IJARIIT INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 7, Issue 4 - V7I4-1750)
Available online at: https://www.ijariit.com

Domain-agnostic Customer Sentiment Analysis Platform for MSMEs in E-Commerce

Veer Kejriwal
veer.kejriwal11@gmail.com
Jamnabai Narsee International School, Mumbai, Maharashtra

Aqsa Temrikar
aqsatemrikar@gmail.com
JBCN International School, Mumbai, Maharashtra

ABSTRACT

This paper aims to tackle the problem of MSMEs being unable to gauge customer sentiment through text-based reviews on e-commerce platforms. The paper suggests a front-end algorithmic solution to the problem which is simple yet impactful in that it allows MSMEs to understand consumer behavior and psyche with minimal or no resources. The platform, which can be adopted by any e-commerce enterprise, uses web scraping frameworks and natural language processing-based sentiment analysis to extract, demystify, and present various metrics and actionable insights essential to analyze underlying sentiment. The proposed end-to-end solution aids MSMEs in fostering sustainable customer relationships and boosting business growth and prosperity.

Keywords: *MSME, e-commerce, Web scraping, Sentiment Analysis, Natural Language Processing*

1. INTRODUCTION

Micro, small, and medium enterprises are increasingly relying on user feedback to drive their understanding of demographics in order to make various business decisions to model design thinking, product development, sales, and marketing. This has been an integral component of their business function for years and an extremely important factor in business success.

However, in the new age of digitalization, and a wide scale shift to e-commerce driven sales, it has become tougher for MSMEs – with limited expertise and resources – to understand customers through text-based reviews published on e-commerce platforms. As compared to their bigger competitors, MSMEs have limited financial resources for branding - which means that their only way to build a brand is by sustaining customer relationships.

This platform aims to bridge the gap between MSMEs and customers – and lead to enhanced customer understanding and behavioral patterns. The 3-step plug and play scrape-filter-visualize process of the platform fulfils the needs of these businesses, which can use this despite limited expertise, to gauge sentiment.

The platform is domain-agnostic. This means that the platform can be adopted by e-commerce enterprises in any industry – as the nature of the products sold, customer feedback, or e-commerce platform does not matter. This expands the platform's reach and enables wider adoption by any enterprise.

Simultaneously, it can also be adopted by customers. Customer reviews have become a dominant factor driving purchase considerations when shopping online. However, it is cumbersome for the user to read through and analyze a large volume of reviews on e-commerce platforms.

2. METHODOLOGY

The solution needs to be user friendly, allowing the user to extrapolate accurate, actionable insights from a concise and easily interpretable form. Additionally, it is imperative that these insights should be derived from an unbiased point of view. Before designing the platform as per these requirements, it is essential to understand the existing ways in which user's interpret customer sentiment.

E-commerce websites have multiple pages under the review section of each product, where the metrics of each review is shown (Reviewer name, review date, review comment, star rating, etc.). The data is presented in a top-to-bottom, or left-to-right format as opposed to a tabular or graphical one. This makes it time consuming for enterprises to skim through reviews and get an overall idea of sentiment. The user can only sort reviews by date, and this limits analysis of sentimental trends.

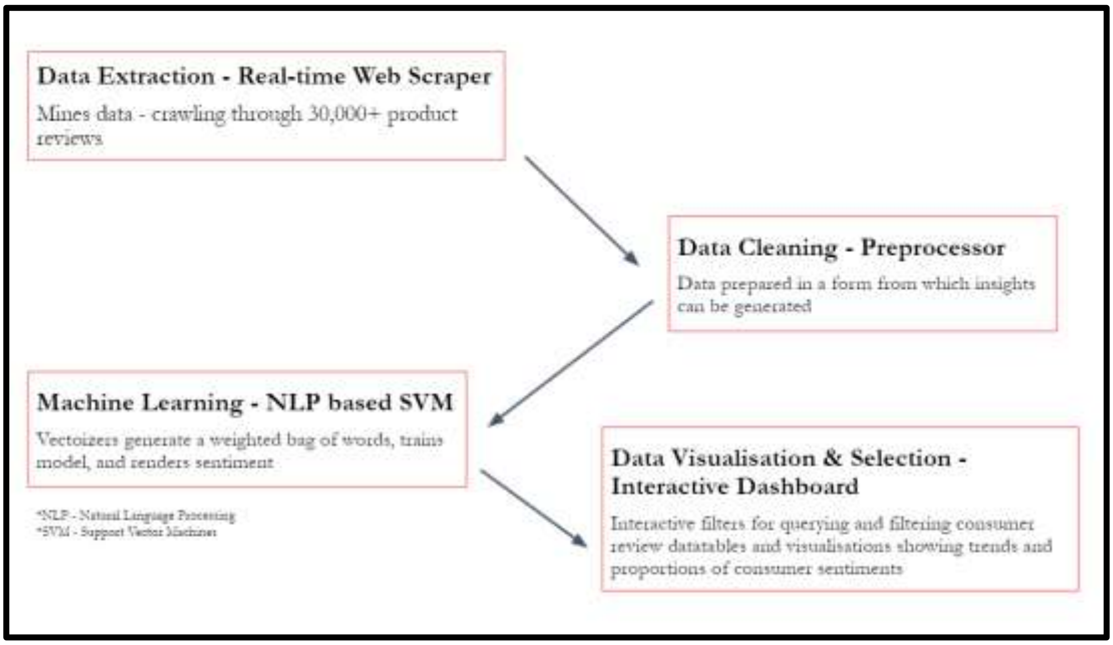

Data Extraction - Real-time Web Scraper
Mines data - crawling through 30,000+ product reviews
Data Cleaning - Preprocessor
Data prepared in a form from which insights can be generated
Machine Learning - NLP based SVM
Vectoizers generate a weighted bag of words, trains model, and renders sentiment
*NLP - Natural Language Processing
*SVM - Support Vector Machines
Data Visualisation & Selection - Interactive Dashboard
Interactive filters for querying and filtering consumer review datatables and visualisations showing trends and proportions of consumer sentiments

Figure 1: Key Components of the Platform

It is apparent that interpreting multiple reviews on e-commerce websites is a lengthy, laborious process and any solution that is designed must enable easier extrapolation of insights. Additionally, it is important for the solution to enable a historical analysis of customer sentiment.

The implementation suggested in this paper consists of a domain-agnostic web scraper that collates all customer review metrics in a tabular format which then passes the data to a machine learning model in order to extract unbiased sentiment from the raw data. The implementation consists of the following primary components:

1. **Real time Web Scraper:** The scraper uses the Scrapy framework in Python to enable navigation of web-pages. During operation, it uses pagination (Chauhan *Scraping Amazon reviews using Scrapy in Python - Datahut*) to iterate through each product in the category selected by the user. Further navigating through all the pages of customer reviews for every product it collates the data row-wise in a .csv file.
2. **Data Processor:** This component is connected to the web scraper - from where it takes the .csv file as an input. The processor is a pipeline which sequentially standardizes the raw, unprocessed data. This is an important stage as NLP requires the input of vectorized, processed forms of data labels. By integration of Pandas and NumPy in Python, the data is prepared into a format which is easy to read for the user and interpretable for the machine learning model.
3. **NLP and Machine learning based Sentiment Analysis pipeline:**

a. **Noise Removal -** The first step required the removal of irrelevant data which interferes with lexical analysis. For example, each review would be passed in order to detect and remove digits, header, footer, HTML, XML, and markup data.

b. **Tokenization -** A fundamental step in NLP is to break down sentences into bags of words (lexical analysis of data)**.**

c. **Vectorization -** The sentiment of each review could only be deduced after each word in the review was assigned a weight. That is, the degree of how positive, negative, or neutral the text is. The bag of words frequency uses word occurrence in order to assign the importance of each word in contributing to the overall sentiment (Tran *NLP sentiment analysis for beginners.*) of a review. This was done by mapping each word to a corresponding vector of real numbers.

Once these NLP techniques normalized the data, it could be used in order to train and test machine learning models. A range of supervised machine learning models from sklearn were tested to evaluate which model was the best fit for the chosen dataset.

A supervised algorithm is that which is given an input-output pair (train set) and based on that, maps an input pair to an output pair. In the case of sentiment detection for - positive, negative, and neutral - tri classification (*IBM data science*) is used where the algorithm classifies review cases by finding a separator by mapping the inputs and output labels to a high (n) dimensional feature space.

This works in sentiment analysis as the data here is not linearly separable and hence this model transforms the data in a way that a separator could be drawn as a hyperplane.

The models used were: Naive bayes, Logistic regression, Support vector machines, Decision tree classifier, and random forest classifier. When testing the model, the test set reviews were used to generate a sentiment prediction which was then compared with the actual sentiment. Although decision tree and random forest classifiers had a higher train set accuracy, the test set accuracy showed the greatest fall and hence showed that the model was over fitted with the data trained and hence Support vector machines was a better model to be used. It had the best performance with the test dataset and did not, relatively, show a wide gap between train and test accuracy as shown in the table below:

Table 1: Accuracy results for the testing machine learning models

Model Tested	**Train Accuracy**	**Test Accuracy**
Naive Bayes	88.80%	93.40%
Logistic Regression	89.90%	93.6%
Support Vector Machines	96.20%	93.60%
Decision Tree Classifier	99.90%	89.00%
Random Forest Classifier	99.90%	92.50%

These are the iterations made to the Support Vector Machines (SVM) machine learning model:

1. Initially, the model was trained on a 70:30 Train test split where 70% of reviews were being used for training the model and 30% were used to make predictions and compare it with the actual labels to see how well the model did. In this split, Support vector machines gave 85.7% test accuracy.
   Solution adopted: To enhance accuracy, a stratified train-test split with a ratio of 80:20 was used so that the model could be trained better to avoid any underfitting or overfitting. A stratified split was used instead of a normal one because more of the dataset reviews are positive and hence a stratified split was needed to ensure that the classifier is not trained on imbalanced data. It is done to ensure that the model is trained on the same proportion of class examples as the original dataset.
2. There was an imbalance in the dataset as it contained more positive reviews than negative ones.
   Solution Adopted: Added 3000 more negative reviews only so that the imbalance could be further eliminated. The data split ratio was also changed to 78:22 from 80:20. These changes helped increase the SVM test accuracy to 91.2%.
3. The model was unable to classify edge case reviews, such as sarcastic and ironic ones.
   Solution Adopted: Used a specialized dataset containing 1000 edge case reviews for training. This gave an accuracy of 91.7%.
4. For fine tuning the model as the last step, a Grid Search optimization was used to find the best parameters on a grid of possible values for the SVC classifier pipeline, parameters and CPU core maximization, instead of tweaking the parameters of various components of the chain. This was then fit to the y training data set which rendered a test accuracy of 92.3%.

The final performance of the model is as follows:

Table 2: The final train-test ratio for the model and respective accuracy

	Initial	Final
Train set (Review, sentiment)	(24,238,24,238)	(31701, 31701)
Test set (Review, sentiment)	(10,389,10,389)	(6926, 6926)
Train: Test Split Ratio	70:30	78:22
Test Set Accuracy	85.7%	92.30%

4. **Interactive dashboard:** This is the front-end of the platform which allows the user to view insights according to tailored requirements (Koehrsen *Interactive controls for Jupyter notebooks*). A detailed operation of this is discussed in the “Operation of platform” section.

3. ALGORITHM

Web-scraper -

1. Run Web Scraper for Marketplace URL,
2. Identify and target the “Brand” (Brand-ID) selected for data collection,
3. Identify and target the “Product” (Product-ID) selected for data collection,
4. Collect data (for (3)) –
a. Product name
b. Username (of user posting the review),
c. Date of posting of the review,
d. Rating of review,
e. Comment of review,
f. URL of review
5. Save data to database,
6. Repeat Steps (4) and (5) until all reviews for the (3) are exhausted

User model selection -

1. Selects “Model”,
2. Model name passed as parameter to main dataframe,
3. Dataframe filtered based on model selected and returned

Sentiment quantification -
1. Train the Machine Learning model the dataset, evaluate its accuracy with a pre-defined dataset; repeat until the model achieves a high degree of accuracy,
2. Pass the review comment from each row in the 'Review' column as a parameter to the machine learning model,
3. Create a new column called 'Sentiment',
4. Return sentiment of 'Positive', 'Negative' or 'Neutral' depending on prediction
5. Save the updated data frame as a CSV file

Cleaning for Data and time -
1. Store full date in the Store "Date Published", "Day", "Month", and "Year". Column
2. Store day number in the "Day" column
3. Store month in the "Month" column
4. Store the year in the "Year" column

Time period range generation -
1. Find MIN year in dataframe
2. Find MIN month from MIN year
3. Find MAX year in dataframe]
4. Find MAX month from MAX year
5. Generate a dictionary with all year and month names between the min and max year
6. Return dictionary

Dataframe filtering -
1. Selects "Year"
2. Dataframe filtered for year
3. Selects "Month"
4. Dataframe filtered for month
5. Dataframe displayed as a preview
6. Clicks button "Export to CSV"
7. Selects destination
8. Uses file path as a parameter to the "save" function
9. Saves dataframe in the directory

Visualisation previewing -
1. Select "Visualisation type" using the radio button. Example taken: Pie chart selected
2. Pie month function called
3. Selects "Year"
4. Selects "Month"
5. "Year" and "Month" values passed to pie chart plotting function
6. Dataframe filtered as per Year and Month
7. Visualisation plotted using plotting function
8. Clicks button "Download PNG"
9. Selects destination

10.Uses file path as a parameter to the "save" function

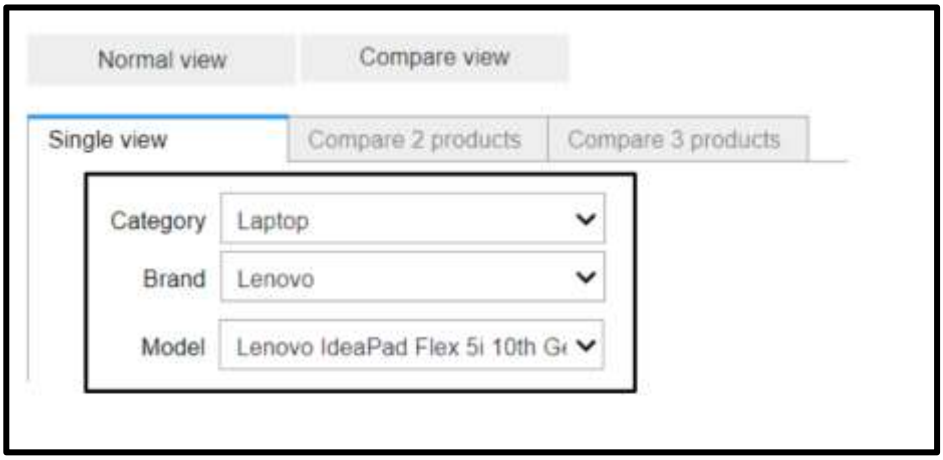
11.Saves dataframe in the directory

Note: The algorithm shown in this section is an extremely simplified version of the source code of the platform, and there are much more complex data preprocessing and structural changes which take place before the final output can be generated.

Operation of platform

The platform's dashboard was designed such that the interface is intuitive and easy for the user to navigate. The following is the procedure to gain customer sentiment insights on a product using the front-end interface delineated in this paper:

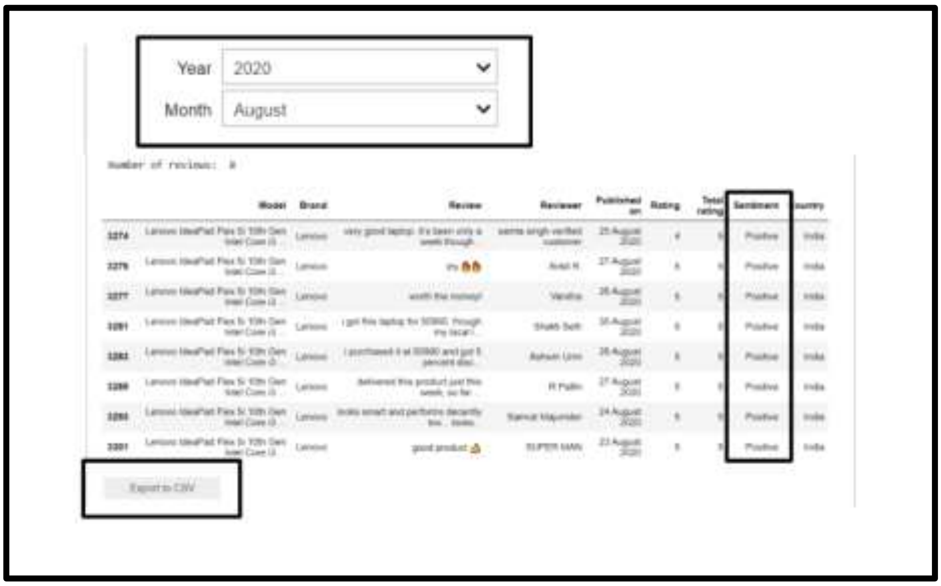
● The user selects the product category they wish to explore - depending on which they choose a brand and respective model.



Category: Laptop
Brand: Lenovo
Model: Lenovo IdeaPad Flex 5i 10th G

Figure 2: Selection of product category, brand, and model

- They filter the results scraped (dataframe) by selecting a specific year and month as per their preference.



Year: 2020

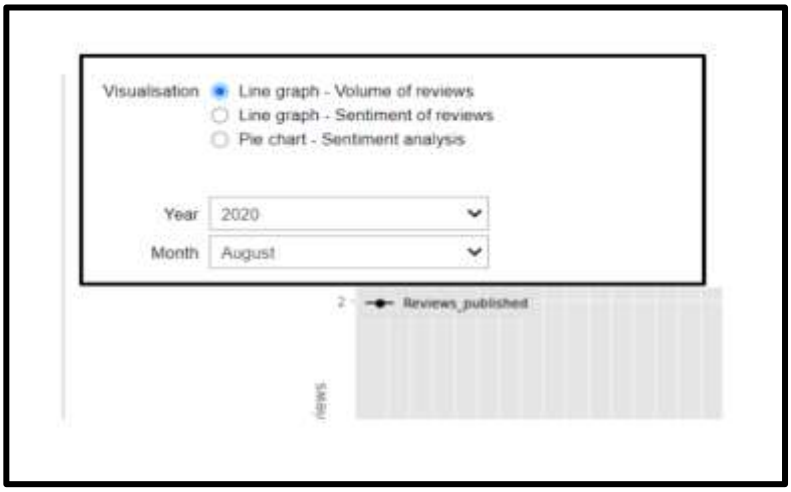
Month: August

Number of reviews: 8

	Model	Brand	Review	Reviewer	Published on	Rating	Total rating	Sentiment	Country
3274	Lenovo IdeaPad Flex 5i 10th Gen Intel Core i3 ...	Lenovo	very good laptop. it's been only a week though...	samta singh verified customer	25 August 2020	4	5	Positive	India
3276	Lenovo IdeaPad Flex 5i 10th Gen Intel Core i3 ...	Lenovo	[illegible]	Ankit K	27 August 2020	5	5	Positive	India
3277	Lenovo IdeaPad Flex 5i 10th Gen Intel Core i3 ...	Lenovo	worth the money!	Varsha	28 August 2020	5	5	Positive	India
3281	Lenovo IdeaPad Flex 5i 10th Gen Intel Core i3 ...	Lenovo	i got this laptop for 50990, though my local i...	Shakti Seth	30 August 2020	5	5	Positive	India
3282	Lenovo IdeaPad Flex 5i 10th Gen Intel Core i3 ...	Lenovo	i purchased it at 50990 and got 5 percent disc...	Ashwin Unni	26 August 2020	5	5	Positive	India
3288	Lenovo IdeaPad Flex 5i 10th Gen Intel Core i3 ...	Lenovo	delivered this product just this week, so far ...	R Palin	27 August 2020	5	5	Positive	India
3295	Lenovo IdeaPad Flex 5i 10th Gen Intel Core i3 ...	Lenovo	looks smart and performs decently too... looks	Samrat Majumder	24 August 2020	5	5	Positive	India
3301	Lenovo IdeaPad Flex 5i 10th Gen Intel Core i3 ...	Lenovo	good product 👍	SUPER MAN	23 August 2020	5	5	Positive	India

Figure 3: Filtering of data frame as per year and month

- Once they view the results displayed by the dataframe, they scroll down to the visualization selection in order to see a historical record of customer sentiment in a graphical manner. To do this, they select the type of visualization from the radio button panel.
- They select the timeframe of the visualization specific to year and month as per their preference.



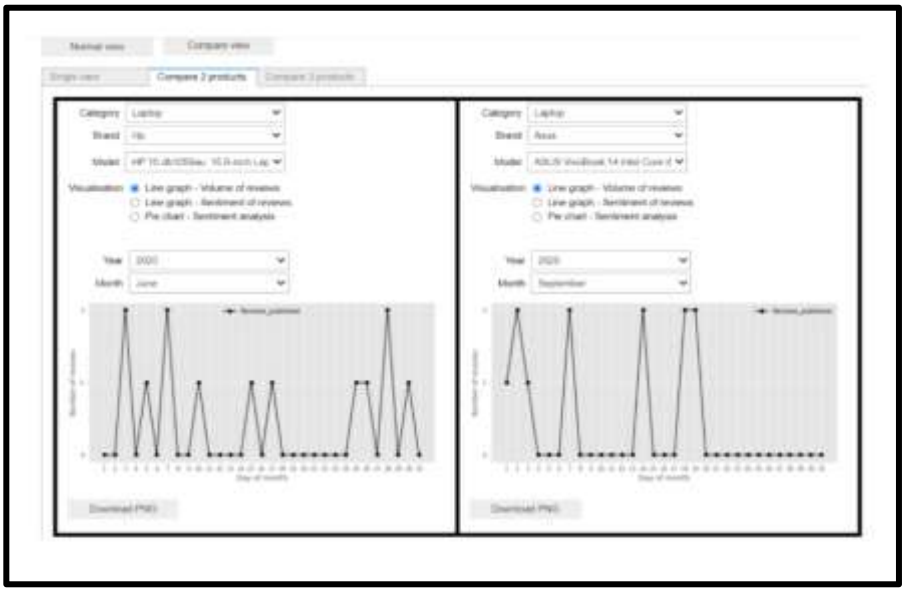
Visualisation
● Line graph - Volume of reviews
○ Line graph - Sentiment of reviews
○ Pie chart - Sentiment analysis

Year: 2020

Month: August

Figure 4: Selecting type of visualization and time period as per year and month

- Users can switch to the “Compare 2 products” tab in order to compare insights via visualizations for 2 different products.



Category: Laptop

Brand: Hp

Model: HP 15 db1069au 15.6-inch Lap

Visualisation
● Line graph - Volume of reviews
○ Line graph - Sentiment of reviews
○ Pie chart - Sentiment analysis

Year: 2020

Month: June

Category: Laptop

Brand: Asus

Model: ASUS VivoBook 14 Intel Core i

Visualisation
● Line graph - Volume of reviews
○ Line graph - Sentiment of reviews
○ Pie chart - Sentiment analysis

Year: 2020

Month: September

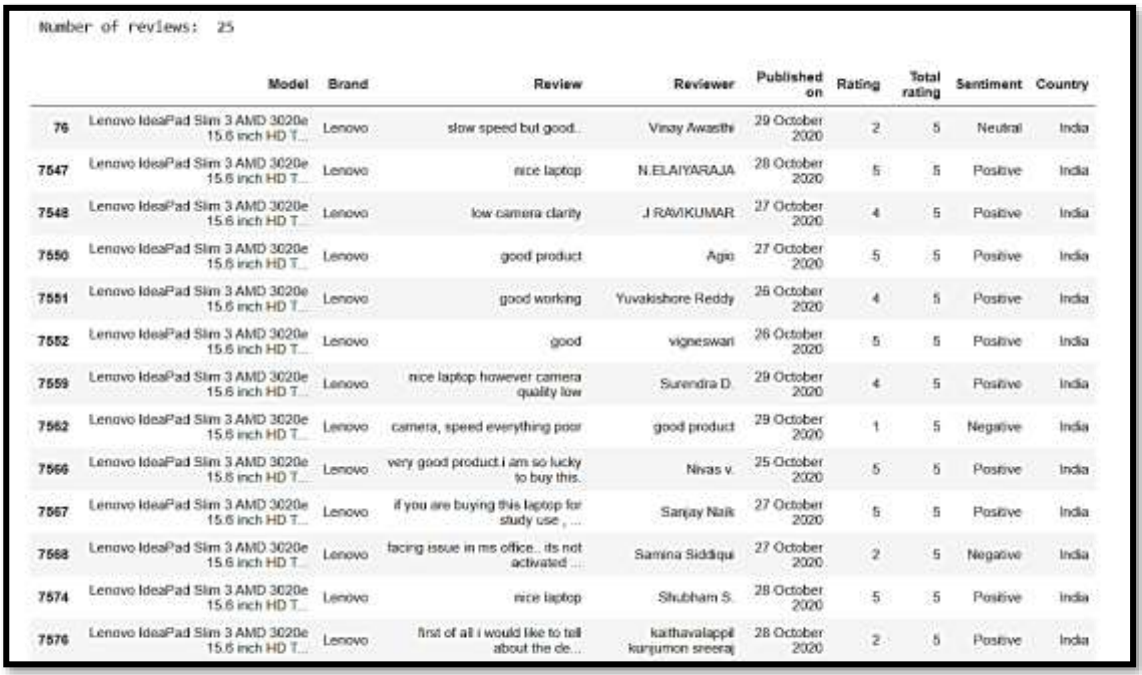
Figure 5: Dual compare view

- As per need, they can download all generated insights from the dashboard to their desktop directory.

4. RESULTS

1. Web scraping - Consumer Review Metrics & Sentiment mining

- This result shows the details of each part of the review scraped. Users can use this dataframe for further, deeper analysis after gaining insights from visualizations. The tabular format enables quicker interpretation and the dataframe can be filtered as per preference.
-



```
Number of reviews:  25
```

	Model	Brand	Review	Reviewer	Published on	Rating	Total rating	Sentiment	Country
76	Lenovo IdeaPad Slim 3 AMD 3020e 15.6 inch HD T...	Lenovo	slow speed but good..	Vinay Awasthi	29 October 2020	2	5	Neutral	India
7547	Lenovo IdeaPad Slim 3 AMD 3020e 15.6 inch HD T...	Lenovo	nice laptop	N.ELAIYARAJA	28 October 2020	5	5	Positive	India
7548	Lenovo IdeaPad Slim 3 AMD 3020e 15.6 inch HD T...	Lenovo	low camera clarity	J RAVIKUMAR	27 October 2020	4	5	Positive	India
7550	Lenovo IdeaPad Slim 3 AMD 3020e 15.6 inch HD T...	Lenovo	good product	Agio	27 October 2020	5	5	Positive	India
7551	Lenovo IdeaPad Slim 3 AMD 3020e 15.6 inch HD T...	Lenovo	good working	Yuvakishore Reddy	26 October 2020	4	5	Positive	India
7552	Lenovo IdeaPad Slim 3 AMD 3020e 15.6 inch HD T...	Lenovo	good	vigneswari	26 October 2020	5	5	Positive	India
7559	Lenovo IdeaPad Slim 3 AMD 3020e 15.6 inch HD T...	Lenovo	nice laptop however camera quality low	Surendra D.	29 October 2020	4	5	Positive	India
7562	Lenovo IdeaPad Slim 3 AMD 3020e 15.6 inch HD T...	Lenovo	camera, speed everything poor	good product	29 October 2020	1	5	Negative	India
7566	Lenovo IdeaPad Slim 3 AMD 3020e 15.6 inch HD T...	Lenovo	very good product i am so lucky to buy this.	Nivas v.	25 October 2020	5	5	Positive	India
7567	Lenovo IdeaPad Slim 3 AMD 3020e 15.6 inch HD T...	Lenovo	if you are buying this laptop for study use , ...	Sanjay Naik	27 October 2020	5	5	Positive	India
7568	Lenovo IdeaPad Slim 3 AMD 3020e 15.6 inch HD T...	Lenovo	facing issue in ms office.. its not activated ...	Samina Siddiqui	27 October 2020	2	5	Negative	India
7574	Lenovo IdeaPad Slim 3 AMD 3020e 15.6 inch HD T...	Lenovo	nice laptop	Shubham S.	28 October 2020	5	5	Positive	India
7576	Lenovo IdeaPad Slim 3 AMD 3020e 15.6 inch HD T...	Lenovo	first of all i would like to tell about the de...	kaithavalappil kunjumon sreeraj	28 October 2020	2	5	Positive	India

Figure 6: Preview of data frame generated from web scraping and sentiment analysis

2. Line graph - Volume of Reviews

- The number of reviews published for a particular product listing is a strong metric for the engagement of the product. I.e. - How much is it talked of in the media. This is a strong indicator of the success of advertisements and marketing. It is symbolic to the buyer's purchase behavior patterns. From this result, the time-wise trends can be extrapolated for further analysis.
-

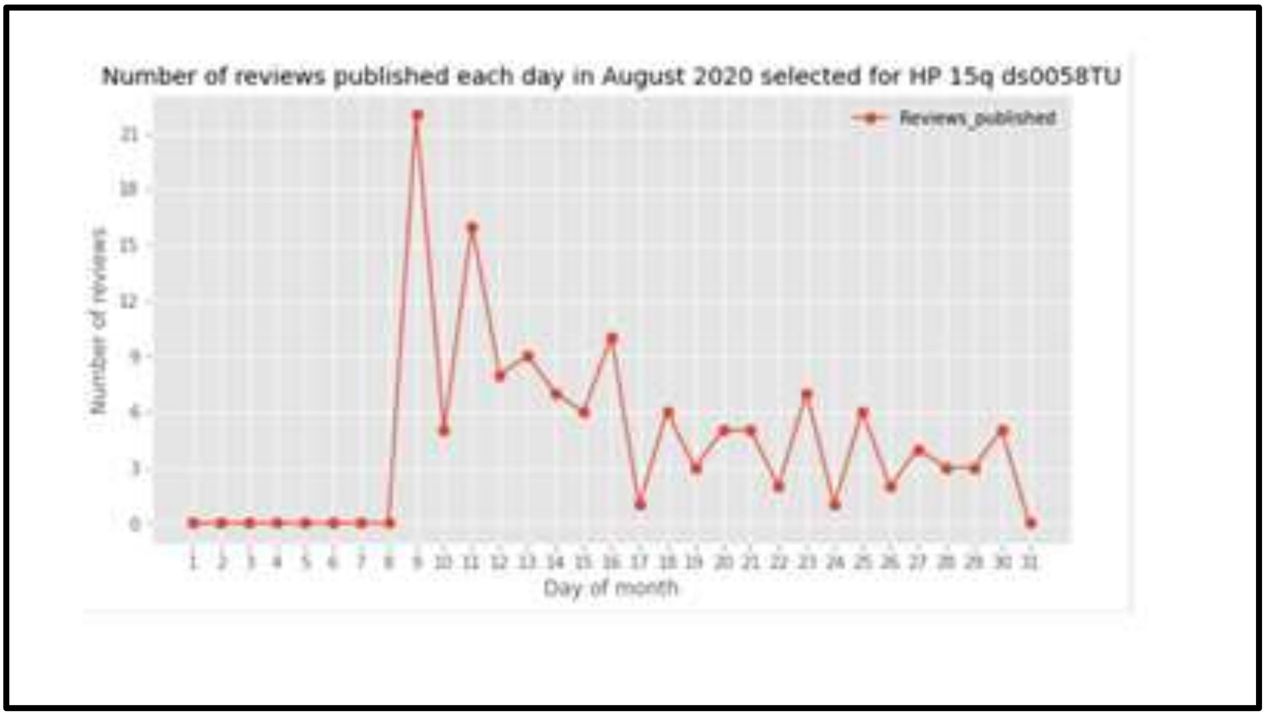


Figure 7: Preview of line graph generated for volume of reviews

3. Multi-line graph - Fluctuation of Sentiment

- In addition to the number of reviews, it is important to gauge the underlying, cumulative sentiment of the reviews. This helps enterprises quantify feedback with respect to new services, products, or other changes. This is particularly helpful to compare with competitors' results.

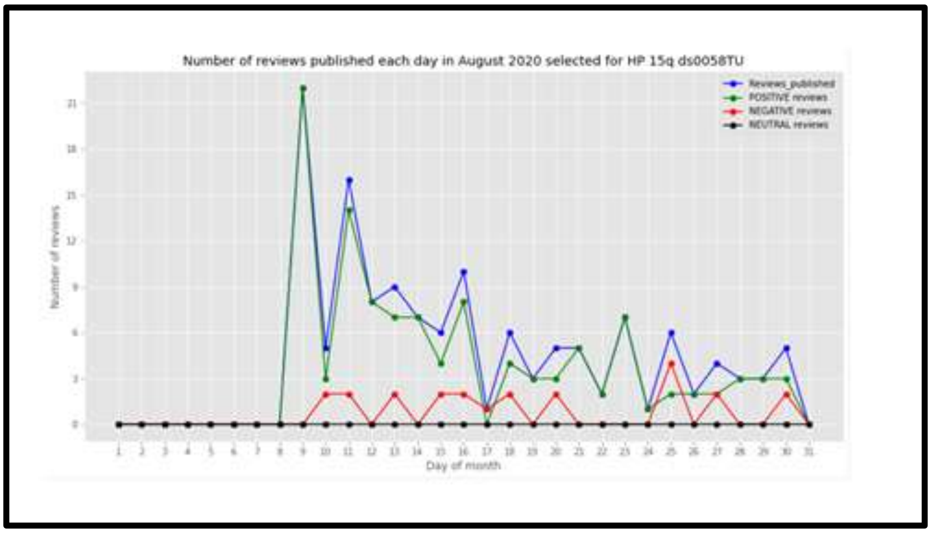


Figure 8: Preview of multi-line graph generated for sentiment fluctuation of reviews

4. Pie chart - Distribution of Sentiment

- Companies with better average ratings are more likely to see views converted to traffic and sales. This overall, quick analysis is provided by this result by not only showing the sentiment distribution, but also the absolute number of positive, negative, and neutral reviews over a selected time period.

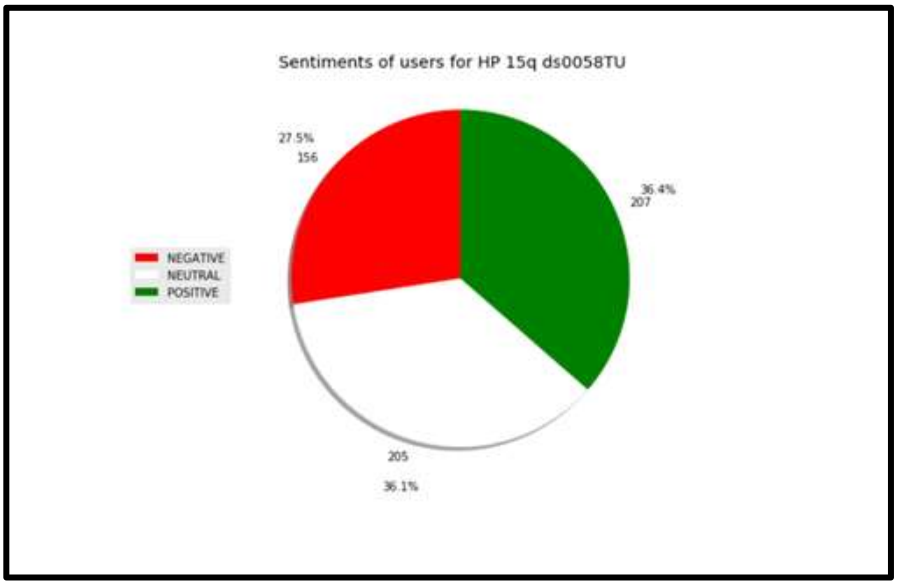
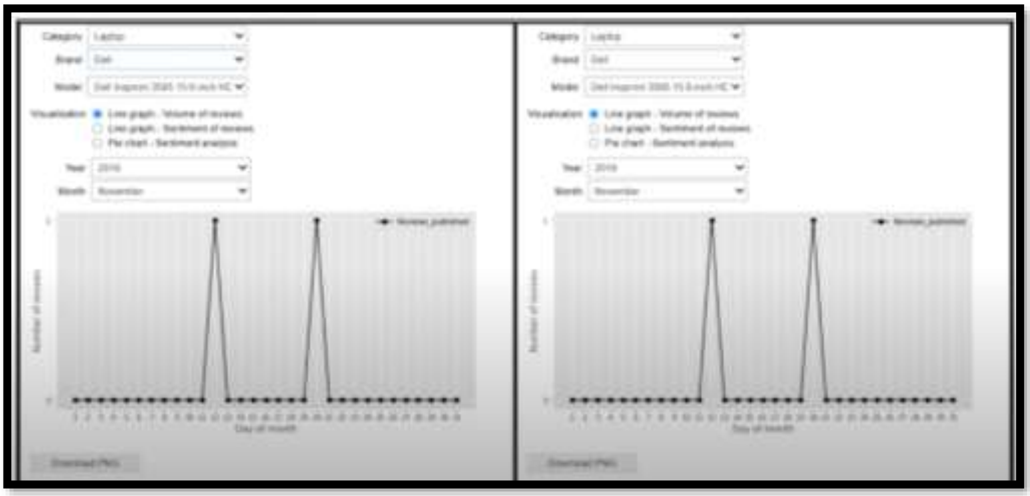


Figure 9: Preview of pie chart showing overall distribution of sentiment

5. Compare view

- Each chart generated on the main dashboard can be viewed on the 'Compare' tab with an additional product as well. This supports enterprises in evaluating their performance and brand image vis-a-vis other competitors.
-


Category: Laptop
Brand: Dell
Model: Dell Inspiron 3585 15.6-inch HD
Visualization: ● Line graph - Volume of reviews ○ Line graph - Sentiment of reviews ○ Pie chart - Sentiment analysis
Year: 2019
Month: November

Category: Laptop
Brand: Dell
Model: Dell Inspiron 3585 15.6-inch HD
Visualization: ● Line graph - Volume of reviews ○ Line graph - Sentiment of reviews ○ Pie chart - Sentiment analysis
Year: 2019
Month: November

Figure 10: Compare view from the "Dual Compare" tab on the dashboard

6. Downloadable options

- All insights generated via the dashboard can be saved by the user directly on the desktop. This aids in sharing of images and excel tables, as well as viewing directly through a device.

5. CONCLUSION

This paper discussed the development and operation of an easy-to-use Sentiment Analysis platform that can enable MSMEs to extract and analyze actionable insights from customer reviews. The system is built to be domain-agnostic, scalable, and unbiased such it that it can be used by any type of MSME.

The paper looked at the typical problems faced by the manual analysis of customer reviews and identified key places where a solution could be integrated to allow MSMEs to improve customer interactions and establish their product's brand with minimal resources. Following this, the solution proposed was broken down through various components of the proposed platform.

To extend the discussion beyond the basic design of the platform, five machine learning models were investigated for performance. After evaluating the train-set accuracy and test-set accuracy, Support Vector Machines was chosen as the model of choice. Furthermore, it was decided that the model would have a 78:22 Train: Test ratio split to prevent the creation of imbalance between datasets. The paper also presented a step-by-step detailed breakdown of the NLP and machine learning approach equipped for the platform which would allow extraction of sentiment from lexical datasets. Grid search optimization was also implemented, and the final accuracy of 92.30% was achieved.

Lastly, the mode of operation of the dashboard was explained, and the ease of operability of the platform was made evident. For future development of this platform, further investigation into the model choices and dashboard designs can be undertaken.

5. REFERENCES

[1] Tran, Joe. "NLP Sentiment Analysis for Beginners." *Medium*, Towards Data Science, 15 June 2020, towardsdatascience.com/nlp-sentiment-analysis-for-beginners-e7897f976897.

[2] Koehrsen, Will. "Interactive Controls for Jupyter Notebooks." *Medium*, Towards Data Science, 27 Jan. 2019, towardsdatascience.com/interactive-controls-for-jupyter-notebooks-f5c94829aee6.

[3] "IBM Data Science." *Coursera*, www.coursera.org/professional-certificates/ibm-data-science.

[4] Chauhan, Bhagyeshwari. "Scraping Amazon Reviews Using Scrapy in Python - Datahut." *Datahut Blog*, Datahut Blog, 14 July 2021, www.blog.datahut.co/post/scraping-amazon-reviews-python-scrapy.