Trend detection in fashion via Text Analytics

A

Thesis

Presented to

the faculty of the School of Engineering and Applied Science

University of Virginia

in partial fulfillment

of the requirements for the degree

Master of Science

by

Will Tx Xiao

May 2021

APPROVAL SHEET

This

Thesis

is submitted in partial fulfillment of the requirements
for the degree of

Master of Science

Author: Will Tx Xiao

This Thesis has been read and approved by the examing committee:

Advisor: Yangfeng Ji

Advisor: Natasha Foutz

Committee Member: Gustavo Rohde

Committee Member:

Committee Member:

Committee Member:

Committee Member:

Committee Member:

Accepted for the School of Engineering and Applied Science:

Craig H. Benson, School of Engineering and Applied Science

May 2021

Trend Detection in Fashion via Text Analytics

Will Tx Xiao Natasha Foutz Yangfeng Ji

***Abstract*—Fashion trend prediction is a complex problem due to a vast number of latent drivers and fashion products' highly temporal and dynamic characteristics. It requires a reliable and time-sensitive means to detect fashion trend. Analyzing social media data is well-suited for the task of trend detection from the consumer perspective. In this project, I leverage text analytics on time-series social media data. I first collected data across multiple categories of fashion over a span of ten years via mining, API and third-party sources. Then I extract fashion related keywords and conduct yearly and quarterly trend detection via keyword clustering and topic modeling. The ultimate goal is to discover hidden semantic structures in a text body at scale and detect types and drivers of different fashion life cycles, such as classic, cyclical, and fad.**

I. Introduction

It became painfully obvious very early on that I seem to have an innate inability to grasp the very concept of fashion. Growing up I have always been puzzled by what my peers were wearing around me and the talks surrounding their clothing choices. Some regard fashion as an art equal to the others, while others point to the rise of fast fashion, in which fashion manifest itself ever more apparently in the form of consumer goods. Some are adamant fashion started when humans started wearing clothes. While undoubtedly that it makes for a valid argument, it's important to dissect the meaning of the word. If we were to pick apart the word of fashion into style, the artistic, physical attribute of apparel, and fashion design the industry aspect, a clearer understanding of modern fashion could be formed. The history of fashion design is widely believed to begin in the 19th century, when Charles Frederick Worth, an English fashion designer, have his label sewn in the clothes that he made. Dubbed the father of haute couture, which literally translates to "high dressmaking", Charles Frederick Worth marks the turning point into an age where the actual style and brand intertwine into a single entity, a commercial product that consumers would love and buy. Fast foward to 2021, fashion design has made leaps and bounds in terms of its reach, impact, and volume. Instead of the House of Worth's single brand, the global fashion market is estimated to be worth $2.5 trillion in 2019. In the United States, consumer spending in the apparel and footwear industry hit nearly $380 billion, and the industry employs more than 1.8 million people in 2017 [1].

That begs the question, what drives fashion? The advance and subsequent sensation of social media further alter and drive public reception of different fashion trends over periods of time outside of the seasonal and traditional domain. Furthermore, technology, globalization, shifts in demographics, business models or mainstream culture can also be among the latent drivers of fashion trend. These latent drivers combined with differences in consumers' adoption rate, i.e. their receptivity to a new fashion trend, social impact, and digital footprint make the fashion industry a highly sophisticated and dynamic market, and its trends, hard to predict. Like many other industries, the COVID pandemic has flip the fashion industry on its head as it causes major shifts to consumer behavior and disruption in supply chains [9]. Nevertheless, the fashion industry is now expecting a surge in recovery, and interests in predicting fashion trends are ever so abundant.

In order to answer the previous question and attempt to predict trend in fashion, it's crucial to identify fashion trends accurately in the first place. In this study, we attempt to determine and further classify fashion trends into categories such as classic (basic), whose popularity stands the trial of time, cyclical (fashion in the figure), whose popularity waxes and wanes within a large window, and fads, whose popularity peaks and fades much more quickly than the cyclical. All three types share the same stages of the typical life cycle of a product, but differ in its rate and volume of adoption over time. To avoid confusion and guarantee consistency, in this paper, classic, cyclical and fad will be used to categorize the three types of fashion products. [4] Due to the temporal, diverse

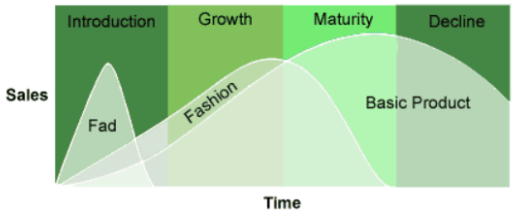


Fig. 1: Life stages of fashion products

and commercial nature of fashion and various factors driving fashion trends, studying fashion trends via historical social media data is uniquely beneficial because social media reacts and reflects hidden drivers for fashion trend in near real time. In an effort to investigate and reveal hidden trends, a number of methods from the Natural Language Processing (NLP) toolbox are used, including topic modeling,the unsupervised learning approach and frequency analysis using bag of words (BoW).

The thesis is part of a larger body of private research that aims to detect fashion trend through a multi-modal approach, encompassing both text analytics and image recognition. Part of the data collected include images of tweets. This is a work in progress.

II. Methods

A. Topic modeling with LDA

Topic modeling is an unsupervised machine learning and NLP technique that uses a statistical, generative model that's capable of scanning through large corpus of text and cluster its words or phrases to assign each document in the corpus a topic, a generated cluster of words. The technique is based on the intuition that within a large corpus of text that one cannot afford time to read, the text can cover a small set of topics, and within each topic, it can cover a small set of specific words. An example would be news about the Flint Michigan water crisis would invariably include the words such like 'Flint', 'Michigan', 'water', 'health', and 'lead'. Fundamentally, it allows us to gain abstract insight or perform qualitative analysis without having to bat an eye on an arbitrarily large number of text documents. Hence ever since its inception, it's seen wide application in fields where large corpora of text are abundant, such as history, academical publications, fiction and literature, social science, machine translation, bioinformatics and many more.

From a high level, in order to use a topic model, one inputs N documents, V words and after training, i.e. iterating over the corpus some set number of iterations with certain statistical inference method, such as the Gibbs Sampling algorithm. And it outputs K topics per document, along with a multinomial distribution of the probability of document n being topic k. In actual implementation, one would also need to pass in the number of passes, and the number of topics one wish to generate. This would introduce the important caveat of hyperparameter tuning, requiring humans to spend a significant amount of effort before achieving satisfying results. We will expand on this in the Discussion section.

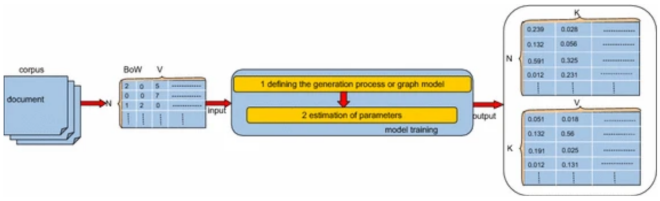


Fig. 2: Dirichlet Distribution parameterization, [8]

Latent Dirichlet allocation is one of the statistical, generative models that has seen wide adoption in topic modeling since Blei's et al. (2002) [2] proposal. LDA is a three-level hierarchical Bayesian model where document in the corpus "is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities" [2]. To understand LDA, one must first understand Dirichlet distribution. One easy way to visualize and understand Dirichlet distribution is by simplices. For our purposes, we can safely ignore the fraction as the coefficient, which is a generic beta function consisting of gamma functions and focus on the probabilities themselves. The key is by changing values of α and m here,

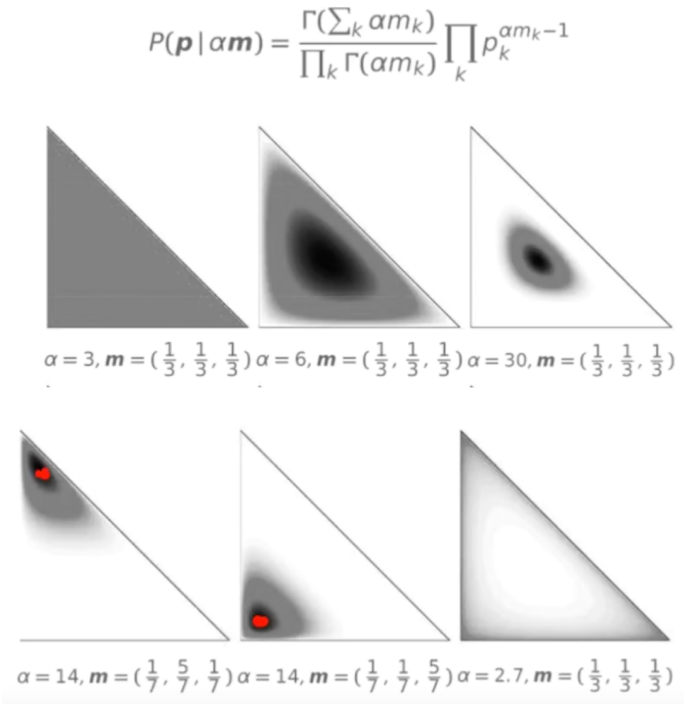


Fig. 3: Dirichlet Distribution parameterization, [3]

we can obtain from uniform distribution to an infinite amount of multinomial distribution within the simplex. The higher α is, the more concentrated the distribution around the simplex's mean, forming this hump around it. And by manipulating m, we can shift around the distribution to make it likelier to find a document or words belonging to a certain topic, with the vertices representing topics. The bottom right simplex shows the ideal parameterization for topic modeling's use case, where it's unlikely to find a particular document, e.g., to be in an ambiguous topic around the mean. Rather, it's denser around the vertices, i.e. higher probability to belong to one of three distinct topics.

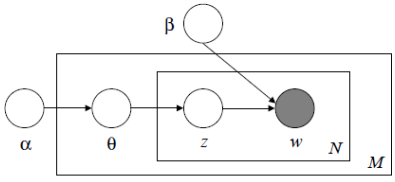


Figure 1: Graphical model representation of LDA. The boxes are "plates" representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document.

Fig. 4: LDA model, [2]

Here,M stands for the number of documents, and N each word within the M documents. For each topic, a multinomial distribution β is drawn with some parameterization λ. Then, for each document, a multinomial distribution θ is drawn with some parameterization α. Next, For each word in the document, select a hidden topic z from yet another multinomial

distribution with some other parameter. And finally chooses the observed word w_n from β and z. This is what it means for LDA to be hierarchical.

B. BERTopic

BERTopic [7] circumvents the problem of the cumbersome hyperparameter tuning LDA faces, and attempts to simplify the process with Google's widely acclaimed pre-trained deep neural net BERT [5] with a hundred million of parameters. It addresses the traditionally difficult problem of evaluating topic models. Entire papers were made on visualizing the output of topic models to make them easier to train and evaluate [6], in technical terms, the eyeball test. How would one begin to practically quantitatively evaluate a topic model, whose objective is to discover hidden semantic structures within large corpus of unseen texts? How does one know if the output didn't miss any topics within documents? Or whether if it misses any important words within topics? Some methods have been proposed but it remains one of the challenges of topic modeling. We will dive into more details in Discussion.

BERTopic harnesses the power of BERT to extract contextual embeddings from the documents. It then runs a dimension reduction algorithm (UMAP) and then a clustering algorithm (HBDSAN) to cluster similar documents together. It created topics by a variant of class-based TF-IDF (term frequency-inverse document frequency), a classic tool in information retrieval and document search. It's a quick and easy way to preview what is the most important, or relevant information in the data. Unlike LDA, BERTopic doesn't have the issue of having to input number of topics or number of passes. Instead it has the opposite problem of potentially having too many closely related topics, at which point a either manual or another round of dimension reduction would be required, merging similar topics and recalculate words within them.

C. BoW

The simple yet efficient BoW is implmented with a countvectorizer. We first take quarterly top two hundred unigrams and top one hundred bigrams. Then take the union of all the quarterly top keywords. Next, each bigram is split into two unigrams, a set difference is done, unigram minus split bigrams tokens, resulting in unique unigrams. This is done to minimize the information overlap between them, and the choice to keep bigram and only unique unigram stems from the observation that we made early on, most frequent/top bigrams tend to be more informative and have a higher probability to be relevant than the unigrams of same ranks. We then sort the overall frequency of ten years to obtain an overview of top X keywords over ten years. Lastly, we perform an additional step, to search back in our data for any missing top keywords. This is accomplished via more research and compiling a comprehensive list of fashion attributes and another set difference. Then we focus on trends that were discovered this way that wasn't in the original top keywords and analyze and categorize those.

D. Datasets

The primary dataset of our experiment is composed of tens of millions of historical tweets containing mentions of seven clothing and apparel categories over the last ten years. Out of the seven categories, they are further divided into two groups: Shirt and Dress v.s. Pants, Jeans, Trousers, Skirt,and Blouse. Both groups are collected via keyword and boolean-based queries but they differ in the type of keyword used. For shirt and dress, both the attribute and category keyword itself are used. An example of some of the early keywords used for collecting shirt: Twitter is filled with random people

(Attribute) AND Entity
((white OR black OR linen OR cotton OR tan OR beige OR yellow OR gold OR orange OR red OR pink OR salmon OR purple OR blue OR navy OR denim OR green OR brown OR grey OR charcoal OR vintage OR chrome OR "punk rock" OR message OR graphic OR striped OR military OR political OR printed OR tagless OR crew OR neck OR statement OR oversize OR band OR logo OR V-neck OR plain OR "make your own" OR slim-fit OR band OR logo OR long OR sleeveless OR slim OR medium OR large OR turtleneck OR diy OR collared OR dress OR uncollared OR flannel OR office OR "short sleeve" OR button OR pocket OR plaid OR metal OR ribbon OR band OR feminist OR thermal OR logo OR camo OR vneck OR space OR hooded OR jersey OR retro OR comfortable OR floral OR sweat OR synthetic OR polyester OR weave OR short OR polo OR collar OR classic OR chambray OR XL OR XXL OR color OR size OR outfit OR material OR fabric OR fit OR pattern OR print OR gabanna OR crewneck) AND (shirt OR tee OR "t-shirt" OR tshirt OR Tees))

Fig. 5: Keyword-based Query

expressing themselves randomly. The majority of tweets in our experience contain no useful information at all, such as "the weather is great today, so I put on a shirt and went for a walk". By this method, we can effectively improve the quality of the dataset by leaving out many tweets that are of no use to us, at the data collection phase. This was necessary due to the sheer volume of tweets containing shirt or dress and not for the rest, because the mention volume of those categories are relatively small and we were able to collect them all. In addition, group one contains mostly original tweets while the five categories contain retweets as well, again because of their low volume.In order to achieve this volume, a system of web scraping and automation scripting is developed and used to maximize rate limits through official API and third party sources. Control variables regarding each category such as daily tweet volume over ten years are collected afterwards through third party sources, including Net Sentiment, number of unique authors, total followers, impression and retweets. They are saved as long-format daily time series.

E. Preprocessing

Preprocessing is an unglamorous yet vital step in analysis of large social media datasets. Our preprocessing pipeline involves removing urls, emojis, punctuations and special characters, converting to lowercase, lemmatization. Stop word removal is an art to its own, as we've discovered, and can have a tangible impact on all subsequent results. Scouring online for common libararies' stop word sets is quickly proven to be inadequate. Custom stop word list of hundreds of words are crafted through trial and error to include everything from curses, slangs, netspeak, to common words occurred in tweets etc. Still getting sub-optimal results we look for larger dictionaries to serve as stop word set. Eventually we added entire

subcategories of dictionaries such as curses and slangs which helped tremendously. We experimented with Part of Speech (PoS) tagging and filtering, however the runtime increase was prohibitively large at scale, which led to its abandonment. Efforts were made to mitigate the lack of PoS tagging such as adding verbs to the set of stop word. Multiple rounds of loose data collected are cleaned, deduped, concatenated before being portioned into yearly and quarterly data per category. Tokenization is also performed for both unigrams and bigrams, for n-grams larger than two tend to yield uninterpretable results in our experiments.

III. Results

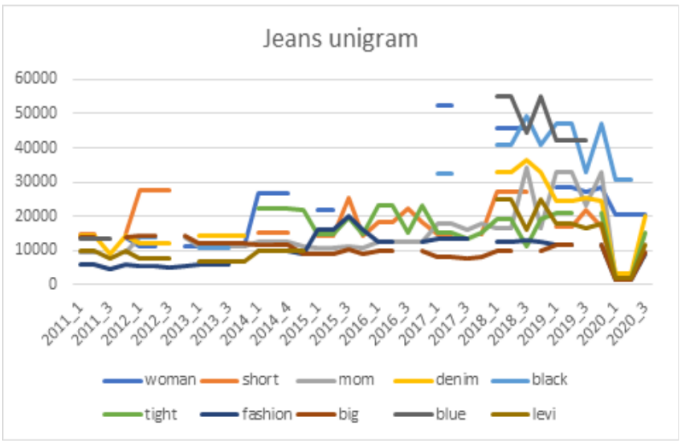
	Topic # 01	Topic # 02	Topic # 03	Topic # 04	Topic # 05
0	perfect	girl	halloween	men	woman
1	store	baby	homecoming	prom	black
2	style	fashion	costume	michael	fashion
3	tie	party	fancy	wore	kim
4	summer	flower	print	zara	men
5	online	princess	free	shoe	watch
6	bad	deal	floral	use	kardashian
7	black	summer	hill	white	nwt
8	hope	impress	sherri	code	white
9	loving	ebay	plus	dad	small
10	fashion	toddler	shipping	kor	kimkardashian
11	code	wedding	amazing	blue	deal
12	link	cute	dolce	black	blue
13	sexy	clothes	gabbana	woman	sz
14	blog	child	adult	asked	ebay
15	different	pink	whats	somebody	red
16	hair	month	sleeve	bout	x
17	ootd	school	short	finally	dog
18	hot	tutu	doll	food	sale
19	beautiful	shopping	maxi	ago	taylor

Fig. 6: Result of LDA over 1 year, Dress

In our experiments so far, topic modeling using either LDA or BERT yielded unsatisfactory results. LDA produced some convincing clusters of words but fell short of getting more than a few passable topic. Under topic 2, 'baby', 'girl', 'cute', 'child', 'princess', 'toddler', 'child', the topic is clearly capturing baby and girl dresses. A quick internet search about 'tutu' would reveal 'Tutudumonde' is a popular brand that designs fancy dresses for toddler, up to teenage girls. Topic 3 seems to capture 'Halloween', 'costume', 'doll', 'homecoming', some Halloween themed dresses and the renowned Italian luxury brand 'dulce' 'gabbana'. Topic 4 seems men themed, 'blue', 'black', 'Micahel', 'Kor', 'Zara' ('Kors' presumably removed 's' from lemmatization), the word 'shoe' revealed it could be more about dress shoe instead of dress. Along with the rest, we can be reasonably confident topic 4 is about men's dress shoes. While topic 5 seems to be surrounding Kim Kardashian and possibly the black, blue or red dress she wears. The BERTopic's outputs on the other hand seems utter gibberish. Possibly due to tweets' pithy structure and small sample size. The reason it was ran over a quarter in the first place was because it was throwing memory error otherwise, presumably due to the mammoth scale of the pre-trained BERT under the hood.

Fig. 7: Result of BERTopic over 1 quarter, Skirt

Seeing disappointing results and facing challenges with topic modelling, we decided to pause the topic modelling and focus our attention onto the fundamentals. The BoW approach overall provided better, interpretable results. Shown below are some of our line charts of their frequency over time (40 quarters over 10 years), separated into unigrams and bigrams. Note that we not only have the time series of the top ten, but also the top one hundred and one thousand. In fact the union of top quarterly keywords end up being thousands of keywords over ten years. The examples shown here are previews of initial results:



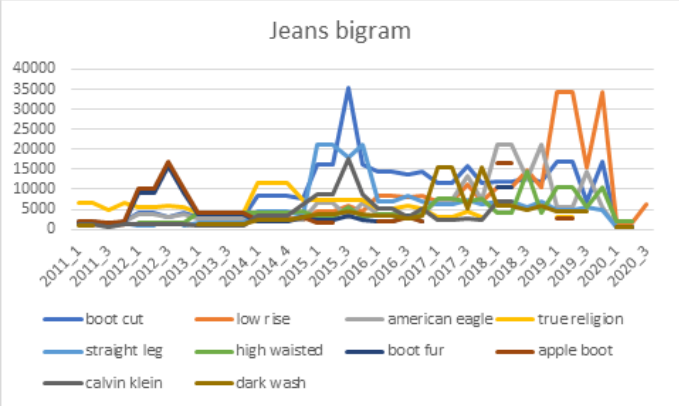


Fig. 8: Result of BoW, 40 quarters, Jeans

One can clearly see that 'black' and 'woman' jeans surge of popularity the middle two quarters of 2017. A decline in women's jeans popularity in the following years but a rise for black jeans. A cursory look at the bottom of the charts would easily tell one that it seems classic jeans styles would include mom, tight, big jeans. While levi's popularity rose and fell - indicating it would be closer to a cyclical product. Similarly, black jeans and woman jeans seem to share traits between cyclical and a fad. On the bigrams side, boot cut saw a sudden rise of popularity the quarter between 2015 and 2016, yet it didn't disappear from the radar and still manage to maintain a steady level of mentions, becoming part of the classic. 2019 saw low rise jeans growth in popularity and fits the bill of a fad. Again looking near the bottom, we can identify dark wash, high waisted, Calvin Klein etc. as classic fashion. Attributes more akin to fads that see trend spikes for a short period of time are supported by anecdotal accounts of fashion-savvy consumers (e.g. "boot-cuts was really popular in 2015").

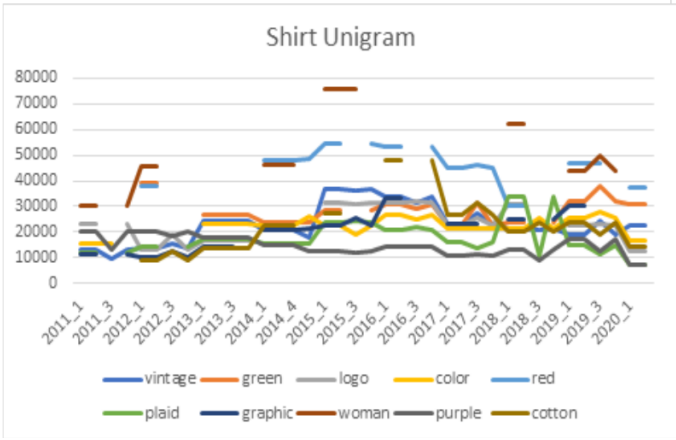


Fig. 9: Result of BoW, 40 quarters, Shirt

For shirt, classic styles seem to include plaid, purple, green, vintage cotton, color shirt. Women shirt see gaps and jumps of popularity over years but always seem to bounce back, so it would safe to call it a cyclic. Red shirts were popular 2014 to 2017 but has seen since a decline, potentially showing a classic falling out of fashion in action.The results shown here haven't adjusted with the aforementioned controlled variables. We were able to make initial judgement on what style belongs to each category but to get a true measure of the mention, one should be able to factor in, e.g., whether the mention of a style has increased, or was there simply just a influx of tweets because of technical reasons, or maybe has there been disproportionate amount of retweets, perhaps by bots. There is work left to be done.

IV. Discussion

Fashion trend detection, preceding prediction, requires an ideally interpretable, reliable and low-latency method to be able to keep up with fashion trends, as fickle as they may be. In our experience, the simplest method ticks all the boxes and is by far the fastest method in comparison. It has the additional benefit of reserving the option to go more fine-grained as well. We have chosen our top keywords to be quarterly but we could do monthly or daily or even hourly if we choose, say in trying to spot some evanescent fad among some sub-culture or sub group. That is not to discount the powerful topic models in any way, shape or form. They can be tuned to be sophisticated to be capable of much more in the right hands, under little computational time or resource constraints. The major challenge with topic modeling with LDA has been its troublesome hyperparameter tuning. In its most widely available package on python, this requires one to enter the $eval\_every = N$ which produces a log file that the programmer would have to step through and calculate cohesion score and in turn slows down the already long runtime. Not only that, the cohesion score only tells half the story; as long as the model has no clearly defined objective function, on evaluating a model's ability to capture high quality topics within the semantic space, there would be no way of learning automatically how many passes it should run or how many number of topics would be suitable, and would always require constant human supervision. One must not only keeps evaluating whether the model has converged or not, but also, at every other pass, take time to actually read the chosen words and multinomial probabilities to check if the model not only has converged but actually makes sense.

V. Conclusion

For our purposes of trend detection, frequency analysis over a large corpus of social media data was able to capture fashion trends' ebb and flow over forty quarters, providing valuable insight into their diverse life cycles, paving the way to investigate fashion trends' latent drivers. Topic modeling with LDA shows promising results but requires more time and guidance in tuning the model. And although preprocessing may be tedious, decisions made at the phase are vital because they impact everything that follows. For one, having a comprehensive and cleverly chosen set of stop words can enable the simple BoW method to achieve surprising results on large-scale social media mining.

REFERENCES

[1] R. D. Beyer, "The economic impact of the fashion industry," Joint Economic Committee, Senate, Tech. Rep., 2019.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[3] J. Boyd-Graber. Computational linguistics. [Online]. Available: https://sites.google.com/umd.edu/2021cl1webpage/

[4] Cornell. The cutting edge apparel business guide. [Online]. Available: https://courses.cit.cornell.edu/cuttingedge/lifeCycle/03.htm

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.

[6] A. Ganesan, K. Brantley, S. Pan, and J. Chen, "Ldaexplore: Visualizing topic models generated using latent dirichlet allocation," 2015.

[7] M. Grootendorst. Bertopic. [Online]. Available: https://github.com/MaartenGr/BERTopic

[8] L. Liu, L. Tang, S. Yao, and W. Zhou, "An overview of topic modeling and its current applications in bioinformatic," *Springerplus*, vol. 5, no. 1608, 2016.

[9] McKinsey, "The state of fashion 2021," McKinsey Company, Tech. Rep., 2020.