**PROJECT 2 – StyleMatch: Unleashing the Power of Machine Learning to Connect Influencers and Fashion Brands**

**GROUP 4**

Group Members: Jeremy Banas, Parthik Davda, Ryan DiBiase, Qochiwe Nyirenda, Jing Wang

School of Business, University of Connecticut

OPIM 5671: Data Mining & Business Intelligence

Professor Sudip Bhattacharjee

April 23rd, 2023

# Executive Summary

The purpose of this report is to use a text-mining strategy to categorize tweets made by fashion and beauty influencers and then pair these influencers with brands or companies in the fashion and beauty industry that have similar core values. Our research will use Twitter as a text source, Python as a data gathering and cleaning tool, and SAS Enterprise Minor Workstation as a text mining technique to develop the model and perform the statistical analysis.

For our analysis, we do not have a specific target variable, but we are aiming to help brands navigate the growing influencers space while also increasing their efficiency in selecting influencers to engage with. Through this model, our aim is to successfully match each one of our fashion brands with influencers that best match their brand images using the commonality in language detected in their tweets via text mining in SAS Enterprise.

Results & Conclusion

Our project allowed us to build that was able to:

* Successfully matched influencers to brand names based on tweets.
* The model was able to distinguish brands/influencers by text topics.
* Used a vector scale to measure brand subject importance and use it to match brands and influencers based on similarity in language.

Based on the results from our model output, we would like to include the following recommendations:

* Inclusion of a more detailed User Topic dataset to add more terms to topics and extract even more subjects.
* Organizations utilizing our model should combine their findings with market/trend analytics to understand the current market trends & what the target audience is interested in.
* Utilization of a more accurate sentiment analysis to better understand the context of the tweets being analyzed.
* Organizations making use of our model should also include predictive analytics to identify influencers likely to be successful in promoting their brand & which content will resonate with the brand’s target audiences.

We would also like to add the following as regards which influencers each brand in our dataset should collaborate with:

|  |  |
| --- | --- |
| BRAND  NAME | INFLUENCER NAMES |
| Nike | Alexa Chung, Shini Park, Margaret Zhang |
| Gucci | Chiara Ferragni, Alexa Chung, Gala Gonzalez |
| MACcosmetics | Camila Coelho, Jenn Im, Nikita Madhani |
| Maybelline | Camila Coelho, Jenn Im, New Darlings |
| Louis Vuitton | Leandra Cohen, Olivia Palermo, Pandora Sykes |
| Reebok | Nikita Madhani, Elma Beganovich, Lauren Conrad |
| Versace | Chiara Ferragni, Olivia Palermo, Elma Beganovich |

Table of Contents

[Executive Summary 1](#_Toc132755320)

[1.0 INTRODUCTION 3](#_Toc132755321)

[2.0 DATA ACQUISITION & DATA PREPARATION 4](#_Toc132755322)

[2.1 Data Acquisition Using Python 4](#_Toc132755323)

[2.3 Data Cleaning & Data Preparation 4](#_Toc132755324)

[3.0 MODEL DEVELOPMENT 5](#_Toc132755325)

[3.1 Model Diagram 5](#_Toc132755326)

[3.2 Output in Excel 5](#_Toc132755327)

[4.0 SENTIMENT ANALYSIS 6](#_Toc132755328)

[4.1 User Topics 6](#_Toc132755329)

[4.2 Output in Excel 6](#_Toc132755330)

[5.0 MODEL INTERPRETATION 8](#_Toc132755331)

[5.1 Results 8](#_Toc132755332)

[5.2 Interpretation 8](#_Toc132755333)

[6.0 PROJECT INSIGHTS 9](#_Toc132755334)

[7.0 CONCLUSION & RECOMMENDATIONS 10](#_Toc132755335)

[7.1 CONCLUSION 10](#_Toc132755336)

[7.2 RECOMMENDATIONS 10](#_Toc132755337)

[8.0 REFERENCES 11](#_Toc132755338)

# **1.0 INTRODUCTION**

The purpose of this report is to use a text-mining strategy to categorize tweets made by fashion and beauty influencers and then pair these influencers with brands or companies in the fashion and beauty industry that have similar core values. Our research will use Twitter as a text source, Python as a data gathering and cleaning tool, and SAS Enterprise Minor Workstation as a text mining technique to develop the model and perform the statistical analysis.

The text set to be analyzed and modeled was from Twitter and was divided into two groups: Influencer tweets and Parent company/Brand mission derived from tweets. The following URL is included as part of our references:

* https://twitter.com/

By gathering a set of texts from selected influencers and brands' daily tweets, we will be able to gain initial insights into the content they are sharing. By applying text mining techniques, we can group influencers who have similar traits based on the words they use in their tweets, which will result in a list of clusters and groups. We will then use these groups to identify connections between influencers and brands based on shared core values.

Since the COVID-19 pandemic, marketing in fashion industries has changed with the increase in social media use making social media marketing one of the key ways to increase brand visibility with influencer-driven marketing generating $6.5 for every $1 spent. At the same time, the marketing dollars available to spend on influencers are decreasing as the number of influencers available for brands to select from is increasing. This means brands within the fashion and beauty industry have an appetite for data analysis or tools that can increase the effectiveness of their marketing campaigns and achieve better marketing-to-revenue dollars ratios.

For our analysis, we do not have a specific target variable, but we are aiming to help brands navigate the growing influencers space while also increasing their efficiency in selecting influencers to engage with. This report will first cover the data preparation and acquisition process through Python, before detailing our text mining process and how we selected the most appropriate parameters for our purpose.

As part of our analysis, we will make the closest connections between influencers and brands. Once we identify ideal sets, we will provide the best-fit influencers back to the brands contracting our team, allowing them to optimize their marketing dollar spend by only focusing on influencers that already fit their brand.

Finally, the project insights will be discussed before we make our conclusion and recommendations.

# 2.0 DATA ACQUISITION & DATA PREPARATION

## 2.1 Data Acquisition Using Python

The first step in the project was to collect data on brands and influencers from Twitter using Python scripts and the Twitter API. We scraped 1000 tweets for each brand and influencer. The scripts were designed to search for specific keywords and hashtags related to each brand and influencer and collect data such as usernames, follower counts, and tweet content.

**2.2 Data Description**

Below is a description of all the data variables used in our dataset.

Table 1: Data Variables

|  |  |  |  |
| --- | --- | --- | --- |
| **ID No.** | **VARIABLE NAME** | **VARIABLE DESCRIPTION** | **VARIABLE ROLE** |
| 1 | Date | Date of tweet | Rejected |
| 2 | Uname | Username (Influencer Name or Brand Name) | Input |
| 3 | Tweet | Actual tweet as downloaded from Twitter | Text |

## 2.3 Data Cleaning & Data Preparation

Below are the steps followed to clean and prepare our data for data modeling.

*2.3.1 Data Cleaning -* After the data was collected, it was cleaned up using several steps. First, all irrelevant hyperlinks were removed from the tweet content to streamline the data. Second, emojis were converted to text to make the data more accessible. Third, similar emojis were clustered together and categorized to simplify the data. Fourth, all tweets in languages other than English were removed to ensure consistency in the data.

*2.3.2 Data Categorization -* In addition to clustering similar emojis, the data was also categorized into separate Excel files for each brand and influencer. All flags were categorized under "flags" to simplify the data and make it easier to analyze.

*2.3.3 Data Upload -* Once the data was cleaned up and categorized, it was uploaded to SAS Enterprise Miner for analysis. Separate Excel files were created for each brand and influencer to ensure consistency in the data and make it easier to analyze.

# 3.0 MODEL DEVELOPMENT

## 3.1 Model Diagram

Diagram

Description automatically generated

Figure 1: SAS Model Diagram

## 3.2 Output in Excel

# 4.0 SENTIMENT ANALYSIS

One of the key features of an accurate influencer marketing campaign is sentiment analysis. Sentiment analysis allows data scientists to gauge user perception of a brand. For our project, we used sentiment to gauge the tones of tweets from influencers and understand whether our dataset contained more negative or positive sentiment as regards our topics.

## 4.1 User Topics

We defined a user topic dataset that contained both positive and negative tones and used it to measure the tone of a particular tweet.

Graphical user interface, application

Description automatically generated

Figure 2: User Topics for Sentiment Analysis

## 4.2 Output in Excel

After running our sentiment analysis, our excel output showed that our dataset was heavily skewed towards positive sentiment. For our project, we assumed that any tweets not assigned a positive or negative score were considered to have a neutral tone.

Table 2: Sentiment Breakdown of Tweets

|  |  |
| --- | --- |
| TONE | COUNT |
| Negative | 1,473 |
| Positive | 8,392 |
| Neutral | 12,448 |

The skew towards positivity is to be expected as influencers talking about a product are generally trying to promote it. From the negative scored tweets, we noticed that our model does not understand the context of the tweets and therefore deems a tweet negative with no understanding of the context in which a word was used. We concluded that the sentiment analysis proved more than adequately that our influencers generally spoke positively about the issues we were interested in.



Figure 3: Excel Output of Sentiment Analysis on Tweets



Figure 4: Sum of Negative & Positive Tweets for each Influencer

# 5.0 MODEL INTERPRETATION

Model interpretation is the process of using different evaluation metrics to understand a machine learning model's performance, as well as its strengths and weaknesses. In this section, we will assess the efficacy of our models and perform model comparison.

We took two approaches to evaluate our model. For both approaches, we took the SAS output binary score values for each tweet by influencers and brands and calculated the percentage of the total tweets that met each category. From there our two methods diverged. For our initial evaluation, we set a threshold of 15% and said any brand or influencer would belong to a topic if they had their tweet percentage above this threshold for that topic vector. We then partnered Brands and Influencers by evaluating what percentage of their final topics overlapped. This approach we called the “Percentage Threshold” approach, since we used 15% as a filter to eliminate topics that were infrequently talked about. For our second approach, the “Percentile Threshold”, our goal was to increase the quality of our matches. To do this, we took the top 20th percentile of topic vectors from the Brand data set and the top 10th percentile of topic vectors from the Influencers data set. To pair the two sets, we eliminated any match that shared less than 50% of topics and only considered the top 2 percentage scores for each brand. For the most part, this approach led to less matches, but overall a higher quality of match.

## 5.1 Initial Results – Percentage Threshold

The goal of the percentage threshold approach was to use numerical methods to identify which topics were primarily used to represent each

Table 3: Vector Percentages by Brand – Percentage Threshold

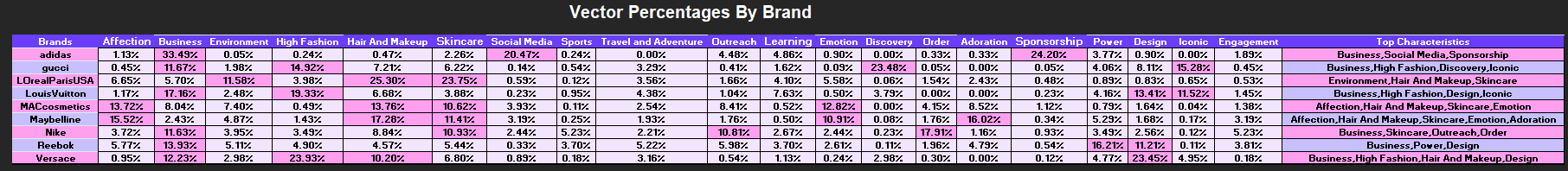
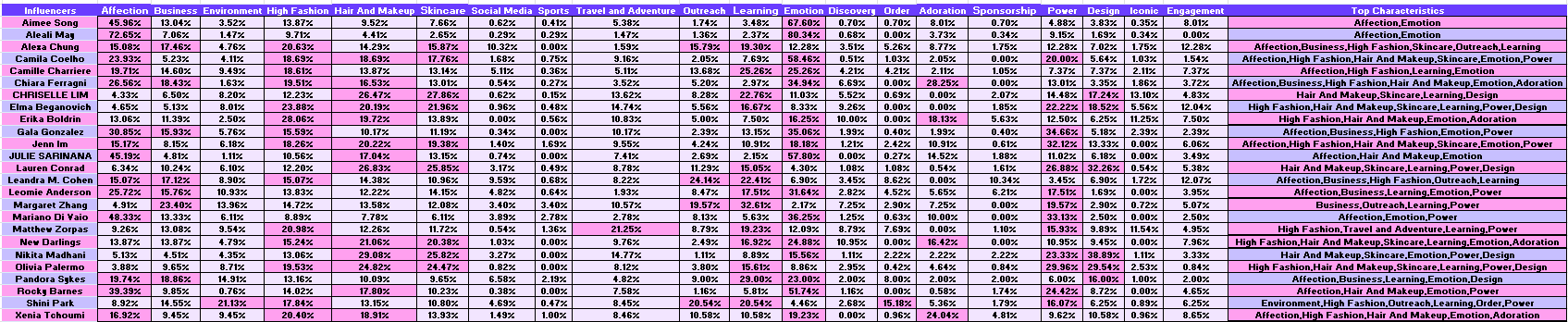


Table 4: Vector Percentages by Influencer – Percentage Threshold



## 5.2 Updated Approach – Percentile Threshold

Table 5: Vector Percentages by Brand – Percentile Threshold

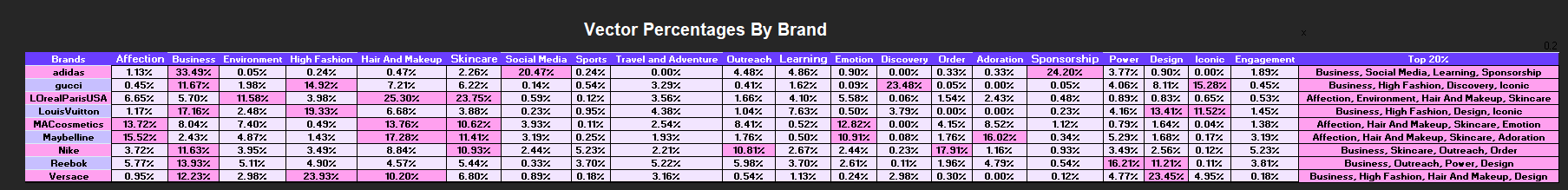
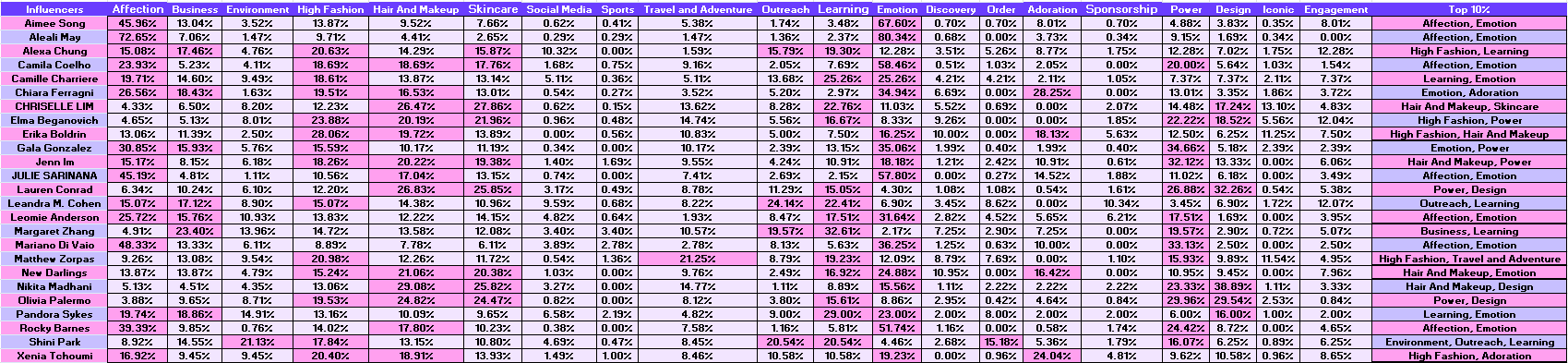


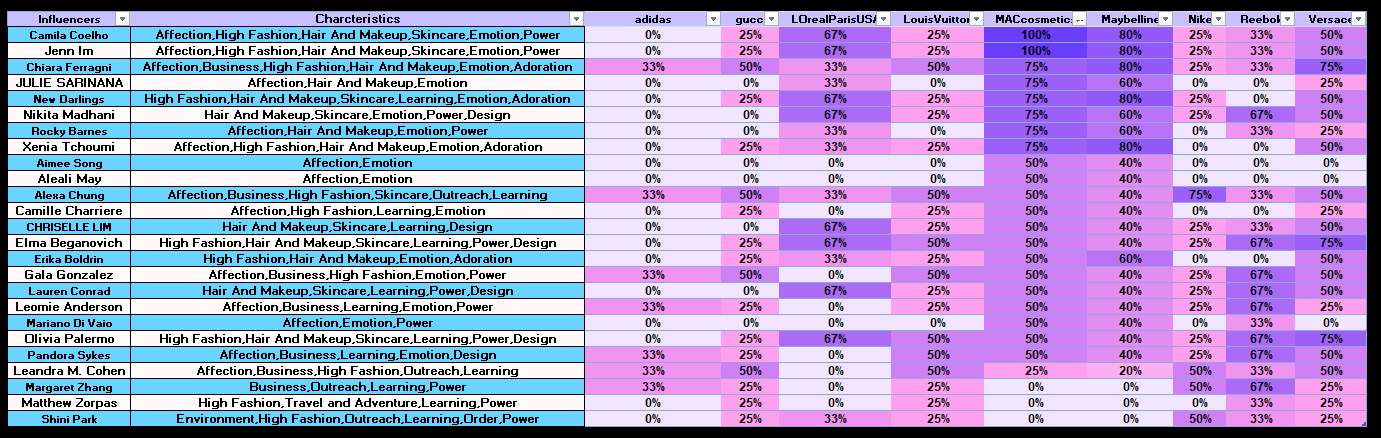
Table 6: Vector Percentages by Influencer – Percentile Threshold



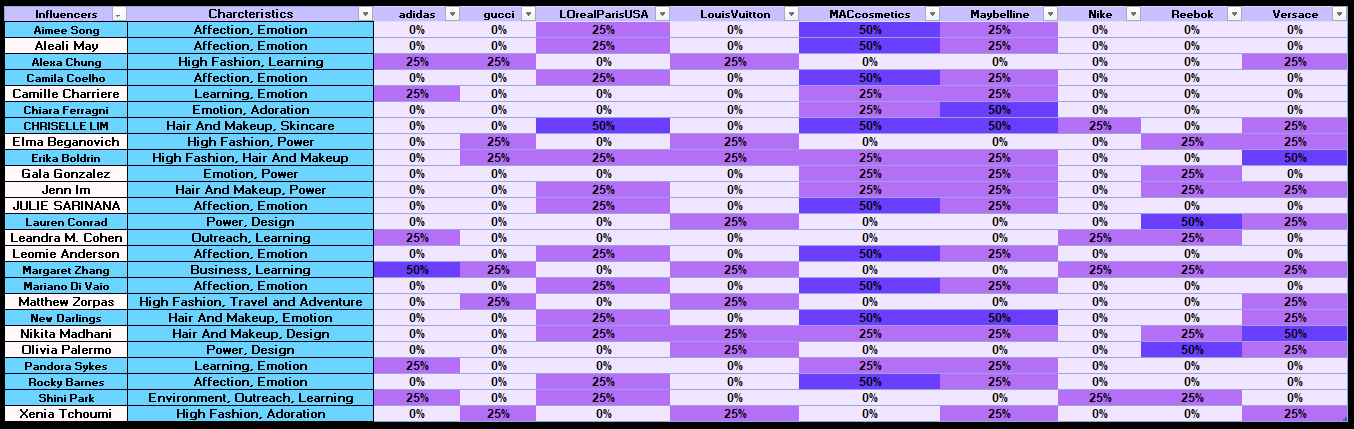
## 5.3 Interpretation

Our models were evaluated and accepted based on the following:

Table 7: Brand and Influencer Matching Matrix – Percentage Threshold



*Table 8: Brand and Influencer Matching Matrix – Percentile Threshold*



# 6.0 PROJECT INSIGHTS

Insights allow us to better understand a business and optimize it based on our data analysis. While exploring our data, the following was noted.

6.0.1 Impact of Language: While conducting our analysis, we noticed the following regarding the terms and topics assigned a high weighting in SAS. As we learnt, terms too rare or too common in the corpus were eliminated from our dataset. From this dataset we made the following conclusions:

* Words too general or too specific would not accurately identify influencers for brands.
* Words too negative or too positive would not accurately identify influencers that fit brands. Our sentiment analysis was not accurate it did not recognize context. Therefore, it is important to use other criteria such as audience demographics and interests when identifying influencers that fit brands.
* The target audience & influencer must speak the same language. In our case, we used English tweets only.

6.0.2 Effects of Matching Brands and Influencers: Our research showed that some of the effects of accurately matching an influencer to a brand are:

* Digital marketing campaigns will thrive.
* Brands are given an opportunity to connect with their influencers’ dedicated fan base.
* Advertising campaigns that strategically aligned influencers and brands involve fewer resources but lead to faster and more astounding results.

6.0.3 Language Similarities Already Exist*:* Another startling realization as we conducted our model interpretation was the evident similarities in language among some influencers. We concluded that this could be attributed to the influencers having a preset script as a reference from specific brands or it could simply be that influencers working in the same industry have a similar manner of speech.

# 7.0 CONCLUSION & RECOMMENDATIONS

## 7.1 CONCLUSION

Our project allowed us to build that was able to:

* Successfully matched influencers to brand names based on tweets.
* The model was able to distinguish brands/influencers by text topics.
* Used a vector scale to measure brand subject importance and use it to match brands and influencers based on similarity in language.

## 7.2 RECOMMENDATIONS

Based on the results from our model output, we would like to include the following recommendations:

* Inclusion of a more detailed User Topic dataset to add more terms to topics and extract even more subjects.
* Organizations utilizing our model should combine their findings with market/trend analytics to understand the current market trends & what the target audience is interested in.
* Utilization of a more accurate sentiment analysis to better understand the context of the tweets being analyzed.
* Organizations making use of our model should also include predictive analytics to identify influencers likely to be successful in promoting their brand & which content will resonate with the brand’s target audiences.

We would also like to add the following as regards which influencer each brand in our dataset should collaborate with:

|  |  |
| --- | --- |
| BRAND  NAME | INFLUENCER NAMES |
| Nike | Alexa Chung, Shini Park, Margaret Zhang |
| Gucci | Chiara Ferragni, Alexa Chung, Gala Gonzalez |
| MACcosmetics | Camila Coelho, Jenn Im, Nikita Madhani |
| Maybelline | Camila Coelho, Jenn Im, New Darlings |
| Louis Vuitton | Leandra Cohen, Olivia Palermo, Pandora Sykes |
| Reebok | Nikita Madhani, Elma Beganovich, Lauren Conrad |
| Versace | Chiara Ferragni, Olivia Palermo, Elma Beganovich |

# 8.0 REFERENCES

1. <https://www.amraandelma.com/100-top-fashion-influencers-in-2020/>
2. <https://www.businessinsider.com/influencer-marketing-report?IR=T>
3. <https://twitter.com/>