Journal of New Media

Tech Science Press

DOI: 10.32604/jnm.2023.045792

ARTICLE

Topic Modelling and Sentiment Analysis on YouTube Sustainable Fashion Comments

Hsu-Hua Lee and Minh T. N. Nguyen*

Department of Management Sciences, College of Business and Management, Tamkang University, New Taipei City, 251201, Taiwan

*Corresponding Author: Minh T. N. Nguyen. Email: nguyenthinhatminh1234@gmail.com

Received: 07 September 2023 Accepted: 17 November 2023 Published: 27 December 2023

ABSTRACT

YouTube videos on sustainable fashion enable the public to gain basic knowledge about this concept. In this paper, we analyse user comments on YouTube videos that contain sustainable fashion content. The paper's main objective is to help content creators and business managers effectively understand the perspectives of viewers, thus improving video quality and developing business. We analysed a dataset of 17,357 comments collected from 15 sustainable fashion YouTube videos. First, we use Latent Dirichlet Allocation (LDA), a topic modelling technique, to discover the abstract topics. In addition, we use two approaches to rank these topics: ranking based on proportion and Rank-1 method. Second, we apply sentiment analysis to identify the user's emotional tone in the comments. As a result, 14 topics were identified. The most common positive and negative scores are 1 and −1, respectively. In total, there are 28.42% positive comments, 22.35% negative comments and 49.23% neutral comments.

KEYWORDS

Topic modelling; sentiment analysis; latent dirichlet allocation; natural language processing; sustainable fashion; YouTube comments

1 Introduction

Sustainable fashion, or eco-fashion, is a movement in the fashion industry focused on sustaining long-term growth while causing less environmental damage. Customers are increasingly conscious of the Earth's ecosystem, and the fashion industry responded by considering sustainability in every decision [1]. When consumers decide about an innovation, they first gather information, and sustainable fashion has been a trending topic in recent years. According to the Google Trends app [2], from January 2014 to May 2023, the interest-over-time value of the term 'sustainable fashion' term increased significantly. In January 2014, there were 17 searches for sustainable fashion for every 100 searches on Google. The number continued to increase and peaked in April 2022 at 100 searches. It is undoubtable that people are interested in obtaining knowledge about sustainable fashion.

One of the most trafficked online sources of information is YouTube. In 2022, YouTube was the largest online video-sharing platform, with over 2.56 billion monthly active users [3]. YouTube videos are uploaded by companies and customers and include a broad range of content. Due to its availability and affordability, people use YouTube as a source of information to answer their questions.

YouTube had approximately 74.8 billion monthly visits in November 2022, which made it the second-most-visited website worldwide after Google [4]. One of YouTube's characteristics is allowing user-generated content, which is original content created by customers. In addition, YouTube encourages interaction between content creators and content consumers. The comments section under videos with the 'reply' function and other features, such as the 'like'/'dislike' button and video sharing, make it easier for YouTubers (content creators) and viewers (content consumers) to communicate. Viewers leave their opinions in the comments, which can help YouTubers improve their channel's quality. YouTube comments also provide information that business managers could use for business development.

However, reading and collecting all the comments is time-consuming. In addition, YouTubers also must analyse the comments to understand their underlying meaning. Instead of performing these steps manually, we can use natural language processing (NLP) to process these massive amounts of data. NLP is the branch of artificial intelligence (AI) that gives computers the ability to understand text and spoken words in much the same way human beings can [5]. At first, computational linguistics and statistical/machine learning and deep learning models are combined. Then, human language will be interpreted in the form of text or voice data. Next, computers can understand the data's meaning with the speaker's sentiment. In NLP, topic modelling and sentiment analysis are two techniques that can be performed in R to summarise text and mine opinion.

Most of previous researches have focused on either content creator's opinions or content consumer's comments which are relevant to sustainable fashion. Moreover, there are few studies in the literature that combine topic modelling using LDA and sentiment analysis using SentiStrength to read and summarise the comments of YouTube videos on sustainable fashion. Therefore, this paper aims to apply natural language processing methods, such as topic modelling and sentiment analysis, to describe and classify the comments and their sentiment. This paper is structured as follows: Section 2 reviews related work on sustainable fashion, topic modelling and sentiment analysis in short texts on the social web. Section 3 applies the approaches–Latent Dirichlet Allocation (LDA) and sentiment analysis using SentiStrength software–to the data. Section 4 presents the results, and Section 5 presents the conclusion and discusses future research.

2 Literature Review

In recent years, there has been a considerable amount of literature on sustainable fashion that analyses social networking services. Greco investigated activists' Twitter protest against the clothing brand Primark in the public debate about sustainable fashion [6]. Blasi et al. [7] studied consumer perception of fashion and eco-friendliness by collecting data from fashion brands and their followers on Twitter. Testa et al. [8] explored the reasons for sustainable fashion's popularity by analysing 25 fashion retailers' Instagram posts. As one of the most widely used social networking services, YouTube was also selected for research on sustainable fashion. Kim et al. [9] studied the purchase intention of sustainable fashion products by surveying 230 YouTube users, while Haines et al. [10] examined YouTube comments related to sustainable fashion in fashion haul videos. However, researchers have tended to focus on content creators or videos where the topic is not exclusively about sustainable fashion. Therefore, the information related to sustainable fashion on YouTube should be studied.

YouTube comments have been extensively studied in NLP with different approaches. Poché et al. [11] used Naïve Bayes and Support Vector Machines (SVM) to classify user comments on coding tutorial YouTube videos. Lee et al. [12] identified YouTube's role in online self-directed learning by applying extraction software to educational videos' comments. Interestingly, various studies used

topic modelling and sentiment analysis techniques in their work, such as Alshamrani et al. [13]. They applied topic modelling to news videos' comments to investigate the relationship between various toxic behaviours and news subjects. Bhuiyan et al. [14] performed an NLP-based methodology of sentiment analysis on YouTube users' comments to identify the most relevant and popular videos.

Topic modelling is used to identify topics from massive unstructured data, which is inefficient when done manually. Topic modelling establishes topics and provides meaningful insights from a collection of documents. There are two main topic modelling methods: Latent semantic analysis (LSA) and LDA. Compared to LSA, LDA has been more thoroughly studied, especially in analysing social data short text such as Douban online reviews [15], tweets [16], Facebook comments [17] and others. Albalwi et al. [18] recommended using LDA to obtain meaningful topics and good results with short text data. Therefore, we used LDA to detect topics discussed in the comment sections.

Another commonly used technique in NLP is sentiment analysis. Sentiment analysis, or opinion mining, is the process of finding the emotion behind a text or speech. It provides an estimation of the content, which can be in binary (positive/negative) or trinary (positive/neutral/negative) [19]. Among the sentiment strength detection software available, SentiStrength was considered suitable for the short informal text of an SNS post [20]. For example, Thelwall et al. [21] classified YouTube comments using SentiStrength to identify patterns and recommend future research. Vilares et al. [22] used the software to analyse Spanish tweets that referred to Spanish politicians and parties. Given that SentiStrength software is highly recommended for performing sentiment analysis for short texts like YouTube comments, we decided to apply it in this study. Table 1 shows the detailed comparison between previous works and our research.

Table 1: Summary of the related work and our research

Related work	Dataset	Methodology
Greco [6]	Twitter tweets about sustainable fashion, brand's website contents	Manual
Blasi et al. [7]	Twitter tweets about fashion brands	Ordinary Least Squares regression
Testa et al. [8]	Instagram post from fashion retailers	Qualitative coding, quantitative insights
Kim et al. [9]	YouTube users survey about sustainable fashion	Regression
Haines et al. [10]	YouTube comments relevant to sustainable fashion on fashion haul videos	Leximancer, Linguistic Inquiry and Word Count
Poché et al. [11]	YouTube comments on coding videos	Naïve Bayes, Support Vector Machines
Lee et al. [12]	YouTube comments on educational videos	Sentiment Analysis, Qualitative content analysis
Alshamrani et al. [13]	YouTube comments on news videos	Deep Neural Network (DNN)-based Architecture, Dataset Handling and Splitting Topic Modelling using LDA

(Continued)

Table 1 (continued)

Related work	Dataset	Methodology
Bhuiyan et al. [14]	YouTube comments	Sentiment Analysis using SentiStrength
Our approach	YouTube comments on sustainable fashion videos	Topic Modelling using LDA, Sentiment Analysis using SentiStrength, Term Frequency-Inverse Document Frequency algorithm

3 Methods

3.1 Data Collection

We built our data by collecting YouTube comments from 15 videos about sustainable fashion. Before scraping comments, the computer history and cookies were deleted. On March 18, 2023, we used the search bar on YouTube with the keywords 'sustainable fashion'. We then used the 'sort by relevance' filter to get the most relevant videos. The chosen videos had to be in English, informative, and have good audio-visual quality. The video content must

- Explain the concept of sustainable fashion;
- Mention the benefits of sustainable fashion;
- Discuss how to apply sustainable fashion in daily life.

We reviewed the search results and chose the 15 videos that met the criteria and had the most views which were shown in Table 2.

Table 2: Video information, including the release date, the number of views and comments

#	Date	Length (mins)	Views	Comments
1	Nov 25, 2019	29:01	5,262,218	6176
2	Feb 12, 2022	42:26	2,974,208	1830
3	Jun 11, 2022	4:26	1,480,123	315
4	Nov 30, 2018	6:50	989,749	538
5	Dec 16, 2020	3:52	915,795	774
6	Jan 8, 2021	12:34	553,749	419
7	Aug 5, 2020	13:50	521,482	3106
8	Jan 17, 2021	16:15	378,629	1679
9	Oct 24, 2021	9:18	378,504	1297
10	Mar 9, 2019	17:07	281,826	499
11	Jun 15, 2016	17:43	195,706	98
12	Feb 17, 2019	2:52	176,882	219
13	May 16, 2022	11:27	174,790	22
14	Sep 4, 2016	6:16	134,338	276
15	Nov 20, 2019	2:56	81,738	109

We collected the comments using the YouTube Data API and Stevesie. YouTube Data API allows the user to access the comments on YouTube, which then can be scraped by Stevesie, the comment extractor. A total of 17,357 comments were collected and stored in a CSV file. Fig. 1 shows the scraping process.

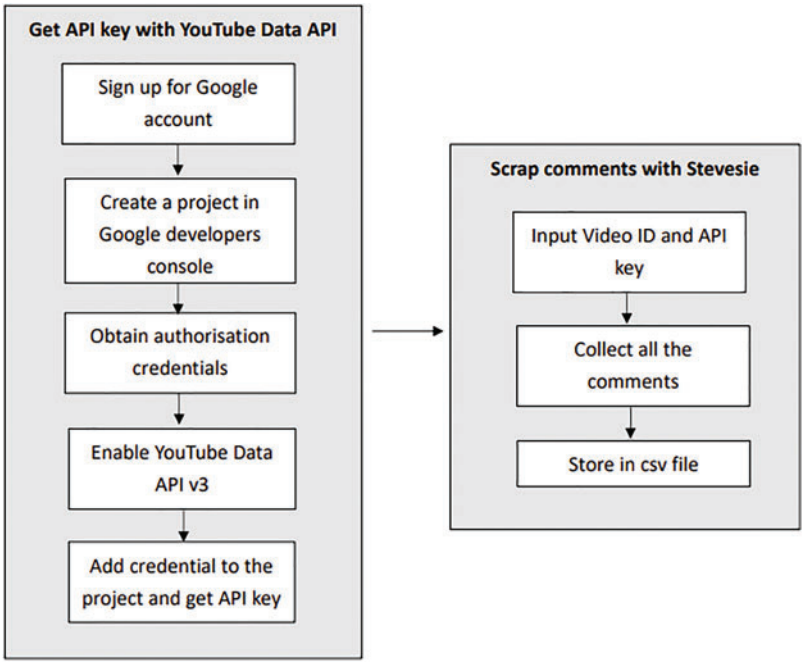

Figure 1: The YouTube comment scraping process

3.2 Text Pre-Processing

While it was easy for humans to understand the meaning of each sentence, a machine would have difficulty understanding the sentences in their raw forms. Potential noise, such as emotions, punctuation and words in different tenses, needed to be cleaned from the text data. We used various text pre-processing strategies, including:

- Expanding contractions;
- Removing punctuations, special characters, non-ASCII characters and numbers; Expanding contractions;
- Lowercasing all texts;
- Removing stop words and additional words;
- Stemming (removing prefixes and suffixes).

These steps were executed by the *tm* package in R. In this study, we only analysed the English comments; therefore, non-ASCII characters, which are limited to 128 characters, were removed. Similarly, punctuation, special characters and numbers were removed. Stop words are words that appear frequently but provide no or little information, such as ‘the’, ‘I’ and ‘it’. The *tm* package in R consists of 174 common English stop words, which can be applied to the data.

We also used the stemming technique in this study. In a document, there are often different forms of one base word, for example, ‘year’ and ‘years’. Stemming helps us to identify the base word

and treat these different forms as the stemmed form. As a result, the feature space of the text data reduced dramatically, and the machine learning model's performance improved. It increased accuracy in finding the most frequent words in the next step. Table 3 shows an example of a sentence before and after pre-processing.

Table 3: Example of before and after pre-processing of a comment

Before pre-processing	After pre-processing
My carbon footprint in terms of clothes ain't that bad, buying clothes in my country is EXPENSIVE so I tent to use my clothes for a longer time.	*carbon/footprint/term/cloth/not/bad/buy/cloth/country/expensive/tend/use/cloth/longer/time*

3.3 Calculating Term Frequency-Inverse Document Frequency (tf-idf)

In information retrieval and machine learning, the *tf-idf* technique is used to determine the importance of words in a document amongst a collection of documents (a corpus). The *tf* in *tf-idf* stands for term frequency, which is how often a specific word occurs in a document:

$$tf\ (t, d) = \frac{number\ of\ occurence\ of\ t\ in\ d}{number\ of\ total\ words\ in\ f} \tag{1}$$

where t is the specific term, and d is the document in a corpus.

On the other hand, *idf* stands for inverse document frequency. The *df* calculates the occurrence of a word in multiple documents, not only in one document. Therefore, *idf*, which is an inverse of the *df*, can be identified with this formula:

$$idf\ (t, D) = \log\left(\frac{N}{number\ of\ documents\ contain\ t}\right) \tag{2}$$

where N is the number of documents d in the corpus D. A high *idf* value implies uncommon words in all documents, following the increase in importance. By multiplying *tf* and *idf*, we have the final *tf-idf* value:

$$tf - idf = tf\ (t, d) * idf\ (t, D) \tag{3}$$

The combination of two measures, *tf* and *idf*, shows how often a word occurs in a document and how rare that word is across all documents. When the *tf-idf* score of a term approaches 0, that term is considered less relevant than others. The higher the score, the more important the term.

3.4 Topic Modelling with LDA

As mentioned previously, topic modelling is a common procedure used to discover the abstract topics in a collection of documents. There are two main topic modelling methods, LSA and LDA. LDA was developed in 2003 by Blei et al. [23]. The method tells us what topics are present in any given document by observing all the words and producing a topic distribution. For example, assume there are three topics A, B and C. Each topic is expressed by a list of words with probabilities for them to belong to each topic.

Assuming that any document can be described by a recipe consisting of a topic and how much of it the document should contain–for instance, 50% topic A, 30% topic B and 20% topic C–the model

then generates such a document by taking the right number of words from the specified topics and mixing them. In reality, given how many topics it should make, LDA takes a collection of documents and divides words into different groups. Finally, we name the topics based on those words. Fig. 2 illustrates the LDA model.

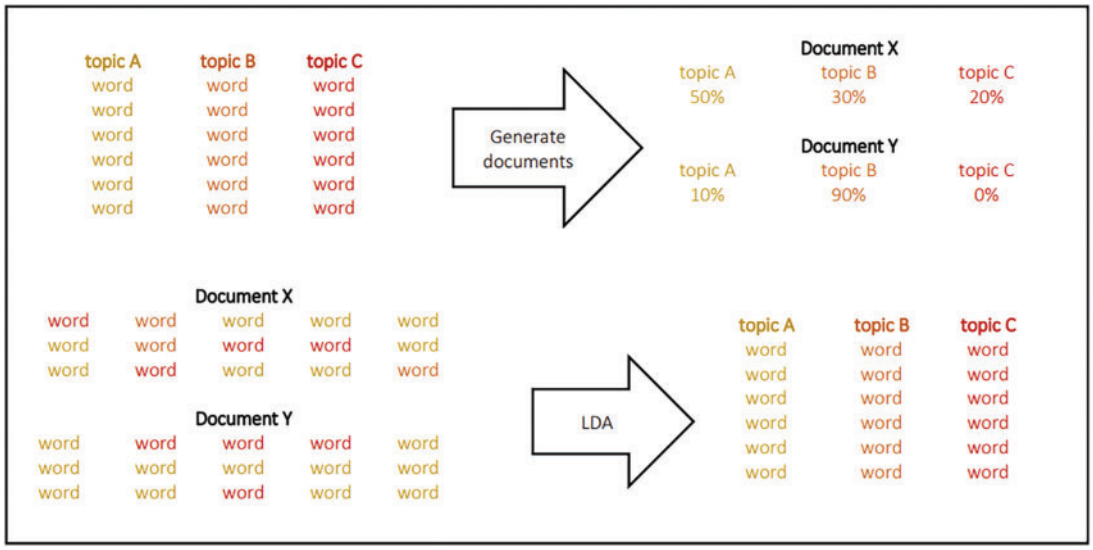

Figure 2: The LDA model

3.5 Sentiment Analysis

We analysed the comments using a sentiment analysis tool to determine their emotional tone as positive, negative, or neutral. In this study, we used the SentiStrength software. It is useful in evaluating English short texts on the social web. SentiStrength reports *two* sentiment strength scores: −1 (not negative) to −5 (extremely negative) and 1 (not positive) to 5 (extremely positive). These two scores are inspired by the psychological reasoning that humans can maintain both positive and negative emotions simultaneously [24]. SentiStrength identifies all the sentiment-related words in a text and scores the overall sentiment based on those words. The sum of the positive score and negative score classifies what categorize the text belongs to. If it is larger than 0, the text is positive. If it is less than 0, the text is negative. The text is neutral when the sum score is equal to 0 and positive score is less than 4 [25].

4 Results

In this study, after using the sorting bar with the 'relevance' filter, a total of 15 videos were chosen. 17,357 comments were collected for this study. After the pre-processing step, we calculated the word frequency and *tf-idf* using the R package. The data which was performed in this step has been pre-processed. Fig. 3 shows the coding for *tf-idf* in R. First, we constructed a term-document matrix with *tf-idf* using function *TermDocumentMatrix*. Second, because the term-document was in the form of a list, it needs to be displayed as a more readable form. We used the function *data.frame* to visualize the result. On the other hand, word frequency is the number of times a word occurs in a document or a corpus, which is different with *tf-idf* scores. To obtain the word frequency scores, we apply the same process as calculating *tf-idf*. In the first step, we removed the component *control* = *list*(*weighting* = *weightTfIdf*) in *TermDocumentMatrix* function. Thus, after the second step,

the output was word frequency values. Fig. 4 illustrates the ranking of the top 30 words in each method. Overall, the top words in *tf-idf* are different with ones in word frequency. For example, 'girl', 'year', 'missed', 'old', 'quality' is the most important words, while 'clothes', 'buy', 'fashion', 'people' and 'wear' is the most frequent word.

```
CorpusDatanew <- tm_map(CorpusData, content_transformer(gsub),
                        pattern = "txt", replacement = "text", fixed = TRUE)
tdm_tfidf <- TermDocumentMatrix(CorpusDatanew, control = list(weighting = weightTfIdf))
m_tfidf <- as.matrix(tdm_tfidf)
headmatrix_tfidf <-head.matrix(m_tfidf,10)
v_tfidf <- sort(rowSums(m_tfidf),decreasing=TRUE)
head(v_tfidf,30)
d_tfidf <- data.frame(word = names(v_tfidf),freq=v_tfidf)
head(d_tfidf,30)
```

Figure 3: Coding for *tf-idf*

4.1 Applying LDA

The next step was to apply the LDA analysis. Topic modelling is an unsupervised machine learning method suitable for discovering the meaning of data. Before running the LDA analysis, we needed to determine the number of topics *k*. *k* is the most important value to define, and it cannot be too small or too large. If *k* is too small, the topics may be too general. If *k* is too large, some of the topics may overlap and cannot be clearly interpreted. To decide on the number of topics, we used the *FindTopicsNumber* function in the *ldatuning* package. This function helps calculate different metrics to estimate the preferable *k* for LDA. In this study, we only use two metrics, *CaoJuan2009* and *Griffith2004*. The best number of topics shows low values for *CaoJuan2009* and high values for *Griffith2004*. According to Fig. 5, the most suitable k value is 14.

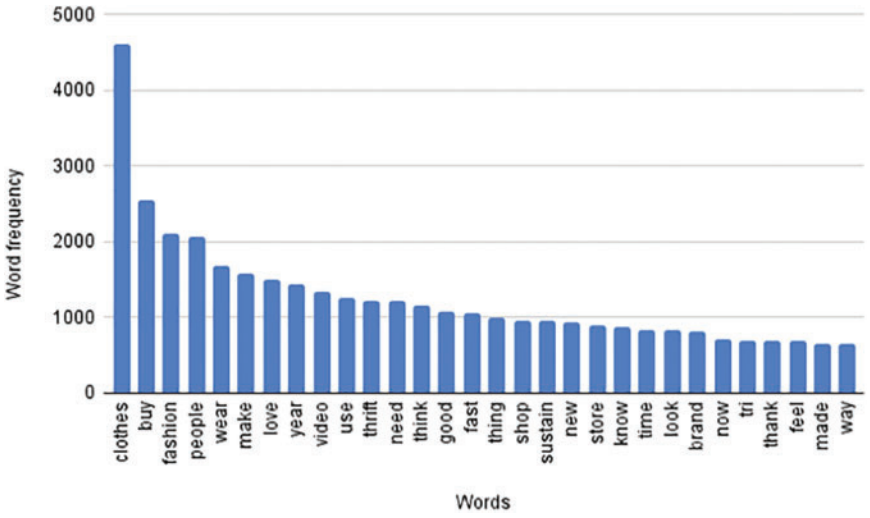

Figure 4: (Continued)

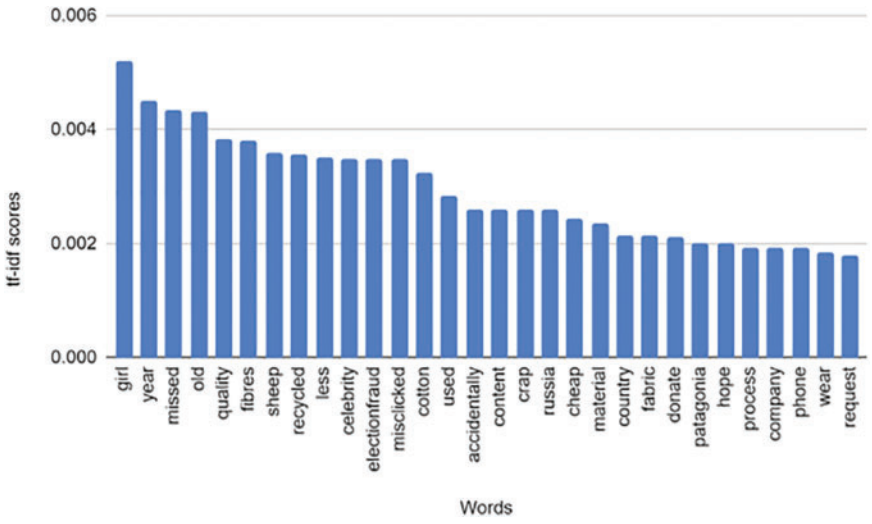

Figure 4: Top 30 words by word frequency and *tf-idf*

```
result <- ldatuning::FindTopicsNumber(
  DTM,
  topics = seq(from = 2, to = 20, by = 1),
  metrics = c("CaoJuan2009",  "Deveaud2014"),
  method = "Gibbs",
  control = list(seed = 77),
  verbose = TRUE
)
```

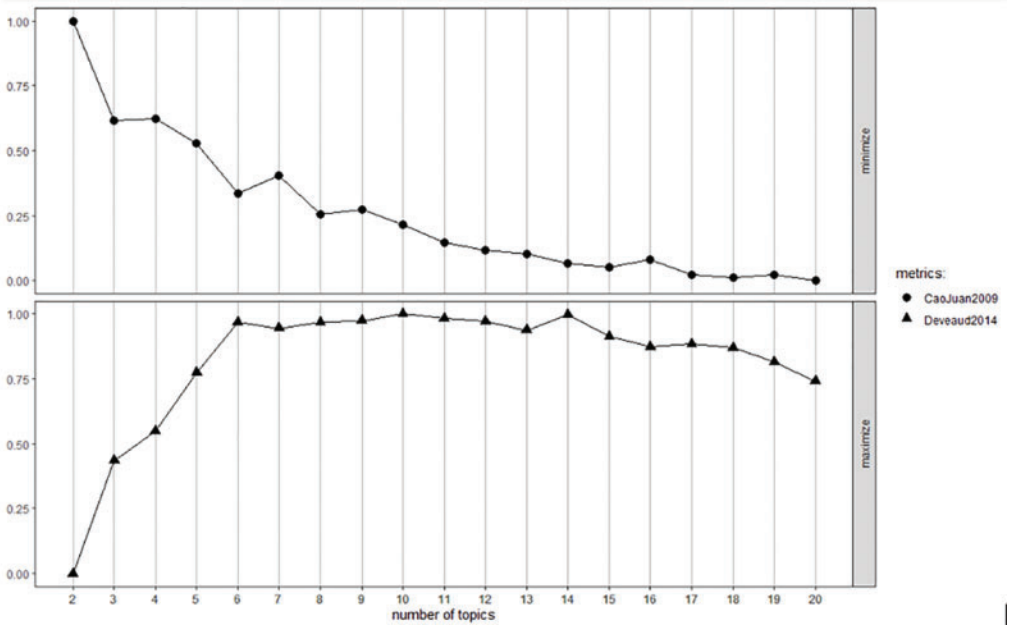

Figure 5: Coding and results of topic tuning

After running the LDA analysis, we had 14 topics with the top 10 most likely terms within the term probabilities. We can now name the topics based on those terms. For example, topic 1 has 'fast', 'people', 'cheap', 'quality', 'money', 'fashion', 'documentary', 'zara', 'stop' and 'expensive'. Therefore, we can name topic 1 'fast fashion'. Table 4 shows 14 topics: Fast fashion, Clothes maintenance, Environmental effects, Wasting management, Ethical brands, Recycling, Media, Thrifting, Second-hand fashion, Fashion industry, Affordability, Durability, Emotional connection, and Purchase intention.

Table 4: LDA topics from YouTube comments data

Topic	Top 10 words
Fast fashion	Fast, people, cheap, quality, money, fashion, documentary, zara, stop, expensive
Clothes maintenance	People, need, make, world, fashion, think, long, care, say, look
Environmental effects	Sustainable, water, cotton, news, women, eco, climate, thinks, certainly, cloth
Wasting management	Quality, away, waste, donate, different, long, shops, time, video, issue
Ethical brands	Video, brands, thank, ethical, shop, great, found, look, thanks, check
Recycling	Wool, recycled, recycling, recycle, every, waste, save, process, hope, make
Media	Zara, episode, show, year, wearing, know, shirts, Netflix, thrift, guy
Thrifting	Thrift, people, think, sustainable, thrifting, videos, fashion, fast, feel, buying
Second-hand fashion	Good, fashion, well, second, start, used, industry, true, comment, nthe
Fashion industry	Fashion, made, clothes, new, many, less, much, problem, world, industry
Affordability	Love, video, thank, sustainable, videos, literally, shop, good, much, affordable
Durability	Clothes, wear, years, buy, people, clothing, fast, use, old, lasting
Emotional connection	Really, much, clothes, stores, love, clothing, going, get, find, see
Purchase intention	Buy, make, know, lot, need, something, many, always, items, want

Another aspect is topic distribution. Fig. 6 illustrates how 14 topics were distributed in 15 documents. Some topics dominate in one document and do not appear in others. For example, the topic 'affordability' appears in documents 3, 7 and 15 yet was barely mentioned in the rest of the corpus. We can see that topic distribution can be uneven between different documents, even though they all have sustainable fashion content.

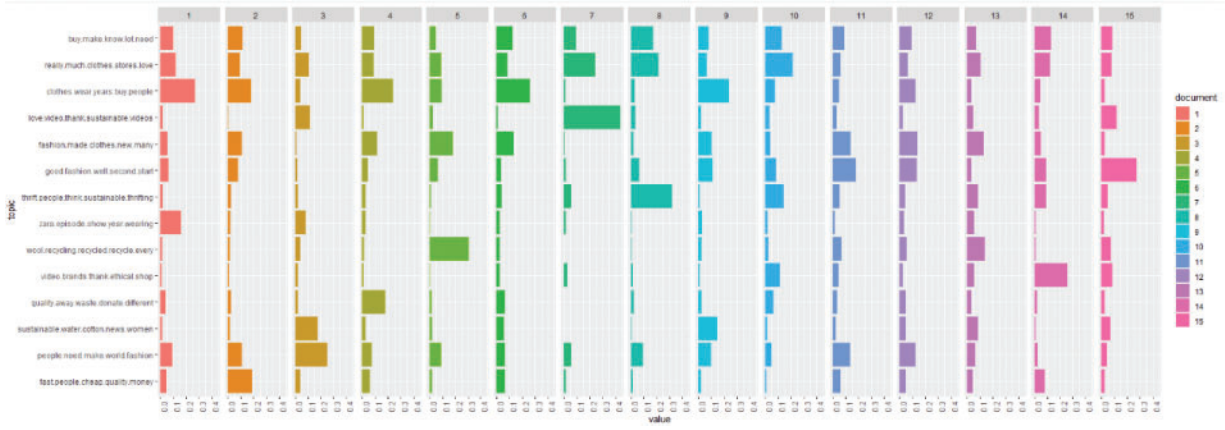

Figure 6: Frequency distributions of the sentiment scores

4.2 Topic Ranking

There are two approaches in topic ranking. The first approach is ranking based on the proportion of each topic in the corpus. Table 5 shows the proportion of each topic in the corpus. The topics are in decreasing order according to their probability. Some topics are more likely to occur than others, such as Purchase intention, Fashion industry and Durability, while Fast fashion and Media did not often occur in the collection of documents.

Table 5: Proportion of each topic in the corpus

Topic	Proportions
Purchase intention	0.20204
Fashion industry	0.16117
Durability	0.10560
Clothes maintenance	0.07510
Affordability	0.06896
Second-hand fashion	0.06557
Recycling	0.06004
Thrifting	0.05316
Wasting management	0.04947
Emotional connection	0.03985
Ethical brands	0.03544
Environmental effects	0.03363
Fast fashion	0.02801
Media	0.02195

The second approach is calculating how frequently a topic is present as a primary topic within a paragraph. This method is also known as Rank-1. The Rank-1 metric shows the number of documents in the corpus in which a topic is the most important. The higher the number of documents, the more

common the topic. If a topic is less common, it might be eliminated in further analysis. As can be seen in Table 6, Durability is the most common topic, followed by Thrifting, Fashion industry, and Clothes maintenance. On the other hand, Thrifting, Fashion industry and Clothes maintenance are less common; while Media, Fast fashion, Environmental effects and Ethical brands does not appear as primary topic in any document.

Table 6: Rank-1 metrics

Topic	Numbers of documents
Durability	3
Thrifting, Fashion industry, Clothes maintenance	2
Recycling, Emotional connection, Affordability, Wasting management, Purchase intention, Second-hand fashion	1
Media, Fast fashion, Environmental effects, Ethical brands	0

4.3 Applying SentiStrength

Before inputting the data into SentiStrength, we needed to clean the data first. Sentences that contained languages other than English or text with obscure meanings were removed manually. SentiStrength program provides scores for whole sentences and a separate score for each word. For example, the sentence 'I wear my clothes till they are full of holes' has positive strength 1 and negative strength −1. The emotion rationale is explained by SentiStrength as follows:

I[0] wear[0] my[0] clothes[0] till[0] they[0] are[0] full[0] of[0] holes[0]
[[Sentence = −1, 1 = word max, 1–5]][[[1,−1 max of sentences]]]

Another example is 'It is amazing how horrible we are doing at reducing waste in general', which has positive strength 3 and negative strength −4, which is explained by:

It[0] is[0] amazing[2] how[0] horrible[−3] we[0] are[0] doing[0]
at[0] reducing[0] waste[−1] in[0] general[0]
[[Sentence = −4, 3 = word max, 1–5]][[[3,−4 max of sentences]]]

Table 7 shows the frequency distribution of the sentiment scores, while Fig. 7 illustrates the sentiment classification. The most frequent positive and negative scores are +1 and −1, respectively. The least frequent positive and negative scores are +5 and −5, respectively, which comprise 0.09% and 0.30% of the whole corpus, respectively. Follow the sentiment classification approach, we discovered that 28.42% are positive comments, 22.35% are negative comments and 49.23% are neutral.

Table 7: Frequency distributions of the sentiment scores

	Scores	%
Positive sentiment scores	5	0.09
	4	2.74

(Continued)

Table 7 (continued)

	Scores	%
Negative sentiment scores	3	14.89
	2	27.57
	1	54.71
	−1	69.98
	−2	18.50
	−3	7.08
	−4	4.14
	−5	0.30

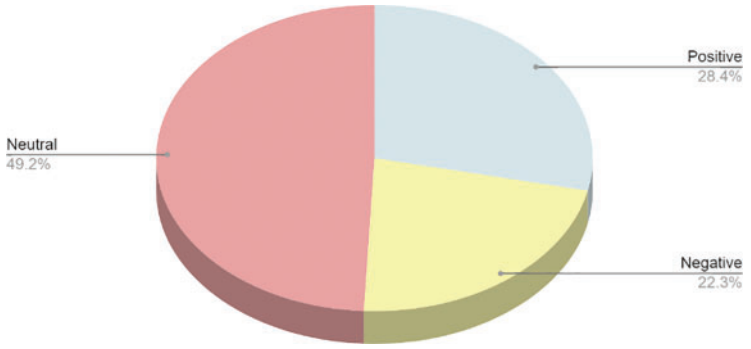

Figure 7: Sentiment classification

5 Discussion

The main goals of this study are to identify and summarise user comments on sustainable fashion YouTube videos. YouTube contains a considerable number of videos on sustainable fashion. The 15 videos selected have significant views and comments, which shows the public is interested in this topic. Because these videos are informative, viewers have gained knowledge from watching them and have shared their opinions.

Our study used 17,357 comments from videos covering specific content, such as the definitions, benefits and applications of sustainable fashion. By calculating word frequency and *tf-idf* scores, we detected the most frequently used and relevant words. The words 'clothes', 'buy', 'fashion', 'people' and 'wear' occur most frequently in the comments section in the 15 videos. On the other hand, the highest *tf-idf* score words are 'girl', 'year', 'missed', 'old', 'quality', 'fibres', 'sheep', 'recycled', 'less' and 'celebrity'. These words are considered the 'keywords' from all comments.

We also conducted the LDA topic modelling analysis to identify the underlying topics discussed across the 15 video comment sections. After determining the number of topics, we generated 14 topics: Fast fashion, Clothes maintenance, Environmental effects, Wasting management, Ethical brands, Recycling, Media, Thrifting, Second-hand fashion, Fashion industry, Affordability, Durability, Emotional connection, and Purchase intention.

Among these topics, Purchase intention, Fashion industry and Durability were the most common in the whole corpus. In addition, based on the Rank-1 method, we discovered that Durability is the most important topic in the three videos' comments sections, while Thrifting, Fashion industry and Clothes maintenance came in second, in the two videos. Surprisingly, Media, Fast fashion, Environmental effects and Ethical brands were not the primary topic in any video's comments. Combining these results, the viewers are more interested in discussing Purchase intention, Durability, Thrifting, Fashion industry and Clothes maintenance. Therefore, YouTubers can improve the video content's quality by further exploring these topics. The content may shift the focus to these topics or the creators may create a video series for each topic. By running a videos series that covers viewer's interest themes or topics, YouTubers can encourage viewers to follow up and engage more in their channel. For example, a series of test on the durability of clothes from different eco-fashion brand is recommended under the topic of Durability. Additionally, content creators can attract more audience's attention by including these topics as keywords in video titles and tags. Furthermore, those keywords can be used in customizing thumbnails for more a more appealing appearance.

These results also benefit for business managers when making decisions. For sustainable fashion brands, these topics are considered as the most thing which potential customers care about when they come in contact with the brand. Therefore, business managers should highlight the product characteristics which are related to these topics for customers. For instance, with the topic of clothes maintenance, when displaying the product both in brink-and-mortar store and online store, an instruction on clothes maintenance for the best result should be included. Moreover, when launching a marketing campaign, the durability of the product may become the key factor to increase brand awareness.

Because YouTube users frequently express their feelings in their comments, we also examined the emotion in the comments expressing the viewers' opinions on sustainable fashion. The SentiStrength analysis produced a negative and positive score for each comment. The most frequent sentiment scores were found at the least extreme scores, which are 1 and -1. While the percentage of positive scores 1, 2 and 3 and negative scores -1 and -2 are high, the percentage of scores 4, 5, -3, -4 and -5 are relatively low. In addition, comments with neutral expressions are almost half the total comments, while positive and negative ones are second and third, respectively. During the text pre-processing for sentiment analysis, we observed that most negative comments belonged to the Fast fashion, Environmental effects and Fashion industry topic. On the contrary, people gave positive comments under Recycling, Thrifting and Second-hand fashion topic. For other topics such as Clothes maintenance, Ethical brands, etc., they shared their experiences and discussed how to apply sustainable fashion.

There are several limitations in this study, which may affect the validity of the results. A potential thread to validity is the software. SentiStrength has been said to outperform other lexical classifiers. However, we are also aware that it may have some limitations. First, the data need to be cleaned before using SentiStrength, which is tedious and time-consuming. Second, SentiStrength cannot interpret text emoticons. For instance, the text ':D' signifies laughter or a big grin in a text message, which expresses happiness. Nevertheless, the software evaluates this into a 1 positive score and a -1 negative score, which is a misclassification. Another limitation is that the cleaning data process before employing in the SentiStrenght was done by human. Human tend to make subjective judgements, which could possibly lead to miss out important data.

6 Conclusion

This study focuses on the analysis of user comments on YouTube videos related to sustainable fashion. The primary objective was to assist content creators and business managers understand viewer opinions, improve video quality and drive business growth. Topic modelling and sentiment analysis were employed on a large dataset of comments from different sustainable fashion videos. The analysis uncovered 14 topics discussed in the comments and revealed the distribution of positive (28.42%), negative (22.35%), and neutral (49.23%) sentiments expressed by users.

In order to help YouTubers and business manager achieve their goals, potential future work for this paper could be:

- Collecting larger data by more advanced tools with a wider range of videos.
- Combining others programming language to provide a comprehensive approach to sentiment analysis.

Acknowledgement: None

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: L.H. is the primary supervisor of the research. L.H. also contributed to the writing and review of manuscript. N.M. contributed to the writing, data collection and analysis. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Data can be accessed through https://doi.org/10.6084/m9.figshare.23787273.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] B. Shen, "Sustainable fashion supply chain: Lessons from H&M," *Sustainability*, vol. 6, no. 9, pp. 6236–6249, 2014.

[2] Trends.google.com.tw, "Google Trends. Sustainable fashion," 2023. [Online]. Available: https://trends.google.com.tw/trends/explore?date=all&q=sustainable%20fashion&hl=en/ (accessed on 23/06/2023).

[3] Statista Insights & Facts, "YouTube-Statistics & Facts," 2023. [Online]. Available: https://www.statista.com/topics/2019/youtube/#topicOverview/ (accessed on 15/06/2023).

[4] Statista Insights & Facts, "Most popular websites worldwide as of November 2022, by total visits (in billions)," 2023. [Online]. Available: https://www.statista.com/statistics/1201880/most-visited-websites-worldwide/ (accessed 15/06/2023).

[5] IBM–United States, "What is natural language processing (NLP)?" 2023. [Online]. Available: https://www.ibm.com/topics/natural-language-processing (accessed on 15/06/2023).

[6] S. Greco, "Twitter activists' argumentation through subdiscussions: Theory, method and illustration of the controversy surrounding sustainable fashion," *Argumentation*, vol. 37, pp. 1–23, 2023.

[7] S. Blasi, L. Brigato and S. R. Sedita, "Eco-friendliness and fashion perceptual attributes of fashion brands: An analysis of consumers' perceptions based on twitter data mining," *Journal of Cleaner Production*, vol. 244, pp. 118701, 2020.

[8] D. S. Testa, S. Bakhshian and R. Eike, "Engaging consumers with sustainable fashion on Instagram," *Journal of Fashion Marketing and Management*, vol. 25, no. 4, pp. 569–584, 2021.

[9] J. Kim, S. Kang and K. H., Lee, "How social capital impacts the purchase intention of sustainable fashion products," *Journal of Business Research*, vol. 117, pp. 596–603, 2020.
[10] S. Haines, O. H. Fares, M. Mohan and S. H. (M). Lee, "Social media fashion influencer eWOM communications: Understanding the trajectory of sustainable fashion conversations on YouTube fashion haul videos," *Journal of Fashion Marketing and Management*, vol. 27, no. 6, pp. 1027–1046, 2023.
[11] E. Poché, N. Jha, G. Williams, J. Staten, M. Vesper *et al.,* "Analyzing user comments on YouTube coding tutorial videos," in *Proc. of 25th Int. Conf. on Program Comprehension (ICPC)*, Buenos Aires, Argentina, pp. 196–206, 2017.
[12] C. S. Lee, H. Osop, D. H. L. Goh and G. Kelni, "Making sense of comments on YouTube educational videos: A self-directed learning perspective," *Online Information Review*, vol. 41, no. 5, pp. 611–625, 2017.
[13] S. S. Alshamrani, M. Abuhamad, A. A. Abusnaina and D. A. Mohaisen, "Investigating online toxicity in users interactions with the mainstream media channels on YouTube," in *Proc. of Int. Conf. on Information and Knowledge Management*, Galway, Ireland, 2020.
[14] H. Bhuiyan, J. Ara, R. Bardhan and R. M. Islam, "Retrieving YouTube video by sentiment analysis on user comment," in *Proc. of IEEE Int. Conf. on Signal and Image Processing*, Kuching, Malaysia, pp. 474–478, 2017.
[15] A. Xu, T. Qi and X. Dong, "Analysis of the Douban online review of the MCU: Based on LDA topic model," *Journal of Physics: Conference Series*, vol. 1437, pp. 012102, 2019.
[16] L. Zou and W. W. Song, "LDA-TM: A two-step approach to Twitter topic data clustering," in *Proc. of 2016 IEEE Int. Conf. on Cloud Computing and Big Data Analysis (ICCCBDA)*, Chengdu, China, pp. 342–347, 2016.
[17] Y. Wang, M. Burke and E. K. Kraut, "Gender, topic, and audience response: An analysis of user-generated content on Facebook," in *Proc. of SIGCHI Conf. on Human Factors in Computing Systems (CHI '13)*, New York, NY, USA, pp. 31–34, 2013.
[18] R. Albalawi, T. H. Yeap and M. Benyoucef, "Using topic modeling methods for short-text data: A comparative analysis," *Frontiers in Artificial Intelligence*, vol. 3, pp. 42, 2020.
[19] M. Thelwall, "The heart and soul of the web? Sentiment strength detection in the social web with SentiStrength," In: J. Holyst (Ed.), *Cyberemotions*, pp. 119–134, Cham: Springer, 2017.
[20] M. Thelwall, K. Buckley and G. Paltoglou, "Sentiment strength detection for the social web," *Journal of the American Society for Information Science and Technology*, vol. 63, no. 1, pp. 163–173, 2012.
[21] M. Thelwall, P. Sud and F. Vis, "Commenting on YouTube videos: From guatemalan rock to EI Big Bang," *Journal of the American Society for Information Science and Technology*, vol. 63, pp. 616–629, 2012.
[22] D. Vilares, M. Thelwall and M. A. Alonso, "The megaphone of the people? Spanish SentiStrength for real-time analysis of political tweets," *Journal of Information Science*, vol. 41, no. 6, pp. 799–813, 2015.
[23] D. Blei, A. Ng, M. Jordan and J. Lafferty, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.
[24] R. Berrios, P. Totterdell and S. Kellett, "Eliciting mixed emotions: A meta-analysis comparing models, types, and measures," *Front Psychol*, vol. 6, pp. 428, 2015.
[25] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai and K. Kappas, "Sentiment strength detection in short informal text," *Journal of the American Society for Information Science and Technology*, vol. 6, pp. 2544–2558, 2010.