ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Master Thesis: Data Science and Marketing Analytics (DSMA)

What Factors Make Fashion Shows Popular?

A Text Analysis Study of Fashion Show Reviews

Name student: Quinten Martis

Student ID number: 494084

Supervisor: Dr. A. Archimbaud

Second Assessor: Dr. M. van Crombrugge

Date final version: 28/04/2023

Abstract:

This thesis empirically investigates fashion show reviews and attempts to identify several factors that influence the positivity/ negativity of said reviews, as well as examining whether these reviews have a significant impact on fashion brands' well-being. Using fashion show review data dating from 2000 to 2014 derived from Style.com and Yahoo Finance, Latent Dirichlet Analysis (LDA) was performed to identify various popular topics such as print, dress, design, jacket, and girl. The following Sentiment Analysis, Welch's t-tests, ANOVA, and Multiple Regression, found that following current fashion cycles, having a female creative director, showing collections during the Spring/Summer fashion month, showing in New York, and avoiding years of financial hardship all have a significant positive impact on the sentiment score of reviews. Furthermore, sentiment score was found to be correlated with company well-being in the form of stock prices, showing the importance for luxury fashion brands to focus on these aspects of their shows.

Table of Contents

1. Introduction

Luxury fashion houses usually release their collections multiple times throughout the year: spring/summer, autumn/winter, and resort/pre-fall. Showing new products this often, leads to an abundance of opinions through journalistic reviews in fashion magazines and websites. There are numerous fashion media outlets that report on each collection such as Vogue, Elle, Harper's Bazaar, Numero, Business of Fashion, and WWD. Furthermore, some of these have as many as 22 editions for different countries around the world. Thus, with each edition writing unique reviews, the sheer size of reviews is so large that it would be highly inefficient to manually study. These reviews furthermore have a heavy influence on the customers and their purchasing habits (Zhu & Zhang, 2010. As a result of all of this, it is important for marketers to have the ability to analyse the, often unstructured, review data. Fashion companies can thus focus on the aspects of the shows that are (un)successful and consider this when going into production or for future collections. Since fashion shows have such a grand number of factors from the fabric of each piece to the details of the set design, it is challenging to know what factors influence opinions in what direction and at which scale. I plan on investigating the characteristics that lead to positive or negative reviews through the research question of:

"What aspects of luxury fashion collections and shows have a positive or negative influence on journalistic reviews?"

Related sub questions include: How does the gender of the designer effect reviews? Does season of the show have an effect on review? Does the city where the show is held affect reviews? Are reviews affected by major events such as 9/11, COVID-19, or the stock market crash of 2008? Are there any trends in what reviewers like and dislike and does this differ during fashion cycles? Do the positive or negative reviews have any association with sales?

The scientific relevance of this study can be seen as there is an increasing importance of computational text analysis in research (Büschken & Allenby, 2016). This importance arises as it is a challenge to analyse unstructured reviews and make sense of the various topics discussed. Furthermore, some of the machine learning methods to be used, are relatively new, and although they should serve as improvements of more traditional models, have not been tested as extensively on different datasets. When looking at the luxury fashion industry specifically, research is slim to none. This could be due to lack of previous data or the fact that fashion shows are so multidimensional that text summaries have been too intimidating.

When looking at the marketing side, it can be seen that the newly timed interest in text analysis is reflected in slim or new research revolving its use for marketing. Amado et al. (2018), for example, studied trends in big data in marketing through an analysis of literature written from 2010 to 2015. Their findings show that big data publications are not aligning cutting edge techniques toward marketing benefit and that these applications are in the "embryotic stage" (Amado et al., 2018). Another article by Berger et al. (2020) does, however, show promising insights on the use of textual analysis for marketing. They find that it can be used for prediction and understanding. They also elaborate on the opportunities for future research including expanding the data sources used, looking at different types of interactions, and different research topics.

In terms of social relevance, the luxury fashion market is currently booming. For example, LVMH has recently become the largest market capitalization in the Euro zone. LVMH owns some of the world's top grossing luxury brands, including Fendi and Givenchy which are included in this dataset. Additionally, Prada Group, owner of Prada and Miu Miu, reported strong growth in sales last year despite the effects of the coronavirus pandemic (Spencer, 2021). Specifically, revenues for 2021 totalled $3.8 billion, increasing by 41 percent from 2020 and eight percent compared to 2019 (Spencer, 2021).

Thus, finding possible effects on financial performance in this market is highly socially relevant. These fashion brands can furthermore heavily benefit from knowing which aspects of

their collections, in particular, are successful or unsuccessful as this can have implications on their reputation and sales.

This paper will investigate fashion show reviews and attempt to identify several factors that influence their positivity/ negativity, as well as examining whether these reviews have a significant impact on fashion brands' well-being. Section 2 will begin with a literature review, where existing studies will be examined for possible mechanisms that affect review sentiment. These include fashion cycles, market disruptions, gender bias, and prestige in relation to location of the show. Moreover, literature on the effect of reviews on company sales and revenue will be examined. Section 3 will give illustrate of the data used for the study including descriptive statistics. This is followed by section 4 which includes a comprehensive description of the methods used, as well as how they will be implemented for this study. This includes Latent Dirichlet Analysis (LDA) to discover which topics are most discussed in the fashion show reviews. Then, to analyse which of these topics are positive and negative, sentiment analysis will be performed on the journalistic fashion show reviews, The resulting sentiments will then be compared with restrictors based on, key words, gender of house founder, and season of the show. These will then be evaluated for their possible statistically significant difference through two sample t-tests. Furthermore, ANOVA will be performed to compare sentiments between cities and years. The years will be further analysed through a multiple linear regression. The section concludes with an investigation on relationship between review sentiment and company stock data will be done through plot comparison. Section 5 will illustrate results, explored by going through the hypotheses, previously drawn in the literature review. The paper rounds off with a conclusion in section 6 that will include discussions on the results, as well as a post hoc reflection of the study.

2. Literature Review

When again reflecting on the literature related to this topic, only a modest amount exists and most is often outdated; especially when looking at the mechanisms in question. Nevertheless, available research does provide some insight into answering the related research question and sub-questions.

When looking at the research question of "*What aspects of luxury fashion collections and shows have a positive or negative influence on journalistic reviews?"* a possible influence on the reviews is whether the shows follow the fashion trends of the time, or rather, does not follow those prior. This can be explained by the phenomenon of fashion cycles.

Fashion cycles refer to the phenomenon of how fashion trends arise, then peak in popularity which leads to an eventual decline over time. Fashion cycles have been observed in many industries, including clothing, footwear, and accessories. In his article, Pesendorfer (1995) finds that the length of fashion cycles can vary widely, from a few months to several years. The factors that then influence the length of these cycles are multifaceted and include factors such as shifts in consumer preferences, advances in technology, and changes in cultural norms (Pesendorfer, 1995). Fashion cycles typically follow a predictable pattern of stages, beginning with the introduction of a new trend or style. Early adopters and fashion influencers then adopt the trend, resulting in its wider adoption among consumers. As the trend becomes more popular, it reaches its peak, and eventually, starts to decline as consumers move on to newer trends. These trends often reflect designers' experimentation and innovation with new materials, colours, patterns, cuts, and silhouettes.

The length of these fashion cycles is thus not fixed. Franck (1997) finds that fashion cycles tend to be shorter in times of economic success, while in times of economic hardship, they tend to last several seasons. He also suggests that trends tend to simply be variations of older trends as innovation is rare. A more recent article by Wu and Song (2019) finds two opposing forces in terms of the course of fashion cycle length. Firstly, that trends in recent are becoming more globalized and fast-paced due to the presence of fast fashion and social media. However, as its

environmental and social impact has increased in attention, there is also a growing trend towards ethical and sustainable fashion. This may thus lead to longer fashion cycles.

When focusing on what aspects of fashion shows might have a negative influence on reviews in the 2000's, one might assume that words associated with 1990's fashion trends would increase negativity due to the fashion cycles. Kass outlines some of these trends in her article, "*The 20th Century of American Fashion: 1900-2000*" (2011). These include the minimalist trend, which emphasized clean lines and simple silhouettes. Additionally, the grunge trend which was characterized by flannel shirts, ripped denim, and combat boots.

This also ties in with the rise of streetwear fashion, which was influenced by hip hop and skate culture, which then also influenced the emergence oversized clothing and logomania, or prominent displays of logos. Other trends include high-tech, exotic elements, and androgyny through more masculine styles.

Although some trends, like logomania and futurism, carried over into the 2000's, the new decade brought with it differing trends. These include the return of glamour through fabrics such as velvet, silk, and embellishments. This also tied into the popularisation of retro themes from the 1970's and 1980's, like high wasted pants and skirts. Additionally, there was the emergence of skinny, form-fitting, clothing, and military inspired clothing such as army jackets and cargo pants. Lastly, a major trend during the time was the bohemian trend, characterized by relaxed, flowy, clothes and earth tones (Donohue, 2022). These factors have led to the first hypothesis of:

"Fashion show reviews with words associated with previous fashion trends will have a lower sentiment than those reflecting those of the current fashion cycle"

A second factor that could affect reviews, is the gender of the fashion designer. Misogyny is still pervasive issue in current day society. This comes in the form of violence, online harassment, underrepresentation in business and politics, and the gender pay gap; all stemming from the mindset that women and men are not equal. This could influence reviews as female designers would be critiqued harsher than their male counter parts. Paustian-Underdahl et al. find evidence for this in their meta-analysis (2014). They find that gender biases and stereotypes influence how

individuals perceive leadership effectiveness. Furthermore, women are more likely to receive critical feedback on their performance, personality, and communication style (Paustian-Underdahl et al., 2014).

Stokes further studied how gender affects people in the fashion industry. In their paper "The glass runway: How gender and sexuality shape the spotlight in fashion design" (2015), finds that women in fashion design face challenges that affect their ability to both succeed and receive recognition in the industry. For example, female designers are less likely to receive funding, even with comparable levels of talent and experience. In addition, women in the fashion industry have reduced access to networks and mentorship. This is particularly effective as these mentors and insiders tend to be predominantly male (Stokes, 2015). Lastly, there seems to be a bias in the recognition that female designers receive. As women may be held to different standards than men, deserving female designers are overlooked for awards and recognition. The combination of these findings has led to the third hypothesis of:

"Reviews on collections by female founded fashion houses have lower overall sentiment than those by their male counterparts"

A third factor that could affect review sentiment, is the time of year the review is written in. Here, the weather, temperature, and sun exposure the journalists experience could have an impact on their mood and in turn, the sentiment of their reviews. The reasoning behind this would coincide with that of seasonal depression, although not necessarily at that severity level. The National Institute of Mental Health (NIMH) finds that seasonal changes to fall, where days are shorter and there is less sunlight, can lead to "feeling sad, depressed, or hopeless", as well as feeling agitated and sluggish (2021).

Other studies find affirming results. Lu and Cooper (1999) did a study on job satisfaction and found that employees reported feeling more satisfied with their jobs and less stressed during the summer months compared to the winter months. The researchers suggested that this could be due to the longer daylight hours and warmer weather during the summer, which can boost mood and increase feelings of well-being.

Terman et al. did a study on seasonal affective disorder and found that depression among workers was highest during the winter months (1989). The cause this shift in mood and well-being, they found, could include colder weather, lower sun exposure, and reduced daylight hours. These findings lead to the third hypothesis of:

"Reviews written for the Spring/Summer collections will have a lower sentiment than those written for the Fall/Winter shows."

The semantics for this hypothesis are important as Spring/Summer collections are shows in the fall and would thus have lower sentiments according to the hypothesis.

Other factors that might influence reviews are global disasters and issues. Here, writers would let their outside strains affect their work. Although it is often said to separate one's work life with their personal life, it is logical that this is not possible in extreme cases.

The first possible mechanism that causes this could be mood. Being surrounded by tragedy can affect a person's mental and physical health, possibly leading to symptoms of anxiety, depression, and chronic pain. Parker et al. found that workers who had experienced trauma had higher levels of work-related strain, as well as lower job satisfaction and performance compared to those who had not experienced trauma (2014).

Another mechanism that causes this could be stress. Here, peoples outside stresses, whether it come from economic hardship, from the global financial crisis for example, would spill over into other aspects of their life. Marco and Suls (1993) found that current-day stress is a major influence of mood across multiple days. This excess of stress and affected mood would thus make writers be overly critical or negative.

In terms of the September 11 attacks, its effect on journalistic reviews could be seen as the offices for Vogue are in New York City. Thus, the proximity to the event is likely to have a direct effect on the employees. Laugharne et al. (2007) studied the effects of the attack 5 years later. They found that posttraumatic stress disorder (PTSD) is associated with indirect and direct

exposure to terrorist attacks. Specifically, the rates for PTSD were around 20-38% in the weeks following terrorist attacks, and symptoms persisted throughout the years in the case of direct exposure. There were also factors that increased risk of PTSD including: geographical closeness to the location of an attack, being female, and a high amount of media reporting of terrorist attacks can be associated with the development of PTSD in vulnerable individuals (Laugharne et al., 2007). These might be significant, particularly female sex, as over 70% of Vogue employees are women (Conde Nast, 2021) and the demographic of fashion brands are largely women. These findings have led to the second hypothesis:

"Market disruptions such as the September 11 attacks or the Global Financial Crisis have a negative effect on journalistic reviews"

A fifth mechanism that could influence reviews is location. Showing at Paris Fashion Week is widely considered as one of the most prestigious events in the fashion industry, due to its history and criterions. Paris Fashion Week is held twice a year in Paris, France, and features the most famous designers and fashion houses (The Guardian, 2019). Not only is it seen as the "the ultimate seal of approval" for a designer, but it can also lead to more recognition and an increase in sales. In her article, Arnold (2018) validates these claims by stating that the prestige of fashion excellence in Paris stems from the 1700's, when Paris was seen as the center for the luxury fashion market. She finds that these traditions were upheld though their expert artisans and their craftsmanship, as well as the traditions and established fashion houses. These findings lead to the fourth hypothesis of:

"Reviews on collections shown in Paris have a higher sentiment than those shown in New York, London, or Milan"

All these mechanisms that could affect runway show reviews would be of high importance to fashion brands given that they affect the well-being of the company. In "Reviews, Reputation, and Revenue: The Case of Yelp.com" Luca finds an effect on the sales of a company based on

reviews. Specifically, that that a one-star increase in a restaurant's Yelp rating led to an increase in revenue of 5-9% (Luca, 2016). Reviews can also affect company reputation, employee performance, and customer loyalty, which can further impact the economic well-being of a company (Liu & Jang, 2009).

When looking at the fashion industry in particular, a study by the Council of Fashion Designers of America (CFDA) looked at the impact of fashion critic reviews on designer sales (2014). They surveyed 50 designers who presented collections during New York Fashion Week and analyzed sales data before and after the publishing of the reviews. They found that positive reviews saw and average increase in sales of 35%. This effect was also magnified for not yet well-established designers (CFDA, 2014). It is a given that there are many other factors that affect sales around the time of New York Fashion Week, but this may be an indication that reviews have influence.

In terms of negative impact, the CFDA found that designers who received negative reviews saw an average sales decrease of 33%. Another study by McKinsey & Company found supports this by finding that negative reviews, as well as social media comments, can quickly spread and undermine a company's reputation, leading to decreased customer trust and a loss in sales. These findings have led to the fourth hypothesis of:

"Sales, as reflected by respective stock prices, mirror the review sentiment score trends"

Altogether, through answer these research questions, this paper differentiates itself from previous research and literature through the by using an updated data set with a greater number of observations and variables. Furthermore, the methods used include machine learning methods which are yet to be publicly used in the luxury fashion area. Lastly, the reviews will be compared to metrics for company well-being to attempt to provide luxury fashion brands with concrete suggestions to improve financial welfare.

3. Data

The dataset in this research consists of runway show reviews written by various journalists commissioned for style.com. The data retrieved from the Harvard Dataverse (Chilet et al., 2017) includes 6629 runway reviews of high-end fashion brands from 2000 to 2014 that were scraped off style.com, which is now vogue.com. The dataset contains a total of 8 variables. These variables are *year, season, designer, author, city, date, image,* and *review.* The *year* and *season* variable represent the year and season for which the show was presented, respectively. Additionally. The *designer* variable represents the name of the fashion brand. There are a total of 817 different brands which include subset brands, for example, the brand Victoria Beckham has a subset brand Victoria by Victoria Beckham. The variables *author* and *city* indicate the journalist who wrote the style.com review while city indicates the city in which the fashion show took place, respectively. Furthermore, the *date* variable shows the date of the review.

The reviews in the dataset were posted between September 12, 1999, and June 12, 2014. As fashion shows are usually shown a few months prior to the season/year they are designed for, the dates are ahead of the *year* variable. The dataset also contains an image variable which contains the header image link accompanying the review, however, these were omitted for this research. Lastly, the *review* variable contained the actual review by the journalist.

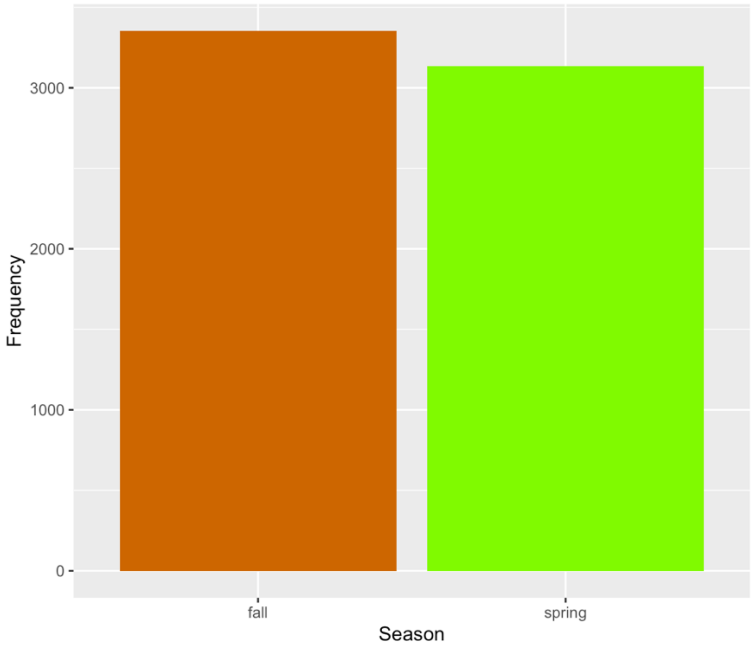

Figure 1: Review frequency per season of collection

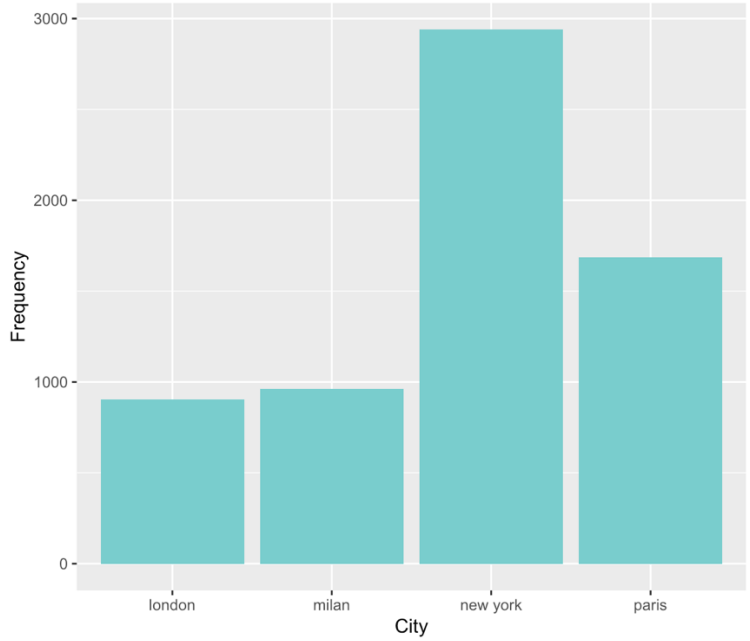

Figure 2: Review frequency per city

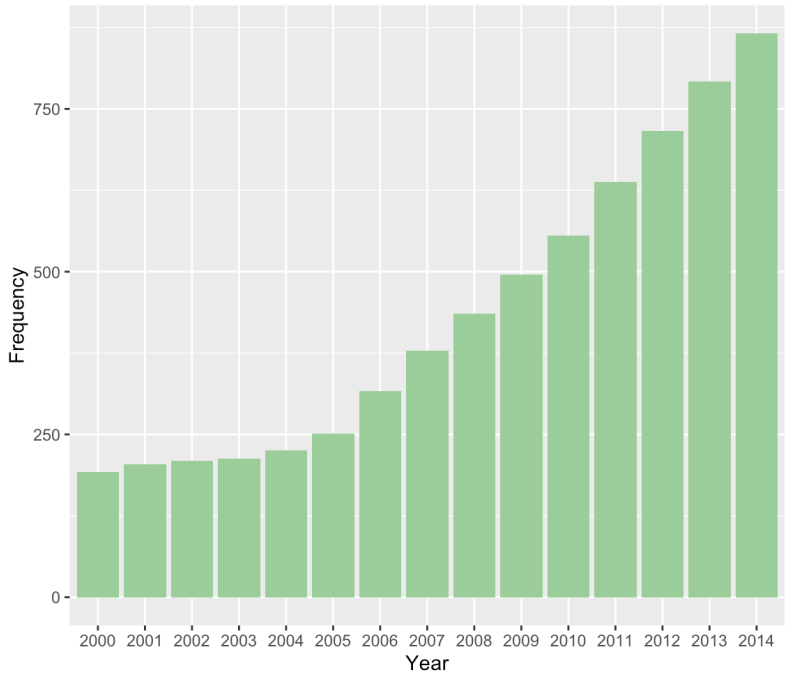

Figure 3: Review frequency per year

Figures 1 through *3* show the frequencies for the *season, city, and year* variables, respectively. It can be seen that the number of reviews seem inconsistent for all variables. For season, there are 3354 reviews written for fall/winter collections while there are 3135 for spring/summer collections. In terms of city, there is a clear majority for reviews written on shows shown in New York City. This majority has a count of 2940, followed by Paris with 1684, Milan with 961, and lasty London with 904. Lastly, in *Figure 3,* the number of reviews seems to rise exponentially throughout the years.

A second dataset includes stock prices for four top French luxury companies. The data is retrieved from Kaggle (Kanawattanachai, P., 2022) and includes containing historical daily stock prices from 2000 till 2022 for Louis Vuitton, Christian Dior, Hermes, and Kering, totalling to 22,941 observations. The dataset contains 8 variables, namely *date, symbol, adj_close, close, high, low, open,* and *volume.* The *date* variable lists the date which ranges from December 31, 1999, till May 19, 2022. The *symbol* variable indicates the company name, while the *adj_close, close, high, low,* and *open* variables represent the adjusted closing price, closing price, high price, low price, and

open price of the stocks selling that day, respectively and in US dollars. Lastly, the volume variable represents the amount of number of shares traded in the stock.

For this analysis, average adjusted closing price will be calculated per year. This variable not only considers the cost of shares at the end of the day, but also considers other factors like dividends, stock splits, and new stock offerings. Additionally, the data used will be cut, but not till the end of 2014, which is when the style.com dataset ends. The stock price data set will end in 2015, this is to consider the dissimilar dates for the showing of the collection and the actual release of the collection in stores and online. For example, spring/summer collections are displayed to the press during fashion week in September for the following summer. In contrast, fall/winter collections are shown in February for the coming fall/winter.

3.1 Data Pre-Processing

The data first underwent a general cleaning that included removing blank reviews, reviews that were simply a caption or stated that the review would be "posted shorty", fixing typos in journalist names, and removing the "by" in front of journalist names as this would falsely classify observations. After this cleaning, the data set is left with 6489 observations for 707 fashion brands.

For answering the various sub-questions, some adjustments needed to be made to the data. This included adding a *gender* variable to indicate the gender of the founder of the brand. This was done manually, where a brand that included a female founder, was classified as female founded. *Figure 4* shows the count of collections per gender of the designer. It can be seen that there are more collections designed by male designers with a frequency of 3836, compared 2653 for women.

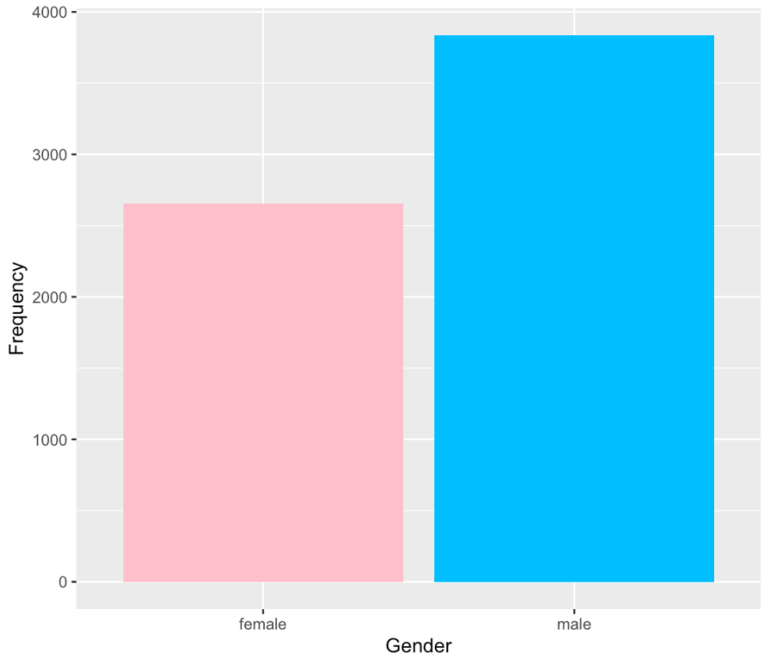

Figure 4: Review frequency per gender of designer

To prepare the data for sentiment and LDA analysis, the review data that is used as the input is first cleaned to ensure accurate analysis. First, unnecessary punctuation is removed. this includes exclamation marks, numbers, dots, commas, times, dollar values, dashes, hyphens, excess spaces, and any other symbols. Stop words are removed next, so that only polarizing terms are to be considered. For the LDA analysis, the words are stemmed and tokenized. Stemming converts the words into their root form. Here words such as “detailed” and “detail” are both seen as “detail” as they are recognized as having the same meaning. This is necessary as LDA analysis if based on word frequency and words that are equivalent should be counted as such. Lastly, a corpus must be made using the stemmed review data. This is a collection of documents where only the frequency of each word is considered and where the order of the words is disregarded, a necessary condition for LDA analysis.

Figure 5 illustrates an example of the transformation a review undergoes through pre-processing. The raw review is taken directly, as is, from Style.com. The first transformation is then made by removing punctuation. It can be seen punctuation such as commas, dashes,

apostrophes, and periods are not included in the second stage of the review. Lastly, the stop words are removed and the words in the review are stemmed. The final form of the review is shorter as words such as “a” and “the” are removed. Furthermore, words like “simple,” “easygoing,” and “carry” are transformed into “simpl,” “easygo,” and “carri,” respectively.

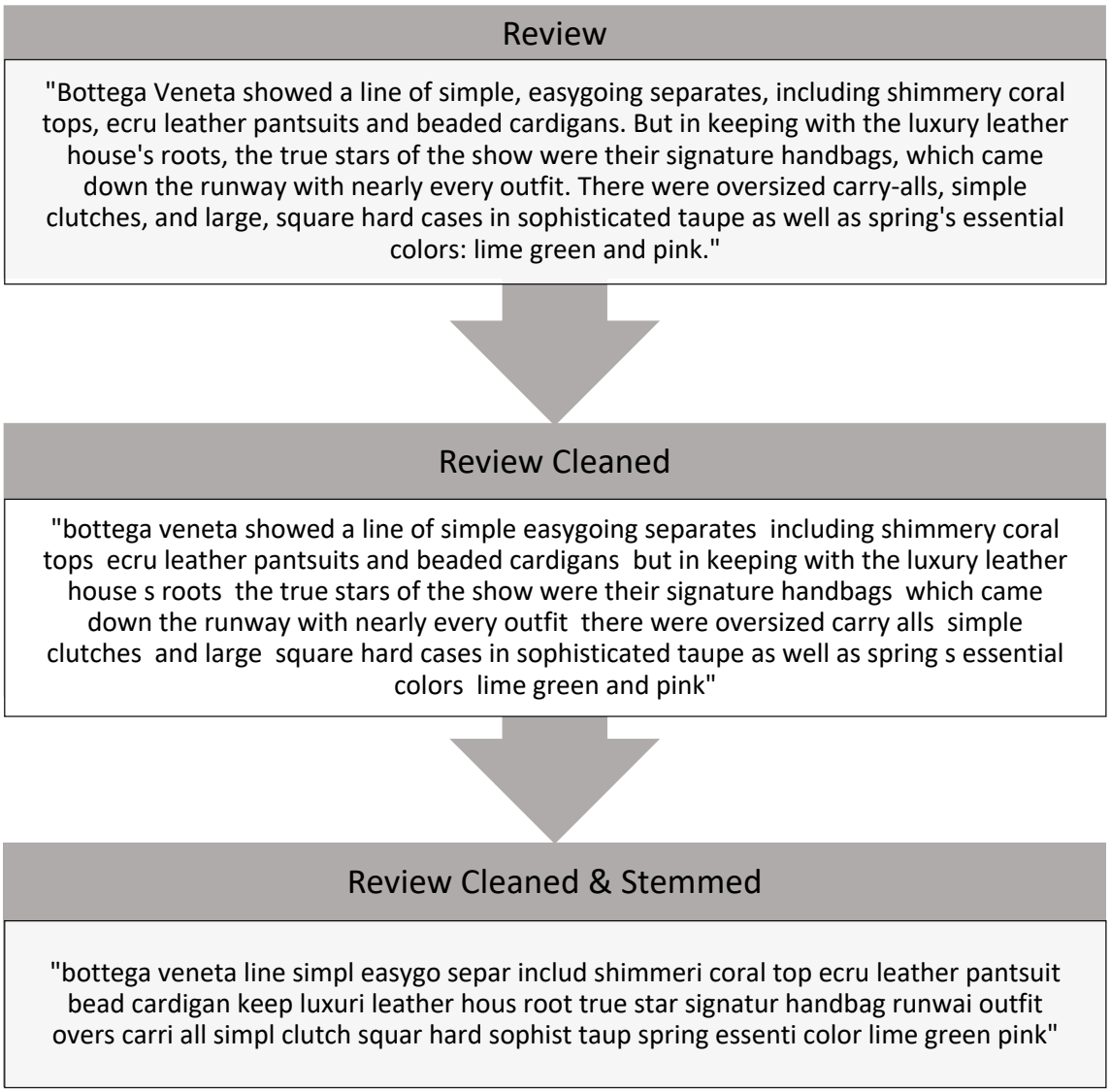

Figure 5: Example of Pre-Processing of a Style.com Review

4. Methods

The research will be done through text/statistical analysis and summarizing methodologies. This includes Latent Dirichlet Allocation (LDA) as well as sentiment analysis. These will be implemented through R, using packages *topicmodels* and *textmineR*, as well as *sentimentr*, respectively. The results from the sentiment analysis will then be compared for all the variables in the data set such as city, gender, and author.

4.1 Latent Dirichlet Allocation

Firstly, LDA will be performed to identify the underlying themes or concepts that are being discussed as it can take an input of a large set of documents and provide a way to organize and summarize the content of those documents. Furthermore, LDA is also known for its positive track record and reliability (Wang, 2018). The topics uncovered using LDA models will give an indication of the overall essence of all reviews.

LDA is a generative probabilistic model for collections of discrete data where the items are modelled into a set of underlying topics (Blei et al., 2003). The topics are formed by identifying collections of words within the corpus with high co-occurrences, meaning that words in the document that occur often together belong to the same topic. The resulting output of the model includes the probability that the topics would appear in the document and the probability that the word appearing in a topic.

Here, the created corpus D is made up of the documents $\mathrm{w} = (\mathrm{w}_1, \ldots, \mathrm{w}_N)$. The document then contains N words. w_i then identifies each word, $\mathrm{w}_i \in \{1, \ldots, V\}$, where V is a dictionary containing the terms (Grün & Hornik ,2011). The steps are then that each word has a probability of occurring in each topic. This is the term distribution, β, which follows a Dirichlet distribution: $Dirichlet(\delta)$. This is a "multivariate generalization of the Beta distribution" (Lin, 2016). Each document then has a probability of occurring in a topic. This is the topic distribution θ which also follows a Dirichlet distribution: $Dirichlet(\alpha)$. Then, for each of the *N* words w_i, a topic z_i, is

selected. This follows a multinomial distribution: Multinomial(θ). Then a word is selected, also from a multinomial distribution, given that it occurs in a given topic z_i : $p(w_i|z_i, \beta)$.

The optimal number of topics must then be found. This is often done by measuring perplexity, or the equivalent to the geometric mean per-word likelihood. In this case, four other metrics will be used for computational reasons. These include “Griffiths2004", "CaoJuan2009", "Arun2010", and "Deveaud2014.” These metrics, found by their respective, namesake, authors are combined by Nikita in his package *ldatuning* (2016). The metrics include Griffiths and Steyvers (2004) which uses the Gibbs sampling algorithm. This sequentially changes the number of topics in a corpus and evaluates the consequences. Secondly, Cao Juan et al. (2009) selects the LDA model based on density. Thirdly, Arun et al. (2010) which measures symmetric KL-Divergence of salient distributions. Lastly, Deveaud et al. (2014) maximises the information divergence between all pairs of LDA’s topics in order to estimate the number of latent concepts.

For this LDA analysis, after the corpus is created during the data-preprocessing, a document-term matrix is then made. This is done by tokenizing the text and then creating a matrix with rows representing the documents and columns representing the terms (Blei et al., 2003). Secondly, the number of topics to be extracted from the matrix must be chosen. For this, all metrics will be calculated at once using the *ldatuning* package, this surpasses the training of multiple models to minimize computational time. The metrics used will include “Griffiths2004", "CaoJuan2009", "Arun2010", and "Deveaud2014.” Of these, to find the optimum number of topics, “Arun2010” and “CaoJuan2009” must be minimized, while “Deveaud2014” and “Griffiths2004” must be maximized (Nikita, 2016). This will be revealed by plotting the metrics against number of topics for a sequence from 2 to 15.

Once the ideal number of topics is determined, the LDA model will be trained with the document-term matrix and the number of topics as input. Furthermore, it will use 80% of the complete data set, the remaining 20% will then be used as the test set. The resulting output are then the most probable words for each topic.

4.2 Sentiment Analysis

After the general topics of the reviews are found, sentiment analysis is performed for further insights regarding the positivity, negativity, or neutrality of the reviews. This analysis is beneficial as it can identify and extract the tone from text data.

Sentiment analysis is the automated process of understanding the opinion, or sentiment, of a given text. The computer does this by adding the sentiment counts of the individual words into the sentiment score of the whole text (Kwartler, 2017). It does this by incorporating the chosen sentiment lexicon, depending on the desired output format, and comparing the words to the polarized words. For this study this would include retrieving sentiment scores and examining how they differ for the types of reviews, how they change over time, and how they compare to sales data. As the data is solely text, this method is unsupervised.

For this analysis, the sentiments of the cleaned but unstemmed reviews were calculated using the sentiment function. First the most used words and their sentiment, either positive or negative, will be visualized for initial insights. This will then be repeated for the entire review by calculating a total sentiment score. Furthermore, the *lexicon* and *syuzhet* packages will be used for the bing (Hu & Liu) and jockers dictionaries, respectively. As these two have both been previous defaults for the sentiment formula in the *sentimentr* package, they will be used as inputs and compared to see which offers the most valuable insights. The jockers dictionary includes 10,738 words and aims to incorporate emotional shifts into text. It has two classification polarity and intensity and scores each on a continuous range from -1 to +1 (Jockers, 2017). The bing dictionary included 6,874 words and uses solely polarity for classification. Additionally, its scale is not continuous but either positive (1) or negative (-1) (Hu & Liu, 2004).

4.3 Welch's T-Test and ANOVA

To answer the sub questions, such as "How do attitudes differ for reviews written about female- and male-designed collections?" or "Do seasons have an effect on reviews?" two sample

(independent samples) Welch's t-tests and ANOVA will be used to see whether there is a significant difference in sentiment score. A Welch t-test tests a null hypothesis to an alternative hypothesis. Here, H_0: $\mu_1 = \mu_2$ and H_a: $\mu_2 \neq \mu_1$. One would reject the null hypothesis if the t-statistic is larger than the critical value of the t distribution, given a certain degree of freedom (Ruxton, 2006). Here, $t > t_\alpha$. The t-statistic is calculated by the equation below:

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Where, $\bar{x}_1$ and, $\bar{x}_2$ are the mean of group 1 and 2, n_1 and n_2 the size of group 1 and 2, and s_1 and s_2 the standard deviation of group 1 and 2, respectively. The degrees of freedom for unequal variances is:

$$v = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)\text{^}2}{n_1 - 1} + \frac{(s_2^2/n_1)\text{^}2}{n_2 - 1}}$$

Welch tests have an advantage over regular independent sample t-tests as it relaxes the assumption of equal variances between the populations. However, normality, independence, and random sampling are other assumptions that must still hold for the reliability of the test (Ruxton, 2006). Since this is a relatively large sample size of 6489 observations, the normality assumption is not very critical. Secondly, independence holds as it is presumed that the reviews do not influence each other and that each is unique to the collection it is about. Random sampling also holds as the data included all reviews written on style.com.

ANOVA also works by testing a null and alternative hypothesis. However, since there is a greater number of groups tested, the hypotheses are different. Here, the null hypothesis is that

all the group means are equal, and the alternative hypothesis is that that at least one group has a mean that is significantly different from the others. The ANOVA coefficient, F, is then calculated. This is the mean sum of squares due to treatment (MST) divided by the mean sum of squares due to error (MSE) (Ståhle & Wold, 1989):

$$\mathrm{F} = \frac{\mathrm{MST}}{\mathrm{MSE}}$$

The MST represents the average amount of variation between the group means that is explained by the model, while the MSE represents the average amount of variation within each group that is not explained by the model. A larger MST, or smaller MSE, thus results in a large F-statistic. This indicates that the difference between the group means is statistically significant (Ståhle & Wold, 1989). The p-value is then calculated from this value and represents the probability of getting the result by chance if the means are equal, or if null hypothesis were true. The p-value is then compared to the chosen significance level. If the p-value is smaller than the chosen significance level, one must reject the null hypothesis.

For this analysis, first, two sample t-tests will be performed to compare the sentiment scores for the two different collection seasons and the two different designers' genders. The observations will be split to perform the tests and various other descriptive statistics, such as mean and standard error, will be calculated. After this, ANOVA will be performed to see whether review sentiment changes significantly depending on the city where the show is held as well as over time.

The assumptions of the ANOVA analysis must be met in order to maintain a valid ANOVA analysis. These include continuity of the dependent variable, two or more categorical independent groups, independence of observations, no significant outliers, and lastly a normal distribution of the dependent variable. Continuity holds as the sentiment score as it is continuous between -1 and 1. The independent groups are the four cities and the 14 separate years, so this assumption also holds. Independence of observations was also an assumption for the Welch's t-

tests, so this is assumed to hold. It is also assumed that there are no outliers as the sentiment score lies between -1 and 1. Lastly, the large number of data points, the normality assumption is relaxed.

4.4 Multiple Regression

To gain further insights of the directional effect that the variables have on review sentiment, a multiple OLS cross-time linear regressions will be performed. Here, the independent variables will include dummy variables for gender, season, city, and year.

Multiple linear regression is a statistical method that illustrates the relationship between a response variable and multiple independent variables. In multiple linear regression, the relationship between the response variable is modeled as a linear accumulation of the predictor variables as well as an error term that represents noise. The objective is to find the regression coefficients that give the least sum of squared errors when looking at the difference between the predicted values and the actual response values (Eberly, 2007). In this model, review sentiment is a function of gender, season, city, and year. The single-equation demand model, as seen in Equation 1, can estimate review sentiment, while controlling for gender, season, city, and year:

$$Y_{jkpt} = \beta_0 + \beta_j\left(Gender_{jkpt}\right) + \beta_k\left(Season_{jkpt}\right) + \beta_p(City_{jkpt}) + \beta_t(Year_{jkpt}) + \mathcal{E}_{jkpt} \quad (1)$$

Where Y is review sentiment score. Where j refers to gender, k refers to season, p refers to city, and t refers to year. Parameters β_x are coefficients for their respective dummy variables and ε is the error term.

The coefficients in the model represent the amount that the response variable is expected to change for a one-unit positive change in each predictor variable, holding all other predictor variables constant. In this case, the predictor variables are dummy variables, so the one-unit increase is simply the presence of said variable. The estimated regression coefficients are

calculated using least squares regression, where the sum of squared errors between the actual and predicted values are minimized.

4.5 Stock Prices

Lastly, to assess whether there is any correlation between sentiment score and company well-being, a visual analysis will be performed. Series graphs with average sentiment score, as well as average adjusted closing price will be created, and their variations compared. Furthermore, this will be refined by looking at three specific brands, namely Hermes, Louis Vuitton, and Christian Dior.

5. Results

The first analysis through LDA gives us an indication as to what the 'aspects' in the research question of, "What aspects of luxury fashion collections and shows have a positive or negative influence on journalistic reviews?" refers to. This is done by identifying latent topics in the reviews.

For the analysis, first the number of topics for the model were chosen by comparing various metrics. *Figure 6* illustrates a plot that includes the metrics for picking the number of topics in the LDA analysis. These are "Griffiths2004", "CaoJuan2009", "Arun2010", and "Deveaud2014." Although the lines do not perfectly minimize and maximize at the same number of topics, it can be seen that around 13 topics, all metrics are relatively optimized, thus this will be the number of topics chosen for this LDA analysis.

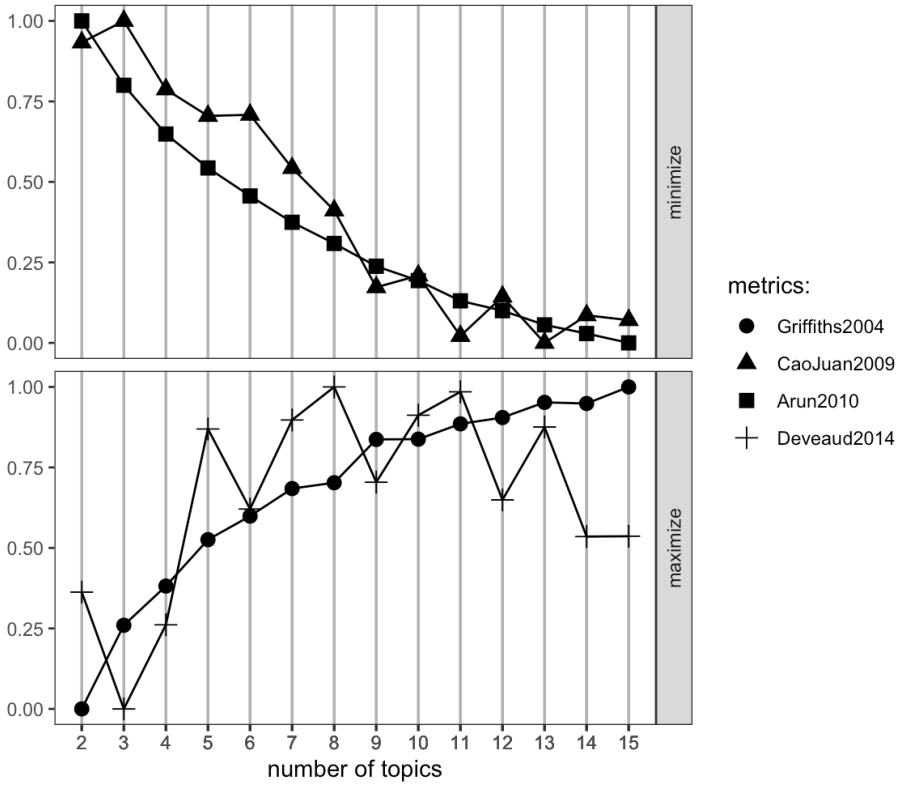

Figure 6: LDA Metrics for Number of Topics Selection

Figure 7 illustrates the prevalence of each topic and appears to be proportional to alpha, with a linear pattern and a slope nearing 1. This an important performance measure for LDA (Reisenbichler & Reutterer, 2019). Two seeming outliers appear at the top right of the plot around 20% and 25% topic prevalence. *Figure 8* is a histogram for maximum topic probability. Here, the distribution seems to skew right. Meaning that the majority of observations are on the lower side of maximum topic probability, namely around the peak of about 0.4.

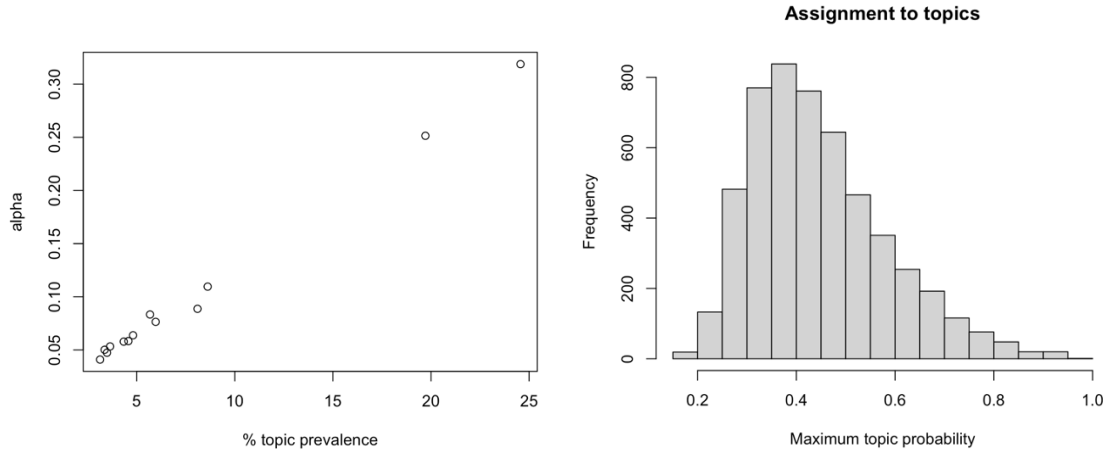

Figure 7 & 8: LDA Prevalence of each Topic and LDA Assignments to Topic, respectively

Table 1 shows the summary statistics for the topics. The 13 topics have their respective label, prevalence, coherence, top term topic, and top terms document. The latter two variables refer to the *top_terms_phi* and *top_terms_gamma* variables, respectively. The phi summarizes the topics by listing the most probable words for each topic, based on the topic-term distribution. Meaning that the words in each topic are ranked based on their probability of being associated with that respective topic (Arun et, al., 2010). The gamma lists the most probable documents for each topic, and summarizes the topics that way, based on the document-topic distribution. This shows which documents are most relevant to each topic.

The LDA analysis performed shows that the topics are: jacob, de, print, girl, fashion, fashion, jacket, brand, dress, print, fashion, model, and collect. In term of prevalence, topic 9 “dress” has the highest score with 24.6. This is followed by jacket with 19.7, topic 5 “fashion” with 8.62. 13 “collect” score of 8.1. This means that these topics have the greatest proportion of words in the given document collection made by the LDA model (Blei et al., 2003).

The “dress” topic included most probable terms such as dress, black, skirt, white, and collect. Most probable document terms include other designer/brand names like dell, mendel, acqua, herrera, and Lhuillier. Topic 7 is “jacket” and includes top terms like jacket, leather, coat, collect, and design. Designers included karan, kor, panichgul, pillov, and dkny. The fifth topic, “fashion,” has top topic terms like dress, fashion, collect, jacket, and design. The brands in the top document

terms include, Chalayan, ghesqui, kokosalaki, wantanab, and theysken. “Collect” is the 13th topic and includes the top terms collect, design, dress, season, and cloth. Brands included in the top document terms are azria, costa, mouret, bartlett, and maier.

Table 1. LDA Summary Statistics

Topic	Label	Prev.	Coherence	Top Terms Topic	Top Terms Document
1	jacob	3.64	0.137	jacob, girl, dress, marc, model	jacob, caval, furstenberg, johnson, vuitton
2	de	4.34	0.019	design, dress, de, collect, la	som, renta, basso, blass, pilotto
3	print	4.81	0.012	print, dress, collect, girl, short	smith, mccartnei, miu, philo, walker
4	girl	4.58	0.009	girl, print, dress, leather, jacket	missoni, lim, rykiel, taylor, giannini
5	fashion	8.62	0.004	dress,fashion, collect, jacket, design	chalayan, ghesqui, kokosalaki, watanab, theysken
6	fashion	3.36	0.016	design, fashion, collect, london, mcqueen	mcqueen, kane, yamamoto, margiela, saunder
7	jacket	19.7	0.022	jacket, leather, coat, collect, design	karan, kor, panichgul, pullov, dkny
8	brand	3.49	0.029	design, brand, collect, label, fashion	macdonald, deacon, burberri, ungaro, beckham
9	dress	24.6	0.022	dress, black, skirt, white, collect	dell, mendel, acqua, herrera, lhuillier
10	print	5.97	0.016	print, collect, color, dress, design	williamson, pucci, cornejo, baptista, dunda

11	fashion	5.68	0.023	collect, fashion, cloth, design, world	lagerfeld, owen, lemair, marra, rick
12	model	3.13	0.018	model, cloth, design, armani, fashion	armani, burch, jensen, viktor, rolf
13	collect	8.1	0.005	collect, design, dress, season, cloth	azria, costa, mouret, bartlett, maier

Notes: Prev. refers to prevalence

An extension of the "Top Terms Topic" in *Table 1* is shown in *Figure 9* where the top ten terms of each topic are plotted from highest probability at the top to lowest. The terms discussed previously range in probability from around 0.005 to 0.04. The highest probabilities seem to be unevenly large. These appear for the term "dress" in topic dress and "print" in the print topic.

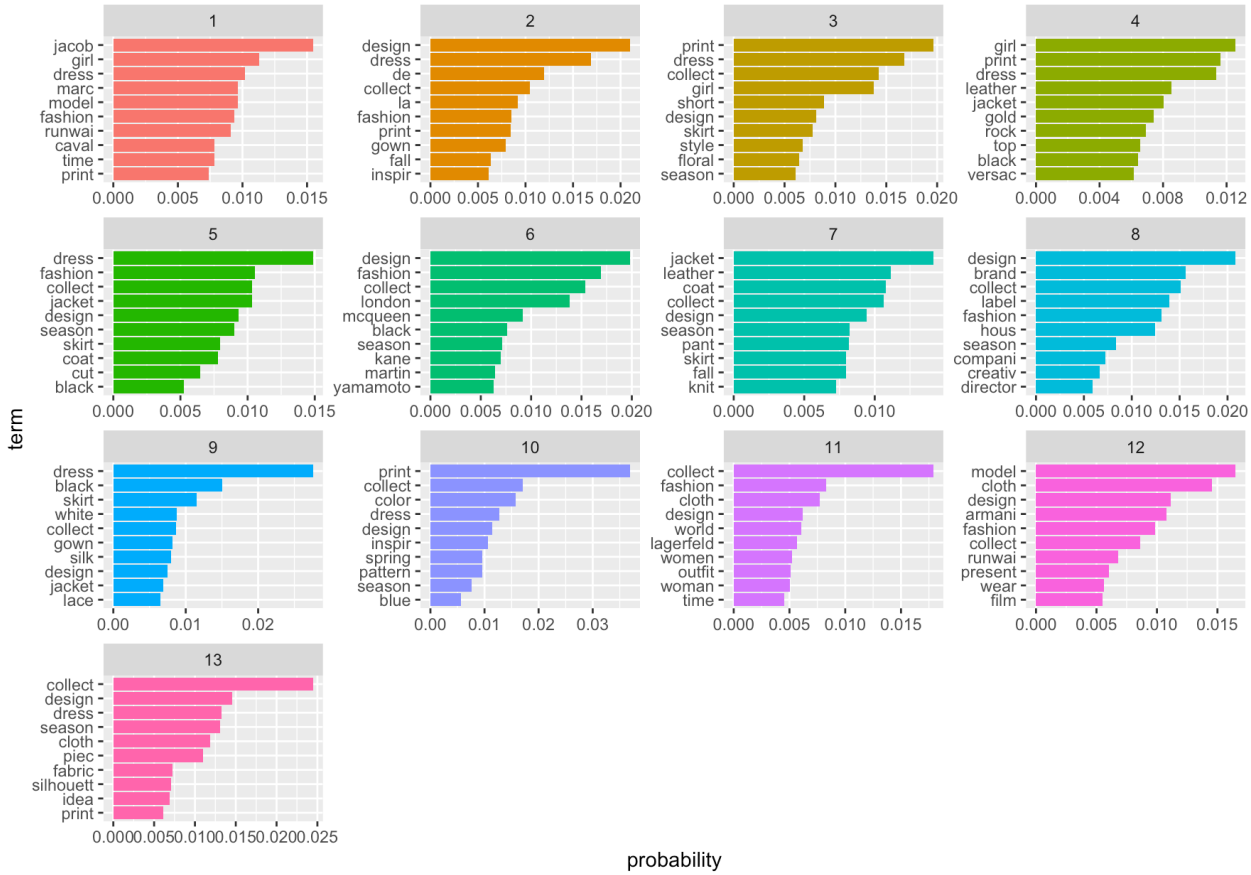

Figure 9: Plots Displaying Top 10 Terms for each Topic in the LDA Model

5.1 Hypothesis 1

To answer the first hypothesis of: "Fashion show reviews with words associated with previous fashion trends will have a lower sentiment than those reflecting those of the current fashion cycle," we look at the results of our sentiment analysis.

Figure 10 and *Figure 11* show frequencies for words that occur more than 400 times within all reviews and categorize them between positive and negative sentiment scores. *Figure 10* and *Figure 12* use the jockers and bing dictionaries, respectively. It can be seen that the Jockers dictionary has a greater array of words in its library as there is a higher number of words that have a frequency over 400.

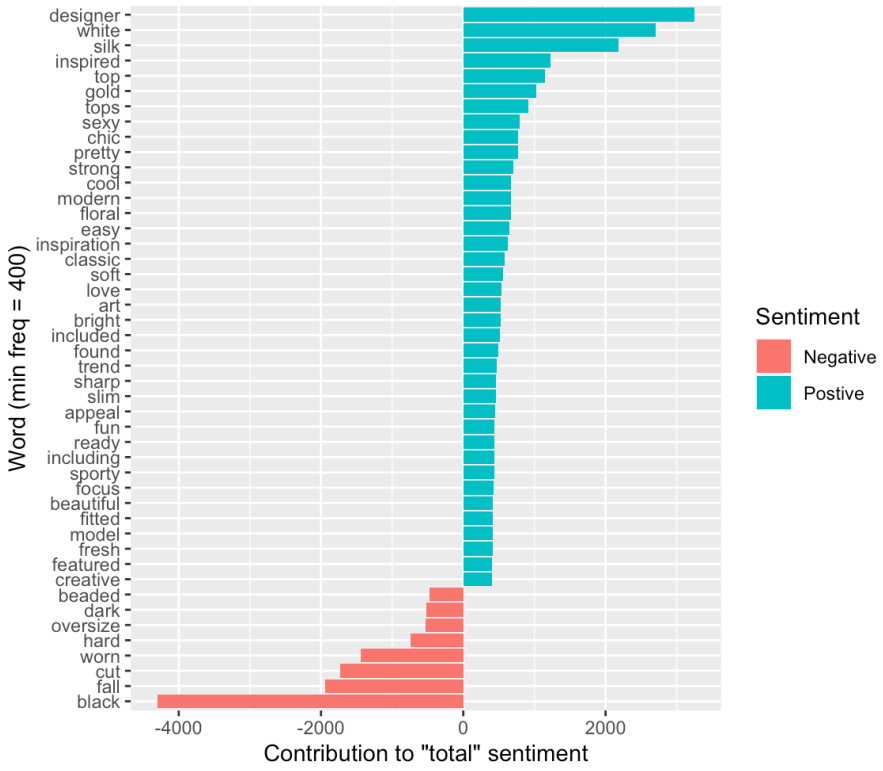

Figure 10: Frequency of Positive/Negative Words using Jockers Dictionary

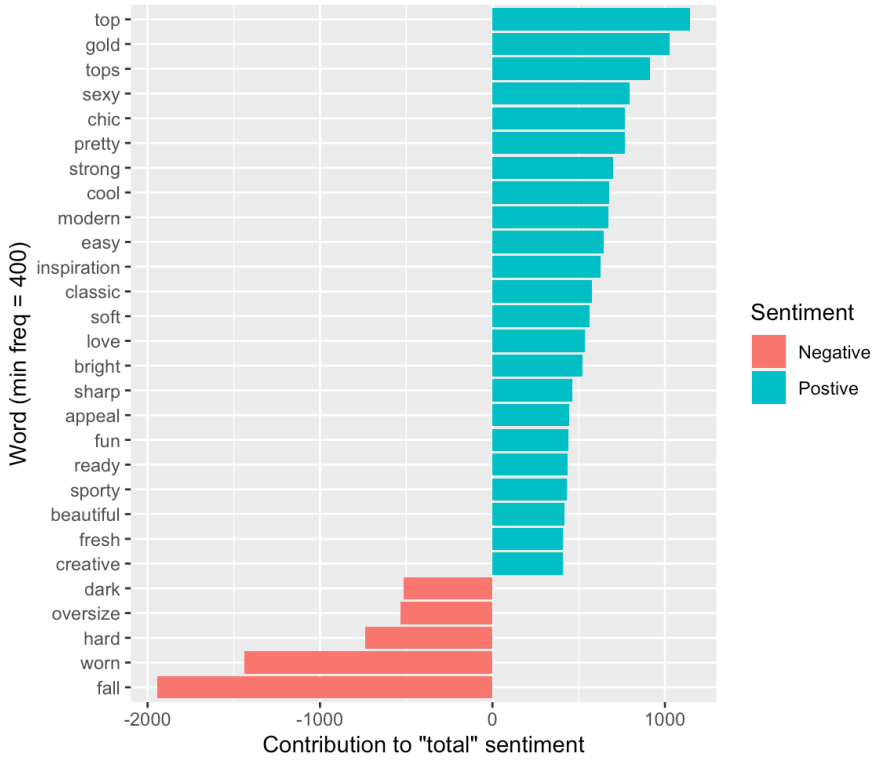

Figure 11: Frequency of Positive/Negative Words using Bing Dictionary

In terms of fashion cycles, some of the positive words associated with the 1990's fashion cycles include "chic", "sporty," and "strong" which coincide with the minimalist, streetwear, and androgynous trends of the 1990's. Additionally, words such as "white," "sexy," "pretty," "classic," "dark,", and "black" can be categorized within either decade.

However, words such as "slim" and "fitted" specifically reflect the 2000's trend of skinny and form-fitting clothing. Additionally, "silk" has a high positive contribution to the total sentiment and is considered a 2000's staple (Donohue, 2022). Lastly, the bohemian trend is reflected in words such as "floral," "easy," and "soft."

Majority of the negatively classified words that coincide with 1990's trends such as grunge and streetwear. These include "oversize," "hard, "worn," and "cut." One negatively classified word, "beaded," seems to reflect the 2000's trend of glamour which saw the return of embellishments. However, this word has the lowest contribution to total sentiment of all the negative words.

Thus, it can be that there are more negatively classified words associated with 1990's trends, and that their total negative contribution is greater than words associated with 2000's trends. Furthermore, there are more positively classified words associated with 2000's trends and their total contribution to the positively classified words is greater than those associated with 1990's trends. Thus, hypothesis 1 holds: fashion show reviews with words associated with previous fashion trends seem to have a lower sentiment than those reflecting those of the current fashion cycle.

Further insights that the sentiment analysis has given are shown in *Figure 12.* Here, average sentiment score per year is plotted over the 15 years in the data set. It will later be tested whether year has a significant effect on sentiment score, but this graph can offer a preliminary visual analysis. Mean sentiment score has a steady decrease from around 0.250 till the year 2003, where it seems to rise again, until lowering to its minimum in 2006 and 2007. This is followed by a slow increase back up to a sentiment score of around 0.175. This, and all further analyses, will use the bing dictionary as it offers a more concise and explicit terms which is useful for interpretation and computational efficiency.

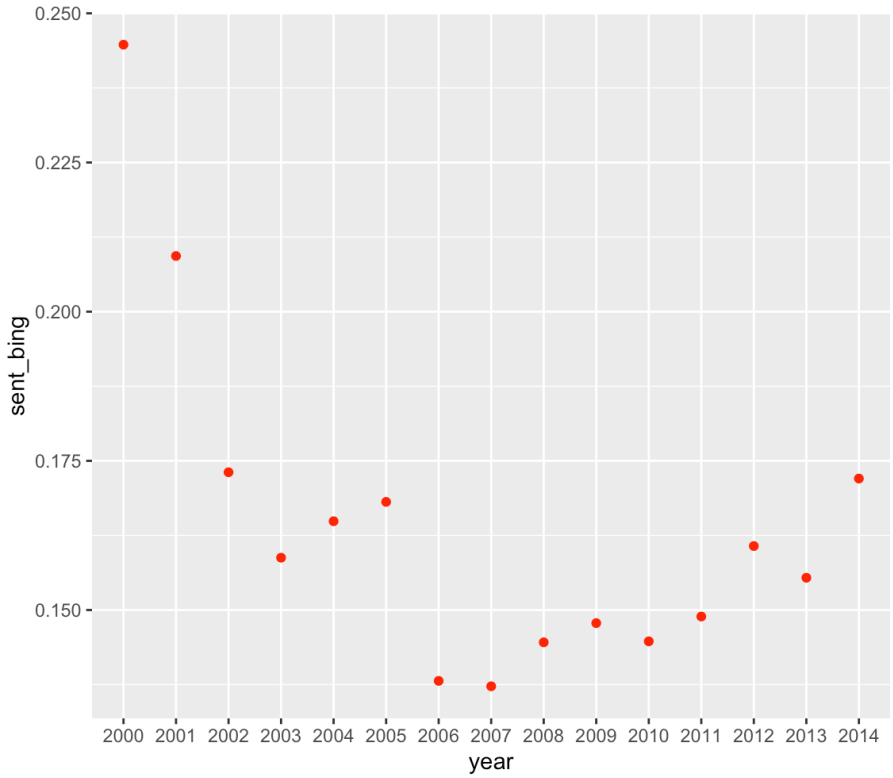

Figure 12: Mean Sentiment Score per Year using Bing Lexicon

5.2 Hypothesis 2

To answer the second hypothesis of: "Reviews on collections by female founded fashion houses have lower overall sentiment than those by their male counterparts," the results of the Welch t-test must be observed.

Table 1 represents the results of a Welch t-test on the gender sentiment. The aim is to find out whether the means of the two groups, 0.156 for male and 0.164 for female, that differ in the dataset, are also differing in real life. In *Table 1*, it can be seen that the p-value of the t-test that compares our two given means is equal to 0.000, which is a highly significant result. We can thus say with 99% confidence that the mean sentiment score of reviews is indeed different between males and females. The null hypothesis is rejected with 99% confidence. Although the reviews are found to be significantly different for the genders, the second hypothesis does not hold as female founded fashion houses have a significantly higher mean than those of their male counter

parts. Thus, reviews on collections by female founded fashion houses have do not have a lower overall sentiment than those by their male counterparts

Table 1. Welch Two-Sample t-test on Gender Sentiment

	N	Mean	Std. Error	Std. Dev	Conf. Int.		T-crit	df	P-value
Male	3836	0.156	0.002	0.095					
Female	2653	0.164	0.002	0.096					
Difference		H_0: diff = 0			-0.012	-0.003	-3.174	5678.1	.000***

Notes: Asterisks indicate significance levels for the p-value, where *: p<0.1, **: p<0.05, and ***: p<0.01. A 95% confidence level is used for the confidence interal. Df refers to degrees of freedom

5.3 Hypothesis 3

The third hypothesis compares sentiment scores of the review written in opposing seasons. Similarly to hypothesis 2, a Welch t-test was again done. *Table 2* shows results of a Welch t-test on the seasonal sentiment. The aim is to find out whether the means of the two groups, 0.155 for Fall/Winter and 0.164 for Spring/Summer, that differ in the dataset, are also differing in real life. The p-value, of the Welch t-test that compares the means, is equal to 0.000, which is a highly significant result. We can thus say with 99% confidence that the mean sentiment score is indeed different between Spring/Summer and Fall/Winter reviews. The null hypothesis is rejected with 99% confidence. Although there appears to be a significant difference in sentiments for the two seasons, the third hypothesis of: “Reviews written for the Spring/Summer collections will have a lower sentiment than those written for the Fall/Winter shows” does not hold. The mean sentiment score for Fall/Winter is significantly lower than that of Spring Summer.

Table 2. Welch Two-Sample t-test on Season Sentiment

	N	Mean	Std. Error	Std. Dev	Conf. Int.		T-crit	df	P-value
Fall/Winter	3354	0.155	0.002	0.097					
Spring/Summer	3135	0.164	0.002	0. 094					
Difference		H_0: diff = 0			-0.014	-0.005	-3.174	6477.1	.000***

Notes: Asterisks indicate significance levels for the p-value, where *: p<0.1, **: p<0.05, and ***: p<0.01. A 95% confidence level is used for the confidence interal. Df refers to degrees of freedom

5.4 Hypothesis 4

After Welch t-tests are done to see whether gender and season influence the sentiment of reviews, ANOVA tests and multiple linear regression is used to see whether these differences continue when looking at different year and cities.

Hypothesis 4 looks at whether sentiments significantly differ between years and attempts to pin-point a cause for the potential differences. Firstly, an ANOVA test is done to see whether one or more of the years differs significantly. *Table 3* gives us the results of this ANOVA test. The F-value of the ANOVA test is equal to 22.95, which is very high and as a result gives us a p-value of almost 0. This means that we can reject the null hypothesis of equal means with a 99% confidence and say that at least one of the 14 cities has a significantly differing mean sentiment score.

Table 3. ANOVA on Yearly Sentiment

	df	Sum sq.	Mean sq.	F-value	P-value
Year	14	2.81	0.200	22.95	.000***
Residuals	6474	56.53	0.009		

Notes: Asterisks indicate significance levels for the p-value, where *: p<0.1, **: p<0.05, and ***: p<0.01. Df refers to degrees of freedom

Although it gives an indication, it does not give sufficient evidence for whether the hypothesis holds. For this, a multiple regression is done to pin-point which years in particular have a significantly different mean sentiment score.

Table 4 depicts the results of our multiple linear regression with the gender, cities and years as dummies. The constant of the regression is the sentiment score when all the dummies take on the value 0, which means that the runway has taken place in London, with a female designer for a Fall/Winter collection releasing in 200. We can see in *Table 6* that most of the dummies are significant to the 99% confidence level. For the years, each year leads to a decrease in sentiment score ranging between 0.035 and 0.111, depending on the specific year, compared to the base year of 2000 when all else is kept constant. Although all years are statically significant, it can be seen that the most impactful years are 2007, 2008/2010, 2009, and 2011, all with coefficients more negative than 0.101.

Based on these findings, the fourth hypothesis of, "Market disruptions such as the September 11 attacks or the Global Financial Crisis have a negative effect on journalistic reviews," holds. Although there is no overwhelming evidence that the September 11 attacks have a distinctive impact on review sentiment, the case for the Global Financial Crisis of 2008 seems to differ. Here, there is a cluster of significant and pronounced coefficients that negatively impact review sentiment.

Table 4. Multiple regression of gender, season, city, and year on review sentiment

Linear Regression	**Coeff.**	**Std. Error**	**t-value**	**p-value**
Male	-0.003	0.002	-1.186	0.236
Spring	0.009	0.002	4.025	0.000***
Milan	-0.004	0.004	-1.007	0.314
New York	0.032	0.004	9.04	0.000***
Paris	-0.007	0.004	-1.716	0.086
Year				
2001	-0.035	0.009	-3.786	0.000***

2002	-0.07	0.009	-7.684	0.000***
2003	-0.086	0.009	-9.42	0.000***
2004	-0.08	0.009	-8.86	0.000***
2005	-0.079	0.009	-8.95	0.000***
2006	-0.11	0.008	-13.126	0.000***
2007	-0.111	0.008	-13.677	0.000***
2008	-0.105	0.008	-13.189	0.000***
2009	-0.102	0.008	-13.05	0.000***
2010	-0.105	0.008	-13.622	0.000***
2011	-0.101	0.008	-13.367	0.000***
2012	-0.09	0.007	-12.016	0.000***
2013	-0.095	0.007	-12.87	0.000***
2014	-0.078	0.007	-10.703	0.000***
Constant	0.234	0.007	31.255	< 2e-16

Notes: Asterisks indicate significance levels for the p-value, where *: p<0.1, **: p<0.05, and ***: p<0.01

After running the regression analysis, it must be checked if the model works well for the data. Figure 13 includes diagnostic plots for the previous regression model.

The first plot shows whether the residuals have non-linear patterns, as this could mean there is a non-linear relationship between the dependent variables and the independent variable that the model doesn't capture. In this case, it seems to be an equal spread of residuals around a horizontal line without distinct patterns, meaning that this is a good indication that there aren't any non-linear relationships.

The second plot indicates whether the residuals are normally distributed. It can be seen that, for this model, they are as they line up well on the straight dashed line. Meaning that the residuals appear to be normally distributed.

The Scale-Location plot can be seen on the bottom left. This graph checks the homoscedasticity assumption by showing if the residuals are equally spread out along the range

of predictors. In this case, the residuals appear randomly spread and the red smooth line is close to horizontal, so the variances can be assumed as equal.

Lastly, the residuals vs leverage plot helps to find influential cases by looking out for outlying values at the upper or lower right corner. These locations are where cases can be influential as they gave high Cook's distance scores. Meaning that they are influential to the regression results (Bommae, 2015). In this case, Cook's distance lines cannot be seen, and it appears that all cases inside the Cook's distance lines.

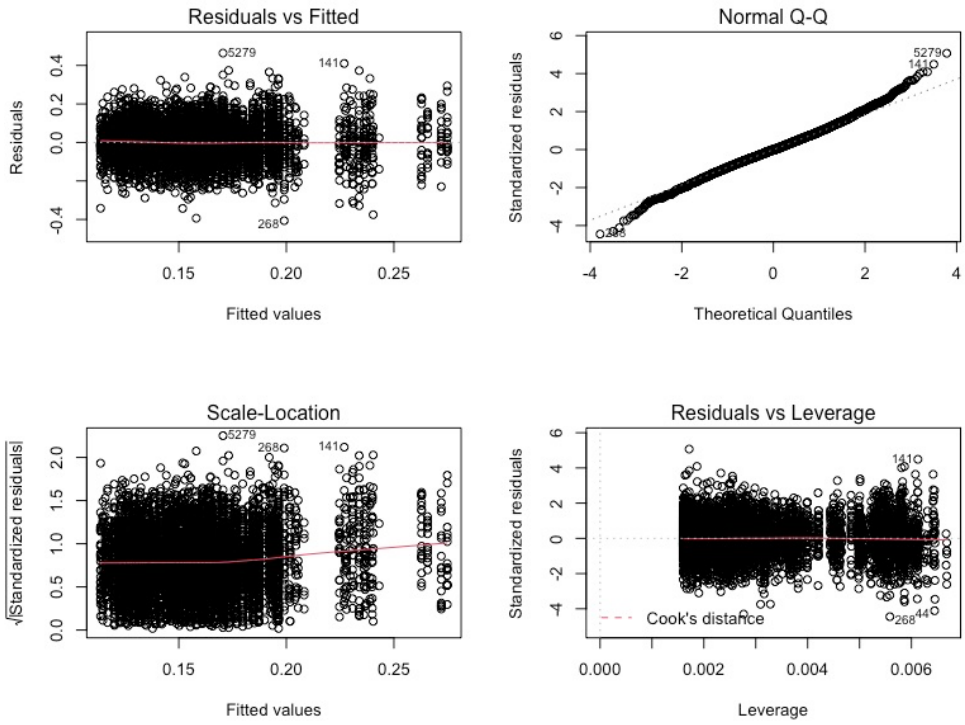

Figure 13: Multiple Regression Diagnostic Plots

5.5 Hypothesis 5

The ANOVA method is repeated to test the fifth hypothesis, namely "Reviews on collections shown in Paris have a higher sentiment than those shown in New York, London, or Milan". Here the effect of show location is of importance and the ANOVA tests whether either one of the four

cities has a different mean. *Table 5* gives us the results for the test. The F-value of the ANOVA test is equal to 68.37, which is very high, and the resulting p-value of close to 0. This means that we can reject the null hypothesis of equal means with a 99% confidence and say that at least one of the cities has an unequal mean.

Although this points into the direction of the hypothesis, it is still unknown whether Paris, in particular, has a positive effect on average sentiment. For this, *Table 4* must be referred to again. It can also be seen that the location of the runway show is statistically important, only when the runway takes place in New York. Specifically, keeping all else constant, when a show takes place in New York the review, on average, has a sentiment score of 0.032 points higher. However, the coefficients for Milan and Paris are not found to be significant, meaning that a show happening in Milan or Paris will not cause a significant difference in sentiment scores compared to the show taking place in London. This means that the fifth hypothesis does not hold and that reviews on collections shown in Paris do not appear to have a higher sentiment than those shown in New York, London, or Milan.

Table 5. ANOVA on City Sentiment

	df	**Sum sq.**	**Mean sq.**	**F-value**	**P-value**
City	3	1.82	0.606	68.37	.000***
Residuals	6485	57.52	0.009		

Notes: Asterisks indicate significance levels for the p-value, where *: p<0.1, **: p<0.05, and ***: p<0.01. Df refers to degrees of freedom.

5.6 Hypothesis 6

Lastly, it must be seen whether is beneficial for luxury fashion brands to influence the sentiment scores through the discussed mechanisms. To do this, there must a correlation between sentiment score and company well-being. *Figures 14* and *15* display the mean sentiment score as well as the adjusted closing price per year, respectively. Although the y-axes are unequal and are scaled differently, the main pattern of the scatterplot is of most importance. Both

variables seem to decrease in value from the year 2000 till 2003. After 2003, adjusted closing price rises till around 2008 while sentiment only does so until 2006. This delay, however, might be explained by the difference in time between fashion shows and the release of the collection. After their respective declines, both variables seem to rise again, with adjusted closing price rising exponentially up to almost 150 dollars. These findings are in-line with the sixth hypothesis of: "Sales, as reflected by respective stock prices, mirror the review sentiment score trends." To further analyse, a more detailed comparison can be done.

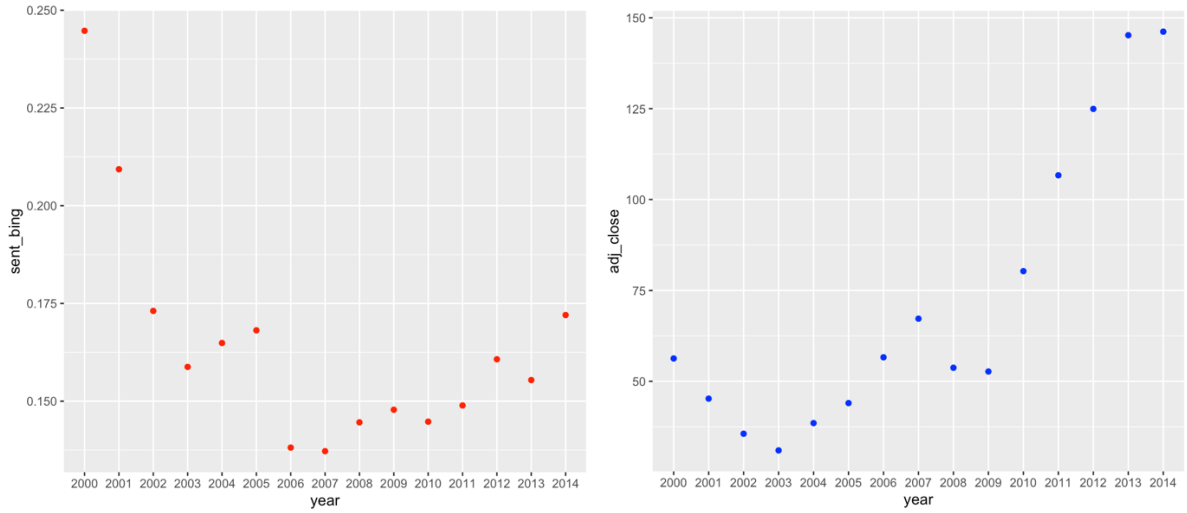

Figure 14 & 15: Mean Sentiment Score (Bing Lexicon) and Adjusted Closing Price per Year, respectively

Figures 16 and *17* shows a more in-depth version of the previous two figures. Here, the mean sentiment score and adjusted closing prices are separated for 3 luxury fashion houses, namely Christian Dior, Hermes, and Louis Vuitton. At first glance, the side-by-side figures look quite different. The sentiment scores seem to be much more unstable while the stock prices for all brands seem to increase steadily until around 2008. This could be an indication of the effect of the 2008 global stock market crash. The reviews, however, do not seem to decrease during this time. Some similarities can be seen where Hermes and Louis Vuitton seem to both steadily increase in both variables after 2008. Additionally, Louis Vuitton sees a decrease in both sentiment and stock prices around 2002 and 2003. Evidence from all four graphs lead to the

conclusion that the sixth hypothesis holds and that sales, as reflected by respective stock prices, do mirror the review sentiment score trends.

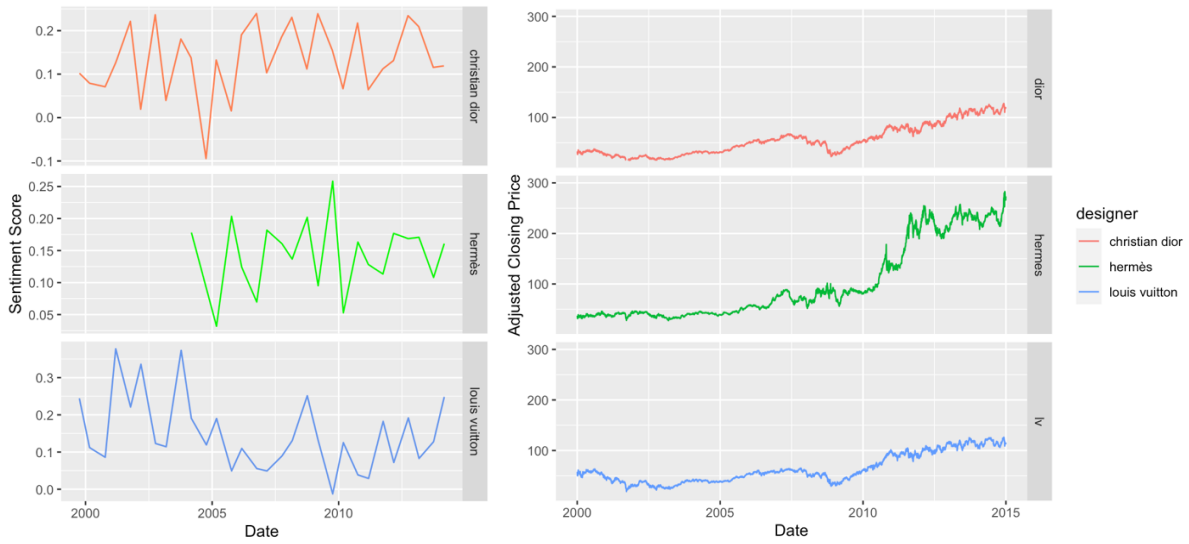

Figure 16 & 17: Mean Sentiment Score (Bing Lexicon) and Adjusted Closing Price per Year, respectively for Christian Dior, Hermes, and Louis Vuitton.

6. Discussion

6.1 Conclusion

This paper investigated fashion show reviews in order to identify several factors that influence their positivity/negativity. Moreover, examined whether these reviews have a significant impact on fashion brands' well-being. The findings indicate that fashion cycles, gender of the designer, which season the collection is designer for, which city the collection is shown in, and which year the collection is designer for are all aspects that significantly influence the positivity/negativity of fashion show reviews.

Particularly, fashion cycles affected reviews as the negative words in the reviews were more often associated with the trends of the previous cycle while the positive words were attributed with the current fashion cycle. This was in line with the first hypothesis.

The second hypothesis, however, did not hold. Here, it was predicted that reviews on collections by female founded fashion houses have lower overall sentiment than those by their male counterparts. The results of the Welch t-test and multiple regression conversely found that female designed collections actually have a higher mean sentiment score. Although there are various reasons that could explain this occurrence, one reason that is specific to this study is that the method of classifying gender for brands led to an overestimation of fashion brands that have a woman at the helm.

Similarly, the third hypothesis also did not hold. The thought behind this was that collections shown in fall would have a lower sentiment due to the effect of fewer hours light and sun exposure on mood and job satisfaction. However, the findings show that collections shown in the fall actually have a higher average sentiment. An explanation could be the opposing force, where clothes shown for the Spring/Summer collections have a brighter colour palette and thus could leave journalist with a more positive outlook.

The year a collection was designer for was also found to be a significant determinant of sentiment. The fourth hypothesis predicted that market disruptions such as the September 11 attacks or the Global Financial Crisis have a negative effect on journalistic reviews. The coefficient cluster around 2008 were significant and pronounced with a negative impact on review sentiment.

The last mechanism tested was the effect of show location on reviews. Here, only New York was found to have a significant impact. This is not in line with the fifth hypothesis which predicted that collections shown in Paris have a higher sentiment than the other three cities. A reason for this could be that there are a disproportionate number of reviews for each of the cities, with New York having nearly double the number of reviews compared to Paris and nearly triple when compared to London and Milan.

Finally, it was tested whether sentiment had any correlation with company well-being. When comparing scatterplots for mean sentiment reviews and fashion brands' mean adjusted closing

price, evidence was found that sales, as reflected by respective stock prices, mirror the review sentiment score trends.

The findings give evidence for luxury fashion brands to influence the studied aspects of their fashion shows to improve financial well-being. Namely, it would first be advised for said brands to study and design in line with the trends of the current decade. Secondly, it would benefit their reviews to hire female designers as creative directors. Thirdly, if they were to show only once a year, it would be recommended to show a collection during the Spring/Summer fashion month. Finally, it would be advised to show collections in New York to have a significant positive effect on reviews.

Furthermore, this study has expanded on the current research of fashion shows by introducing machine learning methods. Additionally, it has tested the methods on a new data set, presenting new performance evaluations and providing evidence for method improvement.

6.2 Limitations

There are a several limitations of this study. Regarding the data, only reviews taken from style.com were used. Although this is the most extensive data set on fashion shows available, it does not reflect all opinions neither does it contain a review for every single fashion show. This could lead to a decrease in external validity when extrapolating these methods and results to additional data. Additionally, not all fashion brands had the same number of reviews, meaning that word frequency when it comes to brand names could be skewed.

Another limitation is the lack of extensive, and available, data on luxury fashion brand sales. This led to the use adjusted closing price as an indication for sales, but this could differ from actual sales data as stock prices are affected by more than solely sales. However, both are seen as an accurate measure of company well-being.

Computational time was also an issue during the research. Primarily for the LDA analysis, where the document term matrix had a size of over 4 GB, when testing for the number of topics to choose, this would often take over 20 minutes or would cause the entire R-studio application

to crash. This could be solved, however, through the use of a smaller sample or a more advanced computer with more computational power.

An additional issue with the used data is the errors in grammar. This leads to an inaccurate count in the frequency of words which is important for LDA and well as Sentiment Analysis. Although this is a possibility, it would be unfeasible to go through all of the reviews to check for errors as there are over six thousand reviews of which each review can have a length of over a hundred words.

Lastly, the gender variable created categorized the designers as female given that the brand included at least one female founder. This was done for computational reasons as some brands included multiple founders. However, this does give a slight overestimation fashion brands that have a woman at the helm. Furthermore, the designers at fashion brands are not constant and are often cycles throughout the years. Thus, categorizing the designers as male or female for the complete 14 years of this data set might not be completely accurate to analyze what the designer's gender is at the time of the review. However, if one were to manually identify every amendment in creative director for every single brand, seeing that there is not available data set, this would take an enormous number of hours.

Future research could be done with more extensive and improved data, when available, in order to receive more managerial insight. With more current data, it would be interesting to see how COVID-19, for example, has affected both reviews and sales. Additionally, other variables such as proportions of minority and plus size models, or where the journalist was seated, might be variables of interest.

Using social media review data would also be an interesting study. This might offer more valuable insights the fashion brands as this would include a much greater number of reviews as well as direct opinions of their customers. Here, emoticons could also play a key role, thus using the correct dictionary would be vital in order to accurately transform them into text and looking at their adjacent sentiment. Additionally, research could be done on this topic but focussing more on set design as there has been a shift in theatricality and gimmicks during fashion shows in attempts to create 'viral' moments. Moreover, as designers have shifted into spokespeople of

the brands they design for, it would be interesting to see how personal situations of the designers affect brand reviews and sales. Some examples of this could be John Galliano who was at Christian Dior when he was arrested for anti-sematic remarks, as well as Demna Gvasalia who is currently being accused of inappropriate use of children in his advertisements for Balenciaga.

Using extended methods could also offer beneficial improvements. Sentence- or topic-based LDA and sentiment analysis could be performed to get more specific results as it looks at a more micro scale. Additionally, methods that look at word embeddings might also offer insight as it would allow to see the meaning of the text. For this the Global Vectors for Word Representation model could be used. Lastly, an event study could also be done to improve the findings related to hypothesis 6, or the comparison between sentiment score and stock prices. This could target specific dates, such as the release of a review article, and see whether they had an effect on the change in stock prices.

7. References

Ale Chilet, Jorge; Chen, Cuicui; Lin, Yusan, 2017, "Runway Reviews and Instagram Activities of High-End Fashion Brands", https://doi.org/10.7910/DVN/BCOSKY, Harvard Dataverse, V1

Amado, A., Cortez, P., Rita, P., & Moro, S. (2018). Research trends on Big Data in Marketing: A text mining and topic modeling based literature analysis. *European Research on Management and Business Economics*, *24*(1), 1-7.

Arnold, R. (2018,). The Evolution of Fashion Capitals: Why Paris Reigns Supreme. The Business of Fashion.

Arun, R., Suresh, V., Veni Madhavan, C. E., & Narasimha Murthy, M. N. (2010). On finding the natural number of topics with latent dirichlet allocation: Some observations. In *Advances in Knowledge Discovery and Data Mining: 14th Pacific-Asia Conference, PAKDD 2010, Hyderabad, India, June 21-24, 2010. Proceedings. Part I 14* (pp. 391-402). Springer Berlin Heidelberg.

Berger, J., Humphreys, A., Ludwig, S., Moe, W. W., Netzer, O., & Schweidel, D. A. (2020). Uniting the tribes: Using text for marketing insight. *Journal of Marketing*, *84*(1), 1-25.

Bobb, B. (2019). *"that's not a plane, that's a hair dryer!": Anna Wintour and Donatella Versace on rafting and riding horses together*. Vogue. Retrieved from https://www.vogue.com/slideshow/anna-wintour-donatella-versace-forces-of-fashion-2019

Bommae, Kim (2015) Diagnostic Plots. University of Virginia Library. Diagnostic Retrieved from https://data.library.virginia.edu/diagnostic-plots/

Büschken, J., & Allenby, G. M. (2016). Sentence-based text analysis for customer reviews. *Marketing Science*, *35*(6), 953-975.

Cao Juan, Xia Tian, Li Jintao, Zhang Yongdong, and Tang Sheng. 2009. A density-based method for adaptive lda model selection. *Neurocomputing — 16th European Symposium on Artificial Neural Networks 2008* 72, 7–9: 1775–1781. http://doi.org/10.1016/j.neucom.2008.06.011

Clinton, L. M. (2015). *Karl Lagerfeld gave Anna Wintour the ultimate BFF gift*. Glamour. Retrieved from https://www.glamour.com/story/karl-lagerfeld-anna-wintour-friends

Council of Fashion Designers of America. (2014). Impact of Fashion Critic Reviews on Designer Sales.

Deveaud, R., SanJuan, E., & Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document numérique*, *17*(1), 61-84.

Donohue, M. (2022). *24 Fashion Trends From the 2000s That Aged Surprisingly Well*. ELLE. https://www.elle.com/fashion/trend-reports/a40910116/best-2000s-fashion/

Eberly, L. E. (2007). Multiple linear regression. *Topics in Biostatistics*, 165-187.

Franck, G. (1997). The fashion cycle: An analysis of trends, cycles, and fashions. Fashion Theory: The Journal of Dress, Body & Culture, 1(4), 337-356.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, *101*(suppl_1), 5228-5235.

Grün, B., & Hornik, K. (2011). topicmodels: An R package for fitting topic models. *Journal of statistical software*, *40*, 1-30.

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 168-177).

Jockers, M. (2017). Package 'syuzhet'. *URL: https://cran. r-project. org/web/packages/syuzhet*.

Kanawattanachai, P. (2022, May 19). *Top French luxury companies daily stock prices*. Kaggle. Retrieved February 16, 2023, from https://www.kaggle.com/datasets/prasertk/french-luxury-companies?resource=download

Kass, A. G. (2011). The 20th Century of American Fashion: 1900-2000 (Doctoral dissertation).

Kwartler, T. (2017). *Text mining in practice with R*. John Wiley & Sons.

Laugharne, J., Janca, A., & Widiger, T. (2007). Posttraumatic stress disorder and terrorism: 5 years after 9/11. *Current opinion in psychiatry*, *20*(1), 36-41.

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788-791.

Li, W. Y., Choi, T. M., & Chow, P. S. (2015). Risk and benefits brought by formal sustainability programs on fashion enterprises under market disruption. *Resources, Conservation and Recycling*, *104*, 348-353.

Lin, J. (2016). On the dirichlet distribution. *Department of Mathematics and Statistics, Queens University*, 10-11.

Liu, Y., & Jang, S. C. (2009). Perceived fit and customer loyalty in chain restaurants: The mediating role of customer satisfaction. International Journal of Hospitality Management, 28(4), 562–570.

López Ortega, A. (2021). Are microtargeted campaign messages more negative and diverse? An analysis of Facebook Ads in European election campaigns. *European Political Science*, 1-24.

Lu, L., & Cooper, C. L. (1999). The effects of season and weather on job satisfaction. Journal of Environmental Psychology, 19(1), 1-15.

Luca, M. (2016). Reviews, Reputation, and Revenue: The Case of Yelp.com. Harvard Business School Working Paper, (16-043).

Marco, C. A., & Suls, J. (1993). Daily stress and the trajectory of mood: spillover, response assimilation, contrast, and chronic negative affectivity. *Journal of personality and social psychology*, *64*(6), 1053.

Mohr, I. (2013). The impact of social media on the fashion industry. *Journal of applied business and economics*, *15*(2), 17-22.

Monks, E. (2022). *The rise of negative comments on brand-owned social media ads*. Crisp. https://www.crispthinking.com/blog/negative-comments-social-media-ads

National Institute of Mental Health. (2021). Seasonal affective disorder. Retrieved from https://www.nimh.nih.gov/health/publications/seasonal-affective-disorder/index.shtml

Parker, C. P., Baltes, B. B., Young, S. A., Huff, J. W., Altmann, R. A., LaCost, H. A., & Roberts, J. E. (2014). Relationships between psychological climate perceptions and work outcomes: A meta-analytic review. Journal of Occupational Health Psychology, 6(2), 107-131.

Pauca, V. P., Shahnaz, F., Berry, M. W., & Plemmons, R. J. (2004, April). Text mining using non-negative matrix factorizations. In *Proceedings of the 2004 SIAM international conference on data mining* (pp. 452-456). Society for Industrial and Applied Mathematics.

Paustian-Underdahl, S. C., Walker, L. S., & Woehr, D. J. (2014). Gender and perceptions of leadership effectiveness: A meta-analysis of contextual moderators. *Journal of applied psychology*, *99*(6), 1129.

Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).

Pesendorfer, W. (1995). Design innovation and fashion cycles. *The american economic review*, 771-792.

Rajkumar Arun, V. Suresh, C. E. Veni Madhavan, and M. N. Narasimha Murthy. 2010. On finding the natural number of topics with latent dirichlet allocation: Some observations. In *Advances in knowledge discovery and data mining*, Mohammed J. Zaki, Jeffrey Xu Yu, Balaraman Ravindran and Vikram Pudi (eds.). Springer Berlin Heidelberg, 391–402. http://doi.org/10.1007/978-3-642-13657-3_43

Reisenbichler, M., & Reutterer, T. (2019). Topic modeling in marketing: recent advances and research opportunities. *Journal of Business Economics*, *89*(3), 327-356.

Ruxton, G. D. (2006). The unequal variance t-test is an underused alternative to Student's t-test and the Mann–Whitney U test. *Behavioral Ecology*, *17*(4), 688-690.

Santana, A. D. (2014). Virtuous or vitriolic: The effect of anonymity on civility in online newspaper reader comment boards. *Journalism practice*, *8*(1), 18-33.

Siddiqui, K. (2022). Brand equity trend analysis for fashion brands (2001-2021). *Journal of Global Fashion Marketing*, 1-18.

Smyth, D. (2021, July 2). *The Average Cost of Advertising in a Fashion Magazine*. Sapling. https://www.sapling.com/10016643/average-cost-advertising-fashion-magazine

Soroka, S., Daku, M., Hiaeshutter-Rice, D., Guggenheim, L., & Pasek, J. (2018). Negativity and positivity biases in economic news coverage: Traditional versus social media. *Communication Research*, *45*(7), 1078-1098.

Spencer, M. (2021). *Global luxury sales set to outpace pre-COVID levels this year, Bain says*, from https://www.reuters.com/business/retail-consumer/global-luxury-sales-set-outpace-pre-covid-levels-this-year-bain-says-2021-11-11/

Ståhle L., & Wold, S. (1989). Analysis of variance (ANOVA). *Chemometrics and intelligent laboratory systems*, *6*(4), 259-272.

Steenstra, I. A., & Knol, D. L. (2014). The impact of weather conditions on occupational injuries and accidents in the construction industry. International Journal of Biometeorology, 58(7), 1517-1524.

Stokes, A. (2015). The glass runway: How gender and sexuality shape the spotlight in fashion design. *Gender & Society*, *29*(2), 219-243.

Terman, M., Terman, J. S., Quitkin, F. M., McGrath, P. J., Stewart, J. W., & Rafferty, B. (1989). Light therapy for seasonal affective disorder: A review of efficacy. Journal of Occupational and Environmental Medicine, 31(8), 741-746.

The Guardian. (2019, September 26). Why Paris Fashion Week is the pinnacle of style. The Guardian.

Wang, W., Feng, Y., & Dai, W. (2018). Topic analysis of online reviews for two competitive products using latent Dirichlet allocation. *Electronic Commerce Research and Applications*, *29*, 142-156.

Wu, J., & Song, W. (2019). The Relationship between Fashion Cycles and Cultural and Economic Contexts. Sustainability, 11(12), 3431.

Zhu, F., & Zhang, X. (2010). Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *Journal of marketing*, *74*(2), 133-148.