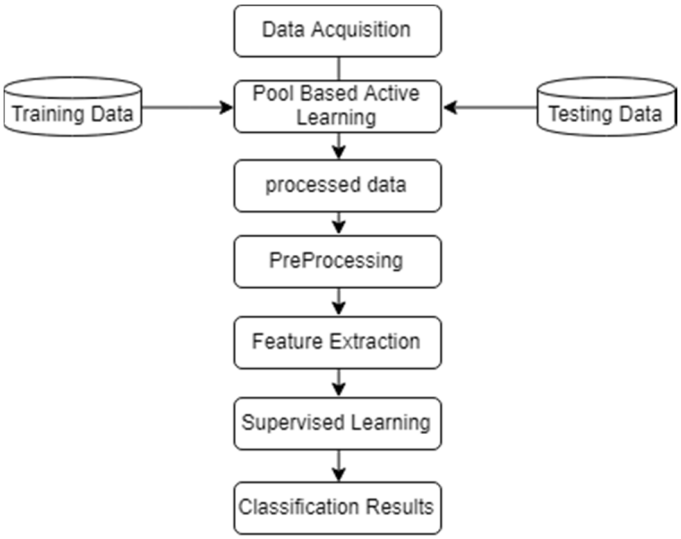

Figure 1: Work Process

*A. **Data Acquisition***

We acquired our dataset of 3 different JSON formats and labeled our dataset. As we have a large amount or reviews manually labeling was quite impossible for us. Therefor we preprocessed our data and used Active learner to label the datasets. As amazon reviews comes in 5-star rating based generally 3 star ratings are considered as neutral reviews meaning neither positive nor negative. So we discard any review which contains a 3-star rating from our dataset and take the other reviews and proceed to next step labeling the dataset.

Pool Based Active Learning:

Active learning is a special case in semi-supervised learning algorithm. The main fact is that the performance will be better with less training if the learning algorithm is allowed to choose the data from which it learns [2]. Active learning system tries to solve data labeling bottleneck by querying for unlabeled instance to be properly labeled by an expert or oracle. As manually labeling the dataset is quite an impossible task so that to reduce time complexity we use a special kind of semi-supervised learning approach known as pull based active

978-1-5386-5696-9/18/$31.00 ©2018 IEEE

learning. In the process of our active learning we need to provide it some pre labeled datasets as training and testing and take unlabeled dataset. For using active learning, we need to provide some manually labeled reviews as training –testing sets. Then from a pool of unlabeled dataset learning method will ask oracle or user to label few data. And it will run some classifiers to calculate the accuracy. Accuracy shows whether the decision boundary is separating most the values in two classes. Higher the accuracy higher the data is being labeled. If the accuracy is greater or equal to 90% then we take those data and combined it with already pre-labeled data to get our labeled dataset. If not, we again consider help from the oracle to label some more data. After the accuracy is greater than 90% we considered the data to be labeled.

B. ***Data Pre-Processing***

Tokenization**:** It is the process of separating a sequence of strings into individuals such as words, keywords, phrases, symbols and other elements known as tokens. Tokens can be individual words, phrases or even whole sentences. In the process of tokenization, some characters like punctuation marks are discarded. The tokens work as the input for different process like parsing and text mining.

Removing Stop Words: Stop words are those objects in a sentence which are not necessary in any sector in text mining. So we generally ignore these words to enhance the accuracy of the analysis. In different format there are different stop words depending on the country, language etc. In English format there are several stop words.

POS tagging**:** The process of assigning one of the parts of speech to the given word is called Parts of Speech tagging. It is generally referred to as POS tagging. Parts of speech generally contain nouns, verbs, adverbs, adjectives, pronouns, conjunction and their sub-categories. Parts of Speech tagger or POS tagger is a program that does this job.

C. ***Feature Extraction***

Bag of Words: Bag of word is a process of extracting features by representing simplified text or data, used in natural language processing and information retrieval. In this model, a text or a document is represented as the bag (multiple set) of its words. So, simply bag of words in sentiment analysis is creating a list of useful words. We have used bag of words approach to extract our feature sets. After preprocessed dataset we used pos tagging to separate different parts of speech and from that we select nouns and adjectives and use those to create a bag of words. Then we run it through a supervised learning and find our results and also the top used words from the review dataset.

TF-IDF:TF-IDF is an information retrieval technique which weighs a term's frequency (TF) and also inverse document frequency (IDF). Each word or term has its own TF and IDF score. The TF and IDF product scores of a term is referred to the TF*IDF weight of that term. Simply we can state that the higher the TF*IDF score (weight) the rarer the term and vice versa. TF of a word is the frequency of a word.

IDF of a word is the measure of how significant that term is throughout the corpus.

When words do have high TF*IDF weight in content, content will always be amongst the top search results, so anyone can:

1. Stop worrying about using the stop-words,
2. Successfully find words with higher search volumes and lower competition.

Chi Square: Chi square(X^2) is a calculation that is used to determine how smaller the difference between the observed data and the expected data .
In this approach we have preprocessed our dataset then we have divided data into training and testing set. We used pipeline method to apply TF-IDF, Chi square and other classifiers onto our dataset and got the results.

Algorithm for proposed approach

Input:
Labeled Data=labeled data obtained after **active learning** process.

Output:
Accuracy of classifiers;
Precisio,Recall,F-1 Measure for positive and deceptive values.
//product review polarity accuracy

1. Load labeled data positive & negative
2. Preprocess labeled data
3. for every X= {X1…Xn} in labeled data
4. Extractfeature(Xi)
5. Cross validate into training & testing set
6. Classifier.train()
7. Accuracy= classifier.accuracy()
8. majority_voting(accuracy) using vote classifier

9. show result(accuracy,precision,recall,f1measure)

10.end

extractfeature(text) return n-gram feature

majority_voting(accuracy) return highest accuracy

D. ***Evaluating Measures:***

Evaluate metrics play an important role to measure classification performance. Accuracy measure is the most common for this purpose. The accuracy of a classifier on a given test dataset is the percentage of those dataset which are correctly classified by

978-1-5386-5696-9/18/$31.00 ©2018 IEEE

the classifier [48]. And for the text mining approach always the accuracy measure is not enough to give proper decision so we also took some other metrics to evaluate classifier performance. Three important measures are commonly used precision, recall, F-measure. Before discussing with different measures there are some terms we need to get comfortable with-

- TP (True Positive) represents numbers of data correctly classified
- FP (False Positive) represents numbers of correct data misclassified
- FN (False Negative) represents numbers of incorrect data classified as correct
- TN (True Negative) is the numbers of incorrect data classified

Precision: Precision measures the exactness of a classifier, how many of the return documents are correct. A higher precision means less false positives, while a lower precision means more false positive. Precision (P) is the ratio of numbers of instance correctly classified from total. It can be defined as-

$$P = \frac{TP}{TP+FP}$$

Recall: Recall calculates the sensitivity of a classifier; how many positive data it returns. Higher recall means less false negatives. Recall is the ratio of number of instance accurately classified to the total number of predicted instance. This can be shown as-

$$R = \frac{TP}{TP+FN}$$

F-Measure: Combining precision and recall produces single metrics known as F-measure, and that is the weighted harmonic mean of precision and recall. It can be defined as –

$$F = \frac{2P.R}{P+R}$$

Accuracy: Accuracy predicts how often the classifier makes the correct prediction. Accuracy is the ratio between the number of correct predictions and the total number of prediction.

$$Accuracy = \frac{Correct\ Prediction}{Total\ data\ points}$$

IV. RESULTS

There were several machine learning algorithms used in our experiment like Naïve Bayesian, Support vector Machine Classifier (SVC), Stochastic Gradient Descent (SGD), Linear Regression (LR), Random Forest and Decision Tree. We have conducted cross validation methods and 10 fold gave the best accuracy. We conduct the best classifiers on 3 categories of product reviews and see the results according to the evaluation measures. The classifiers were applied on different feature selection process where the common features from TF-IDF and bag of words gave best results for all the datasets.

Dataset	Classifier	Accuracy 10 Fold	Accuracy 5 Fold	Precision	Recall	F1 score
CELLPHINE & ACCESSORIES	Linear support Vector machine	**93.57**	**88.34**	0.96	0.97	0.97
	Multinomial Naïve Bayes	90.28	84.41	0.89	0.92	0.91
	Stochastic Gradient Descent	91.88	84.93	0.9	0.93	0.91
	Random Forest	92.72	88.20	0.967	0.967	0.97
	Logistic regression	88.2	81.99	0.87	0.88	0.88
	Decision tree	91.45	83.71	0.95	0.95	0.95

Table-1: Experiment result for cellphone & accessories data

Dataset	Classifier	Accuracy 10 Fold	Accuracy 5 Fold	Precision	Recall	F1 score
MUSICAL	Linear support Vector machine	**94.02**	89.76	0.9889	0.971	0.98
	Multinomial Naïve Bayes	91.57	89.77	0.98	0.93	0.96
	Stochastic Gradient Descent	92.89	88.264	0.99	0.96	0.98
	Random Forest	93.56	88.51	0.98	0.97	0.975
	Logistic regression	91.34	87.14	0.96	0.95	0.95
	Decision tree	92.45	86.27	0.969	0.96	0.96

Table-2: Experiment result for musical Instruments data

Dataset	Classifier	Accuracy 10 Fold	Accuracy 5 Fold	Precision	Recall	F1 score
ELECTRONICS	Linear support Vector machine	**93.52**	**91.72**	0.98	0.99	0.98
	Multinomial Naïve Bayes	89.36	86.89	0.899	0.96	0.93
	Stochastic Gradient Descent	92.61	90.96	0.964	0.988	0.975
	Random Forest	92.89	91.14	0.968	0.988	0.978
	Logistic regression	88.96	87.843	0.919	0.955	0.937
	Decision tree	91.569	87.50	0.962	0.9669	0.96

Table-3: Experiment result for electronics data

From all the experiments it can be seen that support vector machine provided with greater accuracy in every dataset. As the working dataset is quite larger and support vector machine works better with large scale dataset without over fitting it. And from these results highest accuracy was 94.02%.

978-1-5386-5696-9/18/$31.00 ©2018 IEEE

V. COMPARATIVE ANALYSIS

In this section our research was tried to be compared with other related works. The comparative analysis was based on accuracy. The comparison can be seen in the table below-

Paper Title	Year & Citations	Dataset	Accuracy
Amazon Reviews,business analytics with sentiment analysis [11]	2016	Review of cellphone& accessories	72.95%
			80.11%
Sentimetn Analysis in Amazon Reviews Using Probalbilistic Machine Learning [5]	2013 (6)	reviews of books	84.44%
		reviews of Kindle	87.33%
Mining somparative opinions from customer reviews for competitive intelligence [12]	2011 (234)	Customer product reviews	61.00%
Amazing: A sentiment mining & Retrieval System [12]	2009 (125)	E commerce reviews	87.60%
"Feature Selection Methods in Sentiment Analysis and Sentiment Classification of Amazon Product Reviews" [8]	2016	Review on books	70.00%
			70.00%
			80.00%
		Review on music	62.00%
			80.00%
			68.00%
		Review on Camera	62.00%
			80.00%
			68.00%
Proposed Model	2018	Review of cellphone& accessories	93.57%
		Review of Electronics	93.52%
		Reviews of music Instruments	94.02%

Table-4: Comparative Analysis

Different researches listed in the table have conducted different pre-processing steps and feature extraction processes. As in our research we tied to improvise all the extraction processes and preprocessing steps and pick the best accuracy from it. Pull based active learning process have contributed labeling and selecting the best reviews as our training and testing data. Use of different preprocessing process helped sorting out unnecessary words. And finally taking the best features extracted from the datasets and learning through proper classifiers it was possible to attain greater accuracy. From the table it can be decided that the approaches used in approaches our proposed model shows more effectiveness and could achieve a better result than some of the related works.

VI. CONCLUSION AND FUTURE WORKS

In this research we proposed a supervised learning model to polarize a large amount of product review dataset which was unlabeled. We proposed our model which is a supervised learning method and used a mix of 2 kinds of feature extractor approach. We described the basic theory behind the model, approaches we used in our research and the performance measure for the conducted experiment over quite a large data. We also compared our result with some of the similar works regarding product review. We also went through different kinds of research papers regarding sentiment analysis over a text based dataset. We were able to achieve accuracy over 90% with the F1 measure, precision and recall over 90%. We tried different simulation using cross validation, training-testing ratio, and different feature extraction process for comparing varying amount of data to achieve promising results. In most of the cases 10 fold provided a better accuracy while Support Vector Machine (SVM) provided best classifying results. It is hard to gather huge amount of gold standard dataset for this purpose as e-commerce sites have their limitations on giving data publicly. Also scraping data can be a problem as we can't scrape enough data to consider it as real-life public reviews over different products.

Some future works which can be included to improve the model and also to make it more effective in practical cases. Our future works include applying PCA (Principal Component Analysis) in active learning process to fully automate data labeling process with less assistance from the oracle. The model can be incorporate with programs that can interact with customer seeking a score of a particular product. As we used a large scale dataset we can apply the model on local market sites to get better accuracy and usability. And lastly we will try to continue this research until we generalize this model to all kinds of text based reviews and comments.

REFERENCES

[1] Samha,Xu,Xia, Wong & Li "Opinion Annotation in Online Chinese Product Reviews." In Proceedings of LREC conference, 2008.

[2]. Nina Isabel Holleschovsky, "The social influence factor: Impact of online product review characteristics on consumer purchasing decisions", 5 th IBA Bachelor Thesis Conference, Enschede, The Netherlands 2015

978-1-5386-5696-9/18/$31.00 ©2018 IEEE

[3]Elli, Maria Soledad, and Yi-Fan Wang. "Amazon Reviews, business analytics with sentiment analysis." 2016

[4]Xu, Yun, Xinhui Wu, and Qinxia Wang. "Sentiment Analysis of Yelp‘s Ratings Based on Text Reviews." (2015).

[5] Rain, Callen. "Sentiment Analysis in Amazon Reviews Using Probabilistic Machine Learning."Swarthmore College (2013).

[6] Bhatt, Aashutosh, et al. "Amazon Review Classification and Sentiment Analysis." International Journal of Computer Science and Information Technologies 6.6 (2015): 5107-5110.

[7]Chen, Weikang, Chihhung Lin, and Yi-Shu Tai."Text-Based Rating Predictions on Amazon Health & Personal Care Product Review." (2015)

[8]Shaikh, Tahura, and DeepaDeshpande. "Feature Selection Methods in Sentiment Analysis and Sentiment Classification of Amazon Product Reviews.",(2016)

[9]Nasr, Mona Mohamed, Essam Mohamed Shaaban, and Ahmed Mostafa Hafez. "Building Sentiment analysis Model using Graphlab." IJSER, 2017

[10]Text mining for yelp dataset challenge; Mingshan Wang; University of California San Diego, (2017)

[11] Elli, Maria Soledad, and Yi-Fan Wang. "Amazon Reviews, business analytics with sentiment analysis." 2016

[12] Xu, Kaiquan, et al. "Mining comparative opinions from customer reviews for Competitive Intelligence." Decision support systems 50.4 (2011): 743-754.

[13] Miao, Q., Li, Q., & Dai, R. (2009). AMAZING: A sentiment mining and retrieval system. Expert Systems with Applications, 36(3), 7192-7198.

[14] He, Ruining, and Julian McAuley. "Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering." Proceedings of the 25th International Conference on World Wide Web.International World Wide Web Conferences Steering Committee, 2016.

978-1-5386-5696-9/18/$31.00 ©2018 IEEE