**1) Domain:** Ecommerce & Social Media Analytics

**2) Problem statement, Objective and scope:**

**Problem statement:**

Customers find it difficult in deciding on good and bad products, as the process of checking the reviews of product in every other ecommerce sites/social media is cumbersome and time consuming. The impact of which is low customer satisfaction, reduced purchase rate. A successful solution would be displaying all the reviews from different platforms in one repository, then classifying the products into good or bad and readdress those impacts.

**Objective:**

The objective of this project is to show how sentimental analysis and model prediction can help improve the user experience over deciding or classifying a good and bad product

**Scope:**

The opinions must be in English and simple words. It does not support other languages. It may not handle sarcastic or negation well. So, in that case, result may vary or unexpected. Product perspective - This software doesn't depend on any other hardware of software other than resources provided by a system. R setup with respective modules or libraries do all the work required. User characteristics - The users can be anyone who wants to analyse data based on polarity of the sentences and check whether it’s a good or bad product.

**3) Data source, rows & columns, Missing values count, Outliers count and data description:**

**Data Source:**

* Twitter
* Facebook
* Mouthshut.com
* Official websites of the respective product and brand
* [gadgets.ndtv.com](https://gadgets.ndtv.com/)
* Consumerreports.org

**Rows and columns: 13316 and 3(brand, category, review)**

**Missing values count: 138**

**Outliers Count: Nil**

**Data Description:**

Product reviews (review comments) of different categories of products (household products, electronic gadgets, mobiles) has been collected for different brands(Eg: Apple,LG,etc,..)

Data has been collected from social media’s as well from official websites and other product review platforms.

**4) Models, tools and language used:**

**Tools:** Rstudio, Jupyter Notebook, Excel, Tableau

**Languages:** R,Python

**Models:**

Naïve Bayes

Logistic Regression

**5) Comparison of model performance for train& test or cross validation measures with mean and standard deviation:**

|  |  |  |
| --- | --- | --- |
| **S 1NO** | **Machine Learning Model** | **Accuracy** |
| **1** | **Naive Bayes** |  |
|  | 3 groups classification  2 groups classification | 67%  76% |
| **2** | **Logistic Regression (2 group classification)** |  |
|  | Simple Model  Using Pruned Vocabulary  Using N-Gram  Using Hashing Feature | 93.42%  93.66%  93.37%  91.33% |

**6) Main Inferences:**

Negative reviews were only of 15% of overall data and neutral and positive were of 85% of data which has confirmed data is highly biased.

When using 3 class variables we got an overfit model. So we have decided to go with 2 class model where neural comments were also being distributed equally between the 2 classes, which has given much better accuracy.

But 3 class prediction is highly recommended and we had to balance the data by creating subsets of the combinations of data(eg: positive,netral and neutral, negative and negative, positive) and apply SMOTE technique to get the balanced well spread data and then feed into the model for better predictions.

**7) Conclusion and recommendation:**

Categorization of reviews assist the customers to make an informed choice on whether to buy a product or not based on it's True Positive, False Positive, True negative and False negative points by reducing the time that they would have spent reading through a loads of reviews from different social media. The proposed approach in this paper tries to predict sentiments from reviews posted by users on social media like Facebook, twitter and other official websites of that particular brand. The approach used a lexicon-based approach. It tries to evaluate each opinion bearing word based on its intensity. It assigns a valence to words found in the tweets/paragraph using AFINN lexicon dataset. The logistic regression and Naïve Bayes classifiers are used for training and testing the various product review dataset.

From a business aspect, this project built would help understand the product acceptance of customers from various social media platforms irrespective of where the purchase was made. Model with further analysis will help businesses recognize product features which are less liked by users and improve on them for its success. Similarly, highlighting of features/products which are much liked could be done and thereby showcase the product overall success and increase visibility.

In future, we can enlarge the number of users by providing an environment(app) to share their reviews no matter how they purchase the product(though e-commerce websites or offline) which also helps to collect the data directly from users along with other sources. Also, there should be some automatic method that can recommend products based on features. One can use some mining technique or soft computing methods over these features to recommend products for users.