EAI Endorsed Transactions
on Scalable Information Systems

Research Article EAI.EU

Google Maps Data Analysis of Clothing Brands in South Punjab, Pakistan

Muhammad Ahmad, Kazim Jawad, Muhammad Bux Alvi*, and Majdah Alvi

Department of Computer Systems Engineering, Faculty of Engineering, The Islamia University of Bahawalpur, 63100, Punjab, Pakistan

Abstract

The Internet is a popular and first-hand source of data about products and services. Before buying a product, people try to gain quick insight by scanning through online reviews about a targeted product. However, searching for a product, collecting all the relevant information, and reaching a decision is a tedious task that needs to be automated. Such composed decision-assisting text data analysis systems are not conveniently available worldwide. Such systems are a dream for major cities of South Punjab, such as Bahawalpur, Multan, and Rahimyar khan. This scenario creates a gap that needs to be filled. In this work, the popularity of clothing brands in three cities of south Punjab has been assessed by analysing the brand's popularity using sentiment analysis by prioritizing brands based on organic feedback from their potential customers. This study uses a combination of quantitative and qualitative research to examine online reviews from Google Maps. The task is accomplished by applying machine learning techniques, Logistic Regression (LR), and Support Vector Machine (SVM), on Google Maps reviews data using the n-gram feature extraction approach. The SVM algorithm proved to be better than others with the uni-bi-trigram features extraction method, achieving an average of 80.93% accuracy.

Keywords: Sentiment Analysis, Google Maps Data, Clothing Brands, Logistic Regression, Support Vector Machine

Received on 08 September 2022, accepted on 13 November 2022, published on 13 January 2023

doi: 10.4108/eetsis.v10i3.2677

1. Introduction

Customer feedback is vital for deciding a brand's reputation. Most customers willingly share their genuine experiences and thoughts about the product or service. Online reviews are helpful for brands in improving their product quality, customer engagement, credibility, and trust [1]. The common practice of the customers is to check the overall rating of the product or manually skim the previously recorded text feedback, analyse it, and then decide whether to buy a product or not. A manual effort of searching, analysing, comparing, and concluding is a tedious practice that mostly ends with the targeted product's shallow/misleading conclusion [2, 3]. Therefore, an algorithmic method should be introduced to overcome this limitation [4]. This study is an effort to provide the people of South Punjab with a quick fix, reviews-driven sentiment classification system for clothing brands that may assist them in quantifying trends in Bahawalpur, Multan, and Rahimyar khan cities. The data in this research is extracted from their potential customer's brand reviews posted on Google Maps. Google Maps is the fastest-growing review site surpassing Facebook [5]. Many brands and other businesses are enlisted on Google Maps where a user can access the outlet's location and give feedback about products or services. The proliferation of reviews on Google Maps improves the platform's significance. The extracted customer reviews are in free text form. Text data have implied shortcomings that can hinder the sentiment analysis process. Such drawbacks of free-text data include:

- Data is noisy.
- Data contains irrelevant content.
- It contains spelling errors and contractions.
- It often consists of user-improvised language related to computer-mediated social networks.

*Corresponding author. Email: mbalvi@iub.edu.pk

The flow of this research follows data extraction from Google Maps, data integration, data cleaning, preprocessing, feature engineering, and model development. The selected features are fed to two state-of-the-art Machine Learning Algorithms, i.e., LR and SVM to develop the predictive model. The predictive model can successfully classify customer reviews about Bahawalpur, Multan, and Rahimyar Khan clothing brands. This research benefits a vast community, including brand managers, owners, and potential customers, in terms of statistical analysis and descriptive judgment. Through the statistical analysis, brand managers and owners can better understand customer sentiments and market trends, hence financially reaching a better market share. At the same time, customers can scrutinize the brands' descriptive judgments and obtain their services. This study resulted in a machine learning-based model that utilizes Google Maps reviews to assist people in evaluating a clothing brand quickly. The developed model obtained an average of 80.93% accuracy on the validation dataset.

The paper is distributed in a way that section 2 describes the related literature work, dataset elaboration is given in section 3, whereas section 4 is about the research method. The experimental work results are discussed in section 5, and section 6 concludes the work and indicates future direction. Finally, the paper ends with a list of references.

2. Related Work

This section describes the research related to systems that use web technologies to accumulate customer feedback to enhance, improve, or change their service style or product quality. Such feedback analysis system(s) can be developed using python and allied tools such as selenium, Beautiful Soup, sentiment analyzing lexical libraries, and scikit-learn [6]. Natural Language Toolkit (NLTK) is another powerful tool that works with human-generated text data and encompasses a cluster of open-source modules that provide easy-to-use interfaces to over 50 corpora and lexical resources [7]. Each passing second adds Millions of Megabyte data to the digital world using the Internet [8], and Internet data extraction is coined as web harvesting, web scraping, or web extraction. Researchers are taking advantage of the vast amount of available data and using it for research purposes [9]. Beautiful Soup [10] and Selenium WebDriver [11] are among some of the handy tools that make scrapping [12] easier by automating tasks to reduce human intervention. Brand review data is vital and appealing for many researchers and analysts to get insights into the trends using customer-generated product reviews. The authors applied text mining and network analysis to product reviews [13]. They also suggested an approach that allowed managers to control the strengths and weaknesses of brand image effectively. Filipa Rosado-Pinto et al. worked on customers' online restaurant reviews [14]. They explored the brand's authenticity and consumer brand engagement using text-mining techniques.

Online reviews do not directly impact sales, but indirectly, they are very effective for customers to choose the right brand, which can boost the brand's sales. Stephen J. Carson et al. [15] studied the effects and relationship between online customer reviews and sales. Boonyanit M. and Viriya T. [16] examined the customer experience by analysing online reviews of different restaurants present on Google Maps and used the VADER [17] for review classification. They applied a logistic regression algorithm to obtain results.

Supervised learning methods require annotated data samples (Text data is often unlabelled.). Usually man-power is used to label data samples. The overall human labelling accuracy is 82.9%, proved in human labeling experiments [18]. However, human annotation involves a lot of time and expenses. An alternate for text data annotation, especially for sentiment classification, is using sentiment lexicons such as Vader, SentiWordNet, TextBlob, etc.

Text preprocessing plays a vital role with unstructured text data. The authors in [19] presented text preprocessing techniques that eliminated noisy data and improved the model's performance. Another work that reported the impact of text preprocessing on model performance is described by S. Alam and N. Yao in [20]. M. B. Alvi et al. studied the implications of recursive preprocessing on text data [21]. They have used 19 preprocessing techniques and suggested a preprocessing pipeline that works iteratively. Another work by the same authors reported on developing a hybrid model that could perform sentiment analysis on the issue of global warming [22]. They claimed to achieve 86% accuracy using their model. Shuai Liu et al. [23, 24] introduced engineering applications of effective hybrid information and big data processing.

Tokenization is one of the initial and essential processes in preprocessing pipelines, separating input strings into individual tokens. The significance and complexity of tokenization are addressed in [25]. The Authors in [26] have performed a comparative evaluation of tokenization based on quantitative methods. Stemming and lemmatization are similar and effective preprocessing steps that help reduce the vector space of a given data set. Vimala B. and Ethel Lloyd-Yemoh [27] proposed a performance-based comparison between stemming and lemmatization and found that lemmatization techniques produced the best result. Michael W. Browne reported on cross-validation methods in [28]. The research concluded that predictive accuracy depends on sample size and the number of predictor variables.

Ronen Feldman described the applications and challenges of one of the hottest fields in computer science: sentiment analysis [29]. Alessia D'Andrea et al. in [30] explained the classification of approaches, tools, applications, and implementation of sentiment analysis. The performance of various sentiment analysis methods showed that SVM gave higher accuracy than the entropy method, as shown in [31]. The authors of [32] investigated Decision Tree, LR, Naïve Bayes, Random Forest, and SVM classifiers implemented in Apache Spark (an in-memory intensive computing platform). Their findings indicated that the LR for product reviews achieved the highest classification accuracy.

3. Dataset

Data mining is the computational process of identifying, engaging, categorizing, analysing, and maintaining data [33]. People not only take guidance on the location using Google Maps but also share their feedback. Many clothing brands in Bahawalpur, Multan, and Rahimyar Khan are listed on Google Maps. The targeted brands with more than 13 reviews were scrapped, ignoring other brands with a smaller number of reviews. In total, there were 51 brands in the final list: 18 in Bahawalpur, 24 in Multan, and 09 in Rahimyar Khan, with a total number of 4121 reviews in the initial list, of which 1416 belonged to Bahawalpur, 1951 belonged to Multan, and 454 belonged to Rahimyar Khan city where 300 reviews were exempted from the list. The Reviews were examined and exempted by using the following criteria: (1) empty reviews, (2) reviews in non-English, and (3) irrelevant reviews.

Google Maps facilitates the customers to provide star ratings and text feedback. But some of the users give their feedback only in star ratings. Such star rating feedback creates the dataset's blank (empty) text fields. The data extractor unified such blank input into the primary dataset. The scrapped dataset also acquired non-English data samples (Urdu script and Roman Urdu reviews). Additionally, the dataset also includes irrelevant data samples related to business promotions and sports. All these inappropriate reviews were identified and exempted from the dataset for final qualitative analysis to avoid skewed results and peculiar outcomes. The sample of exempted reviews is shown in table 1, while the filtered review sample is given in Table 2.

Table 1. Sample of excluded data from Google Maps Dataset

No.	Review	City	Brand
1	“ ” ~Empty reviews	All Cities	All brands
2	Mjhe sapphire ka suit acha lagta hai	Multan	Sapphire
3	Contact me for private tuitions: +92 302****08	Bahawalpur	Khaadi

Table 2. Sample of included data from Google Maps Dataset

No.	Review	City	Brand
1	Affordable brand and good quality garments	Bahawalpur	Engine
2	Up to 50% sale currently. Nice collection, casual and formal as well. Must visit	Bahawalpur	Charcoal
3	I had visited the outlet a weak before a sale, suppose if a shirt was of 1000 then is now @ 800 & sale tag claiming 30% off... Does that make some sense???	Bahawalpur	Breakout
4	Great place! They recently had a sale of 50 percent of!! Enjoyable for women who are interested in clothes, but it is quite expensive!	Bahawalpur	Sapphire
5	Unnecessarily expensive and cheap stuff for more price. Not recommended	Bahawalpur	Khaadi
6	Not too much excited about prints and designs but stiff and fabric no doubt of the good quality	Bahawalpur	Nishat Linen
7	The stuff is not good, but the stitching is mind blowing. Most shirts' color fades away in sunlight beware.	Bahawalpur	Outfitters
8	The front desk officer/manager's behaviour is loud and not nice	Multan	Beech Tree
9	A good place to shop for trendy clothing and accessories	Multan	Breakout
10	Extremely bad customer service. Unethical staff with zero manners	Multan	Cougar
11	Your dubbta width is so good especially it is accurate for namaz, simply perfect	Multan	Ideas
12	Limelight is bestselling brand ever. Specially for unstitched cloths.	Multan	Limelight
13	Nice international standard outlet	Multan	Sapphire
14	No more on this location. Shifted to another place	Multan	Zellbury
15	A lil bit costly form other brands but have a vast variety of suiting and other casual clothes	Rahimyar Khan	Edenrobe
16	j. Needs to provide latest designs and colors bcz mostly they make cloths with old colors	Rahimyar Khan	J.
17	Good addition in clothing choice for Rahim yar khan	Rahimyar Khan	MariaB
18	Shop to make your event memorable	Rahimyar Khan	Uniworth
19	Nice atmosphere with reasonable rates of female clothes	Rahimyar Khan	Zellbury

20	Good quality fabrics available here with beautiful designs	Rahimyar Khan	Alkaram
21	Second Diner's outlet in Rahim yar khan, very big, great new stuff but more experience!	Rahimyar Khan	Diners

4. Experimental Method

This section describes the method adopted to undertake this work. The method includes four main phases: data extraction and integration, data preprocessing, exploratory data analysis (EDA), and model development. In the first phase, a custom build scrapper was used to extract the data from Google Maps. Data integrator involves the integration of multiple .csv files to develop a single .csv file. Since the review text has no polarity, the next step is annotating the data. After the annotation, the annotated data samples were compared for variance, and the conflicting reviews were considered for discussion. In the data preprocessing step, the integrated data was masked by removing unwanted and trivial reviews. The features in the training data were selected to fulfil the assured bias. EDA involved comparative analysis. Figure 1 shows the proposed sentiment analysis system for brand reviews.

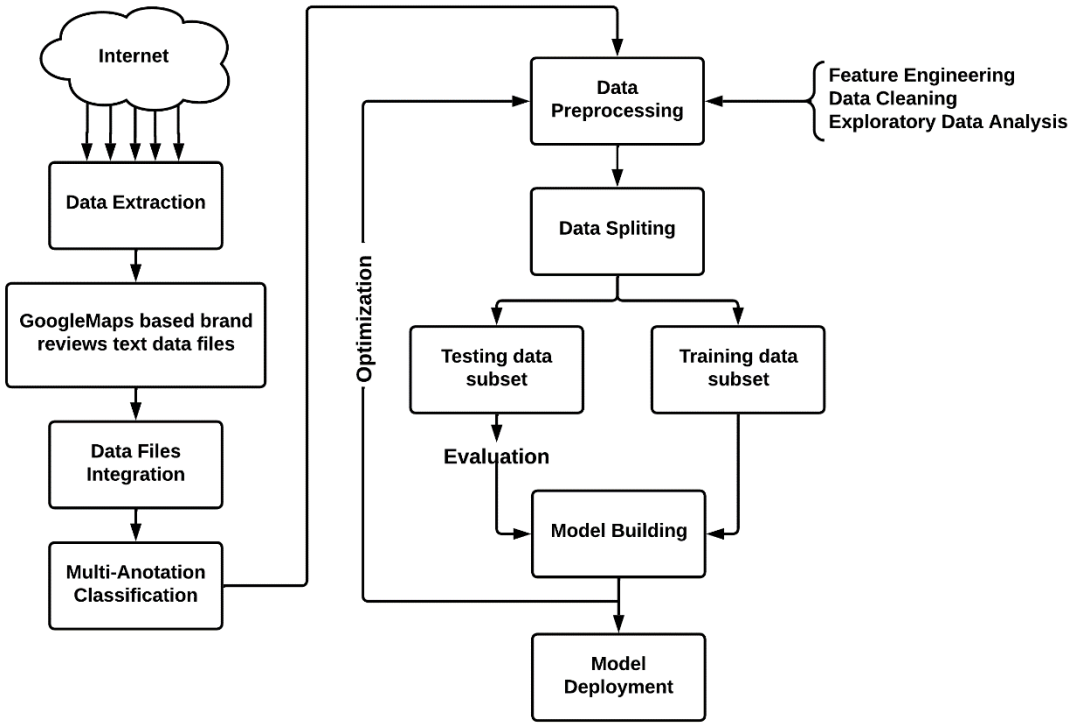

Figure 1. Step-by-step methodology for sentiment analysis system for brand reviews

4.1 Data Extraction

Privacy of user credentials is one of the major concerns in digital civilization. Therefore, only publicly available reviews were extracted that were published on the Internet by the person, and no personal information was gathered. Different social networking sites facilitate through Application Programming Interfaces (APIs) that assist in data collection. Rate limiting, API licensing, and manual data collection are too costly in terms of data and time [34]. The motive of a custom extractor is to provide some additional advantages over traditional extraction methods, i.e., the usage of APIs. Google Maps allows users to search by the name of the brand and the search filters such as "location". There are hundreds of cloth-related brands identified to have reviews. The well-known brands in three major cities of South Punjab were examined on Google Maps to categorize if they contained enough customer reviews. The brands with more than 13 reviews were filtered out to maintain the quality of the analysis. Fifty-one brands in these cities fulfilled all the requirements, having more than 13 reviews. In the second phase, the listed brands from the initial search round were efficiently scrapped by the custom scraper.

The Customer reviews were extracted using the developed Google Maps reviews data extractor (RDE). The extractor was developed using "Selenium" and "Beautiful Soup". Moreover, the utilization of the scrapper depends upon the URL of the required web page. The scrapper parsed through filtered brand URLs, extracting all the

reviews on that page, and saved them in a .csv file with the name of the brand in the next index. Once all the reviews from the URL were scraped, the scrapper proceeds to another brand URL. Subsequently, All the reviews from the list were extracted using the same procedure, moving with three different files based on their cities to analyse their popularity among people.

4.1.1 Data Integration

Algorithm 1 shows the functional approach of the data integrator. The developed data integrator combines the data files (in .csv format) into one file, representing all the brand reviews of respective cities. The file integration provides an easy city-wise comparison of all the clothing brands.

Algorithm 1

Input: *The data files of the multiple brands of a specific city.*
For each data file, do:
1 If (Reviews >= 13), Append it into a new data file.
2 Drop the data file having less than 13 reviews (if any).
3 Fetch the name of a brand as a value.
4 Extend the new data file with reviews and values.
5 Repeat the step for other cities.
Output: *The unified data file.*

4.1.2 Data Annotation

The obtained data set does not have any polarity. Therefore, the Multi Annotator methodology reliably helped to annotate the integrated data set. The annotators were clear about a few perspectives, such as what to annotate (according to contextual perspective). All three annotators independently annotated the data into two sub-categories (positive and negative). After discussion, some contradictory labels over a few data samples were smoothly resolved by adopting the majority vote label.

4.2 Data Pre-processing

Data preprocessing is a rudimentary but essential step of sentiment analysis [19, 35]. Data preprocessing steps convert the raw uncleaned text data into a format compatible with the machine learning algorithms. The textual data cannot be fed directly to machine learning algorithms. In the preprocessing pipeline, the output of one process becomes an input for the forthcoming process.
Tokenization [36, 37] is the second preprocessing step that breaks continuous strings into small fragments: words, keywords, phrases, or symbols. Tokenization or word segmentation is a significant step because it helps in masking non-important words, punctuation, and digits. A raw annotated data set contains impurities, which may cause redundancy/inaccuracy in an algorithm. Text data preprocessing also involves the process of non-trivial term reduction. A Non-trivial term reduction implies case-normalization, stop words/special characters/numbers removal, and stemming. A case-conversion process that brings uniformity to the text. Removal of the special characters and numbers constitutes another step of preprocessing text data. Stemming brings down the different forms of verbs, nouns, and word variants into semantically base words. These initial preprocessing steps not only result in a cleaner dataset but also mitigates feature space, which increases computational efficiency. After following the above steps, a document-term matrix is built, the second stage of preprocessing. The rows of the document term matrix denote reviews, whereas the column represents the (n-gram) features. N-grams are very useful for the development of N-gram Language Models (LMs). N-gram is the frequency of the word sequence that appears in a corpus text. N-gram used to have the improved prediction of a system by aid of probability of occurrence of a particular word. The probability calculation example of the N-gram model is given in Table 3.

Table 3. The probability calculation example of the N-gram model

This clothing brand is good
P (‘this clothing brand is good.’) = *P ('this', 'clothing', 'brand', 'is', 'good')*
P ('this’) P ('clothing '|'this') *P ('brand' | ‘this clothing')* *P ('is'|' this clothing brand')* *P ('good'|' this clothing brand is')*

Unigram, bigram, uni-bigram, and uni-bi-trigram are the features that are used for this work. Considering the training data set of three reviews as an example:

- *Good brand for shopping.*
- *Expensive for shopping.*
- *Worse brand for shopping.*

These three reviews contain six unigrams, i.e., “Good”, “Brand”, “For”, “Shopping”, "Expensive", and "Worse" (refer to Table 4). There are five bigrams, i.e., “Good brand”, “Brand for”, “For Shopping”, “Expensive for”, and "Worse Brand " (refer to Table 5). There are eleven uni-bigrams, i.e., “Good”, “Brand”, “For”, “Shopping”, “Expensive”, “Worse”, “Good brand”, “Brand for”, “For shopping”, "Expensive for”, “Worse brand" (refer to Table 6). Furthermore, there are fifteen uni-bi-trigrams, i.e., “Brand”, “Brand for”, "Brand for shopping", “Expensive”, “Expensive for”, “Expensive for shopping", “For”, “For shopping”, "Good", "Good Brand"," Good brand for", "Shopping", "Worse” “Worse brand”, "Worse brand for" (refer to Table 7).

Table 4. Example of a document-term matrix (Unigram)

Feature	Good	Brand	For	Shopping	Expensive	Worse
1	1	1	1	1	0	0
2	0	1	0	0	1	0
3	0	1	1	1	0	1

Table 5. Example of a document-term matrix (Bigram)

Feature	Good Brand	Brand For	For Shopping	Expensive For	Worse Brand
1	1	1	1	0	0
2	0	0	1	1	0
3	0	1	1	0	1

Table 6. Example of a document-term matrix (Uni-Bigram)

Features	Good	Brand	For	Shopping	Expensive	Worse	Good Brand	Brand For	For Shopping	Expensive For	Worse Brand
1	1	1	1	1	0	0	1	1	1	0	0
2	0	1	0	0	1	0	0	0	1	1	0
3	0	1	1	1	0	1	0	1	1	0	1

Table 7. Example of a document-term matrix (Uni-Bi-Trigram)

Features	Brand	Brand For	Brand For Shopping	Expensive	Expensive For	Expensive For Shopping	For	For Shopping	Good	Good Brand	Good Brand For	Shopping	Worse	Worse Brand	Worse Brand For
1	1	1	1	0	0	0	1	1	1	1	1	1	0	0	0
2	0	0	0	1	1	1	1	1	0	0	0	1	0	0	0
3	1	1	1	0	0	0	1	1	0	0	0	1	1	1	1

Another preprocessing step involves splitting the dataset into training and testing sub-datasets. StratifiedKfold cross-validation is used to separate the data for training testing with the 05 splits. Cross-validation also helps avoid model overfitting and data leakage. The comparative percentages were also calculated in the EDA process, and the positive and negative ratios were categorized. In this research, only a positive rate is selected for further analysis, as shown in Table 8.

Table 8. The brand reviews with their comparative percentage

Cities	Brand Name	No. of Reviews	Positive	Negative	Neutral	Comparative Percentage (Positive)
Bahawalpur	Edenrobe	228	200	15	13	87.72
	ChenOne	186	132	34	20	70.97
	Diners	176	139	18	19	78.98
	J.	159	139	12	8	87.42
	Sapphire	152	136	9	7	89.47
	Khaadi	90	79	8	3	87.78
	Limelight	83	80	2	1	**96.39**
	Outfitters	67	59	7	1	88.06
	Charcoal	47	27	5	15	**57.45**
	Uniworth	43	41	0	2	95.35
	Nishat Linen	41	39	0	2	95.12
	Engine	30	26	2	2	86.67
	Breakout	27	19	3	5	70.35
	Alkaram Studios	21	20	0	1	95.24
	Ideas	20	18	2	0	90.00
	Factory Price	17	15	2	0	88.24
	Jean Junction	16	15	1	0	93.75
	One	13	12	0	1	92.31
Multan	Rang Ali Fabrics	259	180	32	47	69.50
	Khaadi	223	163	30	30	73.09
	Ideas	164	121	28	15	73.78
	Diners	162	113	35	14	69.75
	Nishat Linen	144	99	16	29	68.75
	Insaaf Fabrics	133	96	26	11	72.18
	Alkaram Studios	99	62	15	22	62.63
	Limelight	96	68	9	19	70.83
	Uniworth	83	46	21	16	**55.42**
	Ethnic	77	58	12	7	75.32
	Breakout	76	54	14	8	71.05
	Rahim Centre	70	43	7	20	61.43
	Edenrobe	59	38	19	2	64.41
	Shirt and Tie Shop	52	43	5	4	82.69
	Prince Cloth	43	30	4	9	69.77
	Bonanza Satrangi	39	31	6	2	79.49
	Cougar	29	22	5	2	75.86
	Ismails Gulgushat	26	21	4	1	80.77
	Gravity Inc	24	20	4	0	**83.33**
	Cambridge, Zeen & Guts	22	15	6	1	68.18
	So Kamal	20	14	4	2	70.00
	Beechtree	19	15	3	1	78.95
	International Cloth sales depot	19	15	2	2	78.95
	Maria B	13	8	3	2	61.54
Rahim Yar Khan	J.	131	105	15	11	80.15
	Khaadi	118	81	25	12	68.64
	Diners	55	41	8	6	74.55
	Alkaram Studios	31	27	1	3	**87.10**
	Zellbury	28	20	5	3	71.43
	Edenrobe	28	19	4	5	67.86
	Breakout Kids	27	22	2	3	81.48
	Uniworth	21	17	4	0	80.95
	Maria B	15	10	4	1	**66.67**

4.3 Model Development

Machine learning algorithms are used to build models by using the training dataset. The following algorithms are used for this purpose.

4.3.1 Logistic Regression

Logistic regression is one of the most powerful and widely used algorithms for binary classification problems. logistic regression describes and tests the hypotheses about the relationship between a categorical outcome and predictor variables [38]. In logistic regression, the hypothesis function theta may be determined using Equation 1, which predicts values between 1 and 0. With the threshold set to 0.5, it classifies the output as true (positive) or false (negative).

$$0 \leq h_\theta\,(\theta^T\,x) \geq 1 \quad (1)$$

The hypothesis function is determined using Equation 2.

$$h_\theta(\theta^T x) = \frac{e^{\theta Tx}}{e^{\theta Tx}+1} \quad (2)$$

4.3.2 Support Vector Machine

The support vector machine algorithm has been found to be effective in text mining applications. The large-margin classifier classifies the output as either 0 (negative) or 1 (positive). The SVM detects the optimal separating hyperplane and maximizes the margins (in the training data), which can be especially effective in high dimensional spaces. The margin is the distance between the hyperplane and the nearest data points of different data set classes. The SVC uses various parameters to modify SVC algorithm functionality (such as kernel=”linear”, C=1, loss=”squared_hinge”, max_iter=1000, penalty=”l2”, tol=0.0001). The hypothesis function of SVM can be determined by Equation 3:

$$h_\theta(x) = \begin{cases} 1\ if\ \theta^T x \geq 1 \\ 0\ if\ \theta^T x \leq -1 \end{cases} \quad (3)$$

5. Result and Discussions

The results of two models with four features (unigrams, bigrams, uni-bigrams, and uni-bi-trigrams) are shown in Table 9. The SVM-based model outperformed logistic regression using all four features. Overall, the best results were achieved by using Uni-Bi-trigram, obtaining 83.26% accuracy. The accuracy of the classifiers was also compared, and it was observed that SVM outperforms with an average accuracy of 80.93%, followed by LR with an average accuracy of 79.23%.

Table 9. Results after feature reduction

Data Set	Features	Total number of Features	Logistic Regression %	Support Vector Machine %
Bahawalpur	Unigrams	1482	84.72	83.69
	Bigrams	5213	83.34	83.26
	Uni- Bigrams	12918	83.17	83.94
	Uni-Bi-Trigrams	10513	84.12	82.23
Multan	Unigrams	1439	77.74	79.79
	Bigrams	4323	71.48	71.79
	Uni- Bigrams	5762	77.94	78.87
	Uni-Bi-Trigrams	10083	77.74	83.25
Rahim Yar Khan	Unigrams	666	78.85	79.29
	Bigrams	1673	76.65	77.09
	Uni- Bigrams	2339	77.53	83.70
	Uni-Bi-Trigrams	4055	77.53	83.70

This work analysed the popularity of the clothing brands in Bahawalpur, Multan, and Rahimyar Khan. There was a total of 51 clothing-related brands in this data set. Only brands with 13 reviews or above were extracted and analysed, and others were ignored. After the analysis, it was found that, based on the positive reviews count, "Edenrobe", "J.", and "Diners" were the most popular brands in Bahawalpur. "Rang Ali Fabrics", "Khaadi", and "Ideas" were found to be the most popular brands in Multan. In Rahimyar Khan, "J.", "Khaadi", and "Diners” were found to be the most popular brands, as shown in Figure 2.

A comparative study revealed that in the long run, "Limelight", "Uniworth", and "AlkaramStudios" tend to be more popular brands among the people of Bahawalpur. "Gravity", "Shirt and Tie", and "Ismail Gulgushat" were found to be the most popular brands in Multan. "AlkaramStudios", "Breakout", and "Uniworth" were found to be the most popular brands in Rahimyar Khan, as shown in Figure 3.

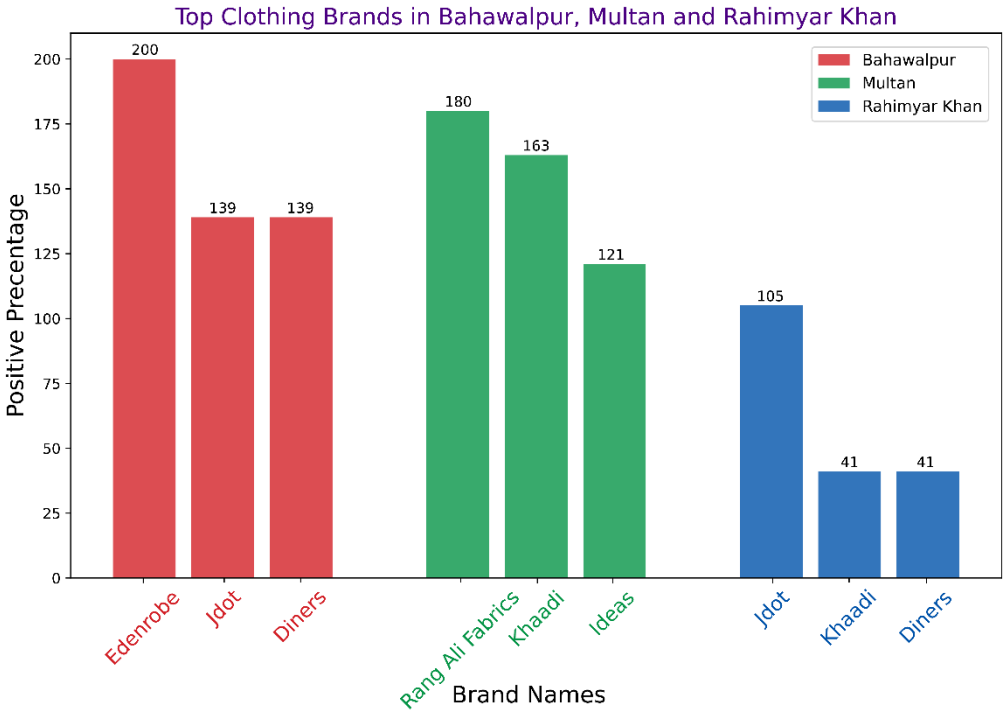

Figure 2. Brand's popularity based on a positive count

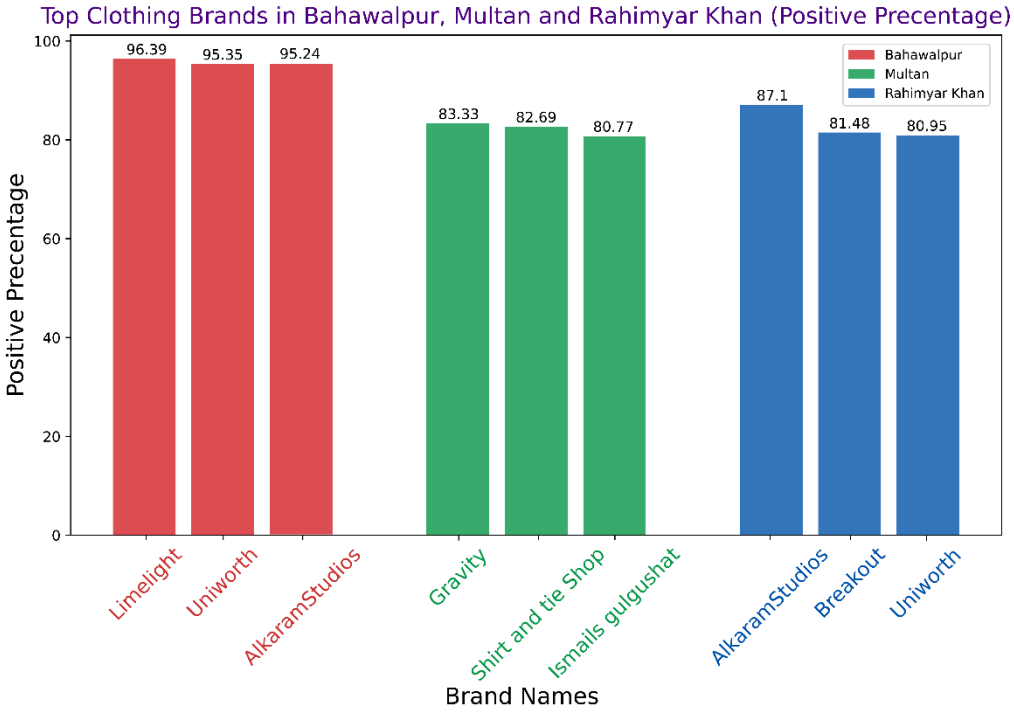

Figure 3. Comparative study on brand's popularity

6. Conclusion and Future Work

In this work, a sentiment analysis system is developed to conclude people's choices about various brands in the three most populated cities of south Punjab, Pakistan using Google Maps data. A custom data extractor was developed to get brand reviews from Google Maps data. The extracted 4121 reviews were integrated from three different files into a single file using a robust custom data integrator. The data was annotated by the multi-annotation method, and then preprocessing was regulated on the data set. Four different features (unigrams, bigrams, and their unions) were utilized with two machine learning algorithms (LR and SVM) for building the classification model. The experiments were conducted using StratifiedKfold cross-validation to suppress model overfitting, and then accuracies were computed. The results showed that the SVM algorithm-based model with Uni-Bi-Trigram features performed better than LR, obtaining an average accuracy of 80.93%. This accuracy is obtained by taking the average of all the accuracies obtained using uni-bi-tri gram with SVM. This comparative study further reveals that "Limelight" is the most popular brand in Bahawalpur, "Gravity" is famous in Multan city, and "AlkaramStudios" is prominent in Rahimyar Khan city.

An extension to this work may include adding multilingual data sets from different data sources (Facebook/Instagram shops, website reviews from the brand's online stores, and related websites) and transforming the city-oriented analysis into brand-oriented across all of Pakistan.

References

[1] G. Salamander. "Why online reviews are so important?" https://eclincher.com/why-online-reviews-are-so-important/ (accessed 24/05/2022).

[2] J. Ha and S. S. Jang, "Effects of service quality and food quality: The moderating role of atmospherics in an ethnic restaurant segment," *International journal of hospitality management*, vol. 29, no. 3, pp. 520-529, 2010.

[3] J. Zhang, W. Zheng, and S. Wang, "The study of the effect of online review on purchase behavior: Comparing the two research methods," *International Journal of Crowd Science*, 2020.

[4] J. Zhao, K. Liu, and L. Xu, "Sentiment analysis: mining opinions, sentiments, and emotions," ed: MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info …, 2016.

[5] B. Ideas. "Comparison of Local Review Sites: Which Platform is Growing the Fastest?" https://www.brightlocal.com/research/comparison-of-local-review-sites/ (accessed 17-05-2022).

[6] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *the Journal of machine Learning research*, vol. 12, pp. 2825-2830, 2011.

[7] S. Bird, "NLTK: the natural language toolkit," in *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, 2006, pp. 69-72.

[8] Live-Counter. "How Big Is The Internet." https://www.live-counter.com/how-big-is-the-internet/ (accessed 27-05-2022).

[9] W. Fu, S. Liu, and G. Srivastava, "Optimization of big data scheduling in social networks," *Entropy*, vol. 21, no. 9, p. 902, 2019.

[10] C. Zheng, G. He, and Z. Peng, "A Study of Web Information Extraction Technology Based on Beautiful Soup," *J. Comput.*, vol. 10, no. 6, pp. 381-387, 2015.

[11] S. Gojare, R. Joshi, and D. Gaigaware, "Analysis and design of selenium webdriver automation testing framework," *Procedia Computer Science*, vol. 50, pp. 341-346, 2015.

[12] B. Zhao, "Web scraping," *Encyclopedia of big data*, pp. 1-3, 2017.

[13] S. Gensler, F. Völckner, M. Egger, K. Fischbach, and D. Schoder, "Listen to your customers: Insights into brand image using online consumer-generated product reviews," *International Journal of Electronic Commerce*, vol. 20, no. 1, pp. 112-141, 2015.

[14] F. Rosado-Pinto, S. M. C. Loureiro, and R. G. Bilro, "How brand authenticity and consumer brand engagement can be expressed in reviews: a text mining approach," *Journal of Promotion Management*, vol. 26, no. 4, pp. 457-480, 2020.

[15] N. N. Ho-Dac, S. J. Carson, and W. L. Moore, "The effects of positive and negative online customer reviews: do brand strength and category maturity matter?," *Journal of marketing*, vol. 77, no. 6, pp. 37-53, 2013.

[16] B. Mathayomchan and V. Taecharungroj, "'How was your meal?' Examining customer experience using Google maps reviews," *International Journal of Hospitality Management*, vol. 90, p. 102641, 2020.

[17] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Proceedings of the international AAAI conference on web and social media*, 2014, vol. 8, no. 1, pp. 216-225.

[18] Y. Kim and S. Ross, "Searching for ground truth: a stepping stone in automating genre classification," in *International DELOS Conference*, 2007: Springer, pp. 248-261.

[19] S. Kannan *et al.*, "Preprocessing techniques for text mining," *International Journal of Computer Science & Communication Networks*, vol. 5, no. 1, pp. 7-16, 2014.

[20] S. Alam and N. Yao, "The impact of preprocessing steps on the accuracy of machine learning algorithms in sentiment analysis," *Computational and Mathematical Organization Theory*, vol. 25, no. 3, pp. 319-335, 2019.

[21] M. B. Alvi, N. Mahoto, M. A. Unar, and M. A. Shaikh, "An Effective Framework for Tweet Level Sentiment Classification using Recursive Text Pre-Processing Approach," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 6, pp. 572-581, 2019.

[22] M. B. Alvi, N. A. Mahoto, M. Alvi, M. A. Unar, and M. A. Shaikh, "Hybrid classification model for twitter data-a recursive preprocessing approach," in *2018 5th International Multi-Topic ICT Conference (IMTIC)*, 2018: IEEE, pp. 1-6.

[23] S. Liu, Z. Li, X. Cheng, and Y. Lin, "Introduction of recent advanced hybrid information processing," *Mobile Networks and Applications*, vol. 23, no. 4, pp. 673-676, 2018.

[24] S. Liu, H. Zhou, and X. Cheng, "Recent Advancement in Hybrid Big Data Processing," *Mobile Networks and Applications*, vol. 25, no. 4, pp. 1514-1517, 2020.

[25] J. J. Webster and C. Kit, "Tokenization as the initial phase in NLP," in *COLING 1992 Volume 4: The 14th International Conference on Computational Linguistics*, 1992.

[26] B. Habert *et al.*, "Towards tokenization evaluation," in *Proceedings of LREC*, 1998, vol. 98, pp. 427-431.
[27] V. Balakrishnan and E. Lloyd-Yemoh, "Stemming and lemmatization: a comparison of retrieval performances," 2014.
[28] M. W. Browne, "Cross-validation methods," *Journal of mathematical psychology*, vol. 44, no. 1, pp. 108-132, 2000.
[29] [29]R. Feldman, "Techniques and applications for sentiment analysis," *Communications of the ACM*, vol. 56, no. 4, pp. 82-89, 2013.
[30] D. Alessia, F. Ferri, P. Grifoni, and T. Guzzo, "Approaches, tools and applications for sentiment analysis implementation," *International Journal of Computer Applications*, vol. 125, no. 3, 2015.
[31] T. Shivaprasad and J. Shetty, "Sentiment analysis of product reviews: a review," in *2017 International Conference on Inventive Communication and Computational Technologies (ICICCT)*, 2017: IEEE, pp. 298-301.
[32] T. Pranckevičius and V. Marcinkevičius, "Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification," *Baltic Journal of Modern Computing*, vol. 5, no. 2, p. 221, 2017.
[33] M. J. H. Mughal, "Data mining: Web data mining techniques, tools and algorithms: An overview," *Information Retrieval*, vol. 9, no. 6, 2018.
[34] K. Petrosyan. "Data extraction using API scraping and main challenges." https://kristinelpetrosyan.medium.com/data-extraction-using-api-scraping-and-main-challenges-de4256c1c146 (accessed 25-05-2022).
[35] M. J. Denny and A. Spirling, "Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it," *Political Analysis*, vol. 26, no. 2, pp. 168-189, 2018.
[36] S. Vijayarani and R. Janani, "Text mining: open source tokenization tools-an analysis," *Advanced Computational Intelligence: An International Journal (ACII)*, vol. 3, no. 1, pp. 37-47, 2016.
[37] T. Hiraoka, H. Shindo, and Y. Matsumoto, "Stochastic tokenization with a language model for neural text classification," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 1620-1629.
[38] C.-Y. J. Peng, K. L. Lee, and G. M. Ingersoll, "An introduction to logistic regression analysis and reporting," *The journal of educational research*, vol. 96, no. 1, pp. 3-14, 2002.