M. Rao Batchanaboyina* and Nagaraju Devarakonda

Design and Evaluation of Outlier Detection Based on Semantic Condensed Nearest Neighbor

https://doi.org/10.1515/jisys-2018-0476
Received December 6, 2018; previously published online May 1, 2019.

Abstract: Social media contain abundant information about the events or news occurring all over the world. Social media growth has a greater impact on various domains like marketing, e-commerce, health care, e-governance, and politics, etc. Currently, Twitter was developed as one of the social media platforms, and now, it is one of the most popular social media platforms. There are 1 billion user's profiles and millions of active users, who post tweets daily. In this research, buzz detection in social media was carried out by the semantic approach using the condensed nearest neighbor (SACNN). The Twitter and Tom's Hardware data are stored in the UC Irvine Machine Learning Repository, and this dataset is used in this research for outlier detection. The min–max normalization technique is applied to the social media dataset, and additionally, missing values were replaced by the normalized value. The condensed nearest neighbor (CNN) is used for semantic analysis of the database, and based on the optimized value provided by the proposed method, the threshold is calculated. The threshold value is used to classify buzz and non-buzz discussions in the social media database. The result showed that the SACNN achieved 99% of accuracy, and relative error is less than the existing methods.

Keywords: UC Irvine Machine Learning Repository, normalization, outlier detection, semantic approach using condensed nearest neighbor (SACNN), Twitter.

1 Introduction

The number users in social media are growing enormously and also the information shared between them; the streaming of data posted on social media is increasing every day [18]. Hundreds of millions of users are contributing on social media and creating a huge volume of data in various blogs [4, 11]. The Microblog, a short form for mini blog, combines the social and information networks using the term of follower–friend relationship [7]. In Twitter, if user A follows user B, then user A gets the update tweets of user B in the microblog. Now, Twitter is one of the most popular social media tools, where people can send and read messages in the form of tweets. Currently, Twitter has 316 million monthly active users, and 500 million tweets are sent every day [15]. The users have the liberty to tweet about anything from their mind; for instance, tweets that are related to quotes, lyrics, ideas, news stories, etc., can be sent in Twitter. The users can integrate the hashtag (#) in their tweets to show tweets that are related with respective events, news, or something else [6, 10]. The users can tweet their message within 140 characters, and the corresponding followers can instantly get the tweets of the users [16].

There are large amounts of tweets and information available in social media, which, in turn, requires features to recommend important events or news occurring across the globe. In this research, the semantic approach using condensed nearest neighbor (SACNN) is proposed to identify the topic trend in social media.

***Corresponding author: M. Rao Batchanaboyina,** Computer Science and Engineering, Achraya Nagarjuna University, Guntur, A.P.-522510, India, e-mail: mallik.mit@gmail.com. https://orcid.org/0000-0003-4329-040X
Nagaraju Devarakonda: Department of Information Technology, Lakireddy Balireddy College of Engineering, Mylavaram, Krishna (DT), A.P.-521230, India

The topic that is trending in social media is identified using an outlier detection. In simple form, a number of discussions about news, events, or others are used to calculate a trending topic. This framework helps to provide better recommendation news that is relevant to the interest of users, which helps in better subscription and experience of the user. The buzz UCI dataset is used to evaluate the proposed SACNN method, and the performance is derived from its function. This dataset is publicly available and consists of two different social media data, namely: (1) Twitter [14, 17] and (2) Tom's Hardware. Tom's Hardware is a worldwide forum that provides information about new technology with distinct features. The pre-processing is done using the min–max normalizing method that eventually provides linear transformation of function and, minimum and maximum transform data. The proposed method is based on outlier detection to measure the topic trending in both social media.

2 Literature Review

The buzz detection in social media is the process of detecting a topic that was discussed by a number of users in social media. An efficient buzz detection technique helps in recommending the topic to the users, and marketing about the campaign or the product. Twitter plays an important role in the current marketing trends about products or services. If more active users discuss their opinion about a product or services, it helps provide the respective companies with feedback. The buzz discussion about a product or services increases the company's awareness of a product's sales and profits. Some of the recent paper related to data mining in social media is considered to understand the latest technique used in buzz detection.

Abrahams et al. [1] utilized the text mining tools for categorizing the data provided by the user and filter the valuable content from the mass posting of user data. The automotive industries were used as an example for executing and changing the parameters of text-mining diagnostics from social media. The user discussion of automobile components in the social media was categorized accurately and automatically with the help of this tool. The most distinctive terms of each component determine the procedure description that helps in marketing and competitive intelligence to manufacturers, service centers, suppliers, and distributors. A large-scale research was required for this technique in vast social media, and it has to be verified manually by an automotive expert to evaluate the technique.

Derek Davis et al. [5] proposed the method of SociRank to extract a topic from news from social media and news media; then, the news were arranged in order of prevalence from the measure of media focus (MF), user attention (UA), and user interface (UI). The prevalence of a topic in the news media is considered as MF, which gives the media insight news. The prevalence of topic in social media is considered as UA, which will provide a topic that the user is interested in on social media mainly in Twitter. Finally, the interaction of topics in the social media and the number of users discussing about the topic was denoted as UI. SociRank helps to improve the quality of automatic system and offers a variety of news topics. This method is needed to be executed in the different types of social media to do a deep analysis of user interest on the topic.

Bichen Shi et al. [13] presented an efficient method for merging news and social feeds, named it as "Hashtagger+", and recommended Twitter hashtags to news articles. The several streaming hashtag recommendation approaches were studied and illustrated that the pointwise learning-to-rank method is more efficient than the multi-class classification, as a complex method. The new technique was included in the pointwise learning-to-rank method for data collection and feature computation, which increases the efficiency and coverage of the existing method. The learning-to-rank method evaluated the performance, and the result showed that it outperforms the state-of-art method. Some errors may present in this text mining technique, and it is not evaluated in this study.

Avudaiappan et al. [3] established a system for monitoring emergent keywords and summarizing the document stream based on dynamic semantic graphs of streaming documents. The emergent keywords were ranked by the notion of dynamic eigenvector centrality and summarizing emergent events with the help of an algorithm, which is based on minimum weight set cover. The demonstration of the method was done with an analysis of streaming Twitter data related to public security events. The summarization of data was discretized, which provide the element repeated in the data.

Kaleel et al. [8] proposed a method that uses the locality-sensitive hashing (LSH) technique for the detection of trending events discussed in Twitter. There are some key challenges in the method, namely, designing the dictionary using an incremental term frequency-inverse document frequency (TF-IDF) in high dimensional data to create tweet feature vector, leveraging LSH to identify the interested event, measuring the discussion of topic depending on time, geo-locations and cluster size, and increase in the efficiency of cluster discovery process, while retaining the cluster quality. The trending detection was conducted using this method for specific events with the help of LSH and k-means, then compared with the group average agglomerative clustering technique.

In order to overcome the issues in the existing methods and to improve the efficiency, the SACNN is used. The description of the proposed method is given in the next section, which helps to understand the technique used here.

3 Proposed Methodology

The social media contain abundant information about public opinions, events, and news, and also, a number of many active users. This causes a number of tweets in social media about different kinds of things or news. The identification of buzz news in social media helps to recommend the trending news topic for the user and also increases the number of tweets. The proposed system of condensed nearest neighbor (CNN) identifies the buzz topic in both Twitter and Tom's Hardware. Twitter is one of the most famous tools in social media, and Tom's Hardware forums cover the information about the latest technology. The buzz UCI dataset is used to evaluate the performance of the proposed CNN method for the detection of buzz. This section contains a brief explanation about the database used in this method and working of the CNN. The architecture of buzz monitoring in social media is presented in Figure 1. The μ denotes the mean of the discussion, and σ gives the standard deviation of the discussion, shown in eqs. (1) and (2). In the eq. (2), $P(x)$ is the probability of the x function.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x - \sigma)^2}{N - 1}} \tag{1}$$

where x gives the number of data observed in the social media, and N means the total number of discussions. Then, the mean is given by:

$$\mu = \sum xP(x) \tag{2}$$

3.1 Min–Max Normalization

Min–max normalization performs a linear transformation on the original data, and the min–max normalization maps a value d of p to d' in the range [*new_min* (p), *new_max* (p)]. The min–max normalization is calculated by eq. (3).

$$d' = \frac{[d - min(p) * [new_{max}(p) - new_{min}(p)]]}{[max(p) - min(p)]} \tag{3}$$

where
$min(p)$ = the minimum value of the attribute
$max(p)$ = the maximum value of the attribute

The min–max normalization maps a value d of P to d' in the range [0, 1], so put $new_{min}(p) = 0$ and $new_{max}(p) = 1$ in eq. (4). Now get the simplified formula of min—max normalization in Eq. (4).

$$d' = \frac{d - min(p)}{max(p) - min(p)} \tag{4}$$

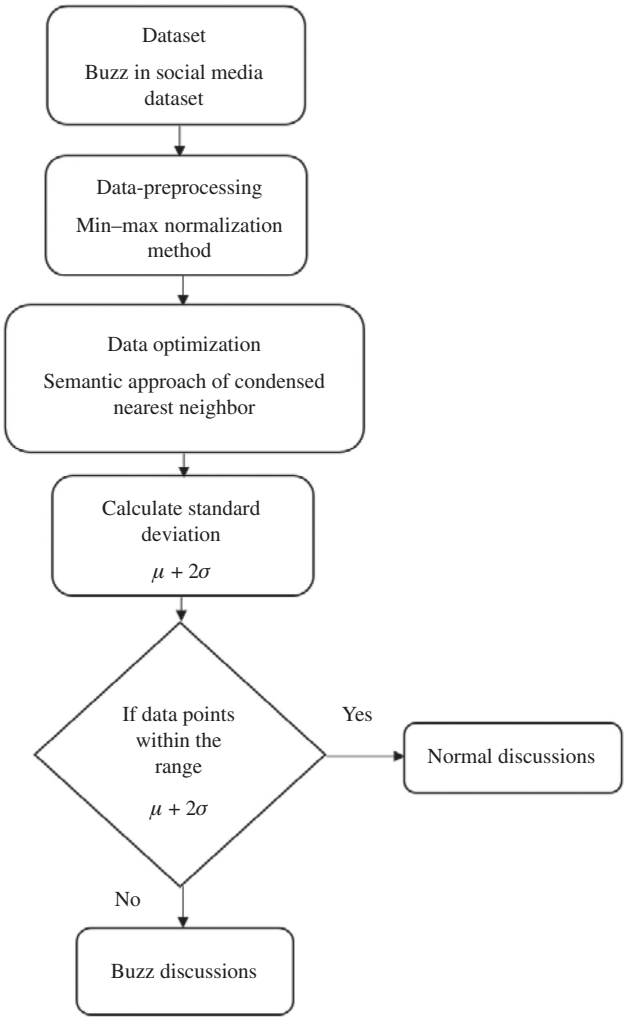

Figure 1: Architecture of Buzz Detection in Social Media.

Min–max normalization preserves the relationship among the original data values. These data WAS used by the SACNN to classify the buzz topic in the given dataset.

3.2 Database Description

The buzz UCI database is used for the evaluation of the proposed method, and CNN is used to find the buzz topic in the social media. The buzz database consists of two different social media data, namely, Twitter and Tom's Hardware. Twitter is micro-blogging platform having a number of active users and also creates an abundant information about various news or events occurring throughout the world. Tom's Hardware is a worldwide forum, which provides the details about the new technology with conservative dynamics but distinctive features of the data. The properties of Tom's Hardware and Twitter [2] are given below:

- Tom's Hardware is available in both German and French, but the only English version of the data is collected in this database.
- There is no direct audience estimator in Twitter, so the nad feature is used as a target feature. Audience estimator is given in Tom's Hardware, where the number of views is used.
- Twitter is more reactive of the exchanges than Tom's Hardware; 80% of the re-tweets take place in daytime. Following the initial tweet, the replies are produced during the following weeks.
- Twitter is a broader network than Tom's Hardware, with 500 million users visiting Twitter, while Tom's Hardware has only 41 million visitors per month.

In Table 1, Twitter is denoted as TW, Tom's Hardware is denoted as TH, the language of French, English, and German are given as FR, EN, and DE, respectively. There are 6671 topics present in the database related to the technology such as over-cloaking, grafikkarten, disque dur, android etc.

3.3 Condensed Nearest Neighbor Rule

The lazy learner like k-nearest neighbor classifier has the disadvantage of lacking in memory that requires storing the whole sample. The larger sample size requires more time for the sequential computer in response. In order to reduce the redundancy and number of free parameters, an editing procedure is followed, and this provides an alternative to abstract and parametric models that selectively discard the redundant part of the training set. The method minimizes the number of stored patterns by storing a subset of the training set only. The training set of the data may be similar, and some extra information is discarded.

Sample S of a subset has to be found, which is small and accurate. To choose the best subset z^*, out of $2^{|s|}$ possible subsets, the error measure is given in eq. (5) using the regularization theory:

$$\begin{aligned} z^* &= arg\min_z E(z) \\ E(z) &= \sum_{x \in s} L(x \vee \mathbb{Z}) + \gamma |z| \\ L(x|z) &= \begin{cases} 1, & if\ \ D(z_c, x) = min_j D(z_j, x) \wedge class(x) \neq class(z_c); \\ 0, & otherwise \end{cases} \end{aligned} \tag{5}$$

$z_c \in z$ is the closest stored pattern to using the distance measure $D(.)$. $L(x \vee z)$ is non-zero when the labels of x and z_c does not match. According to the regularization theory, using the combination of data, and prior smoothness information provides a solution to a problem. From eq. (5), data misfit is measured that is caused due to an error in classification using the 1-NN rule. The second term measures the size of the subset stored and also defines the smoothness of the class boundary. The nearest neighbor classifier categorizes the input space in the form of Voronoi tessellation, and the class boundaries are piecewise linear. There are more examples present that show that this boundary becomes more ragged and less smooth. The smoothest case is obtained in one example per class where linear boundaries between classes have γ, which is the parameter that indicates the trade-off between these two terms. Equivalently, this is a Bayesian approach that has higher prior probabilities to simpler models.

Minimizing eq. (5) is a combinatorial problem, and it is not solved by any known polynomial time procedure for optimal manner. Though no formal proof is known, we believe this problem to be NP-complete. The simple local search computation was first proposed as a CNN, and later as IB2, and growth and learn. The 1-NN rule is used to classify the pattern correctly for all elements of S with a consistent subset z of sample S. There are many consistent subsets present, and one trivial consistent subset is present in the set itself. Usually, many are interested in the minimum consistent subset because it computes with less resource and storage.

Initially, it is started with zero subset, then, one by one is passed, and in case the subset is not classified correctly, then, more subset value is added in the storage. This shows that the error is more important than the size (i.e. $\gamma < 1$ in eq. (1)). This method is for local search and does not guarantee the finding of minimal subsets, and if the training set order is changed, it leads to a different subset. The few subsets are enough

Table 1: Summary Per Language.

	NB. users			NB. discussions			Nb. examples			Nb. topics
	FR	EN	DE	FR	EN	DE	FR	EN	DE	
Twitter	72×10^3	0	8×10^3	50×10^4	0	1×10^4	4879	0	3026	4957
Tom's Hardware	24×10^6	30×10^6	10×10^6	23×10^7	28×10^7	46×10^6	76,292	35,317	29,089	6671

without additional subsets. It can classify correctly, when the stored subset classifies all the patterns in the training set. At each cycle, the classification accuracy increases, and pattern is also added to the storage subset. The number of patterns added at the cycle is reduced to 5% and helps increase the learning rate.

The classification rule is 1-NN, but other nonparametric variants are also possible at the expense of more computation. If the sample is noisy, there is also the possibility of using the k-CNN, but this may be costlier. If $k = 2i + 1$, then, for the correct classification of a new pattern, at least $i + 1$ of its nearest neighbors must be from a correct pattern class, and if this is not true, in the worst case, it would be added to $i + 1$ times z.

The pseudo code for CNN

```
PROCEDURE CNN(S, D, Z)
BEGIN
Z := {};
REPEAT
        additions:=FALSE;
        FOR all patterns in the training set DO
            Randomly pick x from training set, S
            Find z_c ∈ Z such that D(x, z_c) = min_j D(x, z_j)
            IF class(x) ≠ class(z_c) THEN
              Z := Z ∪ x;
              additions:=TRUE
            END IF
        END FOR
UNTIL NOT(additions);
END CNN;
```

The CNN optimizes the data, and the standard deviation is calculated using the equation $\mu + 2\sigma$. Classification occurs based on standard deviation, and this will show that it is a buzz or not. Occurring of a discussion below the standard deviation is a normal discussion, and discussion occurrences above the standard deviation are buzz. Once the CNN operates the classification of the topic in the database, then, the outcome of the proposed method is evaluated. The outcome is evaluated in terms of various parameters and compared with the existing method.

4 Experimental Result

This section gives a brief explanation about input data and outcome of the proposed method. Twitter and Tom's Hardware data are collected and stored in the dataset, and they are presented in the English, German, and French languages. The nad feature is applied as a target feature in Twitter and Tom's Hardware; the number of displays shows the visitors. This dataset is published in the UCI guidelines and are used in this research.

The result obtained from the proposed method is evaluated and compared with the existing method that is performed in a similar dataset. The outlier detection is performed using the proposed method SACNN. This will calculate the outlier value for the topic in Twitter and Tom's Hardware, and based on the value, this will classify the normal discussion and buzz discussion in the social media. The min–max normalization techniques are performed on the data to fill the mission data, and then it is used for classification. This technique is executed in the Python 2.3.5 Jupyter notebook software installed in the system with 4GB of RAM configuration. The formula for the calculating parameter is shown in the eqs. (6)–(8). The accuracy is the measure of trueness and the performance of the proposed method in the buzz detection. The mean squared error is the estimator of the average of the square of the errors, and that is the average value between the expected and estimated values. The RMSE is the square root of MSE, and the square root is introduced to make the scale of the errors the same as the scale of the targets. In eq. (6), TP is true positive, TN is true negative, FN is

false negative, and FP is false positive. Equation (7) measures the MSE value of the proposed method, where n denotes the vector of predictions that is generated from the sample of n data points, and y is the vector of the observed value of the variable being predicted.

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{6}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{y_i}\right)^2 \tag{7}$$

$$RMSE = \sqrt{MSE} \tag{8}$$

Accuracy calculated from the SACNN is compared with the accuracy of the existing method [2] using the same database as shown in Table 2. The research [2] involves the use of hybrid artificial bee colony with KNN to identify the buzz in the social media. The buzz is identified through the increase in user, attention level, user interaction, user count, etc. The database is separated into four sets, the existing and proposed method performs on the sets. The outcome of the proposed and existing method is shown in Table 2. This shows the efficiency of the proposed method over existing methods. The four parameters are measured for the methods such as accuracy, MSE, RMSE, and NRMSE. The error rate of the proposed method is less compared to the existing method. The CNN has the advantage of storing fewer samples compared to the KNN, which causes to reduce the computation of the system. The CNN stores the subset of the training samples and decreases the samples of the data. It reduces the irrelevant data that helps to improve the performance of the detection. The existing method kNN is cluster data based on the voting values. The proposed SACNN is involved in the cluster of the classes based on the distance, and the values of the data that are present in the two classes are calculated. This method helps to reduce the noise of the data and provides a high-accuracy classification. The execution time of the proposed method is also reduced due to the less measuring of the voting value for the data.

The relative error of the proposed and existing methods are shown in graphical representation in Figure 2. The relative error of the SACNN is less than that of the proposed method, and this increases the performance of the system. The relative error is calculated for a number of different test data and is compared with the existing method [9]. This clearly showed that the SACNN technique outperformed all existing techniques. This method has the capacity to provide high accurate buzz detection in Twitter and Tom's Hardware. It is proven that this technique can be applicable to the buzz detection method in social media.

The run time of the SACNN is compared with different methods in the same database, as given in Table 3. The different techniques were processed by Shekar et al. [12] to understand the performance of the feature selection in different datasets. The random forest has a run time of 67.9 s, and the principal component analysis (PCA) has a run time of 19.25 s. The proposed method has a lower run time of 17.26 s compared to the other existing method.

The buzz detection technique is useful in finding the trending topic or event in a given location. This outlier detection is applied to the buzz detection, which helps to particularly identify the event or

Table 2: Comparison of Existing and Proposed Method.

Method	Database	Accuracy	MSE	RMSE	NRMSE
SACNN	Set 1	99.9451	0.02773	0.000234	0.02342
	Set 2	99.9475	0.02724	0.000215	0.02382
	Set 3	99.9368	0.02734	0.000268	0.02322
	Set 4	99.982	0.02683	0.000282	0.02121
ABC-KNN [2]	Set 1	99.83	0.000402	0.040719	0.040719
	Set 2	99.824	0.04095	0.000418	0.04098
	Set 3	99.828	0.04182	0.000473	0.04124
	Set 4	99.834	0.04126	0.000416	0.04057

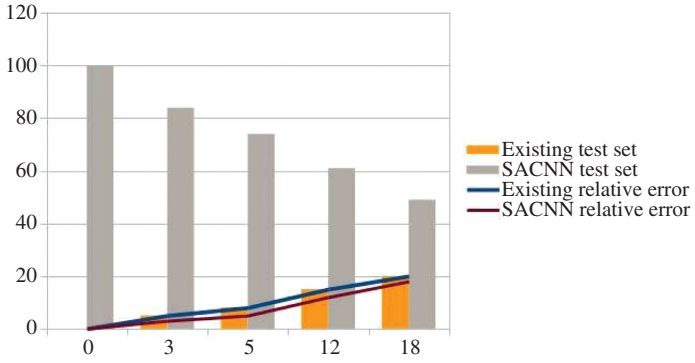

Figure 2: Relative Error for the Proposed Method.

Table 3: Run Time of the Proposed Method Compared with Different Methods.

Methods	Run time (s)
Genetic algorithm + support vector machine – AdaBoost [12]	383.58
Principal component analysis – AdaBoost [12]	19.25
Particle swarm optimization + artificial neural network – AdaBoost [12]	5467.89
Random forest [12]	67.9
SACNN	17.26

trending topic. The proposed method provides the most accurate buzz detection compared with the existing method, and this shows the effectiveness of the proposed method.

5 Conclusion

The social media play a vital role in news transaction and events occurring in all over the world. Twitter has many active users who are posting their own content related to the news or events. There are over millions of users in Twitter who create an enormous amount of data, which makes it difficult for buzz monitoring. The aim of this research is to provide effective buzz monitoring in social media to help the users have a better experience. SACNN helps in identifying the highly discussed topic in social media. The buzz UCI dataset contains the data of Twitter and Tom's Hardware, and this dataset is used for the proposed buzz detection using an outlier. The data is normalized with the help of the min–max normalizer, and SACNN is used in buzz detection for both social media. The proposed method of buzz monitoring system is evaluated, and the experimental results showed efficiency of the technique. The proposed SACNN attained an accuracy of up to 99% in the UCI dataset, which is higher than the existing classification method. The relative error of the proposed system is less compared to the existing technique. The future work of this method has a chance to perform this method with the dimensional reduction technique and to also evaluate the other social media datasets.

Bibliography

[1] A. S. Abrahams, J. Jiao, W. Fan, G. A. Wang and Z. Zhang, What's buzzing in the blizzard of buzz? Automotive component isolation in social media postings, *Decis. Support Syst.* **55** (2013), 871–882.

[2] R. Aswani, S. P. Ghrera, A. K. Kar and S. Chandra, Identifying buzz in social media: a hybrid approach using artificial bee colony and k-nearest neighbors for outlier detection, *Soc. Netw. Anal. Min.* **7** (2017), 38.

[3] N. Avudaiappan, A. Herzog, S. Kadam, Y. Du, J. Thatche and I. Safro, Detecting and summarizing emergent events in microblogs and social media streams by dynamic centralities, in: *2017 IEEE International Conference on Big Data (Big Data)*, pp. 1627–1634, IEEE, 2017.

[4] J. Benhardus and J. Kalita, Streaming trend detection in twitter, *Int. J. Web Based Communities* **9** (2013), 122–139.

[5] D. Davis, G. Figueroa and Y. S. Chen, SociRank: identifying and ranking prevalent news topics using social media factors, *IEEE Trans. Syst. Man Cybern. Syst.* **47** (2017), 979–994.
[6] R. Dovgopol and M. Nohelty, Twitter hash tag recommendation (2015). arXiv preprint arXiv:1502.00094.
[7] D. Gao, W. Li, X. Cai, R. Zhang and Y. Ouyang, Sequential summarization: a full view of twitter trending topics, *IEEE/ACM Trans. Audio Speech Lang. Process.* **22** (2014), 293–302.
[8] S. B. Kaleel and A. Abhari, Cluster-discovery of Twitter messages for event detection and trending, *J. Comput. Sci.* **6** (2015), 47–57.
[9] F. Kawala, A. Douzal-Chouakria, E. Gaussier and E. Dimert, Prédictions d'activité dans les réseaux sociaux en ligne, in: *4ième Conférence sur les Modèles et L'Analyse des Réseaux: Approches Mathématiques et Informatiques*, p. 16, 2013.
[10] S. Khater, D. Gračanin and H. G. Elmongui, Personalized recommendation for online social networks information: personal preferences and location-based community trends, *IEEE Trans. Comput. Soc. Syst.* **4** (2017), 104–120.
[11] N. Shahid, M. U. Ilyas, J. S. Alowibdi and N. R. Aljohani, Word cloud segmentation for simplified exploration of trending topics on Twitter, *IET Software* **11** (2017), 214–220.
[12] A. K. Shekar, P.I. Sánchez and E. Müller, Diverse selection of feature subsets for ensemble regression, in: *International Conference on Big Data Analytics and Knowledge Discovery*, pp. 259–273, Springer, Cham, 2017.
[13] B. Shi, G. Poghosyan, G. Ifrim and N. Hurley, Hashtagger+: efficient high-coverage social tagging of streaming news, *IEEE Trans. Knowl. Data Eng.* **30** (2018), 43–58.
[14] J. Skaza and B. Blais, Modeling the infectiousness of Twitter hashtags, *Physica A* **465** (2017), 289–296.
[15] M. Vicente, F. Batista and J. P. Carvalho, Gender detection of Twitter users based on multiple information sources, in: *Interactions Between Computational Intelligence and Mathematics Part 2*, pp. 39–54, Springer, Cham, 2019.
[16] H. Wang, Y. Li, Z. Feng and L. Feng, ReTweeting analysis and prediction in microblogs: an epidemic inspired approach, *China Communications* **10** (2013), 13–24.
[17] M. C. Yang and H. C. Rim, Identifying interesting Twitter contents using topical analysis, *Exp. Syst. Appl.* **41** (2014), 4330–4336.
[18] U. Yaqub, S. A. Chun, V. Atluri and J. Vaidya, Analysis of political discourse on twitter in the context of the 2016 US presidential elections, *Gov. Inf. Q.* **34** (2017), 613–626.