Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2018) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

The 13th International Conference on Future Networks and Communications (FNC 2018)

Clustering analysis of online discussion participants

Peter Krammer[a*], Marcel Kvassay[a*], Ján Mojžiš[a*], Ivana Budinská[a], Ladislav Hluchý[a], Marek Jurkovič[b]

[a] *Institute of Informatics, Slovak Academy of Sciences, Dúbravska 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

In this paper we perform density-based clustering of discussion participants from online editions of two major Slovak national newspapers, *Sme.sk* and *Cas.sk*. We use language-independent statistical attributes characterising their communication patterns and the content of their posts. In each newspaper, we separately analyse two categories of news (domestic and international). A large majority of participants in each dataset was found to belong to one stable and dominant cluster present in all our datasets. We interpret it as comprising the "standard" or "average" discussion participants. The remaining participants could be viewed as various kinds of "anomalies" or "departures from normal" (not necessarily negative) and were assigned to several minor clusters. The shape and position of some minor clusters generalized well across the datasets. Overall, we found significant structural similarities between the four datasets in terms of histograms of attributes, the existence of one stable and dominant cluster, and the similar shape and location of several minor clusters. This is a significant result given that the four datasets were largely independent and the two newspapers adopted radically different policies for dealing with karma and foul language. The proposed approach therefore looks very promising as a means of identifying anomalous behaviour on diverse online discussion platforms.

Keywords: online discussion forum; social media analytics; data analysis; clustering; machine learning; user segmentation

1. Introduction

Modern social networking services and technologies like Facebook, Skype or ICQ effectively abolish borders and connect people around the world in real time. They represent the technological basis around which diverse online

* *E-mail address:* *{peter.krammer, marcel.kvassay}@savba.sk*

communities may form. Community members enter into various kinds of relationships, which in turn give rise to the rich internal structure and dynamics of social networks. Like networks of computers, social networks too can be studied with the mathematical tools of graph theory and network science.

One of the most intensely studied topics in network science is *community detection*. Here, the term *community* denotes a group of network nodes more densely connected to each other than to the rest of the network. Many clustering methods for their detection already exist [1], [11]. Their successful application to social networks, however, depends on the accuracy with which we identify and extract relationships among their members from available data. In some situations, the task is fairly straightforward. For example, social platforms like Facebook or Twitter provide means for users to indicate their preferences for others, e.g. by following them. Here, if A follows B and B follows A, we can reasonably surmise the existence of a deeper social tie between them, at least for the purpose of online community detection. For other tasks, even less information may suffice, e.g. Rowe et al. found out in [2] that the number of followers on Twitter ranked among the top three features determining whether a given user's tweets would elicit any response. In other contexts, e.g. on online forums that do not provide means to tag and follow other users, community detection can be very demanding and may require advanced natural language processing capability [3], whose quality and availability varies greatly from language to language.

Sometimes the primary focus of research is not on community detection, but rather on the roles and behaviours of communicating individuals regardless of the presence or absence of strong social ties among them. Thus, for example, Morrison et al. in [4] distinguished four overarching user roles: *popular users* who regularly contributed useful content eliciting replies; *ignored users* whose posts rarely elicited any reply; *joining conversationalists* who communicated intermittently with few others; and *elitists* who communicated intensely in small circles. These roles were defined on the basis of nine features formulated by Chan et al. in [5], which characterised users by various statistics, such as how many of their posts received replies, how many posts they produced per thread, how many threads they initiated, how many bi-directional neighbours they had, etc.

Administrators and moderators of online communities naturally have a vested interest in their smooth and productive functioning. Ideally, such communities should foster the positive potential of all their members. In practice it is often found that the anonymity of the net also encourages undesirable and antisocial behaviours [6]. One very frequent form is *trolling*. Summarising the perceptions of real users rather than academic theoreticians, Claire Hardaker in [7] defined *troll* as someone who constructs the identity of sincerely wishing to be part of the group in question, including professing, or conveying pseudo-sincere intentions, but whose real intention is to cause disruption and to trigger or exacerbate conflict for their own amusement. This definition, however, is highly inconvenient because of the onus (and difficulty) of objectively proving someone's trolling intention. Other researchers, therefore, preferred to search for more easily measurable and practicable criteria. As an example, Cheng et al. in [8] concentrated on so-called "future banned users" (i.e. prospective trolls) and managed to identify them on the basis of their first ten posts with 74% mean accuracy and 71% mean F1.

In this paper we build on our previous work [9] which explored online discussions in major Slovak national newspapers and mapped keywords to the most extensive ones. Our present focus is on the "typology" of discussion participants on these forums and to what extent it might generalise across different newspapers (*Sme.sk* vs. *Cas.sk*) and news categories (domestic news vs. international news). Our approach is akin to that of Morrison et al. in [4], except that we do not rely just on the communication patterns of discussion participants, but include some content-oriented features as well. The following section describes our data acquisition and preparation procedures.

2. Data Acquisition and Preparation

We crawled and extracted publicly available data from online discussion forums of two major Slovak national newspapers (*Sme.sk* and *Cas.sk*) for two article categories (domestic news and international news). Unfortunately, neither the traditional nor the latest state-of-the-art crawlers were directly usable for our purpose. Our crawler had to traverse paginated lists of news articles and their discussions and, at the end of each list, to recognise that there were no more pages to traverse, and finish. For that, the crawler had to be able to accept specific input parameters telling it how to paginate and how to recognise the end of the list, because these might differ for each data source. Such considerations eventually forced us to design and develop a dedicated crawler for the task.

This crawler accepts several regular expression-type arguments for recognizing the end of the list and extracting the title, URL and date of each article. Article URLs are then passed to *wget* software (www.gnu.org/software/wget) in order to retrieve their full text. In the next step, we extract the discussion post count for each article, if available, and select a subset of the most intensely discussed articles (those with the highest post counts). Their discussion posts are then crawled and processed. For discussion post crawling, our crawler accepts regular expression-type patterns for their date, nickname (ID) of their author and their full text. In this way we obtained four separate datasets, each covering online discussions for news articles from 2000 till 2017. We list their features in Table 1.

Each dataset can be represented as a homogeneous table in which one row corresponds to one discussion participant (identified by a unique *ID* or nickname) and each column stands for one of their characteristic attributes. Table 2 lists the most important original "raw" attributes extracted and calculated for each participant in each dataset. Although some discussion participants might have been operating under two or more distinct nicknames, we decided to accept that as a kind of noise in our data and did not try to identify and filter out such occurrences.

Table 1. An overview of our four news-related datasets.

Data source & category	Article count	Discussion count	Discussion post count	Nickname (ID) count
Cas.sk domestic	34 989	6 089	306 064	26 030
Cas.sk international	67 863	6 078	457 843	29 958
Sme.sk domestic	173 488	1 500	968 698	34 998
Sme.sk international	144 333	1 500	489 238	22 764

Table 2. List of attributes for each online discussion participant. Their values are summed over all discussions in a given dataset.

Attribute Name	Attribute Description
ID	Unique identifier of a given discussion participant ("the user") in a given dataset.
Posts	Total number of the user's posts in the dataset
Hoax_count	Total number of hoax links (i.e. URLs referring to websites classified as controversial at *www.konspiratori.sk/en/*) in the user's posts
Vulg_count	Total number of expletives in the user's posts
Words_count	Total number of words in the user's posts
Kval_Positive	The user's positive karma (cumulative total)
Kval_Negative	The user's negative karma (cumulative total)
Other_reacted_to_me	Number of distinct participants who responded to this user (totalled across all discussions in the dataset)
Days_active	Number of days elapsed between the first and the last post of this user in the dataset
Posts_violated_codex_cnt	Number of the user's censored posts (at *Sme.sk*)

Note: Regarding expletives, our approach depended on the data source. *Cas.sk* is relatively benevolent and expletives could be found in its discussions by plain matching. *Sme.sk*, on the other hand, strictly censors foul language, so we had to come up with a different approach. When a post on *SME.sk* violates the rules, its text is replaced by a notice. Guessing that this happens mostly due to foul language, we counted such notices (*Posts_violated_codex_cnt*) and used that as an approximation for the count of expletives (*Vulg_count*). Despite this ad-hoc improvisation we observed high similarity between *Sme.sk* and *Cas.sk* regarding foul language (see Fig. 3).

It would be a mistake, however, to use the attributes from Table 1 directly to compare the participants, because they only carry "raw" information in the form of absolute numbers without regard to how long each of them had been active or how many posts he or she managed to produce during that time.

A forum member who had been active for ten years is likely to have produced a far greater number of posts, words, URLs, etc. than another user of the same type who has started just recently. In order to detect their similarity, we have to transform these absolute attributes into relative ones, i.e. to divide them by the total number of the user's posts or by their days of activity or by some other appropriate quantity. Because there were several options open to us in this regard, we could produce a great variety of relative attributes and then select those that turned out to be the most effective and relevant for the task at hand. When we applied this process to our data, we finally arrived at the set of most effective relative attributes listed in Table 3.

In the next step we tried to reduce various detrimental influences and side-effects in our data. This primarily consisted in removing the participants with less than a pre-specified number of posts in the dataset from further analysis. There were two reasons for this. First, such “reticent” participants were not sufficiently characterised by their infrequent posts and, second, the calculation of relative attributes for them often lead to certain numerical values, such as 0.25, 0.33, 0.5, 0.75, etc. which manifested as dense horizontal or vertical lines in the corresponding scatter plots, such as the one shown in Figure 1a. At first these lines gave us the impression of carrying valuable information (and even managed to confuse our clustering tools), but upon closer scrutiny we found them to be just artefacts caused by the paucity of posts from the “reticent” participants.

Table 3. List of most effective relative attributes for online discussion participants

Attribute Name	Calculation Definition	Attribute Description
rel_vulg	Vulg_count / Posts	Average number of expletives per the user’s post
rel_word	Words_count / Posts	The user’s average post length.
rel_hoax	Hoax_count / Posts	Average number of hoax URL links per the user’s post
rel_kval	(Kval_Positive - Kval_Negative) / Posts	Average karma per the user’s post
rel_post_per_day	Posts / (Days_active + 1)	Average number of the user’s posts per day for a given dataset (corrected against division by zero)
rel_react_to_me	Other_reacted_to_me / Posts	The user’s response elicitation factor: the number of distinct discussion participants who responded to this user, divided by the total number of his or her posts in the dataset
rel_violation	Posts_violated_codex_cnt / Posts	Ratio of the user’s censored posts to all his or her posts.

We have therefore decided to analyse only the participants for whom we had at least 27 posts in the dataset. We determined this limit empirically: we wished to keep as many users as possible so that our clustering results were representative, and this value turned out to be the smallest one for which the linear artefacts observed in Figure 1a effectively disappeared. After removing the “reticent” participants as well as approximately ten others, who were obvious outliers, we obtained a much smoother Figure 1b.

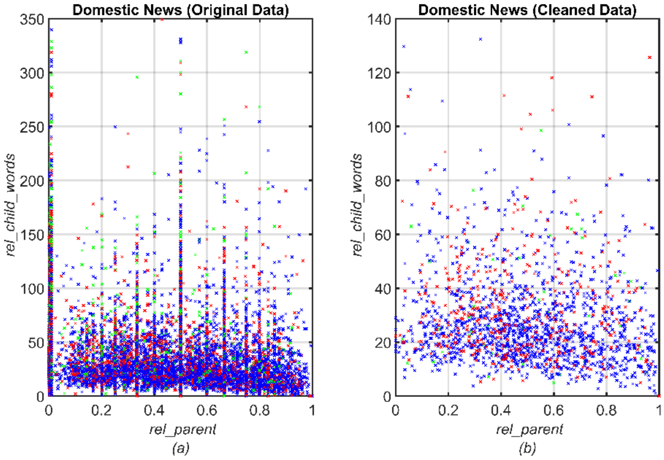

Fig. 1. Dependence of the average “child” post length (*rel_child_word*) on the user’s “parent” post ratio (*rel_parent*) from *Cas.sk*. Data point colour represents the user’s gender (blue=Male, red=Female, green=Unknown). Data cleaning removed the linear artefacts and, by removing the outliers, also reduced the vertical scale.

User and post counts for our datasets, both before and after the cleaning, are listed in Table 4. We can see that while the user counts dropped rather dramatically (by one order of magnitude), the drop in the post counts was relatively modest (about 20% for international news and about 33% for domestic news). This is because we eliminated the reticent participants.

Table 4. User counts and post counts in the original and the cleaned datasets

Original Data Sets	User Counts (Cas.sk)	Post Counts (Cas.sk)	User Counts (Sme.sk)	Post Counts (Sme.sk)
Domestic News	29 958	306 064	34 998	968 698
International News	26 030	457 843	22 764	489 238
Cleaned Data Sets	(Cas.sk)	(Cas.sk)	(Sme.sk)	(Sme.sk)
Domestic News	1 885	206 660	5 773	776 501
International News	2 201	368 337	3 522	370 072

The next step in data pre-processing was to standardize them so that the mean value of each attribute became zero with standard deviation equal to one. This levelled the playing field and equalized their chances for attribute selection. Alternatively, we might have simply normalized the data, i.e. to project them into the interval <0, 1>, but this would be driven exclusively by the minimum and the maximum value of each attribute, which might well belong to outliers. Standardization -- by taking into account the overall distribution of attribute values -- appealed to us as a more prudent and suitable choice.

3. Data Analysis

3.1. Comparison of Datasets

As a preliminary to clustering, we visually compared sample density in our datasets in various two-dimensional projections, hoping to find common shapes and outlines indicative of deeper structural similarities. We quickly managed to identify a group of features (*rel_post_per_day*, *rel_react_to_me*, and *rel_word*) exhibiting significant similarities across two or more datasets. Figure 2 shows the scatter plots for the pair (*rel_react_to_me*, *rel_word*). Considering that our datasets were not of equal size, the likeness of their spatial distribution was quite striking.

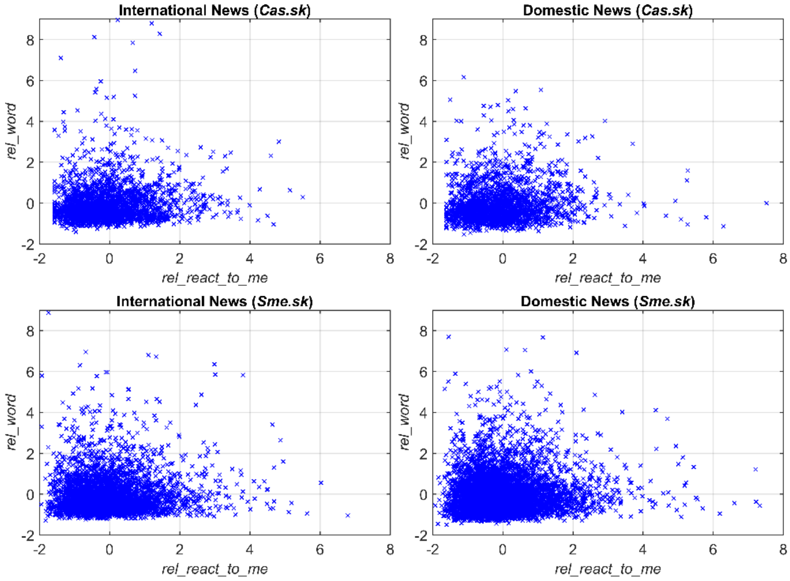

Fig. 2. Scatter plots of our datasets for the feature pair (*rel_react_to_me*, *rel_word*)

In the bottom left portion of all the charts there is an irregular oval (or blob) with maximum sample density. It is three to four standard deviations wide and two to three deviations high, with the center of gravity near the origin (i.e. the mean). Its bottom border is quite sharp, but its top and right borders are distinctly blurred. In fact, as we move away from the origin along the main diagonal, at about (1, 1) we seem to enter a strip of lesser but relatively even density, which extends roughly up to the antidiagonal and is parallel to it. Finally, barring a few isolated samples hovering just above the antidiagonal, the top right portion of all the charts is empty. This would seem to signify that longish posts have negligible chance to elicit replies from many distinct people.

We observed the same measure of likeness for the pair (*rel_react_to_me*, *rel_post_per_day*), but do not show the scatter plots here in order to save space. We were extremely delighted to see the likeness between the scatter plots which used the pair (*rel_react_to_me*, *rel_vulg*) for *Cas.sk* and its approximation in the form of (*rel_react_to_me*, *rel_violation*) for *Sme.sk*. We show these scatter plots in Fig. 3. Their likeness means that we can take *rel_violation* as equivalent to *rel_vulg* for analytical purposes.

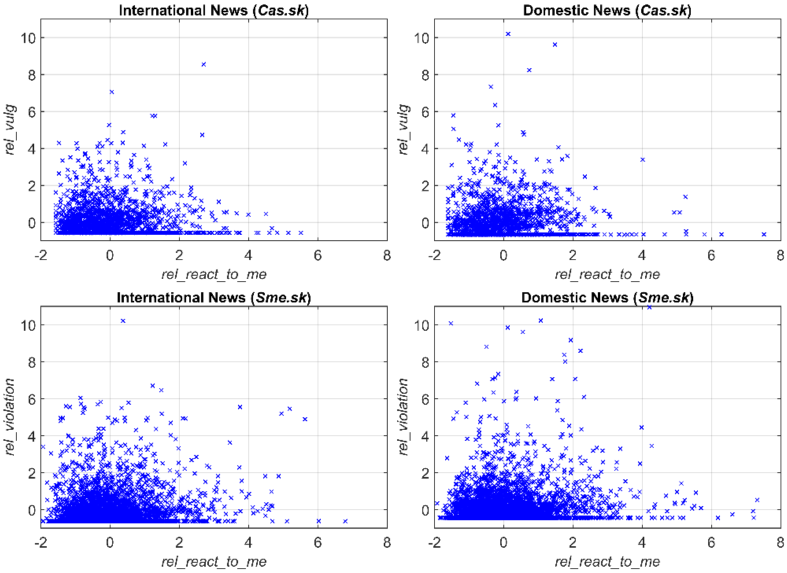

Fig. 3. The top row shows scatter plots for *Cas.sk* and the feature pair (*rel_react_to_me*, *rel_vulg*), while the bottom one for *Sme.sk* and the approximated feature pair (*rel_react_to_me, rel_violation*)

Unfortunately, not all the attributes exhibited such striking similarities across our datasets. Notable among them was the average karma per post (*rel_kval*): *Sme.sk* uses a proprietary algorithm which differs significantly from the simple summation of user up-votes and down-votes employed by *Cas.sk*. Because karma is the second of only two attributes that characterize each participant indirectly through the reactions of others, it made our work of searching for a common clustering structure much more difficult. Nevertheless, even without karma we obtained encouraging results which we describe in the following subsections.

3.2. Feature Selection for Clustering Analysis

In feature selection for clustering we have considered several aspects:

- We avoided using highly correlated attributes. (High correlation was typical for relative attributes derived from the same base attribute.)
- We preferred attributes describing different facets of the participants' posts (their average length, the presence of foul language, hoax links, etc.) or of their communication patterns (their posting frequency, how many unique people ever responded to them, etc.).
- We also preferred attributes whose characteristics (histograms) were similar between two or more datasets.

We relied on these criteria and on visual inspection of scatter plots in selecting a small group of the most effective attributes listed in Table 2. Of these, we finally dropped *rel_kval*, because *Sme.sk* defined karma in a way that was incompatible with that of *Cas.sk*.

Since we were dealing with real-life data of human online behaviour, we expected that our samples (representing individual discussion participants) would be distributed over the attribute space relatively evenly, perhaps with varying density but without obvious wide margins or clear delimitations. In fact, any such clear borders would more likely indicate that our data lacked representativeness or were contaminated with noise or somehow distorted, as in Fig. 1a. These concerns and considerations prompted us to rely on density-based clustering.

3.3. Clustering Parameters and Main Results

Clustering was performed along five dimensions or features. (Table 3 lists seven, but we dropped *rel_kval* and used *rel_violation* only for *Sme.sk* as an equivalent of *rel_vulg* in *Cas.sk*.) We used Canopy clusterer [10] in Weka with the following settings: Periodic Pruning Rate = 10000, maximum Number of Canopies in Memory = 100. T1 and T2 radii as well as the number of clusters were determined heuristically by the algorithm itself. In our clustering experiments, we manually varied Minimum Canopy Density and random seeds.

As we increased Minimum Canopy Density from 1.0 to 10.0 for *Sme.sk International news* dataset, the number of clusters decreased from 13 to 4 as shown in Table 5. The most important finding was the presence of one dominant and stable cluster comprising around 92% of samples, whose size remained almost constant throughout the process. However, Minimum Canopy Density strongly affected the number and the shape of the remaining minor clusters: with its decrease they would split into smaller ones and, on occasions, a brand new one would emancipate itself at the edge of the dominant cluster.

Table 5. Clustering results for *Sme.sk* from International News dataset (left side) and Domestic News dataset (right side).

Minimum Canopy Density	# identified clusters in International news dataset	Relative size of the dominant cluster	Number of samples in the dominant cluster	# identified clusters in Domestic news dataset	Relative size of the dominant cluster	Number of samples in the dominant cluster
1.0	13	92%	3 224	12	96%	5 548
1.5	10	92%	3 224	8	96%	5 548
2.0	10	92%	3 224	8	96%	5 548
2.5	9	92%	3 224	7	96%	5 548
3.5	7	92%	3 227	7	96%	5 548
4.5	5	92%	3 241	6	96%	5 565
5.5	5	92%	3 241	6	96%	5 565
6.5	5	92%	3 241	6	96%	5 565
8.0	5	92%	3 241	6	96%	5 565
10.0	4	93%	3 283	5	97%	5 612

Analogous and even more stable results were observed for *Sme.sk Domestic news* dataset (also shown in Table 5). Clustering of *Cas.sk* along the same five dimensions produced similar results: the dominant cluster contained 90% of samples in *Cas.sk Domestic news* and 86% in *Cas.sk International news.* Moreover, the effect of varying Minimum Canopy Density from 1.0 to 10.0 on its relative size did not exceed 1 percentage point. High similarity was also observed when we repeated the clustering exercise in four dimensions after removing *rel_hoax* attribute.

To sum up, all four datasets were found to contain one dominant cluster with stable size and well-defined borders. We interpret it as comprising the “average” or “normal” discussion participants. The remaining 7 to 14% of dataset samples, characterised by significantly higher values of some of their attributes, were unevenly scattered around the dominant cluster and could be modelled as belonging to several minor or secondary clusters. We could view them as “atypical” or even “anomalous” in a way, if we can keep that designation free of negative connotations: while the participants who slip into foul language too readily are clearly undesirable, those with unusually high karma are likely beneficial for the forum, provided that their karma scores are genuine.

3.4. Minor Clusters

Our application of density-based Canopy clusterer to selected relative attributes of online discussion participants could be viewed as a method for detecting anomalies, because it was able to discriminate the “typical” or “standard” participants from the “atypical” or “anomalous” quite robustly. It was, however, considerably less robust when trying to distinguish various kinds of anomaly: Table 5 testifies that the number (and the location) of minor clusters was sensitive to Minimum Canopy Density. Both varied with random seeds too. This, we believe, was due to the limited quantities of atypical samples *vis-à-vis* the “standard” ones. Despite these problems, some minor clusters turned out to be similar across two or more datasets, as shown in Figures 4 and 5.

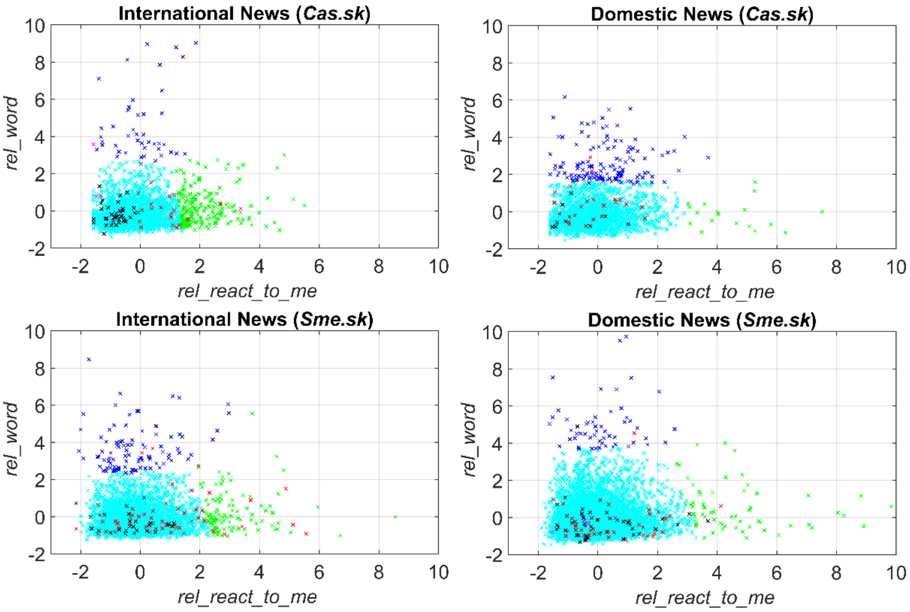

Fig. 4. Cluster visualisations for our datasets in two-dimensional projection (*rel_react_to_me, rel_word*). The cyan cluster is the dominant one, the dark blue one represents the authors of longish posts, the green one the “popular” participants with high response elicitation factor. Other minor clusters are not clearly visible here.

Although our datasets were largely independent and composed of different samples, both figures show similar spatial arrangement of clusters in all of them. Some minor clusters are easily separable in the chosen two-dimensional projections, while others are not. The reason is that the clustering was performed along five dimensions of which only two could be shown in each chart. In all the charts, the cyan cluster is the dominant one. The samples of other (minor) clusters were “preferentially” displayed in the forefront so as to remain visible.

The clusters shown in Figures 4 and 5 correspond to Minimum Canopy Density=4.5, which we found preferable as it did not lead to the formation of too many minor clusters. Moreover, their similarity across datasets was quite obvious. In each dataset we repeated the clustering five times with different random seeds and obtained broadly similar results: the borders of the dominant cluster have hardly changed, and though we did observe some re-shuffling among minor clusters, most of them kept their location and approximate size. We chose that value of random seed as final for which we obtained the best consistence in the assignment of samples to clusters.

Overall, the most significant result for us was the fact that the similarity in the shape and the location of clusters generalised well across different servers (*Sme.sk* vs. *Cas.sk*) despite their differing policies regarding foul language (*rel_vulg* vs. *rel_violation*). We therefore expect that other discussion forums might also exhibit such similarities in the distribution of their attribute values, in the existence of one dominant cluster and in the gradual decrease of sample density with increasing distance from it, or in the similar size, shape and location of minor clusters. This would also imply the applicability of our method of “anomaly” detection to them.

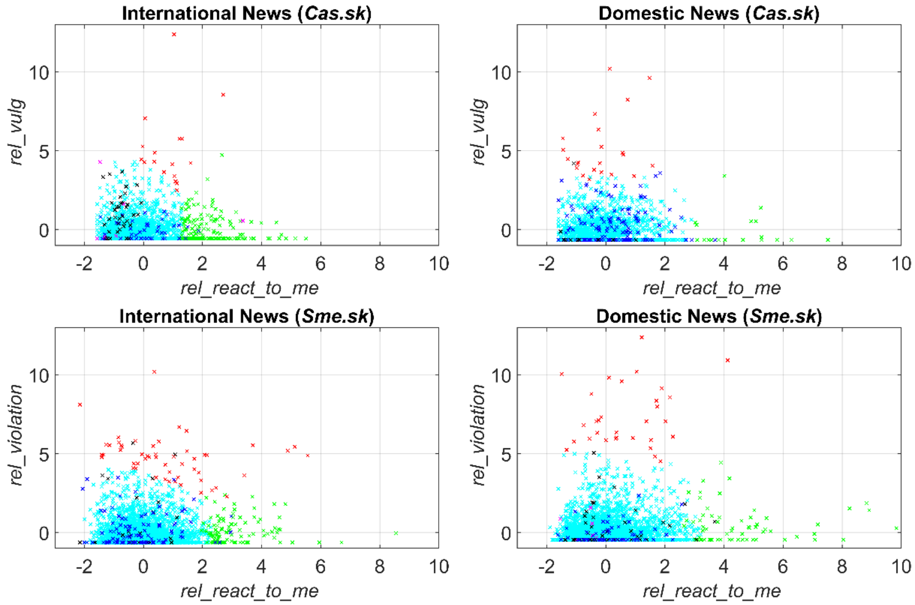

Fig. 5. Cluster visualisations in two-dimensional projection (*rel_react_to_me, rel_vulg*) for *Cas.sk* and in (*rel_react_to_me, rel_violation*) for *Sme.sk*. The cyan cluster is the dominant one, the red one represents the "foul-mouthed" participants, the green one the "popular" participants. Other minor clusters are not clearly visible here.

For evaluation, we used DBScan [12], a Density Based clusterer, with more than 80 different parameter settings. In all the cases, it confirmed the existence of one dominant cluster with 92% or more samples. The remaining samples were distributed among several tiny clusters and noise (a notion which DBScan supports). The number of these tiny clusters strongly depended on DBScan parameter settings (roughly in line with Table 5). A similar problem arose in [13], where multiple clustering methods gave different solutions, which the authors consolidated into a hierarchy of clusters using their domain knowledge. Additionally, the dominant cluster in our datasets was also correctly detected by the hierarchical clusterer with Complete link type [14], which assigned to it more than 95% of samples.

For our future work we envisage two promising methods of determining the number and shape of our minor clusters more robustly and consistently:

- After reliably determining the stable "inhabitants" of the dominant cluster, we would remove them from the dataset and repeat the clustering only for the remaining "anomalous" samples. In this way the standard samples would not influence the clustering, thus giving the clusterer a greater freedom to determine the best possible borders between the minor clusters.
- Alternatively, we could also choose such values of Minimum Canopy Density and random seed that would lead to the formation of numerous minor clusters, and then merge the neighbouring ones manually (if they are significantly similar) until we end up with the desired or manageable number of final clusters.

Conclusions

In this paper we performed clustering analysis of online discussion participants from two major Slovak national newspapers, *Sme.sk* and *Cas.sk*, on the basis of language-independent statistical attributes characterising their communication patterns and the content of their posts. In each newspaper, we separately analysed two categories of news -- domestic news and international news.

We found out that a large majority of participants in each dataset belonged to one stable and dominant cluster. We interpret it as comprising the "standard" or "typical" discussion participants. The same dominant cluster was identified after removing one of the clustering features (*rel_hoax*), which we consider a proof of its stability. Sample density in this dominant cluster gradually decreases as we move away from its centre, which makes the visual detection of its borders a bit problematic. However, density-based Canopy clusterer was able to identify its borders quite precisely and robustly.

In the vicinity of the dominant cluster there were areas with somewhat lower and uneven sample density. Discussion participants residing there could be viewed as “atypical”, “anomalous” or “extreme” in a sense, but this should not be construed as something negative. In fact, some “extremes” might be highly desirable and beneficial for the forum (e.g. extremely high karma, as long as it is genuine). Despite their small proportion (7% to 14% of our datasets), we succeeded in modelling and assigning such “anomalous” participants to several minor clusters, and some of those clusters generalised well across two or more datasets.

Overall, we found significant structural similarities between the four datasets in terms of histograms of attributes, the existence of one stable and dominant cluster, and the similar shape and location of several minor clusters. This is a significant result given that the four datasets were largely independent and the two newspapers adopted radically different policies for dealing with karma and foul language. The proposed approach therefore looks very promising as a means of identifying anomalous behaviour on diverse online discussion platforms. In the future we plan to apply these methods to other large datasets and analyse the structure and behaviour of minor clusters in more detail.

Acknowledgements

This work was supported by the projects VEGA 2/0167/16, VEGA 2/0154/16 and PROCESS EU H2020-777533.

References

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

[2] Rowe M, Angeletou S, Alani H. Predicting discussions on the social semantic web. In Extended Semantic Web Conference 2011 May 29 (pp. 405-420). Springer, Berlin, Heidelberg.

[3] Goldberg Y. A Primer on Neural Network Models for Natural Language Processing. J. Artif. Intell. Res.(JAIR). 2016 Jan 1; 57:345-420.

[4] Morrison D, McLoughlin I, Hogan A, Hayes C. Evolutionary Clustering and Analysis of User Behaviour in Online Forums. In ICWSM 2012 Jun 4.

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

[6] Suler J. The online disinhibition effect. Cyberpsychology & behavior. 2004 Jun 1; 7(3):321-6.

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

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

[9] Mojžiš J, Budinská I. Analyzing news articles from the side of discussion threads. In Intelligent Engineering Systems (INES), 2017 IEEE 21st International Conference on 2017 Oct 20 (pp. 297-300).

[10] A. McCallum, K. Nigam, L.H. Ungar: 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 ACM-SIAM symposium on Discrete algorithms, 169-178, 2000.

[11] Monowar Hussain Bhuyan, Dhruba K. Bhattacharyya, Jugal K. Kalita: Towards an Unsupervised Method for Network Anomaly Detection in Large Datasets, Computing and Informatics, Vol 33, no 1 (2014).

[12] Ester, Martin, et al. "A density-based algorithm for discovering clusters in large spatial databases with noise." Kdd. Vol. 96. No. 34. 1996.

[13] Chan, Jeffrey, Conor Hayes, and Elizabeth M. Daly. "Decomposing Discussion Forums and Boards Using User Roles." ICWSM 10 (2010): 215-218.

[14] Defays, Daniel. "An efficient algorithm for a complete link method." The Computer Journal 20.4 (1977): 364-366.