Mining social sentiments for demand analysis in Footwear industry

A.M.A.N.P.W.M.R.P.D.B. Athurupane

198742X

Dissertation submitted to the Faculty of Information Technology, University of Moratuwa, Sri Lanka, for the partial fulfilment of the requirements of the Degree of Master of Science in Information Technology.

July 2022

Declaration

We declare that this thesis is our own work and has not been submitted in any form for another degree or diploma at any university or other institution of tertiary education. Information derived from the published or unpublished work of others has been acknowledged in the text, and a list of references is given.

Name of Student

A.M.A.N.P.W.M.R.P.D.B. Athurupane
…………………………………………

Signature of Student

UOM Verified Signature

Date: 30/07/2022

Supervised by

Name of Supervisor

Dr. Saminda Premaratne
……………………………………….

Signature of Supervisor

UOM Verified Signature
…………………………..

Date: 30/07/2022

Acknowledgment

This dissertation would have been impossible without the tremendous support and motivation of those who assisted me throughout this journey. First and foremost, my heartfelt gratitude goes to my supervisor and Programme coordinator, Dr. Saminda Premaratne, who took responsibility for supervising me without hesitation. I am eternally grateful for him, for always being patient, flexible and empathetic. Also, I am thankful for his expertise, encouragement, moral support and guidance which he provided with generosity to facilitate my task.

Next, my heartfelt gratitude is extended to my friend Ms. S. N. Rambukkanage for the immeasurable support provided to me through these two years of period.

It is with great pleasure I thank the University of Moratuwa, Sri Lanka, for all its efforts and facilities that it has taken to contribute towards this postgraduate Programme. Specially, I would like to thank my colleagues for their support to complete this research in many ways.

Finally, all my thanks and respect go to my family and friends for all the unceasing love and support they provided me.

Abstract

The public tends to express their thoughts about particular goods and or services through popular social media networks such as Twitter and Facebook, while the firms also use social media to communicate with their consumers. As a result, this beneficial information can be used to make marketing and business decisions. However, due to the vast, noisy, and dynamic nature of these information, capturing the true public opinion has become a key challenge. Sentiment analysis is one of the methods that is employed to extract positive or negative attitudes from this social media information and thus, it has drawn the attention of the scholars during last decade. Different scholars have used a variety of techniques and methods to capture accurate results. Data mining is one of the recent approaches adopted by them to obtain better results from sentiment data. Moreover, some scholars have extended their studies to gain more insights from different topics. Such studies conducted to predict future results, analyze trends, detect anomalies etc. It can be beneficial for massive industries like footwear to understand their market through these approaches and streamline their product and service catalogs to meet the needs of their customers. This research aims to analyze previous studies conducted on this area, identify their contribution, challenges and limitations, and build a new comprehensive demand prediction model for footwear industry using data mining techniques.

Key words : *Sentiment Analysis, Data Mining, Demand prediction, Footwear Industry, Polarity detection, Lexicon, CrowdTangle, Twitter API, Postman, RapidMiner*

Table of Contents

Page

Declaration ..... **i**
Acknowledgment ..... **ii**
Abstract ..... **iii**
Table of Contents ..... **iv**
List of Figures ..... **vii**
List of Tables ..... **vii**
1. Introduction **1**
1.1 Background of the study 1
1.2 Problem Statement 3
1.3 Aim and Objectives 4
1.3.1 Aim 4
1.3.2 Objectives 4
1.4 Proposed solution 4
1.5 Summary 5
2. Literature Review **6**
2.1 Introduction 6
2.2 Data mining in online social media 6
2.3 Social sentiment analysis 7
2.4 Demand analysis using social media sentiments 8
2.5 Summary 9
3. Technology adopted **10**
3.1 Introduction 10
3.2 Tools and software used 10
3.2.1 RapidMiner Studio 9.10 10
3.2.2 CrowdTangle 10
3.2.3 Upscene Advanced Data Generator 4 11
3.2.4 Twitter API 12
3.2.5 Postman API 12
3.2.6 Microsoft Excel 13
3.3 Data mining algorithms and other techniques used 13
3.3.1 Naïve Bayes classification 13
3.3.2 K-NN Algorithm 14
3.3.3 Random forest 14

3.3.4 Deep learning 15
3.3.5 Rule induction 15
3.3.6 Decision tree 15
3.3.7 SentiWordNet 16
3.4 Summary 16
4. Approach for mining social sentiments for demand analysis in Footwear industry 17
4.1 Introduction 17
4.2 Inputs to the model 17
4.3 Output from the model 17
4.4 Process in brief 17
4.5 Platform selection 18
4.6 Data collection 18
4.6.1 Collection of Facebook post reactions data 19
4.6.2 Collection of Twitter post data (Tweets) 20
4.7 Summary 23
5. Analysis and Design 24
5.1 Introduction 24
5.2 Analysis of the study 24
5.3 Design of the study 25
5.3.1 Attribute selection 25
5.3.2 Text mining 26
5.3.3 Sentiment Analysis 26
5.3.4 Prediction models 30
5.4 Final design of the study 30
5.5 Summary 31
6. Implementation of the model 32
6.1 Introduction 32
6.2 Data preprocessing and Text mining 32
6.2.1 Facebook post reaction data 32
6.2.2 Twitter post data (Tweets) 35
6.3 Sentiment analysis 35
6.3.1 Sentiment analysis in Facebook post reaction 35
6.3.2 Sentiment analysis in Tweets 35
6.4 Process of creating aggregated dataset 37

6.5 Sentiment prediction 38
6.6 Identifying trend patterns and predict demand... 40
6.7 Summary 40
7. Evaluation 41
7.1 Introduction 41
7.2 Evaluation of the Sentiment predictions 41
7.2.1 Decision Tree algorithm 41
7.2.2 Random Forest algorithm 42
7.2.3 Naïve Bayes algorithm... 42
7.2.4 Deep learning algorithm 42
7.2.5 Rule induction algorithm 43
7.3 Evaluation of identifying trend patterns and predict demand 45
7.4 Summary 46
8. Conclusion and Future work.. 47
8.1 Conclusion of the study.. 47
8.2 Limitations of the study.. 48
8.3 Future work 49
9. Reference.. 50
Appendix A... 52
Appendix B 53
Appendix C... 54

List of Figures

Page

Figure 3-1: CrowdTangle Dashboard 11
Figure 3-2: Upscene Advanced Data Generator 11
Figure 3-3: Postman dashboard 12
Figure 4-1: Raw data collected from CrowdTangle 22
Figure 4-2: Raw data collected from Twitter 22
Figure 5-1: Proposed design of the model 25
Figure 5-2: Actual design of the model 31
Figure 6-1: Process of extracting attribute values from FB data 33
Figure 6-2: RapidMiner Operators used to analyse sentiment in Tweets 36
Figure 6-3: Parameter setting in Extract Sentiment operator 36
Figure 6-4: Example of the Final Dataset 38
Figure 6-5: Process of Predicting Sentiments 39
Figure 6-6: Field configuration of Pivot table 40
Figure 7-1: Sentiment trend for Flip Flops 45
Figure 7-2: Sentiment trend for Sandals 45

List of Tables

Page

Table 4-1: Additional Attributes and values 19
Table 4-2: Filters used in CrowdTangle 20
Table 4-3: Query parameters used in Postman API 21
Table 5-1: Sentiment score of Facebook reactions (Source: [21]) 29
Table 6-1: Tasks performed by RapidMiner Operator in attribute value extraction ... 34
Table 6-2: Tasks performed by RapidMiner operators in sentiment prediction 39
Table 7-1: Accuracy percentages of Decision Tree algorithm 41
Table 7-2: Accuracy percentages of Random Forest algorithm 42
Table 7-3: Accuracy percentages of Naive Bayes algorithm 42
Table 7-4: Accuracy percentages of Deep Learning algorithm 42
Table 7-5: Accuracy percentages of Rule Induction algorithm 43
Table 7-6: Suitable ML algorithms for Predict Sentiments 43

Chapter 1

1. Introduction

1.1 Background of the study

Millions of individuals across the world today use the internet to share their opinions or feelings on products and brands, whether it's via a Facebook reaction, a Twitter thread, or a Facebook post/comment. Reactions or comments on social media sites such as Twitter, Instagram, LinkedIn, Facebook, and others are common ways for people to express their feelings. Such feeling or attitude people express on social media about a given business, service, or product is known as social media sentiment, or simply "social sentiment." Companies are keen on capturing these online conversations in order to learn more about their customers and users, as well as the customers of their market competitors.

In order to obtain insights from these conversations, companies need to analyze them carefully. Practically, companies cannot manually search for reviews and comments, analyze them, and categorize them as good or bad due to the vast amount of data available. To cater this issue, we can use sentiment analysis and data mining techniques, algorithms to handle enormous amounts of data and obtain consumer insights in order to understand more about consumers' sentiment towards the brand or product. In addition to analysis, predictions are also made using these data mining approaches.

The footwear industry is a multibillion-dollar business worldwide. The footwear sector, which is a subset of the clothes and apparel industry, consists of shoes, sporting shoes, sneakers, athletic shoes, designer shoes, and as well as other related products. Footwear is frequently made of leather, fabric, and various man-made materials. Footwear can be considered a basic necessity item, standing with food, shelter, and clothes in terms of importance and nowadays most people have a footwear collection to meet a variety of criteria and expectations. There are many different styles, colors, and sizes of formal, casual, and sporting footwear. Like apparel, footwear, have evolved from simple items to true prestige symbols for consumers. Through footwear manufactures' different products and ranges, as well as their advertising efforts, footwear companies and stores play a significant part in catering to this market.

[1] states that the leading footwear producers are India, China, Turkey, Vietnam, Italy, Mexico, Indonesia, Spain, Brazil, and Thailand, whereas the top footwear consumers and importers are the United States, France, the United Kingdom, Japan, Italy and Germany. Formerly, the ignored group of footwear producers from Korea and Taiwan dominated the global business arena. Despite the COVID-19 issue, the worldwide market for footwear is predicted to expand at a compound annual growth rate (CAGR) of 2.3 percent to reach US$440 billion by 2026 from an estimated US$384.2 billion in 2020. By the end of the analysis period, sales of casual footwear are anticipated to increase at a CAGR of 2.6 percent to reach US$213.3 billion. The Athletic Footwear section's growth is lowered to an updated 1.9 percent CAGR over the next seven years after a thorough analysis of the pandemic's commercial effects and the ensuing economic crisis. This market sector currently holds a 37.6 percent segment of the global footwear market. According to [2], Nike Inc., Reebok, Puma, Adidas AG, Under Armour Inc., Geox, Skechers USA, ECCO Sko AS, Timberland, Deichmann SE, BATA, Jack Wolfskin, The Aldo Group, Columbia Sportswear, Polartec, and Asics Corp. are the leading players/brands in the worldwide footwear market.

In this massive industry, it is vital for footwear manufactures and retailers to know what their current and potential customers think about their products. Regardless of whether or not a firm has a social media presence, its customers and users almost certainly do. Individuals express their opinions about businesses, shopping, technology, and other topics on social media; many people seek recommendations and trade ideas on Facebook. Furthermore, a large number of users use social media to assess and review businesses and services. When the customers expressing their social sentiments, they are not always positive; many of them are likely to be negative or critical, and their words can be highly damaging to organizations in some situations.

Collecting and analyzing these huge unstructured conversations, ratings and reactions, in other words, social sentiments can be used to reveal important patterns or trends of products or brands, and they are very important for footwear manufactures and retailers to identify how their products and brands performed in the market over time. These historical data can be used to predict the future demand for certain products or brands. Identifying these patterns will help to improve customer service and satisfaction and, also it helps to monitor product or brand reputation, and analyzing consumer experience.

Thus, in this research, special attention should be given for data preprocessing since extracting sentiments from social media post engagements is not an easy process. Several methods have been used which were presented by few scholars. This study focuses on gathering Facebook post reactions and Twitter posts published regarding footwear categories. After identifying the sentiments of each post, proper data mining techniques will be used to predict the demand of each footwear category over time.

1.2 Problem Statement

Understanding clients' opinions about a product or service is the greatest approach for a firm to determine whether or not they like it. When it comes to demand prediction for a certain product or brand, these opinions are vital for an organization to analyze the trend for their product portfolio. As a result, social sentiment analysis of a product or brand is one of the most common methods that businesses employ to understand their clients' opinions nowadays. Traditional statistical techniques will take too long and be too expensive to analyze these social media sentiments since they are vast, noisy, and dynamic. Therefore, polarity must be extracted from these sentiments using data mining approaches.

Though numerous studies have been conducted in the area of social sentiment analysis, most of the studies have been used only one social network site or text related data for sentiment analysis. However, after social media sites such as Facebook introduced the new emoticons to interact with posts instead of one emoticon (Like button), new aspects of research were opened for social sentiment analysis. Recently, few scholars have conducted such research based on these Facebook reactions. However, forecasting models developed based on sentiment analysis using multiple social media platforms and multiple data types (reactions and comments) are limited, and a comprehensive demand prediction model based on social sentiments is yet to be developed. Even though those limited researches covered some trend pattern identification and forescasting, building a model for predicting sentiments and predicting demand for huge industries like footwear is also limited in the field of sentiment analysis. Therefore, such a model is critical for massive industries like footwear (or any other industry) to understand their market and streamline their product and service catalogs to meet the needs of their customers.

1.3 Aim and Objectives

1.3.1 Aim

The aim of this research is to design, implement and evaluate suitable models using data mining techniques to analyze trends and demands in the footwear industry through social sentiment prediction.

1.3.2 Objectives

01. To identify the most suitable attributes for trend analysis and demand prediction in Footwear industry by analyzing the nature of data.
02. To evaluate existing approaches for sentiment analysis and use best approaches to extract sentiment from collected data.
03. To construct a model to predict sentiment patterns using suitable data mining techniques.
04. To extend the model for trend analysis and demand forecasting.

1.4 Proposed solution

This study proposes finding a solution to the above problem by evaluating the current practices which were being used by other researchers for sentiment analysis and constructing a model to predict sentiment patterns using suitable data mining techniques. The model has been extended for trend analysis and demand forecasting using the predicted sentiments. As the first step, this study has gathered two different types of post data (Reactions and Text - from 2017 to 2021) from two different social media platforms (Facebook and Twitter) related to different footwear compositions. Then, the sentiment has been extracted from the collected data using appropriate sentiment extraction methods. After that, these data sets were aggregated to create unique compsitions of footwear categories and different types of data sets were created and used for the analysis. Different machine learning algorythms have been tested to select appropriate techniques for sentiment prediction for each of the data sets. Finally, using the sentiment predictions, a model was further extended for analyzing trend patterns and demand predictions for each footwear composition.

1.5 Summary

This chapter describes the research topic in general. Also, this discusses the reasons for carrying out this research in the current context and explains the importance of the identified problem and the need of a solution. Furthermore, solid aims and objectives were described in this chapter to construct a strong research area in the selected field.

Chapter 2

2. Literature Review

2.1 Introduction

This chapter describes about the different approaches for social media sentiment analysis and how other researchers have used these approaches for demand prediction in various domains. Especially this chapter emphasizes on how different social media platforms enable the users to show their feelings on a given topic and the different data mining techniques used to extract those feelings. Finally, this chapter explains the various data mining algorithms used by previous scholars to predict demand or forecast on different topics using the extracted sentiments of social media users.

2.2 Data mining in online social media

Organizations must clearly exploit their information resources to get a full understanding of markets, consumers, goods, employees, competitors, suppliers, and more in order to succeed in a globally-integrated and more consumer-empowered economy. Organizations will gain value by efficiently analyzing and managing the fast expanding latest and existing large amount of data (Big data) [3]. Big data is defined as a collection of diverse, enormous, and complicated data from various business and technology areas that cannot be effectively addressed by traditional technologies, infrastructure, or skills [4]. Therefore, according to [5] we need to utilize data mining to extract useful information from this enormous and complicated data collection that isn't totally visible and isn't always easy to obtain. In other words, data mining is the act of uncovering meaningful patterns and correlations in high volume of data, utilizing advanced data analysis tools and techniques such as artificial intelligence, machine learning, and statistics [6]. The growth of online social media has resulted in a flood of data from social networks. Data mining techniques can be used by researchers and professionals to analyze these vast volumes of complex, constantly changing social media data in order to find useful information. [5].

2.3 Social sentiment analysis

The capacity to tap into public sentiment on social media data is becoming a more essential tool for consumer segmentation, market analysis, and stock price forecasting in strategic market planning and positioning. Sentiment Analysis is most commonly used method to examine people's feelings, views, assessments, attitudes, evaluations, and emotions about companies' goods, services and their qualities as expressed online via text, video, and other communication methods. These messages might be classified as good, neutral, or negative [7]. We can call this as 'polarity detection'[8].

Researchers have conducted many studies previously to review different sentiment analysis techniques to obtain useful information from social media data. One research on this area, was discussing the overview and potential applications of big data sentiment analysis techniques [7]. Here, the authors discuss different sentiment analysis approaches and their suitability for the big data context. This research examines the advanced sentiment analysis methods, including sentiment classification algorithms through machine learning (ML) techniques such as Naive Bayes (NB), Support Vector Machine (SVM) and Maximum Entropy (ME). They state that these ML techniques can produce outstanding results in text categorization. In [8], addition to Naïve Bayes classification and Maximum Entropy classification, they have used another method called 'SentiWordNet', a lexical database that assigns scores to words and therefore finds an emotion score for a whole phrase. Some researchers have used semantic dictionary based approaches such as WordNet [9]. [10] used classification techniques. Likewise, there are many other techniques used by many scholars for opinion mining and sentiment analysis. Some of the other techniques are Bayesian networks (BN), K-nearest neighbor (KNN), Artificial neural network (ANN) and Decision tree (DT) [11].

These studies mainly focus on sentiment extraction from various text sources and languages through different social media networks and web pages. It can be a Facebook comment/post, Twitter comment/post, online product review etc. Even if the information published on social media is believed to be genuine public input because people offered their thoughts without being asked [12], these information are large, noisy and dynamic in nature [5]. As a result, the effectiveness of a better sentiment analysis technique depends greatly on the depth of analysis, the number of documents, the variety of the targeted domains, the genre, and the language of the texts. [13].

Even though the scholars define sentiment analysis as the sentiment identification and public opinion analysis that is considered as a research area in the field of text mining [11], a new approach added after Facebook releasing six emoji (like, sad, love, angry, haha and wow) options (reactions) in February 2016 as an extension of old “Like” button and allowed people to express their feelings on a post in more diverse ways [14]. Many scholars have recently conducted more research purely or partially based on this novel approach for sentiment analysis. One research completed in 2020, studied users' reactions on Facebook reflect the results of the election held in State of Mexico in 2017. This study has collected 4,128 Facebook posts and reactions to tackle the research problem [15]. Authors have developed a new method to measure positive and negative sentiments on Facebook reactions and, combined it with ‘Oh and Kumar’ sentiment index [16]. Another study released in 2018, used an emoji-based sentiment analysis and classifier to learn about individuals' emotional responses to marijuana-related Facebook posts. The High Times Magazine Facebook page was used to collect around 15,000 posts and 14 million user reactions for this study [17]. A study in 2019 [18], shows that it is possible to detect the polarity of a subject by Facebook reactions as it is an expression of human feeling in virtual environment.

2.4 Demand analysis using social media sentiments

During the literature assessment, I discovered that, despite the fact that various approaches for sentiment analysis have been proposed, only a few academics have undertaken studies on forecasting using social sentiments. [9] has built a predictive model for automobile sales based on social media sentiment analysis, combined with time-series analysis. This research mainly focused on text sentiments, and they have used dictionary-based approach, WordNet dictionary, to capture sentiments through IBM COBOL system (a system that can gather and examine a variety of user-generated social media content from message boards, blogs, microblogs, news forums and social media sites). Since this study conducted on 2014, any kind of reactions/emoticons have not been considered for the analysis. Another study [4] was undertaken to formulate a methodology to improve demand forecasting in supply chains using social network site data: Facebook and Twitter. The proposed approach uses trend, sentiment, and word analysis findings from social media big data to forecast product demand by combining an enhanced "Bass emotion model" with predictive modeling on historical sales data. A case study in the apparel industry retail supply chain has used to validate the proposed forecasting methodology presented on this paper. Predicting the popularity of TV shows using sentiment analysis (based on text evaluations) using Twitter is another study conducted related to this area [19].

2.5 Summary

Based on these recent studies, this study may conclude that these researchers have worked hard to construct forecasting models based on social sentiments and opinions. However, some research only looked at a few social media platforms, while others only looked at only text data. Therefore, a comprehensive demand prediction model using different type of social sentiment in social media yet to be developed for different industries.

Chapter 3

3. Technology adopted

3.1 Introduction

This chapter discusses different techniques and technologies adopted to complete the study. Based on the research topic, various techniques and technologies have been used in different stages such as the data gathering, data preprocessing, data analyzing and in outcome generation.

3.2 Tools and software used

3.2.1 RapidMiner Studio 9.10

RapidMiner is a powerful data mining tool that can manage all aspects of data mining, model implementation, and model operations. This study employed RapidMiner Studio, which offers a wide library of 1500+ algorithms and functions. Mainly RapidMiner studio was used for three major implementations in this study. First, for text mining and then for the sentiment analysis of mined text data and lastly for demand prediction. Other than that, the RapidMiner studio was used for data preprocessing and cleaning. RapidMiner Studio consists of easy access, drag and drop operations and made this study’s process efficient.

3.2.2 CrowdTangle

CrowdTangle is a Facebook public analytics tool that allows researchers, publishers, fact-checkers, journalists, and others to track, analyze, and publish on what's going on in the world of social media. In this study, Facebook post reactions data have been collected through the CrowdTangle platform. Using its easy-to-use dashboard, different queries have been executed and the relevant post reaction data has been extracted.

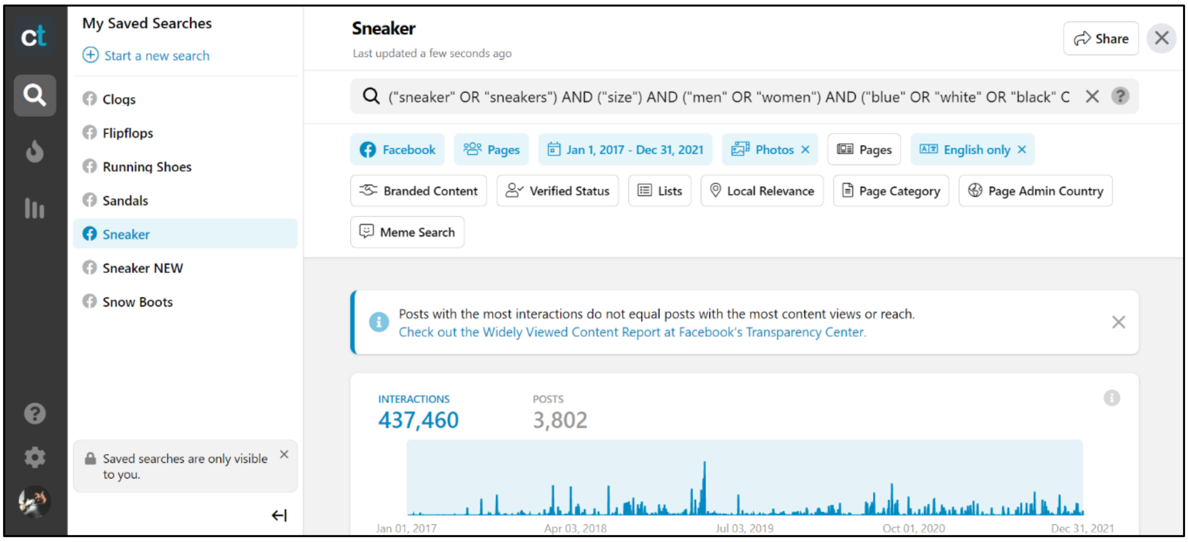


Figure 3-1: CrowdTangle Dashboard

3.2.3 Upscene Advanced Data Generator 4

Upscene Advanced Data Generator 4 is a sophisticated tool for creating test data for applications in many popular databases. It comes with a library of real-like data and many number of features that make the test data as realistic as possible. Since the Facebook post reactions data collected through the CrowdTangle platform were unbalanced at the time of analyzing, this tool has been used to generate a balanced data set to continue the research process. Using this tool, a sample data set of posts with reactions for each post has been created within less amount of time.

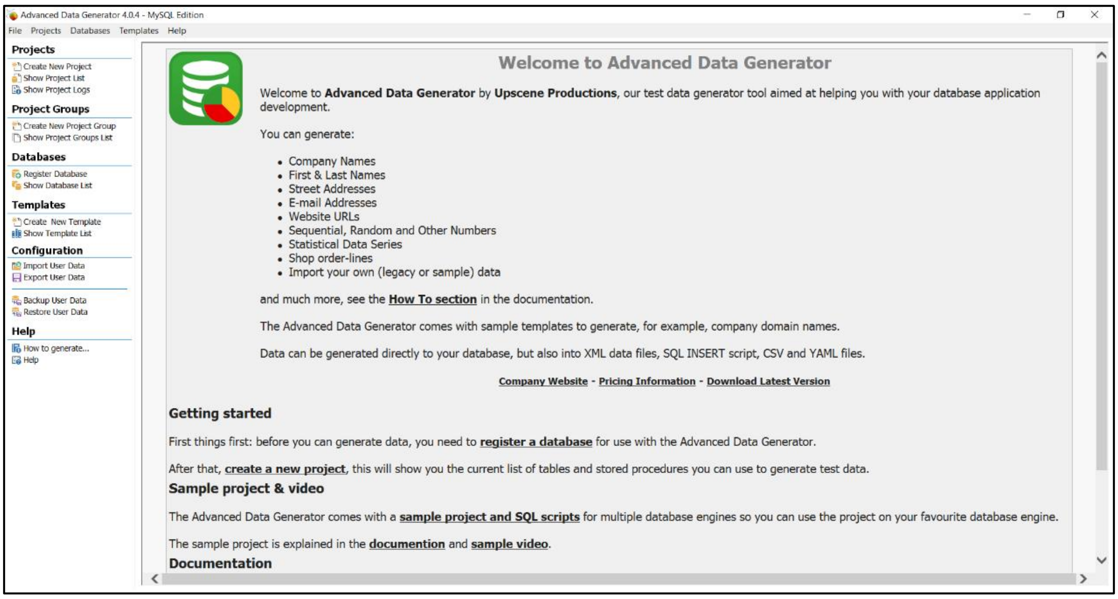


Figure 3-2: Upscene Advanced Data Generator

3.2.4 Twitter API

The Twitter API is a collection of programmatic interfaces for understanding and building Twitter conversations. We can use this API to discover and obtain, interact with, or create a range of resources, including the twitter users, tweets, messages, lists, trends, places and media. Twitter API has been used in this study to extract historical data of twitter, in other words, public twitter posts published by different users in related footwear categories. Then these collected tweets have been preprocessed and mined for sentiment extraction. To use the Twitter API, a developer platform needs to be created by mentioning the requirement of use of Twitter data. Therefore, an 'Academic research access' account has been created and study was granted access for Twitter historical data through the created account. Since this study was needed Twitter post data from 2017 to 2021, an 'Academic research access' account creation was must.

3.2.5 Postman API

Postman API enables the creation and utilization of various APIs. Each stage of the API lifecycle is made easier and more collaborative with Postman, enabling for speedier and better API design. Users can easily design, test, distribute, document APIs using simple and sophisticated HTTP/s queries using Postman. This platform has been used to gather Twitter post data through Twitter API. Different queries have been executed and the relevant Twitter post data has been extracted using this platform.

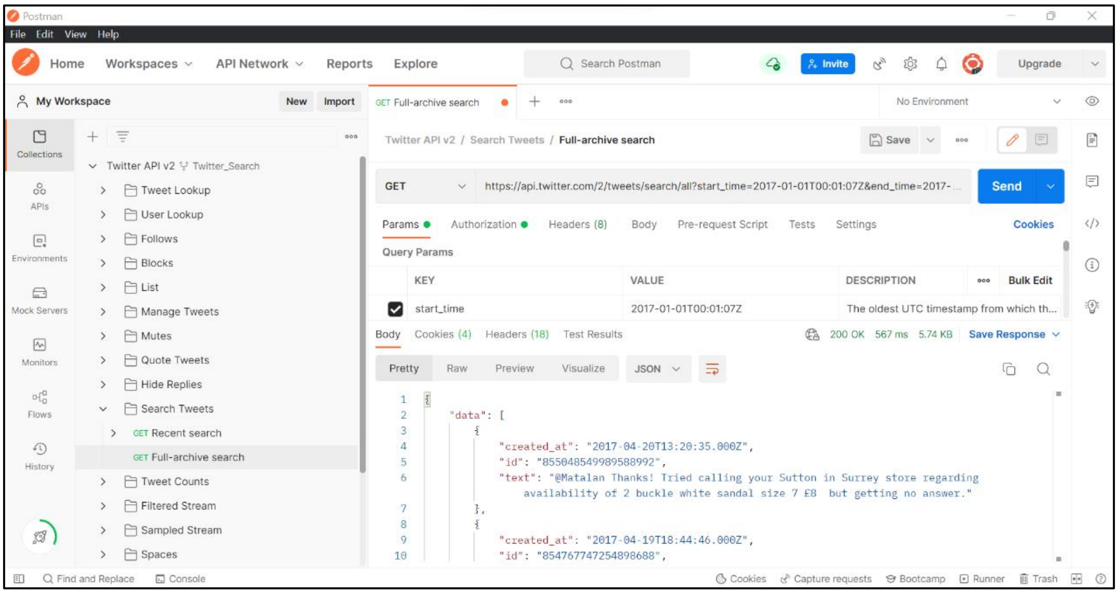


Figure 3-3: Postman dashboard

3.2.6 Microsoft Excel

Microsoft Excel is a spreadsheet application created by Microsoft which can be installed to many operating systems such Windows, iOS, Android etc. It includes computinng and calculation tools, pivot charts, pivot tables, graphing tools, and Visual Basics(VB) for Applications, a macro programming language to manage and process different type of data. Microsoft Excel is another application that was used in this study to preprocess the collected social media post data. Different in-built functions have been used for this purpose and duplicates removing, data aggregations and different calculations performed to smooth out the data set.

3.3 Data mining algorithms and other techniques used

Researchers and practitioners are using data mining techniques to evaluate vast amounts of complicated and fast paced social media data. Data mining is the process of discovering new and useful patterns in data. Knowledge discovery is another term for data mining [5].

3.3.1 Naïve Bayes classification

Naïve Bayes is a probabilistic classifier. It is constructed on probability models with robust individuality expectations. Often, the independence molds have no consequence on realism. As a consequence, they are considered as naïve.

A Naive Bayes model has following dimensions:

- Name of the input field
- Input field value or input field value range.
- Value of the target field

This means that a Naïve Bayesian model archives how frequently a target field value seems together with an input field value. Naive Bayes are extremely scalable, with parameters relative to the number of variables in a learning problem. Maximum-likelihood training can be done in linear time by assessing a closed-form appearance, as opposite to the costly iterative approximation used for many other types of classifiers.

3.3.2 K-NN Algorithm

A fundamental machine learning technique that depends on the supervised learning approach is K-Nearest Neighbor. The K-NN algorithm uses similarity between the new case/data and existing cases to choose which group to place the new case in. The K-NN method provides all information that is available and uses similarity to categorize new data facts. This means that when unique data is received, it may be quickly classified using the K- NN algorithm into a well-matched group. Although the K-NN technique can be used for both classification and regression, it is more frequently employed for classification errors. As a non-parametric algorithm, K-NN does not assume anything about the underlying data. The KNN method does nothing more than supply the dataset during the training phase, and when fresh data is received, it simply classifies it into a category that is highly similar to the fresh data.

3.3.3 Random forest

A supervised machine learning algorithm called random forest is frequently used to fix errors in classification and regression. Its theories are based on decision trees built from a variety of cases, and for classification and regression, it uses the consensus vote and the average. The Random Forest Algorithm's ability to handle data sets with both continuous and categorical variables, as in regression and classification, is one of its most important aspects. It produces better categorization results. As the names imply, the random forest is made up of numerous independent decision trees that collaborate. The class with the most votes becomes the forecast of our model. Each tree in the random forest creates a class prediction.

3.3.4 Deep learning

Deep learning algorithms are prominent in key characteristics and are capable of handling many different operations for both structured and unstructured data. On the other hand, since deep learning algorithms require access to enormous volumes of data in order to function well, they can ignore some obligations that may entail compound issues. In its dataset-driven algorithms, ImageNet, a popular deep learning tool for image processing, has exposure to 14 million photos. For deep learning tools that use images as their dataset, it is a full tool that has identified a new level of presentation.

Deep learning algorithms are very advanced algorithms that process each layer of the neural network to learn more about the previous image. The joint layers use this information to create comprehensive graphics by connecting it to previous data. The layers are extremely sensitive for detecting low-level image properties like edges and pixels. For example, other deep learned layers might be trained to recognize specific objects like cats, plants, gears, and so on, while the center layer could be trained to recognize specific sections of an object in a snapshot.

3.3.5 Rule induction

The procedure of concluding if-then rules from a data set is recognized as rule induction in data mining. These figurative decision rules clarify an ordinary association among the data set's attributes and class labels. Intuitive rule induction reinforces many real-world involvements. For example, we could announce that "if it is 8.30 a.m. on a monday, highway traffic will be heavy" and "if it is 8.30 a.m. on a Sunday, traffic will be light." These strategies are not continuously appropriate. During the holiday season, traffic at 8 a.m. on weekdays may be light. However, these rules grip true in overall and are resultant from realistic practice based on our day-to-day explanations.

3.3.6 Decision tree

The decision tree algorithm is categorized as supervised learning. They can resolve regression and classification difficulties. The decision tree resolves the problematic issues by utilizing a tree illustration in which each leaf node resembles to a class label and attributes are denoted on the tree's interior node. By means of a decision tree, we can embody any Boolean function on discrete attributes. Decision Tree discards the Disjunctive Normal Form, which is another name for the entirety of the Product form. Recognizing the property for the root node in each stage is the main challenge in Decision Tree. The term for this is attribute selection. A decision tree is a branch of a probabilities tree that enables you to select a precise course of action.

3.3.7 SentiWordNet

SentiWordNet is a judgement mining lexical source. SentiWordNet allocates three sentimentality notches to each Synset of WordNet: positivity, negativity, and objectivity. Sentiment classification is the usage of involuntary approaches for forecasting the orientation of subjective content on text documents, with applications fluctuating from recommender and marketing systems to client intellect and information repossession. SentiWordNet is a WordNet-based view lexicon in which each period is related with mathematical scores representing positive and negative sentiment data.

3.4 Summary

In summary, this chapter discussed about different techniques and technologies adopted to complete this study. Based on the research topic, various techniques and technologies have been utilized in different stages such as the data gathering, data preprocessing, data analyzing and in outcome generation.

Chapter 4

4. Approach for mining social sentiments for demand analysis in Footwear industry

4.1 Introduction

This chapter presents the approach for mining social sentiments for demand analysis in footwear industry. Inputs, outputs, process of the model, platform selection and data collection process will be discussed in this chapter. This chapter also discusses how the technologies stated in previous chapter are used to complete the actual work.

4.2 Inputs to the model

One of the main activities of this study is to extract sentiments form social media data regarding footwear. There are two types of data considered for the model,

01. Sentiment extracted from the Facebook post reactions - Emoticon based.
02. Sentiment extracted from the Tweets (Twitter posts) - Text based.

Finally, these two data sets were arranged to one format and aggregated as one data set to train the proposed model. Considerable amount of time has been taken to get these data in to the required format.

4.3 Output from the model

Year wise predicted sentiment of a given shoe category (with other features like size, color, gender) was the outcome of this model. One of the objectives of this study is to use these data for demand prediction of the given shoe categories.

4.4 Process in brief

In the process, posts from two platforms – Facebook and Twitter, was collected first by using relevant search keys and extracted the sentiment of each post. Since the Facebook data were unbalanced, Data Generator has been used to create a balanced data set. Then, using the generated data, sentiment has been identified for each post data row. Tweets' sentiments are extracted and calculated using text mining techniques and "SentiWordNet" libraries. All the collected sentiment data then aggregated and create one aggregated data set. This data set then checked with different Machine Learning techniques mentioned in previous chapter and find out the most accurate predictive model by training and testing the data set.

4.5 Platform selection

Since this study was taken to analyze demand of Footwear categories in the industry using social media sentiments, Facebook and Twitter platforms were selected. Facebook is the most widespread and well-known social network site in the world, with 316 million users as of November 2009 [6] and the number has increased to billion users by now. With introduction of Facebook reactions (Like, Heart, Sad, Angry, Wow, Haha), Facebook has enabled the users to show their opinion on a post by just one click. Twitter was first introduced in July 2006 as a social networking and microblogging program. Microblogging is a type of blogging that allows for more frequent and smaller updates than blogging. Twitter allows users to send public messages (referred to as "tweets") of up to 140 characters in length. That restriction was increased to 280 characters in November of 2017. Twitter is a good source to track how people's feelings evolve over time by looking at a tweet was posted. Twitter is having 1.3 billion accounts, 330 million active monthly users, and 83 percent of the world's leaders are on the platform [20]. As a result, Twitter is a good place to gather text data on a topic like footwear.

4.6 Data collection

Data collection process is carried out under two categories. They are Facebook post reactions and Twitter posts. Under these two categories, trending footwear categories have been used as keywords for search. The following footwear categories are considered in order to narrow down the scope of this study.

01. Sneakers

02. Flip Flops

03. Sandals

04. Clogs

05. Running Shoes

Following attributes of footwear also considered when querying the data from each platform.

Color	Size	Gender
Black, Blue, Grey, Pink, Red, White, Green, Brown, Navy	6, 6.5, 7, 7.5, 8, 8.5, 9, 9.5, 10, 10.5, 11, 11.5, 12	Male, Female

Table 4-1: Additional Attributes and values

First the Facebook post reactions data have been collected through CrowdTangle platform and Twitter posts have been collected through Twitter API using Postman API platform. Due to the privacy policies of each platform, publicly available data were only collected. Historical data of 5 years have been collected (from 2017 to 2021). Two distinct data collection processes are elaborated as follows.

4.6.1 Collection of Facebook post reactions data

According to previous studies, scholars have used Facebook reactions (e.g.- Like, Sad, Wow, Love, Angry and Haha) as a new approach in extracting sentiments from social media posts [15],[17],[18]. In this study, reactions for post data for each footwear category has been gathered using the CrowdTangle platform. CrowdTangle is a Facebook public insights tool that allows publishers, journalists, researchers, fact-checkers, and others to track, analyze, and report on what's going on in the social media world [20]. A total count of each reaction given by the users for a post were gathered. These posts are posted by retailers or manufacturers of Footwear products in their social media pages which is publicly available.

While extracting data from CrowdTangle, the following filters have been used to minimize the capture of irrelevant data.

Social media Platform	Facebook
Account type	Public Pages
Post type	Photos
Time range	1st January 2017 to 31st December 2021
Language	English Only
Page category	Shopping retail, Retail company

Table 4-2: Filters used in CrowdTangle

Search query used to collect posts

("sneaker" OR "sneakers") AND ("size") AND ("men" OR "women") AND ("blue" OR "white" OR "black" OR "red" OR "pink" OR "navy" OR "grey" OR "green" OR "brown") AND ("6" OR "6.5" OR "7" OR "7.5" OR "8"OR "8.5" OR "9"OR "9.5" OR "10"OR "10.5" OR "11"OR "11.5" OR "12")

During data gathering, the following number of posts have been gathered with their specific interaction data.

01. Sneakers – 2,958 Posts

02. Flip Flops – 957 Posts

03. Sandals – 1,136 Posts

04. Clogs - 736 Posts

05. Running Shoes – 620 Posts

Total posts collected – 6,407

4.6.2 Collection of Twitter post data (Tweets)

Many scholars have used different methods to extract Twitter data. One of the ethical and common methods is using Twitter API. Since the Twitter API supports Postman API platform, all the queries and filters executed through Postman API platform. In this platform, tweets, in other words, text data have been collected through the data collection process. These tweets are published by twitter users publicly.

Following query parameters have been used to filter out the tweets.

Parameter	**Value**	**Description**
start_time	2017-01-01T00:01:00Z	The oldest UTC timestamp from which the Tweets will be provided.
end_time	2021-12-31T23:59:59Z	The newest, most recent UTC timestamp to which the Tweets will be provided.
tweet.fields	created_at, text	Fields that need to retrieve.
query	*lang:en (is:reply OR is:quote) (sneaker OR "flip flop" OR sandal OR clogs OR "running shoe") (-SneakerPhetish -mademenFit -info_reseller -SupremeCommerce) (size) (black OR white OR gray OR red OR brown OR navy OR blue OR green OR pink)*	Search term that used to filter out the relevant data. Here, only English tweets were collected. Also some repeating users have been removed. Dash("-") sign represent NOT operator.

Table 4-3: Query parameters used in Postman API

During data gathering, the following number of matching tweets have been collected.

01. Sneakers – 836 Tweets

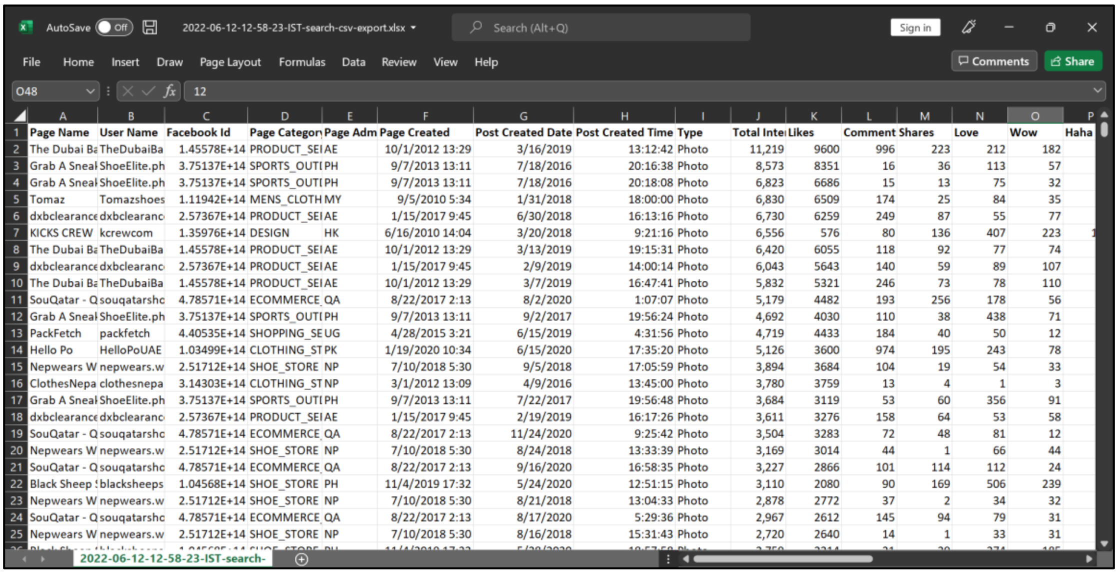
02. Flip Flops – 51 Tweets

03. Sandals – 456 Tweets

04. Clogs - 151 Tweets

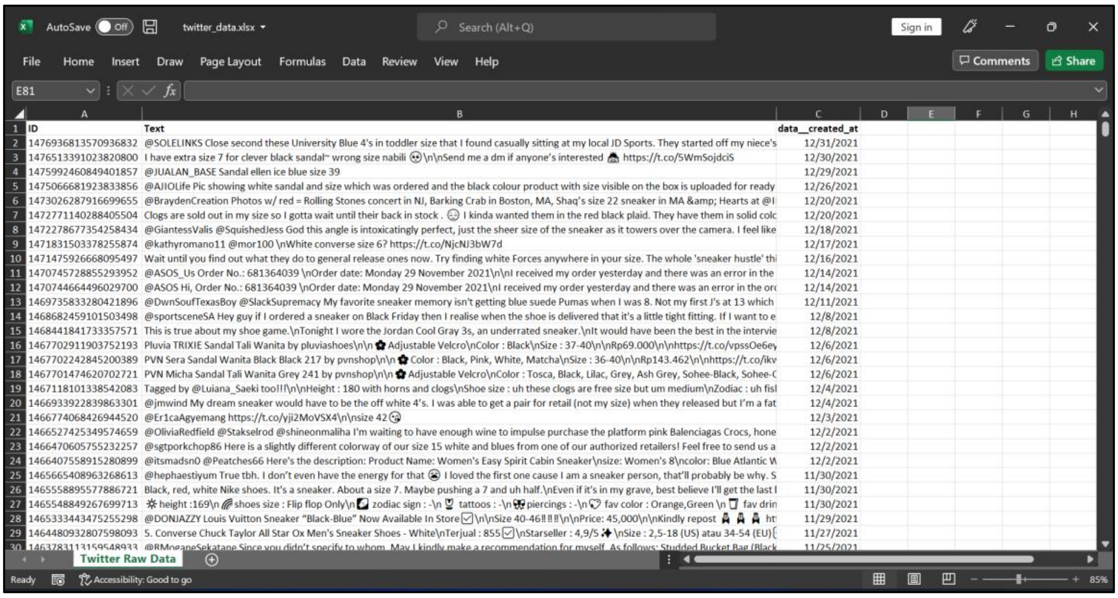
05. Running Shoes – 60 Tweets

Total posts collected – 1,555



Page Name	User Name	Facebook Id	Page Categor	Page Adm	Page Created	Post Created Date	Post Created Time	Type	Total Intei	Likes	Comment	Shares	Love	Wow	Haha
The Dubai Ba	TheDubaiBa	1.45578E+14	PRODUCT_SEI	AE	10/1/2012 13:29	3/16/2019	13:12:42	Photo	11,219	9600	996	223	212	182	
Grab A Sneal	ShoeElite.ph	3.75137E+14	SPORTS_OUTI	PH	9/7/2013 13:11	7/18/2016	20:16:38	Photo	8,573	8351	16	36	113	57	
Grab A Sneal	ShoeElite.ph	3.75137E+14	SPORTS_OUTI	PH	9/7/2013 13:11	7/18/2016	20:18:08	Photo	6,823	6686	15	13	75	32	
Tomaz	Tomazshoes	1.11942E+14	MENS_CLOTH	MY	9/5/2010 5:34	1/31/2018	18:00:00	Photo	6,830	6509	174	25	84	35	
dxbclearance	dxbclearanc	2.57367E+14	PRODUCT_SEI	AE	1/15/2017 9:45	6/30/2018	16:13:16	Photo	6,730	6259	249	87	55	77	
KICKS CREW	kcrewcom	1.35976E+14	DESIGN	HK	6/16/2010 14:04	3/20/2018	9:21:16	Photo	6,556	576	80	136	407	223	
The Dubai Ba	TheDubaiBa	1.45578E+14	PRODUCT_SEI	AE	10/1/2012 13:29	3/13/2019	19:15:31	Photo	6,420	6055	118	92	77	74	
dxbclearance	dxbclearanc	2.57367E+14	PRODUCT_SEI	AE	1/15/2017 9:45	2/9/2019	14:00:14	Photo	6,043	5643	140	59	89	107	
The Dubai Ba	TheDubaiBa	1.45578E+14	PRODUCT_SEI	AE	10/1/2012 13:29	3/7/2019	16:47:41	Photo	5,832	5321	246	73	78	110	
SouQatar - Q	souqatarsho	4.78571E+14	ECOMMERCE_	QA	8/22/2017 2:13	8/2/2020	1:07:07	Photo	5,179	4482	193	256	178	56	
Grab A Sneal	ShoeElite.ph	3.75137E+14	SPORTS_OUTI	PH	9/7/2013 13:11	9/2/2017	19:56:24	Photo	4,692	4030	110	38	438	71	
PackFetch	packfetch	4.40535E+14	SHOPPING_SE	UG	4/28/2015 3:21	6/15/2019	4:31:56	Photo	4,719	4433	184	40	50	12	
Hello Po	HelloPoUAE	1.03499E+14	CLOTHING_ST	PK	1/19/2020 10:34	6/15/2020	17:35:20	Photo	5,126	3600	974	195	243	78	
Nepwears W	nepwears.w	2.51712E+14	SHOE_STORE	NP	7/10/2018 5:30	9/5/2018	17:05:59	Photo	3,894	3684	104	19	54	33	
ClothesNepa	clothesnepa	3.14303E+14	CLOTHING_ST	NP	3/1/2012 13:09	4/9/2016	13:45:00	Photo	3,780	3759	13	4	1	3	
Grab A Sneal	ShoeElite.ph	3.75137E+14	SPORTS_OUTI	PH	9/7/2013 13:11	7/22/2017	19:56:48	Photo	3,684	3119	53	60	356	91	
dxbclearance	dxbclearanc	2.57367E+14	PRODUCT_SEI	AE	1/15/2017 9:45	2/19/2019	16:17:26	Photo	3,611	3276	158	64	53	58	
SouQatar - Q	souqatarsho	4.78571E+14	ECOMMERCE_	QA	8/22/2017 2:13	11/24/2020	9:25:42	Photo	3,504	3283	72	48	81	12	
Nepwears W	nepwears.w	2.51712E+14	SHOE_STORE	NP	7/10/2018 5:30	8/24/2018	13:33:39	Photo	3,169	3014	44	1	66	44	
SouQatar - Q	souqatarsho	4.78571E+14	ECOMMERCE_	QA	8/22/2017 2:13	9/16/2020	16:58:35	Photo	3,227	2866	101	114	112	24	
Black Sheep :	blacksheeps	1.04568E+14	SHOE_STORE	PH	11/4/2019 17:32	5/24/2020	12:51:15	Photo	3,110	2080	90	169	506	239	
Nepwears W	nepwears.w	2.51712E+14	SHOE_STORE	NP	7/10/2018 5:30	8/21/2018	13:04:33	Photo	2,878	2772	37	2	34	32	
SouQatar - Q	souqatarsho	4.78571E+14	ECOMMERCE_	QA	8/22/2017 2:13	8/17/2020	5:29:36	Photo	2,967	2612	145	94	79	31	
Nepwears W	nepwears.w	2.51712E+14	SHOE_STORE	NP	7/10/2018 5:30	8/16/2018	15:31:43	Photo	2,720	2640	14	1	33	31	

Figure 4-1: Raw data collected from CrowdTangle



ID	Text	data__created_at
1476936813570936832	@SOLELINKS Close second these University Blue 4's in toddler size that I found casually sitting at my local JD Sports. They started off my niece's	12/31/2021
1476513391023820800	I have extra size 7 for clever black sandal~ wrong size nabili \n\nSend me a dm if anyone's interested https://t.co/5WmSojdciS	12/30/2021
1475992460849401857	@JUALAN_BASE Sandal ellen ice blue size 39	12/29/2021
1475066681923833856	@AJIOLife Pic showing white sandal and size which was ordered and the black colour product with size visible on the box is uploaded for ready	12/26/2021
1473026287916699655	@BraydenCreation Photos w/ red = Rolling Stones concert in NJ, Barking Crab in Boston, MA, Shaq's size 22 sneaker in MA & Hearts at @I	12/20/2021
1472771140288405504	Clogs are sold out in my size so I gotta wait until their back in stock . I kinda wanted them in the red black plaid. They have them in solid colc	12/20/2021
1472278677354258434	@GiantessValis @SquishedJess God this angle is intoxicatingly perfect, just the sheer size of the sneaker as it towers over the camera. I feel like	12/18/2021
1471831503378255874	@kathyromano11 @mor100 \nWhite converse size 6? https://t.co/NjcNJ3bW7d	12/17/2021
1471475926668095497	Wait until you find out what they do to general release ones now. Try finding white Forces anywhere in your size. The whole 'sneaker hustle' thi	12/16/2021
1470745728855293952	@ASOS_Us Order No.: 681364039 \nOrder date: Monday 29 November 2021\n\nI received my order yesterday and there was an error in the	12/14/2021
1470744664496029700	@ASOS Hi, Order No.: 681364039 \nOrder date: Monday 29 November 2021\nI received my order yesterday and there was an error in the orc	12/14/2021
1469735833280421896	@DwnSoufTexasBoy @SlackSupremacy My favorite sneaker memory isn't getting blue suede Pumas when I was 8. Not my first J's at 13 which	12/11/2021
1468682459101503498	@sportsceneSA Hey guy if I ordered a sneaker on Black Friday then I realise when the shoe is delivered that it's a little tight fitting. If I want to e	12/8/2021
1468441841733357571	This is true about my shoe game.\nTonight I wore the Jordan Cool Gray 3s, an underrated sneaker.\nIt would have been the best in the intervie	12/8/2021
1467702911903752193	Pluvia TRIXIE Sandal Tali Wanita by pluviashoes\n\n ✿ Adjustable Velcro\nColor : Black\nSize : 37-40\n\nRp69.000\n\nhttps://t.co/vpssOe6ey	12/6/2021
1467702242845200389	PVN Sera Sandal Wanita Black Black 217 by pvnshop\n\n ✿ Color : Black, Pink, White, Matcha\nSize : 36-40\n\nRp143.462\n\nhttps://t.co/ikv	12/6/2021
1467701474620702721	PVN Micha Sandal Tali Wanita Grey 241 by pvnshop\n\n ✿ Adjustable Velcro\nColor : Tosca, Black, Lilac, Grey, Ash Grey, Sohee-Black, Sohee-C	12/6/2021
1467118101338542083	Tagged by @Luiana_Saeki too!!!\n\nHeight : 180 with horns and clogs\nShoe size : uh these clogs are free size but um medium\nZodiac : uh fisl	12/4/2021
1466933922839863301	@jmwind My dream sneaker would have to be the off white 4's. I was able to get a pair for retail (not my size) when they released but I'm a fat	12/4/2021
1466774068426944520	@Er1caAgyemang https://t.co/yji2MoVSX4\n\nsize 42	12/3/2021
1466527425349574659	@OliviaRedfield @Stakselrod @shineonmaliha I'm waiting to have enough wine to impulse purchase the platform pink Balenciagas Crocs, hone	12/2/2021
1466470605755232257	@sgtporkchop86 Here is a slightly different colorway of our size 15 white and blues from one of our authorized retailers! Feel free to send us a	12/2/2021
1466407558915280899	@itsmadsn0 @Peatches66 Here's the description: Product Name: Women's Easy Spirit Cabin Sneaker\nsize: Women's 8\ncolor: Blue Atlantic W	12/2/2021
1465665408963268613	@hephaestiyum True tbh. I don't even have the energy for that I loved the first one cause I am a sneaker person, that'll probably be why. S	11/30/2021
1465558895577886721	Black, red, white Nike shoes. It's a sneaker. About a size 7. Maybe pushing a 7 and uh half.\nEven if it's in my grave, best believe I'll get the last I	11/30/2021
1465548849267699713	height :169\n shoes size : Flip flop Only\n zodiac sign : -\n tattoos : -\n piercings : -\n fav color : Orange,Green \n fav drin	11/30/2021
1465333443475255298	@DONJAZZY Louis Vuitton Sneaker "Black-Blue" Now Available In Store \n\nSize 40-46‼‼‼\n\nPrice: 45,000\n\nKindly repost ht	11/29/2021
1464480932807598093	5. Converse Chuck Taylor All Star Ox Men's Sneaker Shoes - White\nTerjual : 855 \nStarseller : 4,9/5 \nSize : 2,5-18 (US) atau 34-54 (EU)	11/27/2021

Figure 4-2: Raw data collected from Twitter

Even though the same type of data needed to be collected for the study, due to the limitation on collecting Facebook related data because of their privacy policies, the following aspects have been considered.

- How users (people) show their sentiments towards a Footwear product via Facebook reactions (emoticons).

- How people actually think about a Footwear product. This is what they actually comment about a product in their own words. These data taken from the Tweets.

4.7 Summary

In summary, this chapter discussed about how the technologies and techniques mentioned in the previous chapter used in early stages of the research process. Also, what type of inputs selected, the brief process of the research, outcome of the study, platforms used and how the data were collected from each platform. Next chapter provides the analysis and design of the research and the top-level model that was used in the study to make use of these collected data.

Chapter 5

5. Analysis and Design

5.1 Introduction

This chapter explains how the problems and solutions of the study have been analyzed with relevant literature and with the new approaches. Then, based on the comprehensive analysis, top-level design of the proposed solution will be discussed.

5.2 Analysis of the study

According to the previous literature (which was discussed in the Chapter 2) many scholars have conducted research on sentiment analysis on different domains. Some of them further extended their sentiment analysis research to build forecasting models or prediction models. Most of them have selected only one social media platform and also it was mainly based on text sentiments. After Facebook was extending their Like button with new emojis, people used it as way of showing their emotions on a particular post. This post represents a topic or domain. Based on that fact, some of the researchers used this as a way of measuring the sentiments towards particular topic. After analyzing this previous literature, this study focused on building a comprehensive sentiment analysis model using both text data and post reaction data. Also, wanted to find answers for "why this broad model can not be used for better demand prediction of a certain domain or industry".

Next, this study analyzed about capabilities of current technologies and techniques on data extraction, data processing, sentiment extraction and predictions. With the help of applications such as RapidMiner, CrowdTangle, Postman API and Excel, the process of this study got feasible from many aspects. Also, premade supervised learning algorithms like Support vector machine, Naïve Bayes, Decision tree etc. ease up the prediction process.

5.3 Design of the study

High-level design of the proposed model is vital before implementation. By designing the model of the study using different elements is crucial for better understanding and good planning.

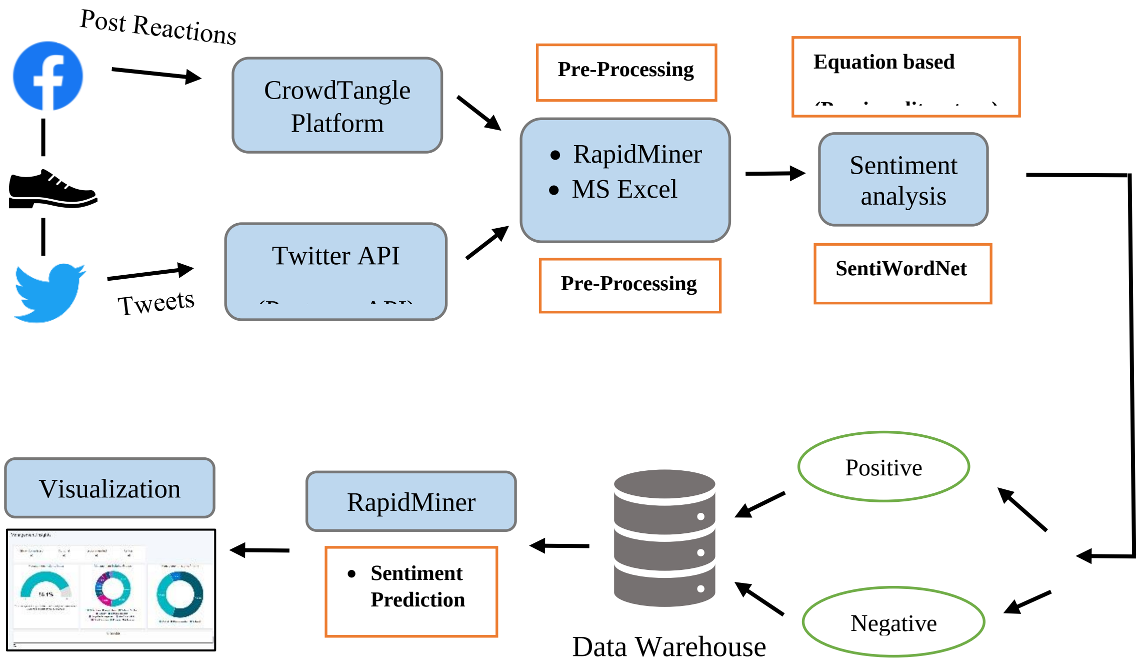


Figure 5-1: Proposed design of the model

5.3.1 Attribute selection

When selecting the attributes of the data set, some important factors had to be considered. Such as,

- Both attributes should be available and able to extract from both platforms.
- Should be frequently used by majority of the social media users.
- The value should be a common word or number and constant on both platforms.
- Should related to the domain(Footwear industry).

By considering the above factors, the following attributes were selected.

1. Shoe category
2. Color of the shoe
3. Size of the shoe
4. Gender
5. Post created date/ year
6. Sentiment (Class label)

Some of the values of these attributes were not directly captured in the data collection. Text mining techniques have been used to extract them.

5.3.2 Text mining

Text mining is the process of converting unstructured text into a structured format in order to find new insights and patterns. Main usage of text mining in this study was to extract attribute values from the two distinct data sets – Facebook post reactions and Twitter posts. The values are the shoe categories, shoe sizes, colors and whether it is a women's or men's shoe. These data were embedded either in Facebook post description or Twitter post. The steps that carried out to extract these attribute values will be discussed on the next chapter – Implementation of the study.

5.3.3 Sentiment Analysis

In this study, sentiment analysis is conducted in two different ways for the two data sets gathered which were post reaction data collected from Facebook posts and post data from Twitter. Since these two data sets produce different data values (reaction data and texts), different sentiment analysis techniques needed to get accurate results. Following two techniques have been used.

01. Sentiment analysis on Facebook post reactions

Step 01 – First step was to find the Total Positive Sentiment and Total Negative Sentiment using reactions(or interactions). In order to perform this, this study has used the formulas given in article [15]. They are,

$$\textit{Total Positive Sentiment} = \textit{Wow} + \textit{Likes} + 2 * \textit{Love}$$

$$\textit{Total Negative Sentiment} = \textit{Haha} + \textit{Sad} + 2 * \textit{Angry}$$

According to [15], Facebook users will "like" a post if it connects with them or makes them feel good. On the other side, Facebook users use the "Love" reaction when they are very enthusiastic about something. Because of this, a "Love" response in this study is scored as being twice as passionate as a "Like" response. The "Angry" and "Sad" reactions are calculated in the same way since a post that provokes negative emotions produces these two emotions; an "Angry" expression signifies a higher negative feeling in the polarization than a "Sad" expression. As a result, an "Angry" expression is regarded to be twice as strong as a "Sad" expression. "Wow" and "Haha" were considered to be neutral reactions. Since "Wow" responses heavily dominated in posts with more "Like" and "Love" expressions, they concluded that they are indicative of a good emotion. The vast majority of the gathered expressions connect with "Sad" and "Angry," suggesting that "Haha" expressions have a negative tone (sarcasm).

Even though this article [15] uses “Wow” and “Haha” expressions, these expressions did not consider in the sentiment calculation considering the nature of the domain (Footwear) of this study. As an example, if a shoe manufacturer creates a funny post to market their product, people will hit more “Haha” reactions. But it is not the actual sentiment that people express on the shoe. So, the final modified formula that was used to calculate the total positive and negative sentiment as follows.

$$\textit{Total Positive Sentiment} = \textit{Likes} + 2 * \textit{Love}$$

$$\textit{Total Negative Sentiment} = \textit{Sad} + 2 * \textit{Angry}$$

Step 02 – The next step is to find the Sentiment Index of each post using the Total Negative Sentiment and Total Positive Sentiment values obtained in Step 01. For this, 'Oh and Kumar' Sentiment Index logarithm has been used to find the overall sentiment on Facebook posts [16]. The correlation between positive and negative posts are depicted by the sentiment index logarithm. The result is a value that represents how negative or positive the emotions expressed by social media followers are [15]. The logarithm is as follows,

$$\text{Sentiment Index} = ln\left[\frac{1 + Total^{\,Positive}}{1 + Total^{Negative}}\right]$$

This study measured the sentiment and the impact value of Facebook users toward each footwear related posts in this way.

Step 03 – Finally, "sentiment value" class label (or attribute) has generated using Sentiment Index value generated in step 02. The "positive", "negative" and "neutral" class labels have generated if Sentiment Index value is greater than zero, less than zero and equal to zero respectively. Additionally, posts which have "neutral" class labels were removed from the data sets since the Positive and Negative sentiments only be considered for the trend analysis and forecasting in this study.

After reviewing the previous literature, another method of calculating sentiment of Facebook reactions was found. It was a database which is consisted with sentiment score for each emoji [21]. Using their database in http://kt.ijs.si/data/Emoji_sentiment_ranking/ website, sentiment score of each Facebook reaction was gathered and the scores are as follows.

Emoji	Sentiment score
Like 👍	0.521
Heart ❤️	0.746
Haha 😄	0.421
Wow 😮	0.269
Sad 😢	0.007
Angry 😡	-0.299

Table 5-1: Sentiment score of Facebook reactions (Source: [21])

Using the following equation (mentioned in one previous study, [14]), the overall sentiment score was calculated.

$$\frac{1}{n}\sum_{i=1}^{n} log((number\ of\ emoji_i) + 1) \times sentiment\ of\ emoji_i$$

$$n = the\ total\ number\ of\ distinct\ emojis$$

Final result was overall sentiment scores were got positive values when negative values should be expected in some cases. One major issue was that the "Sad" emoji has a positive sentiment score even though it has a very less value. Also this issue elaborated by the authors in study [14]. Therefore, this method has not taken into this study by considering the limitations.

02. Sentiment analysis on retweets on Twitter posts

There are many text mining or classification techniques used by many scholars on Twitter post data to extract human thoughts and perceptions. A comprehensive study conducted on "classification techniques for opinion mining and sentiment analysis" [11] shows a high level view on available techniques for sentiment analysis (See Appendix A).

In this study, a lexicon-based approach has been used to extract polarity from the Twitter posts. "SentiWordNet" is a lexical database that assigns scores to words and so determines a sentiment score for a sentence as a whole. This study has used this database to calculate the sentiment score of the collected Tweets. Also, sentiment score calculation of Tweets has been processed on RapidMiner studio using its “Sentiment Analysis” operator.

5.3.4 Prediction models

Main focus of this study is to check whether the proposed model can predict the demand of selected shoe categories by analyzing the past behavior of their social sentiments. For this purpose, different data mining algorithms mentioned in Chapter 3 have been used and tested their accuracy. Those algorithms were selected based on the previous literature and using categorization provided in http://mod.rapidminer.com/#app web site (After selecting the nature of our inputs and outputs, this application suggests the relevant machine learning algorithms that can be used for predictions). The most accurate algorithm has used to predict the future sentiments of the footwear categories. By analyzing this and based on the sentiment trend of the past 5 years, a basic prediction of the selected footwear categories were analyzed. The process of testing and the steps carried out the prediction model will be discussed in detail in the next chapter – Implementation.

5.4 Final design of the study

After identifying the sentiments of Facebook reaction posts which was gathered using CrowdTangle, the study found that the data set was unbalanced and cannot carry out the research process further. The problematic result was 90% of the posts had “Positive” sentiment and only 10% could identified as “Negative”. Therefore, to create a balanced data set, Data Generator (‘Upscene Advanced Data Generator 4’) was used in the study to mimic the actual data set gathered by CrowdTangle and carried out the research further.

One reason of receiving an unbalanced data set (More “Positive” sentiments and few “Negative” sentiments) found in this study was that the limitation of emojis in the Facebook reaction button and only one reaction can hit per post by a single user. As an example, when a person sees an expensive and attractive shoe in a post, and if that person thinks the shoe is not worth for the price but unique attraction is there, person will hit the ‘Love’ or ‘Like’ reaction rather than ‘Sad’ or ‘Angry’ reaction. The person does not have a way of expressing both of his/her feelings at once, which is the Positive sentiment (Attractive shoe) and the Negative sentiment (Not worth the price).

Also, this depends on the topic or the context of the post shared. The way the people react to a political related post is different than this context. When it comes to the text related replies such as comments or Tweets, people have more control and power of expressing their feelings by words or large variety of emojis. So, this limitation was not found when analyzing the Twitter data.

Therefore, the actual high-level model had to change as follows.

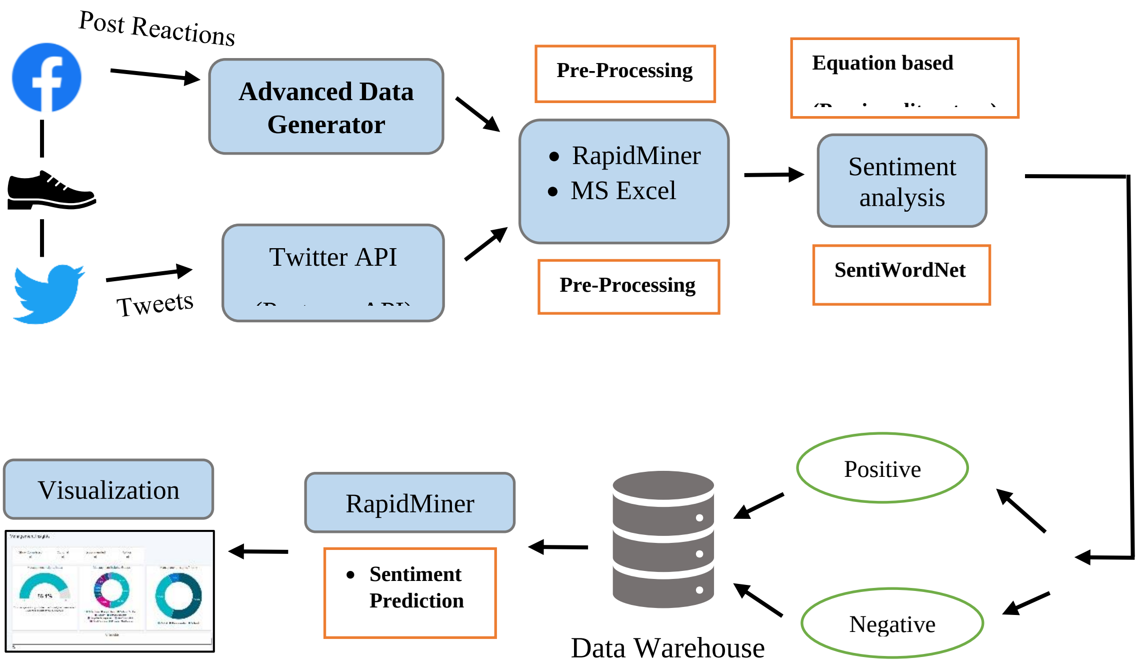


Figure 5-2: Actual design of the model

5.5 Summary

This chapter discussed the problems and solutions of the study that have been analyzed with relevant literature and the new approaches. Then based on the comprehensive analysis, top-level design of the proposed and actual solution has been discussed with limitations. Then in the implementation stage, the design will be executed and analyze the actual implementation.

Chapter 6

6. Implementation of the model

6.1 Introduction

This chapter discusses the implementation details of each and every process discusses in the previous chapter – Analysis and Design. Moreover, this chapter presents how this study used technologies and techniques discussed in Chapter 3 – Technology adopted for each major part of the design.

6.2 Data preprocessing and Text mining

Data preprocessing is important in data mining because the raw data obtained from the actual world are subjected to noise, inconsistency, and incompleteness. To provide consistent and high-quality results which align with the expected outcome, data preprocessing needs to be conducted, which includes stages such as data cleansing, integration, transformation, reduction, and discretization. This section also includes how text was mined to extract the values of attributes.

In this study, the data preprocessing and text mining has been done differently for the two data collections.

6.2.1 Facebook post reaction data

6.2.1.1 Data collected using the CrowdTangle platform

When extracting the data using the CrowdTangle platform, different attributes such as Page name, Page created date, Post created time, Page Admin, Followers, No. of Comments, No. of Shares and Overperforming score were extracted from post data as default. Since these attributes are irrelevant for the study, they were removed to increase the simplicity of the dataset. Furthermore, to calculate the sentiment score and the sentiment index, Total Positive sentiment, Total Negative sentiment, Sentiment Index and Sentiment Value attributes were added to the data set.

- Facebook ID
- Post created time
- Post description (Message)
- Total number of Interactions
- No. of Likes
- No. of Hearts
- No. of Sad
- No. of Angry
- Total Positive sentiment
- Total Negative sentiment
- Sentiment index
- Sentiment value

Next step was to extract attribute values (Shoe category, Sizes, Colors and Gender) from the Post description. For this, RapidMiner studio was used.

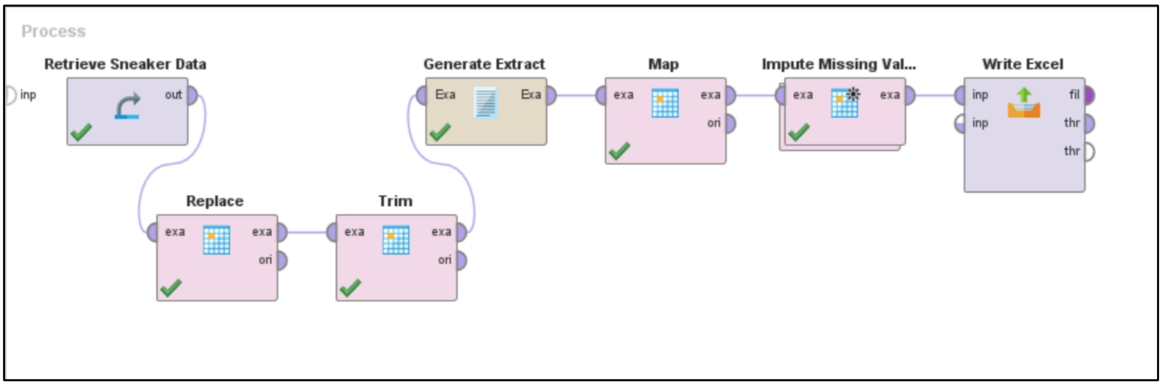


Figure 6-1: Process of extracting attribute values from FB data

Each operator of the RapidMiner studio performed specific task in extracting data. They are as follows.

Operator	Task performed
• Retrieve	Load the data into the process.
• Replace	Replaces parts of the values of selected nominal attributes matching a specified regular expression by a specified

<table>
<tr><td></td><td>replacement. This operator was used to remove URLs, @mentions in the text.

Regular Expression used–
(https?://[-a-zA-Z0-9+&@#/%?=~_|!:,.;]*[-a-zA-Z0-9+&@#/%=~_|]))|(RT\s*@[^:]:\s[A-Za-z]+)|((^|[^@\w])@(\w{1,15})\b)|(\\n)</td></tr>
<tr><td>• Trim</td><td>Used to remove leading and trailing spaces from the text.</td></tr>
<tr><td>• Generate Extract</td><td>This operator extracts values from the text.

Example of the Regular expression used –
(?i)blue|black|white|gray|red|brown|navy|green|pink</td></tr>
<tr><td>• Map</td><td>Used this operator to convert all the data values into one format. Ex- blue, BLUE, BLue → Blue.</td></tr>
<tr><td>• Impute missing values</td><td>Used this operator to fill missing values of the data set. K-NN algorithm was selected.</td></tr>
<tr><td>• Write Excel</td><td>This operator was used to write the processed data set to an excel sheet</td></tr>
</table>

Table 6-1: Tasks performed by RapidMiner Operator in attribute value extraction

The final attributes were taken for the study from these data as follows.

- Post Date – *From 1/1/2017 to 31/12/2021*
- Shoe Category – *Sneakers, Clogs, Sandals, Flip Flops, Running shoes*
- Gender – *Men, Women*
- Color - *Black, Blue, Grey, Pink, Red, White, Green, Brown, Navy*
- Size - *6, 6.5, 7, 7.5, 8, 8.5, 9, 9.5, 10, 10.5, 11, 11.5, 12*
- Sentiment – *Positive, Negative*

6.2.1.2 Data generated using Upscene Advanced Data Generator 4

As mentioned in the Chapter 5, sentiments of Facebook reaction posts gathered using CrowdTangle was unbalanced in terms of the sentiments (class label). Therefore, Upscene Advanced Data Generator 4 used to generate a data set with the selected attributes above. Parameters were configured in the software and a sample data set of 5,000 rows was created (See Appendix B) and this data set was used for this study.

6.2.2 Twitter post data (Tweets)

Using the Twitter API in Postman API platform, tweets were gathered according to the steps mentioned in Chapter 4, Data collection section. Selected attributes were Post_Created_Date and Twitter post. Attribute values were extracted from the text in Twitter posts by following the same process used in attribute value extraction of Facebook post description. Same operators and configurations in RapidMiner studio were used for this (See *Table 6-1: Tasks performed by RapidMiner Operator in attribute value extraction*). Some of the repeated rows have been removed manually.

6.3 Sentiment analysis

Sentiment analysis for both datasets were carried out separately since the nature of data sets and calculating sentiment value is different to each other.

6.3.1 Sentiment analysis in Facebook post reaction

The steps which were discussed in the Chapter 5 (Sub Section 5.3.3) used to calculate the Total Positive sentiment and Total Negative sentiment for Facebook posts considering the reactions of the posts. The modified formula was used. Using these values Sentiment index was calculated. To calculate this value, the equation discussed in Chapter 5 was used. Finally, using a simple IF function in MS Excel, sentiment value (Negative or Positive) was calculated.

6.3.2 Sentiment analysis in Tweets

Sentiment analysis of Tweets carried out in RapidMiner Studio. The “Extract Sentiment” operator was used for this and “SentiWordNet” dictionary was the selected sentiment score measuring approach. Addition to this, “Generate Attribute” operator was used to get the sentiment value of each post data. Finally, the “Write Excel” operator was used to save the final twitter data set in excel format.

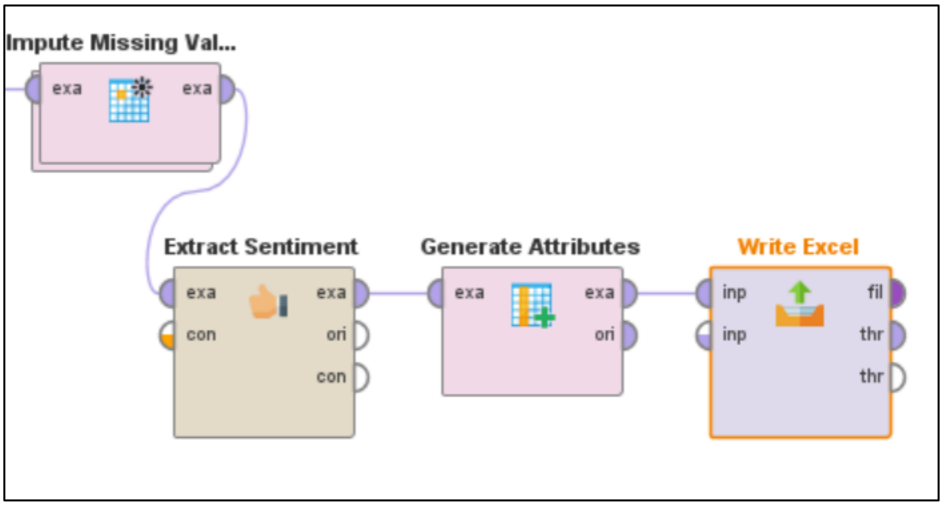


Figure 6-2: RapidMiner Operators used to analyse sentiment in Tweets

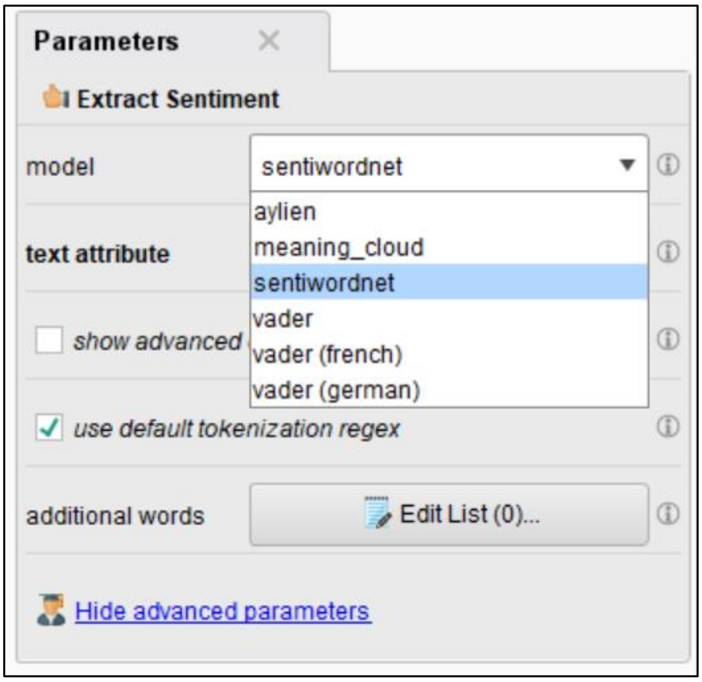


Figure 6-3: Parameter setting in Extract Sentiment operator

Finally, these two datasets were combined together as one dataset after finding the sentiment values. Finally, the dataset was consisted with 6,555 data rows.

From this combined dataset, a new dataset was created by aggregating the sentiment values of each shoe categories. Here, all the unique combinations of shoe categories were considered based on the year. Aggregated dataset creation was required since the previous dataset included repeating data rows and there was no meaning represented by the dataset.

As an example, aggregated sentiment value (Negative or Positive) of *Sneaker, Men, Grey, Size 10* for each year was calculated. Likewise, the same has done for other shoe compositions by considering all the attributes.

6.4 Process of creating aggregated dataset

Step 01 – Extract each unique composition of shoe categories from the main data set. Here, 229 unique compositions were identified. Year also extracted from the post created date.

Step 02 – Calculate the number of Total Positive and Negative sentiments in each composition (Ex- *Sneaker, Men, Grey, Size 10)*. Simple COUNTIFS function in excel was used to calculate the totals. This was helped to identify the total number of Positive and Negative sentiments in each year per one unique composition.

Step 03 – Calculate the difference between Positive and Negative (Positive – Negative) sentiment and identified the overall sentiment for those shoe compositions. Here, if a composition gets an equal value for both Positive and Negative values, that considered as a "Neutral" sentiment towards that shoe composition.

Step 04 – If a composition got zero values for both Positive and Negative, that record has removed from the dataset since they considered as noisy data in the data set. Both Positive and Negative totals get zero because no user has expressed any sentiment towards those posts/composition.

After these steps, the final data set was generated with 1,757 records. This is the final processed dataset that has been used for the predictions and analysis.

	A	B	C	D	E	F	G	H	I
1	Year	Shoe_Category	Gender	Color	Size	Total Positive	Total Negative	Difference	Final_Sentiment
2	2017	Clogs	Men	Black	6	1	2	-1	Negative
3	2017	Clogs	Women	Black	9	1	1	0	Neutral
4	2017	Clogs	Men	Blue	6.5	2	1	1	Positive
5	2017	Clogs	Men	Blue	8	2	1	1	Positive
6	2017	Clogs	Men	Blue	8.5	2	1	1	Positive
7	2017	Clogs	Men	Blue	9	1	2	-1	Negative
8	2017	Clogs	Women	Blue	8	1	1	0	Neutral
9	2017	Clogs	Women	Brown	7	1	1	0	Neutral
10	2017	Clogs	Women	Brown	9	1	1	0	Neutral
11	2017	Clogs	Women	Brown	10.5	1	2	-1	Negative
12	2017	Clogs	Men	Green	6.5	1	2	-1	Negative
13	2017	Clogs	Men	Green	11.5	1	1	0	Neutral
14	2017	Clogs	Women	Green	8	2	2	0	Neutral
15	2017	Clogs	Women	Green	9.5	1	1	0	Neutral
16	2017	Clogs	Men	Grey	7	1	2	-1	Negative
17	2017	Clogs	Men	Grey	10	1	1	0	Neutral
18	2017	Clogs	Men	Grey	11	1	1	0	Neutral
19	2017	Clogs	Men	Grey	12	1	1	0	Neutral

Figure 6-4: Example of the Final Dataset

6.5 Sentiment prediction

Using the final aggregated dataset, prediction of sentiments for footwear categories (or compositions) carried out after discovering the sentiments of the historical post data. RapidMiner studio was used this for process and the following test cases were employed.

- 2 training data sets and 2 test data sets were selected based on the year.
    1. Training data set – From 2017 to 2019 and predict sentiment for 2020
    2. Training data set – From 2017 to 2020 and predict sentiment for 2021
- 5 Machine learning algorithms were selected based on the past literature.
- 6 Datasets – One aggregated dataset and shoe category wise datasets.
- 2 types of class labels – One with the Neutral sentiment and other one after removing data rows which has Neutral sentiments.

Results of these tests and final evaluations are discussed in the next Chapter and only the implementation of these processes are described in this Chapter.

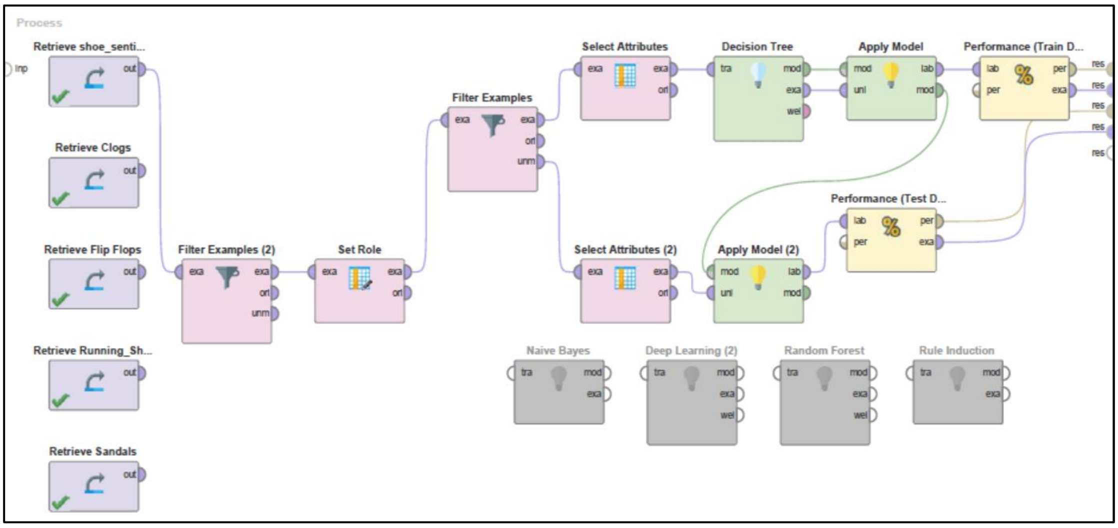


Figure 6-5: Process of Predicting Sentiments

Tasks performed by each operator as follows.

Operator	Task performed
• Retrieve	Load the data into the process.
• Filter Examples	Which Examples are keeping and which Examples are removing were chosen by this Operator. 1st Filter example was used to Filter data with Neutral sentiment and without it. 2nd one was used to select only relevant attributes for the model. Only Shoe Category, Gender, Color, Size and Sentiment.
• Set Role	This Operator is used to change the role of an attribute. This was used to set the role of Sentiment attribute as 'label'.
• Algorithm	Any suitable algorithm can be used. Here the Naïve Bayes, Random Forest, Decision Tree, Deep learning, Rule induction were selected and tested them separately by enabling them.
• Select attributes	This Operator selects out a set of attributes from a dataset and discards the rest.
• Apply model	Applies a model on a Data set.
• Performance	This operator provides a list of the categorization task performance criteria values. Accuracy was selected as the parameter.

Table 6-2: Tasks performed by RapidMiner operators in sentiment prediction

6.6 Identifying trend patterns and predict demand

Identifying the trend patterns and demand predictions of footwear have not been extensively covered using data mining algorithms in this research but have been done separately by using the aggregated final dataset. For this, the study has applied Microsoft excels' "Pivot table" and "Pivot Chart" tools to the aggregated dataset and created a dashboard of sentiment trends. The sum values of Positive and Negative sentiment counts have been taken and created the chart and the full chart of Trend.

The first step was to add a pivot table and configured the table fields as follows in the Figure 6-6: Field configuration of Pivot table

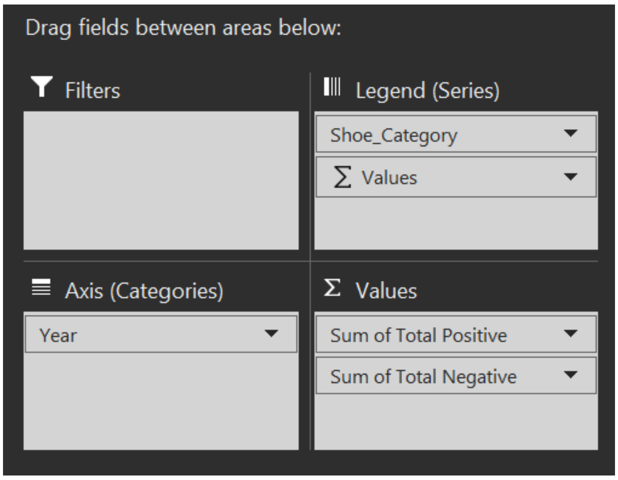


Figure 6-6: Field configuration of Pivot table

Then using the Pivot charts, a "Line chart with markers" was added. Then to filter out the chart, "Slicers' was added. A picture of a full dashboard can be seen in Appendix C. Result of this process is that we can manually come to conclusions on demand of Footwear categories by analyzing the chart. The collected, preprocessed and aggregated dataset was helped for this analysis.

6.7 Summary

In this chapter, how the design of the study has been implemented was discussed. Different activities performed at each level of the implementation stage was elaborated and data preprocessing, text mining tasks and prediction process described. Finally, how the dataset can be used for identifying trend patterns and predicting demand of footwear industry elaborated. Next chapter will evaluate the results of these processes.

Chapter 7

7. Evaluation

7.1 Introduction

The results gained from the implementation phase will be discussed and interpreted in here in order to make the conclusion of the research study in the final chapter. First the model of the sentiment prediction will be discussed and then trend patterns and how footwear demand predictions can be identified from this study will be discussed.

7.2 Evaluation of the Sentiment predictions

Accuracy level of selected five different algorithms have been discussed below with the different test cases mentioned in the previous chapter. All the algorithms were executed on aggregated dataset and each Footwear wise datasets to obtain a comprehensive view on the proposed model.

7.2.1 Decision Tree algorithm

Datasets	Train data from 2017 to 2019. Predict for 2020				Train data from 2017 to 2020. Predict for 2021			
	With 'Neutral' Sentiment		Without 'Neutral' Sentiment		With 'Neutral' Sentiment		Without 'Neutral' Sentiment	
	Train data	Test data	Train data	Test data	Train data	Test data	Train data	Test data
All	57.78%	40.66%	53.45%	53.45%	56.18%	42.31%	53.45%	54.24%
Clogs	72.48%	40%	75.56%	50%	71.43%	53.57%	72.31%	53.33%
Flip Flops	45.47%	41.05%	54.32%	54.74%	45.02%	39.89%	54.42%	51.70%
Running shoes	77%	50%	80.65%	71.43%	75.21%	52.38%	78.95%	47.06%
Sandals	78%	40%	78.26%	75%	69.33%	38.64%	78.57%	50.00%
Sneakers	66.23%	35.19%	80%	57.58%	64.39%	40.32%	80.53%	66.67%

Table 7-1: Accuracy percentages of Decision Tree algorithm

7.2.2 Random Forest algorithm

Datasets	Train data from 2017 to 2019. Predict for 2020				Train data from 2017 to 2020. Predict for 2021			
	With 'Neutral' Sentiment		Without 'Neutral' Sentiment		With 'Neutral' Sentiment		Without 'Neutral' Sentiment	
	Train data	Test data	Train data	Test data	Train data	Test data	Train data	Test data
All	78.11%	38.46%	84.23%	56.03%	73.06%	37.36%	80.23%	49.58%
Clogs	89.91%	42.22%	93.33%	60.00%	84.42%	39.29%	92.31%	40.00%
Flip Flops	68.56%	38.42%	80.04%	50.68%	64.41%	37.23%	75.79%	48.98%
Running shoes	91.95%	40.00%	96.77%	28.57%	88.89%	47.62%	92.11%	47.06%
Sandals	92.37%	37.78%	95.65%	66.67%	82.82%	43.18%	92.86%	55.56%
Sneakers	84.77%	42.59%	91.25%	60.61%	78.05%	38.71%	85.84%	58.97%

Table 7-2: Accuracy percentages of Random Forest algorithm

7.2.3 Naïve Bayes algorithm

Datasets	Train data from 2017 to 2019. Predict for 2020				Train data from 2017 to 2020. Predict for 2021			
	With 'Neutral' Sentiment		Without 'Neutral' Sentiment		With 'Neutral' Sentiment		Without 'Neutral' Sentiment	
	Train data	Test data	Train data	Test data	Train data	Test data	Train data	Test data
All	49.51%	49.18%	54.98%	55.60%	49.93%	46.70%	58.76%	56.78%
Clogs	66.97%	57.78%	68.89%	50%	64.94%	46.43%	69.23%	53.33%
Flip Flops	43.87%	43.16%	54.77%	55.41%	45.68%	43.09%	57.76%	55.10%
Running shoes	64.37%	73.33%	74.19%	42.86%	67.52%	59.52%	71.05%	47.06%
Sandals	61.86%	53.33%	60.87%	62.50%	61.35%	59.09%	67.14%	61.11%
Sneakers	57.62%	51.85%	75.00%	66.67%	57.56%	41.94%	73.45%	64.10%

Table 7-3: Accuracy percentages of Naive Bayes algorithm

7.2.4 Deep learning algorithm

Datasets	Train data from 2017 to 2019. Predict for 2020				Train data from 2017 to 2020. Predict for 2021			
	With 'Neutral' Sentiment		Without 'Neutral' Sentiment		With 'Neutral' Sentiment		Without 'Neutral' Sentiment	
	Train data	Test data	Train data	Test data	Train data	Test data	Train data	Test data
All	51.56%	47.80%	57.58%	52.59%	53.95%	46.98%	61.58%	57.20%
Clogs	58.72%	55.56%	53.33%	55.00%	64.29%	42.86%	58.46%	66.67%
Flip Flops	46.00%	45.79%	51.44%	47.30%	46.08%	42.55%	57.60%	47.62%
Running shoes	65.52%	43.33%	77.42%	28.57%	69.23%	59.52%	68.42%	41.18%
Sandals	61.02%	46.67%	63.04%	62.50%	61.96%	65.91%	68.57%	50.00%
Sneakers	41.06%	31.48%	75%	60.61%	51.71%	38.71%	74.34%	64.10%

Table 7-4: Accuracy percentages of Deep Learning algorithm

7.2.5 Rule induction algorithm

Datasets	Train data from 2017 to 2019. Predict for 2020				Train data from 2017 to 2020. Predict for 2021			
	With 'Neutral' Sentiment		Without 'Neutral' Sentiment		With 'Neutral' Sentiment		Without 'Neutral' Sentiment	
	Train data	Test data	Train data	Test data	Train data	Test data	Train data	Test data
All	52.53%	45.60%	66.77%	55.60%	53.52%	42.58%	68.47%	53.81%
Clogs	62.39%	55.56%	64.44%	65%	62.34%	42.86%	75.38%	60%
Flip Flops	55.95%	42.11%	65.19%	58.78%	56.57%	43.09%	68.45%	48.98%
Running shoes	64.37%	76.67%	58.06%	28.57%	67.52%	59.52%	57.89%	35.29%
Sandals	65.25%	44.44%	65.22%	75%	68.71%	43.18%	61.43%	50%
Sneakers	61.59%	38.89%	63.75%	69.70%	67.32%	37.10%	73.45%	58.97%

Table 7-5: Accuracy percentages of Rule Induction algorithm

After evaluating this comprehensive accuracy percentages, the following findings were obtained.

01. It was difficult to select one algorithm for all shoe categories to predict sentiments, cause the different higher accuracy percentages obtained by different shoe categories in different algorithms.

02. Therefore, based on these training data sets most suitable algorithm to predict sentiment for all shoe categories and for each shoe category were selected as below.

Shoe Category	Most suitable algorithm to predict future sentiments with available training dataset	
	With the "Neutral" class label value	Without the "Neutral" class label value
For All shoe categories	N/A*	Random forest
For Clogs	Naïve Bayes	Deep learning
For Flip Flops	N/A*	Rule induction
For Running shoes	Rule Induction	Decision Tree
For Sandals	Deep Learning	Decision Tree
For Sneakers	Naïve Bayes	Decision Tree

Table 7-6: Suitable ML algorithms for Predict Sentiments

*All the algorithms were not given a sufficient accuracy level.

03. It is notable that datasets which had “Neutral” sentiment value, were given lower accuracy levels in each algorithm than the datasets which were removed the “Neutral” sentiment value.

04. Also, a higher accuracy levels were given by the individual shoe category datasets when compared to aggregated dataset.

So, the final evaluation of this section is that using the algorithms mentioned in Table 7-6: Suitable ML algorithms for Predict Sentiments the sentiments of different Footwear categories (or compositions) can be predicted.

From the predicted sentiments, we can identify that whether there will be a demand for that specific shoe category in the future.

Example –

Assume that there is a Positive sentiment in 2023 for *Sneaker, Men, Grey, Size 10* footwear. This predicts and shows that analyzing the historical sentiment data of this specific shoe, people have a positive feeling to buy this product. Basically, the demand for the product.

7.3 Evaluation of identifying trend patterns and predict demand

As discussed in the previous chapter, to identify the trend patterns of footwear sentiment over years, the MS Excels' "Pivot Table" and "Pivot Chart" were used. Using these two tools, a complete, easy to understandable dashboard was created within minutes. Following are some examples of trend charts.

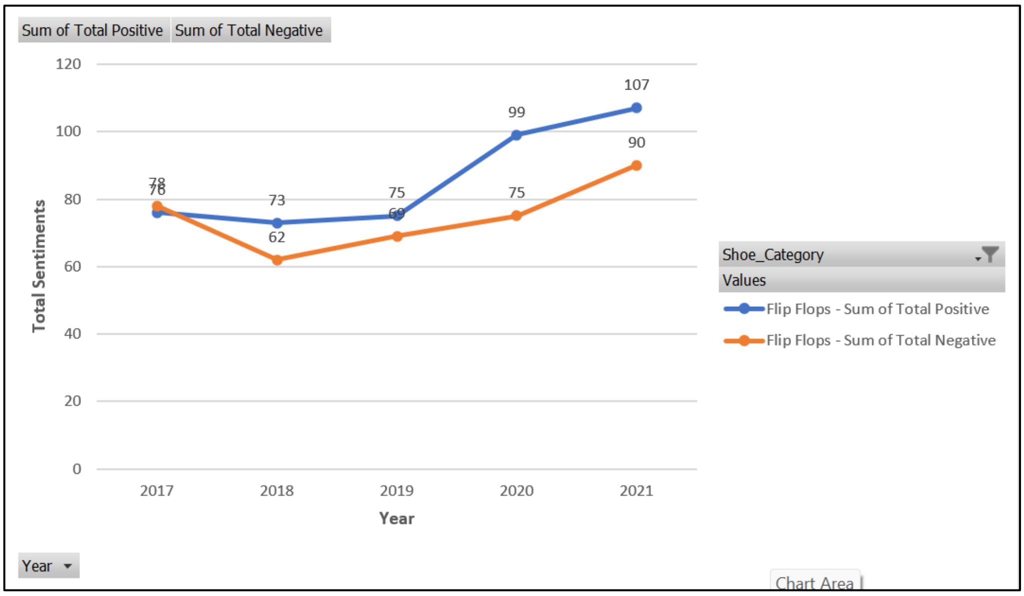


Figure 7-1: Sentiment trend for Flip Flops

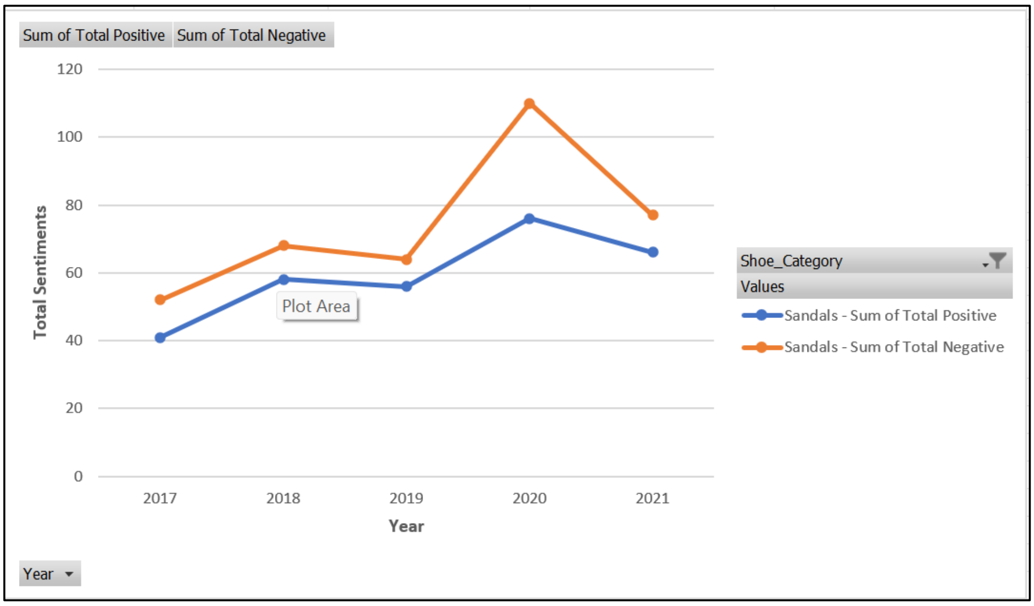


Figure 7-2: Sentiment trend for Sandals

By looking at these trend lines, a Footwear manufacturer or retailer can get a high-level idea of how people liked a certain product in previous years. Also, they can determine that, whether the product had a more Positive sentiment or more Negative sentiment and then can predict whether there will be Positive or Negative sentiment in the future. Likewise, the trend lines created using the dataset which was obtained after utilizing considerable number of efforts in this study, was able to help for footwear demand prediction.

7.4 Summary

This chapter described the results gained from the implementation phase and carefully evaluated the results gained to make proper decisions. In the first section, the model of the sentiment prediction has been discussed and then in next section, how Sentiment predictions and trend patterns helped to forecast demand of Footwear categories were explained.

Chapter 8

8. Conclusion and Future work

8.1 Conclusion of the study

When concludes, first, it is significant to cross check whether this study has met its aims and objectives defined in the Chapter 1. Objectives of this study are as follows.

01. To identify the most suitable attributes for trend analysis and demand prediction in Footwear industry by analyzing the nature of data.
02. To evaluate existing approaches for sentiment analysis and use best approaches to extract sentiment from collected data.
03. To construct a model to predict sentiment patterns using suitable data mining techniques.
04. To extend the model for trend analysis and demand forecasting.

The first objective was partially achieved. Due to the unstructured and vast nature of these social media data, extracting the attribute values were limited. As the result, limited number of attributes had to select such as Shoe category, Size, Color and Gender.

The next objective was completely achieved through referring past literature and the available techniques and technologies. The third objective was achieved, but due to the less amount of data used for the model at the final stage (due to excessive data preprocessing had to carry out to generate a reliable dataset), less accurate results were generated by the algorithms. Even though the last objective was not carried out as a direct process of the proposed model, with results generated by achieving the first three objectives, the last objective also achieved in this study.

Finally, this research is a novel approach for the social sentiment analysis since this study collected two distinct datasets from two social media platforms and was able to create a one common dataset for analysis. Also the analysis was further extended for demand and trend analysis.

8.2 Limitations of the study

The following limitations have been identified while carrying out this study.

01. Limited access on social media data.

The privacy worries of social media users have increased in recent years. Many users have been disturbed by data breach incidents and have been compelled to reevaluate their interactions with social networks and the safety of their personal data. Therefore, social networking sites now restrict the researchers and other people to access bulk data. Even though some sites allow to access the users' data, either they restrict the data can be gathered or need to obatain special permission to access such data. As an example, Facebook does not allow anyone to access individuals' private or public posts. However, accurate results can be obtained for a study like this by collecting more data if the researchers can access to such private and public posts of individuals.

02. Inconsistent nature of the raw text data, which is the various ways that individuals express their feelings on social media.

When it comes to the text based comments, people express their feelings in many different ways. They use different contexts of the languages, mix languages, slang words, different spellings, different emotions etc. As an example, the word "good" can be expressed as "gd", "goodie", "ohh good" or as a thumbs-up emoji. Therefore, according to this example, identifying the word "good" in individuals' comments is difficult because of the inconsistent nature. This problem occurred in several stages in this study. One was when identifying the attribute values. As an example, when extracting the footwear category, people have used the word "sneaker" in many different froms such as "jordans", "tennis shoes", "sneaks" etc. This limitation even has a negative impact on the accurate results generation on sentiment extracting as well.

03. The way people interact with footwear related posts are different.

As an example, when a user is surfing on his/her Facebook news feed, he/she would scroll down to the next post rather than negatively react on a post about a footwear product that he/she does not like.

8.3 Future work

01. This study was taken only four attributes for sentiment analysis. Taking more attributes can generate more accurate results.
02. Collect more historical and large amount of data.
03. Consider the emojis in text posts when analyzing the sentiment.
04. Expand the scope or create a generic model that can be adopted to any similar industry, not just the footwear industry.
05. Use data mining related forecasting models for better trend analysis and forecasting.

9. Reference

[1] R. and M. ltd, “Footwear - Global Market Trajectory & Analytics - Research and Markets.” https://www.researchandmarkets.com/reports/338775/footwear_global_market_trajectory_and_analytics (accessed Jul. 16, 2021).

[2] “Footwear Market Share, Size, Analysis, Forecast (2020-25).” https://www.mordorintelligence.com/industry-reports/footwear-market (accessed Jul. 16, 2021).

[3] D. Turner, M. Schroeck, and R. Shockley, “Analytics: The real-world use of big data in financial services,” p. 16.

[4] R. Iftikhar and M. S. Khan, “Social Media Big Data Analytics for Demand Forecasting: Development and Case Implementation of an Innovative Framework,” *J. Glob. Inf. Manag.*, vol. 28, no. 1, pp. 103–120, Jan. 2020, doi: 10.4018/JGIM.2020010106.

[5] G. Barbier and H. Liu, “Data Mining in Social Media,” in *Social Network Data Analytics*, C. C. Aggarwal, Ed. Boston, MA: Springer US, 2011, pp. 327–352. doi: 10.1007/978-1-4419-8462-3_12.

[6] A. S. Bozkır, S. Güzin Mazman, and E. Akçapınar Sezer, “Identification of User Patterns in Social Networks by Data Mining Techniques: Facebook Case,” in *Technological Convergence and Social Networks in Information Management*, vol. 96, S. Kurbanoğlu, U. Al, P. Lepon Erdoğan, Y. Tonta, and N. Uçak, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 145–153. doi: 10.1007/978-3-642-16032-5_13.

[7] N. M. Sharef, H. M. Zin, and S. Nadali, “Overview and Future Opportunities of Sentiment Analysis Approaches for Big Data,” *J. Comput. Sci.*, vol. 12, no. 3, pp. 153–168, Mar. 2016, doi: 10.3844/jcssp.2016.153.168.

[8] S. C. Jayasanka, T. Madhushani, E. R. Marcus, I. A. A. U. Aberathne, and S. Premaratne, “Sentiment Analysis for Social Media,” p. 7, 2013.

[9] H.-I. Ahn and W. S. Spangler, “Sales Prediction with Social Media Analysis,” in *2014 Annual SRII Global Conference*, San Jose, CA, USA, Apr. 2014, pp. 213–222. doi: 10.1109/SRII.2014.37.

[10] Y.-E. PARK and Y. JAVED, “Insights Discovery through Hidden Sentiment in Big Data: Evidence from Saudi Arabia’s Financial Sector,” *J. Asian Finance Econ. Bus.*, vol. 7, no. 6, pp. 457–464, Jun. 2020, doi: 10.13106/JAFEB.2020.VOL7.NO6.457.

[11] F. Hemmatian and M. K. Sohrabi, “A survey on classification techniques for opinion mining and sentiment analysis,” *Artif. Intell. Rev.*, vol. 52, no. 3, pp. 1495–1545, Oct. 2019, doi: 10.1007/s10462-017-9599-6.

[12] N. H. Mahadzir, M. F. Omar, and M. N. M. Nawi, “A Sentiment Analysis Visualization System for the Property Industry,” *Int. J. Technol.*, vol. 9, no. 8, p. 1609, Dec. 2018, doi: 10.14716/ijtech.v9i8.2753.

[13] V. Hangya and R. Farkas, “A comparative empirical study on social media sentiment analysis over various genres and languages,” *Artif. Intell. Rev.*, vol. 47, no. 4, pp. 485–505, Apr. 2017, doi: 10.1007/s10462-016-9489-3.

[14] Y. Tian, T. Galery, G. Dulcinati, E. Molimpakis, and C. Sun, “Facebook sentiment: Reactions and Emojis,” in *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, Valencia, Spain, 2017, pp. 11–16. doi: 10.18653/v1/W17-1102.

[15] R. Sandoval-Almazan and D. Valle-Cruz, “Sentiment Analysis of Facebook Users Reacting to Political Campaign Posts,” *Digit. Gov. Res. Pract.*, vol. 1, no. 2, pp. 1–13, Apr. 2020, doi: 10.1145/3382735.

[16] C. Oh and S. Kumar, “How Trump won: The Role of Social Media Sentiment in Political Elections,” p. 14, 2017.

[17] T. Tran, D. Nguyen, A. Nguyen, and E. Golen, “Sentiment Analysis of Marijuana Content via Facebook Emoji-Based Reactions,” in *2018 IEEE International Conference on Communications (ICC)*, Kansas City, MO, May 2018, pp. 1–6. doi: 10.1109/ICC.2018.8422104.

[18] F. T. Giuntini *et al.*, “How Do I Feel? Identifying Emotional Expressions on Facebook Reactions Using Clustering Mechanism,” *IEEE Access*, vol. 7, pp. 53909–53921, 2019, doi: 10.1109/ACCESS.2019.2913136.

[19] T. Kadam, G. Saraf, V. Dewadkar, and P. J. Chate, “TV Show Popularity Prediction using Sentiment Analysis in Social Network,” vol. 04, no. 11, p. 3, 2017.

[20] J. Abraham, D. Higdon, J. Nelson, and J. Ibarra, “Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis,” vol. 1, no. 3, p. 22, 2018.

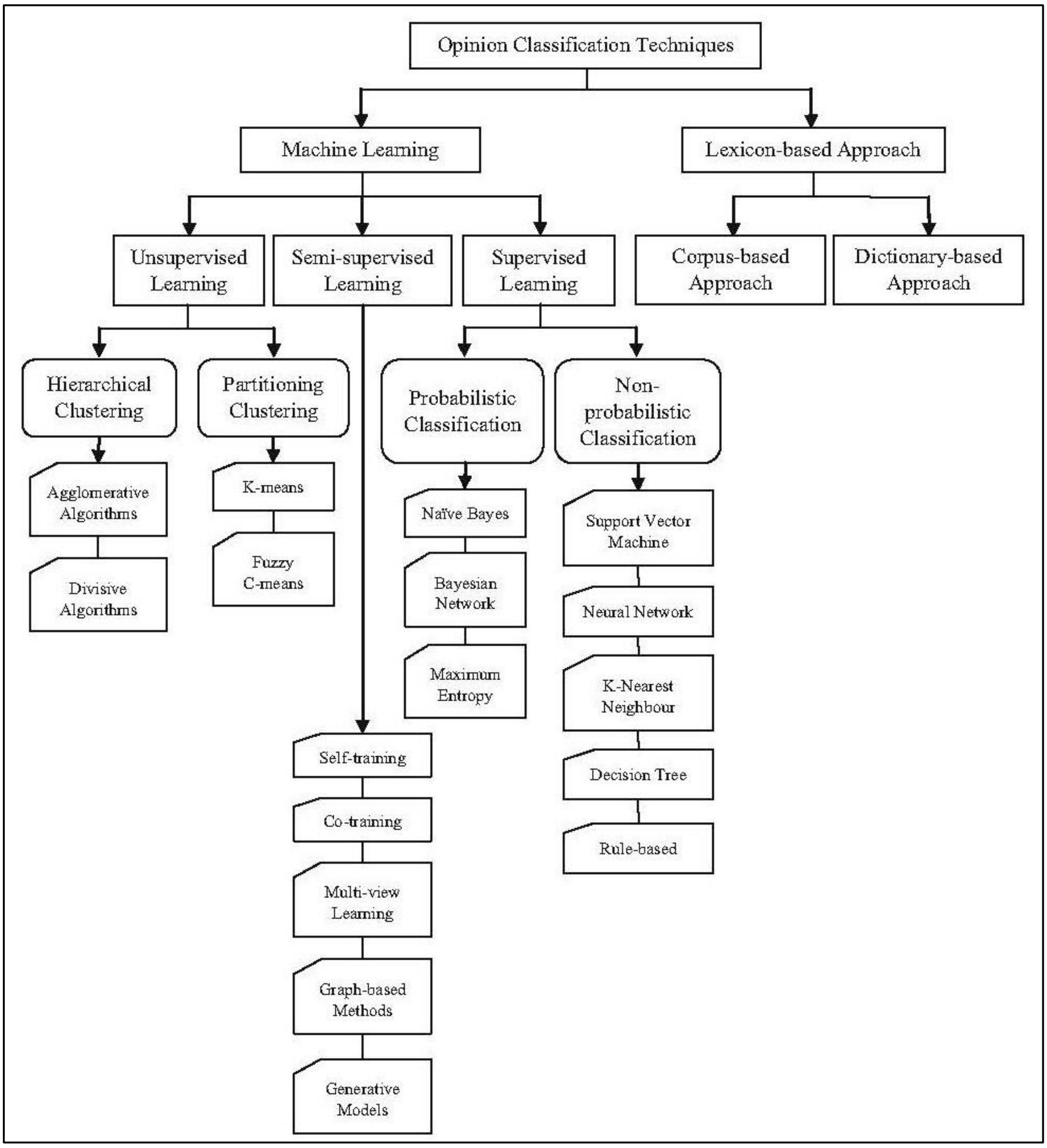
[21] P. Kralj Novak, J. Smailović, B. Sluban, and I. Mozetič, “Sentiment of Emojis,” *PLOS ONE*, vol. 10, no. 12, p. e0144296, Dec. 2015, doi: 10.1371/journal.pone.0144296.

[22] R. Folgieri, T. Baldigara, and M. Mamula, “SENTIMENT ANALYSIS AND ARTIFICIAL NEURAL NETWORKS-BASED ECONOMETRIC MODELS FOR TOURISM DEMAND FORECASTING,” *Congr. Proc.*, p. 10, 2018.

[23] A. Baheti and D. Toshniwal, “Trend Analysis of Time Series Data Using Data Mining Techniques,” in *2014 IEEE International Congress on Big Data*, Anchorage, AK, USA, Jun. 2014, pp. 430–437. doi: 10.1109/BigData.Congress.2014.69.

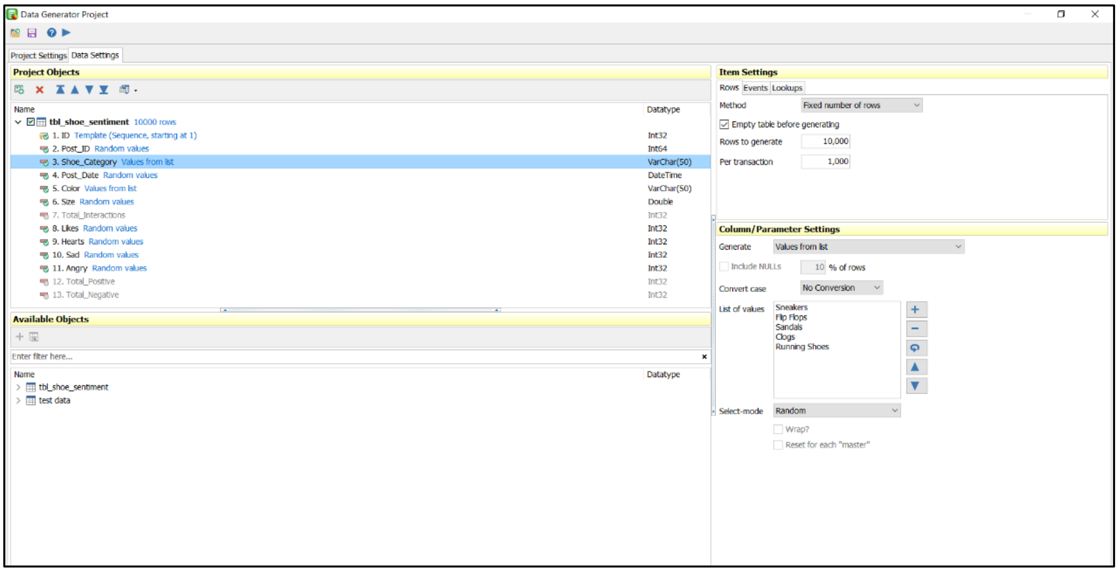
Appendix A

High level view of opinion mining classification techniques (Source: [11])

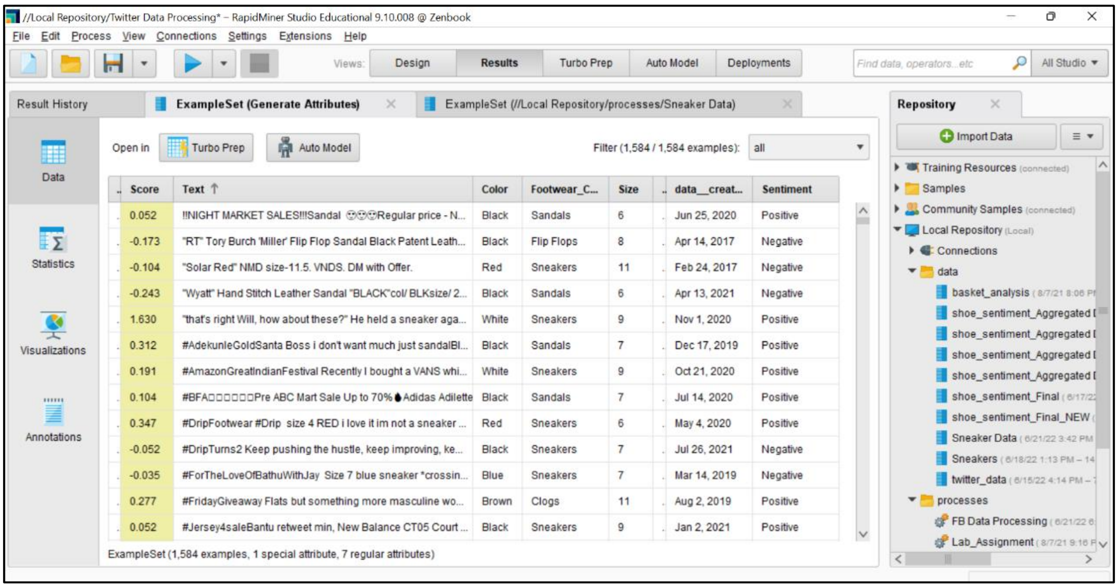


Appendix B

Parameters configured in Upscene Advanced Data Generator 4



Final output of sentiment calculated twitter post data



Appendix C

Sentiment Trend Dashboard