Online Store Growth based on User Reviews using Big Data Analysis

[1]Premkumar T, [2]Karthikeyan S
[1]Student, [2]Assistant professor - Computer Science
Rathinam College of arts and science

***Abstract* - The process of researching, analyzing and revealing hidden data models from Big Data is known as Big Data Analytics. With this knowledge, we can reveal various crucial pieces of knowledge such as market trends, customer preferences and so on. Businesses that incorporate Big Data Analytics tend to reap excellent business benefits. One guaranteed method of getting enough customer reviews to make your product pages more persuasive for shoppers is to use a third party reviews provider. This is a useful way to build up a body of reliable reviews for product pages which could otherwise take some time. These reviews are also authenticated, so customers know that the person leaving the review has actually purchased the product in question. Drawbacks include the fact that such reviews tell other potential customers nothing about buying from your site in particular, as reviews are generally syndicated. Using this big data analysis, we can easily track and identify the original product vendor or customer. Many of the people have focus on product sales and how any peoples are satisfied with their purchase. The prediction of % may vary that depends on market competitions variables substitute of the product inflation and deflation in the retail market.**

***keywords* - Online Store Growth (OSG), OSG - Big Data Analysis (BDA)(OSGBD), User Reviews (UR), Online Store Growth with user Reviews (OSGUR-BDA)**

1. Introduction

In the past success full year, the world is focused on data generating and manipulations, after a few years, the Data is everywhere. We gather and share countless data at every moment, in fact, even our actions are generating data right now. In every moment of making the movie's that all screening, that is showing how we are going to utilize the data. Big data involved all industries like IoT, Artificial Intelligence and part of real-world data as seismic waves. So what is Big Data exactly, and how it is impacting the e-commerce industry? Big Data Analytics is the way of examining a voluminous amount of data in order to unearth hidden patterns, correlations, market trends, consumer preferences and other insights that can help businesses to modify accordingly. The data has many varieties, we have few storage systems to handle the minimal data for structured data, Other has unstructured data like document files; or streaming data from real-time sensors used in the IoT it could be possible to store like SQL, JSON, BSON. Since we are facing to handle huge manipulated data. So that we are using Raw formatted text document and some other files as a Comma-separated document (CSV) it will be more flexible and efficient to handle. Herefore, as Big Data is helping retailers to cater to the customers in a more personalized way via targeted advertising, product recommendations, and pricing, the technology is being increasingly preferred by the retailers.

In the UK, the supply chain Big Data analytics for retail is expected to grow significantly over the forecast period following the manufacturing and energy sector. Big Data plays an essential role in tracking the entire journey of a customer, from end to end. An average online shopper may not realize that every click is being monitored and that all purchases being made are captured from beginning to end. Dividing customers into different segments based on a combination of purchase patterns and demographic details makes it easier to target them better. Such categorizations are even more critical during campaigns and festive sales when companies invest heavily to attract new customers and retain the existing base.

2. Related study:

[1] The study was grounded in a literature review to identify and appraise the current knowledge on the definitional aspects, attributes, types and business value of BDA in e-commerce. In defining e-commerce, Kalakota and Winston (1997) focused on four perspectives: online buying and selling, technology driven business process, communication of information and customer service. However, this definition does not provide adequate focus on transaction cost and other aspects of e-commerce (e.g., B2B, B2G, and C2C etc.) Thus, illuminating these critical aspects, Frost and Strauss (2013) extends the definition focusing on buying and selling online, digital value creation, virtual market places and storefronts and new distribution intermediaries. However, this definition heavily focuses on e-marketing and fails to integrate other important e-business processes. As such, this study puts forward a more holistic definition of e-commerce in big data environment, which aims to achieve both transaction value (i.e., cost savings, improving productivity and efficiency) and strategic value (i.e., competitive advantages, firm performance) in digital markets by transforming production, inventory, innovation, risk, finance, knowledge, relationship and human resource management with the help of analytics driven insights (Wixom et al. 2013).

[2] The excitement surrounding BI&A and Big Data has arguably been generated primarily from the web and e-commerce communities. Significant market transformation has been accomplished by leading e-commerce vendors such Amazon and eBay

through their innovative and highly scalable ecommerce platforms and product recommender systems. Major Internet firms such as Google, Amazon, and Facebook continue to lead the development of web analytics, cloud computing, and social media platforms. The emergence of customer-generated Web 2.0 content on various forums, newsgroups, social media platforms, and crowd-sourcing systems offers another opportunity for researchers and practitioners to "listen" to the voice of the market from a vast number of business constituents that includes customers, employees, investors, and the media (Doan et al. 2011; O'Rielly 2005). Unlike traditional transaction records collected from various legacy systems of the 1980s, the data that e-commerce systems collect from the web are less structured and often contain rich customer opinion and behavioral information.
[3] E-Commerce data is explosively increasing on the scale of Terabytes (TB) to Petabytes (PB) on a daily basis due to the continuous increase in WWW traffic. For instance, to purchase an item on the web, a customer may explore many websites to have satisfactory E-Commerce transaction which not only provides the high quality branded product but also at best possible discounted price or maximum wallet cash back. Hence, as a result, many of the online shopping portals getting big data on daily basis like Amazon or PayTm Mall android based E-Commerce portal, which handles around a million customer transaction logs on a regular basis, resulting into many TB of data generated on a daily basis. This excessive online generated data is commonly known as 'Big Data' with emphasis on high values of various popular V's, i.e., Value, Velocity, Variety, Veracity, and Volume. Big data may be defined as a collection of a huge number of data sets, the speed of incoming data before processing, outgoing data after processing and range of data sources are beyond the capabilities of conventional relational databases systems for processing and management.

3. User Summary

Since, continues growth of the online market, users are very polite and more active to log their own reviews to identify rich products, it helps to increase the selling capacity of the product and users can easily identify the reliability of the products. Users are sharing their own opinion on the products, this particular information could be defined as User Generated Content (UGC). The UGC "refers to media content created by users to share information and/or opinions with other users". In comparison with the traditional methods of advertisement like television and newspaper advertisements, electronic UGC is perceived by potential customers to be more reliable, balanced, and neutral than those provided through private channels. [1] We offered the users to freely enter into GUI. That is more efficient for the users to identify the perfect place to log their valuable thoughts. Also user can share the images to identify how much they are satisfied with the received product from online market. Big Data plays an important role in tracking the entire journey of a customer, from entry to exit. An average online shopper may not realise that every click is being monitored and that all purchases being made are captured from beginning to end. Dividing customers into different segments based on a combination of purchase patterns and demographic details makes it easier to target them better. Such categorisations are even more critical during campaigns and festive sales, when companies invest heavily to attract new customers and retain existing base.

Search and purchase information of buyers three or four months prior to the sale yields a deep understanding of current trends and actual demand. This is extremely useful as sellers can then stock their inventory accordingly and optimise shelf space. For example, based on historical data, Electronics, Automobiles and Apparel & Clothing are some of the categories that see the highest purchase volumes during festivals; hence, demand forecasting and inventory stocking of merchandise before the sale period can be utilised to give optimum discounts, thus maximising revenue.

4. Positive and Negative Valence of User Reviews

User reviews can completely impact purchase inspections. In the past decade, comprehensive analysis in the field of user reviews has concentrated on product levels (e.g., hedonic, utilitarian) and commodity sales. However, analysis of how the components of user reviews (valence, volume, and variance) impact attitudes toward brands is rare, even though brands are among the most valuable corporate assets and corporations use online selling largely to build brand support. Thus, this paper offers a conceptual model that closely examines the relationship between the components of online reviews and brand positions. The model offers a better belief of the influence of contextual factors on brand attitudes inside online conversations. In line with prior analysis, the study imagines volume and variance as moderators of valence. Moreover, the proposed conceptual model integrates brand type (functional, passionate, typical, and lifestyle) and the source of survey (stranger or acquaintance) as possible moderators. Conceptual insights, along with managerial assumptions for online marketing directors, are provided. The earlier investigations have shown an incongruous connection among the valence (positive or negative) of online customer reviews and consumer decision making. With the convenience/diagnosticity entrance, this study attempts to explain this inconsistency by investigating customer expertise as a moderator. Our results from a user to user experiment design indicate that the result of online surveys valence is supervised by customer expertise: The impact contrast between negative reviews and positive reviews is greater for consumers with low expertise than for those with high expertise. Our study adds to the research relevant to the e-WOM effect. And we also provide managerial suggestions for e-marketers.

5. Volume of Consumers Reviews

The expanding popularity of online product review discussions suggests the community of models and metrics that allow firms to harness these new sources of knowledge for decision support. Our work provides in this direction by introducing a novel family of diffusion models that capture some of the unique characters of the online industry and testing their products in the context of making happy to users. We show that the addition of online product review metrics to a benchmark model that includes pre-release marketing, theatre availability and professional critic reviews substantially increases its future estimation; the future estimation of our best model outperforms that of several previously published models. In addition to its contributions in diffusion theory, our study reconciles some inconsistencies among previous studies with respect to what online review metrics

are statistically significant in planning for the future. The future of sales growth could be made by user satisfaction, that the impact or growth must be calculated by end-users review. Consumer product awareness caused by online review volume leads to higher sales. Both theoretical and practical evidence supports the predicting power of online review volume on sales. However, it is found that online review volume is more effective in predicting the sales of experience products than search products. For experience goods, online review volume has a strong prediction power because it can reflect outward cues such as product popularity. However, for all products where consumers can experience the actual product's attributes, online review valance plays an even more important role in predicting sales than online review volume. Based on previous literature, it is clear that online review volume is an important predictor of product sales. Thus we decided to include volume (along with valence) of online reviews as variables in predicting.

6. Impact on Negative reviews

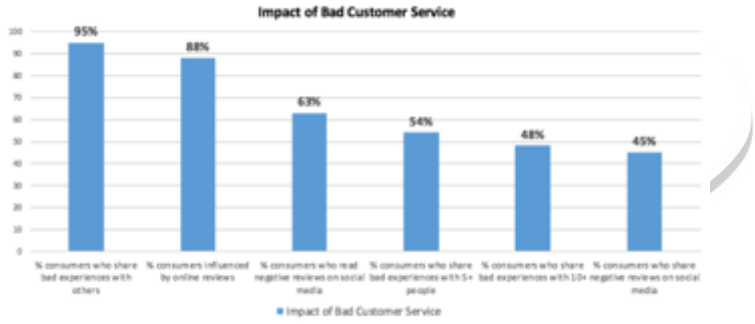
Due to the negative reviews of some customers the whole product is got spoiled. Even If the product is survey good at its selling position, by reading those negative comments and reviews given by one or two customers the total selling of the production is affected.

Negative reviews has the power to turn our profits into loss and these can affect the goodwill of the retailers. Also this negative reviews helps the competitors to race their product in the market even if they are not qualified than the other one. Look for patterns. If you find that customers who come into your restaurant at dinner are having a markedly different experience than lunchtime, it's worth looking into your staff to see what might be the problem. If you find that all of your products are going over beautifully, but there's a regular complaint about one function or detail, test it out yourself to see what could be improved.

Suggest that they leave reviews on the sites that matter most to you, whether that be Face book, Amazon, Yelp, or another site, and consider offering an incentive to customers who do choose to leave a review, whether it be positive or negative.

It's demolishes the reputation of the retailers and also breaks the potential customers trust towards the business. It creates reputation risk. Due to this negative reviews retailers face the problems such as when there is a negative review for product they suffer to repair it too hard. They need to create the rebound trust to the re-branding the product process.

Figure 1. Sales impact given by the user reviews.



6.1 NLP-Natural language processing:

NLP is a branch of data science that consists of systematic processes for analysing, understanding, and deriving information from the text data in a smart and efficient manner. By utilizing NLP and its components, one can organize the massive chunks of text data, perform numerous automated tasks and solve a wide range of problems such as – automatic summarization, machine translation, named entity recognition, relationship extraction, sentiment analysis, speech recognition, and topic segmentation etc. Also (NLP) is an area of computer science and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyse large amounts of natural language data. It is the branch of machine learning which is about analysing any text and handling predictive analysis.

7. Sentiment Analysis with User Reviews

We are introducing the algorithm as Sentiment text analysis, it represents the best businesses understand the sentiment of their customers—what people are saying, how they're saying it, and what they mean. Customer sentiment can be found in tweets, comments, reviews, or other places where people mention your brand. Sentiment Analysis is the domain of understanding these emotions with software, and it's a must-understand for developers and business leaders in a modern workplace.

As with many other fields, advances in deep learning have brought sentiment analysis into the foreground of cutting-edge algorithms. Today we use natural language processing, statistics, and text analysis to extract, and identify the sentiment of text into positive, negative, or neutral categories.

- *Business*: In marketing field companies use it to develop their strategies, to understand customers' feelings towards products or brand, how people respond to their campaigns or product launches and why consumers don't buy some products.

- *Politics*: In political field, it is used to keep track of political view, to detect consistency and inconsistency between statements and actions at the government level. It can be used to predict election results as well!

- *Public Actions*: Sentiment analysis also is used to monitor and analyses social phenomena, for the spotting of potentially dangerous situations and determining the general mood of the blogosphere.

8. Scikit-learn

Scikit-learn provides a range of supervised and unsupervised learning algorithms via a consistent interface in Python. It is licensed under a permissive simplified BSD license and is distributed under many Linux distributions, encouraging academic and commercial use.

The library is built upon the SciPy (Scientific Python) that must be installed before you can use Scikit-learn. This stack that includes:

*NumPy***:** Base n-dimensional array package.
*SciPy***:** Fundamental library for scientific computing.
Matplotlib: Comprehensive 2D/3D plotting.
*IPython***:** Enhanced interactive console.
*Sympy***:** Symbolic mathematics.
*Pandas***:** Data structures and analysis.

Extensions or modules for SciPy care conventionally named SciKits. As such, the module provides learning algorithms and is named Scikit-learn.

The vision for the library is a level of robustness and support required for use in production systems. This means a deep focus on concerns such as easy of use, code quality, collaboration, documentation and performance.

9. Data Analysis in NumPy

NumPy is a Linear Algebra Library for Python and the purpose it's so vital that all libraries in the PyData Ecological system rely on NumPy as the main building block.

It's highly suggested to install Python using Anaconda disposal to make sure all underlying dependencies (such as Linear Algebra libraries) all sync up with the use of a conda install. Numpy arrays are the main reason we use Numpy and they come in two flavors

Vectors: (one dimensional arrays)
*Matrices***:** (two dimensional arrays)

Python is frequently being used as a scientific language. Matrix and vector directions are extremely important for scientific calculations. Both NumPy and Pandas have emerged to be essential libraries for any scientific computation, including machine learning, in python due to their intuitive syntax and high-performance matrix computation capabilities.

NumPy - stands for 'Numerical Python' or 'Numeric Python'. It is an open-source module of Python which provides fast mathematical calculation on arrays and matrices. Since arrays and matrices are an essential part of the Machine Learning ecosystem, NumPy along with Machine Learning modules like Scikit-learn, Pandas, Matplotlib, Tensor Flow, etc. complete the Python Machine Learning Ecosystem and it provides the essential multi-dimensional array-oriented computing functionalities meant for high-level mathematical functions and scientific computation In python, a vector can be represented in many ways, the simplest being a regular python list of numbers. Since Machine Learning requires lots of scientific calculations, it is much better to use NumPy's and array, which provides a lot of comfortable and optimized implementations of essential scientific operations on vectors.

Usages:

```
plot_size = plt.rcParams["figure.figsize"]
print(plot_size[0]) print(plot_size[1])
plot_size[0] = 8
plot_size[1] = 6
plt.rcParams ["figure.figsize"] = plot_size
```

```
airline_tweets.airline.value_counts().plot(kind='pie', autopct='%1.0f%%')

airline_tweets.airline_sentiment.value_counts().plot(kind='pie', autopct='%1.0f%%', colors=["red",
"yellow", "green"])

airline_sentiment = airline_tweets.groupby
(['airline',airline_sentiment']).airline_sentiment.count().unstack() airline_sentiment.plot(kind='bar')
```

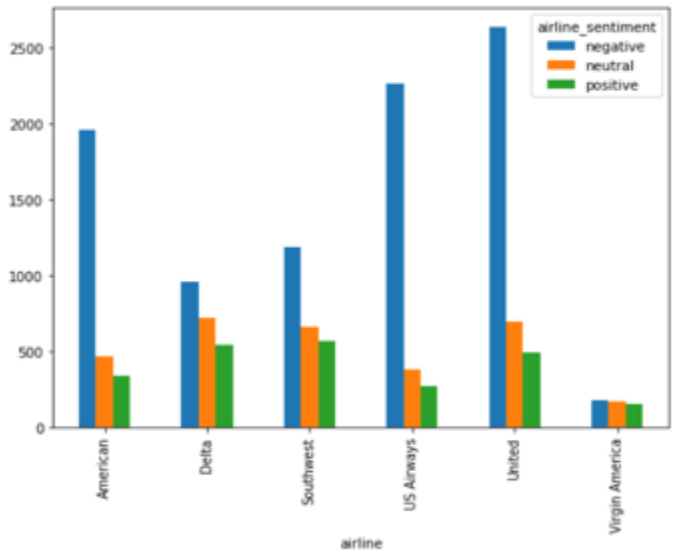


Figure 2. Sample analysis result.

10. Data Cleaning

Clean the Users Review contain many slang words and punctuation marks. We need to clean our reviews before they can be used for training the machine learning model. However, before cleaning the review, let's divide our dataset into feature and label sets. Our feature set will consist of reviews only. If we look at our dataset, the 11th column contains the tweet text. Note that the index of the column will be 10 since pandas columns follow zero-based indexing scheme where the first column is called 0th column. Our label set will consist of the sentiment of the reviews that we have to predict. The sentiment of the reviews is in the second column (index) 1). To create a feature and a label set, we can use the"iloc" method off the pandas data frame.

***Applying the Regex*:**

Using a regex pattern library to replace the symbols and unwanted text to space.

processed feature = re.sub(r'\W', ' ', str(features[sentence]))

processed feature= re.sub(r'\s+ [a-zA-Z]\s+', ' ', processed feature)

processed_feature = re.sub(r'\^[a-zA-Z]\s+', ' ', processed_feature)

single space processed_feature = re.sub(r'\s+', ' ', processed_feature, flags=re.I)

processed_feature = re.sub(r'^b\s+', '', processed_feature)

processed_feature = processed_feature.lower() processed_features.append(processed_feature)

11. Training the Model

Once data is split into training and test set, machine learning algorithms can be used to learn from the training data. You can use any machine learning algorithm. However, we will use the Random Forest algorithm, owing to its ability to act upon non-normalized data.

The sklearn.ensemble module contains the Random Forest Classifier class that can be used to train the machine learning model using the random forest algorithm. To do so, we need to call the fit method on the RandomForestClassifier class and pass it our training features and labels, as parameters.

Work Out:

```
from sklearn.ensemble import RandomForestClassifier

text_classifier = RandomForestClassifier (n_estimators=200, random_state=0)

text_classifier.fit(X_train, y_train)

predictions = text_classifier.predict(X_test)
```

```
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score

print(confusion_matrix(y_test,predictions))

print(classification_report(y_test,predictions))

print(accuracy_score(y_test, predictions))
```

12. Conclusion

With the Help of Big Data Analysis (BDA), online marketer can have the extra hand to sort out the user satisfaction on fair products, the method will help to remove the products, rather than the user purchase. Also it helps to improve purchase growth to the dependent market. Apart from the module implementation, the retail market should have enough sellers to sell similar products. Leading e-commerce firms such as Amazon, eBay and Flipkart have already embraced BDA and experienced enormous growth. Through its systematic review and creation of taxonomy of the key aspects of BDA, this study presents a useful starting point for the application of BDA in emerging retail market research. The study presents an approach for encapsulating all the best practices that build and shape BDA capabilities. In addition, the study reflects that once BDA and its scope are well defined; distinctive characteristics and types of big data are well understood, and challenges are properly addressed, the BDA application will maximize business value through facilitating the pervasive usage and speedy delivery of insights across organizations.

13. References

[1] Shahriar Akter and Samuel Fosso Wamba, Big data analytics in E-commerce: a systematic review and agenda for future research, May 2016.
[2] Hsinchun Chen, Business intelligence and analytics: from big data to big impact, Eller College of Management, University of Arizona, Tucson, May 2018.
[3] Roger H. L. Chiang, Business intelligence and analytics: from big data to big impact, Carl H. Lindner College of Business, University of Cincinnati, May 2018.
[4] Veda C. Storey, Business intelligence and analytics: from big data to big impact, Mack Robinson College of Business, Georgia State University, Atlanta, May 2018.
[5] Dheeraj Malhotra, An intelligent approach to design of E-Commerce metasearch and ranking system using next-generation big data analytics, March 2018.
[6] Durjoy Patranabish, Senior Vice President (Analytics) Blueocean Market, Intelligence, 18th Nov 2016.
[7] Agarwal, R., & Dhar, V. Editorial—big data, data science, and analytics: the opportunity and challenge for IS research. Information Systems Research, 2014.
[8] Duarte Stock - NOVA Marketing Insights, Mar 15, 2018.
[9] Beulke, D., Big Data Impacts Data Management: The 5 Vs of big data, 2011.
[10] Algorithmia - How To Perform Sentiment Analysis With Twitter Data, MARCH 27, 2018.
[11] Journal of Interactive Marketing, Volume 25 - panelLee-YunPan and Jyh-ShenChiou, May 2011.
[12] Michael Svilar, Arnab Chakraborty and Athina Kanioura, Big data analytics in marketing, managing director
[13] Tom, what is Sentiment Analysis and How to Do It Yourself, Feb 2018.
[14] Classification in Python with Scikit-Learn and Pandas - Steven Hurwitt, December 12, 2018.
[15] Python for NLP: Sentiment Analysis with Scikit-Learn - Usman Malik, April 03, 2019.
[16] GAGAN MEHRA, Uses of Big Data for Online Retailers, MARCH 27, 2013.
[17] Christy Bohrer, DATAECONOMY, Ways big data analytics will impact e-commerce, 2018.
[18] Vinu Kumar, Digital Marketing, the Impact of Big Data Analytics in the Retail Industry, May 2018.