# Apparel Brand Sentiment Analysis Using Big Data

Cristina Lawson
University of California,
Riverside
claws004@ucr.edu

Hamsa Gouda Veerendra
University of California,
Riverside
hveer003@ucr.edu

Sridevi Subramanya Raju
University of California,
Riverside
ssubr023@ucr.edu

Varshini Sampath
University of California,
Riverside
vsamp003@ucr.edu

## ABSTRACT

The aim of the project is to analyze the sentiments of the people on the fast fashion brands versus well-established clothing brands. The reviews of people on Twitter form the data for this project. We will evaluate how these fast fashion brands are perceived across the world compared to well-established brands by capturing negative, positive, and neutral sentiments and also visualizing them on maps. The scope of this project is restricted to 10 apparel brands, but this can be scaled up for more clothing brands.

## 1 Introduction

The fast fashion industry faces challenges in satisfying consumer's desire for new products faster which results in the need to continuously investigate the market to quickly understand the demand and to respond accordingly. Social networks such as Twitter, Facebook, Pinterest, etc. are becoming very popular among customers as channels where consumers freely share their voice about their past experiences and their expectations about a certain product.

Fast fashion is cheap and trendy clothing. So fast fashion brands such as Shein, H&M, Boohoo and Asos are very dynamic compared to established popular brands such as Nike and Adidas. Thus the views of the users for fast fashion brands play an important role in the success of the brand. The product quality can be measured by reviews given by the customer. A new client can decide on a given product based on previous reviews.

Sentiment analysis is a popular method to mine consumer's opinions shared online, which requires systematic and automatized procedures .This methodology can collect data on consumers almost in real time and is less costly than traditional techniques based on structured questionnaires.

The aim of the project is to compare the famous clothing brands by analysing the sentiments of the people. The reviews of people on social platforms like Twitter. Further, the project would like to visualize how brands are welcomed across regions and thus help people to make decisions.

### 1.1 Problem Statement

Social media platforms such as Twitter provide space for expression and opinions, where users discuss various events, services, and brands. These reviews are crucial for businesses to thrive and improve their service and quality of the product. However, due to the bulk amount of data, it's difficult to detect the consumer's opinions. The main objective of this project is to explore Twitter data for recognizing customer sentiments about fashion brands and to analyze their overall perception of the brands. Fast fashion brands(Shein, H&M) and established brands(Nike, Adidas) are considered and users' tweets related to these brands are analyzed in three categories - positive, negative, neutral. The results from this project suggest that social media such as Twitter can serve to be the repository of consumer sentiments and opinions.

# 2 **Analytical Framework**

## 2.1 **Data Processing**

### 2.1.1 **Data Collection**

We collected our data from Twitter using Tweepy, a Python library for accessing Twitter API. A total of 175,000 tweets of different brands forms the dataset of this project. Data consists of tweets from 1st December 2021 to 8th December 2021 of 10 different clothing brands such as Nike, Gucci, Adidas, Chanel, H&M, Zara, Shein, Victoria's Secret, Boohoo, and Asos. The brand name was used as a keyword to retrieve tweets.

We stored the results in a data frame.We extracted the following field values from the tweet: author_id,text, geo,created_at,retweets,replies,likes,quote_count and brand.The response from the API is stored in a brand-specific csv file.

We passed the following query parameters to extract data required for our dataset.

**query** = '(Brand_name) -is:retweet lang:en',

**user_fields** = ['username', 'public_metrics', 'description', 'location'],

**tweet_fields** = ['created_at', 'geo', 'public_metrics', 'text'],

**start_time**= 'tweet range start time'

**End_time**= 'tweet range end time'

### 2.1.2 **Data Preprocessing**.

Data from the CSV file was preprocessed by removing Emoticon, symbols, pictograph.Removed rows with null values, filtered out non-english tweets and extracted only the necessary data features.

### 2.1.3 **Data Description**

Sample preprocessed tweet from our dataset looks as follows

| **author_id** | **text** | **geo** | **created_at** | **retweets** | **replies** | **likes** | **quote_ count** | **brand** |
|---|---|---|---|---|---|---|---|---|
| 1225574487091023872 | "OMG OMG, I ACTUALLY HAVE LINGERIE THAT FITS ME AND LOOKS GOOD ON ME NOW !!!!!! thank you sooooo much @SHEIN_official i love eveything that i bought !! | | 2021-12-06 22:51:53+00:00 | 0 | 0 | 1 | 0 | shein |

.

### 2.1.4 **Data Integration**

We create Sparksession which is the the entry point to programming Spark with the Dataset and DataFrame API.We used Spark SQL to read a CSV file into Spark DataFrame and dataframe.write.csv to save or write to the CSV file

We initialize Spark session needs and with the help of SparkSession, DataFrame can be created and registered as tables. Moreover, SQL tables are executed, tables can be cached, CSV data formatted files can be read.

## 2.2 **Semantic Analysis**

### 2.2.1 **Sentiment Analysis Algorithms**

#### 2.2.1.1 **Afinn**

Afinn is the popular lexicon used for sentiment analysis. In python, there is an in-built function for the lexicon. In order to avoid creating different contexts, a spark session was created with the name “TwitterSentimentAnalysis”. With the text classification, the polarity and the subjectivity of tweet content of all the brands were checked. It was noticed that Afinn calculates the sentiment of tweet content from a polarity range of -50(negative sentiment) to +50(positive sentiment). The spread of polarity was much higher in the Zara, Chanel, Gucci, H&M had positive sentiments and Boohoo compared to the Nike, Gucci, Adidas, Chanel, Victoria Secret, Shein, Zara, Asos brands had negative sentiment.

#### 2.2.1.2 **nltk. sentiment.vader**

In order to find the probability of the sentiment to be positive, the probability of the sentiment to be negative, the probability of the sentiment to be positive. NLTK was used to apply data cleansing first where the emojis were removed at first. Analysis was carried out with the data obtained on performing NLTK  by creating the pandas data frame. The columns were added using the lambda function.

## 3**. Visualization**

### 3.3.1 **Geopandas**

Geopandas analyzes geospatial data thus allowing us to analyze twitter data corresponding to its location on a map. For implementing the geopandas we considered two more columns like latitude and longitude. Before plotting the coordinates, a shapefile was used to store the geometric location. Geopandas data frame is used to see where the properties are located in the data frame.  Plotly plots interactive graphs are used to plot the geospatial data on a map of the world. On applying the geopandas to visualize the data we had scrapped for the brand Boohoo, it can be inferred that the positive sentiment is more in the California and East Coast and negative sentiment in Australia and South Africa. For Victoria’s Secret, it can be inferred that the neutral sentiment is in the parts of the USA and high positive sentiment in France, Germany, and UK.

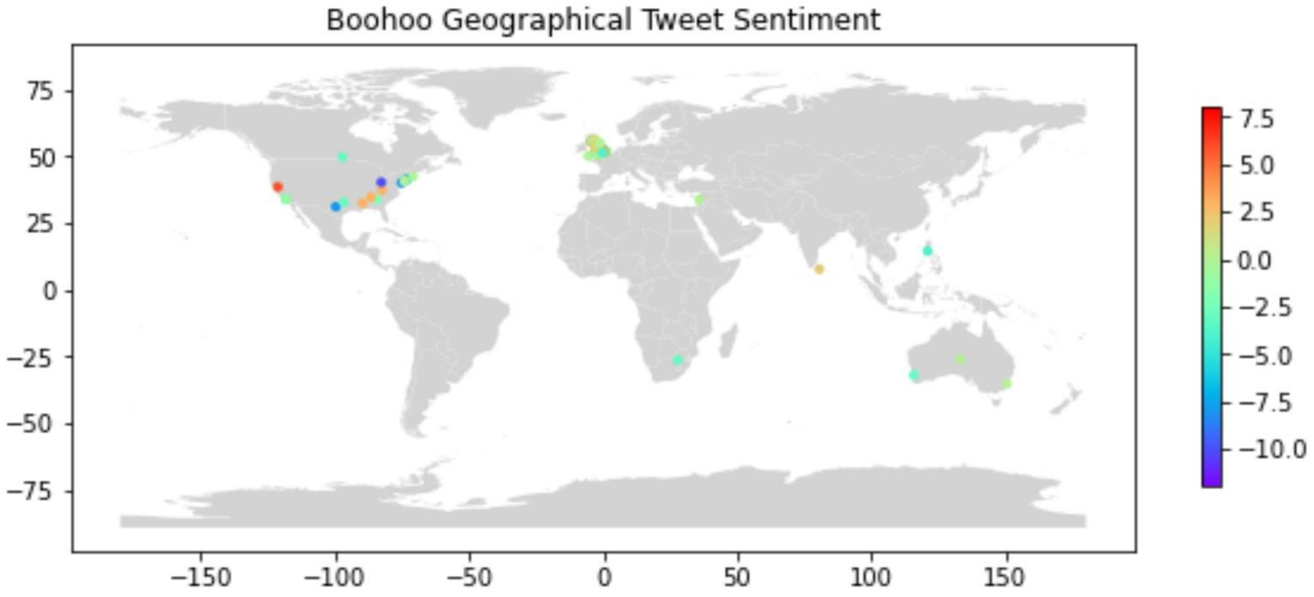

Fig 1: Boohoo Geographical Tweet Sentiment

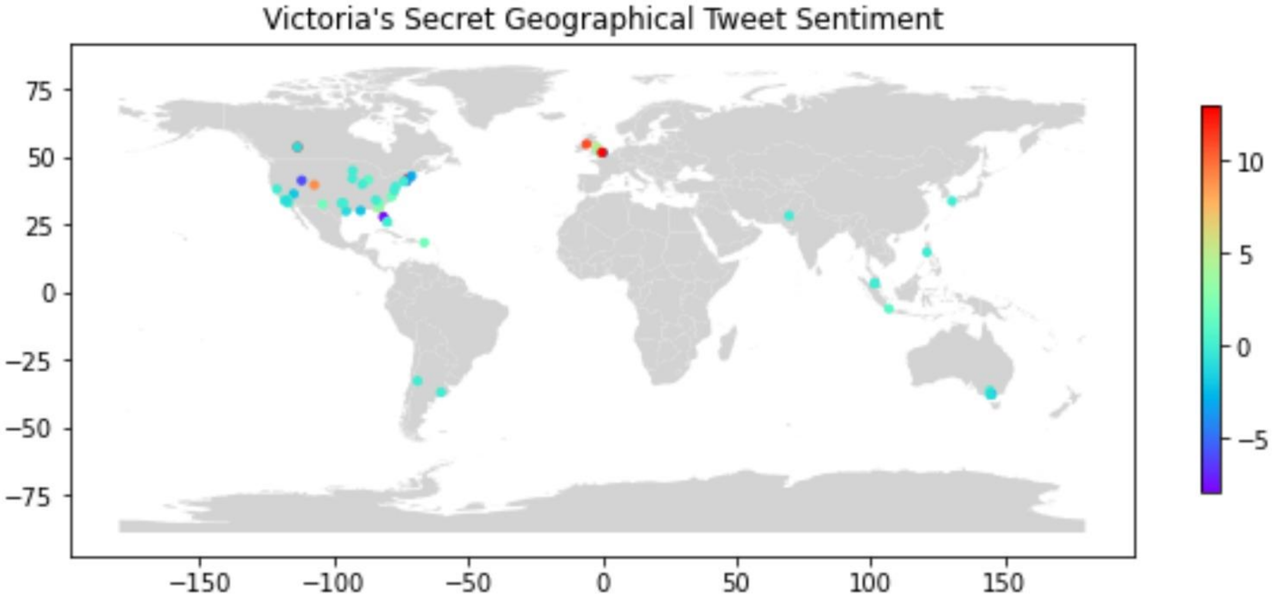

Fig 2: Victoria’s Secret Geographical Tweet Sentiment

### 3.3.2 **Seaborn and Matplotlib**

Python visualization libraries are based wholly or partially on matplotlib. Because of its flexible structure, multiple plots can be drawn into a single image. Seaborn is an open-source python library built to visualize and analyze the data. As visualizing the data is the important part of Big data, by using Seaborn and Matplotlib, we imported graph libraries that are used to visually compare the sentiment analysis of the various brands via graphs.

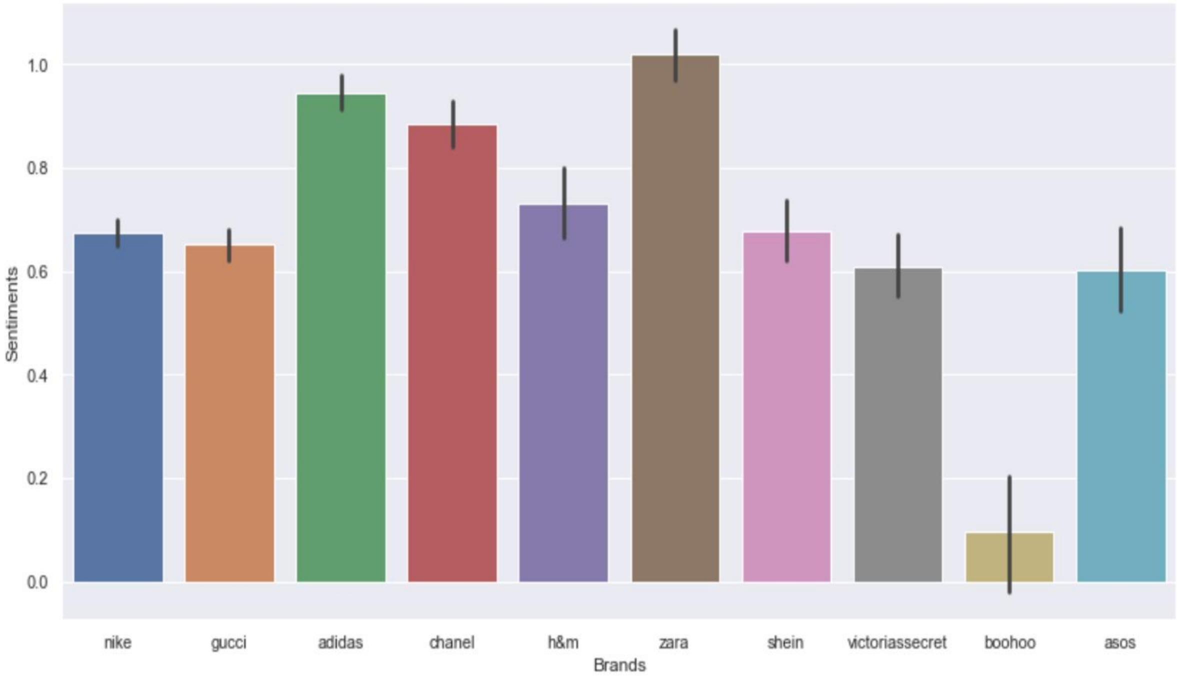

Fig 3: Twitter sentiment Analysis of all the brands

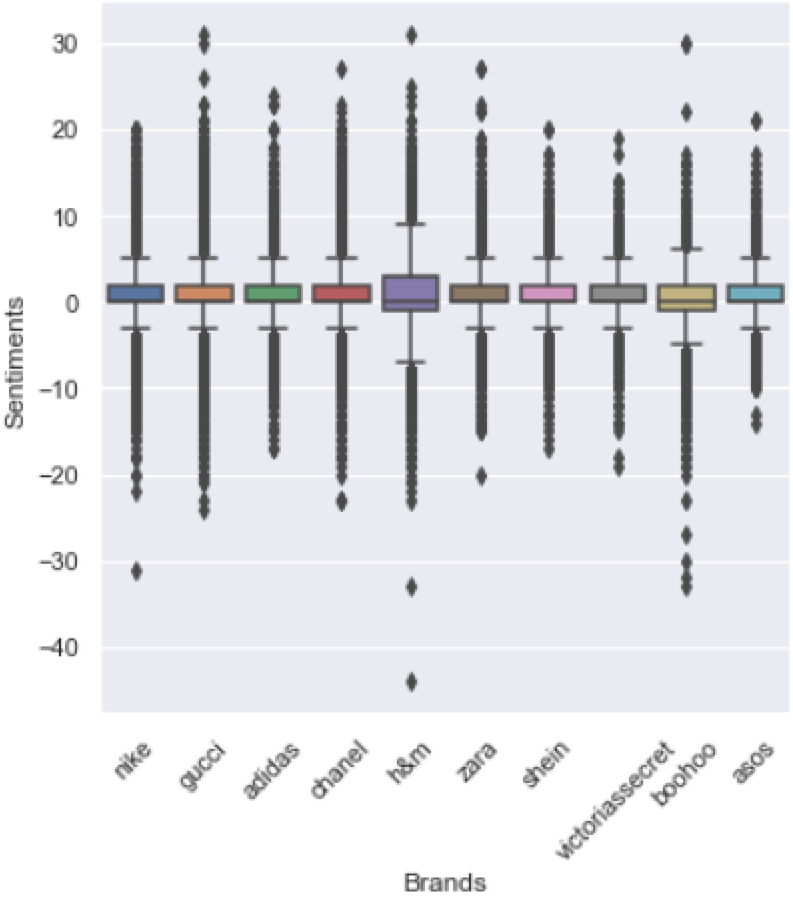

Fig 4: Boxplot for the brands

## 4. **Evaluation**

In the first analysis, we tried to extract tweets using Tweepy for particular fast fashion brands like H&M and Shein. We found that the sentiments were almost similar. So, in order to make it more meaningful and scale the project to a higher level, the number of brands were increased. In the subsequent analysis, we considered about 6 more fast fashion brands like Zara, Forever 21, Uniqlo, Fashion Nova, Boohoo, Missguided. The tweets collected for all the 8 brands including were about 46,343 tweets out of which a couple of brands like Forever21, Missguided, Uniqlo, Fashion Nova showed 100-1000 tweets which were further replaced by mainstream brands like Nike, Adidas, Gucci, Chanel, Victoria Secret, and Asos in the final analysis. This was done for the benefit to measure the scalability of the data collected by increasing the size of the data. With the increased size of data which was about 175,000 tweets, we could get more tweets to work on after pre-processing. Also, the runtime for pre-processing the raw data took about an hour. Thus, this project can be further scaled to compare many brands and since spark is used for data processing, the speed of the processing should also not be affected to a great extent.

## 5. **Conclusion**

This project used around 175,000 tweets. H&M, Zara, Shein, Boohoo, Asos were the fast fashion brands considered along with mainstream brands like Nike, Adidas, Gucci, Chanel, Victoria's Secret. It is seen from the visualizations that the mainstream brands and fast fashion brands have similar sentiments that hover around neutral-slightly positive. One brand with the lowest sentiment was Boohoo which is a fast-fashion, but one of the brands with the highest sentiments was also a fast fashion brand so the comparison between mainstream brands and fast-fashion brands is partially inconclusive. Although, the rest of the fast-fashion brands (H&M, Shein, and Asos) can be seen as having a slightly lower average sentiment than the rest of the mainstream brands (Nike, Gucci, Adidas, Chanel, and Victoria's Secret). So we can conclude that mainstream brands have a more consistent average sentiment over fast-fashion brands which could have either the lowest sentiment, highest sentiment, or be about on par with the mainstream brands.

Based on the geospatial data, we can see that each brand has different sentiments in different parts of the world. In Fig 1 we can see that Boohoo has a more positive sentiment in California, US and has a less positive sentiment in the rest of the world. As for the Victoria's Secret graph in Fig 3, we can see that Victoria's Secret has a more positive sentiment in the United Kingdom than the US.

## 6**. Related work**

Sentiment analysis is a very popular technology in today's world. Now the current works in this area include a mathematical approach which uses a formula for the sentiment value depending on the proximity of the words with adjectives like 'excellent', 'worse', 'bad' or 'positive', 'negative', 'neutral' [3]. Text understanding is a significant problem to solve. Natural language processing can be used to extract Twitter data and the various processes of analysing the sentiments from text data. Public opinion on decision-making is also very important [5]. One of the fundamental processes when extracting twitter data is part-of-speech tagging. It determines tags of words in tweets based on syntactic analysis. In general, the problem in Twitter data can be addressed either through the selection of appropriate tagging methods, the selection of appropriate word representation (feature), or the process of normalizing the informal pattern into a formal pattern [6]. Twitter results are preprocessed by removing duplicate tweets [1]. Thus the sentiment analysis of apparel brands gives a good understanding of participants and the design features preferred by the consumers which in turn improves the marketing strategy of the clothing brands [7]. To filter out the huge number of data we will use the Map-reduce technique [9]. Hadoop is an open-source framework designed to solve problems like processing data and analysis of big data [3]. The geospatial analysis involves georeferenced, GPS, or satellite data that captures surface attributes of a planetary body [11].

.

## REFERENCES

[1] Abdur Rasool, Ran Toa, “Twitter Sentiment Analysis: A Case Study for Apparel Brands” 2019: Twitter Sentiment Analysis: A Case Study for Apparel Brands

[2] “IMPLEMENTATION OF SENTIMENTAL ANALYSIS OF TWITTER DATA-SET FOR APPAREL INDUSTRY” 2020: Implementation of Sentimental Analysis of twitter data-set for Apparel industry

[3] Divya Sehgal, Dr. Ambuj Kumar Agarwal2, “Sentiment Analysis of Big Data Applications using Twitter Data with the Help of HADOOP Framework” 2016: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7894530&tag=1

[4] A Razia Sulthana, A K Jaithunbi, ”Sentiment analysis in twitter data using data analytic techniques for predictive modelling” 2018:
Sentiment analysis in twitter data using data analytic techniques for predictive modelling

[5]Sreeja I, Joel V Sunny, Loveneet Jatian “Twitter Sentiment Analysis on Airline Tweets in India Using R Language” 2020: Open Access proceedings Journal of Physics: Conference series

[6]Endang Suryawati, Devi Munandar, Dianadewi Riswantini, Achmad Fatchuttamam Abka, Andria Arisal “POS-Tagging for informal language” 2018 :Open Access proceedings Journal of Physics: Conference series

[7] Yeong-Hyeon, Seungjoo Yoon,
“Fashion informatics of the Big 4 Fashion Weeks using topic modeling and sentiment analysis” 2019:
Fashion informatics of the Big 4 Fashion Weeks using topic modeling and sentiment analysis - Fashion and Textiles

[8] Rupak Chakraborty, Senjuti Kundu, ”Fashioning Data – A Social Media Perspective on Fast Fashion Brands” 2016: https://aclanthology.org/W16-0407.pdf

[9] Yash Bopardikar, “Twitter data analysis using spark” 2016: https://scholarworks.calstate.edu/downloads/db78tc01m

[10] Silvia Balsi, Lorenzo, “Eco-friendliness and fashion perceptual attributes of fashion brands: An analysis of consumers’ perceptions based on Twitter data mining” 2019: https://www.sciencedirect.com/science/article/pii/S0959652619335711

[11] Alexander Yoshizumi, Megan M. Coffer , “A Review of Geospatial Content in IEEE Visualization Publications” 2020 IEEE Visualization Conference (VIS): https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9331291-