Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2018) 000–000

Procedia
Computer Science

www.elsevier.com/locate/procedia

The 14th International Conference on Future Networks and Communications (FNC)
August 19-21, 2019, Halifax, Canada

Anomaly Detection Method for Online Discussion

Peter Krammer[a]*, Ondrej Habala[a]*, Ján Mojžiš[a], Ladislav Hluchý[a], Marek Jurkovič[b]

[a]*Institute of Informatics, Slovak Academy of Sciences, Dúbravská cesta 9, 845 07 Bratislava 45, Slovak Republic*
[b]*Centre of Social and Psychological Sciences - Institute of Experimental Psychology, SAS, Šancová 56, 811 05 Bratislava, Slovak Republic*

Abstract

The presented article deals with the analysis of users discussing online, on the two most famous Slovak servers Cas.sk and Sme.sk. They provide a sufficiently representative sample of data to detect and compare the essential common behavioral characteristics of users in online discussions. This also makes it possible to identify user partitioning and to develop new methods to detect anomalies, specifically designed to differentiate discussing users with abnormal behavior. In the presented article, such a method is defined, with multiple tuning parameters, using a classification neural network. The proposed method is applied on real data, obtaining encouraging results.

Keywords: Clustering; Classification; Online Discussion; Anomaly Detection; Machine Learning; Neural Network; User Segmentation

1. Introduction

Currently, machine learning methods are used in different areas and domains. They bring new tools and possibilities of more intelligent use of data, modeling, forecasting, increase of efficiency, safety etc. The area of communication, including online discussions on various social networks, online chats, forums, and so on, where user interaction is significant, is not the exception. From a large amount of data contained in statuses, profiles, message contents, as well as technical information about connections, times, message frequencies, and other parameters, interesting and usable information can be obtained. By applying statistical approaches, these can be reflected in improved services in the form of new tools such as lookup of specific unwanted users.

* Corresponding author. *E-mail address:* peter.krammer@savba.sk, ondrej.habala@savba.sk

The present article deals with cluster analysis of online discussions, specifically clustering users into groups based on general characteristics, without using sentiment analysis. Our long-term goal is to identify users who are either violating the discussion rules or make discussion difficult for others. Multiple publications have addressed the topic of clustering users into groups [1, 2, 3], to some extent differing in the number of natural/native user groups as well as the descriptive attributes used. Multiple previous papers are focused on antisocial behavior [4] or trolling [5] in online discussions. Our previous research in this area has tried to detect number and location of clusters [6, 7]. At the same time, we found several interesting properties, primarily the existence of a dominant cluster, and we were able to identify the continuity of the pattern distribution in the attribute space. Our current research builds on these achievements, trying to define a highly stable method suitable for detecting anomalies (users with abnormal behavior in online discussions). Anomaly detection methods, supervised as well as unsupervised, are an important component of (semi-) automated identification of potential threats. They typically stem from general outlier detection techniques [8], [9], [10] but there is a need to distinguish harmless noise, fluctuations or various forms of novelty from malicious intentions, actions or attacks. There are many subtypes of anomaly detection on the nature of the input data, the character of anomalies (point wise, contextual or collective), the form of output and the presence or absence of data labels [11], [12]. Anomaly detection can be used in a wide variety of investigative contexts, such as fraud detection [13], intrusion detection [14], video surveillance systems [15], or forensic investigations in general [16]. In a special context of social networks, process-mining approach [17], and anomaly detection in forum [18] can also be relevant. Large selection of these methods provides a general approach to detection of anomalies. The problem, however, is that in modeling human behavior, respectively in describing specific characteristics of persons, some aspects are often unpredictable or difficult to quantify. For this particular purpose of detecting anomalies, therefore, there is a need for methods with markedly high stability, noise resistance, and/or low sensitivity to selected parameters. The present article defines a method with such desirable properties while focusing on its high stability. A more complete description of clustering, including its definition and properties is available in [19].

2. Previous Research

In the previous research [6, 7], data sets from 2 online discussion servers - Sme.sk and Cas.sk - were analyzed, with data divided between domestic and foreign (international) topics. In these data, significant connections in the distribution(s) of users were revealed, based on the observed relative attributes.

Current research is a follow-up to our previous work, which focused more on revealing common features between different datasets and their cluster analysis.

Our long-term goal is to identify the discussants who spoil the discussion for the rest (by publishing hoaxes, being vulgar, inserting advertisements…) or who in their comments break the law (for example by propagating racism). Currently we possess only some aspects of these forms of expression, we are missing data on blocked user profiles and their comments. We expect to gain access to these data in the future. In our previous analysis we have found out that a significant amount of discussants' characteristics have similar statistical distribution – with continuously changing density. We have also confirmed the existence of a dominant cluster, which covers over 90% of the discussants.

Based also on these aspects we have decided to target within the domain of online discussions a broader set of principles of anomaly detection, or to notice the common characteristics of this domain. The more general principles can be this applied to different selected discussants' characteristics (input attributes) according to the targeted type of anomaly which we want to detect. This approach is convenient also because of very similar distributions of several different attributes.

In this manner the concept of abnormal behavior is simplified into *behavior that is not common* (is not part of the majority clusters). Thus identified discussants are not yet automatically targeted; they are just potential candidates to be submitted for a more thorough analysis of their behavior and detection of rule violations.

This generalized approach allows also for applications in detection of specific types of rule violators on the web, providing their attributes are well defined. If we engage in Anomaly Threshold tuning (described below), together with high method stability we can also expect stable and relevant results. For greater representativeness, the original data sets were enhanced with newly collected data. The significant increase in the number of users and contributions is particularly apparent in the Sme.sk dataset. The number of records before cleaning of data is described in more detail in Table 1.

Table 1. Number of analyzed news and users in discussions on Cas.sk and Sme.sk, before the process of data cleaning.

	Current Research		Previous Research	
Dataset	Discussion posts count	Participant profiles count	Discussion posts count	Participant profiles count
Cas.sk Domestic	339 828	32 531	306 064	26 030
Cas.sk International	577 836	30 756	457 843	29 958
Sme.sk Domestic	3 928 008	70 200	968 698	34 998
Sme.sk International	2 000 968	39 764	489 238	22 764

For all available discussing users, a number of characteristics describing user aspects have been quantified. Such data have been transformed into a homogeneous spreadsheet, where individual rows represent individual discussing users. Table columns represent individual characteristics that we perceive as attributes, also referred to as features, for machine learning. Available data were then cleaned; individual attributes were converted into a standardized form. A more detailed description of the data preprocessing and cleaning phase as well as the importance of each attribute is available in [6], [7].

- Cleaning of data consisted of removing users with locked accounts and users whose post count was below 27. Such users are not accurately represented by the individual characteristics as they have not yet been sufficiently active in the discussions. Their data bring too much noise that needs to be suppressed for further analysis.
- Converting attributes into relative numbers (due to the number of posts/messages) is important in terms of the ability to compare users. The original absolute values cannot be directly compared due to the different numbers of messages, or the length of time a user is active in the discussions.
- Attribute standardization consists of recalculating attributes by linear transformation to attributes whose mean value is 0 and standard deviation 1. The reason for such adjustment is the need to scale ranges to balance the significance of individual attributes for clustering purposes. In this case, the use of normalization to an interval of <0, 1> is inappropriate, as it is significantly affected by two extremes (minimum and maximum) that do not reflect the distribution of data in their given interval.

As can be seen in Table 2, the volume of data after adjustment has been significantly reduced, compared to Table 1. There is a more pronounced decline in the number of users, as many users have had fewer posts. On the other hand, the decrease in the number of posts in the cleaned data set is not so significant.

Table 2. Number of analyzed news and users on servers on Cas.sk and Sme.sk servers, after the data cleaning process.

	Current Research		Previous Research	
Dataset	Discussion posts count	Participant profiles count	Discussion posts count	Participant profiles count
Cas.sk Domestic	235 538	2 068	206 660	1 885
Cas.sk International	474 016	2 671	368 337	2 201
Sme.sk Domestic	3 500 748	15 177	776 501	5 773
Sme.sk International	1 759 539	9 005	370 072	3 522

2.1. Clustering Tendency Testing

Since we are planning to cluster data in the next steps, it is advisable to test whether the data analyzed tend to form a cluster. For this purpose, a statistical test was used - Kolmogorov-Smirnov goodness-of-fit hypothesis test [20]. Using it we tested whether the data corresponded to a uniform distribution, which we simply take as opposed to cluster tendencies. At the significance level of 0.05, it was expected that all the attributes analyzed tend to form clusters. The resulting data, following the process of cleaning and standardization, and the conversion to relative values, contained a number of characteristics - attributes, from which the 4 best were best chosen and used for clustering:

rel_react_to_me - represents the average number of responses/answers from other users attributable to one post.

rel_post_per_day - represents the average number of user posts per day rel_word - represents the average length of a user's contribution, expressed in words

rel_violation - represents the code violation ratio, against the number of posts.

Several aspects have been taken into account when selecting appropriate attributes:

- statistical independence of attributes
- the significance of attributes when identifying abnormal users
- noise of attributes as well as similarity of histograms between datasets

2.2. Visual Comparison

In the cleaned and recalculated attributes, there was a significant mutual similarity across the datasets, as shown in Figure 1. This suggests a suitably chosen representation of properties as well as the existence of a pattern in the data. As apparent in Figure 1, data distribution is largely consistent in a large majority of cases, indicating high representativeness of datasets, but it also makes clusters identification more difficult.

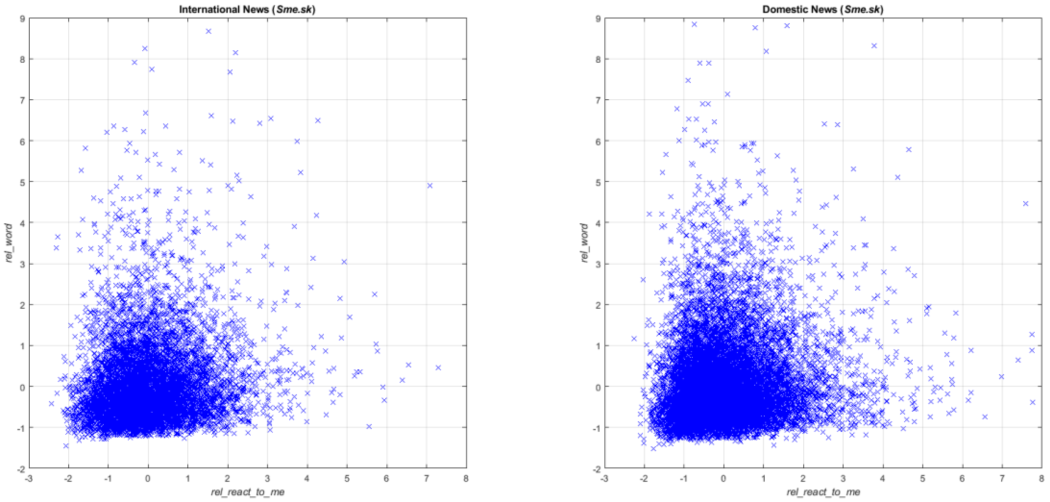

Fig. 1. Illustration of the distribution records in the space rel_react_to_me and rel_word, for posts in the domestic and international discussions

As shown in the work [7], a highly stable dominant cluster with a distinctly continuous sample distribution was successfully identified by the canopy method [21]. This knowledge can be used to define an anomaly-identifying method, using bridged supervised and unsupervised learning methods (when using clustering multiple times, along with a classification model).

3. Defining the Method

The presented anomaly detection method is suitable for the case of continuously varying sample distribution density (indicating problematic cluster identification). It is also suitable for cases where high stability of anomaly resolution is required, which is often the case in modeling specific aspects of human behavior (when the existence of a search pattern is not very pronounced). The method has 4 tunable parameters:

max_density - a parameter representing the maximum density of the canopy clustering method. The recommended interval for this parameter is 5.5 to 10.5, while a value of 6.5 was used in our experiments.

seed_count - an integer parameter representing the number of times the clustering repeats for each density value. In our case the value was 20, the appropriate values are from 5 to approximately 30. These two parameters with increasing values increase the stability of the method, however, also adversely affect the time complexity of the method. significant_ratio - the parameter represents the boundary of the number of records from the total number of records from which the cluster is considered to be significant. For this parameter, we used 0.25, which means that each cluster containing at least 25% of the samples is considered significant.

anomaly_threshold – represents a threshold, below which the record will be considered an anomaly. This allows us to continuously tune the method to the desired sensitivity. In our experiment we used primary a value of 0.65, which means that every record that at least in 65% of the cases (with different seeds and densities of clustering) was not in any significant clusters, is described as an anomaly. The idea behind these 2 parameters is that a record that occurs in any significant cluster, at a frequency greater than or equal to significant_ratio, is not considered an anomaly. The method consists of 3 phases, schematically shown in Figure 2.

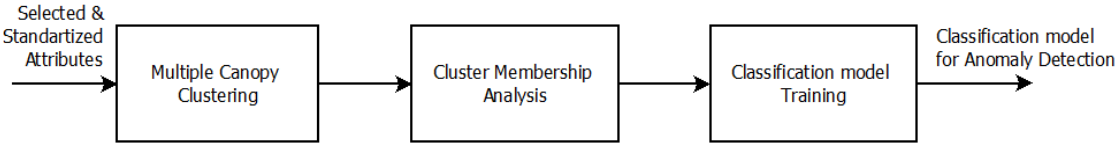

Fig. 2. Simplified block diagram of anomaly detection method

In the first phase of the method, the clustering process is repeatedly implemented using the Canopy Clusterer [20]. The reason for using this type of clusterer is its excellent parallel processing capability and its character - it is a density based clusterer. In the previous research [7] in the domain of online discussions this clusterer showed significantly positive results. It showed repeated stability when detecting the dominant cluster, which has been confirmed with a larger data set. A detailed analysis of the distribution of samples into clusters has also shown that the results in general conform to the logically expected distribution.

Repeating clustering process is carried out for different values of seed and density, which provides the stability of the method as a whole. This multiple clustering can also be described by the pseudocode 1 below. The output of this phase is a multiple division/assignment of individual records into clusters, represented in the pseudocode by the variable cl.

```
i = 0;
for (seed = 1; seed <= seed_count; seed = seed + 1)
{
  for (dens = 1.5; dens <= max_density; dens = dens + 1.0)
  {
    cl[i] = CanopyClustering (Data, seed, dens)
    i = i + 1;
  }
}
```

Pseudocode 1: Pseudocode for Multiple Canopy Clustering Phase

The second phase of the method is aimed at evaluating the clustering. It is necessary for each implemented clustering process to distinguish which identified clusters comprise a sufficiently large number of records (relative to the parameter Significant_ratio) and thus are significant. Subsequently, it is necessary to go through all the records and decide how many times the individual records have occurred in any significant clusters, relative to the total number of clusters. If this ratio is lower than the Anomaly Threshold parameter for a particular record, then the record is declared an anomaly. In the last, third phase of the method, we have trained a classification model for which the target attribute is given information about the records - whether by clustering they have been identified as anomalous or not. The reason for this is the requirement for the method to work for new entries, without the necessity to re-implement the significantly time-consuming clustering. Thus the trained classification model can generalize the boundaries of each identified dominant cluster, and quickly distinguish abnormal cases from ordinary records.

3.1. Applying the Method

This method was applied to the data from the server sme.sk, for domestic and foreign discussion messages. Several Anomaly Threshold values were used. Example of comparison of detected anomalies for thresholds 0.30 and 0.65 are shown in Fig.3. The values of 0.30 and 0.65 for Anomaly Threshold have been experimentally selected to show differences in the visualization in Fig. 3, even with the high stability of the method (which is especially apparent in Table 3). This parameter allows for partial method tuning according to the specified requirements. In general we consider the interval (0.15, 0.75) to be practical, depending on the quality of the analyzed data.

From Table 3, we can see that the frequencies of the identified anomalies change only very slowly with the change of the Anomaly Threshold parameter. When changing the threshold from 0.75 to 0.30, for both the international and domestic dataset the number of detected anomalies changed only by approximately 3%. Given the total range in which this threshold moves - the interval (0, 1), a change of 3% is very small and indicates a strong stability of the method, as well as the identification of the natural generalizing boundary in data, between normal and abnormal behavior.

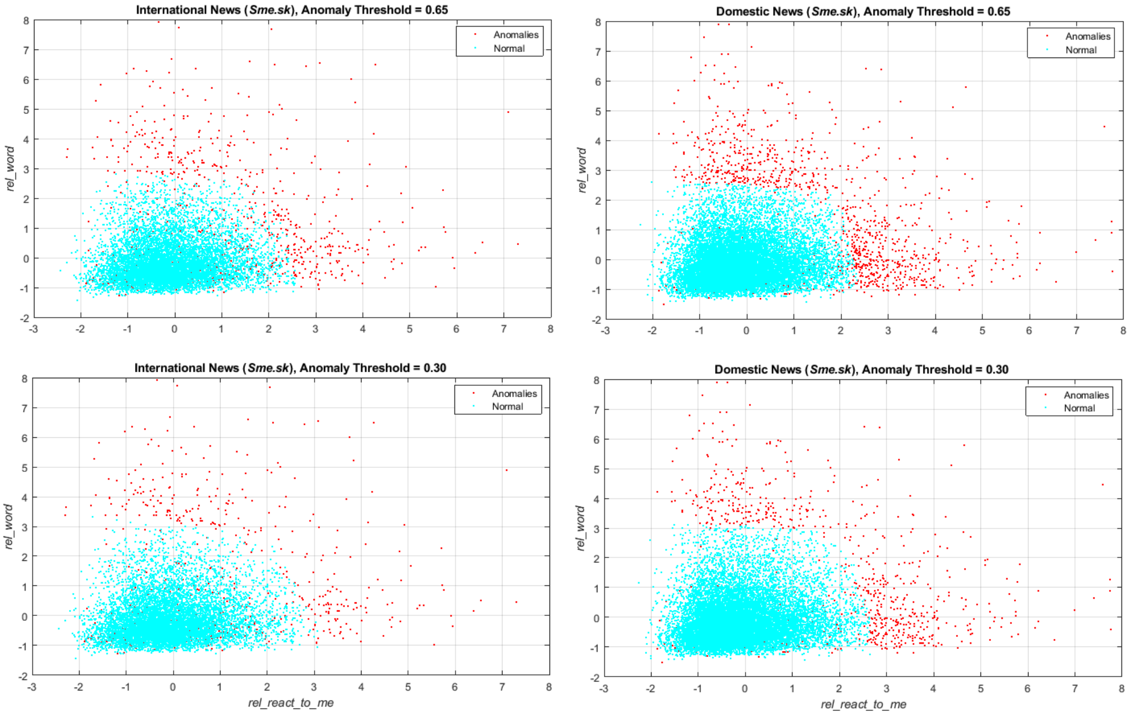

Fig. 3. Visual comparison of anomaly detection for Anomaly Threshold 0.30 and 0.65, domestic and Foreign News

Table 3. Representation of the Anomaly Threshold, the number of anomalies identified as well as the percentage of anomalies identified from all samples.

Anomaly Threshold	Detected Anomalies in International Dataset	Detected Anomalies in Domestic Dataset
0.75	959 (10.65%)	1241 (8.18%)
0.70	917 (10.18%)	1165 (7.68%)
0.65	860 (9.55%)	1128 (7.43%)
0.60	825 (9.16%)	1091 (7.19%)
0.55	805 (8.94%)	1055 (6.95%)
0.50	779 (8.65%)	1019 (6.71%)
0.45	749 (8.32%)	954 (6.29%)
0.40	727 (8.07%)	900 (5.93%)
0.35	704 (7.82%)	837 (5.51%)
0.30	682 (7.57%)	788 (5.19%)

3.2. Training of Classification Model

In the final phase, different types of classification models were trained based on the identified potential anomalies: Support Vector Machine (SVM) with Sequential Minimal Optimization (SMO), Multi-Layer Perceptron (MLP) Classifier, and Radial Basis Function (RBF) Classifier. The following settings were used.

SVM: epsilon 1.0e-12, Complexity parameter = 1.0
RBF: number of functions = 4, tolerance parameter = 1.0e-6, ridge penalty = 0.01
MLP: Activation function = Approximate Sigmoid, Number of hidden units = 2, learning rate = 0.3, momentum = 0.2

For validation, the - 20 Fold Cross Validation method was used, with the accuracy of the classification models being shown in Table 4. The highest accuracy was achieved in multiple cases by the RBF Classifier model, which seems to be the most appropriate for this purpose.

Table 4. Accuracy of the different types of trained classification models for 0.65 and 0.30 Anomaly Threshold values.

		Sme.sk, Domestic News		Sme.sk, International News	
Model Type	Anomaly Threshold	F-Measure	ROC Area	F-Measure	ROC Area
SMO	0.65	0.961	0.789	0.955	0.816
MLP Classifier	0.65	0.993	0.994	0.991	0.987
RBF Classifier	0.65	0.993	0.999	0.990	0.999
SMO	0.30	0.967	0.743	0.960	0.795
MLP Classifier	0.30	0.994	0.983	0.986	0.997
RBF Classifier	0.30	0.994	1.000	0.993	0.999

In order to validate the method as a whole, the DBSCAN clusterer was used for comparison. This is a density-based clusterer, capable in addition to identifying clusters also to identify noise - like anomalies. This clusterer has 2 parameters, of which the first parameter specifies the minimum number of samples required to define a new cluster.

The DBSCAN Method has been shown to be practical in previous research [7] since it is a density-based clustering method. This is very opportune in the case of data distribution with continuously changing density, without evident gaps in the spatial sample distribution. Another positive aspect of the DBSCAN Method is its ability to detect noise. Because of the missing data on the target attribute (whether a sample is anomalous in our context) we were forced to use unsupervised learning. To ensure identical conditions with the proposed method, this parameter was set (according to the parameter significant_ratio) to values 3794 for the domestic dataset and 2251 for the foreign dataset. However, this parameter does not have a significant impact since only 1 - dominant cluster was identified in all cases. However, the identified DBSCAN anomaly number was very sensitive to the second parameter - epsilon (determining the allowed sample distance), as can be seen in Table 5.

Table 5. Number of identified anomalies by the DBSCAN method, for different values of epsilon parameter.

Parameter epsilon	Sme.sk, Domestic News	Sme.sk, International News
0.050	1579 (17.53 %)	1366 (9.00 %)
0.055	1285 (14.27 %)	1139 (7.50 %)
0.060	1054 (11.70 %)	961 (6.33 %)
0.065	896 (9.95 %)	795 (5.24 %)
0.068	824 (9.15 %)	714 (4.70 %)
0.070	749 (8.32 %)	661 (4.36 %)
0.075	637 (7.07 %)	574 (3.78 %)

Thus, the stability of the results achieved by the DBSCAN method is significantly lower compared to the newly designed method, which is also seen by comparing Table 3 and Table 5, when changing the tuning parameter. It has also been found that for suitably selected parameters of both methods (so that both methods identify approximately equal numbers of anomalies) more than 96% of these identified anomalies are identical. Substantially similar results were also achieved for the dataset Cas.sk, which testifies to the qualities of this method. Because of the more generic approach of the method it is evident that the performance evaluation will depend on the concrete case. In this case we use for validation identical weights for both 1st and 2nd type errors – we use the F1 score.

4. Conclusions

In the presented paper, a new anomaly detection method, primarily designed to detect abnormal behavior in online discussions and discussion forums was proposed. Amid advantages of this method are multiple options for its tuning and high stability of the results, despite the significant variability in the behavior of users or significant noise in the input. The method was tested on 2 available datasets, with very encouraging results. In general, the use of multiple clustering over large data is often a time-consuming process. Application of the presented method, however, is based in simple prediction using a trained classification model which represents the boundary between abnormalities and ordinary cases. Therefore, it is not necessary to have the entire dataset of the samples to detect anomalies, the trained model is sufficient. This makes the application of this method very fast, memory-saving, and can therefore be used on less powerful platforms (it is especially suitable for example in mobile platforms). In view of the high speed and low memory requirements as well as the high stability and ability to identify anomalies even in datasets with continuously changing density, this method is well suited for practical application in current problems. It is possible to expect that this method will be applicable also for anomaly detection in other areas of human behavior, not just in discussion forums and chat rooms. This is attested by the fact that in the more complex case of continuously changing density we were able to achieve high stability. In the future, we plan to test the presented method also in other areas and to find out its full potential as well as its other possibilities of use, its properties, resp. possible restrictions. We will also apply the method to data with more precisely defined anomaly properties, using the F2 score for training, to achieve a more precise estimation of the precisions of classification models.

Acknowledgements

This research was supported by the projects VEGA 2/0167/16, VEGA 2/0155/19, U-Comp APVV-17-0619 and PROCESS EU H2020-777533.

References

[1] Chan J, Hayes C, Daly EM. Decomposing Discussion Forums and Boards Using User Roles. In *ICWSM*. 2010 Jan; 10:215-8.

[2] Morrison D, McLoughlin I, Hogan A, Hayes C. Evolutionary Clustering and Analysis of User Behaviour in Online Forums. In *ICWSM* 2012.

[3] Fortunato S, Hric D. Community detection in networks: A user guide. *Physics Reports*. 2016 Nov 11; 659:1-44.

[4] Cheng J, Danescu-Niculescu-Mizil C, Leskovec Antisocial Behavior in Online Discussion Communities. In *ICWSM* 2015 Apr 2 (pp. 61-70).

[5] Hardaker C. Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions. J*ournal of Politeness Research* 2010; 6 (2): 215-42.

[6] Mojžiš Ján, Krammer Peter, Kvassay Marcel, Budinská Ivana, Hluchý Ladislav, Jurkovič Marek: Crawling and Analysis of Online Discussions in Major Slovak National Newspapers. *22nd IEEE International Conference on Intelligent Engineering Systems*: Proceedings. - Spain: IEEE, INES, 2018, p. 119-126. ISBN 978-1-5386-1121-0.

[7] Krammer Peter, Kvassay Marcel, Mojžiš Ján, Budinská Ivana, Hluchý Ladislav, Jurkovič Marek: Clustering analysis of online discussion participants. *Procedia Computer Science*, 2018, vol. 134, p. 186-195. ISSN 1877-0509.

[8] Arthur Zimek, Ricardo J. G. B., and Jorg Sander, "Ensembles for unsupervised outlier detection: challenges and research questions, a position paper", Acm Sigkdd Explorations Newsletter, Vol. 15, No. 1, 2014, pp. 11-22.

[9] Malik Agyemang, Ken Barker and Rada Alhajj, "A comprehensive survey of numeric and symbolic outlier mining techniques", Intelligent Data Analysis, Vol. 10, No. 6, 2006, pp. 521-538.

[10] Victoria Hodge and Jim Austin, "A survey of outlier detection methodologies", AI review, Vol. 22, No. 2, 2004, pp. 85-126.

[11] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection," ACM Comput. Surv., vol. 41, no. 3, pp. 1–58, 2009.

[12] M. Goldstein, "Anomaly Detection in Large Datasets," University of Kaiserslautern, 2014.

[13] Y. Kim, A. Kogan, "Development of an Anomaly Detection Model for Bank's Transitory A.Sys." J.Inf.Syst., vol.28, no.1, pp.145–165, 2014

[14] R. Zuech et al."Intrusion detection and Big Heterogeneous Data: a Survey," J. Big Data, vol. 2, no. 1, pp. 1–41, 2015.

V. Jyothsna, "A Review of Anomaly based Intrusion Detection Systems," Int. J. Com. Appl., v. 28, no. 7, p. 975–8887, 2011.

[15] F. Nater, "Abnormal Behavior Detection in Surveillance Videos," 2012.

[16] A. Mohd Ali, P. Angelov, and X. Gu, "Detecting Anomalous Behaviour Using Heterogeneous Data," in Advances in Comp Intelligence Systems: 16th UK Workshop on Computational Intelligence, Sept. 2016, Lancaster, UK, pp. 253–273.

[17] Aalst, W.V.D., A. Adriansyah, A.K.A.D. Medeiros, F. Arcieri and T. Baier et al., 2012. Process mining manifesto. Bus. Process Manage. Workshops, 99: 169-194. DOI: 10.1007/978-3-642-28108-2_19

[18] Kramer, S. (2010, July). Anomaly detection in extremist web forums using a dynamical systems approach. In *ACM SIGKDD Workshop on Intelligence and Security Informatics* (p. 8). ACM.

[19] Ulrike von Luxburg: Clustering Stability: An Overview, 2010, Foundations and Trends in Machine Learning: Vol. 2: No. 3, pp 235-274.

[20] Justel, Ana, Daniel Peña, and Rubén Z. "A multivariate Kolmogorov-Smirnov test of goodness of fit." *Statistics & Probability L.* (1997).

[21] McCallum, A., Nigam, K., & Ungar, L. H. (2000, August). Efficient clustering of high-dimensional data sets with application to reference matching. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 169-178). ACM.