Amazon Product Reviews: Sentiment Analysis

Vineet Jain [vineet.jain2@mail.dcu.ie],
School of Computing, Dublin City University,
Glasnevin, Dublin 9, IRELAND

***Abstract* – In this paper, we propose a methodology that performs sentiment analysis on product reviews collected from Amazon. Experiments for both classifications of reviews and extraction of narratives from the text are performed with promising outcomes. We also discusses about the existing research in this area and at last, giving an insight about our future work on analysis.**

Keywords: *Sentiment analysis, customer reviews, product reviews, classification, amazon.*

1. INTRODUCTION

The advent of electronic commerce with growth in internet and network technologies has led customers to move to online retail platforms such as Amazon, Walmart, Target, etc. People often rely on customer reviews of products before they buy online. These reviews are often rich in information describing the product. Customers often choose to compare between various products and brands based on whether an item has a positive or negative review, More often, these reviews act as a feedback mechanism for the seller. Through this medium, sellers strategize their future sales and product improvement.

Sentiment analysis or opinion mining is one of the major tasks of NLP (Natural Language Processing) that has gained much attention in recent years. The sentiment is a feeling, expression, thought, or judgment, and using sentiment analysis one can study the target audience's sentiments towards an entity. Its a form of text analysis that senses polarity (e.g. a positive or negative opinion) within whole text, sentence, paragraph or phrase. Knowing people's emotions is important for companies because consumers can communicate their thoughts and feelings more freely than ever before. With the technological improvements in the field of machine learning and automation, Companies can create system that automatically analyzing customer feedback, survey responses and social media interactions. These way marketers can listen to their customers closely, and customize goods and services to suit their needs.

In this paper, we studied the notion of *'aboutness'* through an important type of semantic processing task – 'topic modelling'. Most texts are usually comprised of multiple topics, i.e. topics being talked about in the text. For example, say a product manager at Amazon wants to understand what features of a recently released product (say Amazon Alexa) customers are talking about in their reviews. Say they are able to identify that 50% of people talk about the hardware, 30% talk about features related to music, while 20% talk about the packaging of the product. This would be immensely useful in a similar scenario where a seller on Amazon receives multiple negative reviews and wants to understand the reason behind it.

We also studied the case where buyers are seldom misled by overall collective rating that Amazon without any bifurcation between a service review and product review.
Hence, we trained our model to also classify the reviews as positive, negative and neutral.

Following sections of the paper provide a thorough insight of the tasks that our team undertook, and briefly discuss the previous research/ related work present in the journals worldwide.

2. RELATED WORK

Sentiment analysis has been there for a while, but a lot of active research has happened in the past few years to understand and classify customer reviews. For example, Levent Guner[1] el al from KTH Royal Institute of Technology, Stockholm selected 60,000 random product reviews from Amazon. They used the dataset available in Kaggle that contains 4 million reviews. The performance was compared with three different algorithms namely Multinomial Naïve Bayes (MNB), Linear Support Vector Machine (LSVM) and Long short-term memory network (LSTM). The authors used multiple performance metrics to determine the best performing classification algorithm on the test set. The performance metrics used were Accuracy, Area Under Curve (AUC), Precision, Recall and F1-score. Based on the results of the evaluation, their study concluded that the LSTM model performed the best with precision > 0.90 and AUC = 0.96 for binary classification (positive and negative). Xing Fang and Justin Zhan [2] collected over 5.1 million product reviews in 4 key categories: beauty, book, electronics, and home. They analyzed these reviews with 3 different classifiers, namely, naïve bayes, support vector machines and random forest. Their paper addressed the basic question of evaluating sentiments, categorizing sentiment polarity and concluded with random forest generating more reliable results. As per their findings, for larger data sets SVM worked better than NB.

Wanliang Tan[3] et al performed both traditional machine learning algorithms including naïve bayes, SVM, k-nearest neighbor and deep learning network models such as recurrent network models and LSTM on amazon reviews dataset. They collected 34627 reviews and divided it into 21000 and 13627 records of training and test datasets respectively. In terms of test accuracy, LSTM performed best among all of them with 71.5% accuracy. One of the key reasons for not high enough accuracy was the imbalance in their data, as they concluded. Callen Rain[4] used naïve bayes and decision-list classifiers to classify product reviews (category: books) from Amazon as positive and negative. He used a corpus that includes 50,000 reviews of 15 items that serve as the research dataset. The features such as bag-of-words and bigrams are compared with each other in their usefulness in labelling positive and negative reviews correctly. His analysis showed that naive bayes performed better than the decision-list and bag of words ended up being the best form of feature extraction.

Nishit Shrestha and Fatma Nasoz[5] analyzed the opinions of Amazon.com reviews. They developed a model using recurrent neural networks (RNN) with gated recurrent unit (GRU) that learned low-dimensional review vector representation using paragraph vectors and product embedding. The data used in this analysis is a collection of about 3.5 million product reviews gathered from Amazon.com. Paragraph vectors (PV) are very much inspired by word vectors. PV system learns vectors by predicting the next term, given several sampled contexts from a paragraph. The concatenation of review embedding developed from paragraph vectors and GRU-derived product embedding is used to train a support vector machine (SVM) to classify sentiments. With only review embedding, the anticipated classifier provided 81.29 percent accuracy. The product embedding inclusion improved the accuracy to 81.82 percent. Authors believe that a similar technique can be used to learn user information.

In a research article [7] different approach has been implemented for sentimental analysis in this research an algorithm called a Bow (Bag of words) is used in which the relationship between the words was not considered. To measure the sentiment for the whole sentence, the sentiment of every single word of the sentences has been individually determined and values are collected using some aggregation function. Along with this opinion summarization method based on features driven can be used.For each product a specific feature and their attributes are obtained, and the general feature for each product class is obtained. Then polarity is assigned to each function with the aid of Sequential Minimal Optimization and Support Vector Machines.

3. DATASET AND FEATURES

Our dataset comes from UCSD Design Lab[10]. It contains item reviews and metadata from Amazon, including 142.8 million reviews spreading over May 1996 - July 2014 in a one-review-per-line JSON format. This dataset incorporates reviews (ratings, content, support votes), item metadata (portrayals, class data, value, brand, and picture highlights), and connections. We have picked datasets for the following categories: Musical Instruments, Office Products, Home and Kitchen, Video Games, Beauty, Sports and Outdoors, Toys and Games, Cell Phones and Accessories, Health and Personal Care.

The files contain attributes ‘reviewer ID’, ‘ASIN’, ‘Reviewer Name’, ‘Reviewer text’, ‘Helpful’, ‘Summary’, ‘Rating’, and ‘Review time’.

It demonstrates in that there are 5 group-rankings from 1 to 5. In comparison, these five classes are potentially imbalanced as Class 1 and Class 2 have limited quantities of data while Class 5 has over 500000 ratings.

4. METHODOLOGY

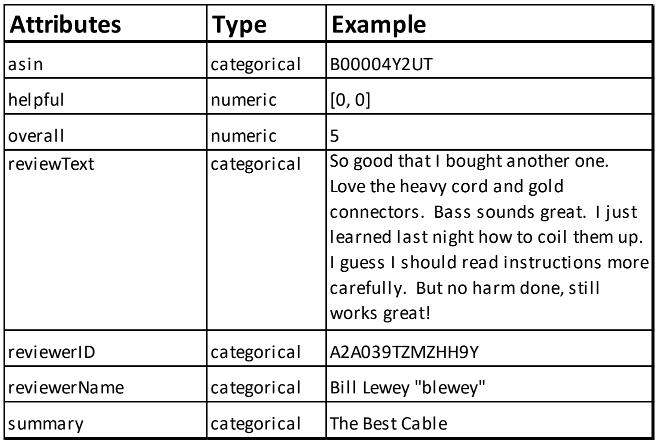
The structure for analytics problem solving is called the CRISP-DM Framework: Cross Industry Standard Process for Data Mining. It involves a series of steps which are quite intuitive: Business understanding, Data understanding, Data Preparation, Data Modelling, Model Evaluation, and finally Model Deployment.

4.1 BUSINESS UNDERSTANDING

We determine our objective by understanding the research questions: Given the ratings of positive, neutral and negative (on a scale of 1-5) for each product review on Amazon, can we predict the ratings of future reviews. Furthermore, can we extract a narrative out of the text reviews for each product.

4.2 DATA UNDERSTANDING

Data contains attributes such as reviewerID (which refers to the unique identification number for each reviewer who has bought the product from the website and provided a rating with some text review), reviewerName (gives the name of the reviewer), reviewText (this is the product review text written by the reviewer on the product page), summary (obtained from reviewText), overall (product rating on a scale of 1-5), asin (this is the ID of the product), helpful (helpfulness rating of the review, e.g. 2/3 5).



Attributes	Type	Example
asin	categorical	B00004Y2UT
helpful	numeric	[0, 0]
overall	numeric	5
reviewText	categorical	So good that I bought another one. Love the heavy cord and gold connectors. Bass sounds great. I just learned last night how to coil them up. I guess I should read instructions more carefully. But no harm done, still works great!
reviewerID	categorical	A2A039TZMZHH9Y
reviewerName	categorical	Bill Lewey "blewey"
summary	categorical	The Best Cable

Exploratory data analysis shows highest word frequency of stopwords, which were removed in the subsequent pre-processing steps. No outliers or blanks were found in the raw dataset.

As shown in Figure 1, rating distribution for the dataset is skewed towards the left.

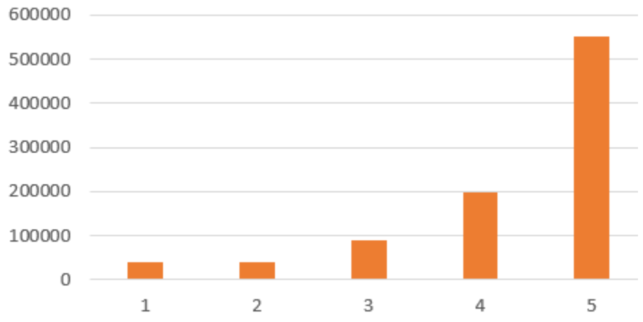


Figure 1. Rating distribution in raw data.

4.3 DATA PREPARATION

Datasets for the 6 different categories were merged in a single data-frame, with below dimensions:

Total records = 920395
Total attributes = 9

The features of interest were found to be ‘reviewText’, ‘summary’ and ‘rating’ as these were most useful and relevant for model building, prediction as well as abstraction. In the ‘reviewText’ field, mixed bag of short and long reviews were found, some of which produced no or little information about the product. In the exploratory analysis, reviewText attribute was explored using the word-cloud visualization for each category. Figure 2 shows word frequencies of tokens in the train set of Office Products category.

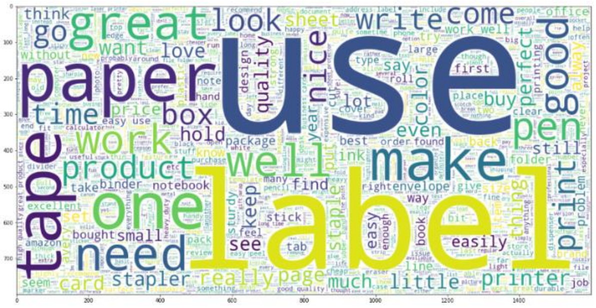


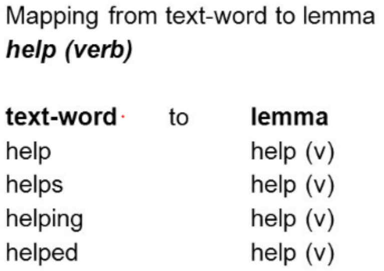
Figure 2. Word-Cloud created on training set for Office Category.

4.4 MODEL

The idea of distributional semantics that is implemented through 'word vectors' has been used heavily in semantic processing for a wide variety of applications. This simple idea has probably been the most powerful and useful insight in creating semantic processing systems. Supervised techniques for word sense disambiguation require the input words to be tagged with their senses. The sense is the label assigned to the word. In unsupervised techniques, words are not tagged with their senses, which are to be inferred using other techniques. In supervised techniques, such as naive Bayes (or any classifier), context-sense set is taken as the training data. The label is the 'sense' and the input are the context words.

In unsupervised techniques, such as the lesk algorithm, you assign the definition to the ambiguous word which overlaps with the surrounding words maximally

Lemmatization[6] is the way toward changing over a word to its base structure. The contrast among stemming and lemmatization will be, lemmatization considers the specific situation and changes over the word to its significant base structure, while stemming just evacuates the last not many characters, frequently prompting off base implications and spelling mistakes.



Mapping from text-word to lemma
help (verb)

text-word	to	**lemma**
help		help (v)
helps		help (v)
helping		help (v)
helped		help (v)

Wordnet Lemmatizer from NLTK library was used to establish structured semantic relationships between words. NLTK POS tagger was used directly without training, and treebank tags were mapped to WordNet part of speech names for tagging the adjectives, nouns and verbs.

Final lemmatized text base was split into train and test datasets for modeling. Using CountVectorizer in NLTK text was converted to word count vectors for both train and test datasets. The CountVectorizer gives a basic method to both tokenize an assortment of content records and construct a jargon of known words, yet additionally to encode new reports utilizing that jargon. The TfidfVectorizer tokenized archives, gained proficiency with the jargon and backwards record recurrence weightings, permit to encode new reports, allowing to now have an educated CountVectorizer that was utilize with a TfidfTransformer to figure the opposite archive frequencies and begin encoding records. Logistic regression and naïve bayes classifiers were compared to classify reviews as positive and negative.

For deriving the narratives, few product IDs were filtered out and sentiments were added to their text vectors. This sentiment completes the whole opinion comprising of both abstraction and expression. For deriving narratives, we used K-means clustering technique through which we clustered together the reviews for an ASIN according to their sentiments such that each cluster contains unique reviews having similar sentiments.

For all the POS tagged reviews or word vectors in a cluster we looked for below patterns or any combination of these

- Noun phrase (2 or more nouns occurring together, for example, United states of America, Steven Tyler
- Number followed by Noun, for example, 90 percent left, 3 months, 2 days battery life)
- Adjective followed by Noun, for example, economic impact, tremendous battery, beautiful screen
- Foreign words/ Internet catch phrases, for example, IMHO, BYOB
- Noun followed by Verb, for example, SmartElectronics cheated

4.5 MODEL EVALUATION

Logistic model showed accuracy of 0.89 on the bag of words and TF-IDF score of 0.88 on Office Products category, while Naiver Bayes model produced similar results of 0.889 and 0.888.

Musical Instruments category produced scores of precision = 1.00, recall 0.01, f1 of 0.03 and support of 228 on 0 and 0.89, 1.00, 0.94, 1825 respectively on 1, with macro average scores of 0.95, 0.51, 0.48 and 2053. Office Products category produced scores of precision = 0.30, recall 0.03, f1 of 0.05 and support of 118 on 0 and 0.90, 0.99, 0.86, respectively on 1 for naïve bayes.

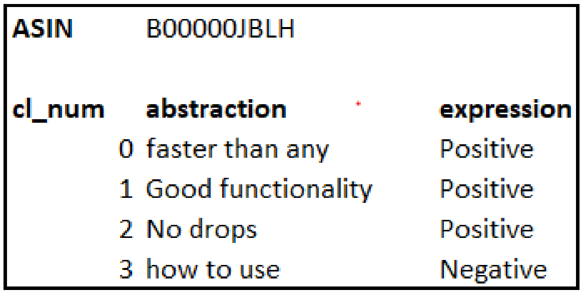
Ration of tp / (tp + fp) is called Precision[6] where tp is the number of true positives and fp the number of false positives. The precision is hence the ability of the classifier not to label a negative sample as positive. The recall is the tp / (tp + fn) ratio whereas tp is the number of true positives and fn the number of false negatives. Intuitively the recall is the classifier's ability to identify all the positive samples.It is possible to view the F-beta score as a weighted harmonic mean of precision and recall, where an F-beta score achieves its highest value at 1 and the worst at 0.

The F-beta score weights recall more than precision by a factor of beta. beta == 1.0 means recall and precision are equally important.

The support is the number of occurrences of each class in y_true.

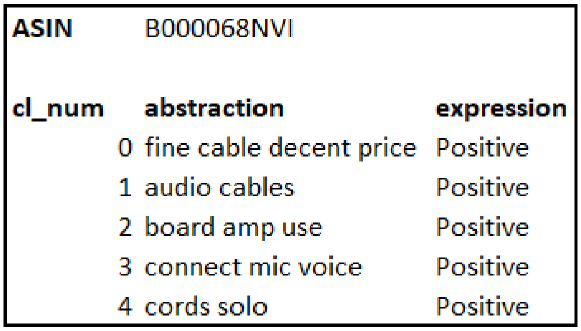
If pos label is None then this function returns the average accuracy, recall and F-measure in binary classification if the average is one of 'micro,' 'macro,' 'weighted' or 'samples.'

We experimented with k-means clustering on the text vectors for k = [4,11] and cluster with best silhouette score value for k. Each text vector was assigned a cluster number accordingly, and further only unique reviews along with their frequencies were kept for each cluster. As ASINs with fewer reviews displayed poor silhouette scores, those with < 0.02 were discarded. This value was tweaked accordingly depending upon the number of reviews for an ASIN. Silhouette scores < 0.02 showed that there is no strong similarity between the reviews in the cluster Figure 3 shows the output narrative (abstraction + expression) for ASIN# B000068NVI.



ASIN	B00000JBLH	
cl_num	**abstraction**	**expression**
0	faster than any	Positive
1	Good functionality	Positive
2	No drops	Positive
3	how to use	Negative

Figure 3. Narrative output for ASIN B0000JBLH



ASIN	B000068NVI	
cl_num	abstraction	expression
0	fine cable decent price	Positive
1	audio cables	Positive
2	board amp use	Positive
3	connect mic voice	Positive
4	cords solo	Positive

Figure 4. Narrative output for ASIN B000068NVI

5. CONCLUSION AND FUTURE WORK

We leave refinement the narratives for accuracy as a part for future work and are also working to incorporate principal component analysis (PCA) that will automate the active learning process of our system and will fully automate the data labellings process and will require less assistance. We will be looking into datasets obtained from local supermarket, This will provide better understanding of our sentiments based on demographics. And lastly we will try to work on this product to achieve a generalize form of this model.

6. REFERENCES

[1] Levent Guner, Emilie Coyne and Jim Smit, “Sentiment Analysis of Amazon.com Reviews”, March,2019,Available:https://www.researchgate.net/publication/332622380_Sentiment_analysis_for_Amazoncom_reviews.

[2] Xing Fang and Justin Zhan, “Sentiment Analysis using product review data”, Journal of Big Data, vol. 2, no.1, 16 June 2015. Available: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-015-0015-2

[3] Wanliang Tan, Xinyu Wang, and Xinyu Xu, “Sentiment Analysis for Amazon Reviews”, Available:http://cs229.stanford.edu/proj2018/report/122.pdf

[4] Callen Rain, “Analysis in Amazon Reviews Using Probabilistic Machine Learning”, 2012.Available: https://www.semanticscholar.org/paper/Analysis-in-Amazon-Reviews-Using-Probabilistic-Rain/f0afe9ea9d286248336ee9dc4e954aecde3475bb

[5] Nishit Shrestha, Fatma Nasoz, “Deep Learning Sentiment Analysis of Amazon.com Reviews and Ratings”,International Journal on Soft Computing, Artificial Intelligence and Applications (IJSCAI), Vol.8, No.1, February 2019. Available: https://arxiv.org/abs/1904.04096

[6]sklearn.metrics.precision_recall_fscore_support Available:https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_recall_fscore_support.html

[7] P. D. Turney, "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews," in Proceedings of the 40th annual meeting of the association for computational linguistics, Dec.2002, pp. 417-424. https://arxiv.org/abs/cs/0212032.

[8] Muhammad Ihsan Zul, Feoni Yulia, Dini Nurmalasari, “Social Media Sentiment Analysis Using K-means and Naïve Bayes Algorithm”, 2nd International Conference of Electrical Engineering and Informatics(Icon EEI 2018), October 2018.

[9] https://www.nltk.org/

[10]Availabe:http://jmcauley.ucsd.edu/data/amazon/

[11]Available:http://datameetsmedia.com/staging/3908/bag-of-words-tf-idf-explained/