Detecting influencer on twitter across different genres:
Capstone 2 Milestone Report 2

By Pravati Swain

Table of Contents

1. Introduction

In Word-of-Mouth marketing, companies use social media influencers to spread information about their new products or brand more effectively. Twitter is one of the most popular platforms available for this purpose. In this project, I propose building a supervised machine learning model to detect whether or not a user is an influencer, or potential influencer, based on tweets across different categories. I will be analyzing tweets containing three different hashtags common amongst influencers: #fashion, #fitness and #travel. In doing so, I will also examine what attributes of a tweet differentiate influencers from non-influencers, which may help inform users on how best to achieve influencer status, and brands on what techniques are most effective for getting attention and followers. For the purposes of this project, influencers are defined as users with greater than 30k followers.

2. Client Profile

There are two primary uses for the findings in this report. The first is understanding better what aspects of a tweet make it more or less likely to come from an influencer. These findings could be useful both to aspiring influencers, current influencers, and any brand seeking to better improve their following on social media.

The second use would be to use the machine learning model I create to predict whether or not someone is an influencer. Clearly, companies can look up users by their follower count already, however this model could be useful to brands to find 1) potential influencers who are doing the right things but are still building a following (and may be cheaper) and 2) to separate users who are influencers from users that simply have high follower counts for other reasons but do not make the types of posts that would be most likely to help their brand.

3. Description of Datasets

For this project, I collected the tweets from Twitter API using Tweepy (a python library for accessing twitter API) for three different hashtags: #fashion, #fitness and #travel. The dataset contains information about user_name, user_description, location, following (the number people the user is following), number of followers, total_tweets (the total tweets of the user), user_create_date, tweet_create_date, retweet_count and text (tweet). All the three datasets contain between 135,000 - 95,000 tweets.

4. Data Wrangling

The datasets used for this project were collected using Twitter API and stored in csv files. The twitter data was then loaded into pandas DataFrame followed by several data cleaning and data preprocessing steps before the EDA and machine learning.

- I removed duplicate tweets, retweets and tweets without hashtags
- I divided the tweets into two groups based on the number of followers. The tweet with more than 30,000 followers are labeled as '1' and tweets with <= 5000 followers are labelled as '0'. I created a new column called 'label' and stored those two groups of tweets. The tweets between > 5000 and < 30, 000 followers are removed from the datasets.
- Then I preprocessed the tweets using several preprocessing packages. I removed URL, Stopwords, mentions, punctuations, numbers and special characters except #tags. I created a new column called clean_tweets and saved the preprocessed tweets.

5. Exploratory Data Analysis

In this data exploration, I checked the distribution of tweets in different labels based on tweet length, % of capital letters per tweets, number of hashtags and the readability level of tweets in three different tweet categories. I divided the tweets in two labels '0' and '1' based on the number of followers. The label '0' is for tweets with <= 5000 followers (non_influencer) and '1' for tweets with > 30, 000 followers (influencer). The three different categories used in this project are fashion, fitness and travel tweets.

5.1 Distribution of tweet length (number of words per tweet) in different labels

I have plotted the distribution of tweet length in different labels (**Fig 1**) and I noticed the tweets by influencers (labeled as 1) are shorter in travel tweets whereas in case of fashion and fitness tweets the average length of tweets are the same for both influencer and non influencer labels. The average tweet length for both influencer and non-influencer categories are about 25 incase of fashion and fitness tweets. In case of travel tweets, the average tweet length for influencer ('1') is about 20 and non-influencer ('0') is about 24.

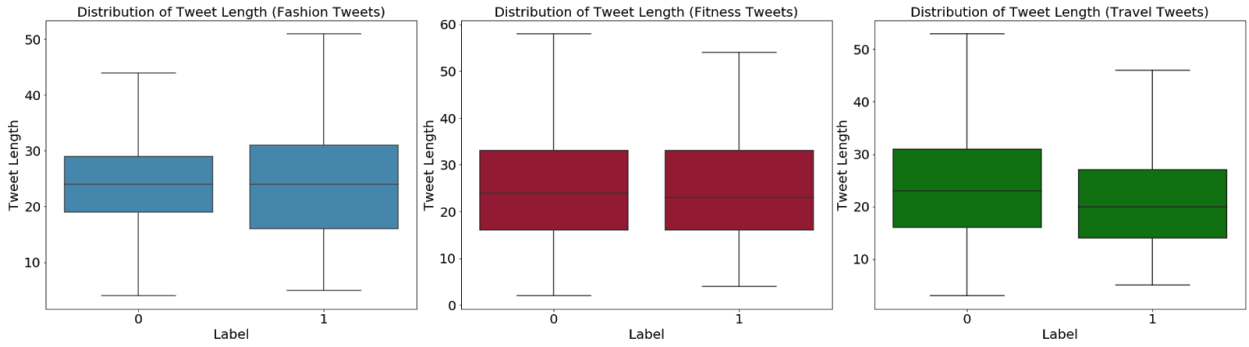


Fig 1 Distribution of tweet length in different labels

5.2 Distribution of % of Caps per tweet in different labels

From the distribution of % of caps per tweet (**Fig 2**) we can see in all the three tweets categories (fashion, fitness and travel) the influencer tweets contain more uppercase letters.

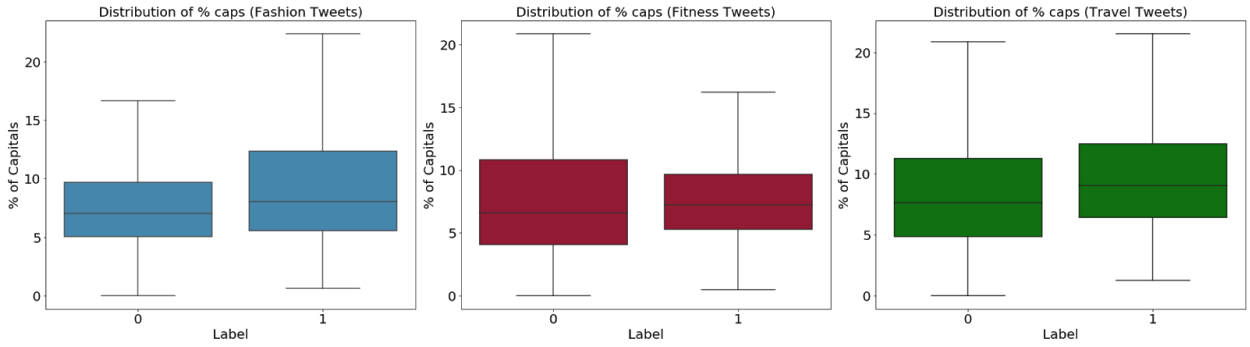


Fig 2 Distribution of % caps per tweet in different labels

The % of capitals shows an unsymmetrical right skewed distribution in the fashion, fitness and travel tweets (**Fig 3**). The number of tweets decreases with increase in % caps. The highest percentage of caps per tweet is about 75%. There are only 100 tweets containing higher than 40% of capitals.

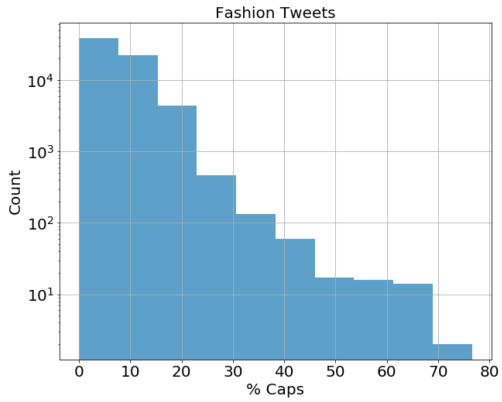


Fig 3 Distribution of % of Capitals

From the distribution of % caps (>20%) by labels (**Fig 4**), we can see more number of non-influencer tweets containing more than > 20% capital letters across all the three tweets categories.

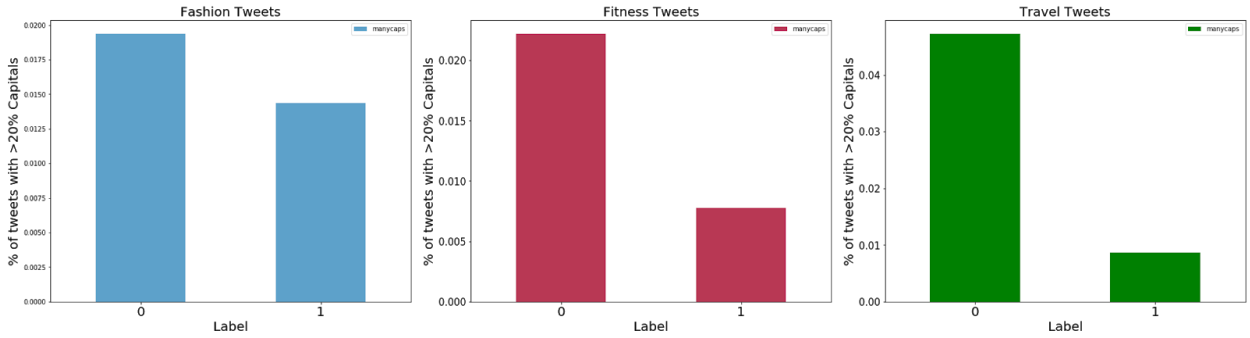


Fig 4 Distribution of % caps (>20 %) in influencer and non-influencer

5.3 Distribution of number of hashtags in different labels

The distribution of number of hashtags per tweets by labels (**Fig 5**) shows that in all the three categories of tweets (fashion, fitness and travel) the influencer tweets contain less number of hashtags.

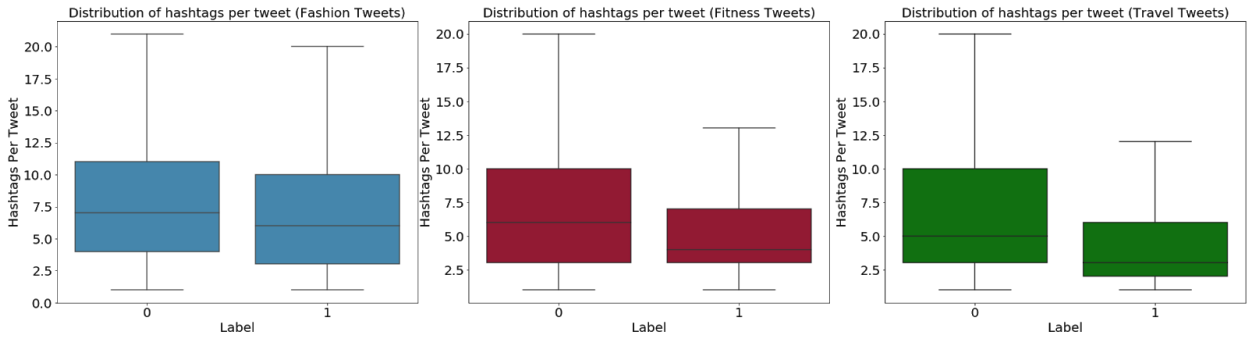


Fig 5 Distribution of number of hashtags per tweets by labels

5.4 Tweet reading level in different labels

The tweet reading level was studied using the Automated Readability Index (**Fig 6**) and it shows the influencer tweets are easy to comprehend in case of travel tweets and difficult to comprehend in case of fashion tweets. The reading level is almost the same for both influencer and non influencer for fitness tweets.

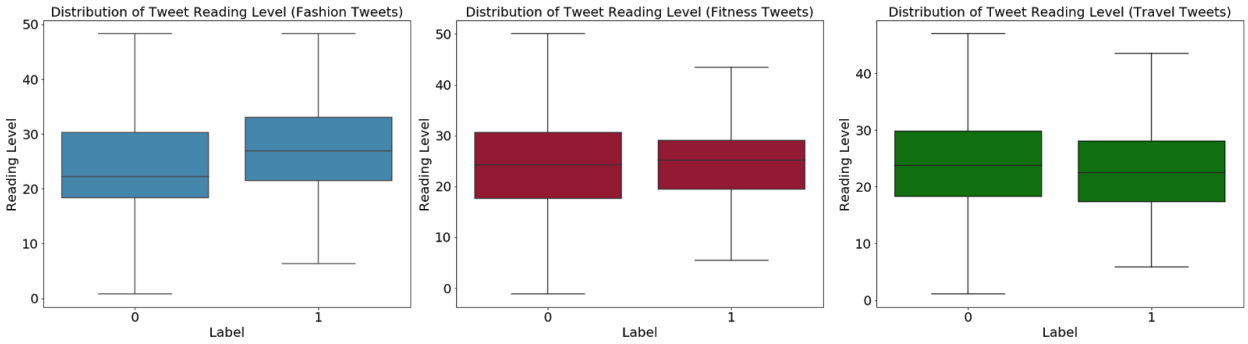


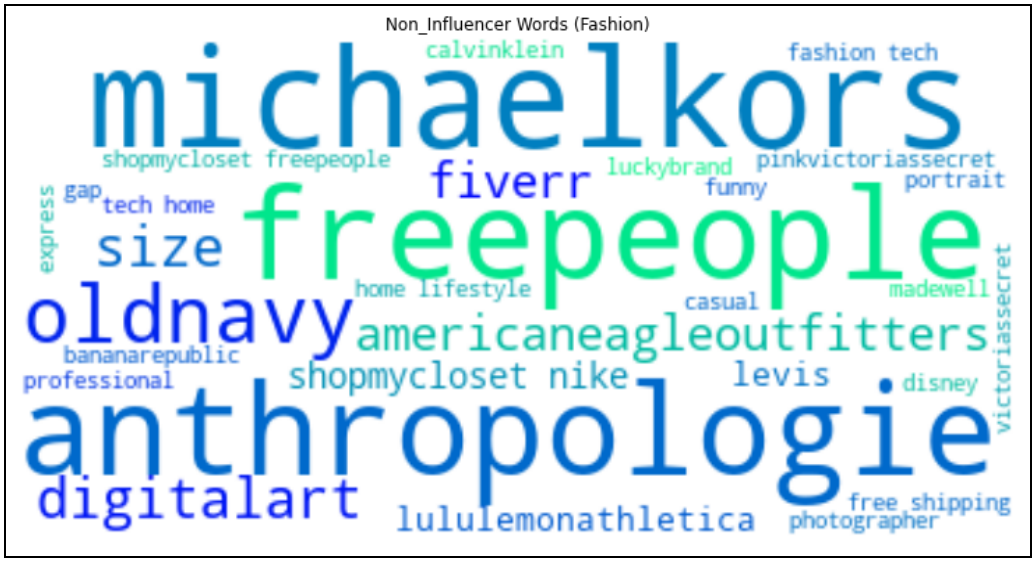
Fig 6 Automated Readability Index Reading in different labels

5.5 Most Predictive words for different tweet categories

To find the most predictive words for the three tweet categories first I preprocessed the tweet text and then created the document term matrix with TfidfVectorizer and trained the Multinomial Naive Bayes model on the matrix. Then I created an identity matrix the length of the vocabulary and used the Multinomial Naive Bayes model to predict on the identity matrix, and get a list of probabilities for each word. Then I sorted for most/least probable words for any given class and created word clouds for most predictive and least predictive words using `.generate_from_frequencies()` method.

I looked at the 30 most predictive words in influencer and non-influencer tweets within the entire corpus of tweets along with their probabilities for all the three different datasets.

Fashion Tweets



As we can see the influencers in fashion tweets varieties of topics like ‘artwork’, ‘photograph’, ‘retailer’, ‘amazonprimeday’, ‘websitedesign’, ‘home architecture’ etc., related to fashion in the tweets. The non-influencers also tweet about varieties of topics but mostly about different brands like: ‘calvinklein’, ‘gap’, ‘levis’, ‘disney’, ‘oldnavy’ etc. In the case of fashion tweets influencers talk about fashion ideas whereas non-influencers talk about brands.

Here are the examples of influencer and non-influencer tweets containing words like ‘websitedesign’ and ‘artwork’. As we can see, a lot of these “fashion ideas” are expressed as a way for influencers to peddle their content.

Influencer Tweets:
'How To Make A Readable Website If you want to be noticed online, you need to have a website that people can enjoy and read easily. We can help step by step: https://t.co/YqAiePd26Y #website #webdesign #marketing #design #digitalmarketing #seo #websitedesign #bhfyp #RT https://t.co/GJ4oxdzqDt'

'HOW TO GROW UP YOUR BUSINESS(EVEN IN LOCKDOWN) https://t.co/gPCllBxk8D #website #webdesign #marketing #design #digitalmarketing #seo #websitedesign #web #business #branding #webdevelopment #graphicdesign #socialmedia #webdesigner #wordpress #ecommerce #webdeveloper #earnmoney'

'Share your Digital Assignments with experts who work with 100+ brands Call +91 98404 13319 https://t.co/9RSW0aqnoN#websiterevamp #websiteredesign #websitedesignservice #godigitell #microsites #cmswebsites #wordpress #digitalagency #digitalmarketing #design #creativedesign https://t.co/06d7NXGgbF'

'Creativity is not a competition 🔄 It has to be felt 🎉 #creativity #art #creative #artist #love #design #photography #handmade #artwork #inspiration #drawing #painting #instagram #artistsoninstagram #instagood #illustration #fashion #create #graphicdesign #nature #photooftheday https://t.co/Y28PTX7XNo'

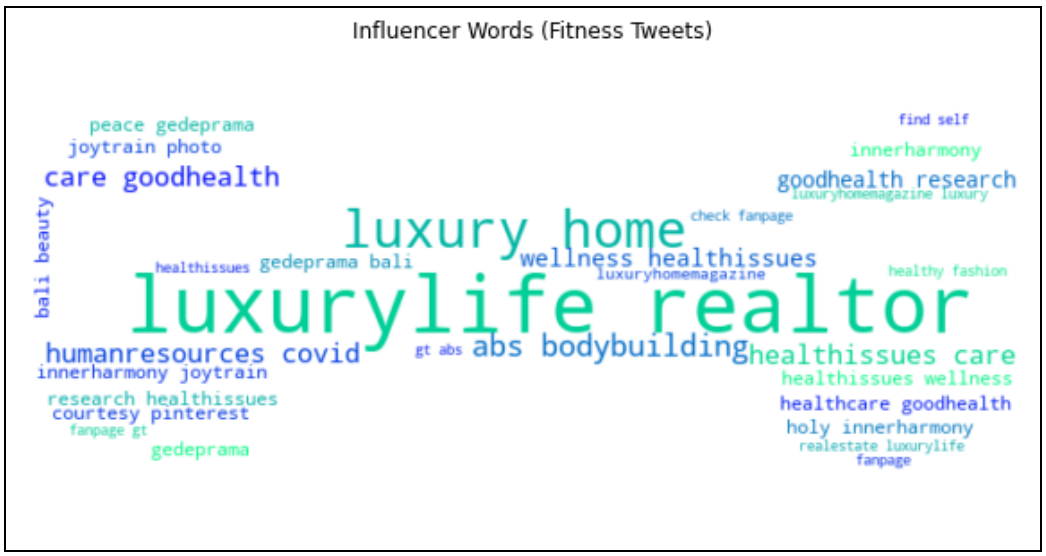
Non_influencer Tweets:
'I’m forever grateful for my opportunities to share some time and photograph the amazing Cameron Boyce. #cameronboyce #thecameronboycefoundatio #xmob_bboytruth #disney #actor #model #photoshoot

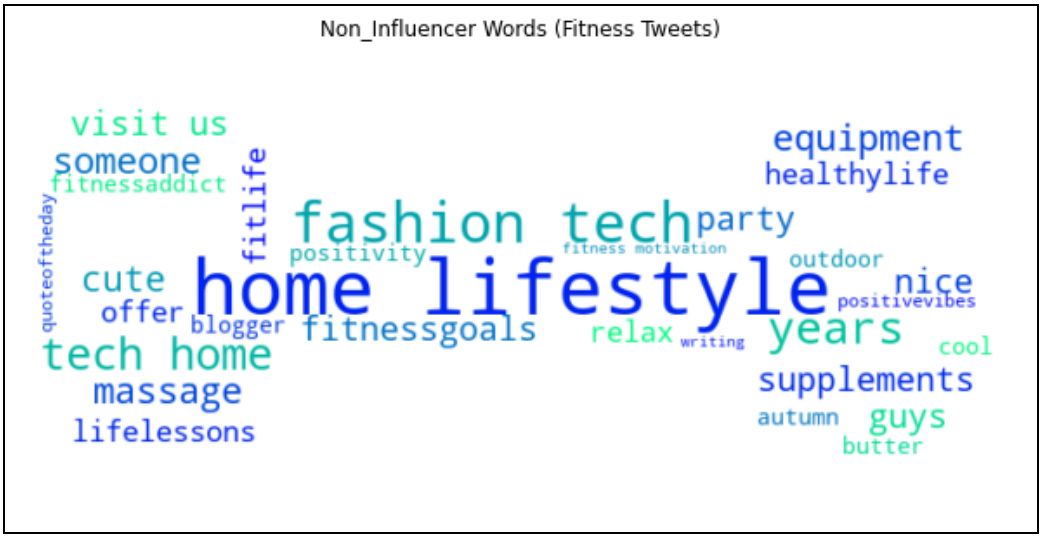
#patrickshipstad #photographer #descendants #paulcbuff #thede… https://t.co/k68UyV7HMN https://t.co/g49mMUyMNO'

"So good I had to share! Check out all the items I'm loving on @Poshmarkapp #poshmark #fashion #style #shopmycloset #blueasphalt #oldnavy: https://t.co/2XugK1t8RO https://t.co/8b2Rs4qjLc"

As we can see in these examples tweets containing the word 'websitedesign' influencer talk about ideas on how to grow up business in lockdown during COVID-19, how to be noticed online and make readable websites etc. Hence people follow the fashion influencer for fashion ideas.

Fitness Tweets





The most predictive words for fitness tweets include health and wellness tweets. The influencer tweets include the words like 'fanpage', 'luxuryhomemagazine', 'gedeprama' to promote the name of the people or website or magazine in their tweets.

Influencer Tweets:
'Greater Nashville Beautifully maintained home in ultra-convenient neighborhood https://t.co/zoATBvpgQB Listed by: Richard Bryan | Fridrich and Clark Realty #luxuryhomemagazine #luxury #home #architecture #design #inspiration #lifestyle #realestate #luxurylife #realtor #tennessee https://t.co/FqcEJvMBuN'

'Greater Sacramento Spectacular estate in the desirable Catta Verdera https://t.co/QcvoIY4NGb Listed by: Sean Work | EXP Realty #luxuryhomemagazine #luxury #home #architecture #design #inspiration #lifestyle #decor #magazine #realestate #luxurylife #realtor #california https://t.co/FEM2tpENrU'

Non-influencer Tweets:
'Our first blog post, the reasons behind wanting to start this blog and improve our health https://t.co/sxrZnqKj9z #healthylifestyle #fitnessgoals #cleaneating https://t.co/Gsfdl4UJkW'

'PB at the gym today, 110kg deadlift 🦵🟧 so proud of myself #fitness #fitnessgoals #personaltraining #gym #bodytransformation #motivated #workout #thisgirlcan #gettingfit #gettinghealthy #weightloss #girlswholift #toning #strongnotskinny #deadlifts #weighttraining #110kgdeadlift'

Luxury Home Magazine promotes new home or real estate companies at various places and helps home buyers. Gede Prama as a spiritual influencer talking about life, mental health, peace, meditation, prayer, kindness etc. Since these tweets scraped during the COVID-19 period the influencer tweets also have a word 'humanresources covid' which might include the health tips regarding COVID-19. The non-influencer tweets include the common words like 'positivevibes', 'fitness motivation', 'fitnessgoals' rather than promoting any health and wellness brands. In the case of fitness tweets, it seems like influencers have lots of promotions, mentioning medical or health brands on their tweets whereas non-influencer tweets are more focused on lifestyle and motivation.

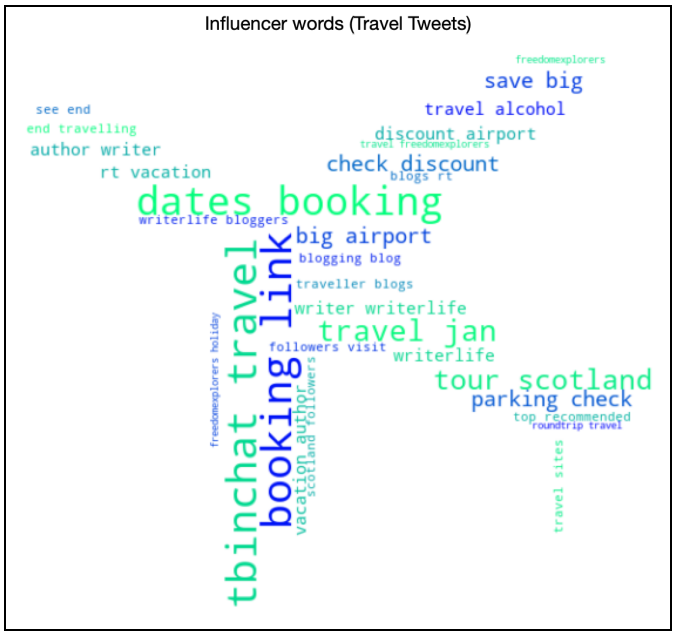
Travel Tweets

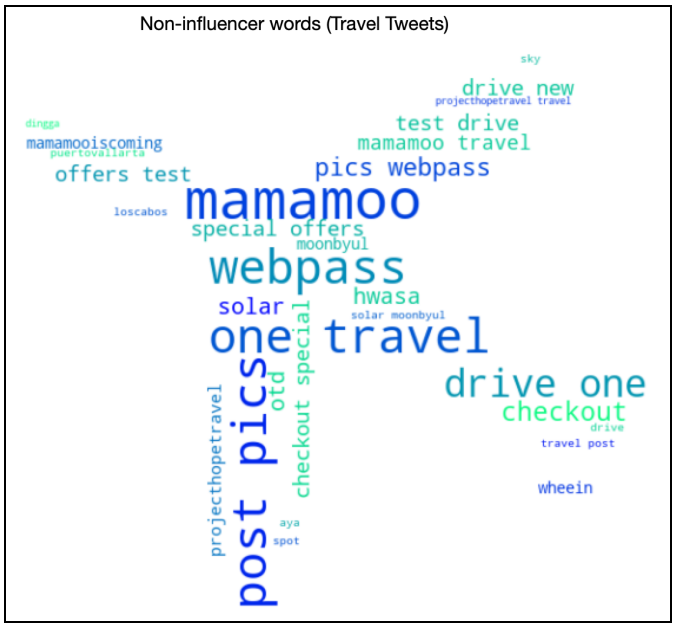
In the travel dataset the influencer tweets include words like 'travel site', 'traveller blogs', 'writer life', 'blogs', 'writer', 'vacation author', 'discount airport', 'check discount', 'save big'. So, the influencers post about the travel website, blogs, bloggers /writers on travel, and about special offers about airports or any travel related businesses.
Below are examples of influencer tweets containing words like 'discount airport', 'travel site' where they talk about discounts at airport parkinga and recommending travel sites.

Influencer Tweets:
"TripAdvisor CEO says Google uses its 'dominance in internet gatekeeping' against travel sites #wanderlust #travelgram #adventure\n#travel #travelling #traveltheworld\n- - Travel to Plan - Plan to Travel - - \n- - Read the Full Story Here - https://t.co/FEZ5d17Tev"

'#airport #aviation #airpllane #airbus #aircraft #travel Save Big on Airport Parking. Check for discount airport parking on https://t.co/Hk21gwU6Ai'
'#entrepreneurs #AffiliateMarketing #tourism #travel #marketing #DigitalMarketing #EmailMarketing #InfluencerMarketing Here my top recommended travel sites! https://t.co/Hk21gwU6Ai'

Non-influencer Tweets:
'#Nebraska @VillagePointeToyota checkout Special Offers and test drive the all NEW @Toyota #Camry #TRD drive one #travel and post pics @MyNextToyota @CarandDriver @Motortrend #WebPass https://t.co/0uRU9rVdJz'





Non-influencer tweets contain the words like ‘test drive’, ‘drive new’, ‘offers test’, ‘special offer’, ‘travel post’ etc. So, the non-influencer mostly tweets about their personal travel experience and also about some special offers about test drives.

The influencer and non-influencer tweets are different across different categories of tweets. At a basic level fitness and travel seem to be more focused on promotion than fashion which seems to be more focused on lifestyle, perhaps because people follow fashion influencers for their artistic merit whereas fitness and travel people follow to get tips.

5.6 Summary of findings from Exploratory Data Analysis

From the exploratory data analysis, we can see in all the three categories of tweets:

- The average tweet length was the same for both influencer and no-influencer in case of fashion and fitness tweets, however for travel tweets the influencers wrote short tweets.
- The % of caps distributions are unsymmetrical right skewed, the average % of uppercase letters was higher for influencers tweets. However, the non-influencers have a larger number of tweets containing more >20% capitals.
- The influencer tweets contain less number of hashtags than the non-influencer tweets.
- The reading level of the tweets measured by Automated Readability Index shows that the influencer fashion tweets has higher reading level, however, for both fitness and travel influencer and non-influencer tweets have similar reading level.

6. Machine Learning

The aim of machine learning is to classify the new tweets as influencer tweets or non-influencer tweets. Before machine learning, we need to pre-process the tweet text followed by splitting the datasets into training and test sets and then do vectorization to encode text data into numeric values. I used two vectorization methods: CountVectorizer and TfidfVectorizer from Scikit-Learn library and selected the best performing method. Then I tuned the hyperparameters and compared three different machine learning algorithms using training datasets and predicted influencer using test datasets.

6.1 Text Pre-Processing

I preprocessed the tweets using several preprocessing packages. I removed URL, Stopwords, mentions, punctuations, numbers and special characters except hashtags. I created a new column called clean_tweets and saved the preprocessed tweets.

6.2 Train and Test Split

I split the data into training and test sets using Scikit-learn's `train_test_split` method with a `test_size` of 30%.

6.3 Vectorizer Selection

Here I used two vectorization methods: CountVectorizer and TfidfVectorizer provided by Scikit-Learn's library to encode the text data to numeric values. First I instantiated both CountVectorizer and TfidfVectorizer with default parameters (min_df=1, ngram_range=(1, 2)) and then saved the document-term matrices from the fit and transform steps of vectorization on the X_train and X_test matrices of all the three categories of tweets datasets. Then I instantiated a simple Naive Bayes classifier, MultinomialNB(), and trained it on the X_train document term matrix and y_train and then made class prediction using X_test document term matrix and calculated ROC-AUC scores to compare both the vectorizers.
Then I gridsearch the min_df value for both the vectorizers on a simple Naive Bayes classifier. Then I recalculated the ROC-AUC scores after hyperparameter tuning. **Table 1** compares the vectorizers with and without hyperparameter tuning for three datasets. Looking at the table we can see TfidfVectorizer worked best across all scenarios and hence I used the document terms matrix from the TfidfVectorizer on other classifiers.

Table 1 Comparison of vectorizers with a Multinomial Naive Bayes Model with and without hyperparameter tuning for fashion, fitness and travel twitter datasets.

Fashion Tweets		
Vectorizer	**ROC-AUC**	**Best Parameters**
CountVectorizer	0.699	min_df = 1, alpha=1
TfidfVectorizer	0.881	min_df = 1, alpha =1
CountVec w/ GridSearch	0.865	min_df = 50, alpha =0.001
TfidfVec w/ GridSearch	0.880	min_df =50, alpha =0.01
Fitness Tweets		
Vectorizer	**ROC-AUC**	**Best Parameters**
CountVectorizer	0.71	min_df = 1, alpha=1
TfidfVectorizer	0.817	min_df = 1, alpha=1
CountVec w/ GridSearch	0.869	min_df = 20, alpha=0.01
TfidfVec w/ GridSearch	0.881	min_df = 20, alpha=0.01
Travel Tweets		
Vectorizer	**ROC-AUC**	**Best Parameters**
CountVectorizer	0.776	min_df = 1, alpha=1

TfidfVectorizer	0.874	min_df = 1, alpha=1
CountVec w/ GridSearch	0.884	min_df = 20, alpha=0.1
TfidfVec w/ GridSearch	0.888	min_df = 20, alpha=0.1

6.4 Model Comparison

I fit and tuned three classifiers: Multinomial Naive Bayes, Logistic Regression, Random Forest Trees using TfidfVectorizer. For all the three classifiers I used ngram_range of (1,2) and min_df = 20 - 50 and compared the models using ROC-AUC scores. For three of the models I used GridSearchCV for cross-validation and hyper parameter tuning.

The score and parameters of each model are shown in **Table 2** along with the best ROC curve (**Fig 7**) for all the three datasets. The highest scoring model turned out to be Random Forest across all categories.

Table 2 Comparison of three machine learning models fitted with TfidfVectorizer for three datasets

Fashion Tweets		
Classifier	**ROC-AUC**	**Best Parameters**
MultinomialNB	0.880	alpha =0.01, fit_prior = True
LogisticRegressionCV	0.879	C= 3.25, l1_ratio=0
RandomForestClassifier	0.892	max_depth= 100, max_feature= auto, n_estimators= 300
Fitness Tweets		
Classifier	**ROC-AUC**	**Best Parameters**
MultinomialNB	0.881	alpha=0.01, fit_prior=True
LogisticRegressionCV	0.861	C= 3.25, l1_ratio= 0
RandomForestClassifier	0.898	max_depth= None, max_features= sqrt, n_estimators=500
Travel Tweets		
Classifier	**ROC-AUC**	**Best Parameters**
MultinomialNB	0.888	alpha =0.1, fit_prior = True
LogisticRegressionCV	0.890	C=3.25, l1_ratio= 0

RandomForestClassifier	0.919	max_depth= None, max_features= sqrt, n_estimator= 300

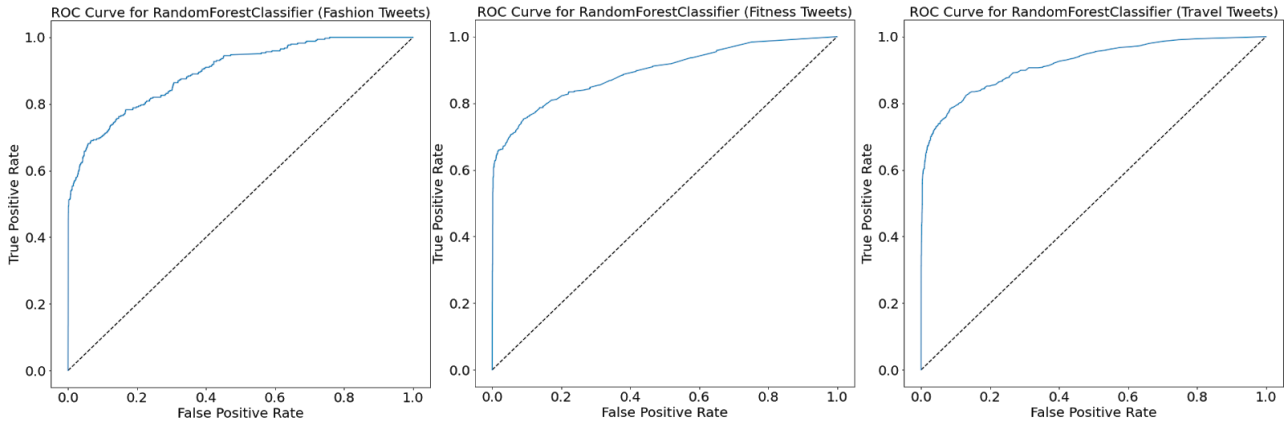


Fig 7 ROC curve for Random Forest (best model) for three datasets

6.5 Thresholding

The default threshold for interpreting probabilities to class labels is 0.5, meaning if the predicted probability of a tweet is greater than equal to 0.5 then that tweet is classified as an influencer tweet else it is classified as an non-influencer tweet, however this isn't necessarily the ideal threshold especially given the class imbalance in this dataset.

Choosing a threshold is a business decision based on the business scenario in which this model might be used. As described in the Client Profile, the clients might be brands who'd want to:

1) find potential influencers who are doing the right things but are still building a following (and may be cheaper)

2) separate users who are influencers from users that simply have high follower counts for other reasons but do not make the types of posts that would be most likely to help their brand.

In both of these cases, the class that we are focusing on is influencers and we want to find a balance between precision and recall to isolate influencers from non-influencers so the client might reach out to them. As such, F1 score, focused on influencers, seems like an appropriate metric to optimize.

For the best classifier, Random Forest, the best threshold and best scores are shown in **Table 3** below for all the three datasets. The precision, recall and f1 curves with thresholding for all the three datasets are shown in the **Figure 8**.

Table 3 Best threshold and F1-score for the three datasets

Dataset	Optimal Threshold	F1-score
Fashion tweets	0.493	0.640
Fitness tweets	0.267	0.709
Travel tweets	0.330	0.729

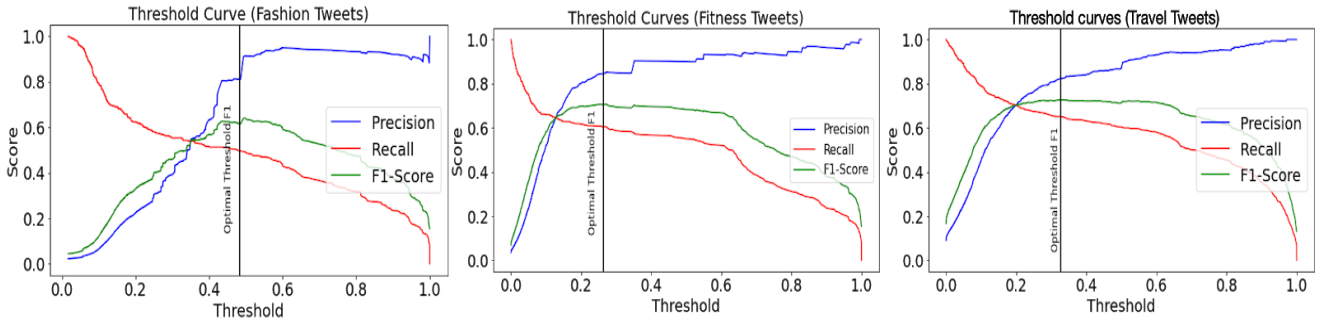


Fig 8 Precision, Recall and F1 Curves with Thresholding

7. Summary

In this project I built a model that could detect and separate influencers from non-influencers based on their tweets. One interesting finding was that influencer tweets contain less number of hashtags than non-influencer in the three categories of tweets, as well as a higher average average % of uppercase letters.

From the most predictive words for different tweet categories, we found that, in the case of fashion tweets influencers talk about fashion ideas whereas non-influencers talk about brands. Hence people follow the fashion influencer for fashion ideas. The influencer and non-influencer tweets are different across different categories of tweets. At a basic level fitness and travel seem to be more focused on promotion than fashion which seems to be more focused on lifestyle, perhaps because people follow fashion influencers for their artistic merit whereas fitness and travel people follow to get tips.

Some possible next steps might include:

Using these models to look at other categories of tweets in which influencers operate.

1) Adapting these models to detect influencers on Instagram, YouTube or other social media networks.
2) Building a Web App to help people in real time to provide feedback on how their tweets could be more similar to an influencer tweet.