Data analysis and mining for Twitter using R

1 Abstract

The purpose of this project is to export and exploit useful information from Twitter. In project was done a data exploration in some fields and performed text mining techniques such as text classification and text clustering implemented on common algorithms such as Naive Bayes, SVM, K-means, LDA, aiming on high-quality information from the text data using the R programming language. An experimental evaluation of the algorithms will be performed in the recall, the precision, and in more evaluation metrics. The final results indicate the Naive Bayes as the best text supervised classifier and LDA as the best unsupervised classifier.

2 Introduction

Below we will present the data that we will process. We will show some interesting insights from the Twitter data using particular attributes and we will attempt to apply machine learning algorithms to the text attribute.

3 Data presentation

The tweets of Twitter were used as a data source and were collected from time December 30, 2016 to Feb 04, 2017 in JSON format. The subject of interest of the project is five different music genres: pop, rock, rap, classical and folk. The first dataset was generated consists of 16800 tweets (observations) (81.1MB) and 9 attributes that include both numerical and categorical attributes. In particular, the attributes utilized are: created_at, lang, time_zone, followers_count, friends_count, favourites_count, statuses_count, retweet_count, favorite_count. The second dataset consists only with the text of tweets and was used for applying text classification More specifically, out of 16800 thousand tweets that were collected, only the unique tweets were used, while there were 6217 deduplicates, accounting for 37% of the total number of tweets and they were removed since they do not help for the analysis. Thus, the dataset includes the unique tweets, which are automatically tagged with

“music_genres” label. The dataset therefore has two columns: the Text, which contains the 10583 tweets and the Music_genres ("classical_music", "folk_music", "pop_music", "rap_music", "rock_music"), among which 2474 tweets were automatically labeled based on keywords retrieved as "classical_music", 1500 tweets were automatically labeled as "folk_music", 2441 tweets were automatically labeled as "pop_music", 2319 tweets were automatically labeled as "rap_music" and 1849 tweets were automatically labeled as "rock_music". A third dataset was generated and was used for applied text clustering. The file contains only one column with the text tweets, that is, the 10583 tweets.

4 Data Analysis and Mining

4a. Data Analysis

The chapter performs a data exploration which involves visualizing and summarizing the data considering that it is vitally important to obtain further information on the dataset.

- **Understanding and summarizing of lang attribute.**

Table depicts the top 10 languages, ordered by decreasing number of tweets. The top 10 languages accounted for 97% of all the tweets. In Table , it is apparent that the most commonly occurring languages are English, Spanish and British English, with the English language constituting 82.8% of the total. The large percentage is confirmed by the fact that we had a limitation of downloading tweets only in the English language. In the dataset, 38 languages were found in the 16800 tweets. In addition, Table shows the distribution of users by the interface languages they choose. The anticipated finding is that the English language dominates with 10740 users constituting 84% of the total. Moreover, the last column shows the number of tweets per the number of users of each language separately.

Table 5.1: Distribution of the top 10 languages in Twitter.

Lang	Tweets	%	Users	Tweets/user
***en*-English**	**13923**	**82.875**	**10740**	**1**
***es*-Spanish**	**793**	**4.720**	**527**	**1**
***en-GB*-British English (United Kingdom)**	**338**	**2.011**	**280**	**1**
***it*-Italian**	**294**	**1.75**	**112**	**3**
***fr*-French**	**292**	**1.738**	**217**	**1**
***pt*-Portuguese**	**233**	**1.386**	**220**	**1**
***de*-German**	**168**	**1**	**117**	**1**
***ru*-Russian**	**153**	**0.910**	**99**	**1**
***ja*-Japanese**	**136**	**0.809**	**103**	**1**
***nl*-Dutch**	**82**	**0.488**	**51**	**2**

- **Understanding and summarizing of time zone attribute.**

With regard to Table , what applies to all tweets is that most of them were declared without a time zone, that is, 6272 tweets are empty, less than 38% of the total. The fields without a time zone probably due to both the lack of explicit user consent and the unavailability of localization services at the moment in which tweets are published. The remaining 62% include time-zone information, where 25% include a city name, such as London, whilst 37% include the name of time zone, such as Pacific Time. In the dataset, 140 time zones were found in the 16800 tweets. Table 5.2 also shows the distribution of users by the time zone they choose. Most users do not choose a time zone, that is, 5195 users that constitute 48% of the total. The cities with the most users are: London that constitutes 6% of the total, Quito constitutes 1.9%, Amsterdam constitutes 1.8%, Rome constitutes 0.3% and Casablanca constitutes 0.69%.

Table : Distribution of the 10 top time zones in Twitter.

Time zone	Tweets	%	Users
None	**6272**	**37.366**	**5195**
Pacific Time(US & Canada)	**3286**	**19.577**	**2221**
Eastern Time(US & Canada)	**1614**	**9.615**	**1157**
Central Time(Us & Canada)	**906**	**5.397**	**673**
London	**771**	**4.593**	**652**
Quito	**345**	**2.055**	**207**
Atlantic Time(Canada)	**261**	**1.554**	**249**
Amsterdam	**232**	**1.382**	**201**
Rome	**184**	**1.096**	**36**
Casablanca	**177**	**1.054**	**75**

- **Understanding, summarizing and visualizing of continuous attributes.**

Table depicts the descriptive statistical measures of the attributes regarding the information on Twitter users of our dataset. Examining the users' information, it appears that the median number of the users' tweets is 6670 and the mode is 4, which means that 6670 users posted a tweet four times. In addition, the median count of the users' followers is 531 with the mode equal to 0, which means that 531 users are not followed by other users, while the median count of the users' favourites is 1445 and the mode is 0, which means that 1445 users that did not do a "like" in a tweet. From Table also seems that the median count of the users' friends is 484 and the mode equal to 0, which means that 484 users do not follow anyone, while the median count of the users' retweets is 83 with the mode equal to 1, meaning that 83 users retweeted a tweet. Furthermore, the median count of the users' favorites is 90 and the mode is 0, which means that in 90 users did not like tweets.

Table : Descriptive statistical measures of the attributes.

Measures of Location					
	Min.	**Median**	**Mean**	**Max.**	**Mode**
Followers_count	0	531	11940.13	28715815	0
Friends_count	0	484	4090.18	5430083	0
Favourites_count	0	1445	10428.19	3555686	0
Statuses_count	1	6670	78040.9	164504281	4
Retweet_count	1	83	1951.56	76915	1
Favorite_count	0	90	2834.23	139460	0

In addition, all the attributes are tabulated in Table are skewed to the right, which means that there are few users with very high numbers of followers, friends, etc,

whilst most have few numbers of followers, friends, etc in the dataset. The plots below also confirm this ascertainment. In particular, the distributions are characterized as right-skewed distributions and the box plots as a positive skew box plot, where as a result of these characteristics the plots have their mean greater than the median. Those box plots do not include extreme values.

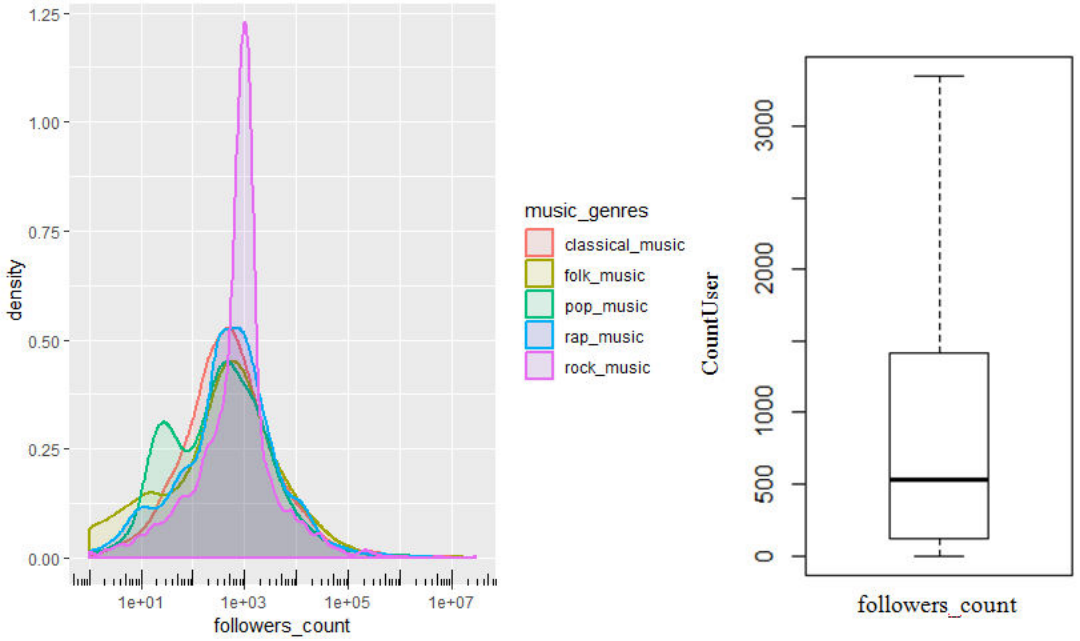

Figure : Distributions and box plot of followers_count attribute.

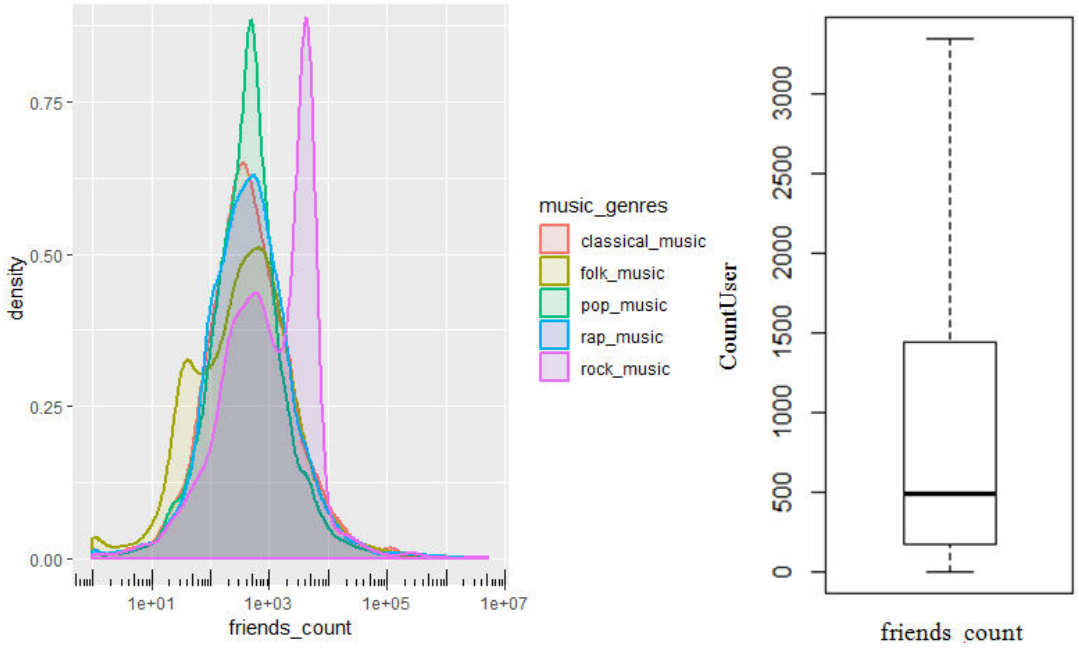

Figure : Distributions and box plot of friends_count attribute.

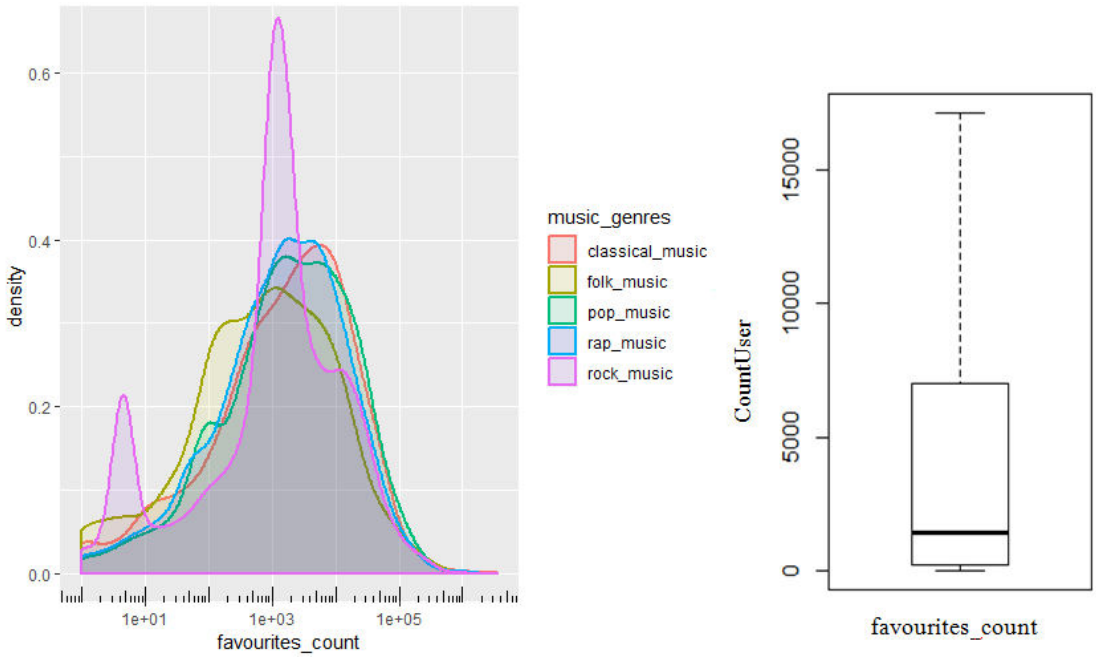

Figure : Distributions and box plot of favourites_count attribute.

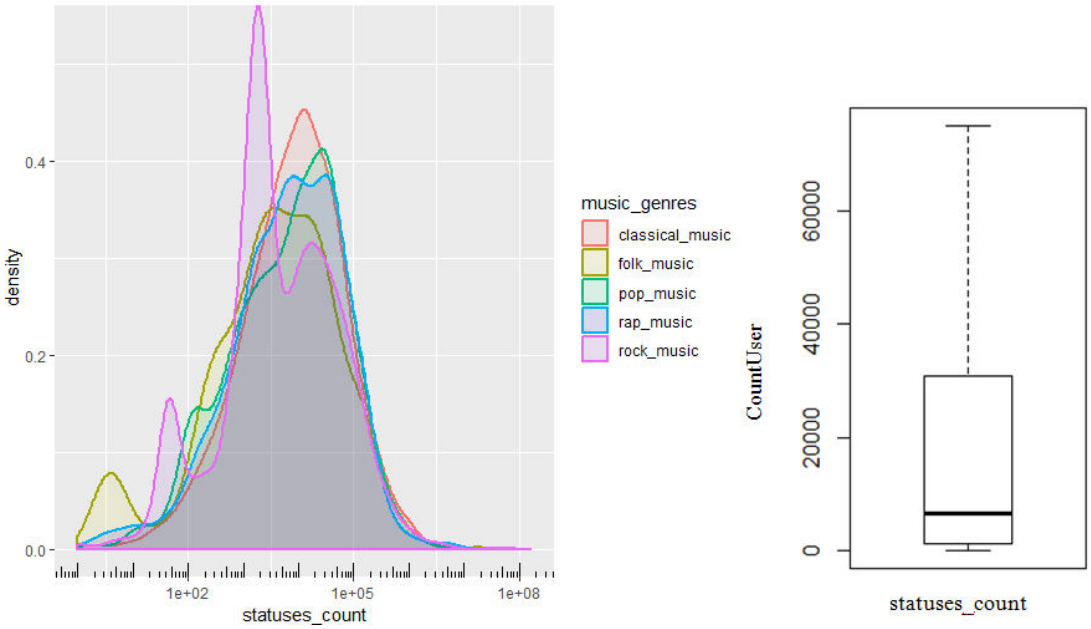

Figure 5.1d: Distributions and box plot of statuses_count attrribute.

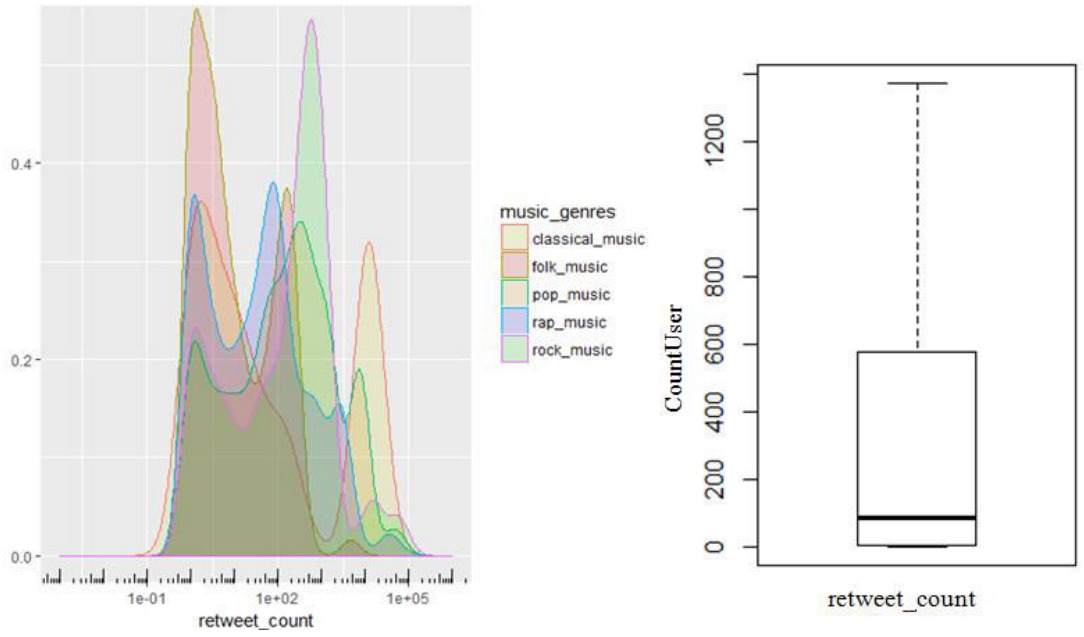

Figure : Distributions and box plot of retweet_count attribute.

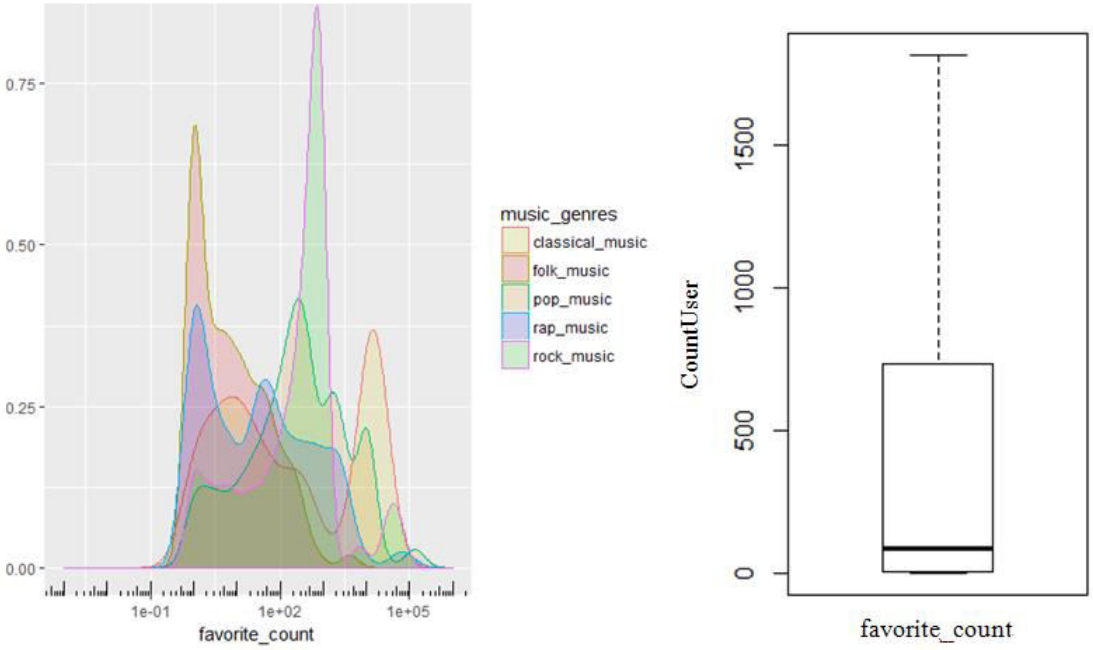

Figure : Distributions and box plot of favorite_count attribute.

Interesting Twitter Statistics

Number of followers, friends and statuses of the users on the dataset

An expected question that can be raised is, how many followers do users have? Every tweet has some impact for the user, and hence the number of followers is of great importance to every user with regard to the direct, faster and greater transmission of information for each tweet. In the current dataset, there are 12781 users (unique users) with an average of 6.10 tweets per user, where 2.75%, namely 351 users are no followed by other users as they have 0 followers. 48.82% of Twitter users have less than 500 followers, 16.76% have less than 1000, 24.24% have less than 5000,4.04% have less than 10000, 2.67% have less than 20000 and 3.45% have more than 20000. Figure 5.2 illustrates the distribution of Twitter users' followers. In addition, Table 5.4 reports the top 5 users with the most followers.

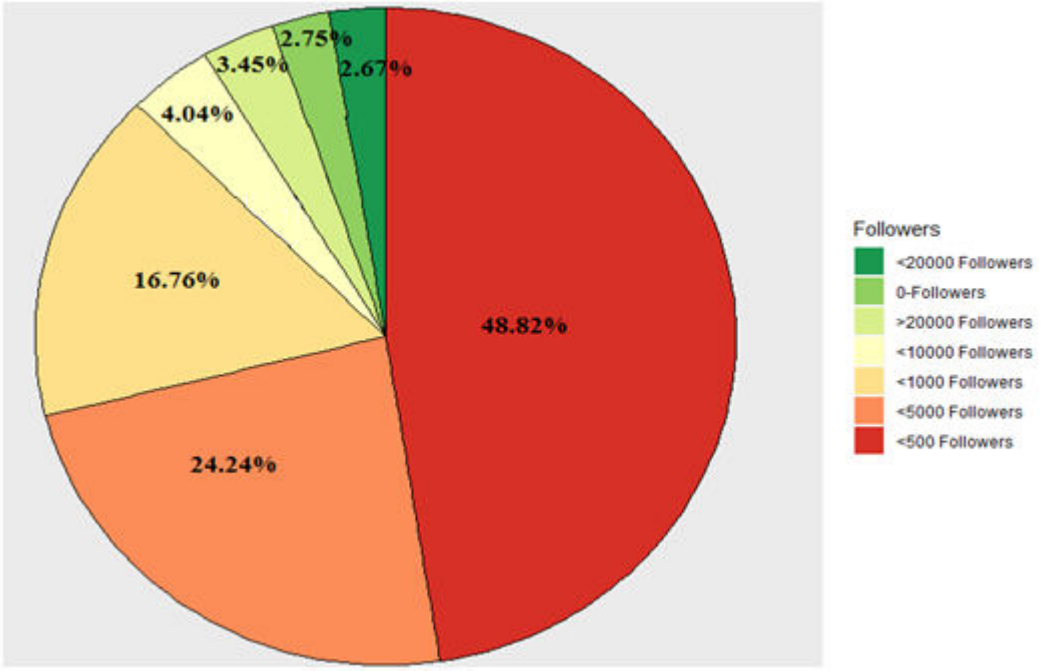

Figure 5.2: Distribution of users by follower count.

Table 5.4: Users with most followers.

Nickname Of user	No.of Followers
User1	28715815
User2	12586176
User3	8447507
User4	5446443
User5	5102585

In the case of number of friends that users have, 1.48%, that is 190 users have 0 friends. 14.86% of Twitter users have less than 500 friends, 17.44% have less than 1000, 25.84% have less than 5000, 2.80% have less than 10000,1.34% have less than 20000 and 1.54% have more than 20000. Figure 5.3 illustrates the distribution of Twitter users' friends. In addition, Table 5.5 reports the top 5 users with the most friends.

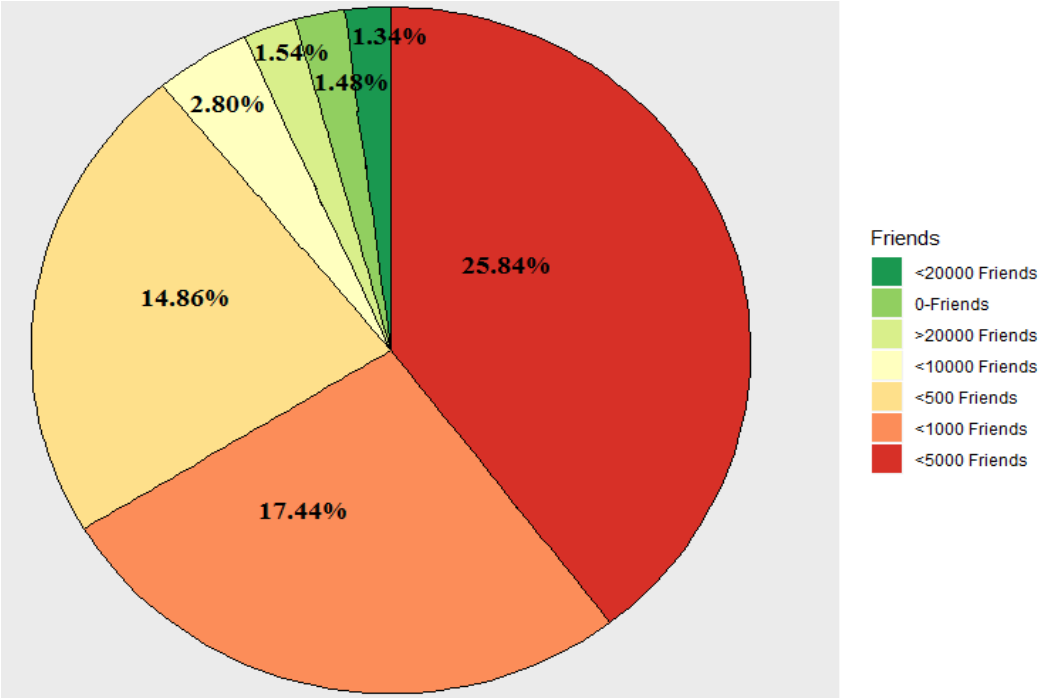

Figure 5.3: Distribution of users by friend count.

Table 5.5: Users with most friends.

Nickname Of user	No.of Friends
User1	5430083
User2	4601143
User3	2504270
User4	1442901
User5	1422981

As regards Figure 5.4, this depicts the distribution of Twitter users by the number of tweets they have sent. The numbers that are shown are quite astounding as well about 100% of all Twitter users have tweeted at least once. 15.10% of Twitter users have made less than 500 tweets, 6.63%% have made less than 1000 tweets, 24.02% have made less than 5000 tweets, 10.57% have made less than 10000 tweets, 11.62% have made less than 20000 tweets and 32.03% have made over than 20000 tweets.

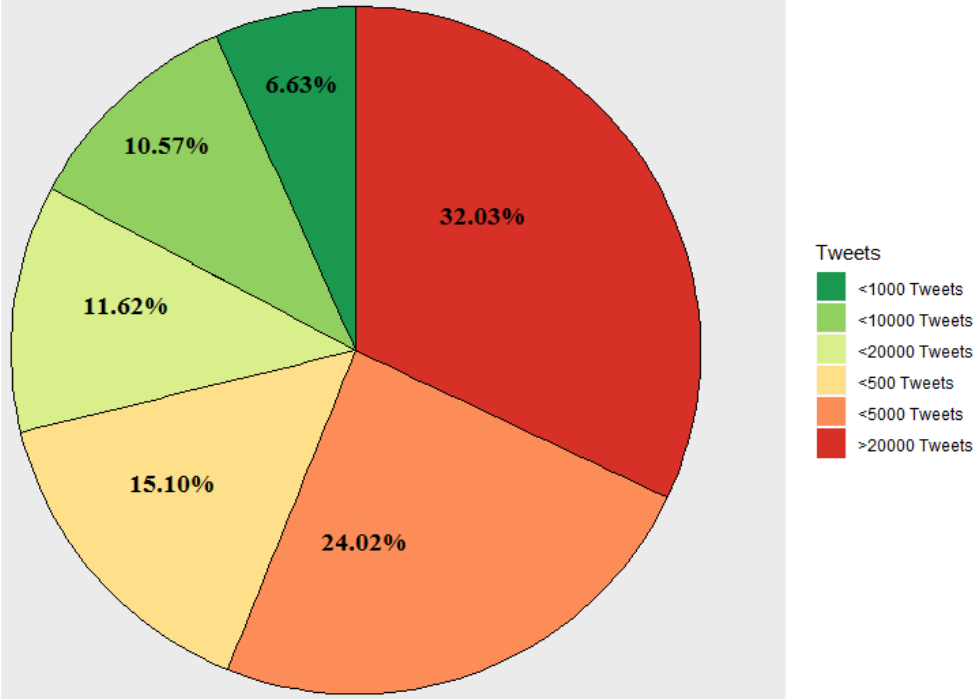

Figure 5.4: Distribution of users by statuses count.

It is apparent that in the vast majority of users, 48.82%, that is 6239 users, have less than 500 followers, 25.84%, that is 3302 users, have less than 5000 friends and 32.03%, that is 4094 users, have posted over 20000 tweets. Thus, it is concluded that a high number of users have few followers and friends, in contrast to the tweets they post. Furthermore, according to the data, the average Twitter user has 1,070 followers 3,124 friends and 6,106 tweets.

4b. Text Mining: Text classification and Text Clustering

The specificity of the algorithms is that they work in many classes, namely, in five. Thus, it's a way to see how they correspond to many classes.

Text Classification

- **Results of Naïve Bayes classifier**

Table reports the evaluation results using k- cross-validation method with k=10. The execution time of the algorithm was 77 seconds. The model has achieved an accuracy of 88.1%, which is quite high accuracy. The high values appear in all evaluation metrics, which proves that the Naive Bayes classifier is effective for text classification.

Table : Evaluation Results of Naïve Bayes classifier.

CA	Precision	Recall	F1	Kappa	Specificity	AUC
0.881	0.879	0.881	0.878	0.851	0.894	0.887

Table presents the results of recall and precision measures of each music genre for the Naïve Bayes classifier. As it is apparent, all the values of both measures are high, but in classical music it's higher. High recall and precision values indicate low false negatives and false negatives. Specifically, this means that in the class of classical music few items are incorrectly labeled and few items that do not belong to this class but should be.

Table : Recall and Precision of each music genre of Naïve Bayes classifier.

Naïve Bayes

	Recall	Precision
classical_music	**0.930**	**0.914**
folk_music	0.901	0.873
pop_music	0.829	0.861
rap_music	0.903	0.906
rock_music	0.843	0.840

- **Results of Support Vector Machine classifier**

Table reports the evaluation results using the k-fold cross-validation method with k=10 and as a kernel parameter the radial. The execution time of the algorithm was 103 seconds. Unlike to Naïve Bayes classifier, the SVM classifier has achieved an accuracy of 85.2%, which is also quite high accuracy. The high values appear in all

evaluation metrics, which proves that the SVM classifier is also suitable for text classification.

Table : Evaluation Results of SVM classifier.

CA	Precision	Recall	F1	Kappa	Specificity	AUC
0.852	0.867	0.833	0.852	0.813	0.962	0.897

Table presents the results of recall and precision measures of each music genre for the SVM classifier. It seems that all the values of both measures are high, but the class of classical music has the highest recall and a lower precision while the class of folk music has the highest precision and a lower recall. Low recall and precision values indicate a large number of false negatives and false positives. In particular, the low value of the class of classical music in the precision measure indicates that many items are incorrectly labeled. The low value of the folk music class in the recall measure indicates that many items do not belong to this class but should be.

Table : Recall and Precision of each music genre of SVM classifier.

SVM

	Recall	Precision
classical_music	**0.917**	0.865
folk_music	0.80	**0.958**
pop_music	0.880	0.760
rap_music	0.863	0.894
rock_music	0.759	0.854

Text Clustering

- **Results of k-Means classifier**

K-means algorithm is going to run with five clusters, fifteen maximum iterations per run and fifty random starts. The k = 5 is selected as it corresponds to the 5 categories of our project. Furthermore, we also test k=6. The plots of the clustering result are achieved with the fviz_cluster() function and the distance metric used is Jaccard. The fviz_cluster() is obtained based on PCA(Principal Component Analysis), which reduces the number of dimensions to two in such a way that the reduced dimensions capture as much of the variability between the clusters as possible, thus enabling better visualization.

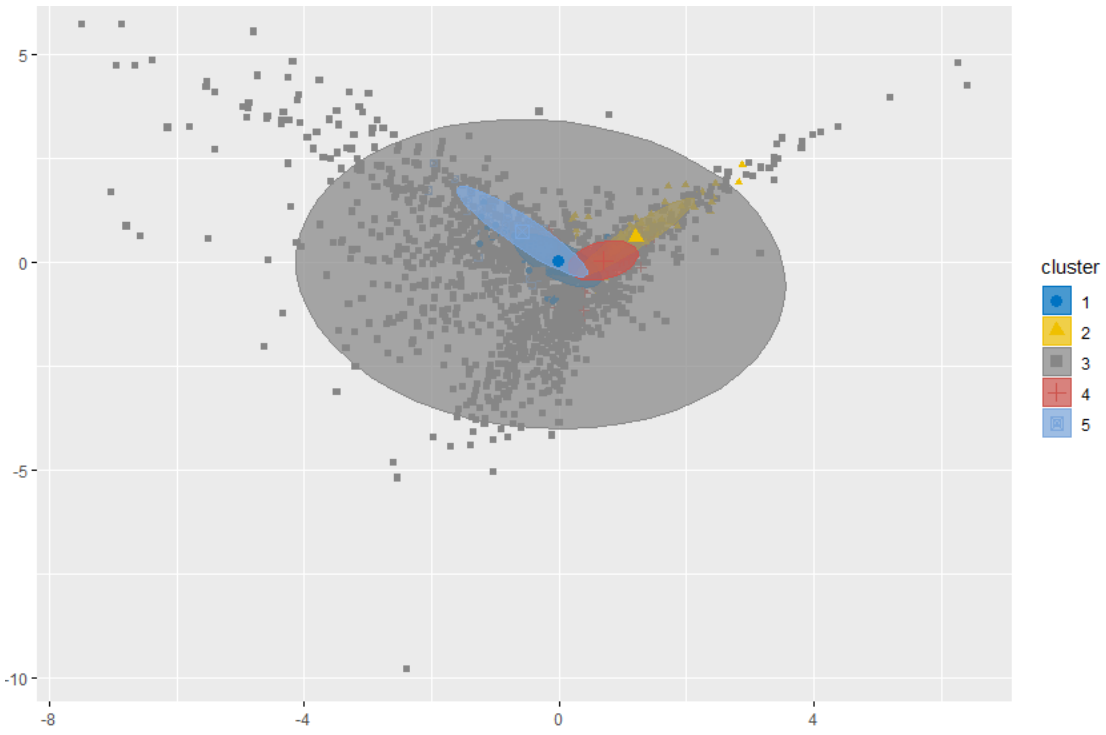

Figure : Cluster plot with k=5 and without the stemming technique.

The execution time of K-means without the stemming technique was 1707 seconds. As it appears from Figure 5.9, the algorithm groups the dataset into five clusters with each of five music genres. In particular, in cluster 1 there are 923 documents of which 37 belongs to classical music, 0 belongs to folk music, 50 belongs to rap music,630 belongs to rock music and 206 belongs to pop music, while in cluster 2 there are 1074

documents of which 834 belongs to classical music, 67 belongs to folk music,55 belongs to rap music, 52 belongs to rock music and 66 belongs to pop music. Additionally, in cluster 3 there are 6164 documents of which 1991 belongs to pop music, 1078 belongs to rock music, 2569 belongs to rap music, 556 belongs to folk music and 0 belongs to classical music. In cluster 4 there are 979 documents of which 101 belongs to classical music, 659 belongs to folk music, 50 belongs to rap music, 93 belongs to rock music and 76 belongs to pop music, while in cluster 5 there are 1399 documents of which 980 belongs to pop music, 128 belongs to rock music, 126 belongs to rap music, 81 belongs to classical music and 84 belongs to folk music.

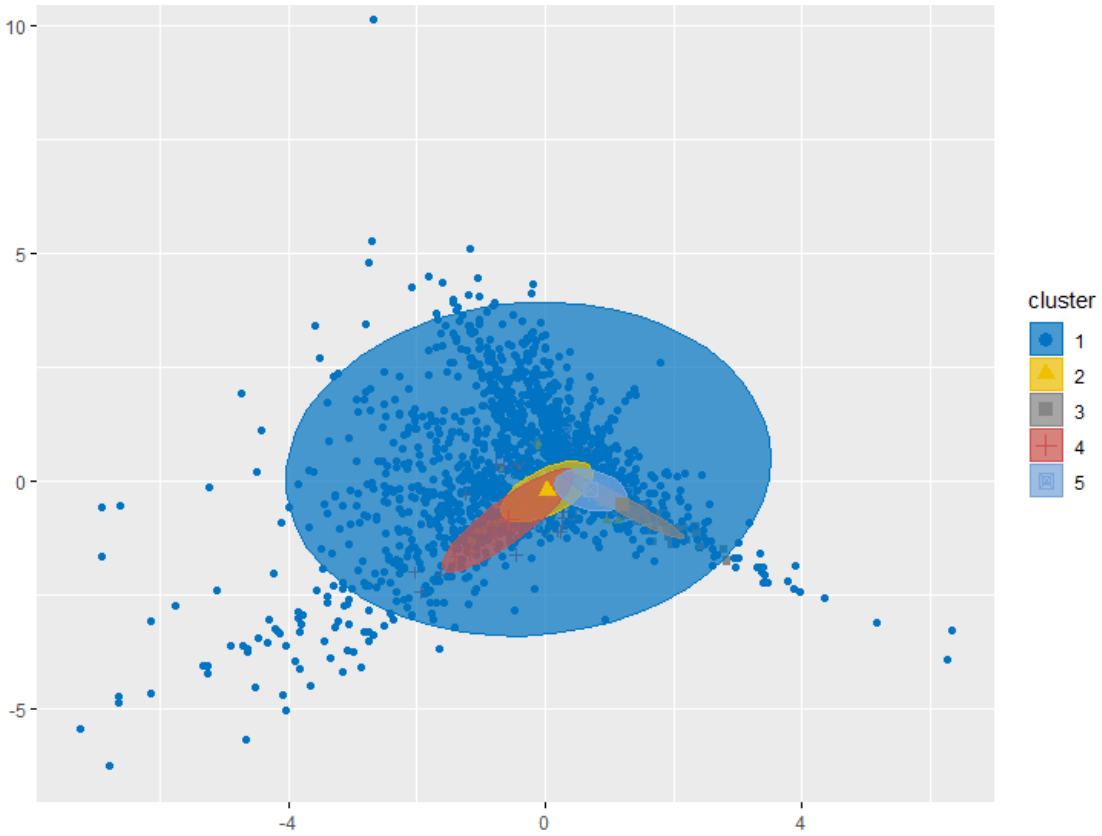

Figure 5.10: Cluster plot with k=5 and the stemming technique.

The execution time of K-means with the stemming technique was 1801 seconds. In Figure 5.10, the implementation of the algorithm returns the following results: in cluster 1 there are 6468 documents of which 1205 belongs to classical music, 622 belongs to folk music, 1972 belongs to rap music, 1120 belongs to rock music and 1449 belongs to pop music, while in cluster 2 there are 1007 documents of which 38

belongs to classical music, 36 belongs to folk music, 59 belongs to rap music, 675 belongs to rock music and 199 belongs to pop music. Furthermore, in cluster 3 there are 1299 documents of which 1055 belongs to classical music,65 belongs to folk music, 61 belongs to rap music, 56 belongs to rock music and 62 belongs to pop music. In cluster 4 there are 885 documents of which 0 belongs to classical music, 33 belongs to folk music, 77 belongs to rap music, 74 belongs to rock music and 701 belongs to pop music, while in cluster 5 there are 910 documents of which 93 belongs to classical music, 607 belongs to folk music, 48 belongs to rap music, 86 belongs to rock music and 76 belongs to pop music.

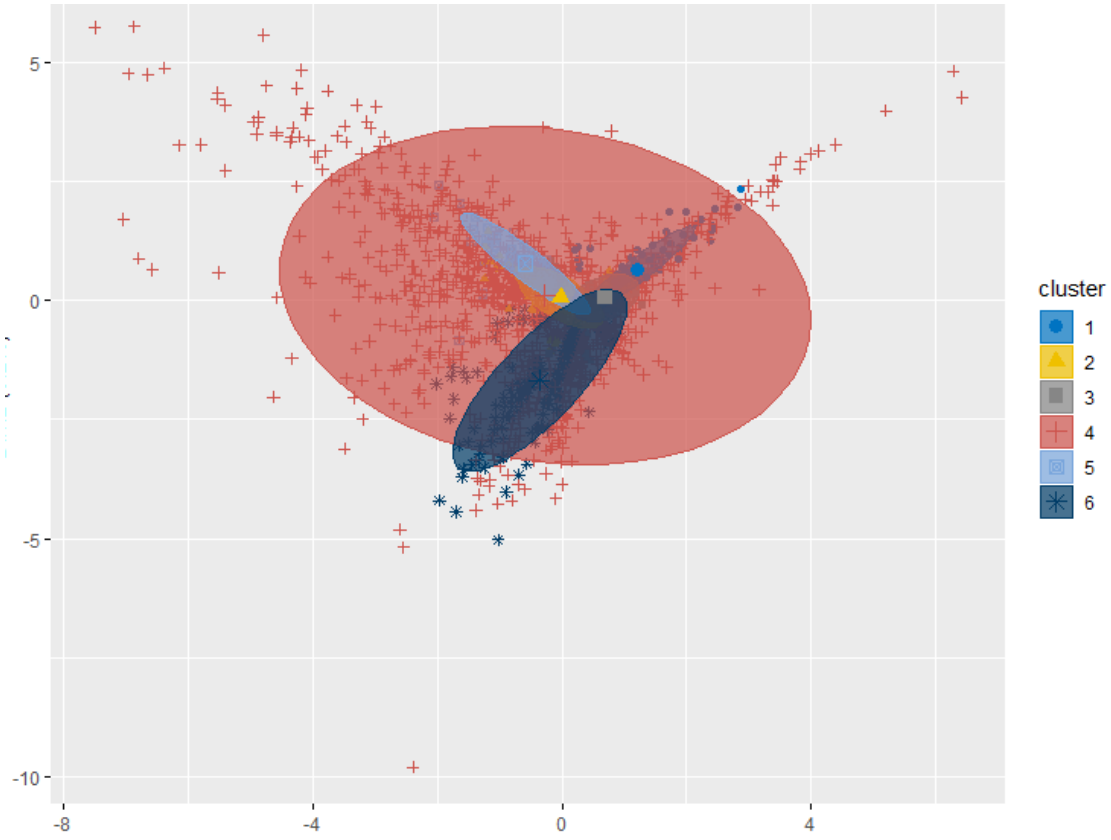

Figure : Cluster plot with k=6 and without the stemming technique.

The execution time of K-means without stemming was 1935 seconds. In Figure 5.11, the implementation of the algorithm returns the following results: in cluster 1 there are 1393 documents of which 1133 belongs to classical music, 69 belongs to folk music, 57 belongs to rap music, 59belongs to rock music and 75 belongs to pop

music, while in cluster 2 there are 1079 documents of which 41 belongs to classical music, 43 belongs to folk music, 63 belongs to rap music, 746 belongs to rock music and 186 belongs to pop music. In addition, in cluster 3 there are 980 documents of which 99 belongs to classical music, 664 belongs to folk music, 50 belongs to rap music, 91 belongs to rock music and 76 belongs to pop music, while in cluster 4 there are 4901 documents of which 1020 belongs to classical music, 525 belongs to folk music, 1056 belongs to rap music, 992 belongs to rock music and 1308 belongs to pop music. In cluster 5 there are 915 documents of which 31 belongs to classical music, 34 belongs to folk music, 75 belongs to rap music, 77 belongs to rock music and 698 belongs to pop music, while in cluster 6 there are 1301documents of which 42 belongs to classical music, 30 belongs to folk music, 1020 belongs to rap music, 70 belongs to rock music and 139 belongs to pop music.

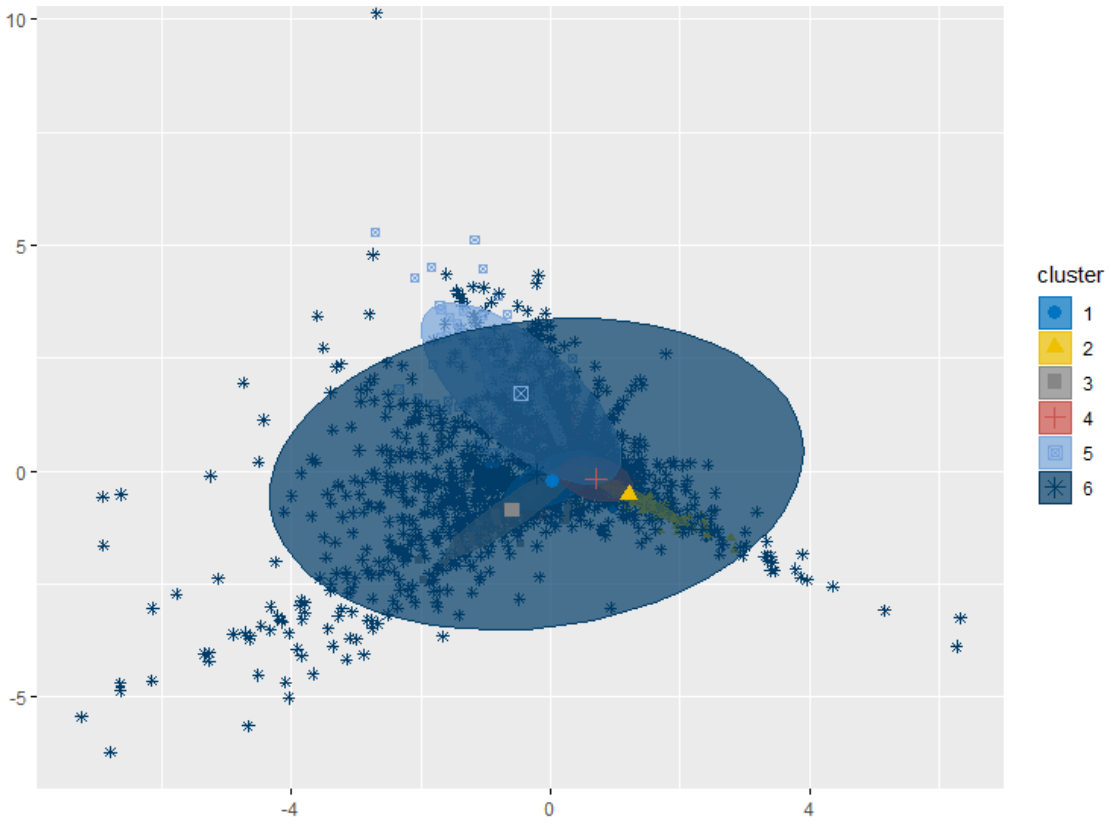

Figure : Cluster plot with k=6 and the stemming technique.

The execution time of K-means with the stemming technique was 1894 seconds.

In Figure 5.12, the implementation of the algorithm returns the following results: in cluster 1 there are 1008 documents of which 38 belongs to classical music, 36 belongs to folk music, 58 belongs to rap music, 661 belongs to rock music and 215 belongs to

pop music, while in cluster 2 there are 1292 documents of which 1058 belongs to classical music, 65 belongs to folk music, 56 belongs to rap music, 56 belongs to rock music and 62 belongs to pop music. Furthermore, in cluster 3 there are 863 documents of which 31 belongs to classical music, 31 belongs to folk music, 70 belongs to rap music, 71 belongs to rock music and 660 belongs to pop music, while in cluster 4 there are 910 documents of which 91 belongs to classical music, 609 belongs to folk music, 48 belongs to rap music, 86 belongs to rock music and 76 belongs to pop music. Additionally, in cluster 5 there are 1250 documents of which 40 belongs to classical music, 49 belongs to folk music, 922 belongs to rap music, 97 belongs to rock music and 142 belongs to pop music, while in cluster 6 there are 5246 documents of which 1119 belongs to classical music, 596 belongs to folk music, 1039 belongs to rap music, 1126 belongs to rock music and 1366 belongs to pop music.

- **Results of Latent Dirichlet allocation classifier**

LDA topic modeling is applied to analyze the topics hidden in these tweets. The LDA() function is going to run with Gibbs sampling technique and k = 5 as the number of topics. The time execution of LDA algorithm is 188 seconds. The top 10 frequent words in each topic are tabulated in Table using the terms () function.

Table : Top 10 terms of each topic.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
rock	music	music	rap	pop
music	folk	classical	music	music
like	free	sad	hiphop	best
great	live	listening	video	love
radio	album	art	youtube	dance
good	awards	piano	star	rnb
band	guitar	track	playlist	jazz
tribute	acoustic	king	play	songs
indie	concert	amazing	beats	hits
metal	blues	friends	artist	soul

The results lead to some observations, such as Topic 1 concerns mainly about the rock music category while Topic 2 corresponds mainly to the folk music category. Topic 3 describes mainly the classical music category, while Topics 4 and 5 concern mainly the rap music category and pop music category respectively. More specifically, Topic 1 consists of 2763 documents and Topic 2 of 2233 documents. Topic 3 consists of 2043 documents, Topic 4 of 1857 documents and Topic 5 of 1672 documents. However, as we saw earlier with the implemention of k-Means, each cluster can provide information about more than one topic. Thus, with the implementation of the LDA classifier, it can be seen that this algorithm can separate better categories. This was achieved after a good cleaning of the tweets using of a stop word list.

Table : Topic probabilities by document.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
9437	**0.359**	0.156	0.156	0.156	0.171
7750	0.142	0.142	0.171	**0.2**	**0.3**
7077	0.190	**0.333**	0.158	0.158	0.158
10333	**0.241**	0.151	0.181	**0.242**	0.181
8380	0.161	0.161	**0.354**	0.161	0.161
10575	**0.272**	0.196	0.151	**0.212**	0.166
2932	0.159	0.159	0.159	**0.376**	0.144
7095	**0.242**	**0.227**	0.196	0.151	0.181
7067	0.158	0.174	0.158	0.158	**0.349**
2932	0.159	0.159	0.159	**0.376**	0.144

The LDA algorithm calculates the topic probabilities associated with each document, and the tidy() function is used to portray those information. Table lists some of the topic probabilities of each document. As it appears from Table , LDA estimates that only 35.93% of the terms in the document 9437 are generated from Topic 1 while document 7750 is drawn almost from Topic 3 with 3% probability. In Table , the

highest probability in each row is bold and the second highest probability is highlighted in red.

Additional Interesting Results from the tweets

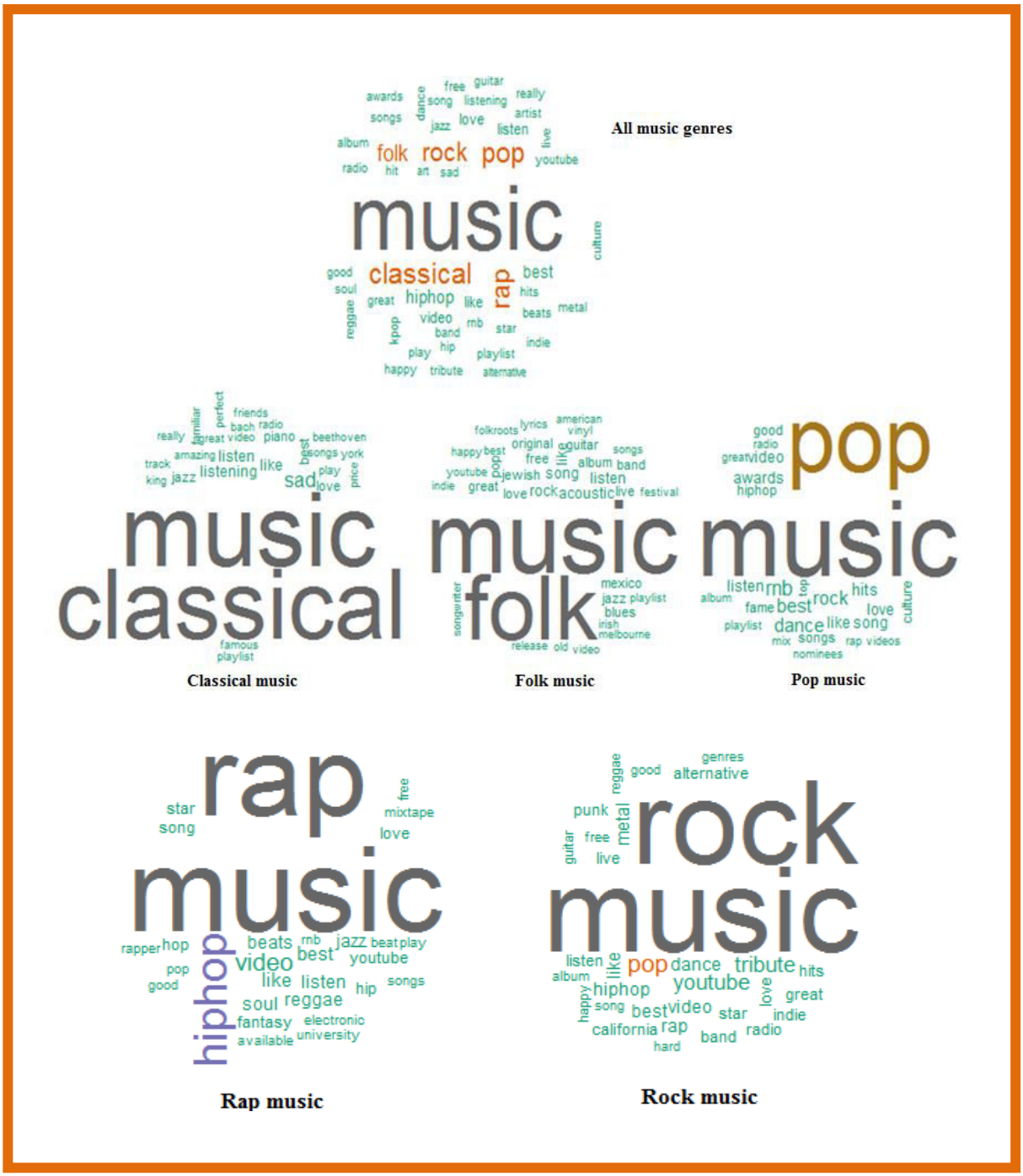

Figure : Word clouds of all music genres.

All music genres

Word	Frequency
music	9950
classical	2356
pop	2256
rap	2211
rock	2001
folk	1350

Classical music

Word	Frequency
music	2428
classical	2304
sad	252
like	129
listen	119

Folk music

Word	Frequency
music	1456
folk	1276
acoustic	84
rock	82
song	78

Pop music

Word	Frequency
music	2162
pop	1870
rnb	228
best	211
rock	206

Rap music

Word	Frequency
music	2070
rap	2008
hiphop	665
video	256
best	163

Rock music

Word	Frequency
music	1637
rock	1607
pop	211
youtube	152
tribute	145

Figure : Frequent terms for all music genres.

Useful information can be extracted and depicted from the text attribute of the dataset. Thus, using the wordcloud() function the frequency of the words was counted and depicted. The word cloud of all music genres is illustrated in Figure and the most used words in each music genre are illustrated in Figure . The word cloud of all music genres clearly shows that ″music″ is the top word with the ″classical″, ″pop″, ″rap″, ″rock″, and ″folk″ words to follow. Some other important words are ″love″, ″sad″, ″good″, ″great″, ″happy″, and ″like″, which are both positive and negative sentiments of the users. Another set of frequent words, ″reggae″, ″indie″, ″alternative″, ″jazz″ and ″metal″ are tweets about other different music genres. Some tweets characterize the music genres such as ″songs″, ″live″, ″artist″, ″band″, and ″tribute″. Finally, it is obvious that the word ″music″ must have appeared in all music genres.

Table : Words associations.

album								
the	folk	music	new	best	good	this	dance	just
0.08	0.04	0.04	0.04	0.03	0.03	0.02	0.01	0.01

song													
songs	rnb	love	pop	jazz	folk	hiphop	new	year	dance	good	listen	rap	this
0.10	0.09	0.08	0.08	0.07	0.06	0.05	0.05	0.05	0.02	0.02	0.02	0.01	0.01

great								
will	songs	make	live	year	time	band	folk	this
0.08	0.06	0.04	0.03	0.03	0.02	0.01	0.01	0.01

An approach for finding associations for a given term, which is a further form of count-based evaluation methods, can be realized with the findAssocs() function, that computes associations between terms in a Term Document Matrix or Document Term Matrix. Table presents the results of applying the function on frequently occurring terms, with at least 0.01 correlation, that is a minimum threshold which has been set. From Table , it is apparent that the term "song" is highly associated with the term "songs", which demonstrates a strong pairwise term association. In particular, the term "songs" was found in 10% of all the tweets.

Conclusion

We did show the high accuracy that the Naive Bayes algorithm has for classifying text data. Regarding the clustering algorithms, we have a poor clustering in the k-means algorithm as it did not make a good distinction between classes. It has been observed that if words with similar content are used in data-selection filters, the algorithm cannot perform proper grouping of clusters. As for the LDA algorithm, it gives good results, as it can separate the classes compared to the k-means algorithm. This was achieved by preprocessing the text using a stopword list to improving the quality of text, as it contains many syntactic features that may not be useful for analysis.