An Analysis of Social Data Credibility for Services Systems in Smart Cities – Credibility Assessment and Classification of Tweets

Iman Abu Hashish(✉), Gianmario Motta, Tianyi Ma, and Kaixu Liu

Department of Electronics, Computer Science and Electrical Engineering, University of Pavia, Pavia, Italy
{imanhishamjami.abuhashish01,tianyi.ma01, kaixu.liu01}@universitadipavia.it, motta05@unipv.it

Abstract. In the "Information Age", Smart Cities rely on a wide range of different data sources. Among them, social networks can play a big role, if information veracity is assessed. Veracity assessment has been, and is, a rather popular research field. Specifically, our work investigates the credibility of data from Twitter, an online social network and a news media, by considering not only credibility, and type, but also origin. Our analysis proceeds in four phases: Features Extraction, Features Analysis, Features Selection, and Classification. Finally, we classify whether a Tweet is credible or incredible, is rumor or spam, is generated by a human or a Bot. We use Social Media Mining and Machine Learning techniques. Our analysis reaches an overall accuracy higher than the benchmark, and it adds the origin dimension to the credibility analysis method.

Keywords: Smart cities · Smart citizens · Social data · Twitter · Twitter bot · Credibility · Veracity · Classification · Social media mining · Machine learning

1 Introduction

Smart cities rely on a wider and wider range of Internet information, which includes sensor data, public data, and human generated data, as social networks and crowdsourced data [1]. In human generated data, relevance and credibility need to be addressed since in social networks, feeds can be propagated without being controlled nor organized.

Our research addresses credibility evaluation techniques for smart mobility support systems. The targeted online social media is Twitter, a widely popular social media as well as a news medium. It enables its users to send and read short messages named Tweets. It is a platform for live conversations, live connections and live commentary. It is accessed daily by 313 million active users with 1 billion of unique visits monthly to websites with embedded Tweets [2]. Twitter users express their opinions, share their thoughts, celebrate religious events, discuss political issues, create news about ongoing events, and provide real time updates about ongoing natural disasters, etc. In addition, Twitter is a rich source for social data because of its inherent openness to public

A. Longo et al. (Eds.): IISSC 2017/CN4IoT 2017, LNICST 189, pp. 119–130, 2018.
https://doi.org/10.1007/978-3-319-67636-4_14

consumption, clean and well-documented API, rich developer tooling and broad appeal to users [3].

We approach the issue of credibility when, in the larger project called IRMA (Integrated Real-Time Mobility Assistant), we started to consider feeds coming from social networks as information sources for mobility information systems. That issue implied credibility assessment, and on another side Big Data technologies given the huge number of feeds in social networks [4].

In section two, we compare the previous implementations. In section three, we illustrate the methodology, and we continue in section four with a comprehensive explanation of our implementation. In section five, we discuss our results, and section six sketches conclusion and future work.

2 State of the Art

To obtain a comprehensive assessment of State of the Art, we used the paradigm of systematic literature review [5]. (See Table 1).

Table 1. Related works

Author/Reference	Approach
Gupta et al. [6]	Used features related to Tweets, users and events to develop an automatic approach for credibility assessment, enhanced by an event graph-based optimization
Gupta et al. [7]	Developed "TweetCred" a real-time, web-based system to assess credibility of Tweets based on 45 features using Machine Learning, specifically, they proposed a semi-supervised ranking model
Skidar et al. [8]	Provided a comprehensive explanation of a better mechanism to extract credible from noisy data and argued the absence of a standard definition of credibility for making such studies more useful for the research community
Namihira et al. [9]	Proposed a method for assessing the credibility of a Tweet automatically based on topic and opinion classification using Latent Dirichlet Allocation and the analysis of semantic orientation dictionary named Takamura
Batool et al. [10]	Proposed a methodology for precise extraction of valuable information from Tweets to facilitate the extraction of keywords, entities, synonyms, and parts of speech from Tweets which are used after for classification

Most of the related works can be divided into two categories: Classification-based analysis as [11–14] adopting supervised classification, [15] or unsupervised classification or a hybrid of the both [16], and Pattern-based analysis as [17].

3 Methodology

Here below we illustrate the steps of our methodology, which includes (Sect. 3.1) Problem Definition and (Sect. 3.2) Proposed Algorithm.

3.1 Problem Definition

The definition of credibility in this work combines different perspectives, namely (Sect. 3.1.1) Users' Perception, (Sect. 3.1.2) Tweets' Content, and (Sect. 3.1.3) Tweets' Origin.

3.1.1 Credibility Based on Users' Perception

Users' perception in defining a credible Tweet varies. Some users trust what is shared on Twitter and start propagating, while other users question what they read, based on the apparent features of a user's profile. With respect to a comprehensive study, users assess credibility from content, creator, and other available features provided by user interface, while neglecting implicit features [18]. Thus, credibility in User's Perception stems from the features of the user interface of Twitter, namely from explicit features.

3.1.2 Credibility Based on Tweets' Perception

A Tweet consists of a user name, text, and possibly a URL, image, and video. Twitter includes four types of API objects, namely Tweets' objects, users' objects, entity objects, and place objects, which are used to extract a set of implicit features corresponding to a single Tweet. In this case, the degree by which a Tweet's content can convey the truthfulness of an event stems from implicit features.

3.1.3 Credibility Based on Tweets' Origin

If we assume that a credible user provides credible information, the origin is a key for credibility. Accordingly, through of a set of implicit features, the user account that originates the Tweets is classified as a Human or a Bot. A human account is a Twitter account whose Tweets are published by an actual user, while a Bot account Tweets are published by a robot. Thus, the truthfulness stems from both explicit and implicit features, which help in identifying the origin of the Tweet.

3.1.4 Spam and Rumor

Spam is unsolicited, unwanted, and malicious content; harmful URLs or simply text-based, mislead, deceive and negatively influence other users on an event. It is directly connected to automated accounts targeting naïve inexperienced users. While Rumor is a widely propagated misinformed content.

3.2 Proposed Algorithm

Our work, which is a part of a wider analysis framework we are working on, includes: Users' score, Tweets network score, and Sentiment score. Our current work focuses on Users' score, calculates features, identifies spam, and detects bots. Accordingly, the related algorithm includes four phases; Features Extraction, Features Analysis, Features Selection and Classification (Features Classification includes credibility and type/origin classification). (See Figs. 1 and 2).

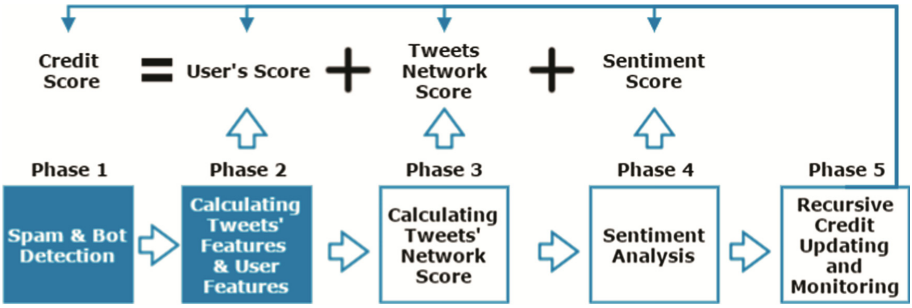

Fig. 1. Overall process

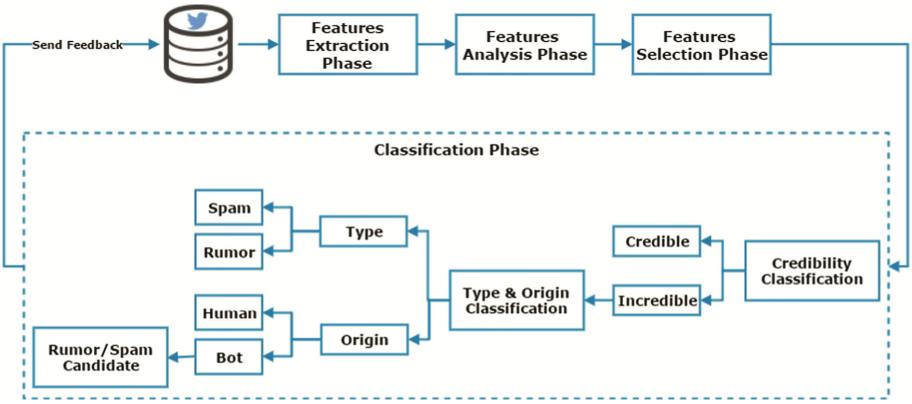

Fig. 2. Details of Phases 1 and 2

3.2.1 Features Extraction

The algorithm starts with a dataset of collected Tweets. When Twitter API is accessed, a target Tweet can be extracted with explicit and implicit features. This phase collects metadata of Tweets to provide contextual information (see Table 2). In Twitter API Documentation [19], the features to be extracted are divided into three sections: Users, Tweets, and Entities. Each is called an object, and each object is composed of a set of fields. Accordingly, those fields are interpreted as features or used for further deriving additional features.

Table 2. Features extraction phase output

Type	Name
From user	created_at, description, id_str, location, default_profile, default_profile_image, favourites_count, following, followers_count, friends_count, geo_enabled, listed_count, protected, screen_name, statuses_count and verified
From tweet	created_at, favorite_count, id_str, in_reply_to_screen_name, lang, possibly_sensitive, retweet_count, retweeted, retweeted_status and truncated
From entity	hashtags, URLs, media and user_mentions

3.2.2 Features Analysis

This phase analyzes the features extracted from the previous phase by considering the nature of Twitter environment as suggested in [20], namely information flow and content propagation, as well as possible interactions such as replying, retweeting, etc. Thus, another set of features can be derived and quantified to address the previously mentioned aspects (see Table 3).

Table 3. Features analysis phase output

Feature name	Description
Friends to followers ratio	Indicates the ratio of the number of people a user is following to the number of people following that user
Followers to friends ratio	Indicates the ratio of the number of people following a user to the number of people followed by the same user
Account reputation	Indicates an estimate of how popular a user account is
Users retweet ratio	Indicates the ratio of the number of retweets propagated by the user to the number of tweets originally published by the same user
External URL ratio	Indicates the ratio of the number of external URLs contained in a Tweet
Value of a retweet	Gives a value based on the deviation of a user's retweet ratio from the average retweet ratio in a target topic

3.2.3 Features Selection

This phase mines the set of features extracted and analyzed to select the features that affect the final judgment on Tweets credibility. Accordingly, a learner-based feature selection technique is applied. It is based on the use of learning algorithms, and the evaluation of the performance on the dataset with different subsets. Accordingly, the subset of features that will achieve the best results will be selected.

3.2.4 Classification

The classification process starts by classifying Tweets in terms of credibility, type as spam and rumor, and finally origin like human accounts and Twitter harmful bots. This phase is divided into two sub-phases, Credibility Classification classifies Tweets into credible or incredible depending on the features obtained. Type and Origin Classification

analyzes the incredible Tweets furthermore by classifying them as rumor or spam, and as humans or bots account.

4 Implementation

The first two phases, Features Extraction and Features Analysis, were implemented using social media mining techniques while the final two phases, Features Selection and Classification, were implemented by using machine learning algorithms explained as follows. (See Fig. 3).

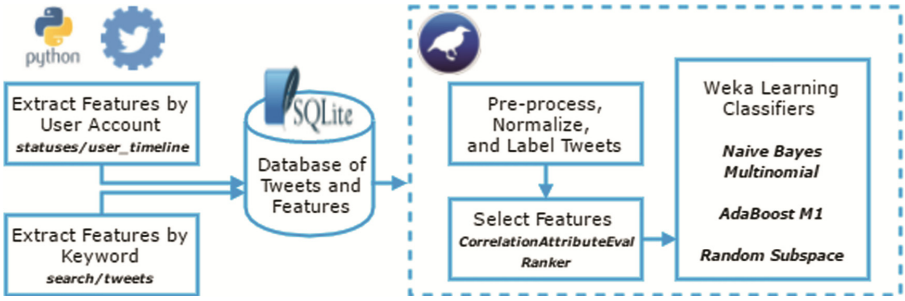

Fig. 3. Overall implementation of the proposed method

4.1 Social Media Mining Techniques

Social data, or human generated data, are big, unstructured and noisy with abundant social relations such as friendships, followers, following, etc. Consequently, using Social Media Mining enables combining social theories with statistical and data mining methods for extracting useful and meaningful data. In our implementation, we used Python and Twitter API.

4.1.1 Features Extraction by User Account

Using this approach, and by providing a list of target user handles, the Tweets of a specific account can be extracted. This approach analyzes the behavior of the target user in terms of posting behavior and Tweets propagated. Twitter proposes a public API, named GET status/user_timeline, which returns a collection of the most recent Tweets posted by the user, indicated either by user_id or screen_name and in our case, screen_name. This API can only return up to 3,200 of users' most recent Tweets, (retweets included).

4.1.2 Features Extraction by Keyword

The second approach is extracting Tweets by a keyword. The keyword is represented by the hashtag. Based on Twitter support [21] a hashtag is used to categorize Tweets by keywords. This approach is implemented by using GET search/tweets API.

4.2 Machine Learning Algorithms

For the last two phases, Features Selection and Classification, machine learning algorithms are exploited using Weka, a data mining software. Features Selection was implemented by applying correlation attribute evaluator, that evaluates the worth of an attribute by measuring Pearson's Correlation between the attribute and the class, using the Ranker in conjunction with the evaluator to rank the features by their individual evaluations, and in our case, their importance in credibility assessment. For Classification, we used several classifiers provided by Weka.

4.3 Twitter Bot Development

As a final step, we created a robot account to enrich the dataset with diverse contents and more robotic behaviors. The developed bot searches Twitter API by using a keyword, once the results are found, the bot retweets them, favorites them, follows their creator and adds the user accounts to a list, thus, reflecting a typical robotic behavior.

5 Evaluation

5.1 Dataset Creation

For the evaluation process, we chose USA 2016 Presidential Elections, then we chose several Twitter accounts covering the elections, along with few related hashtags to create a dataset with a diverse content. We also considered the bot that we have developed. Then, the extracted Tweets were manually labeled in terms of credibility, type and origin. A dataset, of around 2000 Tweets, was created, pre-processed, normalized, and labeled.

5.2 Experiment Setup and Task Preparation

To test performance and efficiency of the proposed algorithm, we performed three tasks: (1) classifying based on credibility, (2) classifying based on type and (3) classifying based on origin, including the following steps:

- Loading the dataset to Weka and assigning the target label as a class.
- Applying Features Selection phase by using correlation attribute evaluator as a subset evaluator and Ranker as a search method.
- Performing the Classification Phase by using the desired classifier.

5.3 Results and Discussion

5.3.1 Credibility Classification

The feature with the highest impact on credibility classification is *from_user_default_profile* which indicates that users have not altered the theme or background of their profiles. When a user first creates an account on Twitter, the default settings are set with an egg picture as an avatar, that is related to *from_user_default_profile_image*. Going on, *from_user_verified*

indicates whether the user account is verified or not. Obviously, a verified account is taken for granted as an official account to propagate credible feeds regarding a specific topic. Other selected features show that, when the user account has a profound network that interacts with what the user propagates, in terms of retweeting or following the user, the level of credibility and trust are higher, which explains the other features that were selected (see Table 4).

Table 4. Features selected for credibility classification

Rank	Feature
1	from_user_default_profile
2	from_user_verified
3	from_user_follower_count
4	from_user_listed_count
5	followers_to_friends
6	user_retweet_ratio
7	from_user_retweet_count
8	from_user_default_profile_image
9	friends_to_followers
10	value_retweet

5.3.2 Type Classification

We modified the original dataset by keeping the Tweets classified as incredible, labeling them as spam, and substituting the credible classified Tweets with a set that was labeled as rumors, while maintaining the total number of instances. However, before going on with the Features Selection, we applied N-Gram Features to consider the text of Tweets. Based on the average weights of the first 10 ranked features, *user_retweet_ratio* is the first attribute that contributes to the final classification, *from_user_retweet_count*, *retweeted_status* and *TXT_rt* behave in the same way. Other network related features were selected, namely *friends_to_followers* and *account_reputation* which also convey

Table 5. Features selected for type classification

Rank	Feature
1	user_retweet_ratio
2	from_user_retweet_count
3	from_user_default_profile_image
4	friends_to_followers
5	from_user_location
6	account_reputation
7	retweeted_status
8	TXT_rt
9	entities_mentions
10	TXT_#election2016

a robotic behavior by following many accounts, leading to a very small ratio between the friends and the followers count. The reason behind this ranking (see Table 5) is that, rumored and spammed Tweets are most likely to be originated by bot accounts. Finally, what distinguishes a rumored Tweet from a spammed one is how fast it gets propagated, thus, the *retweet_count* attribute.

5.3.3 Origin Classification

The features that contribute the most in classifying the origin are almost identical to the features selected in classifying credibility (see Table 6). This proves that the credibility of the Tweet is directly related to its origin.

Table 6. Features selected for origin classification

Rank	Feature
1	from_user_default_profile
2	from_user_verified
3	from_user_followers_count
4	from_user_listed
5	followers_to_friends
6	user_retweet_ratio
7	from_user_default_profile_image
8	from_user_retweet_count
9	friends_to_followers
10	value_retweet

Once the features are selected for each task, the Classification phase directly follows. The results obtained for each task are detailed in Table 7. As can be seen from the detailed accuracy results (see Figs. 4 and 5), Credibility and Type tasks provided higher accuracy measures than our baselines, [20] and [12] respectively. At the best of our knowledge, this is the first classification of Tweets with respect to their origins. Thus, considering only features that can be extracted and analyzed to investigate the robotic behavior, our algorithm looks accurate.

Table 7. Detailed classification results

Criteria	Precision	Recall	F-measure	MCC	ROC area
Credibility	0.902	0.885	0.883	0.783	0.921
Type	0.923	0.918	0.918	0.841	0.984
Origin	0.897	0.879	0.876	0.772	0.926

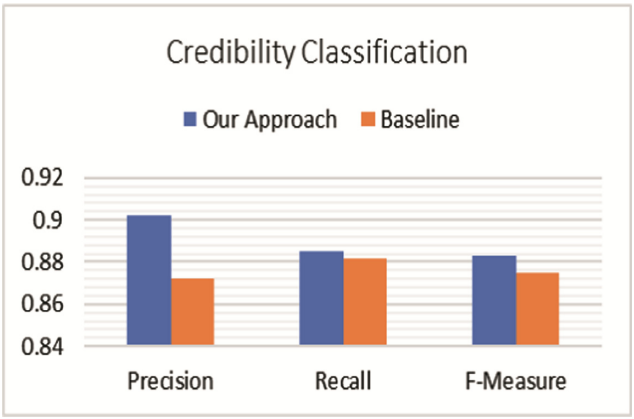

Fig. 4. Credibility task accuracy results vs. baseline

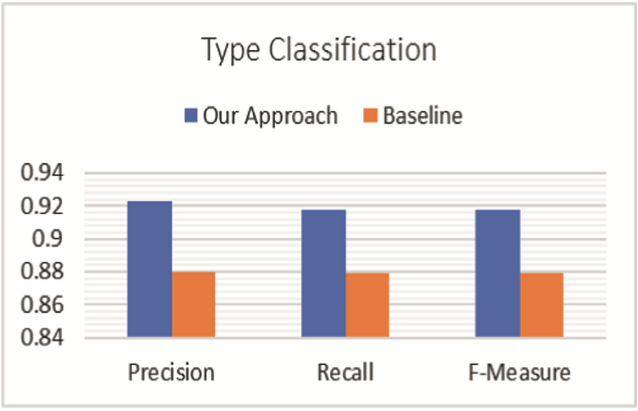

Fig. 5. Type task accuracy results vs. baseline

6 Conclusion and Future Work

Our work intends to provide a profound and comprehensive analysis on social data credibility, and, specifically, on Tweets credibility assessment. The feeds vary very much, they may represent a thought, a mood or an opinion, as well as on-going political news, sports, and natural disasters. Unlike other social networks, Twitter is an open nature that enables everyone to publish thoughts that reach a wide range of people. Because of these elements, credibility is relevant.

Therefore, we propose a comprehensive analysis from three points of view; Tweets' credibility, type and origin. Our analysis, which is implemented using Social Media Mining Techniques and Machine Learning Algorithms with Weka Software, includes four phases;

1. Features Extraction Phase: a set of features, attributes and characteristics of Tweets are extracted;
2. Features Analysis Phase: features are further analyzed and quantified;
3. Features Selection Phase: the list of features that contribute the most to the assessment of Tweets are selected;
4. Classification Phase: Tweets are classified with respect to our viewpoints.

We tested the correctness of our assumptions and the accuracy of our algorithm by conducting an experiment of three tasks that correspond to our three-fold classification phase. To accomplish this, we created a dataset of ~2000 Tweets, each Tweet is associated with 40 features. Our dataset concerned the 2016 USA Presidential Elections.

The accuracy of our algorithm is around ~89% for credibility classification, around ~92% for type classification and around ~88% for origin classification. These results are higher than our baseline for the first two classifications, while the final classification is new at the best of our knowledge.

Of course, our work may be extended by:

- Creating a larger dataset with more diverse content.
- Extending the number of the derived features to investigate the importance in the final assessment of Tweets credibility.
- Deepening the analysis of features from different perspectives such as: the representativeness of the hashtags to the content of Tweets, image processing techniques to explore the media published within Tweets, and Natural Language Processing to process the holistic semantics of Tweets.
- Designing an automatic system that assesses the credibility of Tweets as a browser add-on and a mobile application.

References

1. Motta, G.: Towards the Smart Citizen, New and smart Information Communication Science and Technology to support Sustainable Development (NICST) (2013)
2. Twitter Statistics. https://about.twitter.com/company
3. Russel, M.A.: Mining the Social Web, 2nd edn. O'Reilly, Beijing (2014)
4. Motta, G., Sacco, D., Ma, T., You, L., Liu, K.: Personal mobility service system in urban areas: the IRMA project. In: 2015 IEEE Symposium on Service-Oriented System Engineering (2015)
5. Okoli, C., Schabram, K.: A guide to conducting a systematic literature review of information systems research. SSRN Electron. J.
6. Gupta, M., Zhao, P., Han, J.: Evaluating event credibility on Twitter. In: Proceedings of the 2012 SIAM International Conference on Data Mining, pp. 153–164 (2012)
7. Gupta, A., Kumaraguru, P., Castillo, C., Meier, P.: TweetCred: real-time credibility assessment of content on Twitter. In: Aiello, L.M., McFarland, D. (eds.) SocInfo 2014. LNCS, vol. 8851, pp. 228–243. Springer, Cham (2014). doi:10.1007/978-3-319-13734-6_16
8. Sikdar, S., Kang, B., Odonovan, J., Hollerer, T., Adah, S.: Understanding information credibility on Twitter. In: 2013 International Conference on Social Computing (2013)

9. Namihira, Y., Segawa, N., Ikegami, Y., Kawai, K., Kawabe, T., Tsuruta, S.: High precision credibility analysis of information on Twitter. In: 2013 International Conference on Signal-Image Technology & Internet-Based Systems (2013)
10. Batool, R., Khattak, A.M., Maqbool, J., Lee, S.: Precise tweet classification and sentiment analysis. In: 2013 IEEE/ACIS 12th International Conference on Computer and Information Science (ICIS) (2013)
11. Castillo, C., Mendoza, M., Poblete, B.: Information credibility on Twitter. In: Proceedings of the 20th International Conference on World Wide Web (WWW 2011) (2011)
12. Sahana, V.P., Pias, A.R., Shastri, R., Mandloi, S.: Automatic detection of rumored tweets and finding its origin. In: 2015 International Conference on Computing and Network Communications (CoCoNet) (2015)
13. Zhang, Q., Zhang, S., Dong, J., Xiong, J., Cheng, X.: Automatic detection of rumor on social network. In: Li, J., Ji, H., Zhao, D., Feng, Y. (eds.) NLPCC 2015. LNCS, vol. 9362, pp. 113–122. Springer, Cham (2015). doi:10.1007/978-3-319-25207-0_10
14. Al-Dayil, R.A., Dahshan, M.H.: Detecting social media mobile botnets using user activity correlation and artificial immune system. In: 7th International Conference on Information and Communication Systems (ICICS) (2016)
15. Sivanesh, S., Kavin, K., Hassan, A.A.: Frustrate Twitter from automation: how far a user can be trusted? In: International Conference on Human Computer Interactions (ICHCI) (2013)
16. Gupta, A., Kaushal, R.: Improving spam detection in online social networks. In: International Conference on Cognitive Computing and Information Processing (CCIP) (2015)
17. Wang, S., Terano, T.: Detecting rumor patterns in streaming social media. In: IEEE International Conference on Big Data (Big Data) (2015)
18. Morris, M.R., Counts, S., Roseway, A., Hoff, A., Schwarz, J.: Tweeting is believing? In: Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW 2012) (2012)
19. Twitter API Documentation. https://dev.twitter.com/overview/documentation
20. Kang, B., O'donovan, J., Höllerer, T.: Modeling topic specific credibility on Twitter. In: Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces (IUI 2012) (2012)
21. Twitter Support: Using Hashtags in Twitter. https://support.twitter.com/articles/49309