Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 98 (2016) 596 – 603

Procedia Computer Science

The 2nd International Symposium on Web of Things and Big Data (WoTBD 2016)

Automatic Categorization of Social Sensor Data

Olivera Kotevska[a], Sarala Padi[b], Ahmed Lbath[a]

[a]*University of Grenoble Alpes, CNRS, LIG, F-38000, Grenoble, France, [kotevska.olivera;lbath.ahmed]@imag.fr*
[b]*Indian Institute of Technology, Madras, India, sarala@cse.iitm.ac.in*

Abstract

Today, there is a huge impact on generation of data in everyday life due to micro blogging sites like Twitter, Facebook, and other social networking web sites. The valuable data that is broadcast through micro blogging can provide useful information to different situations if captured and analyzed properly in timely manner. When it comes to Smart City, automatically identifying messages communicated via Twitter can contribute to situation awareness about the city, and it also brings out a lot of beneficial information for people who seek information about the city. This paper addresses processing and automatic categorization of micro blogging data; in particular Twitter data, using Natural Language Processing (NLP) techniques together with Random Forest classifier. As processing of twitter messages is a challenging task, we propose an algorithm to automatically preprocess the twitter messages. For this, we collected Twitter messages for sixteen different categories from one geo-location. We used proposed algorithm to preprocess the twitter messages and using Random Forest classifier these tweets are automatically categorized into predefined categories. It is shown that Random Forest classifier outperformed Support Vector Machines (SVM) and Naive Bayes classifiers.

Peer-review under responsibility of the Program Chairs

Keywords: Random Forest classifier; Micro blogging data; Twitter Data Analysis; Smart Cities; Automatic Categorization of Twitter data;

1. Introduction

Rapid growth of textual information on the web in the past years has influenced the way people communicate, share and get information. Especially, in the context of web, people share their opinions and sentiments for different purposes. People also use different forms of text to express their thoughts or opinions, like pictures, videos, and text. When it comes to social media, it has become attractive source for information access as well as information generation. It has become more popular, and people started using Twitter, Facebook, etc. for writing posts, blogs and events that are happening in everyday life. It also attracts attention for the information sharing capabilities and used effectively in different domains, as well as entertainment and brand related communications. Many significant

* Corresponding author. Tel.: +1-847-404-6900
E-mail address: kotevska.olivera@imag.fr

Peer-review under responsibility of the Program Chairs
doi:10.1016/j.procs.2016.09.093

achievements are accomplished using social networks as data source in different areas like early warning systems for detection of earthquakes, for predicting the German federal elections [1,2,3,4].

In past few years, micro blogging, messages with limited number of characters, has become widely used tool for communication on the Internet. Twitter is one of the first and most popular micro blogging providers with millions of active users. Each user is able to create public posts to initiate discussions, to participate in debates, and to follow the communication of others. As a result, Twitter is widely used communication channel across a wide range of applications for everyday communication purposes. Noteworthy research into some such areas is now emerging, but largely remains in the form of topic, context, and event-related case studies that are able to give substantial light on specific uses of Twitter[5]. There are some efforts in finding the sentiments in order to find the emotions embedded in each tweet by performing linguistic analysis on corpus of tweets[6]. The same linguistic analysis is performed to extract the features for finding the sentiments of Twitter messages[7]. There are other efforts to automatically classify the tweets using clustering method[8].

Monitoring the social sensor activities is a good way to measure customers' loyalty, keeping track of sentiments towards brands, products or just measure their perceptions regarding variety of topics. Having information for what topics people are interested can help to improve service recommendations, like traffic routes, air pollution zones, etc. In past few years, significant research have been done on Twitter data for extracting the sentiments. Most of the research include Naive Bayes classifier and with different features for sentiment analysis purposes. In this paper, we compared existing and widely used algorithms for classification of tweets together with different features. Closer to our experiment Aphinyanaphongs et. al used Random Forest classifier for classification of Twitter tweets into two classes[9].

In the Smart City context integrating information from different sources in real-time is a necessity to make smarter decisions for the city as a whole. To provide this, we created a framework that provides integration, aggregation and processing of sensed data which is part of big project based on cloudlet architecture[10] and the Figure 1 shows the general overview of the framework used for analysis.

As described in the Figure 1, in this paper we focused on creating a process for automatically identifying the topics from Twitter as a social sensor that contribute to situation awareness about the city. The work presented mainly focused on sentiment analysis of Twitter data to automatically categorize the tweets into different categories for information retrieval purposes. Automatic categorization also helps to extract knowledge from tweets in order to provide information to users. The machine learning methods and features used in this paper can be adapted to our future work that involves real time processing of tweets for categorization, sentiment analysis and topic extraction for a large quantity of data.

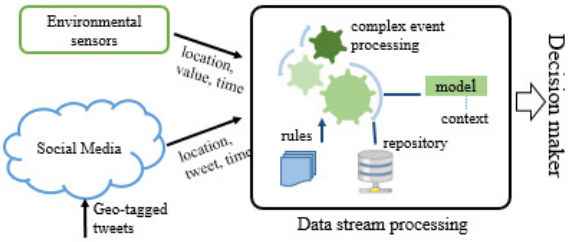

Fig. 1: General overview

The rest of the paper is organized as follows. In Section 2, we give details about the data collection from Twitter social media and discuss details about pre-processing methodology and additional resources used for pre-processing of tweets. In Section 3, features used to represent the text messages in terms of vector space models and different machine learning methods used for categorisation of tweets into predefined categories are described in detail. Section 4 gives complete details of experimental evaluation and classification accuracy on different datasets using different features. Finally, Section 5 concludes the work.

2. Twitter Data Collection for Analysis

2.1. Collecting the Data

For analysis purposes, Twitter data is collected using its Application Programming Interface (API). It is comparatively simple to capture comprehensive data sets of vast majority of all the tweets. Tweets received by Twitter

streaming API are anywhere from 1% of tweets to over 40% of tweets in near real-time. Since the basic idea of this paper is to analyze tweets posted by the people from one location, we collected Twitter data from New York City (NYC). Twitter provides two types of location data, one is using the name of the city and other is using the exact Global Positioning System (GPS) coordinates.

For this study, we choose to use GPS location for NYC data because we can collect tweets in consistent manner for each category. For analysis, two data sets are used and both are from same geo-location coordinates and English language as a filter. As shown in Table 1, *Data set* 1 is a general data set with tweets containing generic terms along with geo-location and English language as filters. Whereas, the *Data set* 2 is composed by tweets that refer to named entities, in this case named entities used for extraction of tweets are sixteen, namely, art, music, film, books, health, sport, food, travel, holidays, tech, weather, religion, news, fashion, shopping, celebrities. These entities are chosen from analysis based on the statistics about most frequently used topics in social media.

Table 1: Number of tweets used for analysis

Training data		Testing data	
Data set 1	Data set 2	Data set 1	Data set 2
1303	1746	1702	8828

2.2. Pre-processing

Twitter text data is unstructured and noisy in the sense that it contains slang, misspelled words, numbers, special characters, special symbols, shortcuts, URLs, etc. The text messages with these special symbols, images may be easier for humans to read and analyze. When the text data is mixed with other types of symbols and images, processing is a major challenging task compared to processing of normal text data. As a result, pre-processing of Twitter data plays a major role in sentimental analysis. The typical characteristics of Twitter data that makes sentiment analysis a challenging research area are: messages are very short and contain less text, message may contain different language text, it contains special symbols with specific meaning, data contains lot of shortcuts, and data has spell mistakes.

These typical characteristics make pre-processing of Twitter data a challenging task for further analysis purposes. This paper discusses the methodology together with Natural Language Processing (NLP) techniques for efficient processing of Twitter messages for analysis purposes. We propose an algorithm implemented in Java Programming language where we incorporated sentiment-aware tokenization [1] while pre-processing the tweets. The proposed algorithm is described as follows:

Algorithm 1 Pre-processing algorithm

```
procedure PRE-PROCESSING OF TWEETS
    for each tweet t_i ∈ T do
        Remove URLs, re-tweets, hash tags, repeated punctuation's
        for each word w_j ∈ t_i do
            miscellaneous symbols
            emotion icons, contractions
            abbreviations, acronyms, smilies
            misspelling words
        end for replace it with full meaningful words
        Remove stop words, punctuation's, non-English words
        Convert to lower case characters
    end for
end procedure
```

It is observed from the collected Twitter messages that emoticons are extremely used in many forms of social media. It is the same case for acronyms, abbreviations or slang words. Because of these reasons, we used implementation

[1] "http://sentiment.christopherpotts.net/tokenizing.html"

functionality to convert smileys[2], emoticons[3], acronyms and abbreviations[4][5][6], contractions [7] and misspelled words[8] to full meaningful words. Table 2 shows the number of conversion inputs used in each category.

Tweets are processed by removing characters like repetitions, punctual characters, stop words, and English stop words[9]. Even though collected tweets are in English language, there were words in other languages, in such cases tweets are ignored for analysis.

Table 2: Dictionary lists used for pre-processing of tweets

List name	Number of lines
Smiles	247
Emoticons	40
Acronyms, Abbreviations and Initials	689
Contractions	51
Misspelling	5875
Stop words	319

Despite the advantages of reducing vocabulary, shrinking feature space and removing irrelevant distinctions and icons is that pre-processing can collapse relevant distinctions, that are important for analysis purposes. Generally, pre-processing of text data improves the quality of text for analysis purposes, whereas coming to twitter data, because of short messages, pre-processing may end up with messages with no text data left for the Twitter message. In many cases, after pre-processing Twitter messages hardly contain one or two words, Table 3 shows Twitter data statistics before pre-processing phase. We can see that tweet messages contain a lot of punctuational marks, stop words, numbers, and non- english words that would not convey any information for the context, and are not useful for any analysis. This becomes a big challenge in pre-processing of Twitter data for analysis purposes.

3. Categorization of Twitter messages

3.1. Extract features

Text data is a sequence of words and these words cannot be fed directly to the machine learning algorithms for analysis purposes. Most of the algorithms expect numerical feature vectors with a fixed size rather than the raw text with variable length. In order to address this, we need to use techniques that provide utilities to extract numerical features from text content. We use the most frequently used features called Bag of Words (BOWs) and Term Frequency Inverse Document Frequency (TF-IDF) vector representations to represent text messages in terms of a feature vector.

Table 3: Data statistics before pre-processing

Statistics/Database Name	Data Set 1	Data Set 2
Tweets	14479	11032
Tokens	137104	138907
Twitter tags, Re-tweets, URLs	14879	18923
Signs	22	2
Contractions	2033	901
Misspell words	668	231
Punctuational marks	53854	76629
Abbreviations, Acronyms, Smilies	1075	3817
Stop words	52696	40897
Numbers	9166	14558
No-English words	10853	13644

[2] "http://www.netlingo.com/smileys.php"
[3] "http://en.wikipedia.org/wiki/List_of_emoticons"
[4] "http://marketing.wtwhmedia.com/30-must-know-twitter-abbreviations-and-acronyms/"
[5] "https://digiphile.wordpress.com/2009/06/11/top-50-twitter-acronyms-abbreviations-and-initialisms"
[6] "http://www.muller-godschalk.com/acronyms.html"
[7] "http://www.sjsu.edu/writingcenter/docs/Contractions.pdf"
[8] "https://en.wikipedia.org/wiki/Wikipedia:Lists_of_common_misspellings"
[9] http://xpo6.com/list-of-english-stop-words/

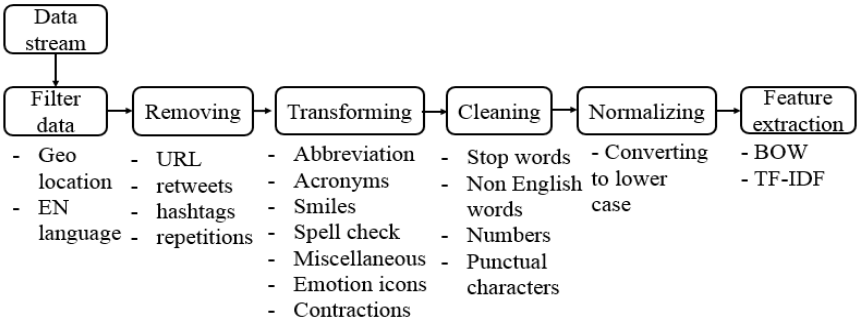

Fig. 2: Data model used to extract the features for Twitter analysis purposes

The data model used to extract these two features
is depicted in Figure 2. This figure also shows the pre-processing steps for tuning the Twitter messages before extracting the BOWs and TF-IDF features. In most of the NLP applications, $BOWs$ and $TF-IDF$ features are frequently used for text processing applications, sentimental analysis on Twitter data, blogs and classification of sentiments from micro-blogs [11,12,13]. Even though these features are extensively used for most of the text processing applications, for completeness purpose, a briefly explanation is included as follows:

1. Bag-of-Words (BOWs): This model represents text as an un-ordered collection of words, disregarding the word order. In the case of text classification, a word in a text message is assigned a weight according to its frequency in the text messages. The BOW representation of Twitter text message 't_n' is a vector of weights '$W_{1n},\ldots,W_{wn}$' where 'W_{in}' represent the frequency of the i^{th} term in the n^{th} text message. The transformation of a text message '**T**' into the $BOWs$ representation enables the transformed set to be viewed as a matrix, where rows represent Twitter text message vectors, and columns are terms in each Twitter text message [14].

2. Term Frequency and Inverse Document Frequency (TF-IDF): It is a feature vector representation method where frequent and rare terms in the text messages are normalised so that rare terms are more emphasised along with frequent terms in the text messages. Term frequency $TF(t_i,T)$ is the number of times the term 't_i' appears in a Twitter text message 't_m', while document frequency $DF(t_i,T)$ is the number of Twitter text messages contains the term 't_i. If we only use term frequency to measure the importance, it is very easy to over-emphasize terms that appear very often but carry little information about the Twitter text message. If a term appears very often across all the Twitter text messages, it means it doesn't carry special information about a particular text message. Inverse document frequency is a numerical measure of how much information a term provides and it is defined as follows:

$$TF-IDF\,(t_i,t_m,T) = TF\,(t_i,t_m)\times IDF\,(t_m,T)$$

$$IDF(t_i,T) = \log\left(\frac{T}{1+|t_m\in T : t_i\in t_m|}\right)$$

where $|T|$ is the total number of text messages in the corpus. Since logarithm is used, if a term appears in all text messages, its IDF value becomes 0. Note that a smoothing term is applied to avoid dividing by zero for terms outside the corpus.

3.2. Categorization of Tweets

Classification of online stream tweets helps to find important information up to date for each type of category. In this paper, tweets are analyzed and classified into predefined categories using supervised learning techniques: Naive Bayes, Support Vector Machines (SVM) and Random Forest (RF) classifiers where Naive Bayes classifier is a probabilistic classifier and SVM is a discriminative classifier. Random Forest classifier is a ensemble method where

more than one decision tree is used for classification purposes based on voting rule[15]. In this paper, these three machine learning models are used for automatic categorization of tweets into predefined categories. Initially, models are trained on training data set as tabulated in Table 1 and these trained models are used to automatically classify the test data set.

For illustration of data, Figure 3 shows the word cloud of **Food** and **Sport** category of tweets. As we can observe from figure, in both of these word clouds, the most dominant words are highlighted. It is interesting to see that the most dominant word in each of the categories are **FOOD** and **SPORT** that are exactly same as category labels. It is also worth noting that, in both the clouds, there are dominant words that are not related with with the category of the tweets like JUST in Food cloud, or NEW in Sport cloud.

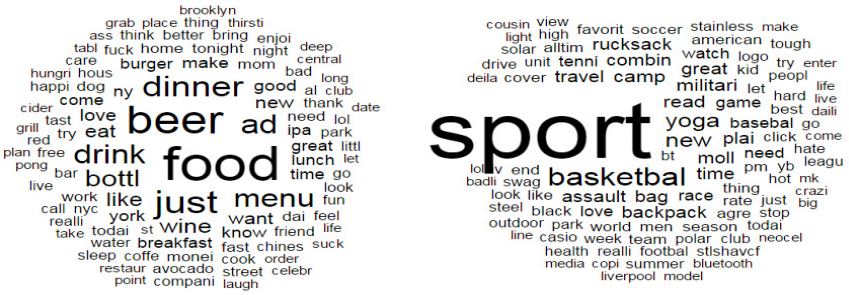

Fig. 3: Word cloud illustration of tweets belong to Sport and Food categories

4. Experimental Analysis

4.1. Data used for analysis

For experimental analysis, we used Twitter social network as a data source and training data sets are created. The ground-truth for training and testing datasets are created manually. The type of tweets and number of categories of tweets used in this study are shown in Table 4. Table 1 shows the data statistics after pre-processing of tweets. There are 1303 tweets for *Data set* 1 and 1746 tweets for *Data set* 2, and 3049 tweets are used for training purposes. For testing purposes, 1702 for *Data set* 1 and 8828 tweets for *Data set* 2 are used.

Table 4: Tweets used for experimental analysis

Label Name	Tweets for Training	Tweets for Testing	Total No. of Tweets
Art	149	4071	4220
Music	286	13877	14163
Film	238	5731	5969
Books	186	3142	3328
Health	134	3304	3438
Sport	151	2400	2551
Food	507	15077	15584
Travel	118	2363	2481
Holidays	18	150	168
Tech	122	1999	2121
Weather	521	8634	9155
Religion	161	1312	1473
News	198	9427	9625
Fashion	122	2383	2505
Shopping	81	2039	2120
Celebrities	55	754	809

The collection of Twitter data tweets and preprocessng of tweets are performed in Java programming language. After pre-processing of tweets, training data sets are used to build the machine learning models for sixteen categories in Python [10] using Scikit-learn package [16] along with NLTK tool [17]. *BOWs* and $TF-IDF$ features are extracted using NLTK tool and Naive Bayes, SVM, and Random Forest classifiers are used from Scikit-learn package in Python. The classification accuracy reported in this paper is calculated as:

$$\text{Accuracy} = \frac{\text{Correctly classified tweets}}{\text{Total number of tweets}} \quad (1)$$

4.2. Experimental Results

Initially, machine learning models are trained individually for sixteen classes. During testing, the trained models are used to automatically categorize the testing tweets. Results are calculated based on ground-truth marked for testing examples. Table 5 shows the classification accuracy with respect to test dataset and from table we can see that Random Forest classifier gives almost 94% accuracy as compared to SVM and Naive Bayes classifiers. The overall accuracy on both data sets are almost 90% accurate and this accuracy has come down due to the fact that **FOOD** class examples are misclassified, it is almost 50% accurate. As a result the overall accuracy is reduced.

Table 5: Classification accuracy of tweets into predefined categories

Classifier	Classification Accuracy (%)					
	BOWs			TF-IDF		
	DS 1	DS 2	Overall	DS 1	DS 2	Overall
Naive Bayes	58.87	82.77	78.90	67.45	91.01	87.20
SVM	51.70	91.68	85.22	65.21	90.79	86.65
Random Forest	66.80	94.27	89.83	69.09	93.50	89.56

DS 1 - Data set 1 and DS 2 - Data set 2

4.3. Discussions

The type of methodology used for processing the tweets is very important and crucial step for sentimental analysis and opinion mining. From the results we can see that the most frequently used topics are about 'FOOD' and 'MUSIC' and less important topics are 'Celebritis' and 'Shopping'. To further clarify the results with respect to *Data set* 1, Figure 4 shows the class wise accuracy for sixteen classes. From this figure, we can notice that **FOOD** class has the lowest accuracy compared to other classes that result in decreas of overall accuracy of the data set. One solution to reduce missclassification in this case is that building hierarchical classification models so that missclassified examples belong to **FOOD** category can be reduced. It is also worth looking at multi-label classification approaches or probabilistic topic models for finding the semantics of tweets for better categorization purposes.

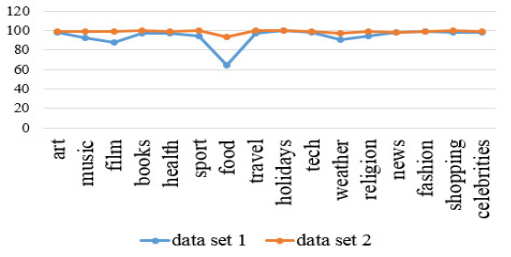

Fig. 4: Accuracy by category for both the data sets using BOWs feature with SVM classifier

5. Conclusion and Future Work

This paper mainly focused on exploring the general patterns of social media usage and presenting a model for automatically categorizing the analytics for a wide range of predefined identifiers over one concrete geo-location, in

[10] http://docs.continuum.io/anaconda

this case, New York City. The paper presents an algorithm for pre-processing of tweets in efficient way where every word in a tweet is important for analysis. The experiments shown that the pre-processing algorithm used to process the tweets really helps to efficiently categorize the tweets into predefined categories. It is shown that Random Forest Classifier combined with TF-IDF feature gives better results compared to SVM and Naive Bayes classifiers.

As a future work, we want to further analyze the tweets in more efficient manner and discover hidden structures. Since the tweets are very short messages, it is worth looking at probabilistic topic models for improving the measures and to better analyze the data for sentimental analysis purposes. It is also worth and useful to look at categorization of tweets in terms of multi labeling perspective for finding the tweets that are belonging to more than one topic. This work can be extended to real time processing of tweets that involve processing of vast amount of data where we need to make use of Apache Hadoop or Apache Mahout frameworks for efficient processing of large amount of social sensor data.

Acknowledgements

This work was supported by the National Institute of Standards and Technologies (NIST), and conducted within a collaboration under Information Technology Laboratory, Advanced Network Technologies Division (ANTD) and University of Grenoble. Our special thanks to Dr. Abdella Battou, ANTD division chief for his support an advises.

References

1. Avvenuti, M., Cresci, S., La Polla, M.N., Marchetti, A., Tesconi, M.. Earthquake emergency management by social sensing. In: *Pervasive Computing and Communications Workshops (PERCOM Workshops)*. IEEE; 2014, p. 587–592.
2. Sakaki, T., Okazaki, M., Matsuo, Y.. Earthquake shakes twitter users: Real-time event detection by social sensors. In: *Proceedings of the 19th International Conference on World Wide Web*; WWW '10. New York, USA: ACM; 2010, p. 851–860.
3. Broniatowski, D.A., Paul, M.J., Dredze, M.. National and local influenza surveillance through twitter: an analysis of the 2012-2013 influenza epidemic. *PLoS ONE 8(12)* 2013;.
4. Tumasjan, A., Sprenger, T.O., Sandner, P.G., Welpe, I.M.. Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM* 2010;**10**:178–185.
5. Bruns, A., Stieglitz, S.. Quantitative approaches to comparing communication patterns on twitter. *Journal of Technology in Human Services* 2012;**30**(3-4):160–185.
6. Pak, A., Paroubek, P.. Twitter as a corpus for sentiment analysis and opinion mining. In: *LREC*; vol. 10. 2010, p. 1320–1326.
7. Kouloumpis, E., Wilson, T., Moore, J.. Twitter sentiment analysis: The good the bad and the omg! *Icwsm* 2011;**11**:538–541.
8. Rosa, K.D., Shah, R., Lin, B., Gershman, A., Frederking, R.. Topical clustering of tweets. *Proceedings of the ACM SIGIR: SWSM* 2011;.
9. Aphinyanaphongs, Y., Lulejian, A., Brown, D.P., Bonneau, R., Krebs, P.. Text classification for automatic detection of e-cigarette use and use for smoking cessation from twitter: A feasibility pilot. In: *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*; vol. 21. NIH Public Access; 2016, p. 480.
10. Kotevska, O., Lbath, A., Bouzefrane, S.. Toward a real-time framework in cloudlet-based architecture. *Tsinghua Science and Technology* 2016;**21**(1):80–88.
11. Saif, H., He, Y., Alani, H.. Semantic sentiment analysis of twitter. In: *The Semantic Web–ISWC 2012*. Springer; 2012, p. 508–524.
12. Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R.. Sentiment analysis of twitter data. In: *Proceedings of the Workshop on Languages in Social Media*. Association for Computational Linguistics; 2011, p. 30–38.
13. Bermingham, A., Smeaton, A.F.. Classifying sentiment in microblogs: Is brevity an advantage? In: *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*. New York, USA: ACM; 2010, p. 1833–1836.
14. Radovanović, M., Ivanović, M.. Text mining: Approaches and applications. *Novi Sad J Math* 2008;**38**(3):227–234.
15. Breiman, L.. Random forests. *Machine learning* 2001;**45**(1):5–32.
16. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research* 2011;**12**:2825–2830.
17. Loper, E., Bird, S.. Nltk: The natural language toolkit. In: *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics-Volume 1*. Association for Computational Linguistics; 2002, p. 63–70.