REVA
UNIVERSITY
Bengaluru, India

A Project Report on

Competitive Analysis of Value Fashion Brands in India with Social Media Analytics

Submitted in Partial Fulfilment for Award of Degree of

Master of Business Administration
In Business Analytics

Submitted By
Tharuka Gallekankanamge
R19MBA82

Under the Guidance of
Dr. JB Simha
CTO, ABIBA Technologies

REVA Academy for Corporate Excellence - RACE
REVA University
Rukmini Knowledge Park, Kattigenahalli, Yelahanka, Bengaluru - 560 064 race.reva.edu.in

August, 2022

Candidate's Declaration

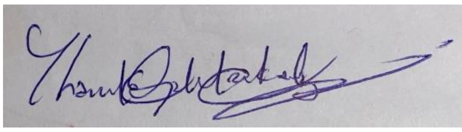
I, Ms. Tharuka Gallekankanamge hereby declare that I have completed the project work towards the first year of Master of Business Administration in Business Analytics at, REVA University on the topic entitled "**Competitive Analysis of Value Fashion Brands in India with Social Media Analytics "** under the supervision of Dr. J. B. Simha. This report embodies my original work in partial fulfilment of the requirements for the award of degree for the academic year 2022.

Place: Bengaluru

Date: 20/08/2022

Name of the Student: Tharuka Gallekankanamge

Signature of Student:



REVA UNIVERSITY
Bengaluru, India

Certificate

This is to Certify that the project work entitled "**Competitive Analysis of Value Fashion Brands in India with Social Media Analytics**" carried out by Tharuka Gallekankanamge with R19MBA82, who is a bonafide student at REVA University, is submitting the first-year project report in fulfilment for the award of first year of Master of Business Administration in Business Analytics during the academic year 2022. The Project report has been tested for plagiarism and has passed the plagiarism test with a similarity score of less than 15%. The project report has been approved as it satisfies the academic requirements in respect of project work prescribed for the said degree.

Signature of the Guide	Signature of the Director
Name of the Guide: Dr. JB Simha	Name of the Director: Dr. Shinu Abhi
Guide	Director

External Viva

Names of the Examiners

1. Harsh Vardhan, Chief Digital Technology Architect, Capgemini
2. Pradeepta Mishra, Director – AI, L&T InfoTech
3. Rajib Bhattacharya, Director – Data & Analytics, Cargill

Place: Bengaluru

Date: 20/08/2022

Acknowledgment

I would like to convey a heartfelt thanks to all my mentors at RACE, Dr. J. B. Simha and Mr. Mithun D J for their continuous support throughout the learning journey. A special mention to Dr. J. B. Simha for his valuable feedback and guidance as a Guide and Mentor throughout the project lifecycle.

I would like to express a special thanks to Dr. Shinu Abhi, Director of REVA Academy of Corporate Excellence for her cordial support, valuable guidance, and information at various stages, which helped in completing the project.

I would like to acknowledge the support provided by the founder and Hon'ble Chancellor, Dr. P Shayma Raju, Vice-Chancellor, Dr. M. Dhanamjaya, and Registrar, Dr. N Ramesh.

I thank my family and my daughter for keeping up with my busy schedule and supporting me in this skill-upgrading journey.

Place: Bengaluru
Date: 20/08/2022

REVA UNIVERSITY
Bengaluru, India

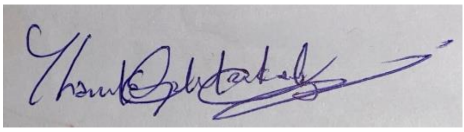
Similarity Index Report

This is to certify that this project report titled "**Competitive Analysis of Value Fashion Brands in India with Social Media Analytics"** in India was scanned for similarity detection. Process and outcome are given below.

Software Used: Turnitn
Date of Report Generation: 19/08/2022
Similarity Index in %: 11%
Total word count: 9685
Name of the Guide: Dr. J.B. Simha

Place: Bengaluru
Date: 20/08/2022

Name of the Student: Tharuka Gallekankanamge
Signature of Student:



Verified by: M N Dincy Dechamma

Signature
Dr. Shinu Abhi,
Director, Corporate Training

List of Abbreviations

Sl. No	Abbreviation	Long Form
1	H&M	Hennes & Mauritz
2	NLP	Natural Language Processing
3	D2C	Direct to Consumer
4	CRISP-DM	Cross-Industry Standard Process for Data Mining
5	CAGR	Compound annual growth rate
6	ARPU	The average revenue per User
7	TF-IDF	Term Frequency-Inverse Document Frequency
8	BoW	Bag of Words
9	HTML	HyperText Markup Language
10	URL	Uniform Resource Locator
11	NLTK	Natural Language Toolkit
12	LDA	Latent Dirichlet allocation
13	NMF	Non-negative matrix factorization
14	BAU	Business As Usual

List of Figures

List of Tables

Abstract

With the rising number of value fashion brands emerging in India in the last decade, it is vital to analyse and understand how the competition engages with their consumer base to understand consumer needs and fashion trends. There is a large amount of content posted by various business establishments and consumers on social media. This study's objective is to extract insights into the "value fashion" industry behaviour in the past four months (January – April 2021) in India, which will help new entrants gain insights and contribute to impactful business decisions.

This study has considered H&M India & Max Fashion and Reliance Trends and utilizes topic modelling machine learning algorithms to evaluate and understand the topics for each of the three brand posts considered for the study. Topics such as Sale, Campaign driven, Product oriented, or a local language is used in the content and have been read and manually labelled. The data set is acquired via web strapping these Brands' Facebook pages.

The suggestions and efforts are for framing a strategic plan for the social media team which is based on the historic data analyzed. Due to the dynamic nature of content created, this study suggests the action be performed once a month by adding the previous month's data to be up to date with what the competition is doing in their respective social media space. The Topic Modelling methods will help the creative teams in understanding the rival brands better and course correct and find niche areas where the new entrants can leverage easily.

Few modelling techniques were explored and the best out of all explored techniques was BoW + LDA. However, to enhance the accuracy, the corpus needs to be enlarged. The scope of this study was limited to four months brand generated content across three major fashion retail giants in India with a total number of posts of 441. Hence the future scope needs to be expanded to make this study more useful.

Keywords: Text Mining, Topic Modelling, Natural Language Processing, Competitive Intelligence, value fashion, Retail, Apparel

Table of Contents

Chapter 1: Introduction

To determine the worthiness of the content posted by brands in the value fashion space it is important to understand the significance and how it helps new entrants to study the mindset of the competition and drive their brand's content, this study has considered three brands.

- H&M: Hennes & Mauritz AB (H&M) is a global conglomerate head quartered in Sweden specialises in apparel retail and well-known for value fashion apparel for men, women, infants, toddlers, and teens. As of per the official website, H&M is present in seventy-four nations with more than five-thousand retail spaces. It is only second to the world's largest global apparel retail giant Inditex from Spain (H&M website, n.d.).

- Max Fashion is part of Landmark group which is head quartered in Dubai. Max is one of many labels owns by this retail giant. Max came into existence in the year 2004, in the Gulf Continent in Abu Dhabi. The Brand was launched in Indian in 2006. Today, Max is the largest fashion brand in the Gulf region, Africa, and Asia. With more than thousand stores spread in nineteen countries. Max fashion's product portfolio includes clothing for men, women, and kids with footwear, home, and accessories (Max fashion India official website, n.d.).

- Trends is truly Indian, home grown, country's biggest fashion retail network present in every nook and corner of India which is a subsidiary of Reliance Industries Limited. Reliance has a vast range of footprint across the country in telecommunication industry, retail, renewable energy, petrochemicals, textiles, and natural resources, etc. Trends put forward a great product range which is trendy, fashionable, pocket-friendly, and high in quality across women's wear, men's wear, kid's wear, footwear, and accessories (relianceretail, n.d.).

The above brand's Facebook posts are analyzed to understand the market gaps & opportunities such as the fashion trends, in-season colours, styles, etc. Also, to keep an eye on how the customer perceives your brand and help in improving business processes and decreasing inefficiencies. Also, identify new market opportunities to expand and know your customers better.

In today's competitive landscape it is not enough to only analyse the numbers and derive whether your business is profitable. It is also important to analyse the competition. It is important to know what the competition is doing and what the pulse of other brands can be directly correlated to the new entrants. Knowing the competition is as important as knowing your customers so your business decisions are not single dimensional though there is a lack of know-how in data gathering in the Indian retail sector it is important to set the steppingstone to open more avenues and opportunities by adapting modern data analytical approaches in Indian fashion retail industry main objective of this project is to study the content posted by the considered brands and conduct topic modelling methods of Facebook data and provide competitor insights.

In the past decade or so the text analytics industry has come a long way, but the Indian fashion retail sector is yet to adopt and is barely scratching the surface when it comes to the methodology of text mining and Natural Language Processing (NLP) to derive competitive intelligence and understand the customers.

The insights derived from the text analytics conducted on the brand posts make this study different from the rest and it gives out vital insights into the fashion industry. This study weighs heavily on the business impact and how techniques such as text analytics, and topic modelling can help brands to move forward and achieve new heights.

The raw data straight from the source have been pre-processed to derive insights and pre-processed to fit the model. An extensive data cleaning process has been conducted to remove/convert special characters such as user tags (@), hashtags (#), emojis, etc. Word features have been dealt with in detail to ensure there are no word repetitions, upper capitalization, word repetition, unwanted punctuation, etc.

Chapter 2: Literature Review

The below articles, journals, and research studies are extensively reviewed to proceed with this study. Both domain and subject-specific research has been conducted to get a holistic view of the main area of focus.

- **State of Fashion 2022:** To succeed in formulating an effective social media strategy which is an exceedingly challenging task since the number of platforms the strategy needs to be adapted to has changed. The target audience on Facebook vs. Instagram is completely different. Generating relevant and timely content for these platforms is a tedious task and it involves a lot of workforce and investment. So, by default, big organizations have the upper hand when it comes to formulating a more impactful social media strategy (McKinsey & Company, 2022).

- **Apparel Trends: 2025 what new business models will emerge?** The Brands can observe how consumers are interacting with their content which directly influences their buying patterns. New-gen companies leverage emerging tech such as Instagram and WhatsApp to sell and generate more revenue. Currently, the fashion industry is run on a year's old trends, and this is changing fast since social influencers are impacting near-time purchase choices via these mediums (Deloitte Digital, 2022).

- **The Road to 2025, Five market, trade, and investment trends that will pave the way for the international textile and apparel industry**: Online domain has emerged strongly in the past few years due to the digital revolution that is shaping up in India. India will be the world's most tech-savvy e-commerce market with exponential growth due to the rapid growth of internet users in the country (Wazir Advisors, 2022).

- **Gaining competitive intelligence from social media data, evidence from the two largest retail chains in the world**: The Social media boom over the past ten years has been the game changer in terms of the way people communicate and organizations carry out their day-to-day operations. The amount of data generated via social media whilst communicating, resharing, and utilizing the content on social media. It is also important to understand this wealth of information generated via various content generated by organizations to derive insights, identify patterns and opportunities, and define new algorithms and approaches to tackle social media content. Social media mining deals with acquiring social media data, analysis of social media networks, and text mining to give an integrated approach to determining social media text mining. This research paves the way to unique problems arising from social media data and uses basic concepts, unfamiliar problems arising due to this data, and effective machine learning techniques in text mining. This research helped in this study since it covers the concepts from basics to advance within text mining (Mohammad et al., 2014).

- **Application of social media analytics: a case of analysing online hotel reviews:** This case study helped this research in identifying the impact of analysing social media content and comparing it on each catalogue listing level. Apart from this, to deeply analyse the digital content, the suggested framework and the results of the case study reveal that there is a strong need to create a pipeline for social media text/content crunching which will in return help in conducting real-time competitive intelligence gathering for social media content. So far there has been only a little research on this area of gathering competitive intelligence. This paper suggested an original approach for gathering competitive intelligence via social media to articulate how companies/brands can positively utilise analytics in social media. (He et al., 2017).

- **A Text Mining Research Based on LDA Topic Modelling:** A massive amount of content is generated every day. It is important to build a system that can effectively search, manage, and explore this content for meaningful insights. This paper has helped in understanding text mining and probabilistic topic model LDA. This paper has used two approaches. Wikipedia articles and users' tweet topic modelling to be exact. The first build up a document topic model, looking for a subject-specific solution on recommending and articles whilst searching. The next one initializes setting up a user-specific topic model by giving a complete study and an analysis of Twitter users' intent. This process contains collecting data, text normalization, and training the model. This process is referred to whilst conducting the modelling phase. This paper was useful research in this study (Tong & Zhang, 2016).

- **Deep NMF Topic Modelling**: Nonnegative matrix factorization (NMF) based topic modelling techniques wouldn't depend on the model or on data statements. Hence, it is typically created as one of the complicated optimization challenges. High computational complexity and bad local minima are two drawbacks that may occur due to this. This research paper has proposed a Deep NMF (DNMF) topic modelling framework to overcome these issues. An unsupervised deep learning method is initially applied to learn latent hierarchical document structures, with the belief that it could have learned a good representation of documents by, e.g., a deep model, post this the topic word discovery problem can be improved. Then, it takes the output of the deep model to constrain a topic-document allocation for the discovery of the discriminant topic words, which not only improves the effectiveness but also reduces the computational complexity over traditional unsupervised NMF models. This study restricts the topic-document allocation in three aspects, which takes the advantage of three major sub-sets of NMF, basic NMF, structured NMF, and constrained NMF, respectively. To overcome the drawbacks of deep neural networks in unsupervised topic modelling, this research has adopted a non-neural-network deep model which is a multilayer bootstrap network. This study has evaluated the suggested method with various representative references covering key sections of topic modelling on a range of actual text content. Investigational results explain the usefulness of the suggested techniques under several evaluation metrics (Wang & Zhang, 2021).

- **Social media competitive analysis and text mining: A case study in the pizza industry:** social media have been embraced by many organizations. Many organizations are utilizing social media techniques such as Twitter and Facebook to offer several facilities and interact with their user base. Due to this, a massive amount of user-generated content is freely available on social media sites. To improve rival gain and efficiently assess competitiveness of organizations, firms must examine and deeply analyse not only the customer-generated content on their social media sites but also the textual brand-generated content on their competitors' social media sites. To help businesses recognize how to achieve a social media competitive analysis and transform social media data into knowledge for all stakeholders and marketing professionals, this paper defines a deeply evaluated case study that applies text mining to analyse structured text content on Facebook and Twitter sites of the three major pizza chains: Pizza Hut, Domino's Pizza, and Papa John's Pizza. The results show the value of social media competitive analysis and the strength of text mining as an effective technique to extract business benefits from an immense amount of available social media data. This research has also given recommendations to assist businesses to improve their social media competitive analysis strategy (He et al., 2013)**.**

- **Social media analytics, competitive intelligence, and dynamic capabilities in manufacturing SMEs:** This research suggests how Small and Medium Enterprises (SMEs) can improve their multi-tasking abilities via learning mechanisms such as social media analytics and competitive intelligence processes that include the preparation, gathering, examination, and knowledge distribution. This research focuses on the impacts of social media analytics on four stages of competitive intelligence to enhance the dynamic capabilities of manufacturing SMEs. Larger companies know how in using social media analytics is better compared to SMEs. However, limited experiential studies have used the multi-tasking abilities approach and examined the fundamental connection between competitive intelligence processes and use of social media analytics. The data collected were examined using structural equation modelling. These research findings show that social media analytics positively influence four phases of competitive intelligence, especially the phases of collection and analysis (Hassani & Mosconi, 2022).

Chapter 3: Problem Statement

Fashion brands like H&M and Trends & Max fashion India are part of this study and engage with their follower base via various social media mediums as a part of their digital marketing strategy to promote their products and acquire new consumers. A new entrant into the market can conduct competitive research to leverage the wealth of information generated by the competition to understand what other brands do as well as this activity can help in getting the flavour of the consumer base.

This study is extracting insights into the Value Fashion Industry behaviour in four months (January – April 2021) in India which will help novel entrants to gain insights as well as contribute to impactful business decisions. The problem this study is addressing is to provide meaningful direction to a new entrant who has recently entered the Indian value fashion industry to get a sense of the competition's social media presence on Facebook.

Consumers follow a brand on social media to get inspired and potentially help in their buying decisions. Hence brands need to keep their followers engaged by constantly developing content and using this as a medium to promote the latest range of products. A new entrant to the industry can easily access this data to understand how the competition is geared to serve its customers and identify the strengths and weaknesses of the competition. There is an ample amount of industry-specific content that gets regularly. As part of this study, the focus is given to extracting valuable insights from the brand-generated posts and developing a model that will keep a close watch on what the near competition is posting which will in return help the new entrant in formalizing its own social media strategy.

This information and insights will help the company's marketing, product, and digital sales teams. The study will focus on analysing historic information posted by the brands and it will help the marketing team in formulating the marketing strategy and pave the way to form a social media plan. The product team will understand what the competition is offering the customers and what are the trends the competition is following. This helps in understanding whether the new entrant is ahead, on par, or lagging compared to the competition regarding our product offering. The Digital sales teams can identify modes of promoting the products and capsule collections to focus on and generate more revenue for the company.

To conclude, the problem covered in this study is to derive insights that would be useful for new entrants in the Indian value fashion domain to understand the competitor's strategy in connecting with their customers. What do they post on social media? Are they highlighting their latest product range? Are they promoting the offers and discounted merchandise? Is it country specific? This study is solving the need of new entrants to understand their competitors better and frame their strategies to enter the Indian market.

Chapter 4: Objectives of the Study

This study focuses on three main objectives. Those are,

- Leverage on text mining and topic modelling methodologies to derive insights for strategic decision-making for new entrants in the Indian value fashion space.

- Derive specific insights related to pricing, branding & range of products.

- Utilize topic modelling techniques to identify the main topics and develop a model that can be used in the future.

By using competitor's Facebook post content. It has become a mandate to be on social media if you want to claim that you are a Direct to Consumer (D2C) Brand in this digital day and age. Also, it is important to have engaging and relevant content so the follower base can refer to and relate to the posted content.

Text mining techniques are used for cleaning the data taking only valuable information that would help in deriving industry-specific insights in topics such as sale/ offer driven, campaign, product specific or uses local language, etc. The sale or offer-driven content would contain communication regarding discounted merchandise and offers (summer t-shirts flat 20% off), Buy 1 get one free, buy products worth INR 999/- get an additional 10% off, etc.). Campaign-specific posts would be about a certain theme (Women's Day, Holi, Independence Day, Diwali campaigns run by various brands), Product related content is the best way for retailers to promote the product range, talk about the latest trends to ensure the products to reach its potential customers. Also, it is important to localize the content to ensure it is more relatable to the audience. All the above diverse types of posts generated by brands will help the new entrants to form r own content as well as a solid social media plan.

Chapter 5: Project Methodology

This study has been conducted as per Cross Industry Standard Process for Data Mining (CRISP-DM) methodology. This method has been adopted to derive insights from brand-generated content and used for topic modelling. The CRISP-DM methodology is a well-adapted process model across various domains, and it has six phases which is a step-by-step approach to any data science problem.

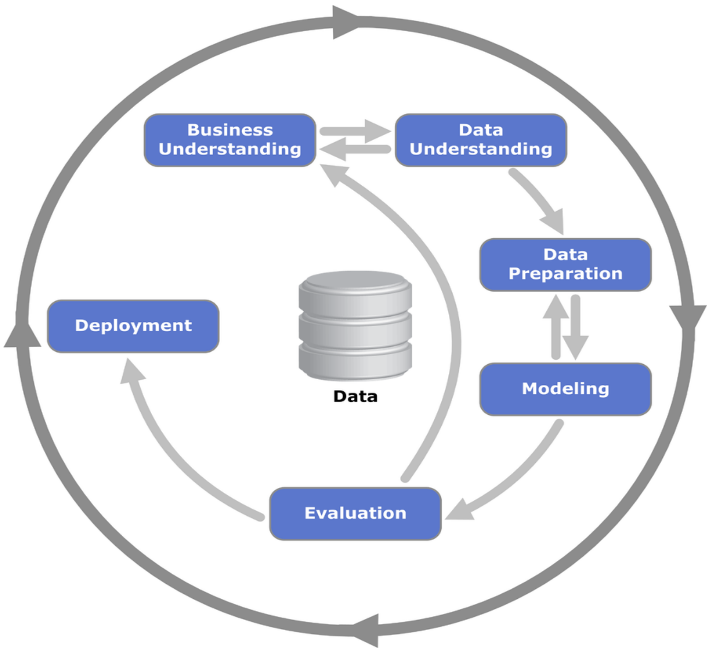


Figure 5.1: Cross-industry standard process for data mining Lifecycle (IBM, n.d.).

The six phases mentioned above are thoroughly explained below.

Business Understanding:

The first step of the Business Understanding phase helps in identifying the objectives and purpose of the project. Once the objective and the scope are framed, it helps in assessing the ground reality, defining the set of goals for text mining and topic modelling, understanding the project plan, assessing how the industry and the domain react, and getting a pulse of the business. This is by far the most crucial step before initializing any data science problems.

Data Understanding:

After thoroughly understanding the business, the next step is data understanding. At this stage, needed information is collected about the data. The goal is to identify, gather, and deeply examine the available data. This means observing the available data, exploring the data, and understating various data points, labels, and features. It also includes analysing the data for potential quality issues which may later hamper the quality of the results and the objective of this study. This includes looking at data types and sources and whether it is structured or unstructured data and planning for the approach.

Data Preparation:

This phase is the most time-consuming and vital as it sets the correct precedent for the next steps and prepares the data for the evaluation. The data gathered are pre-processed, cleaned, and observed for any outliers, anomalies, biases, or missing g information are prepped properly based on the domain. Also, the data is either standardized to get rid of biases and feature engineering is conducted to gather fruitful insights. This would also include combining various other data sources and points and restructuring the data frame to perform the next steps.

Modelling:

Text mining and machine learning techniques are applied in the modelling phase, where the data models are created based on data and the business goals. These models could be based on Topic modelling, Text summarization, or sentiment/ emotion prediction.

Model Evaluation:

At this phase, the strength of the model is determined. Depending on the desired output and business goal, the results and findings of the model are tweaked if needed. This also helps in preparing for the next steps in the approach. The results of the model are assessed against business goals and the desired insights.

Deployment:

This is the final stage of CRISP-DM methodology, where on the successful completion of the evaluation stage, the business deploys the model into live projects to examine the results. This also helps in planning for future projects, looking at the business benefits as they meet the desired goals and the roadmap for the scope. It assesses what is the need of the hour and learns and plans the growth.

Chapter 6: Business Understanding

For any data science project, before assessing the data it is especially important to thoroughly appreciate the industry and the pulse of the business. This will help in formulating a better result. This study is based on a data set that comes from the Indian fashion retail domain. Below are a few facts from the Indian fashion retail industry which was acquired via a fashion e-commerce report (Statista Digital Market Outlook, 2021).

- Projected income in the fashion segment is US$19.69 billion in 2022.
- Estimated Compound Annual Growth Rate (CAGR) is (2022-2025) 18.92%.
- Projected market volume is US$33.11 billion by 2025.
- Expected number of users in the fashion segment is expected to be 446.2m users by 2025.
- User penetration will be at 22.8% in 2022 and hit 30.9% by 2025.
- ARPU (Average Revenue Per User) will amount to US$61.46.

Domain Explanation:

The Online sales market share of fashion includes D2C sale of apparel (menswear, womenswear, and Kidswear), footwear, luggage, and bags, and accessories (socks, caps, and jewellery) by a medium which is online. The mode of sales in this market share includes e-commerce retailers such as Myntra, AJIO, amazon, etc.

The categories considered under value fashion which are in scope for the study are as below.

- Apparel and footwear
- Watches, jewellery, and other accessories (e.g., hats, scarves)
- Eyewear
- Luggage and bags
- Leather goods (e.g., leather bags, shoes, and belts)

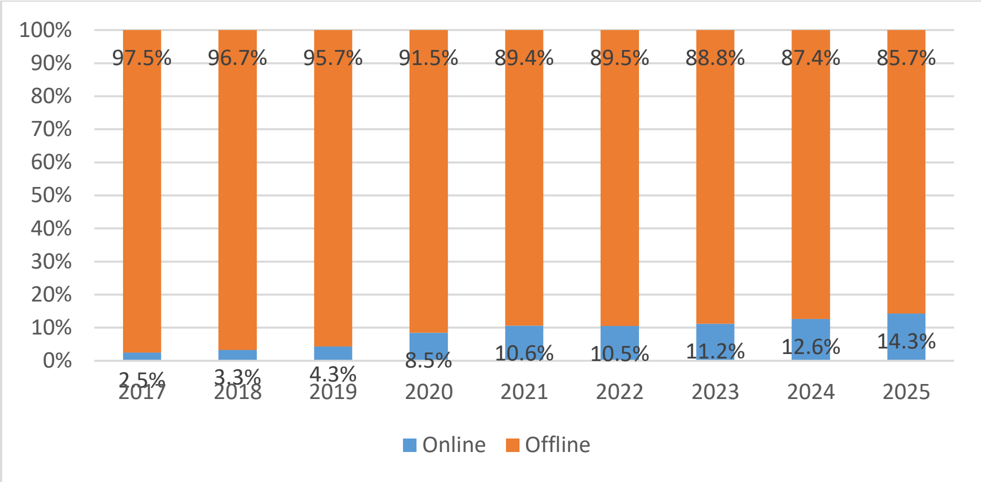


Figure 6.1: Sales Market Share by Chanel, Source: (Statista Digital Market Outlook, 2021)

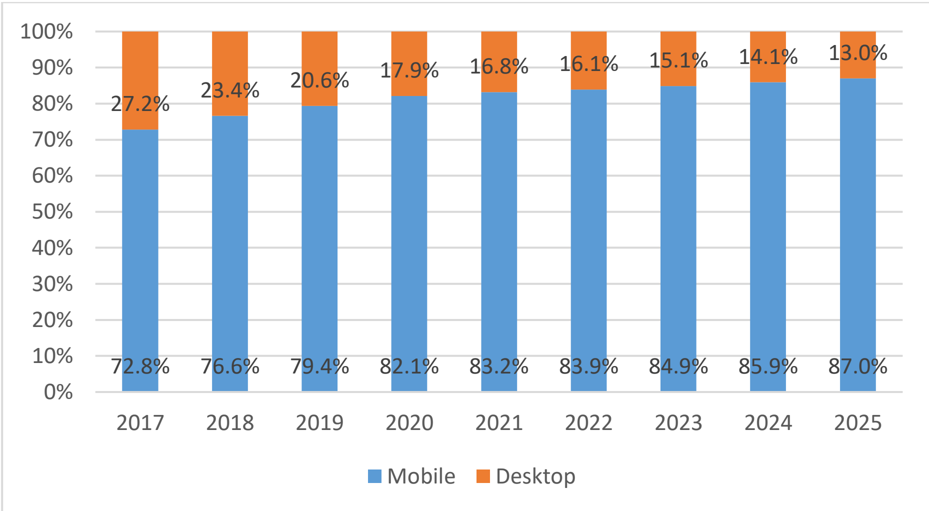


Figure 6.2: Desktop vs. Mobile Market share, Source: (Statista Digital Market Outlook, 2021)

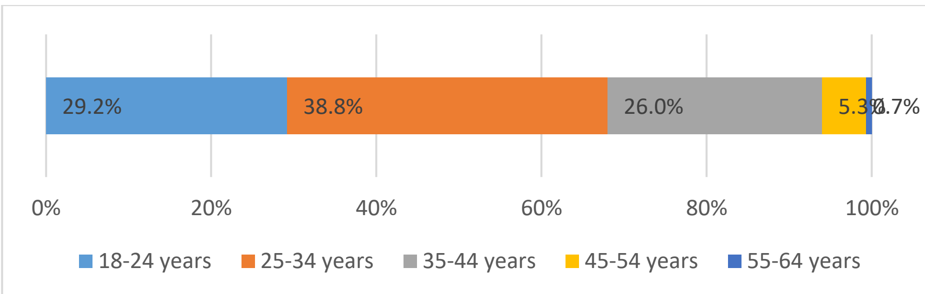


Figure 6.3: Market share by age group, Source: (Statista Digital Market Outlook, 2021)

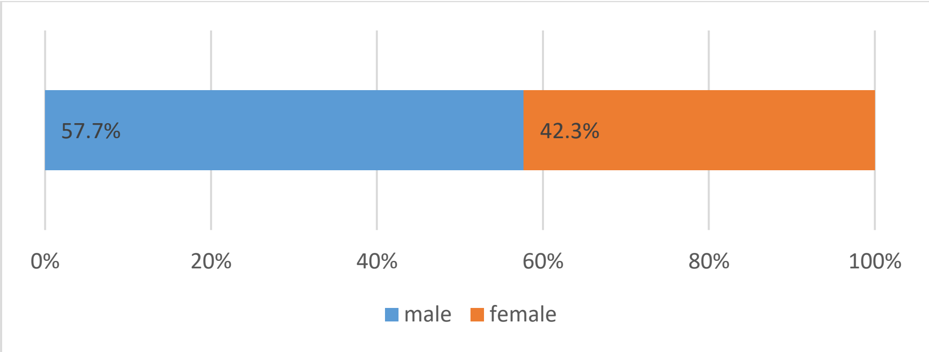


Figure 6.4: Market Share by Gender, Source: (Statista Digital Market Outlook, 2021)

Indian fashion retail domain is expected to achieve great heights by 2025. As per Figure 6.1 Online market share is exponentially growing and continues to grow. Sales via mobile devices are increasing hence the desktop share will keep reducing and this is depicted in Figure 6.2. This is a promising figure for the online sales channel.

Sixty-eight percent of the Indian shoppers are below the age of 34 years as per Figure 6.3 and most shoppers are male which is 57.7% and this is shown in Figure 6.4. This generation is tech-savvy and is most of the workforce hence the focus should be given to these age groups and the male segment.

Post understanding the industry and the domain it is also important to understand the organizations. The data belongs to three retail giants that are operating their business in India. The below points summarise a quick organization introduction post reading about these companies on their official websites.

- H&M: Hennes & Mauritz AB (H&M) is a global conglomerate head quartered in Sweden specialises in apparel retail and well-known for value fashion apparel for men, women, infants, toddlers, and teens. As of per the official website, H&M is present in seventy-four nations with more than five-thousand retail spaces. It is only second to the world's largest global apparel retail giant Inditex from Spain (H&M website, n.d.).

- Max Fashion is part of Landmark group which is head quartered in Dubai. Max is one of many labels owns by this retail giant. Max came into existence in the year 2004, in the Gulf Continent in Abu Dhabi. The Brand was launched in Indian in 2006. Today, Max is the largest fashion brand in the Gulf region, Africa, and Asia. With more than thousand stores spread in nineteen countries. Max fashion's product portfolio includes clothing for men, women, and kids with footwear, home, and accessories (Max fashion India official website, n.d.).

- Trends is a truly Indian, home grown and country's biggest fashion retail network present in every nook and corner of India which is a subsidiary of Reliance Industries Limited. Reliance has a vast range of footprint across the country in telecommunication industry, retail, renewable energy, petrochemicals, textiles, and natural resources, etc. Trends put forward a great product range which is trendy, fashionable, pocket-friendly, and high in quality across women's wear, men's wear, kid's wear, footwear, and accessories (relianceretail, n.d.).

Chapter 7: Data Understanding

Data understanding and transforming the data to fit the requirement is the most time-bound activity in the entire project. This is a vital activity since the data frame is used in consecutive activities to derive insights and topic modelling. It is best to review the data set thoroughly and understand it before moving on to any data transformation.

So, what does it mean when we say “understand” the data? The objective of this step is to understand the attributes of the data and summarize and derive the essence of the data by identifying key characteristics, such as the volume of data and the total number of variables/attributes in the data. Understanding if there are any issues with the data, such as missing values, inaccuracies, and outliers is vital at this stage.

This study has considered the below sub-sections under “Data understanding”

- Data Source
- Data Pipeline
- Data Exploration

Data Source:

The data set is gathered via web scraping the Brand’s (H&M, Max fashion India & Reliance Trends) official Facebook pages. The timeline considered is from January 2021 to April 2021.

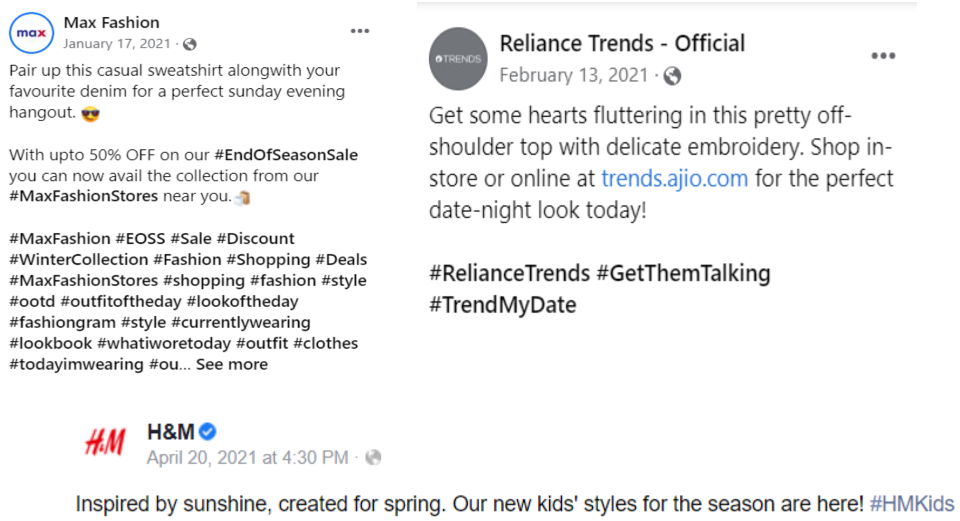


Figure: 7.1: Examples of brand-generated posts (H&M, n.d.; Max Fashion India, n.d.; Reliance Trends, n.d.)

Figure 7.1 highlights a few examples of the brand-generated content that is considered for this study. The original data set gathered has five attributes/columns and it is mentioned below in Table 7.1.

Columns	**Definition**	**Data type**
Brand	Which Brands posted the content	Categorical
Date	Date content was posted	Time
Posted	Content posted by each Brand	Var Char
No. of comments	No. of comments the post received	Numeric
No. of likes	No. of likes the post received	Numeric

Table 7.1: Data explanation of gathered data

The above table summarizes the data considered for the study. Brand Column tells us which brand posted the content and it is categorical, and posts are spread across three Brands (H&M, Max Fashion India & Reliance Trends). The Date column captures the exact date the post went live on social media. The "Posted" column is the most important column where the content posted by brands is captured. This is the column used for text mining and topic modelling. No. of Comments and No. of Links are columns that are capturing user engagement as numerical figures.

Data Pipeline:

The concept of a data pipeline is to summarize how the data gets transformed to arrive at the result. It is always good to jot down the transformation in steps at the "Data understanding" phase to deeply analyses each attribute's connections. The logic behind the data pipeline is to pull all the sentences into a big matrix/ data frame with columns dedicated to each word and either Term Frequency-Inverse Document Frequency (TF-IDF) or Bag of Words (BoW) as the observation's values under each word. Usually, text mining projects such as this study would need to repeat the transformations on the data.

Below Figure 7.2 is a customized data pipeline jotted down by me for a unique representation of the text transformation for this study.

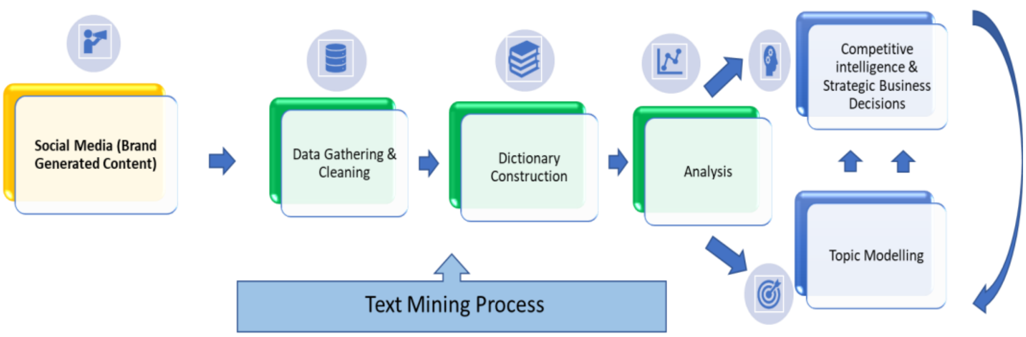


Figure 7.2: Data Pipeline

As depicted in the above pipeline the brand-generated content is used and passed on to the next step which is the data gathering and cleaning phase. Then the dictionary construction and analysis are conducted to derive competitive intelligence and use these insights for strategic business decisions.

After the pipeline is drafted, the direction becomes clear and easy to deploy.

Data Exploration:

This study has derived some valuable insights that would affect the future decision-making process of a new entrant into the Indian retail space.

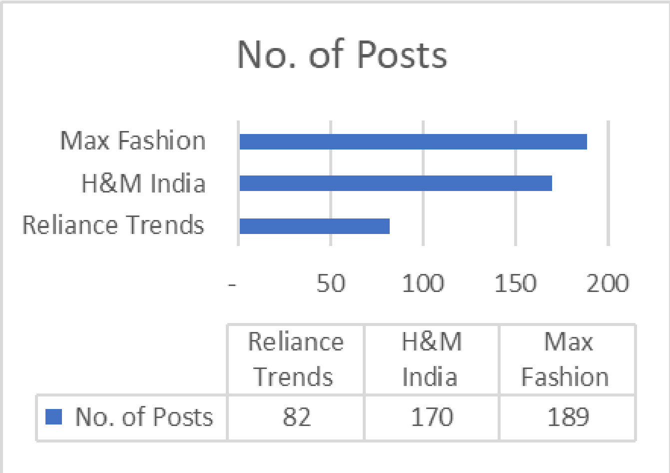


Figure 7.3: Number of posts generated by each brand

As per Figure 7.3. during the same period of four months, Max Fashion India posted most of the content followed by H&M & Reliance trends.

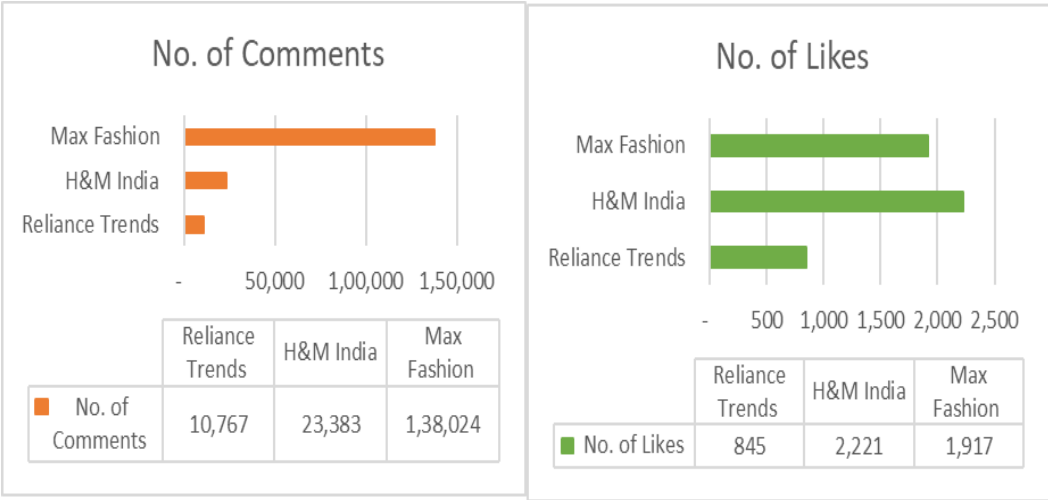


Figure 7.4: No. of comments & likes generated by each Brand

Several comments and likes to depict how much the follower base has engaged with the posted content. More comments and likes translated into posts being viewed and connected to the right audience. As shown in the graph Max Fashion India has cracked the code in engaging with their audience followed by H&M and Reliance Trends.

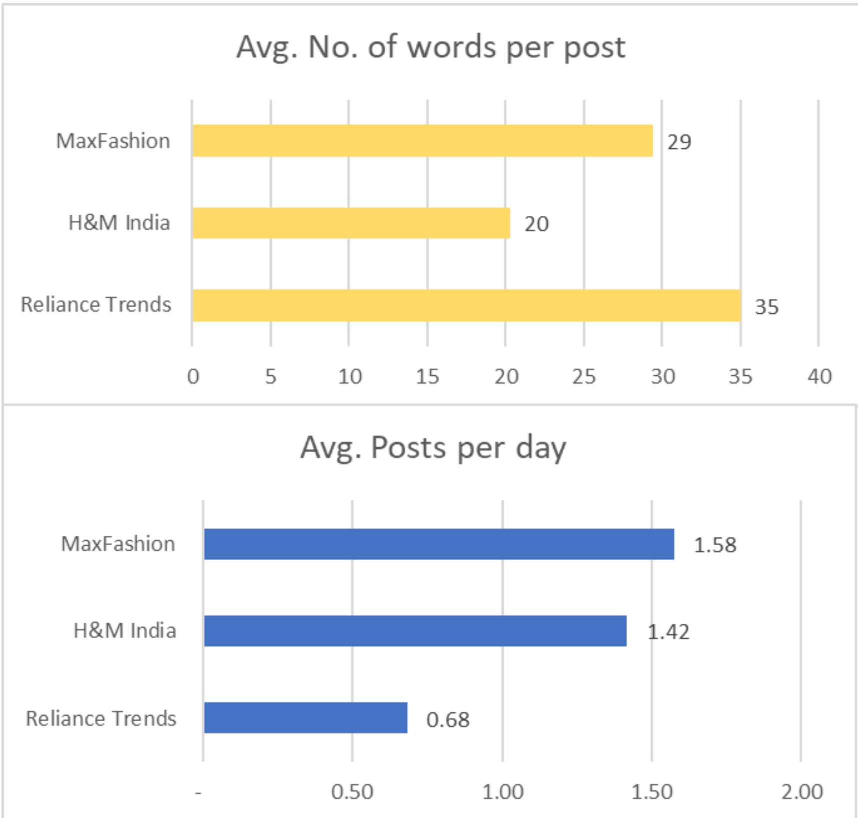


Figure 7.5: No. of words per post and average posts per day by each Brand.

As per Figure 7.5, Reliance Trends posts have the maximum no. of words. However as per Figure 7.4. Reliance Trends have the lowest follower engagement via comments and likes. With this evidence, this study concludes that lengthy content does not get noticed and gets lost within millions of other contents. The new entrant should consider the approach of Max Fashion and H&M where the posts are crisp and precise.

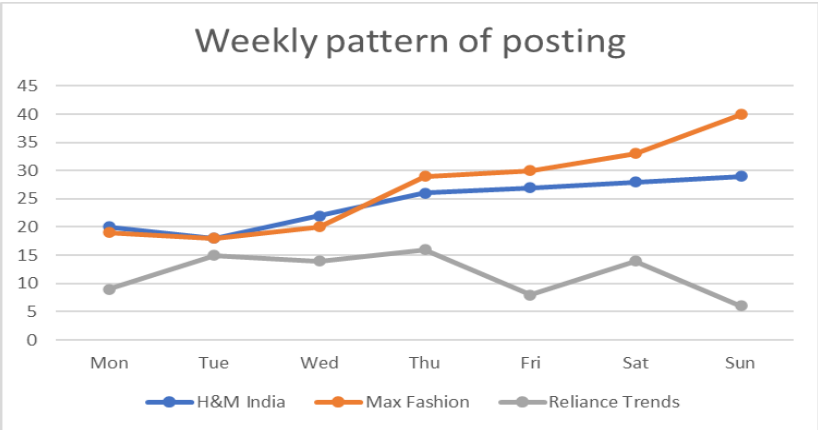


Figure 7.6: Weekly pattern of posting by each brand

The above graph shows the weekly pattern to the number of content posted by the three brands during the four months H&M has a structured approach where the content is posted more towards the weekend. Max fashion India has a similar approach as well. Reliance trends have a contradicting approach where the posting reduces during the weekend. This provides evidence as one of the reasons for the engagement to be low for Reliance trends as per Figure 7.6.

To conclude data exploration the study's findings are summarised below.

- Average word count of H&M India & Max Fashions posts is 20 – 29 words. Reliance Trends posts are lengthy and have more than thirty words. These brand's posts are
- H&M India & Max Fashions posts have more consumer engagement since the no. of comments and likes are more for these brands. Reliance Trends posts are having low consumer engagement.

- The frequency of posting content is high for H&M India & Max Fashions India but low for Reliance Trends which is almost half of H&M India & Max Fashions India.
- Reliance Trends are having low engagement due to a lesser number of posts and due to lengthier posts, which are getting lost in thousands of other posts.
- No. of content posted daily for four months in weekly intervals reveals that H&M and Max Fashion have a structured approach where the no. of content gradually increases towards the weekend and Reliance trends have a contradicting approach where the no. of posts is lesser. This is also evidence that Reliance engagemiso is lesser.

Apart from the above data exploration, to find specific insights, the normalised text contents are fed into a word cloud using python. Please note that the text normalisation is carried out in the “Data Preparation” chapter.

Figure 7.7: Word cloud of brand-generated posts (Normalised text data)

The insights derived from visualizing the normalised posts via word cloud in Figure 7.7 are summarized below.

H&M India:

- H&M focuses on Brand Building via Facebook. They ensure the Brand Name is mentioned in the posts.
- H&M’s Men’s & Kids category is stronger. There is a lot of focus given to promoting ‘’H&M Man”

- Brand focuses on the seasonal collection and monthly new drops. Hence it Is a multi-category brand with a vast range of products it is important to plan what to promote in capsule drops to represent the vast range of products.
- By focusing on capsule collection focus also adds newness and freshness.
- Product-specific attributes are highlighted in the content such as colours, prints, designs, etc.

Max Fashion India:

- Max Fashion also focuses on brand building via Facebook.
- Seasonal collection-centric approach provides the right representation across the larger product range.
- Style, comfort, fashion, and outfit are the key takeaways from Max fashion's content. Max Fashion focuses on the style quotient and gives importance to communicating regarding the comfort of the products.
- Max fashions also promote their offline stores in the content. They are using Facebook as a traffic driver for their 400+ offline stores across India.
- The Women's category is stronger for Max Fashions and their key focus is on promoting the Women's category.

Reliance Trends:

- There is no specific word/topic that is surfacing from Trend's word cloud.
- Reliance Trends uses collection-specific hashtags (example: #getthemtalking) which are campaign specific to get more reach.
- The content is focused both on online shopping via AJIO as well as on their 300+ offline stores across India.

Chapter 8: Data Preparation

Data preparation is vital in text mining, and it addresses the inconsistencies in data. Before starting pre-processing the data, the source must be identified and reviewed. Then the attributes are labelled as numeric, categorical and character, etc. to understand how to proceed with data preparation. The objective of this phase is to eliminate outliers, and discrepancies while amending the data to derive insights.

Text Normalisation:

In the text mining process text normalisation is a vital step to extracting the essence of the input. Each brand's post contains various properties that need to be treated. In this study, four parameters are considered when the data cleaning and pre-processing phase is conducted. These four steps helped me in deriving the main essence out of the brand posts.

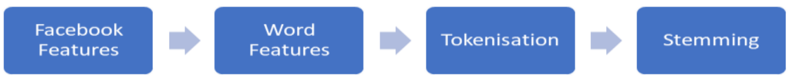


Figure 8.1: Text normalisation process

Facebook Features:

Under each of the above four steps, various activities are conducted. Data acquired from Facebook usually contains HTML (HyperText Markup Language) entities such as "< > &" which comes when you obtain web scraped data on social media. Hence these properties need to be removed. Also, there is @Usertag, and information embedded in the Facebook scraped data. Which need to be removed for further analysis. With the use of python's regex package. these properties have been removed.

Also, there are a few more features that need to be treated when considering Facebook-generated data. Emojis are one of them. Emojis have become a trend and an icon used by the millennial and gen Z population to express emotions such as happiness, dislike, and anger on social media. Hence it is important to derive the meaning of these emojis to extract the essence of these posts. This study utilized an emoji package in python to “demojize” the posts. One more important feature is the URLs 9 Uniform Resource Locator). URL occurrences need to be replaced. The other important feature is the hashtags. The hashtags are a marketing tool to make certain topics that are trendy. These are extremely popular amongst the younger population. Hence a lot of mainstream brands use these as a marketing tool.

Word Features:

After treating the Facebook-generated anomalies, the next step is treating the text anomalies. In text mining, this is a major step to conduct to move on to the next steps. It is important to treat upper capitalization and word repetition. punctuation repetition and word contraction.

Python’s NLTK (Natural Language Toolkit) package was utilised precisely to deal with all text normalisations. All upper capitalized words are converted into lower case words using NLTK. Repetitive words are replaced with a single occurrence. Stop words are considered as a set of single words. To shorten the data and ensure unnecessary space is not allocated for unneeded words a “Stop Word Dictionary” is utilized. All stop words are taken out since it is not useful for this study. Punctuation symbols such as “.”, “,”,”?” are vital to understanding the emotion of a post and other punctuations need to be removed.

“Word contraction” is an important function in text normalisation. In a sentence, there can be certain words such as “they will” which is contracted as “they’ll” which should be normalised. Contractions are a set of words that are shortened by removing certain letters and replacing them with an apostrophe and this will lead to standardization.

To conduct this exercise, the “contractions” package in python was utilised.

Tokenization:

To move onto “Topic modelling” it is important to tokenize the “normalised” data. Tokenization is one of the most popular methods when it comes to normalisation of text data. Tokenization takes care of separating a paragraph, a phrase, or a sentence, or in this case a post generated by the Brand into smaller sections which are words. All these separated words will be called tokens. Tokens can be in the form of words, punctuations, and even n number form.

Original Post (input): Lookback for a moment, thank yourself for getting through this year and get ready for new 365 days, only and only filled with happiness!!

Tokenized Post (output): ['lookback', 'moment', 'thank', 'get', 'year', 'get', 'ready', 'new', 'day', 'fill', 'happiness']

Figure: 8.2: Example of pre- and post-tokenization.

Figure 8.2 highlights the original post and the final tokenised version.

Stemming:

As the last step of text normalisation, “stemming” is conducted and to carry this out the porter Stemmer in NLTK was utilised. Stemming takes care of removing prefixes and suffixes from a word. As per figure 8.2 porterStemmer has stemmed the word “happiness” to “happi”

Under data preparation, the below activities mentioned in figure 8.3 are carried forward to normalise the text data.

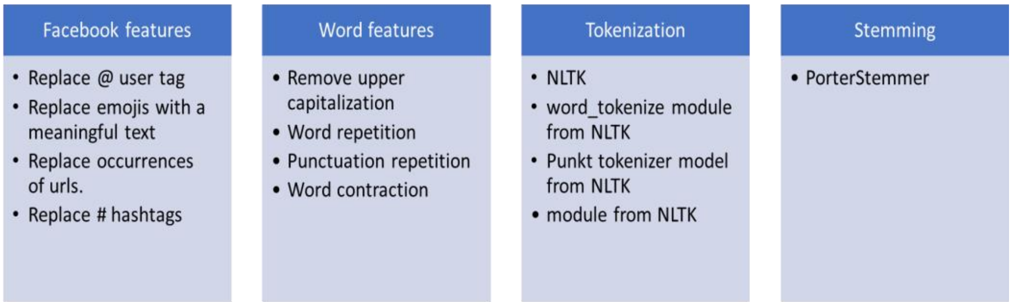


Figure 8.3: Text normalisation process in detail

To carry out topic modelling the brand-generated posts are labelled into four subjects.

- Sale: Posts that were about offers, and discounted merchandise communication was labelled as "Sale"

  Example: It is never too late to grab amazing deals.

  With #NewBeginnings revamp your wardrobe with #MaxEndOfSeasonSale only at our #MaxFashionStore near you.

  Hurry! Get up to 50%

- Season/campaign: Posts that were dedicated to seasonality and marketing communication such as spring and summer collection women's day campaigns were classified under "season/campaign"

  Example: With #NewBeginnings it is time to give your closet a fresh makeover. 🛍

  Shop at the nearest #MaxFashionStore now and avail the amazing collection.

- Product: This label is dedicated to posts that are about the product and the silhouettes.

  Example: Pair up a Hooded Sweatshirt with your favourite Blue Denims, and you are set for the #NewBeginnings.

  Check out the collection at our #MaxFashionStore near you. 🛍

  #HappyWeekendShopping

- Local language: Some posts were posted in local languages such as Hindi, Kannada, etc. Mainstream brands do this to capture their audience and connect with them better. Also, some communications become better in the regional language than translating into English.

    Example: Logo ka kaam hai kehna, log toh kuch na kuch kahnege hi. But all I want to say to all the women out there is #YouAreSoBeauitfull.

    Check out @maxfashionindia to celebrate and appreciate the smart, kind, funny, and lovable woman that is you.

All the Brand generated posts are labelled as per the above topics to carry out topic modelling.

Chapter 9: Modeling

The data set used pertains brand generated post content and topics identified by manually labelling each 441 brand posts of H&M, Max fashion, and Reliance Trends. There are four topics identified whilst labelling the posts. Those are "Product," "Sale," "Season / Campaign" and "Local language." The approach taken in this study is to convert the normalised texts of brand-generated posts to a vector using Bag of Words (BoW) and TF-IDF post this activity each topic is modelled individually with the help of Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF). All the modelling concepts used are extensively explained below.

Bag of Words (BoW):

Bag of Words (BoW) is an NLP tool for transforming any sort of text into numbers that can be processed by a machine learning algorithm. In this study, BoW is used by calling the NLTK python package. The fundamental approach of BoW in the dictionary is set to a word, and each value is set to the number of times the word occurs in the corpus.

BoW models are involved with whether a recognized word appears in a document and the number of times it appears and not the sequence in which it occurs, nor its meaning is analysed. BoW is an important technique in natural language processing, information retrieval from documents, and document classification. The name of this model which is 'Bag of Words' is said to be derived from the well-known word game, Scrabble. The importance of each tile in a Scrabble bag was defined by how often a particular letter emerged on the front page of the New York Times in 1938 (Techopedia, n.d.) Figure 9.1 depicts how BoW works on the text data considered in this study. this study.

Figure 9.1: Example of BoW technique with source data

TF-IDF:

TF-IDF is an abbreviation called as term frequency-inverse document frequency and it is an indicator, used in the NLP domain to retrieve valuable information to conduct machine learning algorithms such as LDA and NMF that can compute the importance of string character representations in the brand post which is the focus of this study.

TF-IDF must be looked at in two angles. TF (term frequency) and IDF (inverse document frequency).

What is TF (term frequency)?

Term frequency functions by seeing the frequency of a certain term related to a certain document. There are several methods, or approaches, for defining frequency:

- Frequency of the word emergence in a document (unprocessed count).
- Term frequency tweaks as per the length of the document (unprocessed count of incidents divided by the number of words in the document).
- Logarithmically scaled frequency (e.g., log (1 + unprocessed count)).
- Boolean frequency (e.g., one of the terms appears, or zero if the term does not appear, in the document)

What is IDF (inverse document frequency)?

Inverse document frequency sees how unique or not a word is within the corpus. IDF is formulated where it is the term (word) used to measure the commonness of and N is the number of documents (d) in the corpus (D). The denominator of this calculation is the number of documents in which the term, t, shows up in the corpus (Capitalone, n.d.).

This study has extensively utilized python’s sklearn package and nltk to conduct the below machine learning algorithms

- BoW + LDA
- TF-IDF + LDA
- BoW + NMF
- TF-IDF + NMF

Latent Dirichlet Allocation (LDA):

In natural language processing, LDA is a generative statistical model that describes a group of observations through unobserved groups, and each group explains why some parts of the data are related. The LDA is an example of a topic model. In this, observations which are words are collated into documents, and each word's existence is attributable to one of the document's topics. Each document will contain a small number of topics (Wikipedia, n.d.-a).

Non-negative Matrix Factorization (NMF):

NMF or NNMF, also known as non-negative matrix approximation is a collection of algorithms in multivariate analysis and linear algebra where a matrix V is factorized into two matrices W and H, with the property that all three matrices have no negative elements. This non-negativity makes the subsequent matrices simpler to examine. Also, in applications such as treating audio spectrograms or muscular activity, non-negativity is essential to the facts being considered. Since the issue is not precisely resolvable in general, it is estimated numerically (Wikipedia, n.d.-b).

All the above natural language processing techniques and machine learning algorithms are executed by using the inbuilt features and packages of python. sklearn and nltk packages came in handy whilst executing the modelling step.

BoW + LDA and TF-IDF + LDA:

Important parameters need to be tweaked whilst executing the LDA machine learning algorithm to get the best outcome. The components = y.shape[1], random_state=11, earning_method="online" were considered to ensure the best outcome. Both BoW and TF-IDF techniques are separately fitted into LDA to see if the results can be better with either one of the techniques.

BoW + NMF and TF-IDF + NMF:

Execute NMF machine learning algorithms n_components=3 and random_state=42 parameters are tweaked and applied. Even in NMF both BoW and TF-IDF approaches were used. This is conducted to see if the results can be better in one method than the other.

Chapter 10: Model Evaluation

In this study, four variation machine learning algorithms i.e., BoW + LDA, TF-IDF + LDA, BoW + NMF and TF-IDF + NMF were built. A topic model algorithm typically has a set purpose. This study is to rightly predict the topic of brand-generated content. Post model building you would want to know how effective and accurate the models are and if it is doing the job it is destined to perform. Hence it is important to conduct the evaluation step.

Evaluating a topic model can help in determining if the model was successful in capturing the essence and the internal structure of the dictionary.

Once you build the model(s) you must thoroughly evaluate and review it to make sure that the model can achieve the desired business objectives.

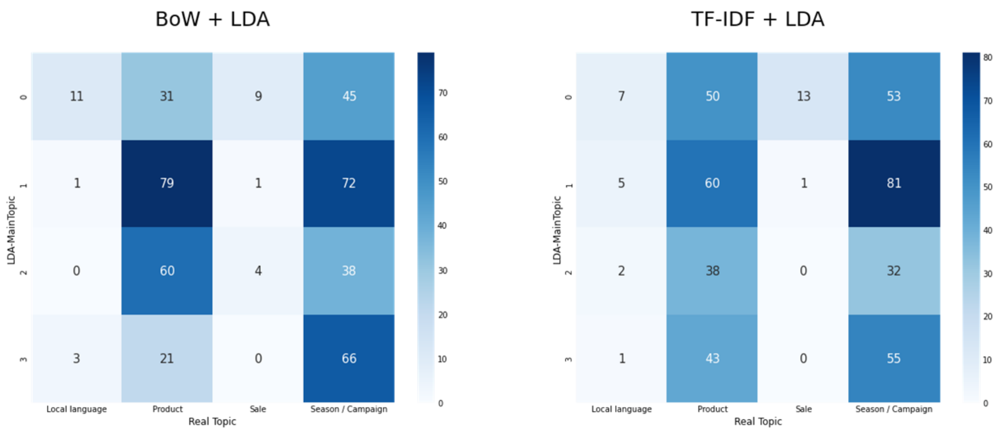


Figure 10.1: Confusion Matrices of BoW + LDA and TF-IDF + LDA

Figure 10.1 depicts the confusion matrices of the BoW + LDA topic model as well as TF-IDF + LDA topic models. Out of these two models, it is evident that BoW + LDA is the best model since the accuracy is higher as depicted in Table 10.1.

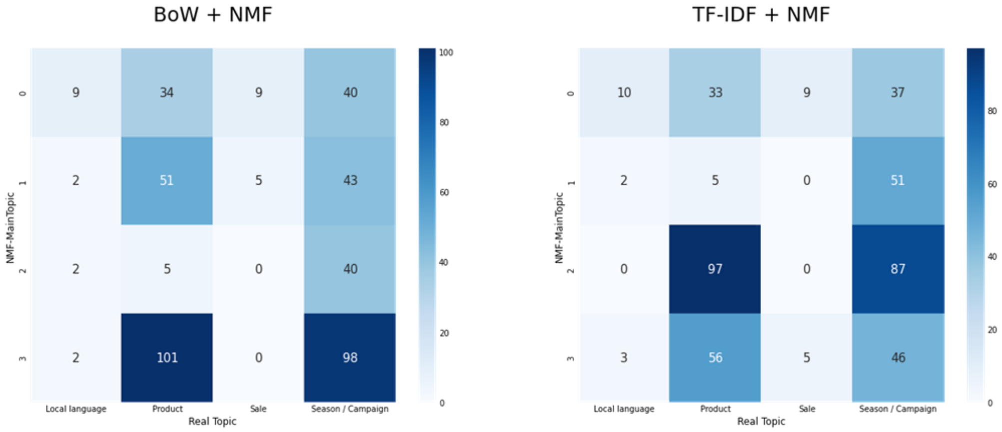


Figure 10.2: Confusion Matrices of BoW + NMF and TF-IDF + NMF

Figure 10.2 portrays the confusion matrices of the BoW + NMF topic model as well as TF-IDF + NMF topic models. Out of these two models, it is evident that BoW + NMF is the top model since the accuracy is greater as shown in Table 10.1.

As per the four variations of models built, the confusion matrix for multiclass classification has been derived to evaluate the performance of the models.

Topic	Accuracy			
	TF-IDF + LDA	BoW + LDA	TF-IDF + NMF	BoW + NMF
Local Language	71.88%	79.82%	80.95%	79.82%
Product	50.57%	57.82%	45.80%	56.92%
Sale	80.50%	75.51%	55.10%	86.17%
Season / Campaign	52.38%	59.41%	45.80%	48.75%

Table 10.1: Model Performance Results

As depicted in Table 10.1 BoW + LDA is giving the best outcome for Product (57.82% accuracy) and Season / Campaign (59.41% accuracy) related posts and comparatively BoW + LDA is giving satisfactory results without a considerable variance for local language (79.82%) and sale related posts (75.51%) as well.

Hence, this study concludes that BoW + LDA is the best approach with the current data set utilised.

Chapter 11: Analysis and Results

To conclude, this study summarises its findings and recommendations below. The below recommendations are passed on to the Marketing team of the "new entrant" to formulate a social media plan as per the below competitive intelligence provided.

- Average word count of a post should be 20 – 30 words.
- The frequency of posting content should be 2 - 3 posts per day.
- As per the analysed weekly posting pattern There should be a gradual increase in posts towards the weekend (Monday being the least and Sunday being the maximum no. of content posted) Hence Suggest posting two posts from Monday to Thursday and Business as Usual (BAU) days and Friday, Saturday, and Sunday to post 3, 4 & 3, respectively.
- A new entrant must focus on brand building via Facebook.
- The stronger category of the new entrant (men's wear, kid's wear, etc.) needs to be the hero of the content to ensure maximum traction is built on the same.
- Customised Seasonal collections and monthly new drops are highlighted to promote the vast range of products. This will also add newness and freshness.
- Product-specific attributes are highlighted in the content such as colours, prints, designs, etc.
- Use this medium as a mode to promote other channels such as offline stores, Ecommerce, etc.
- Utilise brand/collection-specific hashtags (example: #getthemtalking) which are campaign specific to get more reach.

The Supporting data analysis for the above recommendations is depicted in chapter 6, Data understanding.

The above recommendations and inputs are for formulating a strategic plan for the social media team which is based on the historic data analysed. Due to the ever-changing nature of content created this study recommends the activity to be conducted once a month by adding the previous month’s data to be up to date with what the competition is doing in their respective social media space.

The Topic Modelling techniques will help the creative teams in understanding the competition better and course correct and find niche areas where the new entrant can leverage easily.

Chapter 12: Conclusions and Future Scope

As covered in the “Evaluation” phase out of the four approaches taken, BoW + LDA is the best approach. However, to make the model accuracy better in phase two of the same study I would be working towards enhancing the accuracy of this model. To enhance the performance, the corpus needs to be expanded. The scope of this study was limited to four months brand generated content across three major fashion retail giants in India with a total number of posts of 441. Also, the Indian consumers' buying patterns are heavily influenced by regional festivities as well as regional weather situations.

As per the Study conducted the below scope is captured as the future points of discussion.

1. Acquire data for at least three years to ensure all seasonality is covered.
2. Apart from Facebook posts, collect data from Instagram & Twitter.
3. Expand the competitive intelligence analysis to other value players such as easy buy, Zudio, westside, etc.

Bibliography

Capitalone. (n.d.). *Definition of TF-IDF - Capitalone*.

Deloitte Digital. (2022). *Apparel Trends 2025*. https://www.deloittedigital.com/content/dam/deloittedigital/us/documents/blog/blog-20200610-apparel-trends.pdf

Hassani, A., & Mosconi, E. (2022). Social media analytics, competitive intelligence, and dynamic capabilities in manufacturing SMEs. *Technological Forecasting and Social Change*, *175*, 121416. https://doi.org/https://doi.org/10.1016/j.techfore.2021.121416

He, W., Shen, J., Tian, X., Li, Y., Akula, V., Yan, G., & Tao, R. (2015). Gaining competitive intelligence from social media data Evidence from the two largest retail chains in the world. In *Industrial Management and Data Systems* (Vol. 115, Issue 9, pp. 1622–1636). Emerald Group Holdings Ltd. https://doi.org/10.1108/IMDS-03-2015-0098

He, W., Tian, X., Tao, R., Zhang, W., Yan, G., & Akula, V. (2017). Application of social media analytics: A case of analyzing online hotel reviews. *Online Information Review*, *41*(7), 921–935. https://doi.org/10.1108/OIR-07-2016-0201

He, W., Zha, S., & Li, L. (2013). Social media competitive analysis and text mining: A case study in the pizza industry. *International Journal of Information Management*, *33*(3), 464–472. https://doi.org/https://doi.org/10.1016/j.ijinfomgt.2013.01.001

H&M. (n.d.). *H&M official facebook page*. Retrieved August 18, 2022, from https://www.facebook.com/hm/

H&M website. (n.d.). *H&M official website*.

IBM. (n.d.). *Cross-industry standard process for data mining Lifecycle* .

Max Fashion India. (n.d.). *Max fashion official facebook page*. Retrieved August 18, 2022, from https://www.facebook.com/maxfashions/

Maxfashion India official website. (n.d.). *Maxfashion India official website*.

McKinsey & Company. (2022). *The State of Fashion 2022*. https://www.mckinsey.com/~/media/mckinsey/industries/retail/our%20insights/state%20of%20fashion/2022/the-state-of-fashion-2022.pdf

Mohammad, R. Z., Abbasi, A., & Liu, H. (2014). *Social Media Mining: An Introduction*. http://dmml.asu.edu/smm,

Reliance Trends. (n.d.). *Reliance Trends official facebook page*.

relianceretail. (n.d.). *Relianceretail official website*.

Statista Digital Market Outlook. (2021). *Fashion eCommerce report 2021*.

Techopedia. (n.d.). *Bag of Words (BoW)*.

Tong, Z., & Zhang, H. (2016). *A Text Mining Research Based on LDA Topic Modelling*. 201–210. https://doi.org/10.5121/csit.2016.60616

Wang, J., & Zhang, X.-L. (2021). *Deep NMF Topic Modeling*. http://arxiv.org/abs/2102.12998

Wazir Advisors. (2022). *Wazir Report - The Road to 2025*. 1–32. https://wazir.in/pdf/Wazir%20Report%20-%20The%20Road%20to%202025.pdf

Wikipedia. (n.d.-a). *Latent Dirichlet Allocation (LDA)*. Retrieved August 18, 2022, from https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation#:~:text=In%20natural%20language%20processing%2C%20Latent,example%20of%20a%20topic%20model.

Wikipedia. (n.d.-b). *Non-negative matrix factorization*.

Appendix

Plagiarism Report[1]

Competitive Analysis of Value Fashion Brands in India with Social Media Analytics

by Tharuka G

Submission date: 19-Aug-2022 10:57AM (UTC+0530)
Submission ID: 1884245545
File name: ashion_Brands_in_India_with_Social_Media_Analytics_-Tharuka.docx (1.61M)
Word count: 8834
Character count: 47094

[1] Turnitn report to be attached from the University.

Competitive Analysis of Value Fashion Brands in India with Social Media Analytics

ORIGINALITY REPORT

11%	10%	4%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.researchgate.net Internet Source	3%
2	www.capitalone.com Internet Source	1%
3	ukcatalogue.oup.com Internet Source	1%
4	github.com Internet Source	1%
5	en.wikipedia.org Internet Source	1%
6	Submitted to University of South Africa Student Paper	<1%
7	export.arxiv.org Internet Source	<1%
8	www.statista.com Internet Source	<1%
9	Submitted to University of College Cork Student Paper	<1%

10	www.ilsul6ana.com Internet Source	<1 %
11	repository.sharif.edu Internet Source	<1 %
12	www.alianzo.com Internet Source	<1 %
13	Submitted to Tampere University Student Paper	<1 %
14	Submitted to Massey University Student Paper	<1 %
15	mdpi-res.com Internet Source	<1 %
16	Submitted to University of Huddersfield Student Paper	<1 %
17	He, Wu, Shenghua Zha, and Ling Li. "Social media competitive analysis and text mining: A case study in the pizza industry", International Journal of Information Management, 2013. Publication	<1 %
18	core.ac.uk Internet Source	<1 %
19	Submitted to Army Institute of Technology Student Paper	<1 %
20	Submitted to SP Jain School of Global Management	<1 %

Student Paper

21	www.preprints.org Internet Source	<1 %
22	www.science.gov Internet Source	<1 %

Exclude quotes On
Exclude bibliography On
Exclude matches < 10 words

GitHib Link(s):

https://github.com/TharukaG/value_fashion1.ipynb.git &

https://github.com/TharukaG/topic_modelling_bow_tf_idf_lda_nmf.ipynb.gi
t