ASPECT DETECTION IN SPORTSWEAR APPAREL REVIEWS FOR OPINION MINING

Rajapaksha Wasala Mudiyanselage Polwatte Gedara Sampath Rajapaksha

(179345X)

Degree of Master of Science in Computer Science

Department of Computer Science and Engineering

University of Moratuwa

Sri Lanka

May 2021

ASPECT DETECTION IN SPORTSWEAR APPAREL REVIEWS FOR OPINION MINING

Rajapaksha Wasala Mudiyanselage Polwatte Gedara Sampath Rajapaksha

(179345X)

This Dissertation submitted in partial fulfillment of the requirements for the Degree of Master of Science in Computer Science Specializing in Data Science, Analytics and Engineering

Department of Computer Science and Engineering

University of Moratuwa

Sri Lanka

May 2021

DECLARATION

I declare that this is my own work and this dissertation does not incorporate without acknowledgement any material previously submitted for Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Name: R.W.M.P.G.S Rajapaksha

Signature: UOM Verified Signature Date: 28/05/2021

The above candidate has carried out research for the Masters dissertation under my Supervision.

Name of the Supervisor: Dr. Surangika Ranathunga

Signature of the Supervisor: UOM Verified Signature Date: 28/05/2021

ABSTRACT

As a result of the growth of social media sites and e-commerce websites, most of these websites provide platforms for people to express their opinion about their products or services. Main purpose of these platforms is to improve customer shopping experience. Moreover, these websites can use customer reviews to improve their products or services. In the sportswear apparel industry, almost all e-commerce websites provide these platforms for customers to leave their feedback. Since manual analysis of huge number of reviews is practically impossible, the automated approach of sentiment analysis/opinion mining has got the attention.

Sentiment analysis can be classified into 3 categories such as document-level sentiment analysis, sentence-level sentiment analysis and aspect-level sentiment analysis. Document-level or sentence-level sentiment analysis does not give the complete information as reviews consist with multiple entities and may have different opinions for different entities. This issue has inspired the aspect level opinion mining.

There are two core tasks involve with aspect level opinion mining. Those are aspect extraction and aspect sentiment analysis. This research aim at the first task of aspect level opinion mining which is aspect extraction task for sportswear apparel reviews as none of pervious works consider a clothing review dataset. A new data set will be produced with manual annotations by domain experts. This study used different deep learning models and achieved state-of-the-art performance for sportswear apparel reviews. It serves as the baseline for future research.

Keywords

Sentiment Analysis, Opinion Mining, Multi-label classification, Aspect Based Opinion Mining, Aspect Extraction, BERT, RoBERTa, Sentence Pair Classification, Multi-label classification.

ACKNOWLEDGEMENT

My sincere appreciation goes to my family for the continuous support and motivation given to make this thesis a success. I also express my heartfelt gratitude to Dr. Surangika Ranathunga, my supervisor, for the supervision and advice given throughout to make this research a success. I also thank Mr. Hansa Perera from MAS Holdings who provided the dataset for this research. Last but not least I also thank my friends who supported me in this whole effort.

Table of Contents

List of Figures

List of Tables

CHAPTER 1 : INTRODUCTION

1.1 Background

Buying behaviors and buying patterns of consumers are changing dramatically in the fashion industry. Sportswear apparel can be considered as one of the major subfields in the fashion industry. As per the survey done by Statistica [1] in 2016, below are the major reasons which motivate consumers to purchase new items especially for sports.

1. To have the appropriate clothing for a new level of performance
2. To improve the performance with the product
3. To look better while playing sports
4. To practice a different or new sport

The online apparel market is expected to grow at a very high face over the next decade. The major players of the online apparel market are Amazon, Gap, Walmart, eBay, Staples Inc, Kroger and Alibaba [2]. To improve customer satisfaction and shopping experience, all of these e-commerce sites encourage customers to leave feedback [3] and hence people rely on available online reviews when they purchase apparels. Further, consumers focus on different features (aspects) such as Fit, Size, Durability, Fabric, Comfort, Appearance, Price, Defect Free, Variety, Construction, Overall quality,…etc. Study [4] done in the UK, showed that 4.6% of conversion rate has increased if a product has more than 50 customer reviews. Further, 63% of consumers like to buy products from websites that have customer reviews. According to research done by Lightspeed Research, UK, they have found that 1 or 2 or 3 negative online reviews sufficient to prevent the 67% of customers form making a purchase.

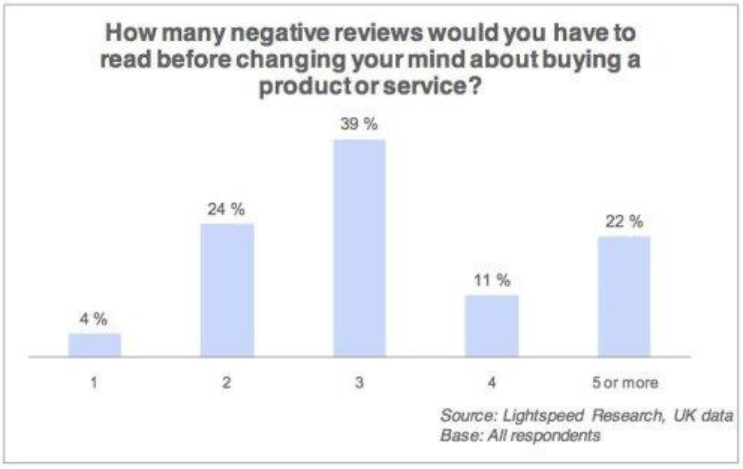

Figure 1: Importance of customer reviews
Source: Lightspeed Research, UK

These stats clearly show that customer reviews are sales drivers and success, or failure of e-commerce Sales highly depend on customer reviews.

1.2 Opinion Mining

Consumers and retailers/manufacturers use reviews in different ways. Consumers can seek opinions from other consumers when purchasing products or services and make decisions based on that. For retailers and manufacturers, they can identify consumer's opinion about their products and service and act based on those feedbacks. Due to these advantages, organizations and researchers have paid high attention to extracting valuable information from customer reviews. But with innumerable online reviews, manual information extraction is a really difficult task. Therefore opinion mining is used to extract information from numerous customer reviews.

1.3 Problem and Motivation

Although there's much research taking place in opinion extraction from customer reviews, few research works can be found for the apparel domain. Bao et al. [5] proposed 'aspect based positive center (ABPCS)' model for feature detection, opinion extraction and polarity classification for cloths and hotel domain reviews which were in the Chinese language. Other than this research, there is no research works that can be

found in aspect-based opinion mining for the apparel domain. But there are various research works in aspect-based opinion mining that can be found for domains like restaurants [6,7,8,9], laptops [8,9], cameras [8], phones [8]. Hence, having no system for sentiment analysis at the aspect level for apparel industry for English reviews can be identified as a problem.

1.4 Objectives

Mere Sentiment analysis will not be helpful in this scenario as aspects are the most important features for both manufacturers and customers who are looking for sportswear apparel product, hence rather than just considering the complete reviews, getting the customer opinion about certain aspects will be highly beneficial. Thus, finding an effective methodology to extract explicit and implicit aspects in sportswear apparel reviews will be the key objective of this research. This research focuses only on aspect extraction, and extracted aspect can be easily used for customer opinion mining using available classification algorithms. The objective of this research work can be summarized as below

1. Identify Critical Aspects of Sportswear apparel.
2. Create an annotated data set for sportswear apparel reviews.
3. Implement a deep ensemble method for aspect extraction from consumer reviews.

1.5 Contribution

This research focuses on developing a machine learning-based model to identify explicit and implicit aspects of sportswear apparel reviews. The contribution of this research can be summarized as follows.

1. Identify the critical aspects for sportswear apparel focusing on both consumers and manufacturers.
2. Created an annotated dataset for sportswear apparel reviews which can be generalized to all type of apparels and use for future research.

3. Achieved state-of-the-art results for ensemble model using transformer-based BERT and RoBERTa models for aspect extraction task of sportswear apparel reviews.

1.6 Report Organization

Chapter 01 gives the introduction to aspect detection in sportswear apparel reviews for opinion mining. Chapter 02 discuss the literature review which includes detailed analysis of levels of opinion mining, aspect extraction methods, data preprocessing techniques and features selection. Final chapter discuss the methodology of this research work.

CHAPTER 2 : LITERATURE REVIEW

2.1 Opinion Mining

Opinion mining, which is also known as sentiment analysis, represents same field of study. Further, opinion extraction, sentiment mining, subjectivity analysis, affect analysis, emotion analysis, review mining' all these terms can be used to define Opinion mining [10]. Moreover, Liu [10] defined the opinion as "An opinion is quadruple (g, s, h, t) where g is the opinion (or sentiment) target, s is the sentiment about the target, h is the opinion holder and t is the time when the opinion was expressed". Opinion mining is considered as a difficult problem to solve considering the unstructured nature of texts used in documents and sentences [6].

Following review from apparel domain can be used to illustrate his basic definition.

Posted by: Sarah

Date: 2017.05.12

Product: Nike Mens Legend Short Sleeve Tee

Review: *(1) My boyfriend loves the way this shirt fits. (2) I even steal it from him sometimes because I love the way it feels! (3) It doesnt shrink-which is a great plus! (4) The color has stayed rich through washing several times as well! (5) I will definitely be ordering more.*

Below are the important points for this review

1. This has several opinions about Nike Men's Legend Short Sleeve T-Shirt. Sentences 1, 2, 3, and 4 express positive opinions toward aspect fits, comfortability, quality, and color respectively whereas sentence (5) express positive opinion about the T-shirt as a whole.
2. Above review consists with opinions of two persons which are known as opinion sources or opinion holders [10]. Sentences 2, 3, 4 and 5 are the opinions of the author and sentence 1 is the opinion of author's husband.

3. Date of this review is 2017.05.12. This is important as it helps to identify how opinions change over time and opinion trends.

2.2 Opinion mining at different levels

Opinion mining can be classified into 3 levels such as document-level opinion mining, sentence-level opinion mining and aspect-level opinion mining [6].

2.2.1 Document-level opinion mining

For the document-level opinion mining, overall opinion about the document is considered and classified as positive, negative or neural [6]. Document-level analysis assumes that it expresses an opinion on a single entity or aspect [11]. Most of the time, document or reviews consist of multiple entities and may have different opinions for different entities. Hence document level opinion mining does not perform well for documents or reviews with multiple entities or multiple opinions.

2.2.2 Sentence-level opinion mining

Sentence-level opinion mining use to identify opinion expressed for each sentence. But much difference cannot be identified between sentence and document-level opinion mining as sentence are just short documents. Sentence level opinion mining assume that usually sentence contains only one opinion. But multiple opinions may include in a sentence. This is still useful if we know what are the entities that are discussed within the review [10].

2.2.3 Aspect-level opinion mining

Reviews contain positive, negative and neutral opinions about different aspects [6]. But both document-level and sentence-level opinion mining do not identify opinion targets or assign sentiments to those targets [10]. Hence it requires more detailed analysis to find these different opinions about different aspects. Aspect based opinion mining extract specific aspects from reviews and determine the polarity towards each aspect

separately. Aspect based opinion mining can be illustrated with the below example shown in table 1.

Review: *"Firstly, the pictures are totally deceptive. I ordered this product primarily because I loved the color and it seems comfortable. It is true this shirt is really comfortable. However, I was a bit disappointed when it arrives since it was not the size I expected. It is like totally different. Secondly, there are couple mentioned about the trademark at the left chest. I laundered only once but one of three has trademark gone. Very disappointing. And lastly, before I found out the trademark gone, I wasnt that feel bad about the size. However, now it makes this shirt really negative on it. I normally wear medium of all most every product. but this product is WAY small for my son. I think he is gonna wear this less than others."*

Table 1: Aspect Based Sentiment Analysis

Aspect	Sentiment	Sample Sentence
Comfortability	Positive	"*It is true this shirt is really comfortable*"
Size	Negative	"*I was a bit disappointed when it arrive since it was not the size I expected but this product is WAY small for my son*"
Quality	Negative	"*I laundered only once but one of three has trademark gone. Very disappointing*"

Aspect can be explicit or implicit. In explicit aspects, explicit aspect words are directly use in the texts while implicit aspects do not use direct words to represent the aspect [7]. Below example exhibit the usage of explicit and implicit words in a review. "*Perfect Shirt, I love the high quality and it is very comfortable and light for the summer! I gave 4 stars because is a little large for me.*" In this example aspects 'quality' and 'comfortability' are explicitly mentioned, whereas size is implicitly mentioned.

Implicit aspect identification is much difficult than explicit aspect identification [4]. The reason is that aspect can be associated with multiple entities, or there can even be aspects that do not directly attach to any entity. Compared to explicit aspect identification, there is high complexity involved with tracing implicit aspects from

reviews [13]. As an example, "*Very comfortable and perfect fit with regards to size chart. I take great care of it and wash it by hand. Also better to avoid using a tambour dryer*", in this review the aspect size mentioned explicitly. In review: "*They are a little thinner than some of the others I have purchased in this brand.,* this review also discuss about aspect size, but word 'thinner' has been used to express the idea about the aspect size. Hence, it's a complex task to extract the aspect in the second review compared to the first review.

Most of the researchers have identified two core tasks in aspect-based opinion mining. Liu [10] identified them as Aspect extraction and aspect sentiment analysis. Panchendrarajan et al. [7] identified the same as detecting aspects and classifying sentiment score for each aspect. As per Hercig et al. [8] Aspect Based Sentiment Analysis (ABSA) identifies the aspect mentioned in sentences and then identify the polarity of opinion. Another research, Chinsha and Joseph [6] has identified the same tasks as aspect detection, aspect-based opinion word detection and orientation detection of the aspect. These three tasks can be described with below

Example review: "*Awesome and comfortable shirt! I wear when working out, playing tennis, and just around the house. Bit expensive"*

Explicit aspect comfortability and implicit aspect price (expensive) can be identified as the first step. The second step is to identify the opinion words , which are awesome and expensive. In this case word expensive, express both implicit aspect and opinion word. Finally, detect its orientations, which are positive and negative.

2.3 Aspect Extraction

Aspect and opinion extraction tasks are considered as the most important and challenging tasks [13] among above identified three tasks. Hence most of the researchers have given their attention to aspect extraction. Compared to explicit aspect extraction, implicit aspect extraction considered as a much difficult task [7]. Rana and Cheah [13] have produced a comprehensive review paper for aspect extraction. More than 50

techniques were discussed for explicit aspect extraction while found only 11 studies for implicit aspect extraction. Moreover, authors have been divided explicit aspect extraction into three main categories, supervised, semi supervised and unsupervised.

2.3.1 Unsupervised Methods

Supervised aspect extraction requires a set of annotated training data. But it is expensive and requires much human labour involvement to label training data. Since generally publicly available data are unlabeled and there are various reviews available for different languages, products, services, and domains, it's desirable to develop unsupervised models that are robust and easily transferable between different languages and domains. Most of the time, supervised approaches achieve reasonable effectiveness compared to unsupervised methods [14].

2.3.1.1 Frequency or Statistical

Frequency or statistical methods identify frequent aspects which are expressed by most of the reviewers. Researchers have observed that [9], [15], usually noun and noun phrases represent the aspects. Hence noun and noun phrases are extracted from sentences and identify as frequent aspects if they exceed the defined support threshold. After identifying nouns as aspects, nearest adjectives are extracted as opinion words [13]. After extracting noun and noun phrases, Hu and Liu [15] used association rule miner, which was based on the Apriori algorithm to identify the frequent item sets. As per them, not all frequent item sets generated by association rule minor were genuine aspects. To remove the words which are not actual aspects, the authors used two pruning approaches. Those are compactness pruning, which checks features containing at least two words, and remove those that are likely to be meaningless and redundancy pruning, which focuses on removing redundant features containing only one word. Bafna and Toshniwal [16] further improved Hu and Liu's approach by using a combination of association rule mining and probabilistic approaches. Features were extracted using the same approach used by Hu and Liu. However, they have used probabilistic technique to identify only the relevant aspect words. Although, frequency

or statistic approach is simple and effective [17], it is challenging to implement for the languages with a lack of reliable NLP [18].

2.3.1.2 Bootstrapping (Unsupervised)

Bootstrapping provides a solution to manual annotations like other unsupervised algorithms. Bootstrapping algorithms initiate with carefully chosen seeds and these seeds are used to collect other similar words (terms) from unannotated corpus. This repeats until the algorithm meets certain stropping criteria [19].

Zhu et al. [20] introduced multi-aspect bootstrapping (MAB) based approch to identify aspects and experiments were done on real chinese restaurant reviews. There may be several aspects on a review and MAB was used to identify all aspects from a review without utilizing annotated data. Bagheri et al. [14] proposed three steps unsupervised approach to extract aspects. As the first step, a generalized approach is used to identify multi-word aspects. Then, a set of heuristic rules were used to identify aspects using relevant opinion words. Finally, They used a bootstrapping algorithm to assign a score to aspects with a new mutual information based matric. unsupervised seed set was used for the bootstrapping algorithm.

There is a major weakness of bootstrapping methods. That is, after multiple iterations, the bootstrapper walks away from the initial semantic meaning of the seeds and selects incorrect words [19].

2.3.1.3 Heuristic or Rule-based

In heuristic or rule-based approaches rules are used to extract natural language processing (NLP) information such as nouns, verbs and adjectives. Later, these information (features) can be used for classification tasks.

Liu et al. [21] introduced a new approach to extract aspects using word-based translation model (WTM). They found out relationship between aspects(opinion targets) and opinions using monolingual word alignment model. After identifying associations, a graph-based algorithm were used to extract aspects. Like frequency or statistical method,

authors considered noun and noun phrases as aspects and adjectives as relevant opinions. The difference of selecting opinion words with frequency or statistical method is, nearest adjectives were not considered as opinions; instead, they used a graph-based model to detect opinions. A new algorithm ASPECTATOR was introduced by Bancken et al. [22] to automatically detecting and rating review aspects. This approach matched few syntactic dependency paths to identify candidate aspects and did not require seed words required by bootstrapping algorithms. They have defined ten handcraft dependency paths to detect relevant aspects and relevant opinion words.

Poria et al. [9] introduced a rule-based approach which is based on common-sense knowledge and sentence dependency trees to extract both explicit and implicit aspects. Implicit aspects were identified using implicit aspect clue (IAC). This approach first identifies the IACs in reviews and then map them to relevant explicit aspects. Several dependency rules were used to identify such aspects and WordNet, SenticNet were used to identify synonyms and semantics of those IACs.

2.3.1.4 Topic Modeling

Topic modelling is a probability distribution-based approach which assumes that a document formulates by a mixture of topics. Topic modelling produces a collection of word clusters in which each word cluster forms a topic [10]. These topics are the aspects of documents (or reviews) and hence topic modelling is used to extract aspects. Mainly there are two basic models, PLSA (Probabilistic Latent Semantic Analysis) and LDA (Latent Dirichlet allocation) [10], [17]. Mei et al. [23] have suggested a probabilistic mixture model ‘Topic Sentiment Mixture (TSM)’ to extract subtopics and sentiments in a collection of weblogs. PLSA model has been used in their approach. Brody and Elhadad [24] introduced a local topic model LDA to work with sentence-level and employs a small number of topics. They identified aspects with the topic model and then adjectives were used to identify relevant opinion words. Zhao et al. [25] proposed MaxEnt-LDA, a hybrid model to jointly discover both aspect words and aspect-specific

opinion words. This model is an extension of LDA and captures both aspect words and sentiment words.

Even though topic modelling is a powerful aspect extraction method, there are few weaknesses. It can identify only some general aspects and difficult to identify precise aspects [17], [26]. Furthermore, topic modelling requires a large volume of data and high tuning [18].

2.3.2 Semi -supervised Methods

Semi-supervised approaches combine a small labelled dataset along with a large unlabeled dataset [10]. Semi-supervised methods partially depend on user inputs and require initial seed words to start the algorithm [13]. Different categorization may be needed based on the specific application. Since the semi-supervised approaches requires initial seeds from users, it's able to provide what user exactly needed [27]

2.3.2.1 Bootstrapping (Semi-supervised)

Weng and Weng [28] proposed a bootstrapping iterative learning strategy that take seed words as input to extract both aspect and opinion words. Moreover, infrequent aspects were identified using a linguistic rule. Zhao et al. [29] introduced a joined model of automatic refinement and bootstrapping based framework to detect opinion words and relevant aspects. To start the bootstrapping, they were used a small set of words as seeds and pre-defined rules. Statistical and dependency patters were used to identify the relationship between opinion words and targets.

2.3.2.2 Double propagation

As per Qie et al.[30] always there exist a relationship between aspect words and opinion words. In double propagation, it transmit information between these words and targets of interests. Based on this property Qiu et al. [30] proposed an approach based on double propagation to extract opinion words and targets. Usually as used in frequency or statistical methods and heuristic or rule-based, opinions are considered as adjective words and nouns are the respective aspects of a sentence. Using the relationship between

aspects and opinion words, they proposed a method to detect aspect from already known opinions and extract opinion from known aspects. Since this approach transfer information back and forth between targets and opinion words, it considered as double propagation. The proposed approach requires some words as seed inputs to identify opinions. Once a few opinion words were extracted, targets can be extracted using extracted opinion words and then both extracted opinion words and targets can be used to extract new targets and opinion words respectively. This approach will continue until no opinion word or target is found. In [18], authors have followed the same approach proposed by Qiu et al. [30]. They used double propagation to arabic reviews and to their english translation. To use the Stanford NLP tools, they have converted arabic reviews in to english using microsoft bing translator. Stanford POS tagger has been used to identify nouns, noun phrases and adjectives. Zhang et al. [31] used the [23] approached with two improvements based on part-whole and no patterns.

Although double propagation is an effective method, due to the flexibility of neural languages, it is hard to define rules to achieve high precision and recall [26].

2.3.2.3 Dependency parsing

Dependency parsing extracts dependency parses of sentences and identify the relationship between head word and other word of the sentence. Wu et al.[32] introduced a novel approach called phrase dependency parsing, which improves traditional dependency parsing to phrase level. Tree is generating using dependency parser and it gives a connection between different words in the review sentence. Since 98% of their review aspects were noun, noun phrases or verb phrases, they considered noun phrases and verb phrases as aspects and pruned some aspects which did not achieve a defined threshold. Further, they used dictionary-based approach to extract relevant opinion words. Then dependency parser was used to identify the relationship between aspects and opinion words. Yu et al.[33] also used the same approach to detect aspects and opinion words of customer reviews. Unlike other semi-supervised algorithms discussed

earlier, this approach assume that opinion words and aspect words are located next to each other.

2.3.2.4 Lexicon-based

In lexicon-based approach a sentence is represents as bag of words. These words are identified as positive or negative words considering a dictionary of positive and negative words. Sum or average value is used to come up with the final prediction for overall sentiment of the sentence.

Wei et al.[34] proposed a semantic-based product feature extraction (SPE) technique to extract product features from consumer reviews. They have used positive and negative adjectives to identify opinion words and detect product aspects. Ma et al.[35] introduced a new product feature-oriented approach to analyze customer reviews to get feature-based inquiries and review summarization. Their approach combines LDA (Latent Dirichlet Allocation) which used in topic modeling and a synonym lexicon to extracts aspects from consumer reviews. Like other researches, noun and noun phrases considered as aspects and those have incorporated with an aspects list which was obtained using LDA.

2.3.2.5 Word alignment or graph-based

By assuming the relationship between nouns and aspects, word alignment or graph-based approaches find the opinion and aspects from reviews. Liu et al. [36] proposed an approach to extract review aspects using a semi-supervised word alignment model (PSWAM). First, they have extracted opinion relations in reviews and estimated the association between words using the proposed model. Noun and noun phrases have been considered as the aspects of reviews. For the detection of actual aspects, graph-based algorithm has been used and estimated the confidence of candidates. Defined threshold was used to extract the relevant aspects. Xu at al. [30] have used a two-stage approach to extract opinion words and aspect words. Sentiment graph walking algorithm was used to

identify pattern between opinion and aspect words. Then, semi-supervised TSMV approach has been used to remove nouns that do not represent actual aspects.

2.3.2.6 Statistical based

Statistical or frequency based approaches identify the frequent aspects and hence find the opinion words aligned with them. Liu and Mukherjee [27] proposed a statistical-based semi-supervised model for aspect extraction. Two separate statistical models were used to detect and classified aspects based on some seed words. Two models are called as Seeded Aspect and Sentiment model (SAS), and Maximum-Entropy Seeded Aspect and Sentiment model (ME-SAS).In ME-SAS model, priors helped to separate aspects and opinions. These models are related to unsupervised topic modelling and joint models of aspects and sentiments. The major difference is, this approach requires seed words as inputs.

2.3.3 Supervised Methods

Supervised methods are the most popular when it comes to traditional information extraction tasks [17]. Supervised methods require manually labelled training data which are not required by unsupervised or semi-supervised methods. Hidden Markov Models (HMM) and Conditional Random Fields are considered as the two main techniques for sequential learning [10].

2.3.3.1 Hidden Markov Model (HMM) based

HMM is considered as a directed sequence model. These models have been used in many sequential labelling tasks like name entity recognition (NER), POS tagging and information extraction (IE) .

To model the joint distribution, two independents assumptions are made for HMM. These assumptions may not be suitable for practical problems and will give a low performance. This is considered as one limitation of HMM [17].

Jin and Ho [50] introduced a novel machine learning model using lexicalized HMMs to detect the aspects and opinions from customer reviews. Two tag sets have been used in this research work. Different categories of entities defined using a basic tag set, whereas patterns for different entities defined by the second tag set. Then they have used these two tags sets to manually tag sentences that representing the patterns between aspects and opinion words. Using the HMM and maximum likelihood estimation (LSE), found out the appropriate sequence of hybrid tags that maximize the conditional probability. As the next step, sentences which contain aspect opinion pairs have been identified and removed all sentences which not include any opinion word [13], [50]

2.3.3.2 Conditional random field (CRF)

CRF is considered as an undirected sequence model. For the aspect-based sentiment analysis (ABSA) task of SemEval 2016, Hercig at al. [16] have used conditional random field (CRF) for opinion target extraction. The task of ABSA was to extract the aspect of a given target entity and estimate the sentiment polarity for each mentioned aspect. They have done the experiments for Chinese, English, French, and Spanish languages. CRF has been used for the Sentence level (SB1) opinion target expression (OTE)

Based on the previous work [15], Li et al.[39] used skip-chain CRFs to extract aspect and opinions by considering the long-distance dependency with conjunctions. Tree-CRFs was used to learn the systematic structure of the sentence. For the aspect based sentiment analysis task 12 of SemEval-2015, Toh and Su [38] used a system based on two supervised learning algorithms. For aspect category classification, sigmoidal feedforward network and, for opinion target detection, conditional random fields have been used.

Jakob and Gurevych [40] have used the CRF-based approach to address the domain portability problem. Their goal was to extract opinion targets from user-generated discourse. As per their observations, a word can give different meaning for different domains.

Although CRF is a powerful approach for aspect extraction, since it's a linear model, it requires a large number of features to give good results [41].

2.3.4 Deep Learning

Deep learning-based approaches have given remarkable results for NLP tasks such as text classification and sentiment analysis in the recent past [42]. But none of the researches have used deep learning-based approaches for aspect extractions until Wang and Liu [43] have proposed a deep neural net-based approach for the data set given for SemEval-2015 Task 12 aspect-based sentiment analysis (ABSA). They have developed a framework with two deep learning models for aspects and sentiments. This architecture is consisting of an aspect model and a sentiment model. The set of word vectors is sent to the aspect model and it gives probabilistic distribution over the aspects. In contrast, the sentiment model takes a sentence and gives the sentence's sentiment. The aspect model has used two layers of neural network with a fully connected first layer and the second layer outputs a softmax distribution. Authors have used Google news 300-dimensional word vectors to train word vectors because the training set had only 2000 sentences. A recursive neural tensor network (RNTN) was used for the sentiment model. Their model has achieved competitive or better performance compared to the best results of SemEval 2015 all subtasks.

Ruder at al. [42] have used convolutional neural network (CNN) for both aspect extraction and aspect-based sentiment analysis in SemEval 2016 task 5 of aspect-based sentiment analysis. They have considered aspect extraction as a multi-classification problem and have used a convolutional neural network to get probability distribution over aspects.

Even though most of the previous aspect extraction tasks used CRF and HMM based methods, they have various limitations. As a solution, Poria at al [41] introduced convolutional neural network (CNN) to aspect extraction. They have used 7-layer deep CNN to tag words in opinionated reviews (sentence) as either aspect or non-aspect words. Further, they have used back-propagation to train the neural network to

maximize the likelihood of training sentences. They identified that the feature of an aspect word depends on its surrounding words and hence used a window of 5 words around each word in a sentence. Two words of left and right of a particular word considered as features and fed them into CNN.

2.3.4.1 Transformers

In 2017, researchers in Google introduced a new deep learning model known as Transformer in the paper 'Attention is all you need' [54]. Transformers are based on the concept of attention mechanism and encoder-decoder architecture. Transformers are designed for sequential data, but does not require sequential data to be process in the order. This allows much more parallelization. Furthermore, transformers can achieve state-of-the-art in translation quality with comparatively limited time training.

Using the architecture of transformers, in 2018, Researchers at Google AI Language, introduced a new model, Bidirectional Encoder Representations from Transformers (BERT) [55]. Instead of restricted existing unidirectional language models, they introduced bidirectional pre-training for language representations [55]. Pre-training and fine-tuning are the two steps that they used in their framework. For the pre-training task they were used two training methods. The first method is Masked Language Model (MLM). 15% of words of a sentence were masked and fed to the model. Then masked words were predicted using the context of unmasked words. The second method is Next Sentence Prediction (NSP). This was used the concept of understanding the relationship between two sentences. For this, they used two sentences as input to the model and predicted the order of sentences. For the fine-tuning task, they used the self-attention mechanism and fine-tuned all the parameters end-to-end. The below figure shows the BERT input representation.

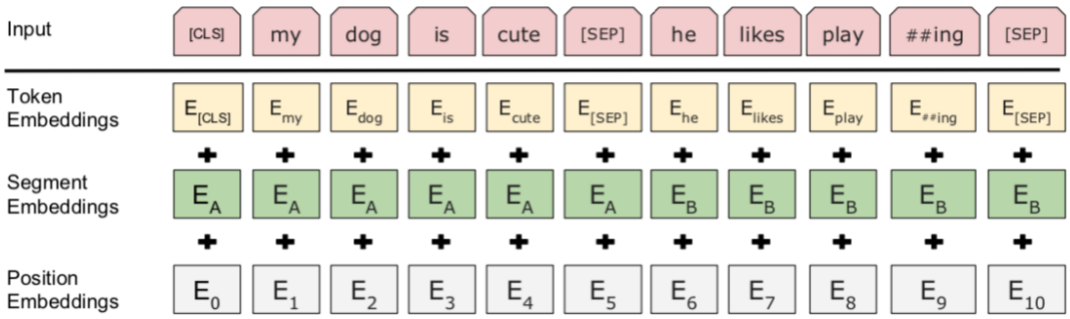

Figure 2: BERT input representation.

(Source : J. Devlin et al.[55])

Four tasks have been discussed in this paper. Those are the sentence pair classification task, single sentence classification task, question answering task and single sentence tagging task. These tasks were performed by incorporating BERT with one additional output layer. The below figure shows the fine-tuning of sentence pair classification and single sentence classification tasks.

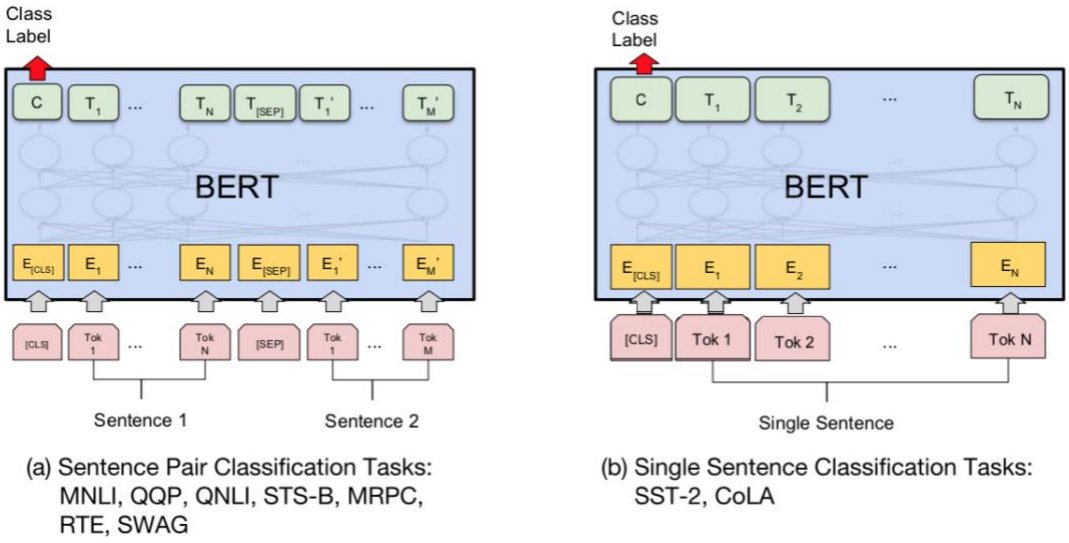

Figure 3: BERT fine-tuning

(Source: J. Devlin et al. [55])

Sun et al.[56] used the BERT sentence pair classification for the aspect-based sentiment analysis (ABSA). This task has two subtasks which are aspect detection and aspect polarity detection. To convert the ABSA to a sentence pair classification task, they construct an auxiliary sentence from the aspect. They could achieve the state of the art results by fine-tuning the pre-trained model from BERT for SemEval-2014 task 4 dataset for both aspect detection and aspect polarity detection tasks. Further, this approach outperforms the BERT single sentence classification for the ABSA task. In [57], the authors implemented the ABSA task as a question answering problem. They used BERT as the base model with a novel post-training approach. In this approach, the authors used a technique called Review Reading Comprehension (RRC). Experimental results which used SemEval 2016 Task 5 ABSA dataset shown that this technique is highly effective. Hoang and Bihorac [52] used BERT and treated the ABSA as a sentence pair Classification Task. This approach can handle out-of-domain aspects as well. Experimental results have shown that proposed model outperforms previous state-of-the-art results for SemEval 2016 task 5 (ABSA).

Researchers at Facebook AI, introduced an improved version of BERT which is called as RoBERTa or Robustly optimized BERT approach [58]. This modification to BERT includes 4 steps which are, longer training period with more data and bigger batch size, removing next sentence prediction task, use longer sentences for training and dynamically changing the masking patterns which used in BERT. RoBERTa achieved state-of-the-art results on GLUE, RACE and SquAD. The authors highlight the fact that this model can be used to improve the performance of every BERT model.

Tain and White [59] introduced a pipeline of aspect detection and sentiment analysis for E-commerce customer review. gate-RoBERTa based sentiment classifier were used for the classification task. But for the aspect extraction task, they used noun based approaches similar to techniques used in frequency or statistical models.

2.3.5 Implicit Aspects

Implicit aspect extraction is considered as a much more complex task than explicit aspect extraction. Hence limited studies can be found for implicit aspect extraction compared to explicit aspect extraction [44]. Some of the papers which were discussed in earlier sections capable of extracting both explicit and implicit aspects.

2.3.5.1 Unsupervised Methods

Poria et al. [9] introduced a novel rule-based approach that uses common-sense knowledge and sentence dependency trees to extract explicit and implicit aspects. They have defined implicit aspect clue (IAC) that is similar to Implicit Aspect Indicators (IAI) which was defined by Cruz et al. [44]. IAC use to express aspects indirectly. As per the authors, implicit aspect extraction is a two-step process which include identification of IAC and related them to relevant aspect words. Cruz at al[44] have done only the first step in their approach. But Poria et al. [9] proposed solutions for both steps. To evaluate the explicit aspect extraction algorithm, they have used a corpus provided by Hu and Liu and SemEval 2014 dataset while corpus developed by Cruz at al [44] was used for implicit aspect extraction. Using this corpus, authors have extracted sentences that have implicit aspects. Then IACs were extracted from each sentence, with their respective explicit aspect (label). Synonyms and antonyms for each IAC were obtained using WordNet, while semantics were obtained using SenticNet. Further, they have obtained a dependency parse structure for each sentence using Stanford parser and then manually defined rules were used on the parse tree to extract aspects.

2.3.5.2 Semi-Supervised Methods

Weng and Weng [28] used a semi-supervised, bootstrapping-based approach to identify explicit and implicit aspects. Mutual information was used to calculate the association between aspect and opinion words and that was used to identify implicit aspects. For the sentences which have opinion words, but it does not have explicit features, they used a defined mapping functions to obtain implicit words for each opinion word.

2.3.5.3 Supervised Methods

Cruz at al. [44] presented a novel method to implicit aspect indicator (IAI) extraction using CRF which is based on sequential labeling. In a review, words such as 'lightweight', 'sleek', 'attractive', 'user-friendly', 'easy to manipulate' used as clues to infer the implicit aspects and these words considered as Implicit Aspect Indicators (IAI). Most of the other explicit and implicit aspect extraction methods consider aspect as noun or noun phrases. Implicit aspect extraction is a two-phase approach. First, IAI are identified and then they are mapped to the relevant aspect.

Panchendrarajan at al [7] proposed a method to detect multiple implicit aspects which are mentioned in restaurant reviews. This approach considered as each opinion word implies an implicit aspect. They have used a manual tag data set to train a model and used that model to identify the aspects which implied by these opinion words. Then calculated a score for each aspect using the co-occurrence between opinion words and other words. Based on the highest score, the potential candidate aspect has been selected. Finally, opinion targets and opinions were extracted and check the relationship with the predicted aspect. Opinion targets were extracted using double propagation approach. To identify the relationship with the predicted aspect, they have modelled different entities with different aspects as a hierarchy.

Schouten and Frasincar [45] developed a supervised learning-based approach to identify whether an implicit aspect is present in a sentence and, if it is present, identify the relevant implicit aspect. They have used a threshold to identify the actual implicit aspects and discard others. Further, the authors have listed two limitations of their approach which are, this method only extracts one implicit aspect even though there might be many implicit aspects within the same sentence and the other limitation is, this requires a sufficient amount of annotated data as this is a supervised approach. Various techniques have used in the literature to identify implicit aspects in reviews and most of them relied on pre-defined set of rules or words. However, the approach used by

Schouten and Frasincar [45] did not require any manual rules and pre-defined words to identify implicit aspects.

2.3.6 Summary

In aspect extraction, there are two types of aspects that can be identified as explicit and implicit aspects. Unsupervised, semi-supervised and supervised approaches can be used to identify aspects. But deep leaning based approaches have given good results in the recent past. Implicit aspect extraction is considered as a much more complex task than explicit aspect extraction. Some techniques can identify both explicit and implicit aspects in one step, whereas others handle these separately. Further, some approaches assume that there's only one aspect per sentences and others assume that there are multiple aspects in a sentence. Most of the researches have considered that aspects are noun and noun phrases. It seems, there's no perfect aspect extraction approach available for all aspect extraction tasks as different approaches have their own limitations.

2.4 Data Pre-Processing

Noisy of text data is considered as one of the major challenge of NLP tasks [46]. This is due to the online communication data (reviews, blogs, comments, posts…etc.) are informally written. Those are full of spelling mistakes, grammatical errors, punctuation errors and irrational capitalization [46]. But most of NLP tools requires clean data to give desired results. Therefore, data preprocessing is an essential task that need to do before any analysis. Below are some of the frequently used data preprocessing techniques and those are depending on the data set and NLP task.

- Spelling correction – isolated spell correction and context sensitive spell correction
- Tokenization
- Lemmatization
- Stop words removal
- Handling/Removing special symbols/numbers and characters

- Removing unnecessary HTML tags
- Convert to lower case or upper case

Lower casing and lemmatization were the data pre-processing technique that Hercig et al. [16] used in their SemEval 2016 target extraction approach. All sentences are tokenized and passed using Stanford parser as the first step in Jin and Ho [50] machine learning framework using lexicalized HMMs to extract the aspects and opinions from customer reviews. Xu at al. [37] have used a two-stage technique to extract opinion words and opinion targets and as the data pre-processing technique, they have removed HTML tags from the texts.

2.5 Feature Selection

Feature selection is a key step to increase the accuracy of sentiment analysis. Like data pre-processing, feature selection also highly depends on the dataset and task. Below are some of the frequently used features for sentiment analysis tasks.

- Bag of words

  This is considered as one of most simple features to extract and has achieved great success in document classification and language modeling. But it has few limitations like vocabulary, sparsity and meaning [47].
- Word N-grams

  Most of authors have used unigram as features to classify reviews. Even though this has provided good results, it has failed in some situations. In such situations, higher order n-gram expected to give better results. Other than review classification, word N-gram has been successfully used as a feature to aspect identification [7] and extract opinion words [48].
- Character N-grams

  Most of the times, this feature has been used in name entity recognition tasks [44]. One limitation of this feature is, with larger n-grams the algoritm training time will increase.

- Word embedding

  Recently, deep learning approaches have used this feature. In SemEval 2016 task 5, word embedding has been used with aspect vector to determine aspect towers an aspect [42]. Moreover, this can be used to increase the accuracy of document classification along with bag of words [49].

- Part of Speech tags (POS)

  Most of the NLP tasks have used POS as a feature. Usually aspects are nouns or noun phrases, hence POS has used as a feature to identify aspects in most of aspect extraction tasks [14], [17], [40].

- Head word (head noun)

  This has successfully used in SemEval 2015 Aspect based sentiment analysis task [38] for both laptop and restaurant domains. They have got high F1 when head word included to existing features (Word, Bigram and name list)

- Word cluster

  This also has been used in SemEval 2015 Aspect based sentiment analysis task [38] and they could increase the F1 score when they included word clusters as a feature.

- Named Entity recognition (NER)

  NER has used to identify names of things such as person, location, and organization in documents or reviews.

- TF-IDF

  In SemEval 2016 aspect-based sentiment analysis task, Hercig et al. [8] used large number of features including TF-IDF.

  Jin and Ho [50] proposed a novel machine learning framework using lexicalized HMMs to extract the aspects and opinions from customer reviews. In their approach they have used six features which another researcher proposed. Those features are, current word, bigram, name list, head word, word cluster and name list generated using double propagation.

Li at al [39] summarized reviews based on object features. They have used word features, dictionary features, sentence features, syntactic features, and edge features. Below is the detailed view of selected features.

Cruz at al. [44] presented a method to implicit aspect indicator (IAI) extraction using CRF. They have used a feature vector contacting, word features, character n-gram features, POS tags, context features, class sequence features and word bi-gram features. Alawami [18] has used six features for his CRF model. Those features are, Part of Speech tags (POS), Super Part of Speech (SPOS), Named Entity (NE), Short dependency path (SP), Word distance (WD), Sentiment words (SW).

CHAPTER 3 : RESEARCH METHODOLOGY

The main objective of the research is to extract pre-identified aspects from sportswear apparel reviews. This chapter will discuss data collection, data annotation, exploratory data analysis and algorithms used for aspect extraction.

3.1 Customer Reviews Data Collection

Most of the e-commerce sites ask customers to leave feedback/reviews for their purchases. When it comes to sportswear apparel, there are numerous online clothes shopping sites selling sportswear appeal products. Amazon[1] and DICK'S[2] Sporting Goods are the two major e-commerce sites with a huge customer base and are receiving an equivalent number of reviews from customers for their purchases. 4234 extracted customer reviews from Amazon, DICK'S Sporting Goods and Nike.Inc will be used in this research. Further, only the reviews received for sportswear brand 'Nike' will be considered here, and the outcome can be used to any apparel brand as aspects are the same for all sportswear apparel regardless of the brand. Dataset is available in .CSV format. Below is the content of the dataset

- Product Name
- customer review summary
- complete customer review
- Star rating
- Number of helpful votes
- Review date

Sample of extracted dataset is shown in below table 2.

[1] https://www.amazon.com/
[2] https://www.dickssportinggoods.com/

Table 2 : Customer reviews

Product Name	customer review summary	complete customer review	Star rating	Number of helpful votes	Review date
Nike Womens Long Sleeve Legend Shirt	Nike wins again.	Comfortable and fits well. I bought it late spring and have only worn it for a short run, but it breathed well. Would highly recommend it.	5		on July 27, 2017
Nike Pro Womens Training Tights	they look good!	These pants look nice. very comfortable the only thing i dont like is that it clearly shows sweat.	3	9 helpful votes	on June 4, 2016
Nike Mens Legend Short Sleeve Tee	Buy a Size Up	Does not fit as expected. XL is usually baggy on me across brands and styles, especially Nike. This is not. Its not supposed to be tight like UA or other DriFit items, its supposed to be a t-shirt, right? Either way, its cut weird and I dont like it.	2	1 helpful vote	on March 26, 2015
NIKE Mens Legend Dri-Fit Poly S/S Crew Top	Dri-fit Tee is worse than cotton t-shirts for long high intensity exercise	It may be designed for light exercise as I was wetter after my exercise session than when using cotton t-shirts. It lost its cooling effects when completely wet. I have discontinued using it, but may test it when the weather cools down for the winter.	2	1 helpful vote	on August 11, 2015

3.2 Product Aspects

There are various product types in this dataset. A few of them are, men's t-shirts, women t-shirts, men's crew top, running top, tank top, sports bra, shorts and pants. Even though, there are various product types, customer concern aspects are almost the same for all product regardless of the product type or brand. Hence pre-defined aspects which belong to the second level of aspect hierarchy will be considered in this research as in figure 4. These aspects were selected focusing on consumers (customers), brand owners and apparel manufactures. All of these aspects are relevant to any kind of apparels. Ozkan and Meric [61] identified that fabric types (materials) and comfortability are critical properties in athletic apparels. As per [62], fabric threads used to make the garment, size of the garment , design of the garments and body movement decide the pressure for human tissue. Freedom of movements depends on the fit of the apparel [63]. Hence material, size and fit of the sportswear apparel may affect the performance of the player. Brand owners should prioritize apparel quality if they care about increasing the level of

consumer satisfaction [64]. Aspect quality is relevant to any kind of apparel, and manufactures are highly responsible for the quality of the apparel. Color is a key aspect that falls under the identity and recognition of sportswear apparels [65]. Some consumers are willing to pay premium prices for high-performance products [65], while some consumers are highly concern about the price of the product. Due to the above reasons, these aspects are considered the most critical aspects of this research by focusing consumers, brand owners and apparel manufacture's points of view.

In addition to the above aspects, there are numerous reviews available without any specific aspects. Hence Aspect 'General' will be used to represent the reviews that do not discuss the above aspects.

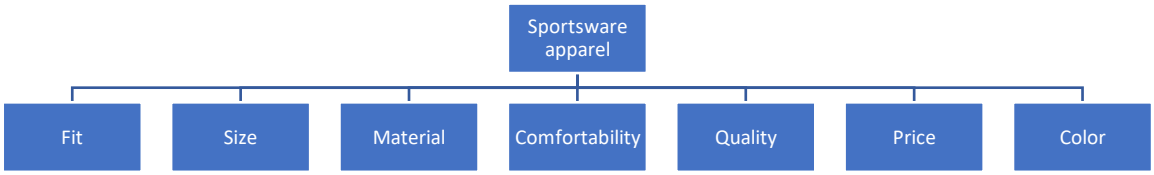

Figure 4: Apparel product aspects

Within a review, these aspects might discuss explicitly or implicitly. But implicit word usage is much higher compared to explicit words for almost all aspects.

- Fit :

  Fit represents how clothes are fitting to the body [51]. Words like fit, tight, loose use to express the fit of a cloth. Even though it has some relations to the aspect Size, these are completely two different aspects for clothing domain. This is one critical factor that need to focus when annotating apparel reviews for aspects like fit and size. Furthermore, there may be a few complications to identify the aspect 'fit' accurately. Some people use the word 'fit' to describe the 'size'. This can be easily identified by reading the full review and those should be considered as 'Size' regardless of the word 'fit'. Another scenario is, there's one material type

called ‘Dri-Fit’ and this should belong to the aspect ‘Material’ despite the word ‘Fit’.

- Size :

  Size represents how large or small the cloth to the body. Words like size, large, small, length, height or S, M, L, XL,…etc use to talk about the size of a cloth. The majority of reviews which talk about the aspect ‘Size’ has been used implicit words over explicit words.

- Material :

  Material represents the fabric which use to produce the cloth. Words like material, fabric, thickness, cotton, dri-fit, nylon…etc. use to express the feelings towards the aspect ‘Material.’

- Comfortability :

  This word represents how comfortable to wear the cloth. This is a relative measurement based on the customer’s expectation or previous cloths. Words like comfortable, comfy, feel good and domain specific words like stay cool, remove sweat, wicks moisture (this is a special property which sportswear apparel use to improve the comfortability)…etc. use to represent comfortability.

- Quality :

  This represent the build quality of the cloth. Various terms like quality, well made, poorly made…etc. use to express the quality of a cloth. Many quality-related complex implicit word combinations like ‘shrink in the dryer’, ‘wash and dry up beautifully’, ‘shrink easily’,...etc. make this a complex aspect to annotate. Only a few reviews talk about durability. But durability highly depends on the quality of the cloth. Hence durability has been considered under the aspect ‘Quality’ during the data annotation.

- Price :

  This word describes how expensive or cheap the cloth. Word like price, expensive, cheap, $, bargain, reasonable…etc. use to express the opinion of price.

- Color :

  This talks about the color of the cloth. Names of different colors and word color, mat, shine use to represent the color. Usage of explicit words are much higher for this word and hence easier to annotate.

- General :

  If a review does not discuss any of above aspects, then it considers as a general review and word ‘General’ will be used to classify those reviews

Table 3 shows the examples of explicit and implicit aspects for each aspect.

Table 3 : Explicit and Implicit aspects

Aspect	Explicit Review	Implicit Review
Fit	*I know that its just a T Shirt, but it fits nicely and is comfortable to wear.*	*A little snug but love it*
Size	*If for yoga, maby get a size bigger than usual.*	*I wear between a 8 or 10 in pants and the med.*
Material	*The fabric and quality are better than most.*	*Ordered Pro Fitted dry fit shirt , received dry fit shirt.*
Comfortability	*Lightweight, good fit, very comfortable.*	*Looks and feels great.*
Quality	*Very disappointed in the quality.*	*shrunk a bit when washed dried.*
Price	*The price is reasonable.*	*Upon research the boys size XL is the same as mens size small, but for half the cost.*
Color	*I love all the different color options that 90 Degree offers.*	*My other shirts are all mat, including the Dri Fit Nike.*

3.3 Data Preprocessing

Data preprocessing plays a vital role in aspect extraction and sentiment analysis. Different preprocessing techniques have been used in previous literature to improve the accuracies [16, 37]. But recent deep learning approaches [42, 52] have used a few data preprocessing steps like convert to lower-case, tokenize the corpus and convert 'entity#aspect' pairs to sentence-like structure. In this research, dataset creation motivated by SemEval-2016 Task 5 Subtask 1 [53]. Hence text level reviews were converted into sentence-level reviews. Below, figure 5 shows the data preprocessing steps employed in this study.

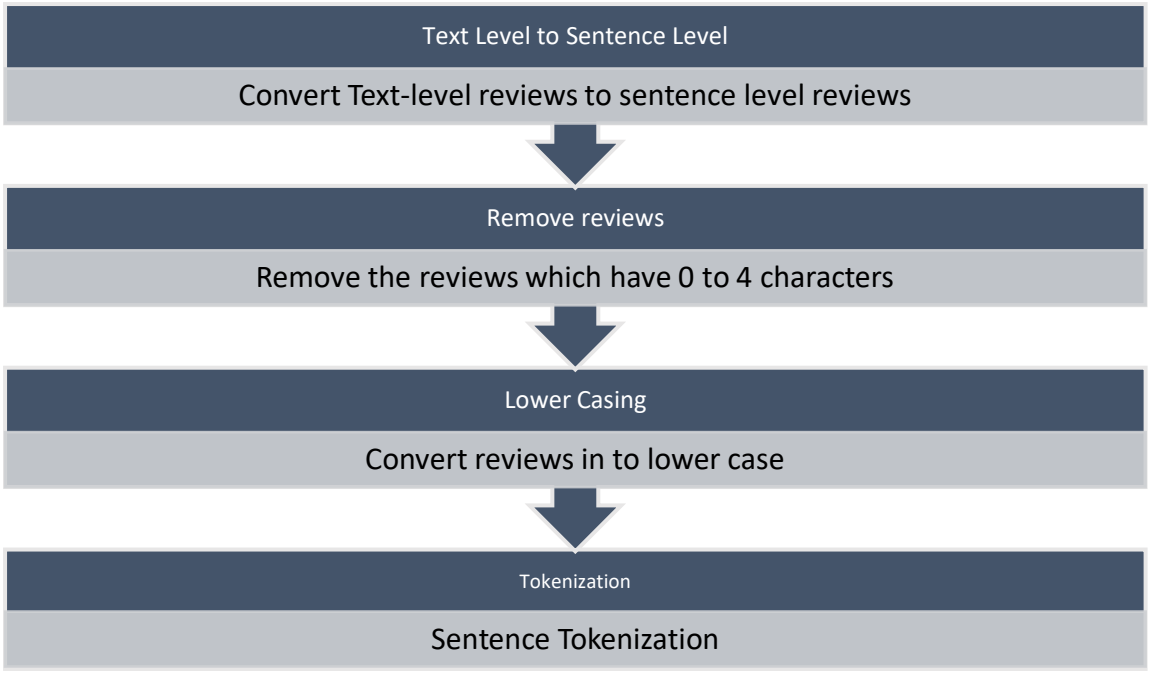

Figure 5: Data Preprocessing

Most of the reviews with 0 to 4 characters are just one word reviews like good, nice, ok, fine…etc. and does not include any meaningful aspect. They were removed from the final dataset. Next, these pre-processed customer reviews used for the data annotation.

3.4 Data Annotation

Data Annotation is a critical task for both supervised learning and performance evaluation. Sentence level manual data annotation carried out for this task. Given a sentence level customer review, the goal was to identify mentioned aspects in the review sentence. Review sentence may consist of one aspect, more than one aspect or no aspect at all. The author created an annotation manual with the support of an apparel domain expert, and then three annotators were used for the annotation task. In order to calculate inter-annotator agreement, randomly selected 500 review sentences in CSV format (Considering text level) were used. Three annotators annotated the 500 reviews independently with the support of the annotation manual and awareness session, which was conducted by the author. Evaluation of inter-annotator agreement will be discussed in the Chapter 4. Since, good agreement was achieved with this task, the other data set was equally distributed among annotators. Shown below figure 6 is a sample of annotated dataset.

Text Level Review	Sentence Level Review	Fit	Size	Material	Comfort ability	Quality	Price	Color	General
My son loved his sweater but it was a bit too fitted for his liking.. Looks very nice and he definitely wears it. I would suggest ordering a size up!	My son loved his sweater but it was a bit too fitted for his liking.. Looks very nice and he definitely wears it.	1							
	I would suggest ordering a size up		1						
Cannot go wrong with Nike products for the most part. Fit and finish is excellent. Also surprised how water resistant the pullover is. Spilled water on myself yesterday and literally bounced off me	Cannot go wrong with Nike products for the most part.								1
	Fit and finish is excellent.	1							
	Also surprised how water resistant the pullover is.					1			
	Spilled water on myself yesterday and literally bounced off me					1			
This is not Nike quality its fake !	This is not Nike quality its fake					1			
very comfortable, fits well, came on time. loving it so far	very comfortable, fits well, came on time.	1			1				
	loving it so far								1

Figure 6: Sample of Annotated dataset

3.5 Exploratory Data Analysis

After completing the data annotation task, exploratory data analysis was carried out to understand the nature of the dataset. Below table 4 shows the summary of selected data set.

Table 4 : Summary of dataset

Number of text level reviews	4234
Number of sentence level reviews	10032
Total word count	98316
Total unique word count	7604
Minimum sentence level word count	1
Maximum sentence level word count	91

There are 8 classes in the dataset, including the aspect ‘General’. Each aspect has a different distribution. Below, figure 7 shows the distribution of the aspect within the dataset.

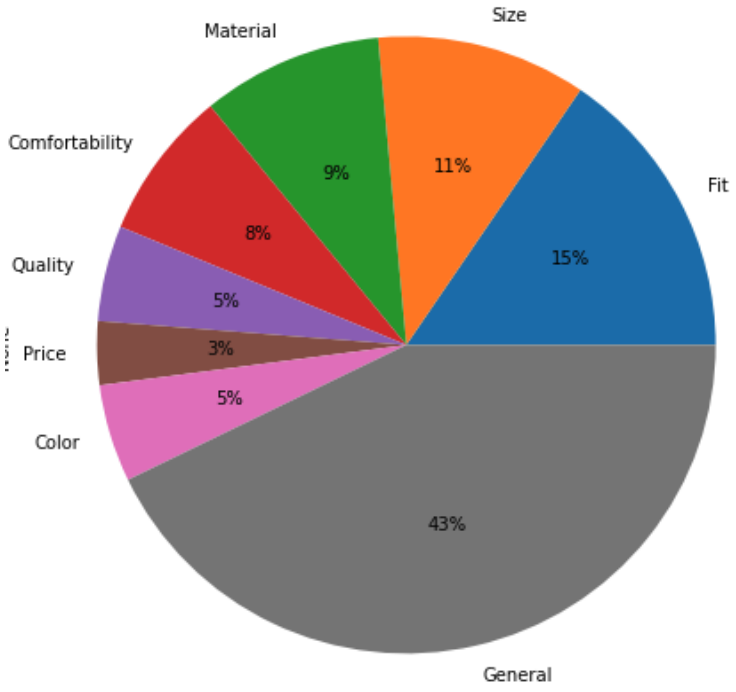

Figure 7: Aspect Distribution

As per the above figure 7, 43% of the review sentences do not mention any of the defined aspects. From the defined aspects, most customers discuss aspects Fit and Size while Price, Color and quality are among the least discussed aspects.

Below, figure 8 shows the distribution of the number of words in review sentences. There are 164 reviews with just one word and one review with 91 words. Most of the reviewers has used around 2 to 12 words per review sentence. Overall this follows a right-skewed distribution.

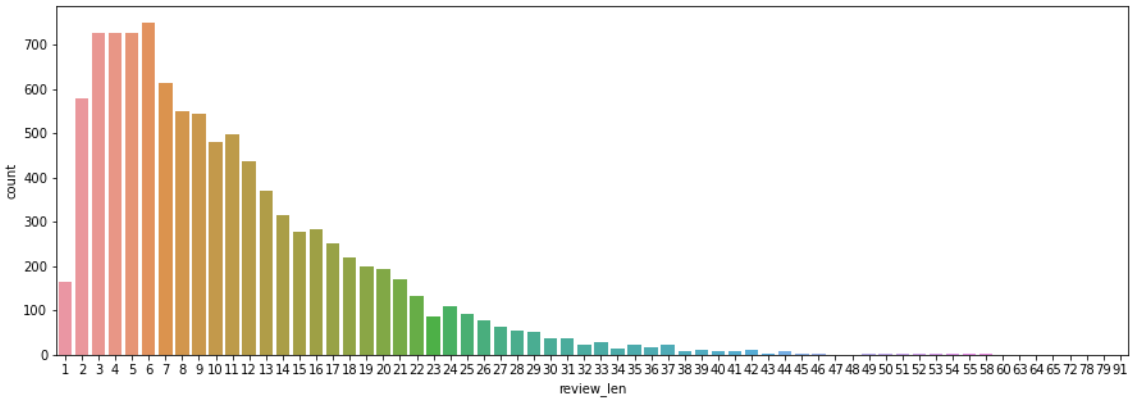

Figure 8: Number of words distribution

3.6 Classification Algorithms

Aspect extraction task can be considered as a classification problem. Furthermore, it can be modeled as a binary classification by considering individual aspects or multi-label classification by considering all aspects. This section discusses the classification algorithms and respective hyperparameters which were used to aspect extraction of sportswear apparel reviews. Out of 10032 sentence reviews, 8032 use for the training and validation purpose. The rest of 2000 sentence reviews will be used for the testing and performance evaluation. Google Colab with GPU hardware accelerator were used for the BERT and RoBERTa implementations. Comparison of the results of each approach will be discussed in the Chapter 4.

3.6.1 Convolution Neural Network (CNN)

S.Ruder et al.[42] considered the aspect extraction task of SemEval-2016 Task 5 as a multi-label classification problem. Since they have achieved promising results for their CNN model, this study employed the same model for sportswear apparel reviews.

Mini-batch size of 10, maximum sentence length of 100 tokens, the dropout rate of 0.5, and 100 filter maps with filter lengths 3, 4, and 5 considered as the hyperparameters of the model. This was trained for 15 epochs using mini-batch stochastic gradient descent, the adadelta update and early stopping. 300-dimensional GloVe vectors trained on 840B tokens were used for the word embeddings as the feature. Softmax activation function

was used to get the probabilities over 8 aspects and then the threshold of 0.2 observed for the highest F1 value. For the CNN implementation, Keras library with the Tensorflow backend was used.

3.6.2 BERT

M.J Hoang and O.A Bihorac [52] treated the aspect extraction task of SemEval-2016 dataset as a sentence pair classification task and achieved state-of-the-art performance. Hence it is considered as the baseline for Aspect extraction for sportswear apparel reviews. In their approach, review and aspect considered as the two sentences and prediction task was to, classify aspect is related to sentence or not using two labels, related and unrelated. Same approach used in this research by considering the review sentence and aspect as two sentences. 'aspect name' and 'not_aspect name' used as the class label, as an example for the aspect Fit, class labels are 'fit' and 'not_fit'. There are 24 BERT models are available for the english language. Out of them BERT-Base Uncased (12-layer, 768-hidden, 12-heads, 110M parameters) model was used in this implementation with 5e-5 as the learning rate and 4 Epochs. Minimum batch size (8) was used to avoid out of memory (OOM) situations. 'ktrain' was used for the implementation. ktrain is a lightweight wrapper for TensorFlow Keras which allows quick and easy implementation with only a few lines of codes.

In addition to considering ABSA task as a sentence pair classification task, this study model it as BERT single sentence classification and BERT multi-label classification for the performance comparison. 1 and 0 used as the class labels for both models. BERT multiclass classification gives probabilities for each class as the output. 0.5 selected as the optimum threshold as it gave the best results.

3.6.3 RoBERTa

Authors of [58] claimed that RoBERTa could improve the performance of every BERT model. Hence RoBERTa model was used considering this as a sentence pair classification task. For this, Roberta-base model which have 12-layer, 768-hidden, 12-

heads, 125M parameters, was used. Same parameters which were used in the BERT model, used with RoBERTa.

3.6.4 Ensemble Methods

Ensemble learning is creating multiple models and then combine them to achieve higher performance than individual models. There are various ensemble methods are available such as voting, stacking, bagging and boosting. Voting is considered as one of the simplest ensemble method which can be used for class labels. Two types of voting methods available are, majority (plurality) voting and weighted voting. For the majority voting, consider the predictions for each observation from all models and then take the majority vote (maximum occurrence) as the final prediction. As an example, three prediction models are A, B and C and respective predictions are 1, 1, and 0. In this case, the majority prediction is 1. Hence 1 is considered as the final prediction of the ensemble model. In the weighted voting approach, different weights assigned to different models. In other words, different model gets difference importance.

In this research, majority voting based ensemble model was developed considering the best three models which described above.

3.7 Evaluation Metrics

For the aspect extraction of SemEval-2016 task 5, F1 Score was used as the evaluation matric. Almost all previous researchers have also used the F1 score for the performance evaluation. In this research, precision, recall, f1-score, and confusion matrix was obtained for each aspect. But only F1 score will be used as the evaluation metric in the Chapter 4.

CHAPTER 4 : SYSTEM EVALUATION

Chapter 3 discussed the collected dataset, selected aspects, data annotation, data preprocessing steps used, exploratory data analysis carried out and the classification algorithms used for the aspect extraction tasks. In this chapter, evaluation results will be discussed for each classification algorithm.

4.1 Inter-Annotator Agreement

Cohen's kappa coefficient was used to calculate the inter-annotator agreement. This was calculated for 3 annotators using 500 sentence reviews. All annotators had the domain knowledge and given prior training for the annotation task. cohen_kappa_score which is available on python Scikit-learn library, was used to calculate the kappa coefficient for each annotator pairs. Below table 5 shows the calculated kappa coefficient for all annotator combinations.

Table 5 : Inter-Annotator Agreement

	Fit	Size	Material	Comfortability	Quality	Price	Color	General
A1 vs A2	0.85	0.83	0.68	0.87	0.55	0.95	0.87	0.79
A1 vs A3	0.91	0.77	0.81	0.78	0.72	0.89	0.96	0.85
A2 vs A3	0.80	0.70	0.69	0.79	0.56	0.89	0.90	0.82
Average	0.85	0.76	0.72	0.81	0.61	0.91	0.91	0.82

A1 – Annotator 1
A2 – Annotator 2
A3 – Annotator 3

As per Cohen, if the cohen kappa score is 0, then it indicates no agreement. If the value is in between 0.01 and 0.2, it indicates none to slight agreement. 0.21-0.40 indicates fair agreement between annotators. 0.41-0.60 is considered a moderate agreement and 0.61-08 is a substantial agreement. If the score is between 0.81 and 1.00, then it indicates the almost perfect agreement [60]. As per this definition, agreements for most aspects

among three annotators belong to a substantial or perfect agreement except two cases where agreement belongs to a moderate agreement. When considering the average agreements between all annotators, 5 aspects belong to perfect agreement while 3 belong to substantial agreement. The lowest score can be observed for the aspect of ‘Quality’. This may due to the large number of explicit aspects and uncertainty of reviews belong to the aspect ‘Quality’. For example, ‘*even better, they passed the squat test, so theyre not see through at least in the black color*’, this review can be considered in different angles. Someone can argue that, when considering the complete sentence, it discuss the aspect 'quality'. But others can argue that it does not discuss the quality specifically. This also highly depends on the domain knowledge of the annotators. Even if they have the required domain knowledge to annotate this, their level of domain knowledge may vary.

4.2 Evaluation Results

This section will discuss the results of each classification algorithm. Six models were used to extract 8 aspects including the aspect ‘General’. These models are

1. CNN : Convolution Neural Network
2. BERT SS : BERT Single Sentence Classification
3. BERT ML : BERT Multi Label Classification
4. BERT SP : BERT Sentence Pair Classification
5. RoBERTa SP : RoBERTa Sentence Pair Classification
6. Ensemble Model : Majority voting of BERT SS, BERT SP and RoBERTa SP

Below table 6 shows the results (F1 Score) obtained by above six models for each and every aspects and overall F1 for all aspects. Test set of randomly selected 2000 sentence reviews were used to obtain below F1 scores.

Table 6 : : Evaluation Results

Model	Fit F1	Size F1	Material F1	Comfortability F1	Quality F1	Price F1	Color F1	General F1	Overall F1
CNN	0.85	0.82	0.73	0.87	0.81	0.66	0.90	0.91	0.850
BERT SS	0.89	0.83	0.82	0.95	0.89	0.90	0.95	0.93	0.908
BERT ML	0.89	0.82	0.83	0.93	0.85	0.91	0.93	0.93	0.900
BERT SP	0.92	0.82	0.83	0.95	0.87	0.92	0.95	0.94	0.910
RoBERTa SP	0.93	0.84	0.82	0.97	0.90	0.91	0.96	0.93	**0.915**
Ensemble Model	0.93	0.83	0.84	0.96	0.91	0.92	0.96	0.94	**0.919**

BERT SP was considered as the baseline model as it was the state-of-the-art results obtained for SemEval-2016 ABSA task. As expected, this model given the best overall F1 score among CNN, BERT SS, BERT ML and BERT SP. CNN model which achieved first or second in 5 out of 11- language domains in SemEval-2016 task gave the lowest F1 for sportswear apparel review dataset.

As [58] highlighted, RoBERTa sentence pair classification model marginally outperforms the previous state-of-the-art results achieved by BERT sentence pair classification model. Out of five models, CNN, BERT SS, BERT ML, BERT SP, RoBERTa, best overall F1 achieved by RoBERTa, BERT SP and BERT SS respectively. These three models were used for the Ensemble model which is based on majority voting. Ensemble model outperformed the RoBERTa model by 0.4%.

These results show that RoBERTa and Ensemble models give the best results for the aspect extraction task of sportswear apparel review dataset.

4.3 Error Analysis

This section provides the error analysis for all aspects for BERT SS, BERT SP, RoBERTa and ensemble models.

4.3.1 Fit

Table 7 : Error Analysis - Fit

Review Sentence	Class	BERT SS	BERT SP	RoBERTa	Ensemble
No stars becaus the size did not fit and now I have to mail it back and pay.	0	1	1	1	1
For everyone wondering its size, its supposed to be slim fit.	0	1	1	1	1

As per above examples in table 7, all algorithms predicted the class as ‘Fit’. Both reviews talk about the size which is different from the aspect ‘Fit’. In this case the word ‘fit’ has caused for the misclassification.

4.3.2 Size

Table 8 : Error Analysis - Size

Review Sentence	Class	BERT SS	BERT SP	RoBERTa	Ensemble
I think it runs big a tiny bit, but it is not see through at all	1	0	0	1	0
Ill be giving it to my smaller friend.	0	1	0	1	1

For the above first example in table 8, word ‘big’ represents the size. However the word ‘big’ was not always used to represents the size of the cloth. Annotation was done considering the context of the sentence. Only the RoBERTa algorithm has classified it accurately. For the second example above, it talks about a size of a person and not about

the cloth. Hence it was annotated as class 0 and both BERT SS and RoBERTa algorithms has classified it incorrectly.

4.3.3 Material

Table 9 : Error Analysis - Material

Review Sentence	Class	BERT SS	BERT SP	RoBERTa	Ensemble
I ordered one size larger because they usually shrink after washing, so the shirt was slightly bigger for me.	1	0	0	1	0
These are NOT Nike Dri fit shirts.	0	1	1	1	1

This aspect has more implicit words. In the first example in table 9, complete sentence represents the aspect 'material'. 'Shrink' is a word to represent the characteristic of material. However, only RoBERTa has classified it accurately. For the second sentence, the word 'Dri fit' represents a special type of cloth. However, all algorithms have misclassified this sentence due to this specific word.

4.3.4 Comfortability

Table 10 : Error Analysis – Comfortability

Review Sentence	Class	BERT SS	BERT SP	RoBERTa	Ensemble
It fits very well and feels confortable as well.	1	0	1	0	0
Great with everything ,very comfotable	1	0	0	0	0

This is an aspect which has highest F1 score. However, spelling mistakes in the reviews have caused the misclassification. Spelling correction step in the data preprocessing stage will address this issue easily. Two examples are listed in table 10.

4.3.5 Quality

Table 11 : Error Analysis - Quality

Review Sentence	Class	BERT SS	BERT SP	RoBERTa	Ensemble
cheaply made	1	1	0	0	0
The material seems well made.	0	1	0	1	1

Aspect ‘Quality’ also has more implicit words. In the first sentence in table 11, cheaply represents the aspect ‘Quality’. However, this word is rare and does not have enough examples for algorithms to learn. Hence both BERT SP and RoBERTa have misclassified it as aspect ‘General’. In the second sentence, it talks about the material and not about the overall cloth. Word ‘well made’ caused BERT SS and RoBERTa for the misclassification.

4.3.6 Price

Table 12 : Error Analysis - Price

Review Sentence	Class	BERT SS	BERT SP	RoBERTa	Ensemble
100 for sweat pants is ridiculous but the quality is there.	1	0	0	0	0
the pants are too small and the material is too cheap.	0	1	0	1	1

In the first sentence of table 12, aspect ‘Price’ represent by the number 100. Even though this is easy to annotator, there’s no any information for algorithms to classify it as ‘Price’. In the second sentence, the word ‘cheap’ has been used to describe the material. It is not related to the ‘Price’. However, both BERT SS and RoBERTa have classified it as ‘Price’.

4.3.7 Color

Table 13 : Error Analysis - Color

Review Sentence	Class	BERT SS	BERT SP	RoBERTa	Ensemble
Only problem was that I ordered gold and got silver.	1	0	0	0	0
Its more of a ugly turquise.	1	1	0	0	0

Words like silver and gold were rare in reviews and it caused algorithms not to learn those words. In the second review of table 13, spelling mistake caused the misclassification.

4.3.8 General

Table 14 : Error Analysis - General

Review Sentence	Class	BERT SS	BERT SP	RoBERTa	Ensemble
Little be tieght .	0	1	1	1	1
Super light weight.	1	0	0	0	0

Due to the spelling mistake, first sentence of table 14 which belongs to the aspect ‘Size’ has been misclassified as ‘General’. For the second sentence, the word ‘light’ is associate with the aspect ‘Color’. Hence it has been classified by the algorithms as ‘Color’.

CHAPTER 5 : CONCLUSION

With the rapid growth of e-commerce, Customer reviews for products and services play a vital part in the e-commerce domain. Sentiment analysis helps to identify the overall opinion of customers towards a product or service. However, customers may express their opinion towards a specific aspect of the product without giving the opinion about overall product. Aspect based sentiment analysis helps to identify aspects and particular polarity towards the aspect. Hence aspect extraction is a pivotal task of aspect-based sentiment analysis. To the best of the author's knowledge, there's no previous research available for aspect-based sentiment analysis for the apparel (clothing) domain. In order to fill this research gap, this research proposes a model to extract aspects from sportswear apparel reviews.

Even though datasets are publicly available for aspect extraction tasks, there is no dataset available for the apparel domain. Most of the datasets are available for domains such as laptops, restaurants, cameras and phones. Hence, a new dataset was created using sportswear apparel reviews which were extracted from popular e-commerce websites. Dataset were manually annotated with the support of three annotators who has the domain knowledge. 7 key aspects, Fit, Size, Material, Comfortability, Quality, Price, and Color, were selected considering consumers, brand owners and manufacture perspectives.

Five deep learning approaches based on transformers were considered for the aspect extraction task. Then ensemble model was created using the best three models. Both RoBERTa and ensemble methods were outperformed the state-of-the-art model whereas RoBERTa model achieved the state-of-the-art results for aspect extraction task of sportswear apparel reviews.

Certain models gave better results for different aspects. RoBERTa model gave higher F1 score for both size and comfortability aspects than the ensemble method. This is due to all three individual classifiers give equal weights (priority) in majority voting approach.

RoBERTa model has learnt the context well than the BERT SS and BERT SP models. This caused RoBERTa model to reduce misclassifications compared to other models.

At the end of this research, the author was able to implement a deep learning model to extract aspects from sportswear apparel reviews. This study shows the effectiveness of transformer-based deep learning language models to extract the aspects from customer reviews. Moreover, this study has contributed a new dataset with 10032 sentence-level customer reviews, which can be utilized for future research in aspect extraction and sentiment analysis. Even though this study specifically used sportswear apparel reviews, this dataset can be considered as a general dataset to represent the apparel (clothing) domain as selected aspects and nature of reviews are the same for almost all clothes regardless of the specific function or brand of the cloth. Moreover, this study focus on the sentence level aspect extraction task. However, annotation task was carried out in a way that the dataset can be used for both aspect level and text level aspect extraction tasks.

5.1 Future Improvements

Below list presents the potential future enhancements and improvements to this research problem.

- Further data preprocessing steps such as spelling correction, removal of stop words, stemming, lemmatization and combination of them.
- Use weighted voting ensemble classifier to improve the F1 score of ensemble model.
- Hyperparameter optimization using a strategy such as random search, grid search or bayesian optimization. However, this will be computationally expensive.
- Instead of using only the aspect name as the second sentence of the sentence pair classification task, create a separate sentence which consist of more complex implicit words to reduce to false negative (to capture the missed aspects).
- Extend the study to identify sentence polarity. Clothing reviews have much more implicit aspects compared to explicit aspects. Some of these implicit words are

represents the polarity as well. Since some of these aspects are quite different from domains like laptop, restaurants, previous approaches might not give the best results.

References

[1] "What motivates you to purchase new items specifically for sport?" Internet: https://www.statista.com/statistics/630020/consumer-motivation-to-buy-sports-apparel-sports-shoes [Nov. 12,2017].

[2]. "Online Apparel Market: Global Industry Analysis and Forecast 2016 – 2026", Internet: https://www.persistencemarketresearch.com/market-research/online-apparel-market.asp, [Nov.15,2017]

[3]. G. Charlton. "Ecommerce consumer reviews: why you need them and how to use them", Internet: https://econsultancy.com/blog/9366-ecommerce-consumer-reviews-why-you-need-them-and-how-to-use-them, March. 20,2012 [Dec 2,2017]

[4] J. Rampton. "The Relationship Between Customer Reviews and Ecommerce Sales", Internet: https://www.shopify.com/content/the-relationship-between-customer-reviews-and-ecommerce-sales, Jan. 9,2017 [Nov. 12,2017]

[5] Y. Bao, H. Xu, F. Jia, and X. Bai, "Aspect-based sentiment analysis using abpcs model and svmp perf in chinese reviews," in *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 3208–3215, IEEE,2017 .

[6] T. Chinsha and S. Joseph, "A syntactic approach for aspect based opinion mining," in *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015)*, pp. 24–31, IEEE, 2015.

[7] R. Panchendrarajan, N. Ahamed, B. Murugaiah, P. Sivakumar,S. Ranathunga, and A. Pemasiri, "Implicit aspect detection in restaurant reviews using cooccurence of words," in *Proceedings of the 7th Workshopon Computational Approaches to Subjectivity, Sentiment and Social MediaAnalysis*, pp. 128–136, 2016

[8] T. Hercig, T. Brychc´ın, L. Svoboda, and M. Konkol, "Uwb at semeval-2016 task 5: Aspect based sentiment analysis," in *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pp. 342–349, 2016.

[9] S. Poria, E. Cambria, L.-W. Ku, C. Gui, and A. Gelbukh, “A rule-based approach to aspect extraction from product reviews,” in *Proceedings ofthe second workshop on natural language processing for social media(SocialNLP)*, pp. 28–37, 2014 .

[10] B. Liu (2012, April 22) “*Sentiment Analysis and Opinion Mining*” *Synthesis lectures on human language technologies* , vol. 5, no 1, pp 1-167, 2012.

[11]. I. Jayasekara and W. Wijayanayake, “Opinion mining of customer re-views: Feature and smiley based approach,”*International journal of Datamining & Knowledge Management process (IJKDP)*, vol. 6, no. 1, pp. 1–11, 2016 .

[12]. S. Moghaddam & M. Ester, “Aspect-based Opinion Mining from Online Reviews”, In *Tutorial at SIGIR Conference*, 2012.

[13]. T. A. Rana and Y.-N. Cheah, “Aspect extraction in sentiment analysis: comparative analysis and survey,”*Artificial Intelligence Review*, vol. 46,no. 4, pp. 459–483, 2016

[14]. A. Bagheri, M. Saraee, and F. de Jong, “An unsupervised aspect detection model for sentiment analysis of reviews,” in *International conference on application of natural language to information systems*, pp. 140–151,Springer, 2013 .

[15]. M. Hu and B. Liu, “Mining and summarizing customer reviews,” in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 168–177, 2004.

[16]. K. Bafna and D. Toshniwal, “Feature based summarization of customers’reviews of online products,”*Procedia Computer Science*, vol. 22,pp. 142–151, 2013 .

[17]. L. Zhang and B. Liu, “Aspect and entity extraction for opinion mining,”inData mining and knowledge discovery for big data, pp. 1–40,Springer, 2014. .

[18]. A. Alawami, *Aspect extraction for sentiment analysis in Arabic dialect*. PhD thesis, University of Pittsburgh, 2017 .

[19]. D. Waegel. “A Survey of Bootstrapping Techniques in Natural Language Processing” [online]. Available: https://www.eecis.udel.edu/~vijay/fall13/snlp/lit-survey/Bootstrapping.pdf

[20]. J. Zhu, H. Wang, M. Zhu, B. K. Tsou, and M. Ma, “Aspect-based opinion polling from customer reviews,”*IEEE Transactions on affective computing,* vol. 2, no. 1, pp. 37–49, 2011 .

[21]. K. Liu, L. Xu, and J. Zhao, “Opinion target extraction using word-based translation model,” in *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pp. 1346–1356, 2012. .

[22]. W. Bancken, D. Alfarone, and J. Davis, “Automatically detecting and rating product aspects from textual customer reviews,” in *Proceedings of the 1st international workshop on interactions between data mining and natural language processing at ECML/PKDD,* vol. 1202, pp. 1–16,2014 .

[23]. Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai, “Topic sentiment mixture: modeling facets and opinions in weblogs,” in *Proceedings ofthe 16th international conference on World Wide Web,* pp. 171–180,2007 .

[24]. S. Brody and N. Elhadad, “An unsupervised aspect-sentiment model for online reviews,” in *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, pp. 804–812, 2010 .

[25]. X. Zhao, J. Jiang, H. Yan, and X. Li, “Jointly modeling aspects andopinions with a maxent-lda hybrid,” ACL, 2010 .

[26]. Q. Liu, B. Liu, Y. Zhang, D. S. Kim, and Z. Gao, “Improving opinion aspect extraction using semantic similarity and aspect associations,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30,2016 .

[27]. A. Mukherjee and B. Liu, “Aspect extraction through semi-supervised modeling,” in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 339–348,2012 .

[28]. B. Wang and H. Wang, “Bootstrapping both product features andopinion words from chinese customer reviews with cross-inducing,”in *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*, 2008 .

[29]. Q. Zhao, H. Wang, P. Lv, and C. Zhang, “A bootstrapping based refinement framework for mining opinion words and targets,” in *Proceedingsof the 23rd ACM international conference on conference on information and knowledge management*, pp. 1995–1998, 2014. .

[30]. G. Qiu, B. Liu, J. Bu, and C. Chen, “Opinion word expansion and tar-get extraction through double propagation,”*Computational linguistics*,vol. 37, no. 1, pp. 9–27, 2011 .

[31]. L. Zhang, B. Liu, S. H. Lim, and E. O’Brien-Strain, “Extracting and ranking product features in opinion documents,” in *Coling 2010: Posters*,pp. 1462–1470, 2010 .

[32]. Y. Wu, Q. Zhang, X.-J. Huang, and L. Wu, “Phrase dependency parsing for opinion mining,” in *Proceedings of the 2009 conference on empirical methods in natural language processing*, pp. 1533–1541, 2009. .

[33]. J. Yu, Z.-J. Zha, M. Wang, and T.-S. Chua, “Aspect ranking: identifying important product aspects from online consumer reviews,” in *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pp. 1496–1505, 2011 .

[34]. C.-P. Wei, Y.-M. Chen, C.-S. Yang, and C. C. Yang, “Understanding what concerns consumers: a semantic approach to product feature extraction from

consumer reviews," *Information Systems and E-Business Management*, vol. 8, no. 2, pp. 149–167, 2010 .

[35]. B. Ma, D. Zhang, Z. Yan, and T. Kim, "An lda and synonym lexicon based approach to product feature extraction from online consumer product reviews,"*Journal of Electronic Commerce Research*, vol. 14,no. 4, p. 304, 2013 .

[36]. K. Liu, H. L. Xu, Y. Liu, and J. Zhao, "Opinion target extraction using partially-supervised word alignment model.," *inIJCAI*, vol. 13,pp. 2134–2140, 2013.

[37]. L. Xu, K. Liu, S. Lai, Y. Chen, and J. Zhao, "Mining opinion wordsand opinion targets in a two-stage framework," inProceedings of the51st Annual Meeting of the Association for Computational Linguistics(Volume 1: Long Papers), pp. 1764–1773, 2013 .

[38]. . Toh and J. Su, "Nlangp: Supervised machine learning system for aspect category classification and opinion target extraction," in *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval2015)*, pp. 496–501, 2015. .

[39]. F. Li, C. Han, M. Huang, X. Zhu, Y. Xia, S. Zhang, and H. Yu,"Structure-aware review mining and summarization," in *Proceedings of the 23rd International Conference on Computational Linguistics (Coling2010)*, pp. 653–661, 2010. .

[40]. N. Jakob and I. Gurevych, "Extracting opinion targets in a single and cross-domain setting with conditional random fields," in *Proceedings of the 2010 conference on empirical methods in natural language processing*, pp. 1035–1045, 2010.

[41]. S. Poria, E. Cambria, and A. Gelbukh, "Aspect extraction for opinion mining with a deep convolutional neural network," *Knowledge-Based Systems*, vol. 108, pp. 42–49, 2016 .

[42]. S. Ruder, P. Ghaffari, and J. G. Breslin, “Insight-1 at semeval-2016 task5: Deep learning for multilingual aspect-based sentiment analysis,”*arXiv preprint arXiv*:1609.02748, 2016 .

[43]. B. Wang and M. Liu, “Deep learning for aspect-based sentiment analysis,”*Stanford University report*, 2015. .

[44]. I. Cruz, A. F. Gelbukh, and G. Sidorov, “Implicit aspect indicator extraction for aspect based opinion mining.,”*Int. J. Comput. Linguistics Appl.*, vol. 5, no. 2, pp. 135–152, 2014 .

[45]. K. Schouten and F. Frasincar, “Finding implicit features in consumer reviews for sentiment analysis,” in *International Conference on Web Engineering*, pp. 130–144, Springer, 2014. .

[46]. L. Dey and S. M. Haque, “Opinion mining from noisy text data,”*International Journal on Document Analysis and Recognition (IJDAR)*,vol. 12, no. 3, pp. 205–226, 2009.

[47]. J. Brownlee. “A Gentle Introduction to the Bag-of-Words Model”, Internet: https://machinelearningmastery.com/gentle-introduction-bag-words-model/ , Oct. 09,2017 [Jan. 04, 2018].

[48]. T. Chinsha and S. Joseph, “A syntactic approach for aspect based opinion mining,” in *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015)*, pp. 24–31, IEEE, 2015 .

[49]. F. Enr´ıquez, J. A. Troyano, and T. L´opez-Solaz, “An approach to the useof word embeddings in an opinion classification task,”*Expert Systemswith Applications*, vol. 66, pp. 1–6, 2016 .

[50]. W. Jin, H. H. Ho, and R. K. Srihari, “A novel lexicalized hmm-based learning framework for web opinion mining,” in *Proceedings of the 26thannual international conference on machine learning*, vol. 10, Citeseer,2009 .

[51]. “Stitch fix”, Internet: https://www.stitchfix.com/women/blog/ask-a-stylist/difference-between-fit-and-style [Nov.05,2020].

[52] M. Hoang, O. A. Bihorac, and J. Rouces, “Aspect-based sentiment analysis using bert,” in *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pp. 187–196, 2019 .

[53] M. Pontiki, D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Man-andhar, M. Al-Smadi, M. Al-Ayyoub, Y. Zhao, B. Qin, O. De Clercq,et al., “Semeval-2016 task 5: Aspect based sentiment analysis,” in *International workshop on semantic evaluation*, pp. 19–30, 2016 .

[54] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez,L. Kaiser, and I. Polosukhin, “Attention is all you need,”*arXiv preprint arXiv*:1706.03762, 2017 .

[55] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,”*arXiv preprint arXiv*:1810.04805, 2018. .

[56] C. Sun, L. Huang, and X. Qiu, “Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence,”*arXiv preprintarXiv*:1903.09588, 2019. .

[57] H. Xu, B. Liu, L. Shu, and P. S. Yu, “Bert post-training for review reading comprehension and aspect-based sentiment analysis,”*arXivpreprint arXiv*:1904.02232, 2019 .

[58] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis,L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized bertpretraining approach,”*arXiv preprint arXiv*:1907.11692, 2019. .

[59] H. Tian and M. White, “A Pipeline of Aspect Detection and Sentiment Analysis for E-Commerce Customer Reviews”, *In Proceedings of ACM SIGIR Workshop on eCommerce (SIGIR eCom’20)*, 2020.

[60] M. L. McHugh, “Interrater reliability: the kappa statistic,”*Biochemia medica*, vol. 22, no. 3, pp. 276–282, 2012

[61] E. T. Ozkan and B. Meric, “Thermo physiological comfort properties of different knitted fabrics used in cycling clothes,” *Textile Research Journal*, vol. 85, no. 1, pp. 62–70, 2015

[62] R. Rossi, “High-performance sportswear,” *High-Performance Apparel*, pp. 341–356, 2018

[63] . Manshahia and A. Das, “High active sportswear–a critical review,”2014

[64] C. Saricam, A. Aksoy, and F. Kalaoglu, “Determination of the prior-ities of customer requirements and quality in apparel retail industry,” *International Journal of Business and Social Science*, vol. 3, no. 16,2012

[65] Y. E. Elmogahzy, “13 - performance characteristics of traditional textiles: Denim and sportswear products,” in Engineering Textiles (Second Edition)(Y. E. Elmogahzy, ed.), *The Textile Institute Book Series*, pp. 319–346, Woodhead Publishing, second edition ed., 2020.

Appendix

Precision and recall values of different models for all aspects

Model		Fit	Size	Material	Comfortability	Quality	Price	Color	General	Overall F1
CNN	Precision	0.84	0.80	0.70	0.83	0.80	0.60	0.89	0.90	0.82
	Recall	0.86	0.84	0.76	0.90	0.82	0.69	0.91	0.91	0.87
BERT SS	Precision	0.86	0.82	0.80	0.93	0.88	0.98	0.94	0.92	0.89
	Recall	0.91	0.84	0.83	0.95	0.89	0.91	0.96	0.94	0.90
BERT ML	Precision	0.82	0.82	0.82	0.92	0.83	0.90	0.91	0.89	0.90
	Recall	0.98	0.82	0.84	0.95	0.91	0.94	0.95	0.97	0.91
BERT SP	Precision	0.88	0.77	0.78	0.92	0.81	0.90	0.92	0.96	0.94
	Recall	0.97	0.88	0.89	0.98	0.94	0.95	0.98	0.91	0.90
RoBERTa SP	Precision	0.87	0.95	0.77	0.97	0.92	0.95	0.95	0.97	0.93
	Recall	0.99	0.76	0.88	0.96	0.89	0.87	0.97	0.90	0.91
Ensemble Model	Precision	0.88	0.92	0.80	0.95	0.89	0.88	0.98	0.92	0.90
	Recall	0.98	0.76	0.88	0.97	0.93	0.95	0.94	0.97	0.94