Detecting influencer on twitter across different genres:
Capstone 2 Milestone Report 1

By Pravati Swain

Table of Contents

1. Introduction
2. Client profile
3. Description of Dataset
4. Data Wrangling
5. Exploratory Data Analysis
6. Machine Learning
7. Summary

1. Introduction

Word-of-mouth (WOD) business strategy has been regarded as an important mechanism by which information about new products or brands or topics can reach large populations. Social media helps people to connect with other people anywhere and let them share their opinions regarding any matter. For the WOD marketing, companies need to find the social media influencers to spread information about their new products or brand more effectively. Twitter is one of the most popular microblogging platforms allowed to publicly discuss various topics or products using tweets. In this project, I propose building a supervised machine learning model to detect the Twitter influencer based on tweets across different categories. For this project I will be using tweets from three different hashtags: #fashion, #fitness and #travel tweets. Number of followers of Twitter users provide the labels for training ML models to predict the influencer based on their tweets regarding a particular product or brand.

2. Client Profile

Social media companies like Twitter, Instagram, YouTube etc can use my machine learning model to predict the influencer based on their social media account information.

3. Description of Datasets

For this project, I scraped the tweets from 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 scraped from 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 used the df.shape attribute to find out the number of rows and columns and df.info() method to check for missing tweets.
- I checked for duplicate tweets and retweets and removed them.
- There are some tweets without search hashtags and I looked for them and removed them.
- 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.
- After preprocessing the datasets were saved as a pickled file to further use in machine learning.
- All the above steps were carried out for the datasets.

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 label ('1') is about 20 and non-influencer ('0') is about 24.

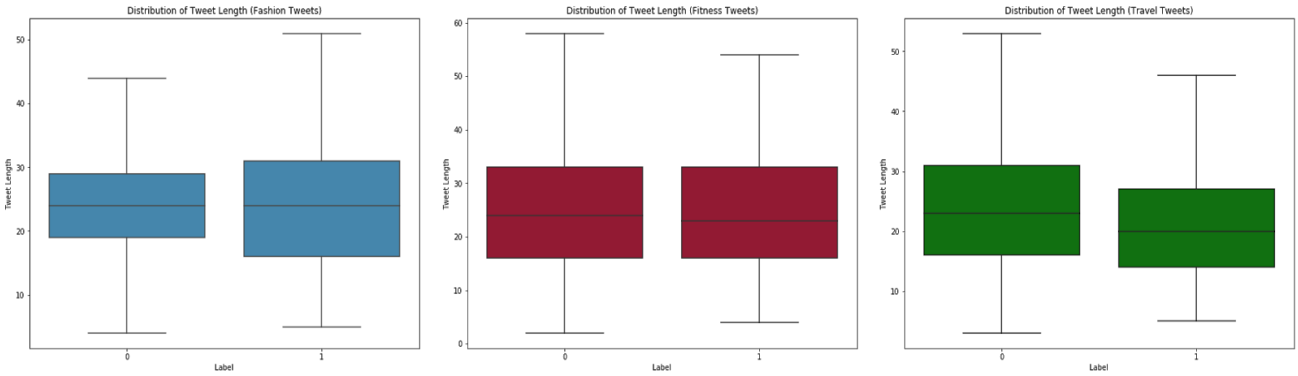

Fig 1 Distribution of tweet length in different labels

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 influencers tweets contain more uppercase letters.

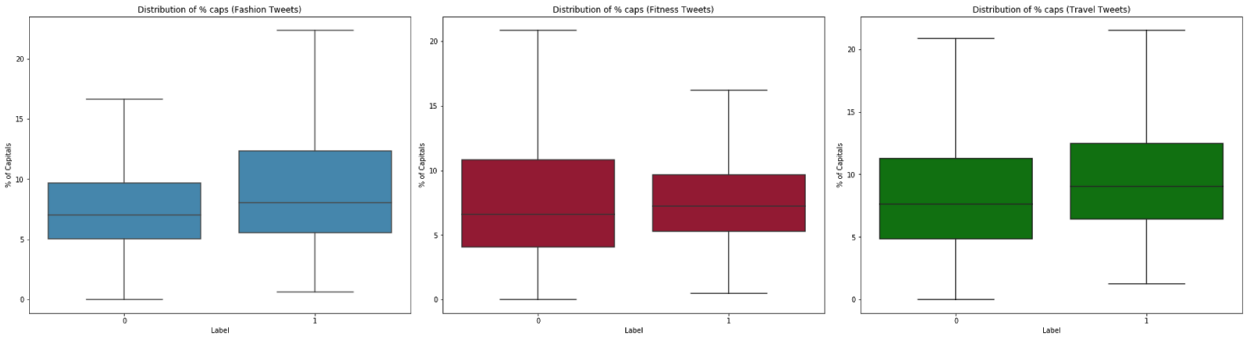

Fig 2 Distribution of % caps per tweet in different labels

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.

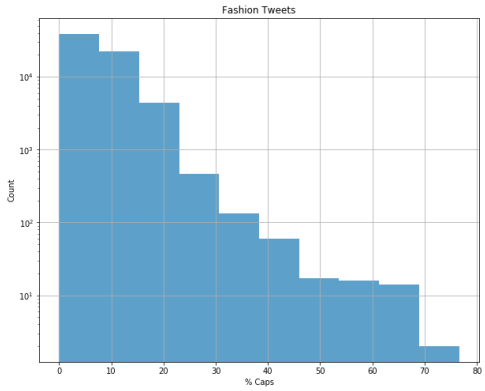

Fig 3 Distribution of % of Capitals

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.

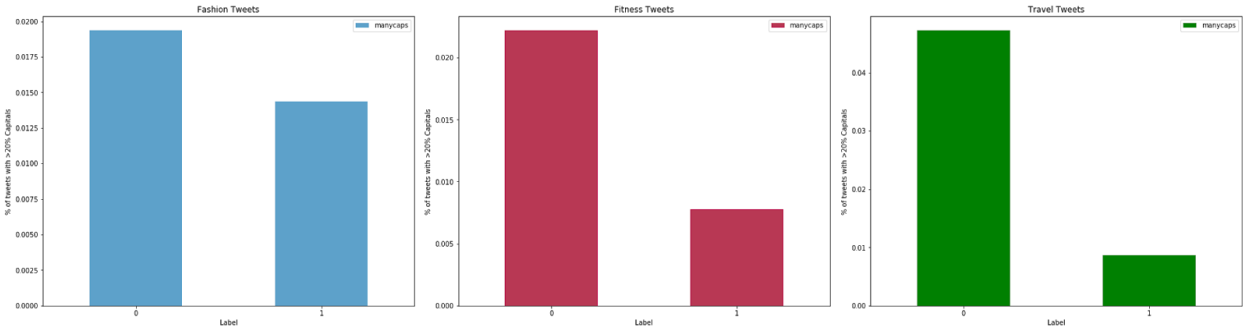

Fig 4 Distribution of % caps (>20 %) in influencer and non-influencer

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.

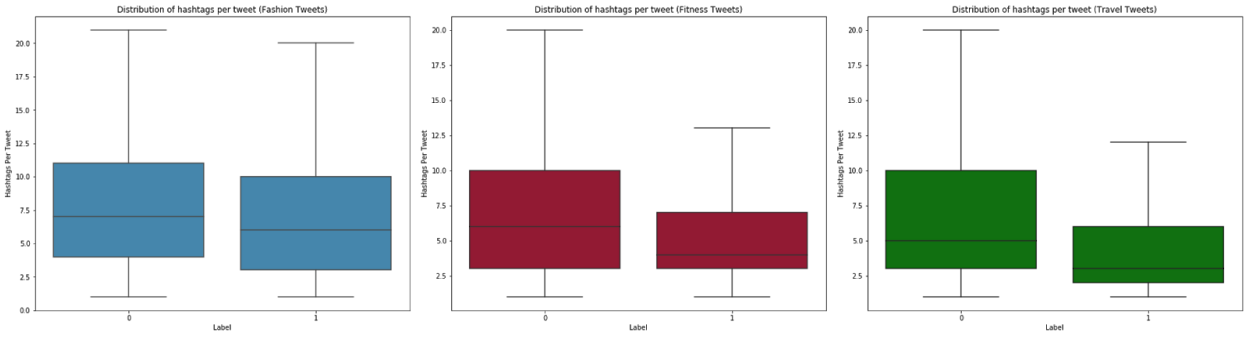

Fig 5 Distribution of number of hashtags per tweets by labels

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.

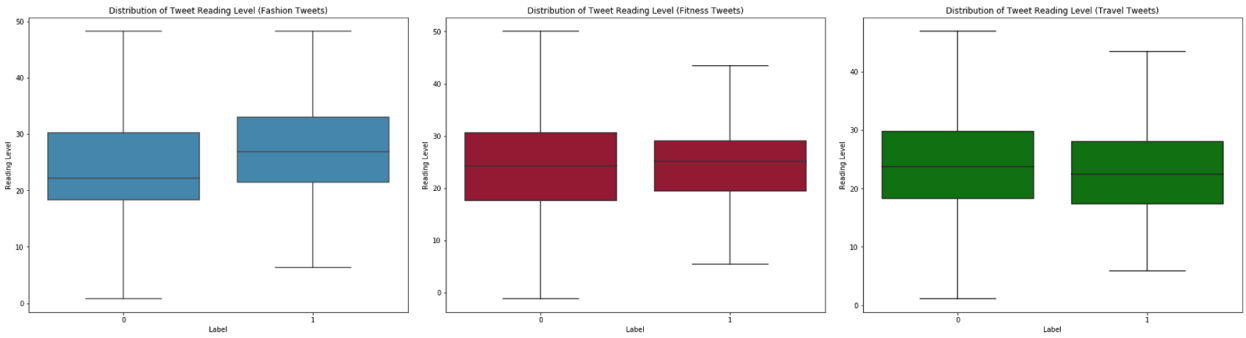

Fig 6 Automated Readability Index Reading in different labels

5.5 Most Predictive words for different tweet categories

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

Influencer words	P(influencer | word)
home architecture	0.80
courtesy	0.77
luxurylife	0.62
listed	0.61
artworks	0.53
stuff	0.53
pinterest	0.52
photograph	0.52
design inspiration	0.51
healthy	0.42
naturelovers	0.40
followers	0.36
famous	0.36
uiux	0.32
peace	0.31
alert	0.28
entertainment	0.27
mugs	0.26
realtor	0.25
dribbble	0.24
websitedesign	0.23
amazonprimeday	0.23
beautiful fashion	0.23
estate	0.23
template	0.23
webdesigner	0.22
information	0.22
websites	0.21
enter	0.20

info	0.20

Non Influencer words	P(Influencer | word)
calvinklein	0.00
madewell	0.00
tech home	0.00
professional	0.00
photographer	0.00
home lifestyle	0.00
funny	0.00
portrait	0.00
luckybrand	0.00
fashion tech	0.00
casual	0.00
gap	0.00
express	0.00
disney	0.00
free shipping	0.00
Pinkvictoriassecret	0.00
bananarepublic	0.00
victoriassecret	0.00
shopmycloset freepeople	0.00
levis	0.00
lululemonathletica	0.00
shopmycloset nike	0.00
fiverr	0.00
americaneagleoutfitters	0.00
size	0.00
digitalart	0.00
oldnavy	0.00
michaelkors	0.00
anthropologie	0.00
freepeople	0.00

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.

Fitness Tweets

Influencer words	P(influencer | word)
fanpage	1.00
check fanpage	1.00
gt abs	1.00
fanpage gt	1.00
healthy fashion	1.00

healthissues	1.00
find self	1.00
luxuryhomemagazine	1.00
luxuryhomemagazine luxury	1.00
realestate luxurylife	1.00
peace gedeprama	1.00
gedeprama bali	1.00
innerharmony	1.00
innerharmony joytrain	1.00
bali beauty	1.00
courtesy pinterest	1.00
gedeprama	1.00
holy innerharmony	1.00
joytrain photo	1.00
goodhealth research	1.00
healthcare goodhealth	1.00
research healthissues	1.00
healthissues wellness	1.00
healthissues care	1.00
care goodhealth	1.00
wellness healthissues	1.00
humanresources covid	1.00
abs bodybuilding	0.97
luxurylife realtor	0.97
luxury home	0.97

Non Influencer words	**P(Influencer | word)**
writing	0.00
fitness motivation	0.00
quoteoftheday	0.00
positivevibes	0.00
blogger	0.00
cool	0.00
fitnessaddict	0.00
outdoor	0.00
butter	0.00
autumn	0.00
positivity	0.00
offer	0.00
relax	0.00
lifelessons	0.00
fitlife	0.00
healthylife	0.00
visit us	0.00
party	0.00
nice	0.00

guys	0.00
supplements	0.00
massage	0.00
someone	0.00
equipment	0.00
cute	0.00
fitnessgoals	0.00
years	0.00
tech home	0.00
fashion tech	0.00
home lifestyle	0.00

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. 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

Travel Tweets

Influencer words	**P(influencer | word)**
roundtrip travel	1.00
freedomexplorers holiday	0.99
freedomexplorers	0.99
travel freedomexplorers	0.99
followers visit	0.99
scotland followers	0.99
see end	0.99
end travelling	0.99
travel sites	0.99
top recommended	0.99
writerlife	0.99
writerlife bloggers	0.99
traveller blogs	0.99
blogging blog	0.99
blogs rt	0.99
travel alcohol	0.99
writer writerlife	0.99
rt vacation	0.99
author writer	0.99
vacation author	0.99
discount airport	0.98

parking check	0.98
big airport	0.98
save big	0.98
check discount	0.98
travel jan	0.97
tour scotland	0.97
booking link	0.97
tbinchat travel	0.97
dates booking	0.95

Non Influencer words	**P(Influencer | word)**
loscabos	0.00
puertovallarta	0.00
travel post	0.00
drive	0.00
sky	0.00
aya	0.00
solar moonbyul	0.00
spot	0.00
dingga	0.00
projecthopetravel	0.00
projecthopetravel travel	0.00
moonbyul	0.00
wheein	0.00
mamamooiscoming	0.00
hwasa	0.00
solar	0.00
otd	0.00
mamamoo travel	0.00
test drive	0.00
checkout special	0.00
offers test	0.00
drive new	0.00
special offers	0.00
checkout	0.00
pics webpass	0.00
webpass	0.00
post pics	0.00
drive one	0.00
one travel	0.00
mamamoo	0.00

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.

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.