Detecting influencer on twitter across different genres

Springboard DS Career Track Capstone 2

Pravati Swain, 06 November 2020

Introduction

- In Word-of-Mouth marketing, companies use social media influencers to spread information about their new products or brand more effectively.

- Twitter: most popular social media platform

- Built a ML model to detect an influencer or potential influencer based on their tweets

- Examined what attributes of a tweet differentiate influencer from non-influencer

- Help users to achieve influencer status, and brands on what techniques are most effective for getting attention and followers.

What Companies Care ?

Two primary uses for the findings

1. **Understanding what aspects of a tweet comes from an influencer**
    - Aspiring influencers, current influencers and brands
    - To improve their following on social media

2. **Using ML model to predict whether or not someone is an influencer**
    - potential influencers (and may be cheaper)
    - To separate users who are influencers from users that simply have high follower counts for other reasons

Dataset

- Collected tweets containing 3 different hashtags: #fashion, #fitness and #travel from Twitter API using Tweepy Library
- All the three datasets contain between 135,000 to 95, 000 tweets.
- Mostly used features: Tweets, Followers Count
- Binary Classification: tweets with >30,000 followers (‘1’) and tweets with <= 5000 followers(‘0’)

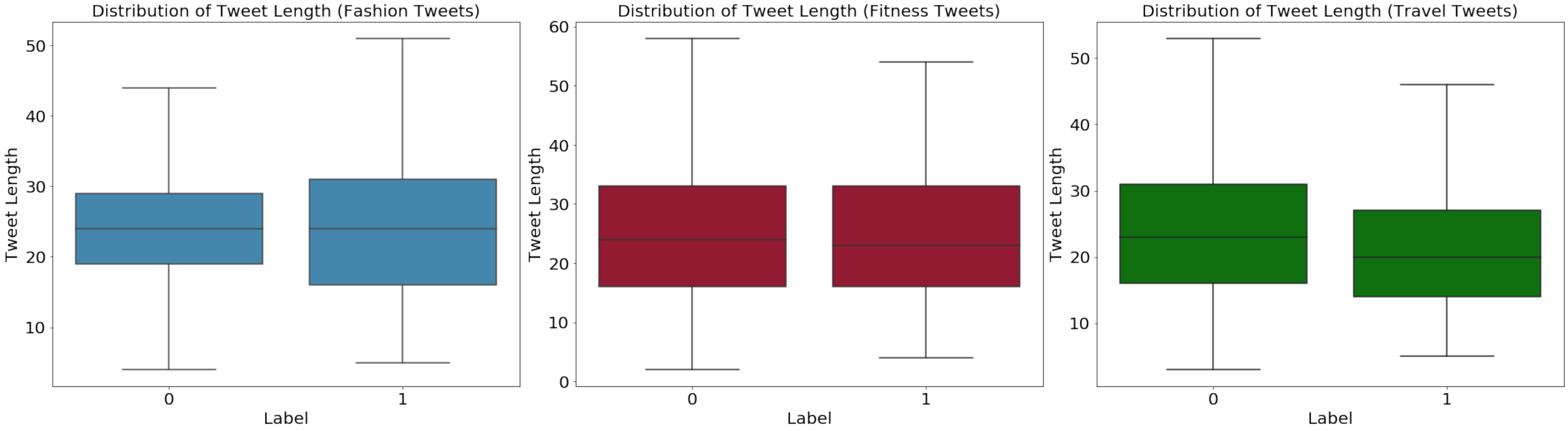
Data Wrangling

- **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)**

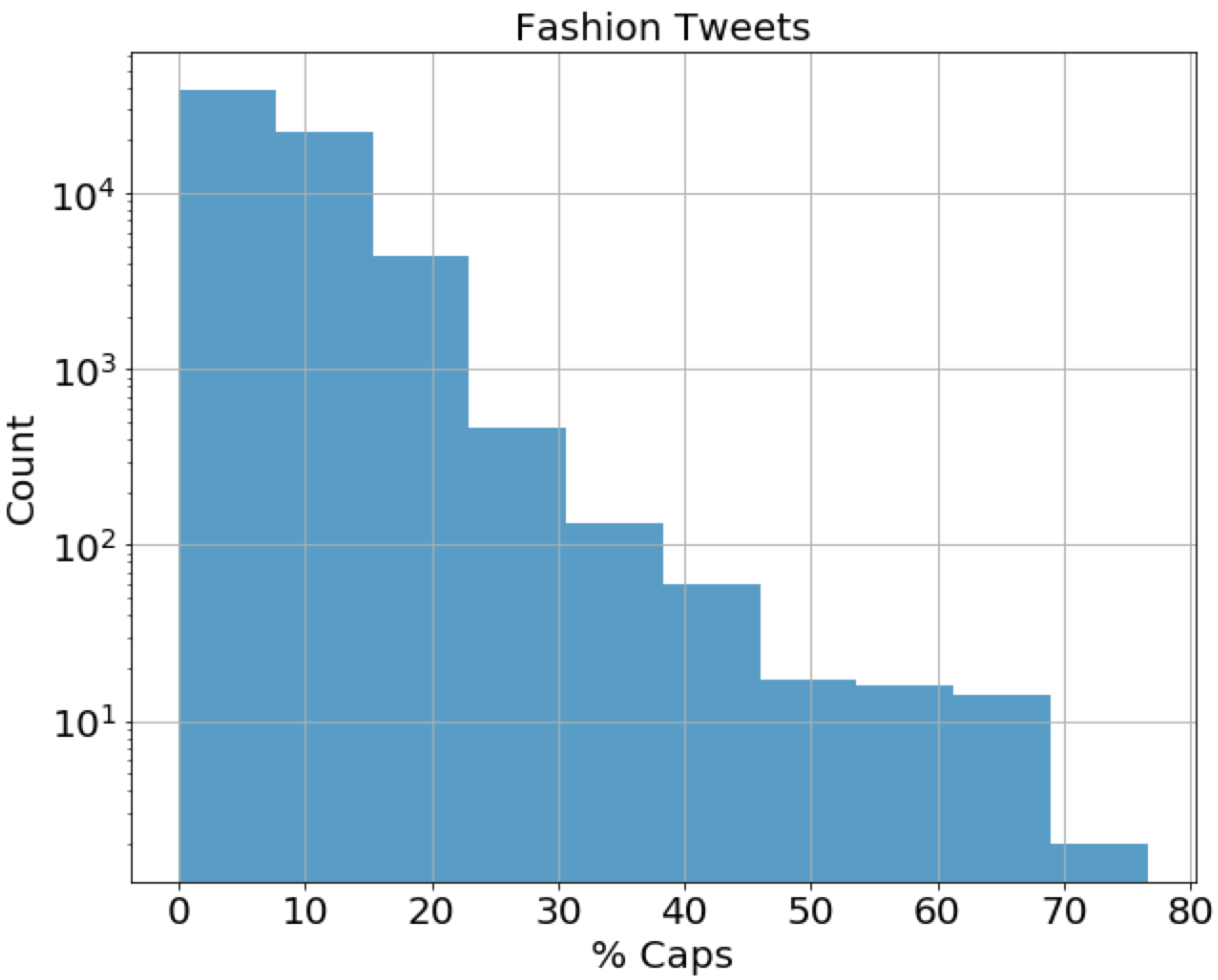
- **Fashion and Fitness Tweets: influencer and non-influencer tweets are about the same length (25 words)**

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.**

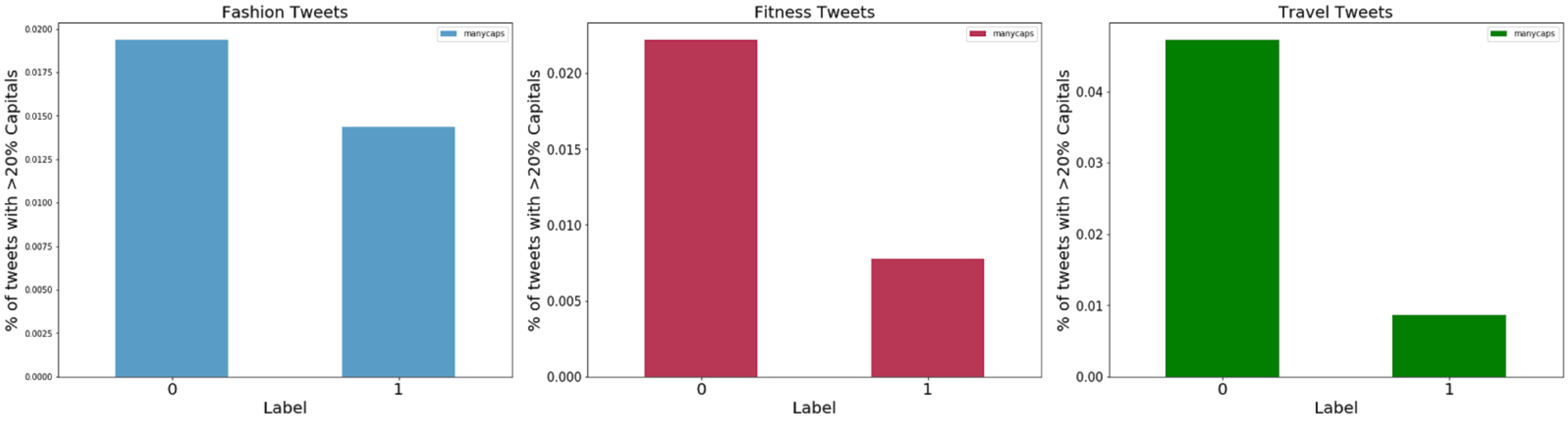
Exploratory Data Analysis



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**

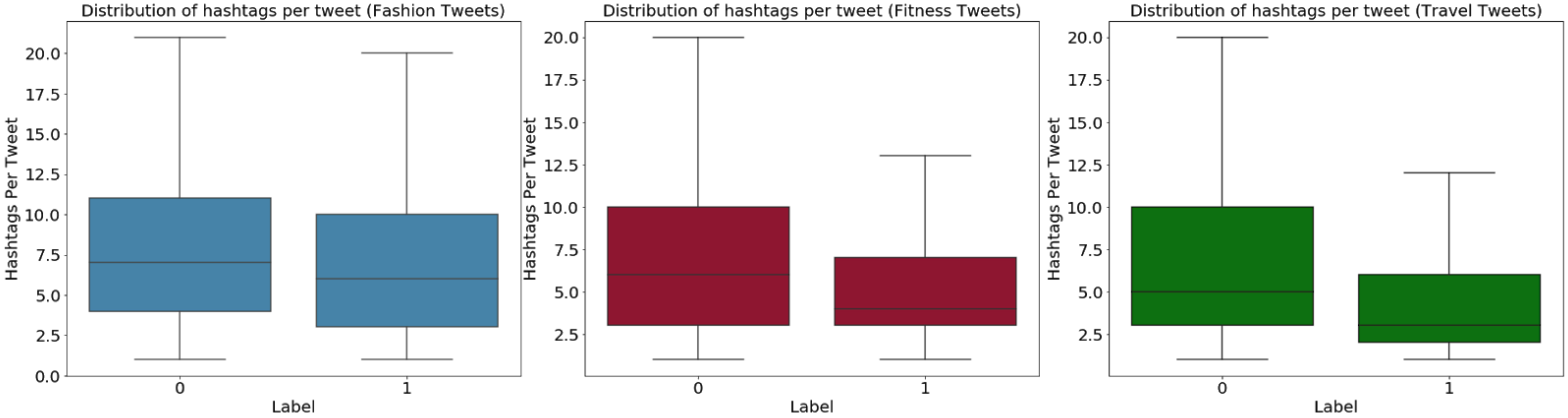
Exploratory Data Analysis



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**

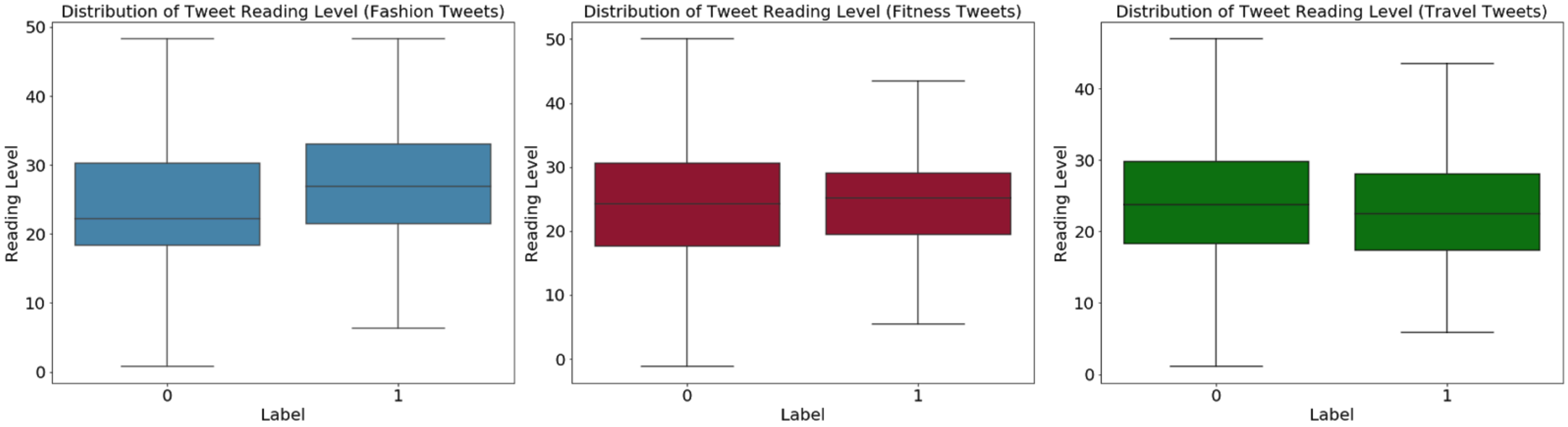
Exploratory Data Analysis



Distribution of number of hashtags per tweets by labels

The influencer tweets contain less number of hashtags.

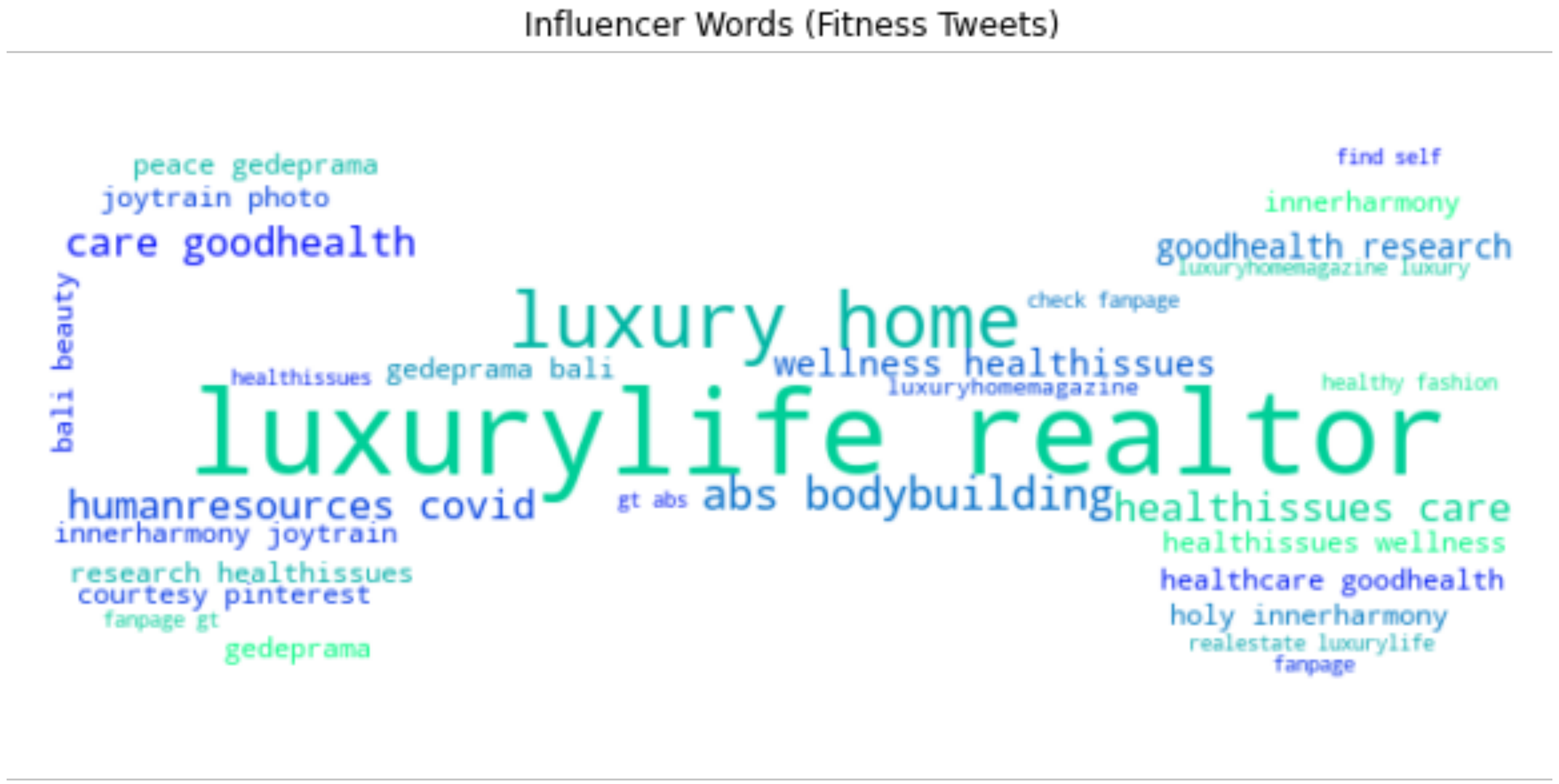
Exploratory Data Analysis

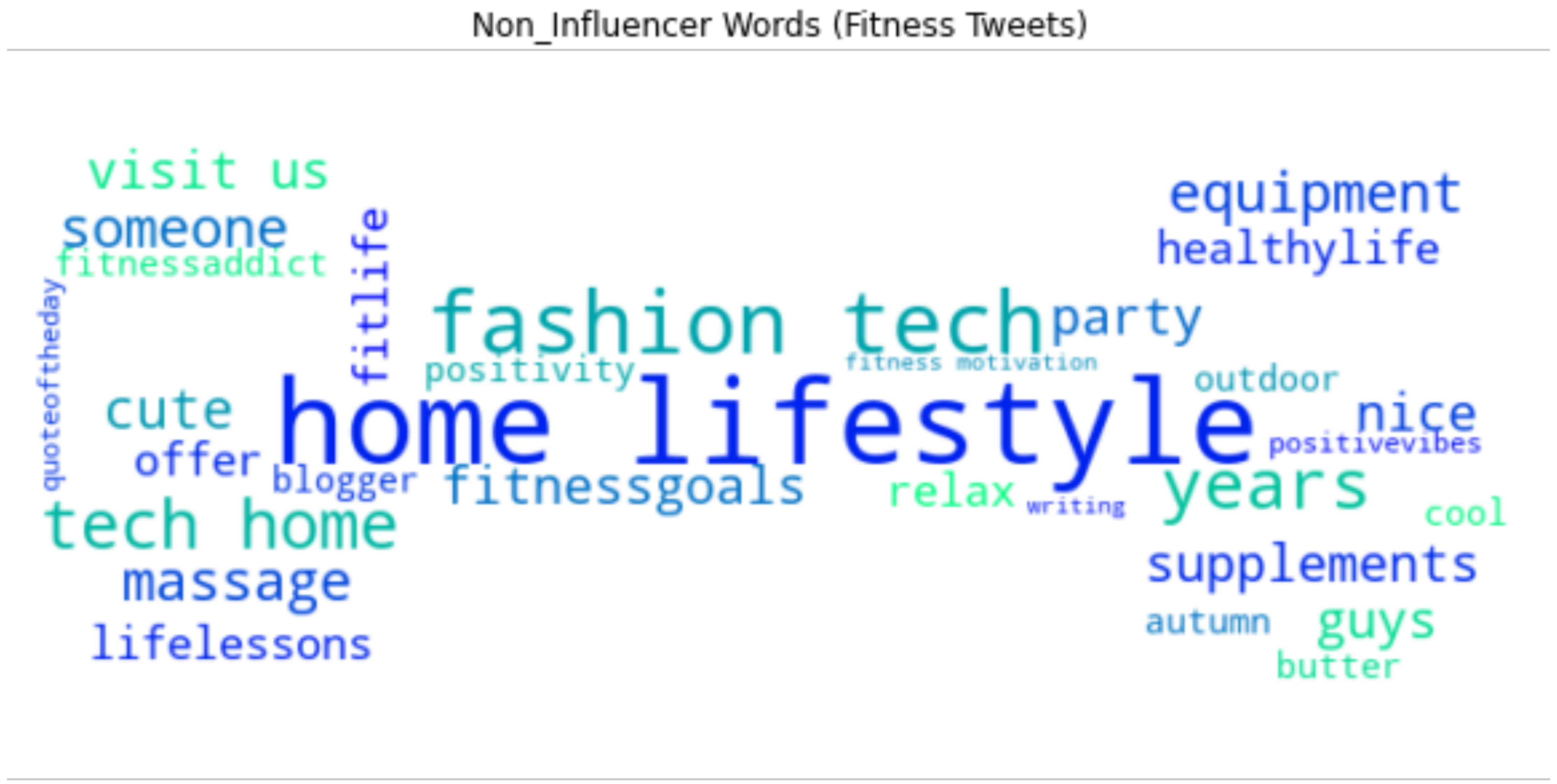


Distribution of Tweet reading level

- **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





- **Influencer tweets: include the words like ‘fanpage’, ‘luxuryhomemagazine’, ‘gedeprama’, “‘humanresources covid’” to promote the name of the people or website or magazine in their tweets.**

- **Non-influencer tweets: include the common words like ‘positivevibes’, ‘fitness motivation’, ‘fitnessgoals’ rather than promoting any health and wellness brands**

- **influencers have lots of promotions, mentioning medical or health brands on their tweets whereas non-influencer tweets are more focused on lifestyle and motivation.**

Most Predictive words for different tweet categories

Influencer Words (Fashion)

Non_Influencer Words (Fashion)

- 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.

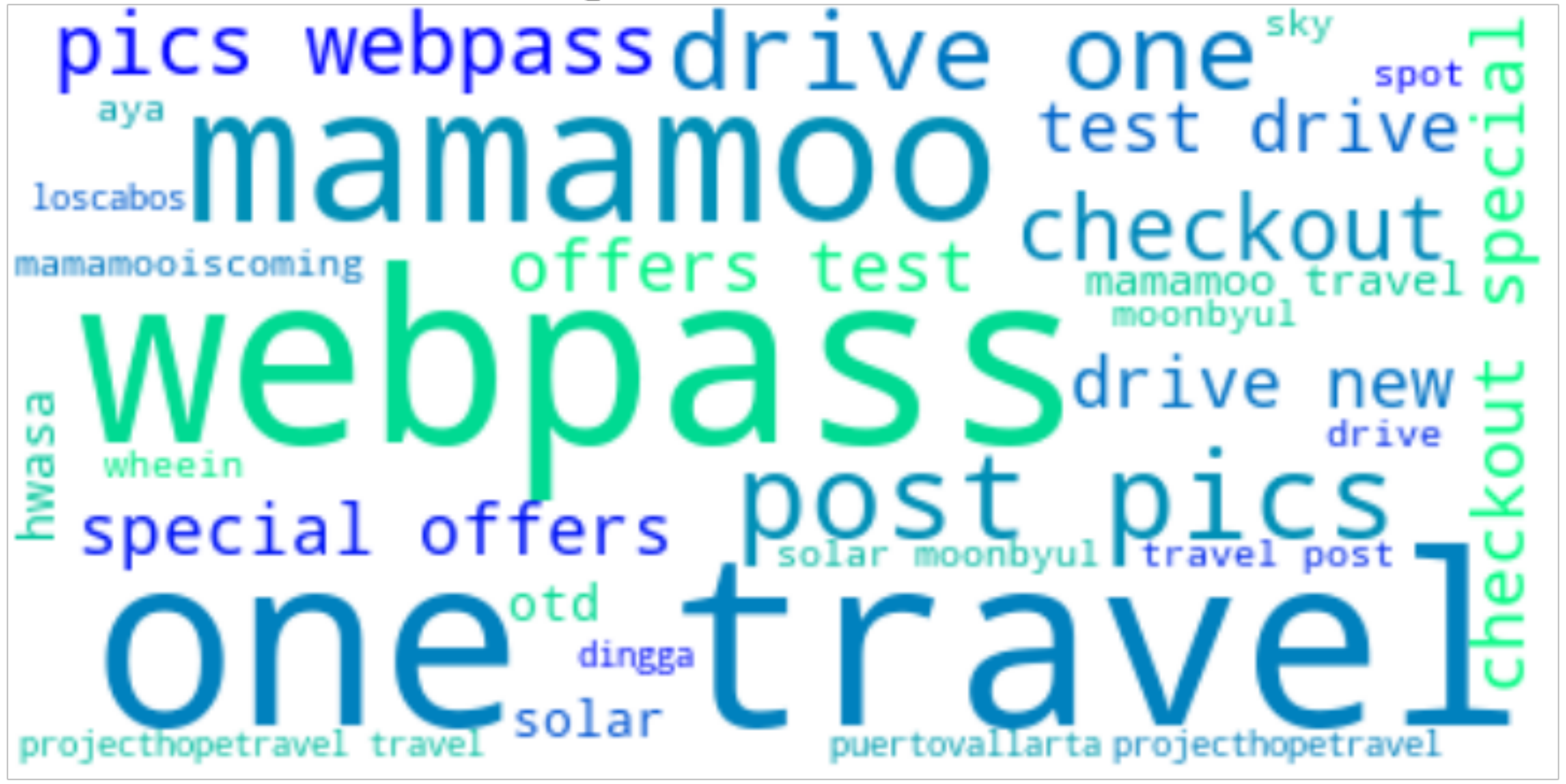
- The non-influencers also tweet about varieties of topics but mostly about different brands like: ‘calvinklein’, ‘gap’, ‘levis’, ‘disney’, ‘oldnavy’ etc.

- In the case of fashion tweets influencers talk about fashion ideas whereas non-influencers talk about brands. Hence people follow the fashion influencer for fashion ideas.

Most Predictive words for different tweet categories

Influencer (Travel Tweets)

Non-influencer (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.
- 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 and also some kind of special offers.

Machine Learning Highlights

Preprocessing → Vectorization → Model Tuning

Preprocessing

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

Vectorization

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

Model Tuning

- 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

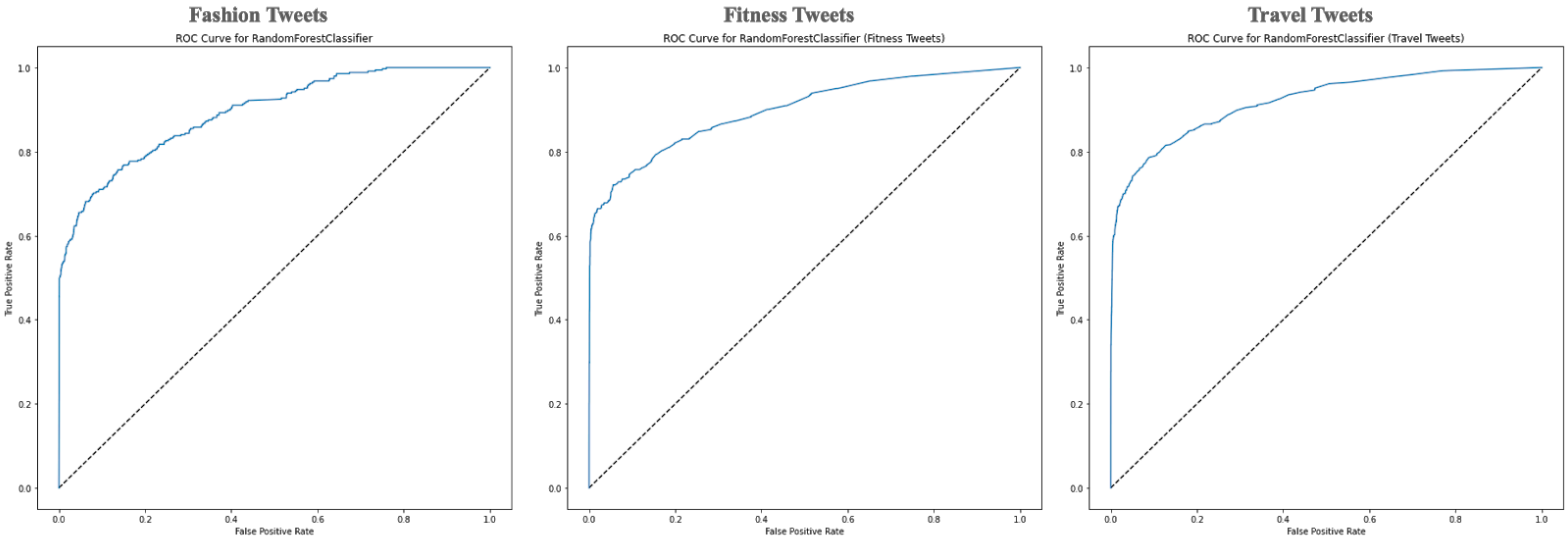
Comparison of three machine learning models fitted with TfidfVectorizer for three datasets

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

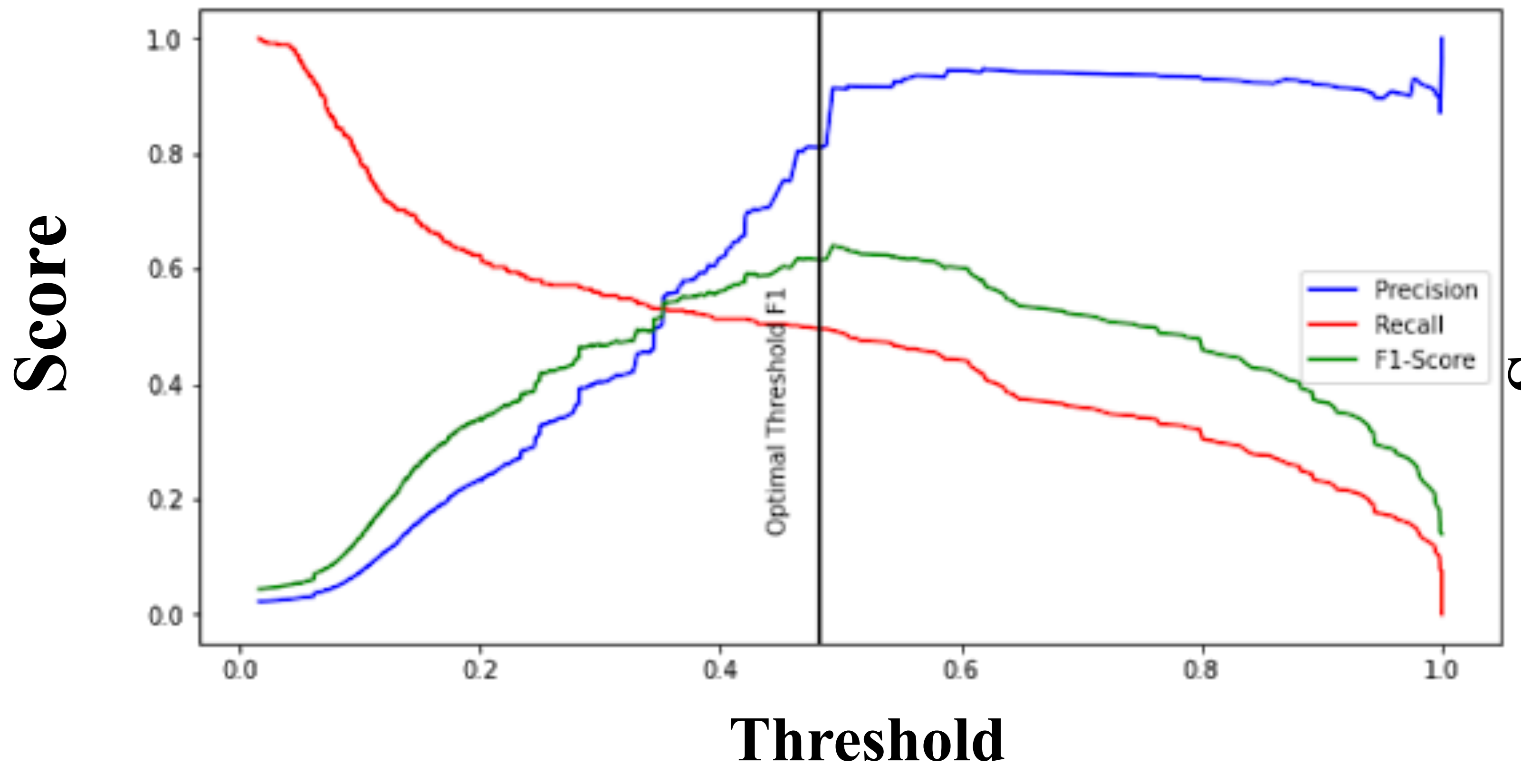
for all the 3 tweet categories

Best Classifier: Random Forest
ROC Curve

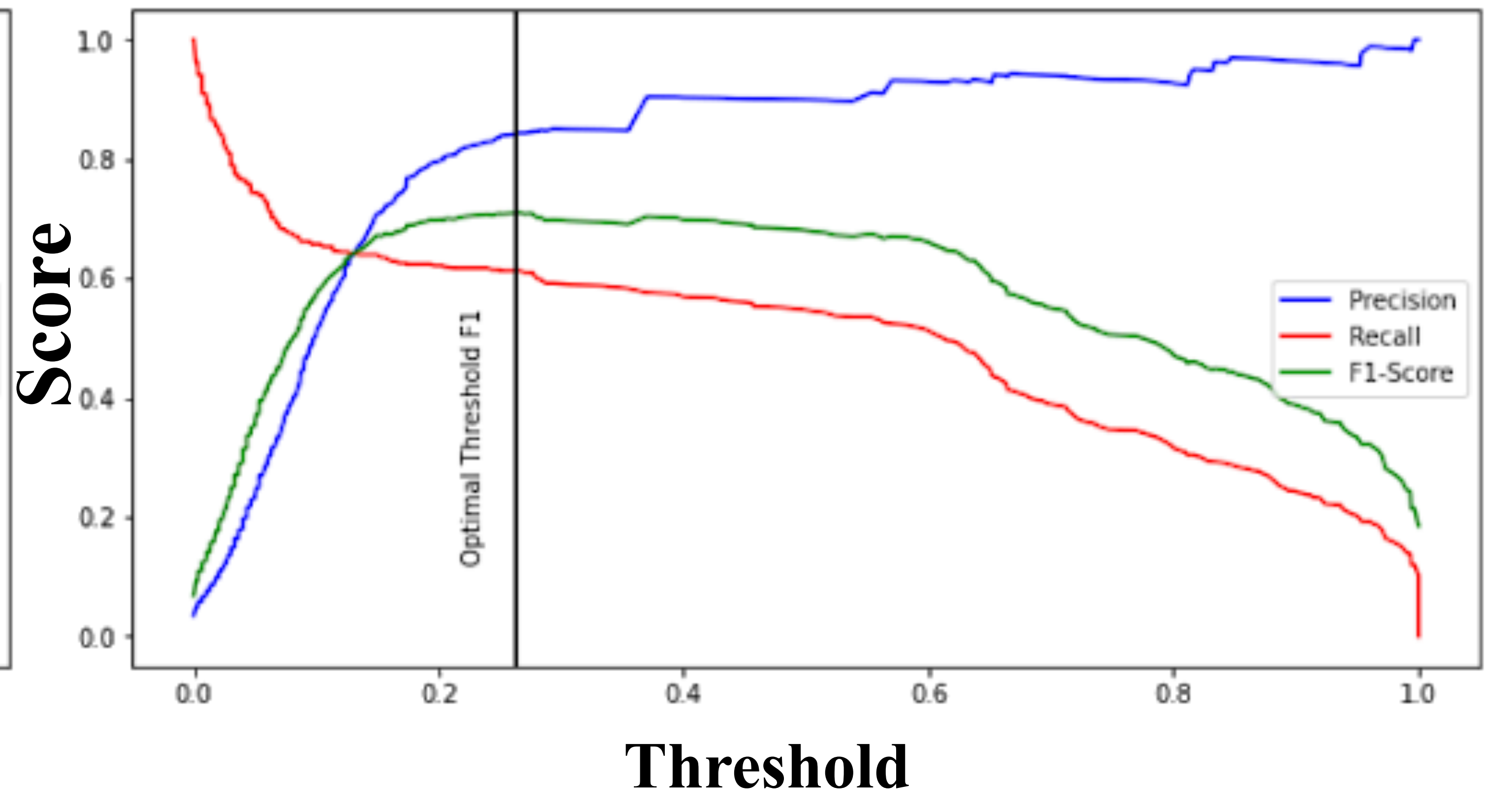


Improve Classification: Thresholding

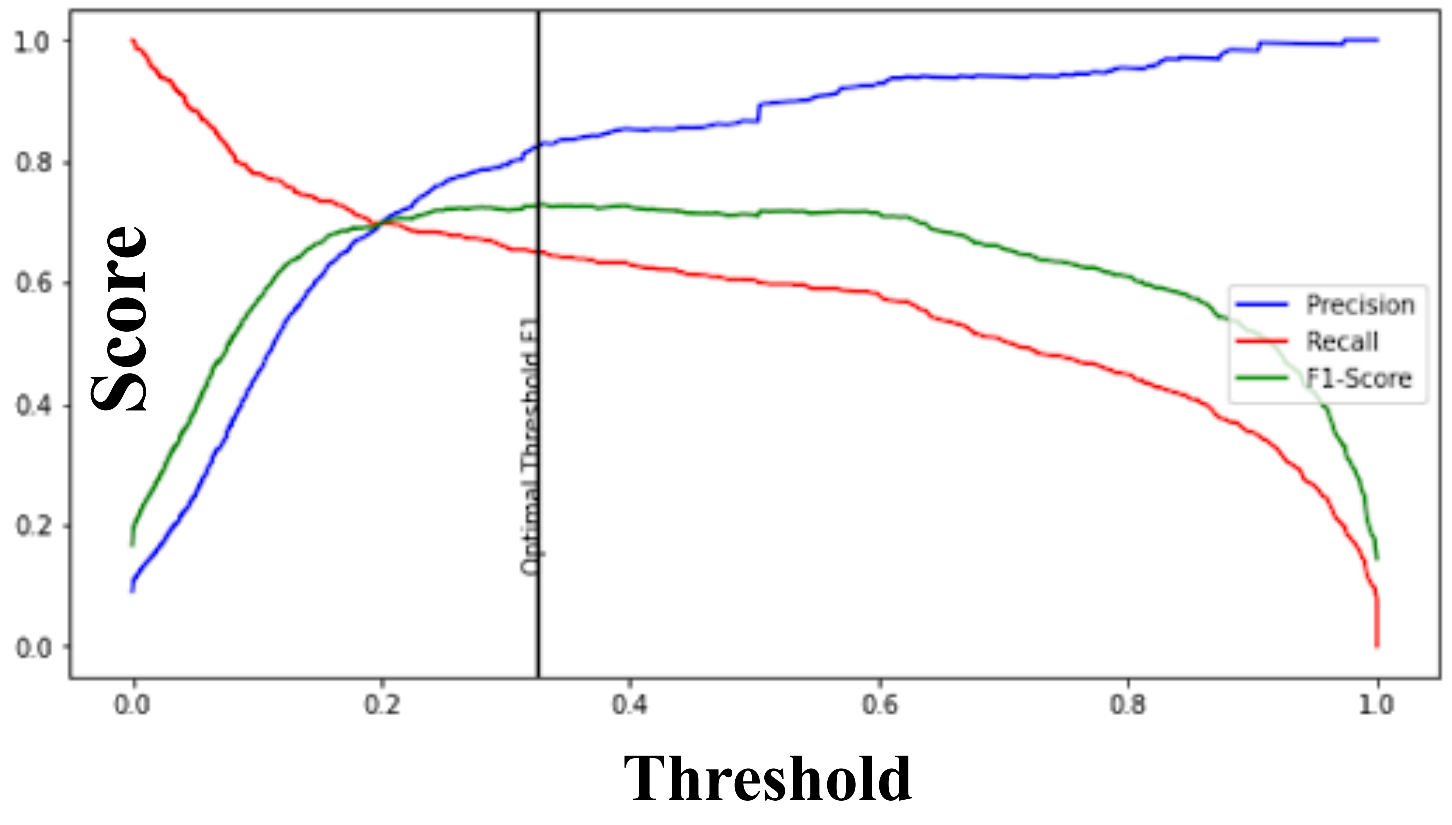
Threshold curves (Fashion Tweets)



Threshold curves (Fitness Tweets)



Threshold curves (Travel Tweets)



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