Detecting influencer on twitter

Springboard DS Career Track Capstone 2

Pravati Swain, 06 November 2020

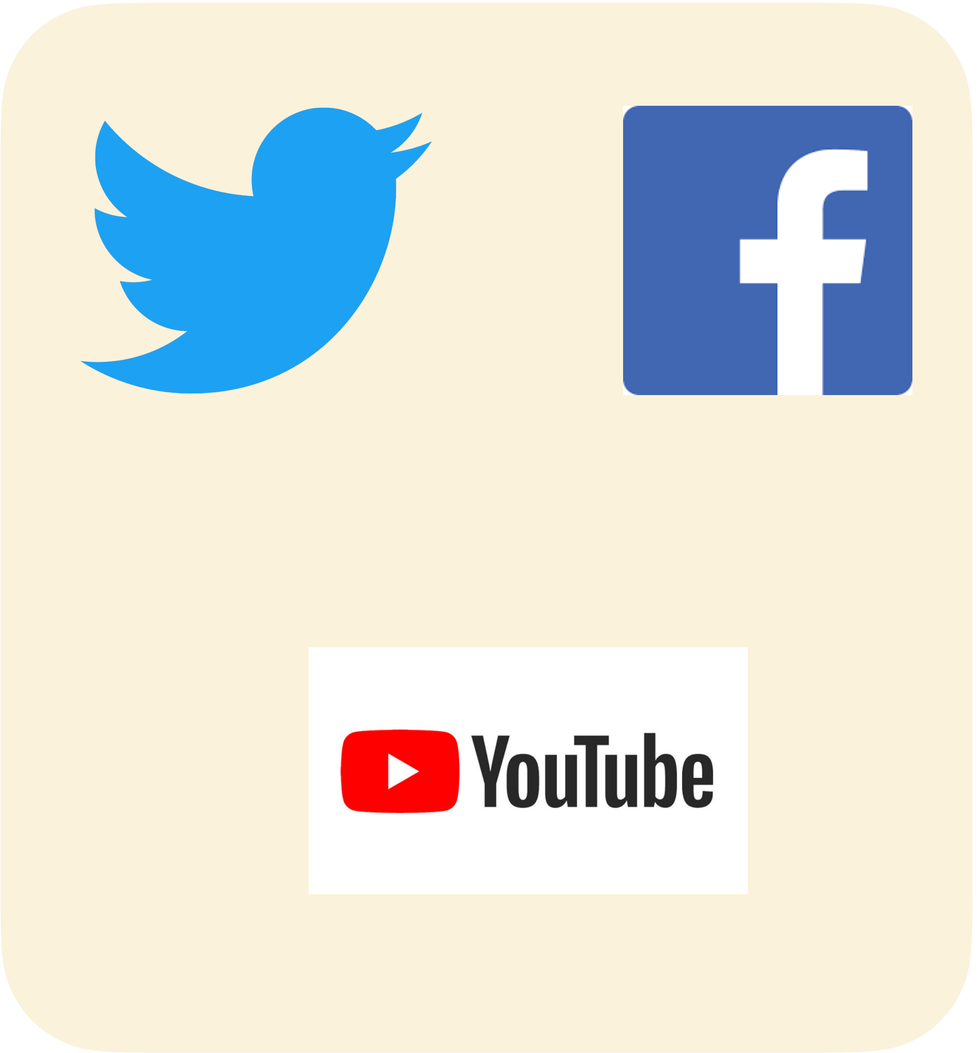
Introduction

- Word-of-mouth (WOM) business strategy is important to spread information about new products or brands or topics to a large population.
- Social media helps people to connect anywhere in the world
- Companies need to find social influencer to spread information about a new product or brand
- Twitter is one of the most popular social media platform

Aim: Analyze and build an ML model to detect Twitter influencer across different categories of tweets

What Companies Care ?

- **Social media companies**: to find out social media influencer across a particular category of product or brand



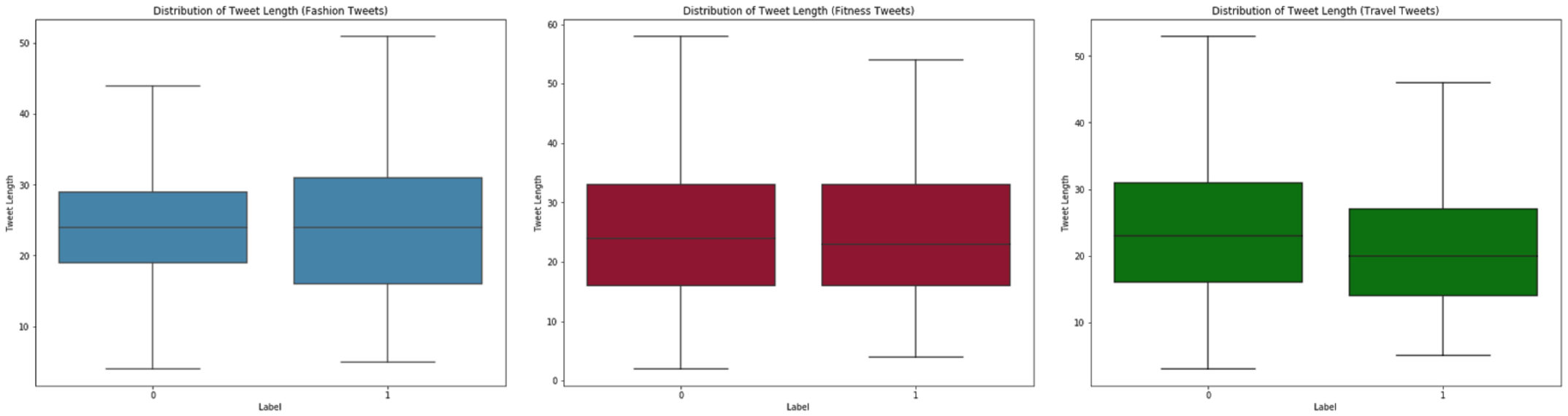
Dataset

- **Fashion Tweets, Fitness Tweets and Travel Tweets scraped from Twitter API using Tweepy Library**
- **All the three datasets contain between 135,000 to 95, 000 reviews.**
- **Mostly used features: Tweets, Followers Count**
- **Binary Classification: tweets with >30,000 followers (‘1’) and tweets with <= 5000 followers(‘0’)**

Data Wrangling

- Removed the duplicate tweets, retweets and tweets without search hashtags
- Removed URL, Stopwords, mentions, punctuations, numbers and special characters except #tags from the tweets and saved in the new column ‘clean_tweets’
- Added new feature ‘label’: tweets > 30,000 followers (‘1’) and tweets <= 5000 followers (‘0’)

Exploratory Data Analysis

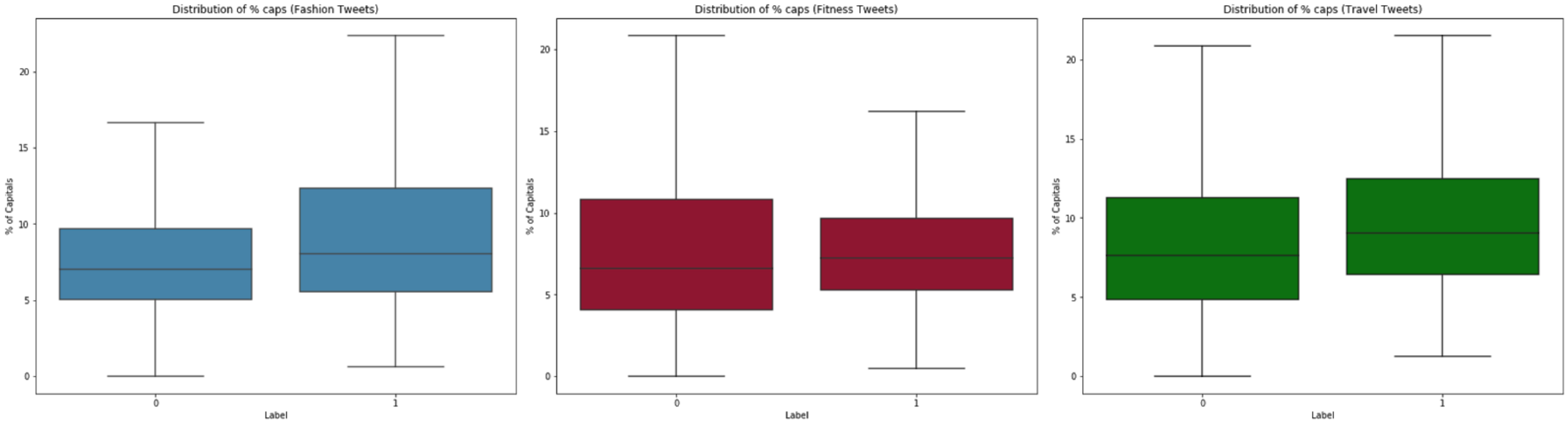


Distribution of tweet length in different labels

- **Travel tweets: Influencer tweets (20 words) are shorter than non-influencer (24 words)**

- **Fashion and Fitness Tweets: influencer and non-influencer tweets are about the same length (25 words)**

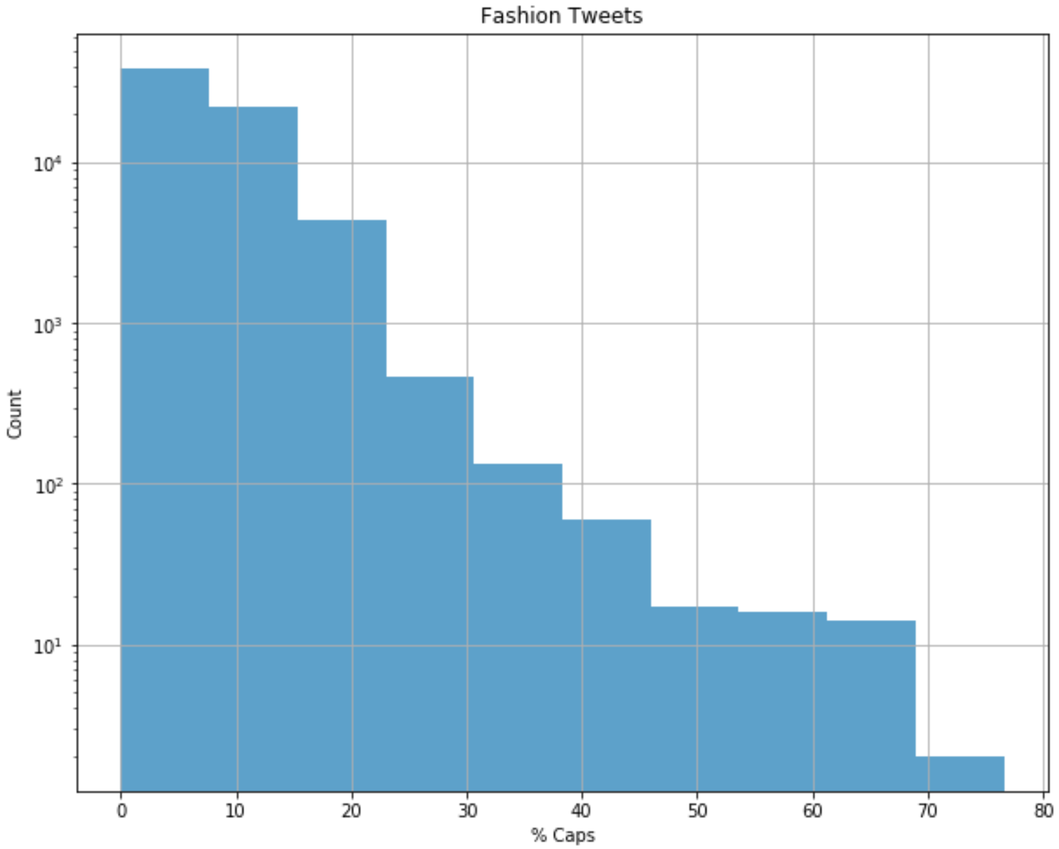
Exploratory Data Analysis



Distribution of % caps per tweet in different labels

- **The influencer tweets contain more uppercase letters in all the three categories of tweets.**

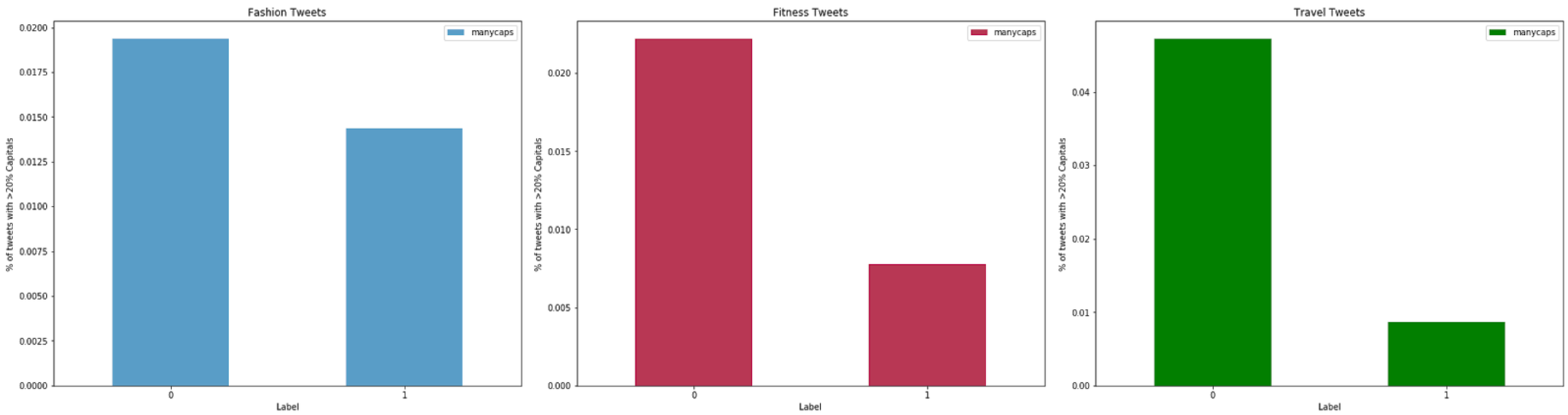
Exploratory Data Analysis



Distribution of % of Capitals

- **The distribution of % of capitals is unsymmetrical right skewed in 3 of the tweets categories**
- **The number of tweets decreased with increase in % caps**
- **There are only 100 tweets with higher than 40% capitals**

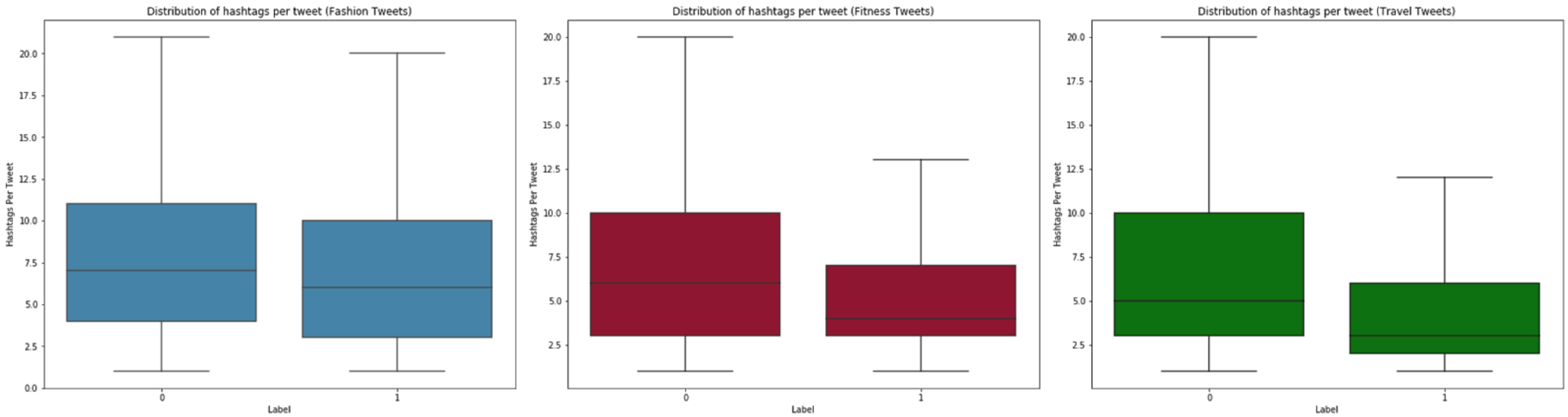
Exploratory Data Analysis



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

- **More number of non-influencer tweets contains more than > 20% capital letters, in all three categories of tweets**

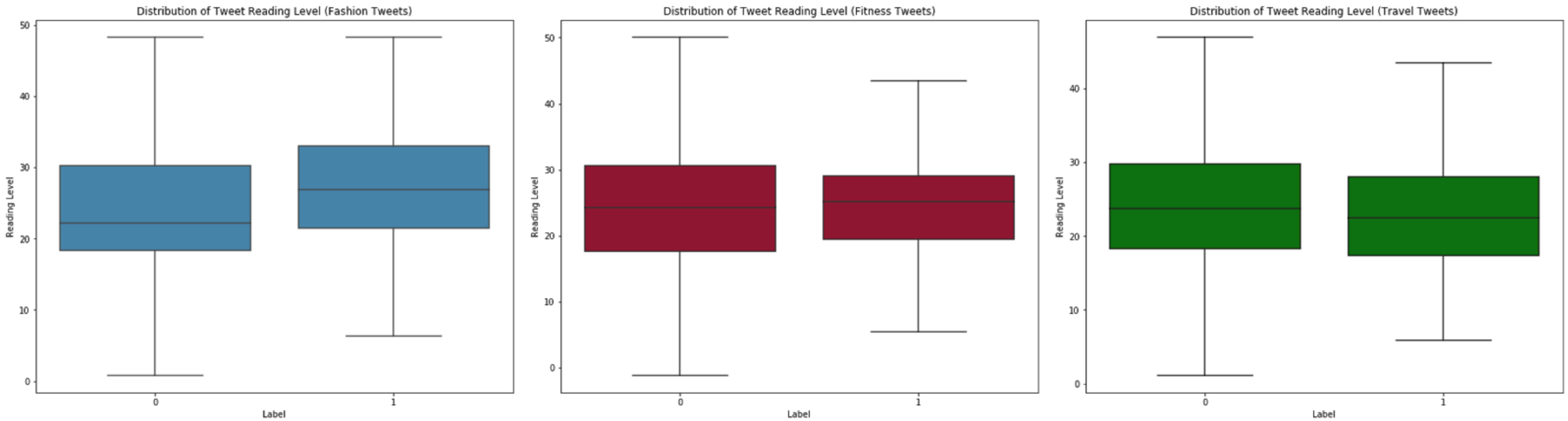
Exploratory Data Analysis



Distribution of number of hashtags per tweets by labels

The influencer tweets contain less number of hashtags.

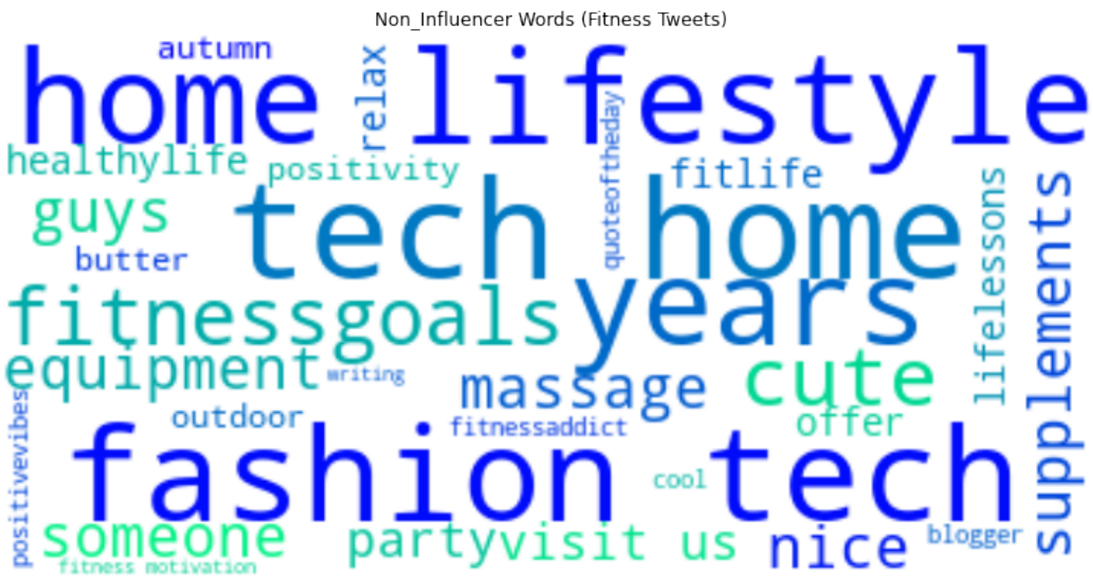
Exploratory Data Analysis



World Clouds for high and low rated reviews

- **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.**

Most Predictive words for different tweet categories



- **The most predictive words for fitness tweets include health and wellness tweets**
- **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 also have a word ‘humanresources covid’ which might include the health tips regarding COVID-19**
- **Non-influencer tweets: include the common words like ‘positivevibes’, ‘fitness motivation’, ‘fitnessgoals’ rather than promoting any health and wellness brands**

Machine Learning Highlights

Preprocessing → **Vectorization** → **Model Tuning**

- Removing URL
- Keeping only alphabets
- Removing mentions
- Removing stopwords

- Vectorizer selection
- Compared CountVectorizer and TfidfVectorizer with a Multinomial Naive Bayes Model
- Select the vectorizer with highest ROC-AUC score

- Fitted and tuned 3 classifiers: Logistic Regression, Multinomial Naive Bayes and Random Forest Trees.
- Tune with GridSearchCV
- Compare ROC-AUC scores

Vectorization

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**

Comparison of vectorizers with a Multinomial Naive Bayes Model

TfidfVectorizer worked best and used it for all the classifier

Model Comparison

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

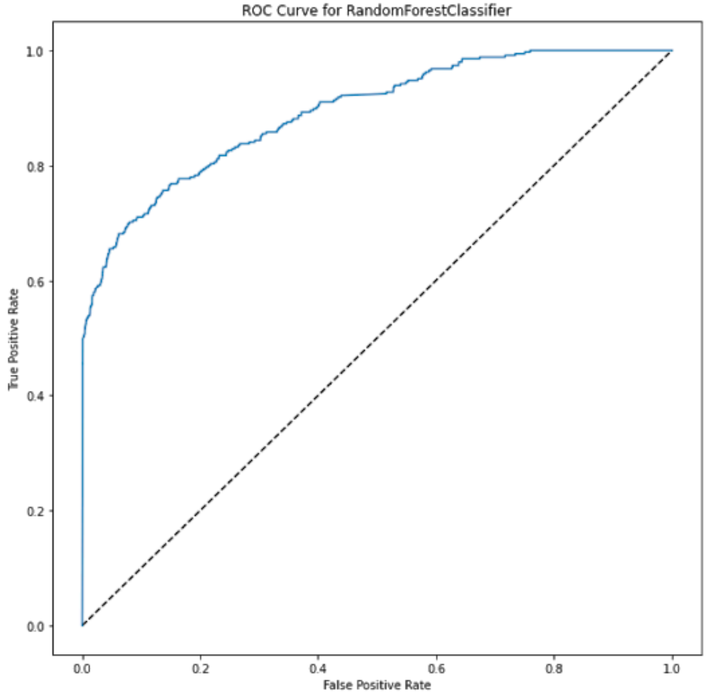
Fashion Tweets		
Classifier	**ROC-AUC**	**Best Parameters**
MultinomialNB	0.880	alpha =0.01, fit_prior = True
LogisticRegressionCV	0.879	C= 3.25, l1_ratio=0
RandomForestClassifier	0.892	max_depth= 100, max_feature= auto, n_estimators= 300
Fitness Tweets		
Classifier	**ROC-AUC**	**Best Parameters**
MultinomialNB	0.881	alpha=0.01, fit_prior=True
LogisticRegressionCV	0.861	C= 3.25, l1_ratio= 0
RandomForestClassifier	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

Best Classifier: Random Forest

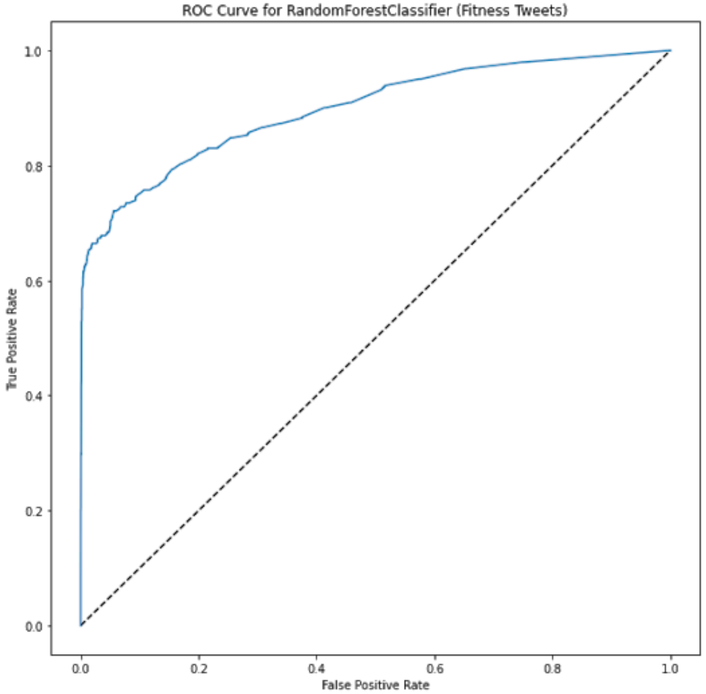
for all the 3 product categories

Best Classifier: Random Forest
ROC Curve

Fashion Tweets



Fitness Tweets



Travel Tweets