Social Media Based Behavior Prediction for Pakistani People

Nasir Amin[1], Muhammad Awais[2], Asad Cheema[3], Muhammad Mohsin Ashraf[4], Ayesha Shahid[5]

[1]Department of Software Engineering, Government College University, Faisalabad, Pakistan
[2]Department of Software Engineering, Government College University, Faisalabad, Pakistan
[3]Department of Information Technology, Government College University, Faisalabad, Pakistan
[4]Department of Software Engineering, Government College University, Faisalabad, Pakistan
[5]Department of Software Engineering, Government College University, Faisalabad, Pakistan

Abstract

Behavior prediction from sentiment analysis is a field of Natural Language Processing which addresses the problem of extracting sentiment or more generally, opinion from text. In this research work, predict the behavior specifically people of Pakistan from tweeter dataset, the Sentiment polarity classification in twitter to classify the messages and polarity of the sentiment towards behavior. Behavior data are extracted from twitter. In the methodology section, machine learning models are provided to support this research study. Machine learning models that are used is Random Forest Model (RFM), Decision Tree (DT) and Naïve Bays (NB). Python is used as developing tool to implement above mentioned models.

Key words:

Social Media, Prediction, Behavior, PTA, CSV, Machine Learning, Sentiments.

1. Introduction

Prediction of human behavior is a big factor to study the behavior in a social media network. Recognition of human behavior is a big challenge in intelligent supervision systems with different applications in crime prevention, abnormal case detection and emergency recognition. In earlier study, they indicated the approximate review on the structure using machine learning without experimental results, and the strategy for combining data of predicted and recognized behavior was not presented.

A. Social Media

Social Media played a huge role in shaping everyone's daily life, cultural norms and trends in the whole world. Many people used social media as the primary source of information related to hot issues and discussions in the country. News and social media play very important role in politician discussion and the behavior related to other subjects. In 2017-2018 Pakistan social media users were more than 44 Million with 30 Million having Facebook accounts as per the stats provided by Pakistan Telecommunication authority (PTA). Most people are used Facebook and Twitter in Pakistan. The raw information available on these platforms can be used to study our society's interests, trends and sentiments. Aim to study the behavior structures of Pakistani social media user interactions and the sentiment behind the conversations [1][2][3]. We are living in the age of microblogging, where changes the behavior of people about their interest and goals. Some social networks sites (SNS) are:

- Twitter (since 2006)
- Facebook (since 2004)
- LinkedIn (since 2002),
- Instagram (since 2010)
- Google Plus (since 2011)

Social networking sites (SNS) are represented in Figure No. 1 that are increasingly present in our daily activities. Most conventional web-based social media network is Facebook and Twitter. For example, it is common to find people who used social network frequently than e-mail or cellphone (voice) [4].

Fig. 1 Social Network Sites (SNS) [5].

Manuscript received October 5, 2018
Manuscript revised October 20, 2018

Facebook and twitter are most visited sites indicated by Alexa positioning [6]. Facebook had eight hundred million unique consumers in March 2011, 140 million information were around on the twitter. The rise of this new kind of data presents behavior prediction via social media with new techniques [7].

a) Twitter

Twitter is a famous application of social networking where users can write messages and shares their opinion called "tweets". Tweeter service quickly grow in world. In 2012, more than one hundred million users shared 340 million tweets in a day [8]. Twitter processed an average of 1.6 billion search queries on daily basis [9]. Figure no 2 represent that how tweeter users can `follow' each other's, be followed by others, or follow somebody. Coordinated social diagram of people (nodes) and their connections (edges) can be developed. Likewise, users can `tweet' or `retweet'.

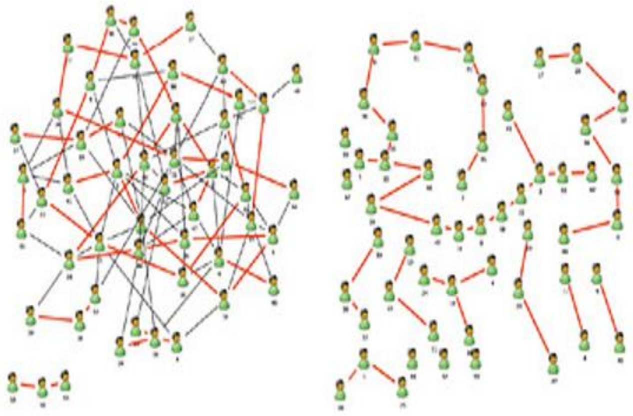

Fig. 2 Twitter Social Graph

b) Face Book

The biggest social organization where people can make `friend' and speak with each other. Figure no 3 representing the Facebook structure of social graph where peoples are node and their connections are edges in the chart.

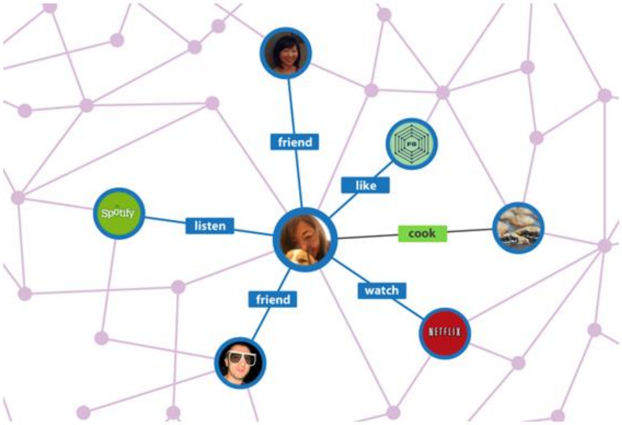

Fig. 3 Facebook Social Graph

The first step of this study to gather the data from social media network. Two option are chosen from Facebook and Twitter as these are the most widely used social networking platforms in Pakistan. After some studies and proofs it is realized that the data available on Facebook has many dimensions and contains a lot of attributes whereas the data on twitter are mainly just tweets from each user. However, the data available on Facebook may seem attractive but it's really hard to get the data and it give away very few public fields via their API and the only solution to get useful data is to register an app which is a very long and hectic process and usually takes months of correspondence.

c) Behaviors

There are different types of the users behavior in online social media network. Different peoples have different communities and different opinion and they share their thoughts, opinion and feelings on Twitter and Facebook etc. There are three types of the behavior that are predict.

- Positive Behavior
- Negative Behavior
- Neutral Behavior

d) Application of Machine Learning and Artificial intelligence in Behavior Prediction

As indicated by John McCarthy, AI is used to develop the algorithms that depend on the intelligent behavior [10]. Throughout the decades, development in artificial intelligence turned out to be satisfactory today and we have robots and robotics development around us [11].

Due to lack of human experts in the derivation of behavior and the limitation of existing behavior systems, there is a need for an efficient approach for the prediction of the behavior based on modern knowledge extraction and representation machine learning techniques. There are various machine learning techniques. Among them, Naïve Bayes (NB), Random Forest Model (RFM) and Decision

Tree (DT) are provided to support. Python developing tool are used to build above mentioned models.
Figure No. 4 represents machine learning workflow. Twitter datasets used in the models to predict the behavior. Later, gather the features that are used by machine learning algorithms especially for grouping the data. After that, predict the behavior people of Pakistan through above mentioned method by taking a trained dataset. At the end, checking the prediction results through test datasets. To build an accurate behavior prediction through social media network sites (Twitter).

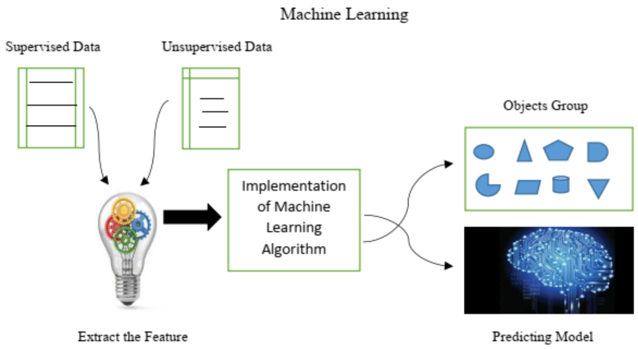

Fig. 4 Machine learning Work Flow Example

Main objectives;

1. To predict the behavior of peoples of Pakistan through social media network.
2. To build accurate models for behavior prediction.
3. To find out importance of machine learning techniques and effectiveness in the behavioral knowledge.

2. Methedology

In this section, machine learning models are provided to support.

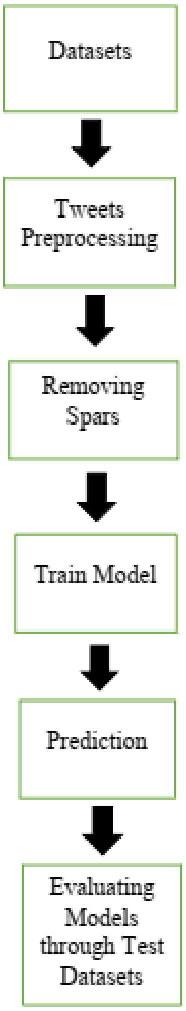

Fig. 5 Methodology Adopted for this study

A. Data sets

Datasets are used to analyze the sentiment of tweets and to predict the behavior of people of Pakistan into three categories:

1. Positive; tweet has positive sentiment.
2. Negative; tweet has negative sentiment.
3. Neutral; tweet has neither positive nor negative sentiment.

Twitter provides standard API without any charges which is used to extract the data of different users from a particular region. So, it’s decided to use Twitter data. However, the data from twitter are extracted and saved in a CSV file containing raw and unprocessed tweets. Three different data sets are used for experimental results, each data set contain 10 k tweets.

B. Data Preprocessing

Data preprocessing is the most important step in any analysis involving real world data. It is a bit easier to format numerical or categorical data. The real-world data has all sorts of anomalies and mistakes. It is necessary to format the data and modify it according to our use case and convert it to a structured form. Unstructured data cannot be used to feed in any analysis pipeline because it would not give us appropriate results. It is observed to check missing values, duplicates, unformatted strings and any other potential flaw in the dataset. As it is dealing with tweeter dataset, there are many irrelevant chunks of text which don't need in our analysis. Step by step process are used to eliminate such words and sentences from our tweets corpus. Pandas library provides all the right tools for data preprocessing and formatting.

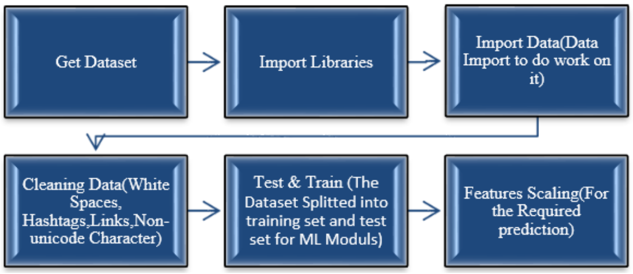
Get Dataset
Import Libraries
Import Data(Data Import to do work on it)
Cleaning Data(White Spaces, Hashtags,Links,Non-unicode Character)
Test & Train (The Dataset Splitted into training set and test set for ML Moduls)
Features Scaling(For the Required prediction)

Fig. 6 Data Preprocessing

C. Features of Data Set

Word cloud module provide to support for each three categories by including all the tweets belonging to that category like

- Positive Tweets
- Negative Tweets
- Neutral Tweets

To achieve this functionality python module named Word Cloud used. Word cloud module is famous python Pillow library for image generation. It also customize as per our preferences, setting the styles and colors etc. By writing a python function, any chunk of the text can be visualized easily. It is observed that the negative sentiment cloud contains, abusive language or hate speech whereas positive sentiment cloud words contains like “good”, “love”, “hope” etc. and the neutral sentiment is general words like day, human etc. that are shown in the below figures 7, 8 and 9 respectively.

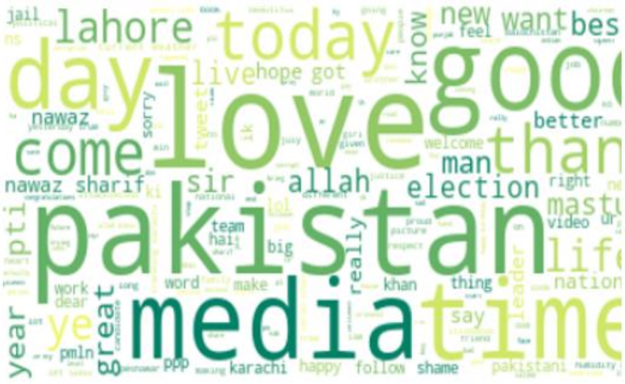

Fig. 7 Positive Word Cloud

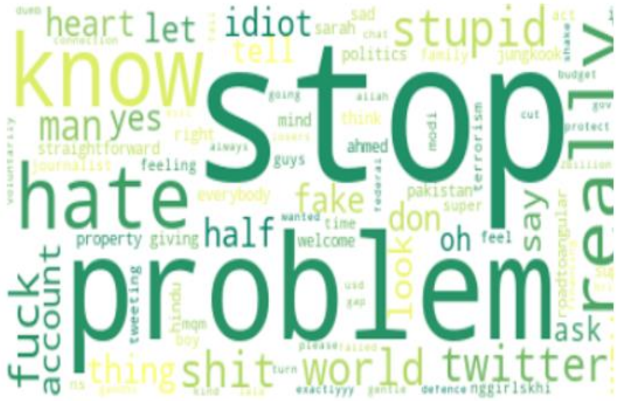

Fig. 8 Negative Word Cloud

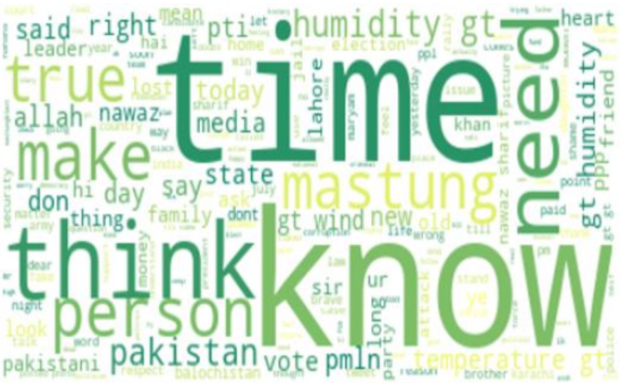

Fig. 9 Neutral Word Cloud

a) Machine Learning Training classifier

After completing all preprocessing steps, trained the models using SciKit Learn. SciKit Learn provided support for model Training. Features are extracted from tweets and train the negative, positive or neutral behavior through models. Figure no 10 show machine learning algorithm framework for these three models.

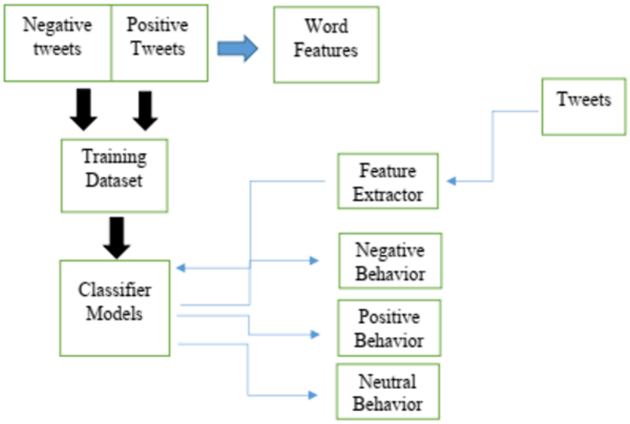

Fig. 10 Machine learning algorithm classifier Framework.

D. Tools

Required tools that are used for this research work are as following:

- Python
- Jupter
- Anaconda

E. Libraries Required

Some libraries are used for this research work are as follows:

- Panda
- TFIDF
- SciKit
- Pillow
- TextBlob
- Tweepy
- NLTK

3. Results and Discussion

Three approaches are used to predict the behavior of people of Pakistan by extracting tweets from twitter API. The sample of tweets datasets are small but some interesting results are got. Each approach showed different results despite being used with the same dataset.

After preprocessing the data, models are trained and showing the results.

Table 1: Adopted Methodology Results

Models	Positive	Negative	Neutral
Naïve Bayes	88.66 %	0.01 %	11.33 %
Random Forest	64.34 %	1.65 %	34.00 %
Decision Tree	27.54 %	6.15 %	66.32 %

Scikit Learn are provided to support the train the data sets, results are better and it can see in the figure no. 11, Naive Bayes (NB) assigned only 0.01 % of behavior the negative label and 11.33 % as neutral while all other 88.66 % behavior as positive. Results are represented in table no 1. Random Forest (RFM) are much more realistic. However it is calculated the Random Forest Model (RFM) predict 64.34 % positive behavior and 1.65 % behavior as negative and the rest of the 34.00 % predict as neutral behavior. In Decision Tree (DT) it is calculated 6.15 % as negative behavior and 66.32 % as neutral and 27.54 % towards positive behavior. Therefore, it can say that the overall behavior of Pakistan people is positive and some of them are neither positive nor negative.

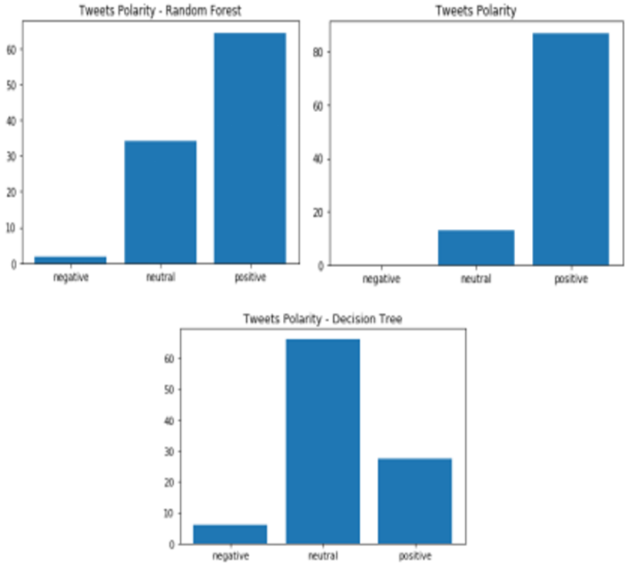

Fig. 11 Adopted Methodology Results

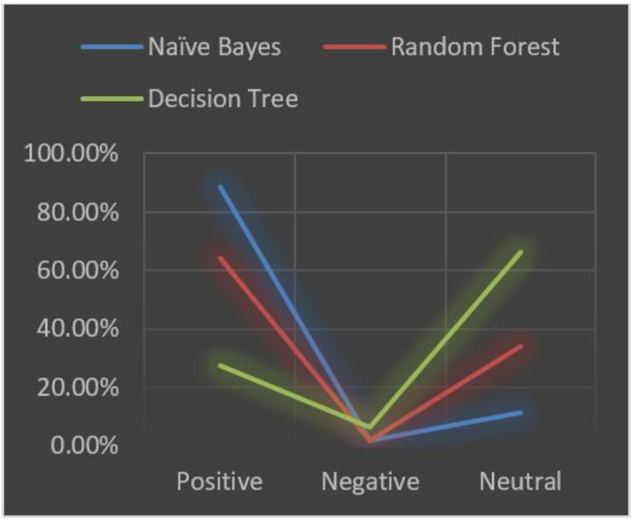

Fig. 12 Adopted Methodology Results

Based on the results and calculations the prediction of positive behavior is rich high in Random Forest and Naïve Bayes classifier. Another side of the results, neutral behavior is very less in Random and Naïve Bayes model and very high in the Decision tree. Based on the results, negative behavior is less in all three classifier that are shown in the figure no 12.

A Test Cases

The accuracy of test results consider in percentage. Here three test datasets are used for all models to predict the behavior and data set contains 10k. Test case 1, 2 and 3 are represented the results. In all three test cases the average accuracy of predict the behavior of Pakistan people in Random Forest Model (RFM) is 64.3 % is positive behavior, 87.8 % positive behavior in Naïve Bayes (NB) and 28.4 % the lowest average of positive behavior in Decision Tree (DT) Negative and neutral behavior accuracy is less in all these models, in Decision Tree only 65.5 % neutral behavior show in the average test cases result table. The people of Pakistan have positive behavior towards other country people and the results are highly support to prove this research work.

1) Test Case 1

Models	Positive	Negative	Neutral
Naïve Bayes	90.1 %	1.1 %	8.8 %
Random Forest	66.8 %	1.1 5	32.1 %
Decision Tree	28.5 %	4.1 %	67.4 %

2) Test Case 2

Models	Positive	Negative	Neutral
Naïve Bayes	88.3 %	0.7 %	11.0 %
Random Forest	63.5 %	2.0 %	34.5 %
Decision Tree	31.2 %	5.8 %	62.0 %

3) Test Case 3

Models	Positive	Negative	Neutral
Naïve Bayes	85.1 %	0.1 %	14.8 %
Random Forest	62.7 %	1.1 %	36.0 %
Decision Tree	25.5	7.0	67.5

4) Test Cases Average Results

Models	Positive	Negative	Neutral
Naïve Bayes	87.8 %	0.63 %	11.5 %
Random Forest	64.3 %	1.4 %	34.2 %
Decision Tree	28.4 %	5.6 %	65.6 %

Another graph are used to show test data sets results in figure no 13. Based on test data sets results and calculations the prediction of positive, negative and neutral behavior in Random Forest, Naïve Bayes classifier and Decision Tree is probably same like train data sets results.

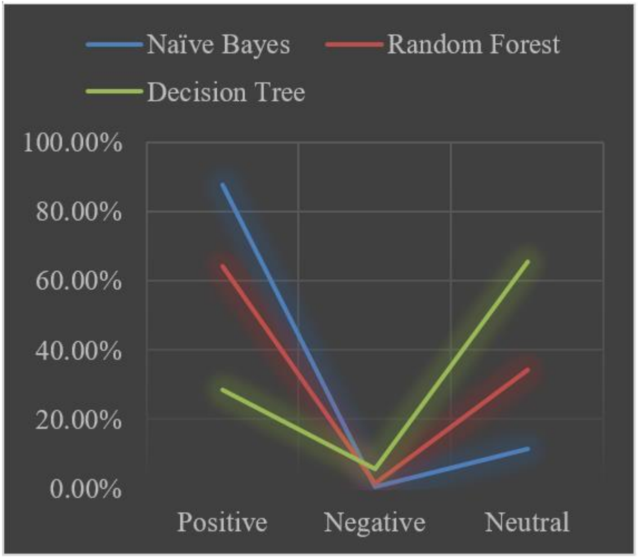

Fig. 13 Test Cases Average Results

4. Conclusion

Machine learning models are helpful in sentiment analysis algorithm, Random Forest Model (RFM), Naïve Bayes (NB) and Decision Tree (DT) provide to support to predict the behavior of Pakistani people over the demonetization datasets on different algorithm that are used from the twitter. Some variations in the results that are obtained from each approach although the tweets corpus are the same. In this research work, Naive Bayes (NB) model biased towards positive sentiment, whereas the Random Forest Model (RFM) and Decision Tree (DT) biased towards the neutral and negative sentiment. The results highly rich towards positive behavior. Test Cases results improve the efficiency of the behavior.

References

[1] W. N. Reynolds, M. S. Weber, R. M. Farber, C. Corley, A. J. Cowell, and M. Gregory, “Social media and social reality,” in 2010 IEEE International Conference on Intelligence and Security Informatics, 2010, pp. 221-226

[2] D. M. Romero, W. Galuba, S. Asur, and B. A. Huberman, “Influence and Passivity in Social Media,” in Proceeding of the 22th international conference on World Wide Web - WWW '11, 2011, p. 113-114

[3] B. D. Loader, “Social Movements and New Media,” Sociology Compass, vol. 2, no. 6, pp. 1920-1933, Nov. 2008.

[4] Lampos V, Cristianini N. Nowcasting Events from the Social Web with Statistical Learning. ACM Transactions on Intelligent Systems and Technology (TIST). 2012; 3(4): 72:1–72:22.

[5] Burlutskiy, N. (2017). Prediction of user behaviour on the web (Doctoral dissertation, University of Brighton). PN 14

[6] Alexa Internet Inc, "Alexa Top 500 Global Sites".http://www.alexa.com/topsites. [Accessed Jan 4, 2012].

[7] Bhattacharya P, Zafar MB, Ganguly N, Ghosh S, Gummadi KP (2014) Inferring user interests in the Twitter social network. In: Kobsa A, Zhou MX, Ester M, Koren Y (eds) Eighth ACM conference on recommender systems, RecSys ’14, Foster City, Silicon Valley, CA, USA—October 06–10, 2014, ACM, 357–360.

[8] Twitter (March 21, 2012). Twitter turns six. Twitter

[9] Twitter Search Team (May 31, 2011). "The Engineering behind Twitter's New Search Experience". Twitter Engineering Blog. Twitter. Archived from the original on March 25, 2014. Retrieved June 7, 2014.

[10] Mesaros, A., Heittola, T., & Virtanen, T. (2016, August). TUT database for acoustic scene classification and sound event detection. In Signal Processing Conference (EUSIPCO), 2016 24th European (pp. 1128-1132). IEEE.

[11] Giannoulis, D., Stowell, D., Benetos, E., Rossignol, M., Lagrange, M., & Plumbley, M. D. (2013, September). A database and challenge for acoustic scene classification and event detection. In Signal Processing Conference (EUSIPCO), 2013 Proceedings of the 21st European (pp. 1-5). IEEE.