International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 6, No.4 (July-2017)

http://dx.doi.org/10.12785/ijcds/0600402

Sentiment Analysis of Social Sensors for Local Services Improvement

Olivera Kotevska[1] and Ahmed Lbath[2]

[1,2] University of Grenoble Alpes, CNRS, LIG, F-38000, Grenoble, France

Received 2 Dec. 2016, Revised 3 Jun. 2017, Accepted 15 Jun. 2017, Published 1 July 2017

Abstract: Today, there is an enormous impact on a generation of data in everyday life due to microblogging sites like Twitter, Facebook, and other social networking websites. The valuable data that is broadcast through microblogging can provide useful information to different situations if captured and analyzed properly promptly. In the case of Smart City, automatically identifying event types using Twitter messages as a data source can contribute to situation awareness about the city, and it also brings out much useful information related to it for people who are interested. The focus of this work is an automatic categorization of microblogging data from the certain location, as well as identify the sentiment level at each of the categories to provide a better understanding of public needs and concerns. As the processing of Twitter messages is a challenging task, we propose an algorithm to preprocess the Twitter messages automatically. For the experiment, we used Twitter messages for sixteen different event types from one geo-location. We proposed an algorithm to preprocess the Twitter messages, and Random Forest classifier automatically categorize these tweets into predefined event types. Therefore, applying sentiment analysis to tweets related to these categories allows if people are talking in negative or positive context about it, thus providing valuable information for timely decision making for recommending local service. The results have shown that Random Forest performs better than Support Vector Machines and Naive Bayes classifiers, and combining sentiment score with cosine similarity of event types provides more detailed understanding for the identified public categories.

Keywords: Automatic Categorization; Microblogging data; Public services; Smart Cities; Twitter.

1. INTRODUCTION

In the city context, participatory sensing can be used to retrieve information about the environment, weather, well-being, traffic congestion, trends in the local economy, dangers or early warnings, as well as any other sensory information that collectively become useful knowledge for the city improvement and smartness. The rapid growth of textual information has influenced the way people communicate, share and get information. Especially, in the context of the web, people share their opinions and sentiments for different purposes. People also use various forms of text to express their thoughts or opinions, like pictures, videos, and text. In the case of social media, it has become an attractive source for information access as well as data generation. It has become more popular, and people started using Twitter, Facebook, and so forth 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 achievements are accomplished using social networks as a data source in various areas like newscasts, early warning systems for detection of earthquakes and predicting the German federal elections [2], [6], [9], [24], [25], [29].

Microblogging, messages with a limited number of characters, has become a widely-used tool for communication on the Internet in the past few years. Twitter is one of the first and most popular microblogging providers with millions of active users. Each user can create public posts to initiate discussions, to participate in debates, and to follow the communication of others. Thus, Twitter is widely used communication channel across a wide range of applications for everyday communication purposes, like in cases when some real-world event happen, for example, a soccer game, adverse weather update, elections, breaking news and so forth.

In this study, each user is considerate as a sensor and tweets are sensor information with the theme, time, and location features. Identifying events from social media

E-mail: kotevska.olivera@imag.fr, lbath.ahmed@imag.fr

presents several challenges:

- Heterogeneity and immense scale of the data
- social media post is short, which means that only a limited content is available for analysis.
- Frequent use of informal, irregular, and abbreviated words, the large number of spelling and grammatical errors, and the use of awkward sentence structure and mixed language.

We are focused on tweets that will result in analyzing the view of the public on generally discussed topics and measure their perceptions regarding a variety of topics. Timely understanding of the tweets reporting various concerns about the city is necessary for city authorities to manage city resources. This information complements similarity and sentiment level measure.

Sentiment analysis (SA) have been used to measure opinions of users about some product if it is satisfactory or not. Marketers and companies use it to understand if their product or service meets users' requirements, while end consumers tend to look at reviews of a product before buying it. SA involves classifying the text into categories like 'positive,' 'negative,' 'neutral,' or even in more detailed levels. SA in Tweeter recently attracted much interest by researchers. It tackles the problem of analyzing the tweets regarding the opinion they express. It has been used to measure sentiment during Hurricane Irene [11] and terrorism [19]. Social media triggered the rise of sentiment analysis which brings new possibilities to city government in general and decision making [1]. SA can contribute to a better understanding of, and appropriate reactions to public's needs and concerns by city governments. Measuring the sentiment at certain area and topic helps to determine the relevant services for the users and promote relevant recommendations (content, collaborative, or hybrid filtering) based on that.

Therefore, in a domain of urban context aware application, we use a case of improving local services by identifying event types and sentiment measurement from social sensors according to contextual information. Zhao et al. [32] classified the event detection on Twitter in three categories: specific event detection, person related event detection, and general event detection. For our analysis, we are interested in specific and general event detection. Therefore, we develop a framework for automatic categorization of social sensor data, and sentiment measurement improves local city services.

We organized this paper as follows; related work in the area of event detection through finding sentiment score and associated public services is presented in section 2. In section 3 we describe functional architecture of the presented system. While in section 4 we give details about experiment setups, like the data collected from Twitter social media and discuss details about pre-processing methodology and additional resources used for pre-processing of tweets. In Section 5, features used to represent the text messages regarding vector space models and different machine learning methods used for categorization of tweets into predefined categories are described in detail. Section 6 shows sentiment label on detected topics, while section 7 gives complete details of experimental evaluation and classification accuracy on different datasets using different features, and similarity between subjects. Finally, Section 8 concludes the work.

2. Related work

Noteworthy research into topic detection and sentiment analysis is now emerging but largely remains in the form of a subject, context, and event related case studies that can give strong light on specific uses of Twitter [10]. In the domain of event detection in Twitter exists a few approaches, based on a type of event (specified and unspecified), detection method (supervised and unsupervised) and detection task (retrospective event detection and new event detection) [5]. In our work, we focused on specified supervised detection approaches. This method was chosen by [9], [25], and [29] for an earthquake, influence and election detection and prediction. While specified unsupervised event detection, was used by [22], he proposed a method for semantic topic extraction and tracking news events and can provide notification and awareness for users.

Most of the research studies include Naive Bayes classifier and with different features for sentiment analysis purposes. However, we compared existing and widely used algorithms for classification of tweets together with various features. Closer to our experiment is [4] used Random Forest classifier for classification of Twitter tweets into two classes. Authors in [27] combine trend detection and sentiment analysis for decision-making purposes based on the Spanish language, while authors in [12] developed a service to detect social events using Twitter messages (tweets) as the input source.

There are some efforts in finding the sentiments from emotions embedded in each tweet by performing linguistic analysis on a corpus of tweets [21]. The same linguistic analysis is conducted to extract the features for finding the sentiments of Twitter messages [16]. Jiang et al. [15] classified tweet polarities by focusing on the syntactic relationships to the target query.

Other authors [26], [3], [30] focus on certain geo-location (city, neighborhood) and identify the topics and sentiments related to them; their output is intended to help city representatives, first responders or citizens.

3. PROBLEM DEFINITION AND PROPOSED SOLUTION

The general high-level overview of the architecture for integrating data from multiple data sources to provide context richer and more accurate city-based services is presented in Fig. 1, and more detailed explained in [17]. On the left side are physical and social sensors they receive the sensed data from the environment. Moreover, on the right side is the data stream processing unit, which is filtering the relevant information, and apply analytics (e.g. detects events, patterns, and relationships between the events) to enable decision making.

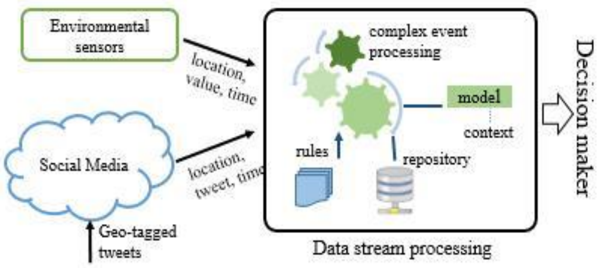
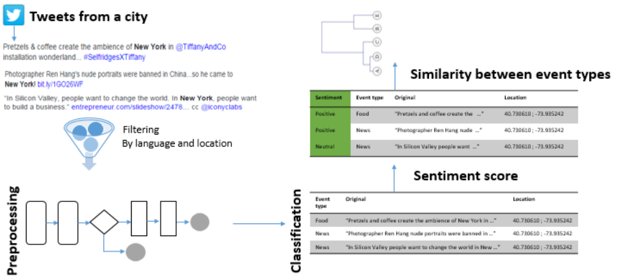


Figure 1. High-level overview of Data Stream Processing system

Our focus is on processing the unstructured data from social sensors like Twitter. The data stream processing unit receives the incoming data and applies the preprocessing which includes: filtering the data by location and language and convert the tweets to the uniform format (where all the characters are translated into letters). The next step is event type detection, where events are processed by a set of rules. Rules can be hand written rules, machine learning algorithms like classification, or sequence models like named entity recognition. We choose the approach of classification techniques, and present the initial results in [18]. Events are saved in a repository to keep a record of original observations for following purposes. The next step is sentiment identification and similarity function between event types. The final output presents the similarity between event types and groups them by the same sentiment level. By identifying the sentiment and similarity relationship between event types, the meaningful relations are highlighted so the decision makers (automatic or humans) and the services related to them can be assigned, see Fig. 2.



Pretzels & coffee create the ambience of New York in @TiffanyAndCo installation wonderland... #SelfridgesXTiffany

Photographer Ren Hang's nude portraits were banned in China...so he came to New York! bit.ly/1GO26WF

"In Silicon Valley, people want to change the world. In New York, people want to build a business." entrepreneur.com/slideshow/2478... cc @iconyclabs

Event type	Original	Location
Food	"Pretzels and coffee create the ambience of New York in ..."	40.730610 ; -73.935242
News	"Photographer Ren Hang nude portraits were banned in ..."	40.730610 ; -73.935242
News	"In Silicon Valley people want to change the world in New ..."	40.730610 ; -73.935242

Sentiment	Event type	Original	Location
Positive	Food	"Pretzels and coffee create the ..."	40.730610 ; -73.935242
Positive	News	"Photographer Ren Hang nude ..."	40.730610 ; -73.935242
Neutral	News	"In Silicon Valley people want ..."	40.730610 ; -73.935242

Figure 2. Functional flow diagram

Using this framework, we are trying to find an answer for the following research questions:

- How to extract the knowledge from data collected from social networks?
- How to identify which services to be deployed based on event type detection, sentiment score, and similarity?

We are mainly focused on automatically categorize the tweets into different event type categories, and enrich the analysis of sentiment identification and similarity function between event types to provide better local services.

For our analysis, we consider events in social sensors as a real-world happening that discuss the associated topic at a specific place and time." Events are characterized by the location, time and organization or person. Each event belongs to event type category or the certain topic. For instance, "*I am at Neil Simon Theatre for Gigi NY in New York*" belongs to the category 'art.'

We use text stream T = (T_1, …, T_n) where T_i is a tweet. Each tweet consists of a set of features (F_1, …, F_k) at location L. The problem of automatic event detection is a problem to identify the facts from a text stream T with the similar set of features at location L, using rules. In our case using supervised machine learning methods. Based on that each T_i belongs to a topic class C = (C_1, ..., C_l), defined as a pair of components T_i.-> C_j. The semantic orientation of the subject states whether the topic is positive, negative, or neutral. Set of sentiment S = (S_1, ..., S_m) are assigned to each pair T_i.-> $C_{j,}$

$$(T_i\text{.-> }C_j)\text{ -> }S_l.$$

Fig.2 shows the functional flow. The output of the functional flow is a knowledge useful for the city services improvement.

4. Experiment setup for analysis

A. Collecting the data

For analysis purposes, Twitter data is collected using its Application Programming Interface. It is comparatively simple to capture comprehensive data sets of a clear 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. 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 consistently collect tweets for each category. For example, we obtain the tweets, from NYC we used the following coordinates:

Latitude: 40.730610 and Longitude: -73.935242

For analysis, two data sets are used, and both are from same geo-location coordinates and English language as a filter was used. As shown in Table 1, Dataset 1 is a dataset with tweets containing generic terms along with geo-location and English language as filters. Whereas, the Dataset 2 is composed by tweets that refer to named entities. In this case, named entities used for categorization of tweets are sixteen, namely, art, music, film, books, health, sport, food, travel, holidays, tech, weather, religion, news, fashion, shopping, celebrities. They represent event types based on the statistics about most frequently topics posted on social media.

TABLE I. NUMBER OF TWEETS USED FOR ANALYSIS

Training data		**Testing data**	
Dataset 1	Dataset 2	Dataset 1	Dataset 2
1303	1746	1702	8828

B. 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, and so forth. The text messages with these symbols, images may be easier for humans to read and analyze. When the text data is mixed with other types of symbols and pictures, 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 the sentimental analysis. The typical characteristics of Twitter data that makes it a challenging are: messages are short and contain less text, the message may contain different language text, it contains special symbols with specific meaning, data contains many shortcuts, and 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 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 Context-aware pre-processing algorithm

```
procedure PRE-PROCESSING OF TWEETS
    for each tweet $t_i \in T$ do
        Remove URLs, re-tweets, hashtags, repeated punctuation’s
        if length($t_i$) > 5 do
        for each word $w_j \in t_i$ do
            miscellaneous symbols
            emotion icons, contractions
            abbreviations, acronyms, smilies
            misspelling words
        the end for replacing it with full, meaningful words
        Remove stop words, punctuation’s, non-English words
        Convert to lower case characters
        end if
    end for
end procedure
```

Following is an example of how the data looks like after some preprocessing steps.

Original: “I'm at Neil Simon Theatre - @nederlanderbway for Gigi (NY) in New York, NY https://t.co/WIGeWlYggy 676 taaaatoooo :)))))))))) aka ILY after #nelisimontheatre”

After removing retweets, URL, hashtags, repeated punctuation: “I am at Neil Simon Theatre for Gigi NY in New York NY 676 tato :) aka ILY after.”

After conversion of smiley symbols, acronyms, abbreviation, contractors, emotion icons: “I am at Neil Simon Theatre for Gigi NY in New York NY 676 tato Smile also known as I love you after.”

The final output, after removing stop words, numbers, punctuation characters: “Neil Simon Theatre Gigi NY New York NY tato Smile known love.”

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 functionality to convert smileys [2], emoticons [3], acronyms and abbreviations [4][5][6], contractions [7] and misspelled words [8] to full, meaningful

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

[2] http://www.netlingo.com/smileys.php

[3] http://en.wikipedia.org/wiki/List of emoticons

[4] http://marketing.wtwhmedia.com/30-must-know-twitterabbreviations-and-acronyms/

[5] https://digiphile.wordpress.com/2009/06/11/top-50-twitteracronyms-abbreviations-and-initialisms

[6] http://www.muller-godschalk.com/acronyms.html

[7] http://www.sjsu.edu/writingcenter/docs/Contractions.pdf

words. Table 2 shows the number of conversion inputs used in each category. Tweets are processed by removing repetitions, special characters, stop words, and English stop words[9].

TABLE II. DICTIONARY LISTS USED FOR PRE-PROCESSING OF TWEETS

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

Even though collected tweets are in the English language, there were words in other languages, in such cases, tweets are ignored for analysis. 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 necessary for analysis purposes. 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

TABLE III. DATA STATISTICS BEFORE PRE-PROCESSING

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

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 many punctuation marks, stop words, numbers, and non-English words that would not convey any information about the context, and are not used for any analysis. This noisy data becomes a big challenge in pre-processing of Twitter data for analysis purposes.

5. CATEGORIZATION OF TWITTER MESSAGES

A. *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. 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 regarding a feature vector. In most of the NLP applications, BOW’s and TF−IDF features are frequently used for text processing applications, sentimental analysis on Twitter data, blogs and classification of sentiments from micro-blogs [2], [7], and [26].

Even though these functions are extensively used for most of the text processing applications, for completeness purpose, a brief explanation is included as follows:

1) *Bag-of-Words (BOWs)*: This model represents text as an unordered 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 ‘tn’ is a vector of weights

‘W_{1n}, ..., W_{wn}’

Where ‘Win’ represent the frequency of the i^{th} term in the nth text message. The transformation of a text message ‘T’ into the BOWs format allows the input stream to be consumed as a matrix, where rows represent Twitter text message vectors, and columns are terms in each Twitter text message [21].

2) *Term Frequency and Inverse Document Frequency (TF-IDF)*: It is a feature vector representation method where shared and rare terms in the text messages are normalized so that rare terms are more emphasized along with successive 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 ‘tm’, while document frequency DF (t_i, T) is the number of Twitter text messages contains the term ‘t_i. Term frequency is over emphasize the terms that appear more often, but that care little information about the content of the Twitter text message. If a term appears very often across all the Twitter text messages, it means it does not 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:

$$TF\text{−}IDF\ (t_i,t_m,\ T\ ) = TF\ (t_i,t_m) \times IDF\ (t_m,\ T) \qquad (1)$$

$$IDF\ (t^i,\ T) = log\left(\frac{T}{1+\ |tm\ \in T:\ tj\ \in tm|}\right) \qquad (2)$$

[8] https://en.wikipedia.org/wiki/Wikipedia:Lists of common misspellings

[9] http://xpo6.com/list-of-english-stop-words/

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 will become zero. It is important to mention that to avoid dividing by zero the smoothing is applied.

B. Categorization of Tweets

Classification of online stream tweets helps to find valuable information up to date for each type of category. We choose machine learning approach because it was successfully applied to several works and achieved great results for classification problems. Tweets are analyzed and classified into predefined categories using supervised learning techniques. For the experiment, we choose two of the most used algorithms [13]: Naive Bayes (NB) and Support Vector Machines (SVM), as well as Random Forest (RF) because it can handle a noisy data well, and compare the performance results. NB classifier is a probabilistic classifier, SVM is a discriminative classifier, and RF classifier is an ensemble method where more than one decision tree is used for classification purposes based on voting rule [8].

This approach relies on using a collection of data to train the classifiers. Initially, models are trained on training dataset as tabulated in Table 1, and these trained models are used to classify the test dataset automatically. For an illustration of data, Fig. 3 shows the word cloud of 'Food' and 'Sports' category of tweets. As we can observe from Fig. 3, in both word clouds, the most dominant words are highlighted. The most dominant word in each of the categories are FOOD and SPORT those are exactly same as category labels. It is also worth noting that, in both the clouds, there are dominant words that are not related to the category of the tweets like JUST in Food cloud, or NEW in Sports cloud.

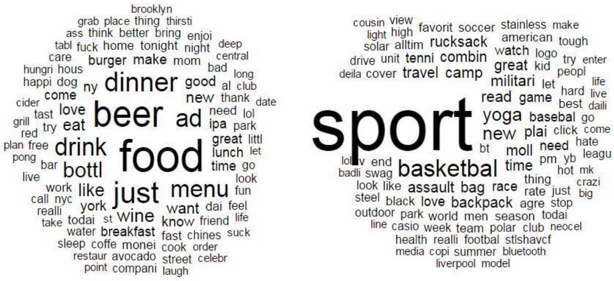


Figure 3. Word cloud illustration of tweets belong to Sport and Food categories

6. Sentiment Analysis

Sentiment analysis on already determined classes of relevant information from online stream tweets helps in determining public opinions. There are many techniques for Sentiment Analysis (SA), some of them are, Machine learning (supervised and unsupervised learning) and Lexicon based (Dictionary, Corpus-based). Also, it depends on the level of doing it: document, sentence or phrase level, aspect or feature level, and word level [1].

For our experiment, we choose a library Stanford CoreNLP [10] that builds up a representation of the whole sentence based on the sentence structure, the order of words is considerate. Also, this library supports five level of sentiment: Strong Negative, Negative, Neutral, Positive, and Strong Positive. We applied SA after categorization step, and now we have a more detailed view in which sentiment context people are talking, positively or negatively about the trending topics. For instance, topic category *Weather* "Rain perfect weather stay read Zenzoris Returns," sentiment *positive.* For illustration Figures 4 and five show the line graph of topic categories with their sentiment measured level.

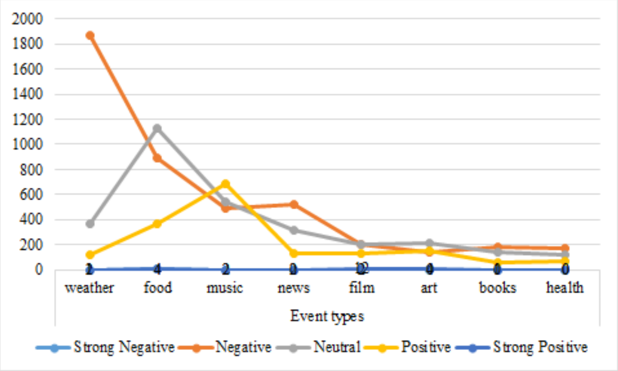


Figure 4. Sentiment measure for the first six categories with the higher number of tweets

Fig. 4 shows that people were mostly talking about the weather but in a negative context, while when they speak of music was to a great positive and few strong positive contexts. Moreover, the opinions of books and health are almost the same.

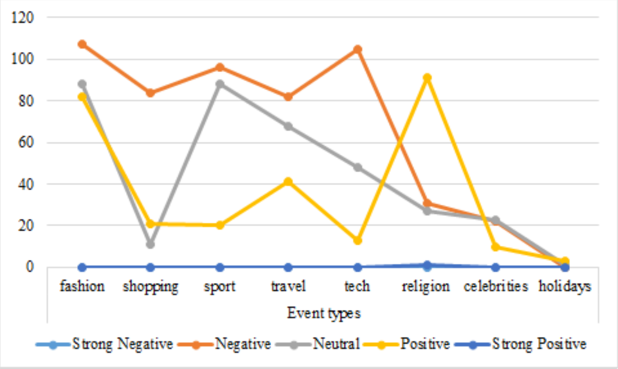


Figure 5. Sentiment measure for the six categories with the lowest number of tweets

[10] http://nlp.stanford.edu/sentiment/code.html

Fig. 5 shows that people were talking about fashion and religion in important positive context, while opinions for sport are mostly negative and neutral and are very close.

A sentiment analysis, in this case, is beneficial, it adds a new value in measuring public opinion as well as know how to best harness the potential benefits of public services. For instance, if an event in central park is detected and the sentiment is negative or neutral, then the services related to navigation for runners or walkers will reroute the paths. Collaborative and personal recommendation services can be activated depending on their settings. The recommender systems, in this case, will adjust their algorithms to include sentiment analysis, and weight differently services that receive a lot of negative feedback or fewer instances. However, the importance and sensitivity of the topic (emergency, earthquake) are highly relevant, in this case, the frequency of the tweets for the negative context can be lower. In the case of real-time processing, as topics and sentiments are changing, service recommendation is changing adequately, too.

7. RESULTS AND EVALUATION

A. *Data used for analysis*

For experimental analysis, we used the Twitter social network as a data source and training data sets are created. The ground-truth for training and testing datasets are created manually. For experiment evaluation, we split the data on testing 80% and training 20%. The type of tweets and number of categories of tweets used in this study are shown in Table 4. The collection of Twitter data tweets and pre-processing 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.

TABLE IV. 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

For the analysis, we used open source programming language Python, and its machine learning package scikit-learn [14] along with natural language processing took NLTK [20]. *BOWs* and *TF−IDF* features are extracted using NLTK tool and NB, SVM, and RF classifiers are used from a scikit-learn package. The classification accuracy reported in this paper is calculated as:

$$\text{Accuracy} = \frac{\text{TP+ TN}}{\text{TP+TN+FP+FN}} \qquad (3)$$

B. *Experimental results*

Initially, machine learning models are trained individually for sixteen classes. During testing, the trained models are used to categorize the testing tweets automatically. Results are calculated based on ground-truth marked for testing examples.

TABLE V. CLASSIFICATION ACCURACY OF TWEETS INTO PREDEFINED CATEGORIES

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

DS 1 = Dataset 1 and DS 2 = Dataset 2

Table 5 shows the classification accuracy on test dataset, and from a table, we can see that RF classifier gives almost 94% accuracy as compared to SVM and NB classifiers. The overall accuracy on both data sets are nearly 90% accurate, and this accuracy has come down because FOOD class examples are misclassified; it is almost 50% accurate. As a result, the overall accuracy is reduced.

C. *Discussions*

High-level topics could be useful for a variety of upstream tasks such as summarization. Visualization diagrams are chosen as representation form for the decision making. The results show that the most frequently used topics are about 'Food' and 'Music' and less relevant topics are 'Celebrities' and 'Shopping.' To further clarify the results on Dataset 1, Fig. 4 shows the class wise accuracy for sixteen categories. From this figure, we can notice that 'Food' class has the lowest accuracy compared to other classes that result in a decrease of the overall accuracy of the dataset. One solution to reduce miss-classification, in this case, is that building hierarchical classification models so that miss-classified 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.

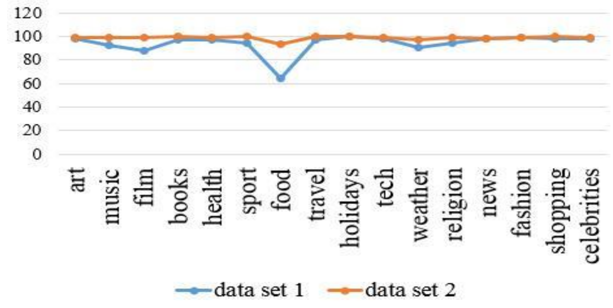


Figure 6. Accuracy by category for both the data sets using BOWs feature with SVM classifier

Furthermore, the accuracy level in important when measuring the sentiment by category because if it is lower than certain acceptable level and we get a high negative output, like in our case for the food we have high neutral and adverse levels, this affects the service recommendations and we can recommend something wrong. However, adding the sentiment detection layer on top of categorization is a step forward for measuring public opinion and making better and appropriate recommendations that satisfy public needs and concerns.

We also gauge the similarity between generated categories to find which of them are more similar. We use it to determine the sentiment with similarity index. We used cosine similarity metric between categories to measure how similar they are and Fig. 7 gives dendrogram visualization output.

This figure shows that some categories are more similar than the rest, like ‘music’ with ‘art,' this is a base for service compositionality that can be used for recommendation. Adding the sentiment measurements gives additional dimensionality to service composition, for instance, we can observe that ‘weather’ and ‘sport’ have high similarity, and they have the same sentiment score (negative), so the decision is to focus on advancing the services related to indoor sports. On the other side, ‘shopping’ and ‘health’ are similar but have different sentiment, while shopping has a positive score, health has negative, so maybe currently there is flu and because of that shopping is increased. Then maybe services related to health advice will be useful to be recommended. How the recommended services will be distributed will depend on of the event types, for instance for the events related to traffic, or safety then maybe some government resources can be allocated. Alternatively, when users want to travel to the particular area, they can have sentiments toward public subjects.

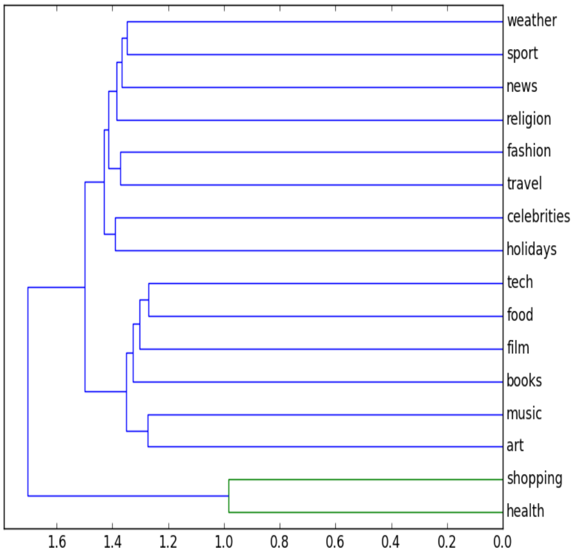


Figure 7. Similarity between categories

8. 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 broad range of predefined identifiers over one concrete geo-location, in this case, New York City. The paper presents a framework that collects only geotagged tweets, taking into consideration the context of the whole crowd, extracts the relevant knowledge from it and use that knowledge for recommending the services.

The experiments showed that the context-aware pre-processing algorithm used to process the tweets helps to categorize the tweets into predefined categories efficiently. It is shown that RF Classifier combined with TF-IDF feature gives better results compared to SVM and NB classifiers. Moreover, sentiment analysis measures provide additional information layer for determining public opinion.
In this paper, we have presented:

- an integrated framework for detecting real-world event types reported on Twitter
- efficient pre-processing algorithm where every word is important for analysis
- the supervised event identification was performed in several stages: data collection, preprocessing, feature selection, classification

- our experiments suggest that RF classifier combined with TF-IDF yields better performance than many leading classifiers
- moreover, sentiment analysis provides more detailed information for previously detected categories in case of service recommendations based on social sensor data

The framework presented here can be easily incorporated into already established service like traffic route recommendations, walking tours around the city, or healthcare application. In other words, it can be used in event management, intelligence gathering, and decision-making.

In future, we want to upgrade this tool with the functionality for monitoring the topic over time, dynamic topic correlation, and based on that identify the right services. Also, we want to work on a dynamic update of events without any predetermined mention of the event. Moreover, the dearth of techniques coupling textual, spatial, and temporal along with social/network structure.

ACKNOWLEDGMENT

This work was financially supported by the University of Grenoble Alpes, France and National Institute of Standards and Technologies (NIST), Gaithersburg, USA and conducted within a collaboration under Information Technology Laboratory, Advanced Network Technologies Division (ANTD).

REFERENCES

[1] Ahmed, K. B., Radenski, A., Bouhorma, M., & Ahmed, M. B. (2016, January). Sentiment Analysis for Smart Cities: State of the Art and Opportunities. In *Proceedings of the International Conference on Internet Computing (ICOMP)* (p. 55). The Steering Committee of the World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).

[2] 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.

[3] Albakour, M. D., Macdonald, C., Ounis, I., Pnevmatikakis, A., & Soldatos, J. (2012, August). SMART: An open source framework for searching the physical world. In *SIGIR 2012 Workshop on Open Source Information Retrieval* (pp. 48-51).

[4] 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.

[5] Atefeh, F., & Khreich, W. (2015). A survey of techniques for event detection on Twitter. *Computational Intelligence*, *31*(1), 132-164.

[6] 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.

[7] 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.

[8] Breiman, L. Random forests. *Machine learning* 2001;45(1):5–32.

[9] 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.

[10] Bruns, A., Stieglitz, S. Quantitative approaches to comparing communication patterns on Twitter. *Journal of Technology in Human Services* 2012;30(3-4):160–185.

[11] Mandel, B., Culotta, A., Boulahanis, J., Stark, D., Lewis, B., & Rodrigue, J. (2012, June). A demographic analysis of online sentiment during hurricane irene. In *Proceedings of the Second Workshop on Language in Social Media* (pp. 27-36). Association for Computational Linguistics.

[12] Ilina, E., Hauff, C., Celik, I., Abel, F., & Houben, G. J. (2012, July). Social event detection on Twitter. In *International Conference on Web Engineering* (pp. 169-176). Springer Berlin Heidelberg.

[13] Injadat, M., Salo, F., & Nassif, A. B. (2016). Data mining techniques in social media: A survey. *Neurocomputing*, *214*, 654-670.

[14] 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.

[15] Jiang, L., Yu, M., Zhou, M., Liu, X., & Zhao, T. (2011, June). Target-dependent Twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1* (pp. 151-160). Association for Computational Linguistics.

[16] Kouloumpis, E, Wilson, T., Moore, J. Twitter sentiment analysis: The good the bad and the omg! *Icwsm* 2011; 11:538–541.

[17] Kotevska, O., Lbath, A., Bouzefrane, S. Toward a real-time framework in cloudlet-based architecture. *Tsinghua Science and Technology* 2016;21(1):80–88.

[18] Kotevska, Olivera, Sarala Padi, and Ahmed Lbath. "Automatic Categorization of Social Sensor Data." *Procedia Computer Science* 98 (2016): 596-603.

[19] Cheong, M., & Lee, V. C. (2011). A microblogging-based approach to terrorism informatics: Exploration and chronicling civilian sentiment and response to terrorism events via Twitter. *Information Systems Frontiers*, *13*(1), 45-59.

[20] 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.

[21] Pak, A., Paroubek, P. Twitter as a corpus for sentiment analysis and opinion mining. In: *LREC*; vol. 10. 2010, p. 1320–1326.

[22]

[23] Rosa, K.D, Shah, R., Lin, B., Gershman, A., Frederking, R. Topical clustering of tweets. *Proceedings of the ACM SIGIR: SWSM* 2011.

[24] Radovanovic M., Ivanovic, M. Text Mining: Approaches and applications. *Novi Sad J Math* 2008;**38**(3):227–234.

[25]

[26] Romsaiyud, W. (2013). Detecting emergency events and geo-location awareness from Twitter streams. In *The International Conference on E-Technologies and Business on the Web (EBW2013)* (pp. 22-27). The Society of Digital Information and Wireless Communication.

[27] 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.

[28] Saif, H., He, Y., Alani, H. Semantic sentiment analysis of twitteIn: *The Semantic Web–ISWC 2012*. Springer; 2012, p. 508–524.

[29] Salas-Zárate, M. D. P., Medina-Moreira, J., Álvarez-Sagubay, P. J., Lagos-Ortiz, K., Paredes-Valverde, M. A., & Valencia-García, R. (2016). Sentiment Analysis and Trend Detection in Twitter. In *Technologies and Innovation: Second International Conference, CITI 2016, Guayaquil, Ecuador, November 23-25, 2016, Proceedings* (pp. 63-76). Springer International Publishing.

[30] Schwartz, R., Naaman, M., & Matni, Z. (2013, June). Making sense of cities using social media: Requirements for hyper-local data aggregation tools. In *Proceedings of the International AAAI Conference on Weblogs and Social Media* (pp. 15-22).

[31] 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.

[32] Zhao, J., Wang, X., & Ma, Z. (2014). Towards events detection from microblog messages. *International Journal of Hybrid Information Technology*, *7*(1), 201-210.

Olivera Kotevska is a Ph.D. student at University of Grenoble Alpes (MRIM/LIG laboratory), France. She is currently a Guest Researcher at ITL Lab at National Institute of Standards and Technologies (NIST), Washington DC Metro, USA. She received a bachelor of academic informatics from the Ss. Cyril and Methodius University, Skopje, Macedonia, and M.Sc. degree in Intelligent Information Systems from the computer science department of the same university. Her research is in the field of data mining, event processing and intelligent information system design.

Ahmed Lbath is a full professor of computer science at University of Grenoble Alpes (MRIM/LIG Laboratory), France and also conducting research in collaboration with ITL Lab NIST in Washington DC metro area where he carried out research activities as visiting professor. He is the IUT Deputy Director, former Head of CNS Department, and former Director of R&D in a French company. He received his Ph.D. degree in computer science from the University of Lyon and held an “Habilitation Diriger des Recherches.” He is currently a project manager coordinating scientific collaborations in the domain of Cyber-Physical Systems and Smart Cities. His research interests include smart cities, cyber-physical-human systems, mobile cloud computing, web services, GIS, recommendation systems, and software design. He published several patents, book chapters, papers in journals, and conferences and he regularly serves as co-chair and member of several committees of International conferences, journals, and research programs.