INES 2018

IEEE 22nd International Conference
on
Intelligent Engineering Systems

PROCEEDINGS

June 21-23, 2018
Las Palmas de Gran Canaria, Spain

Organized by

Óbuda University, Budapest, Hungary

Sponsored by

IEEE Hungary Section
IEEE Computational Intelligence Chapter, Hungary
IEEE Joint Chapter of IES and RAS, Hungary
IEEE SMC Chapter, Hungary

Technical Co-sponsor

IEEE Industrial Electronics Society

Venue

Elder Museum

Part Number: CFP18IES-USB (pendrive);
ISBN: 978-1-5386-1121-0 (pendrive)

WELCOME MESSAGE FROM THE GENERAL CHAIRS

On behalf of INES 2018 Committees, it is my pleasure to welcome you to the *22nd International Conference on Intelligent Engineering Systems* (INES) to be held in Las Palmas de Gran Canaria, Spain.

The growing international competition in the industrial arena has created a demand for the introduction of intelligent techniques to various industrial problems to improve product quality and production process efficiency, as well as reduce production costs. The aim of the INES conference series is to provide researchers and practitioners from industry and academia with a platform to report on recent developments in the area of computational intelligence. INES 2018 focuses on the application of state-of-the-art intelligent techniques to engineering systems. The conference location is Elder Museum.

I would like to acknowledge the efforts of the Technical Program Chairs, the Organizing Committee Chair, the Technical Program Committee members and all those persons responsible for the background activities from local arrangements to conference secretariat. I also want to thank many volunteers who have contributed lots of time and effort to bring INES 2018 to you. My pleasant duty is to gratefully acknowledge the support provided by the sponsors of the conference: IEEE Industrial Electronics Society, IEEE Hungary Section, IEEE Joint Chapter of IES and RAS, Hungary, IEEE Computational Intelligence Chapter, Hungary, IEEE SMC Chapter, Hungary. I hope that all in attendance at INES 2018 will find this event intellectually stimulating and professionally rewarding.

Imre J. Rudas
Óbuda University, Budapest, Hungary

Alexis Quesada-Arencibia
IUCTC, ULPGC, Spain

INES 2018 General Chairs

Table of Contents

Committees

INES 2018 HONORARY CHAIR

***Lotfi A. Zadeh*†,** USA

INES 2018 HONORARY COMMITTEE

Bogdan M. Wilamowski, Auburn University, AL, USA
Mihály Réger, Óbuda University, Budapest, Hungary
Levente Kovács, IEEE Hungary Section chair

INES 2018 GENERAL CHAIRS

Imre J. Rudas, Óbuda University, Budapest, Hungary
Alexis Quesada-Arencibia, IUCTC, ULPGC, Spain

INES 2018 TECHNICAL PROGRAM COMMITTEE CHAIRS

Levente Kovács, Óbuda University, Budapest, Hungary
Rudolf Andoga, Technical University of Košice, Slovakia

INES 2018 TECHNICAL PROGRAM COMMITTEE

Michael Affenzeller, University of Applied Sciences FH Upper Austria, Hagenberg, Austria
Mari Carmen Aguayo Torres, Universidad de Malaga, Spain
Ito Atsushi, Utsunomiya University, Japan
Werner Backfrieder, University of Applied Sciences FH Upper Austria, Hagenberg, Austria
Patricio García Báez, Universidad de La Laguna
Monika Bakosová, Slovak University of Technology, Slovakia
Valentina Balas, "Aurel Vlaicu" University, Arad, Romania
Raquel Barco, Universidad de Malaga, Spain
Ildar Batyrshin, Mexican Petroleum Institute, Mexico
Barnabás Bede, DigiPen, Seattle, USA
Attila L. Bencsik, Óbuda University, Budapest, Hungary
Balázs Benyó, BME, Hungary
Manuel Berenguel, Catedrático de Ingeniería de Sistemas y Automática de la Universidad de Almería, Spain
Uwe Borghoff, Universität der Bundeswehr München, Germany
Wojciech Bozejko, Wroclaw University os Science and Technology, Poland
Agostino Bruzzone, University of Genova, Italy
Klaus Buchenrieder, Universität der Bundeswehr München, Germany
Zenon Chaczko, UTS, Sydney, Australia
Heinz Dobler, University of Applied Sciences FH Upper Austria, Hagenberg, Austria
Eck Doerry, Northern Arizona University, USA
György Eigner, Óbuda University, Budapest, Hungary
Ladislav Főző, Technical University of Košice, Slovakia
Manuel Canton Garbin, Almeria University, Spain
Carlos Godfrid, Universidad de Buenos Aires, Argentina
Gábor Hegedűs, Óbuda University, Budapest, Hungary
Clemens Holzmann, University of Applied Sciences FH Upper Austria, Hagenberg, Austria
László Horváth, Óbuda University, Budapest, Hungary
Junichi Iijima, Tokyo Institute of Technology, Japan
Gerhard Jahn, University of Applied Sciences FH Upper Austria, Hagenberg, Austria
Karel Jezernik, University of Maribor, Slovenia
Zsolt Csaba Johanyák, Kecskemét College, Hungary
Bertold Kerschbaumer, University of Applied Sciences FH Upper Austria, Hagenberg, Austria
George Kovács, CAI of HAS, Hungary
Szilveszter Kovács, University of Miskolc, Hungary
Krzysztof Kozlowski, University of Poznan, Poland
Werner Kurschl, University of Applied Sciences FH Upper Austria, Hagenberg, Austria
Krisztián Lamár, Óbuda University, Budapest, Hungary
Francesco Longo, University of Calabria, Italy
Galina Merkurjeva, Riga Technical University, Latvia

Dimitris Metaxas, Rutgers University
Alajos Mészáros, Slovak University of Technology, Bratislava, Slovakia
Jan Nikodem, Wroclaw University os Science and Technology, Poland
Péter Pausits, Óbuda University, Budapest, Hungary
Béla Pátkai, Tampere University of Technology, Finland
Emil M. Petriu, University of Ottawa, Canada
Miguel Ángel Piera, Universitat Autònoma de Barcelona
Radu-Emil Precup, „Politehnica" University of Timisoara, Romania
Stefan Preitl, „Politehnica" University of Timisoara, Romania
Ales Prochazka, University of Chemistry and Technology & Czech Technical University, Czech Republik
Octavian Prostean, „Politehnica" University of Timisoara, Romania
Ewaryst Rafajlowicz, Wroclaw University os Science and Technology, Poland
Tamás Révai, University National of Public Service, Budapest, Hungary
Jerzy Rozenblit, University of Arizona, Tucson, USA
Andrzej Rys, Kansas State University
Stefano Saetta, University of Perugia, Italy
Jose Sigut, Universidad de La Laguna, Teneryfe
Czeslaw Smutnicki, Wroclaw University os Science and Technology, Poland
János Somló, Óbuda University, Budapest, Hungary
Carmen Paz Suarez Araujo, ULPGC, Spain
Miroslav Sveda, Brno University of Technology, Czech Republik
Sándor Szénási, Óbuda University, Budapest, Hungary
Gábor Szögi, Óbuda University, Budapest, Hungary
Árpád Takács, Óbuda University, Budapest, Hungary
Pavol Tanuska, Slovak University of Technology, Slovakia
József K. Tar, Óbuda University, Budapest, Hungary
Jose Antonio Tenreiro Machado, Institute of Engineering of Porto, Portugal
Annamária R. Várkonyi-Kóczy, Óbuda University, Budapest, Hungary
Mladen Vouk, Northern Carolina University
Jozef Živčák, Technical University of Košice, Slovakia

INES FINCANCE CHAIR

Anikó Szakál, Óbuda University, Budapest, Hungary

INES 2018 ORGANIZING COMMITTEE CHAIR

Tamás Haidegger, Óbuda University, Budapest, Hungary

INES 2018 ORGANIZING COMMITTEE

József Gáti, Óbuda University, Budapest, Hungary
Franciska Hegyesi, Óbuda University, Budapest, Hungary
Gyula Kártyás, Óbuda University, Budapest, Hungary
Krisztina Némethy, Óbuda University, Budapest, Hungary

INES 2018 LOCAL ORGANIZING COMMITTEE

Gabriel De Blasio, IUCTC, ULPGC, Spain
Carmelo Ruben García-Rodriguez, IUCTC, ULPGC, Spain
Roberto Moreno-Díaz Jr., IUCTC, ULPGC, Spain
José Carlos Rodriguez-Rodriguez, IUCTC, ULPGC, Spain

INES SERIES LIFE SECRETARY GENERAL

Anikó Szakál
Óbuda University, Budapest, Hungary
E-mail: szakal@uni-obuda.hu

PROCEEDINGS EDITOR

Anikó Szakál
Óbuda University, Budapest, Hungary

PRODUCTION PUBLISHER

IEEE Hungary Section

Crawling and Analysis of Online Discussions in Major Slovak National Newspapers

J. Mojžiš*, P. Krammer*, M. Kvassay*, I. Budinská*, L. Hluchý*, M. Jurkovič**
* Institute of Informatics SAS, Bratislava, Slovakia
** Center of Social and Psychological Sciences - Institute of Experimental Psychology SAS, Bratislava, Slovakia
*[jan.mojzis, peter.krammer, marcel.kvassay,
budinska.ui, ladislav.hluchy]@savba.sk
**marek.jurkovic@savba.sk

***Abstract*— In this paper we compare the structure of online discussions in two major Slovak national newspapers (*Sme.sk* and *Cas.sk*) for two categories, domestic news and international news. We also perform clustering analysis for one of them (*Cas.sk*). We visualize online discussions through participant profiles, which combine content-oriented features, such as the number of expletives or URL links in the participants' posts, with communication-oriented features, such as how many distinct people ever responded to them. Using five or six most effective features, we identify five relatively stable clusters on the analyzed domain, four of which generalize across the two categories. Each cluster represents a group of discussion participants with similar characteristics and could be interpreted as a kind of participant type or role. We deliberately defined our features in a language-independent way so that they are directly usable in other languages besides Slovak.**

I. INTRODUCTION

World Wide Web is the source of vast quantities of information and continues to grow rapidly. This phenomenon is particularly striking in the area of social networking websites, but other segments of the internet witness similar growth. In our previous work [1] we mapped the evolution of online editions of several Slovak national newspapers (Sme.sk, Cas.sk and Pravda.sk) in the domestic news category. While in 2004 there were about 50,000 domestic news articles, by 2015 their quantity multiplied five times to more than 250,000 and in 2017 their number reached 310,000.

While news articles typically follow certain ethical conventions, fewer restrictions tend to apply to their accompanying comment and discussion sections. Some online forums are quite liberal with respect to profanity and external URL links, while others try to filter these out and sometimes even impose further restrictions. In general, it is in the best interest of forum owners, administrators and moderators to keep their forums in "good health" and free of any antisocial activity. Many forum software packages therefore include some form of automatic filtering of foul language and other prohibited content. There are, however, other tasks that seem to require an almost human-level capability of judgment. To tackle them successfully, more sophisticated analytical approaches are needed. These typically rely either on the content of the posts or on the communication patterns of their authors, or on a combination of both.

For example, Morrison et al. in [2] used ego-centric reply-graphs of forum users to gather nine relevant features (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 neighbors they had, etc.) and on that basis managed to identify four main user roles or clusters: *popular initiators* who regularly contributed useful content eliciting many replies; *ignored users* whose posts rarely elicited any reply; *joining conversationalists* who communicated intermittently with few others; and *elitists* who communicated intensely in small, relatively closed circles. The authors also demonstrated that these roles were not fixed but could change over time.

Of particular concern to online forum moderators and administrators is antisocial behavior, which can take many forms [3]. One very frequent form is *trolling*. In common parlance it is associated with an intention to deceive and disrupt the affected group. As Claire Hardaker put it in [4], a troll is someone who constructs the identity of sincerely wishing to be part of the group in question, but whose real intention is to cause disruption and to trigger or exacerbate conflict for their own amusement. Other researchers, however, noted the difficulty of objectively proving the trolling intention and searched for more practicable criteria. A good example is Cheng et al. in [5, 6], whose effort "paid off" by enabling them to identify so-called "future banned users" (FBU's or prospective trolls) on the basis of their first ten posts with 74% mean accuracy and 71% mean F1.

Online discussion forums attract other kinds of academic research as well. For example, Wood and Douglas [7] tested several hypotheses concerning the differences between the supporters and the opponents of conspiracy theories, and Tan et al. [8] studied persuasiveness of arguments in online discussions.

In our previous work [9] we analyzed online discussions in major Slovak national newspapers and mapped keywords to discussions with the highest numbers of posts. In this article we extend our analysis in a new direction: we investigate whether there is a stable structure or "typology" of discussion participants on these forums and, if so, to what extent it generalizes across different newspapers (*Sme.sk* vs. *Cas.sk*) and categories (domestic news vs. international news). Our approach is akin to that of Morrison et al. in [2], although we do not rely exclusively on the communication patterns of discussion participants, but also include some rudimentary content-

978-1-5386-1122-7/18/$31.00 ©2018 IEEE

oriented features. In the next section we provide more details about the process of obtaining and preparing our data for analysis.

II. DATA PROVISION AND PREPARATION

A. Article and Discussion Crawling

Quite early in the process of exploring our data sources we realised that neither the traditional nor the latest state-of-the-art crawlers [10, 11, 12] were directly usable for our purpose. Our crawler had to be able to traverse iteratively through 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 stop. Thus 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 datasource. Such considerations finally forced us to design and develop a dedicated crawler for the task.

Our crawler accepts several regular expression-type arguments that enable it to recognize the end of the list and to extract the title, URL and date of each news article. Article URLs are then passed to *wget*[1] in order to retrieve their full text. For discussion post crawling, the crawler accepts analogous patterns for their date, nickname (*ID*) of their author and their full text.

B. Data Sources

Our data was extracted from two major Slovak national newspapers (*Sme.sk* and *Cas.sk*) for two article categories (domestic news and international news). Both newspapers were ranked among the most popular in terms of real users by AIMmonitor[2]. We collected the data from the year 2000 to 2017. In order to get the data, we could have used the Common Crawl (CC) service[3] but we found out that their datasets were rather incomplete. They only contained about one third of the existing news articles. This forced us to crawl the concerned newspaper websites directly in order to get the articles and their associated discussions. For each article, we extracted the discussion post count, if it was available. We next selected a subset of most intensely discussed articles, i.e. those with the highest post counts. Their discussion posts were then extracted and processed. The resulting base dataset characteristics are listed in Table I.

TABLE I
AN OVERVIEW OF OUR NEWS DATASETS

Data source & category	# articles	# discussions	# discussion posts	# participant profiles
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

[1] www.gnu.org/software/wget

[2] aimmonitor.sk

[3] commoncrawl.org

C. Basic Features for User Profiling

The data we collected enabled us to create a profile for each discussion participant (identified by his or her *ID*). Like Morrison et al. in [2], we too relied mainly on simple statistics derived from their communication patterns, but we also included a few rudimentary content-oriented features, such as the average number of expletives or URL links in their posts.

The process of creating participant profiles was as follows. First, we retrieved the base attributes of each post, i.e. its *date*, *fulltext*, *karma* and its author's *ID* (nickname). Once we had all the posts, we calculated the total word and post counts for each participant (*words_count*, *posts*). We also counted all expletives (*vulg_count*) and URL links (*links_count*) in these posts. Among URL links, we separately counted those pointing to controversial content (*hoax_count*). For this we relied on a public database of controversial websites at *konspiratori.sk*[4]. Having participants' *ID*s, we could query their forum profiles and retrieve their *gender* (if revealed). This group of features then represented the "content-oriented" aspect of participant profiles.

Their "communication-oriented" aspect was derived from the structure of the discussion threads that they joined, although there were differences in the level of detail available to us at different data sources. In general, each post is a reaction either to another post (then we call it a *child* post), or to the news article at the head of the discussion (then we call it a *parent* post). In this respect we followed the coding methodology of Wood and Douglas in [7]. By collecting the data on who reacted to whom, we were able to characterise the participants by the number of distinct (unique) users to whom they responded (*i_reacted_to*) as well as by the complementary number of distinct users who responded to them (*other_reacted_to_me*). *articles_discussed* is the number of the news articles that the participants actively discussed, while *articles_linked* is the number of those for which they also provided at least one URL link in their posts. Analogously, *article_vulg_discussed* is the number of articles in whose discussion they used at least one expletive. Timeline of participant activity was of interest to us as well. In this study we restricted ourselves to recording the length of their active presence on the forum in days by subtracting the date of their first post from that of their last post (*days_active*).

Finally, there are features that combine content-oriented and communication-oriented information: entities like words, URL links and expletives can be totaled separately for child posts (*child_words_count*, *child_links_count*, *child_vulg_count*) and *parent* posts. Of course, *parent* counts would be redundant, because they can be easily obtained by subtracting the *child* counts from the global ones.

A note on methodology: Regarding expletives, our approach actually depended on the data source. Some newspapers like *Cas.sk* are relatively benevolent, which means that expletives can be identified in its discussion posts by plain matching. Other newspapers, like *Sme.sk*, are rather strict in censoring foul language, so we had to improvise a different approach. When a post on *Sme.sk* violates the rules, its text is replaced by a notice to that

[4] www.konspiratori.sk/en

effect. Surmising that this happens mostly because of foul language, we counted the number of such notices (*posts_violated_codex_cnt*) and used that as an approximation for the count of expletives. In spite of this approximation, we found a high degree of similarity between *Sme.sk* and *Cas.sk* with respect to foul language (see Fig. 2).

D. Derived Features for Data Analysis

Most of the basic features described above are expressed in terms of absolute numbers and as such are not suitable for analysis. In order to detect typological similarities among discussion participants, we need to avoid being misled by superficial differences in their feature values due to the varying duration of their active presence on the forum and other similar causes. Such superficial differences can be effectively removed by calculating derived, relative features with respect to the duration (*days_active*) or some other suitable base attribute (*words_count*, *posts*, etc.). Because there were several options available to us, we tried them empirically and retained those that seemed to work best for our purposes. While doing the clustering analysis we tried to avoid using highly correlated features, especially those derived from the same base feature, because they might skew the results. Our set of selected relative features is listed in Table II.

E. Data Cleaning and Standardisation

In the next phase we cleaned the data by removing the users who produced less than 27 posts in total. These participants were not sufficiently characterized by their limited set of posts, which lead to the formation of distinct linear artefacts (both vertical and horizontal) at special values of our standardized relative features, such as 0, 0.25, 0.5, and 1.0. We determined the minimum of 27 posts empirically by starting with lower values and gradually increasing them until the confusing artefacts disappeared. Since each dataset was stored in one table where records (row vectors) represented participants and column vectors their attributes or characteristic aspects, we simply removed from these tables the rows for users with post counts smaller than 27. While doing so, we did *not* modify the attribute values in the remaining rows, so this cleaning had no influence on the values of attributes like *rel_react_to_me* in the remaining rows.

This cleaning considerably reduced our datasets. Their original sizes in terms of participant profiles can be seen in the last column of Table I. Their sizes after the cleaning are listed in Table III. As part of cleaning, we also discarded about 20 participant profiles whose feature values were obvious outliers.

After the cleaning, the retained features were standardized so that their mean value became zero with standard deviation equal to one. We avoided normalization to the closed interval <0, 1>, because it only uses the feature minima and maxima, which makes it sensitive to outliers. In our opinion, standardization better accounts for the overall distribution of feature values and is more appropriate for the calculation of similarity and distance metrics in clustering algorithms.

With these cleaned and standardized datasets we then commenced our analysis.

III. Data Analysis

A. Comparing Datasets

The first part of our analysis consisted in searching for similarities in the spatial distribution of samples in our datasets. We expected such similarities to be indicative of a common underlying structure that we could then try to uncover through a more focused and detailed clustering exercise.

It did not take us long to find out that some features, notably *rel_post_per_day*, *rel_react_to_me*, and *rel_word*, exhibited similarities in their histograms and distribution functions across two or more datasets. While this kind of similarity does not by itself guarantee a similarity in the deeper structure of the concerned datasets, it does warrant further investigation. In the next step we therefore compared scatter plots of selected pairs of features for all our datasets. Fig. 1 shows the scatter plots for the pair (*rel_react_to_me*, *rel_post_per_day*). Aside from the varying size of our datasets, some measure of similarity in the spatial distribution of samples is quite apparent. The highest sample density is close to the horizontal axis, which means that most participants posted rather infrequently (on average). As we move upwards from the horizontal axis, sample density falls rapidly. It also falls as we move to the right, which means that few participants were able to elicit responses from many distinct people.

TABLE II
LIST OF RELEVANT RELATIVE FEATURES CHARACTERIZING 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 the distinct discussion members who responded to this user divided by the total number of his or her posts in the dataset
rel_violati on	Posts_violated_co dex_cnt / Posts	Ratio of the user's censored posts to all his or her posts.

TABLE III
DATASET SIZES AFTER THE CLEANING PROCEDURE

Dataset name	# participant profiles
Sme.sk domestic	5,791
Sme.sk international	3,534
Cas.sk domestic	1,885
Cas.sk international	2,201

Finally, the top right quadrant in all the four charts is almost empty. It means that frequent posting may not be the best way of engaging others in conversation.

We observed comparable levels of similarity for the pair (*rel_react_to_me*, *rel_word*), but do not show the scatter plots here in order to save space. What really surprised us was the similarity of the scatter plots for the pair (*rel_react_to_me*, *rel_vulg*) for *Cas.sk* and its approximation by (*rel_react_to_me*, *rel_violation*) for *Sme.sk*. We show these in Fig. 2. Their high similarity means that we can consider *rel_vulg* and *rel_violation* interchangeable for the purposes of our analysis.

At this point we would like to emphasize the extraordinary utility of *rel_react_to_me*, which was used on the horizontal axis in all the eight scatter plots. In fact, it was present in the majority of feature pairs revealing interesting patterns in the spatial distribution of our dataset samples. We think its utility stems from the fact that it characterizes each participant indirectly, as if through the eyes of others, and the concerned participant cannot directly alter or distort its value by his or her own actions.

The fact that several features behaved similarly across two or more datasets testifies to their adequacy on one hand, and on the other indicates the presence of deeper structural similarities that might perhaps generalize to other discussion forums as well.

Unfortunately, not all the attributes exhibited such striking similarities across our datasets. Notable among them was the average karma per post (*rel_kval*). At present, on the basis of histogram inspection, we are inclined to believe that there might be a problem with karma at *Sme.sk* and would like to investigate this in more detail in our future work. Because *karma* is the second of only two attributes that characterize each participant indirectly through the reactions of others, this considerably slowed down our advance towards a unified clustering structure for *Sme.sk* and *Cas.sk*. In the rest of this article we therefore perform the clustering exercise only for *Cas.sk*, hoping to correlate its results with those for *Sme.sk* as soon as the problem with its *karma* is fully understood and resolved.

B. *Clustering Analysis of* Cas.sk

We began our clustering exercise by inspecting the scatter plots for various pairs of attributes, hoping to see signs of naturally emerging clusters or at least of unevenly populated areas. Given the source and character of our data, it would be unrealistic to expect clearly visible and non-overlapping clusters with wide margins. This can indeed happen in certain domains, but in real-life online discussions one would rather expect a relatively evenly occupied feature space with varying sample density. This density might naturally and gradually decrease in certain directions, e.g. it is likely that the more posts a user produces, the shorter they are. But even here we cannot a priori exclude the possibility of a prolific author of many longish posts. In this situation, the presence of clear and wide margins between clusters would more likely indicate a lack of representativeness in the data or some kind of distortion or noise. Our primary approach for the identification of clusters was therefore density-based.

Based on the scatter plot visualizations, we selected up to eight most promising attributes and supplied them to various clustering methods. The most interesting spatial structure was observed for the following group of attributes: *rel_vulg*, *rel_words*, *rel_kval*, *rel_post_per_day* and *rel_react_to_me*. The clustering algorithm was EM (Expectation Maximisation), which is density-based and does not require the number of clusters as input. We took advantage of this and let it freely determine both the number and the spatial arrangement of clusters. EM parameters were as follows: distance=Euclidean; minimal standard deviation=0.15; max. iterations=100; number of K-means runs=10.

In both datasets (domestic news and international news), the method identified five clusters of similar relative sizes. Fig. 3 shows the five identified clusters (color-coded) for both datasets from two different perspectives. The top row displays the dependence of the average karma per the user's post (*rel_kval*) on the user's response elicitation factor (*rel_react_to_me*). The bottom row shows an analogical dependence for the user's average post length (*rel_word*). The charts on the left correspond to international news and those on the right to domestic news.

The similarity of cluster positions and shapes between the datasets is significant given that they are largely

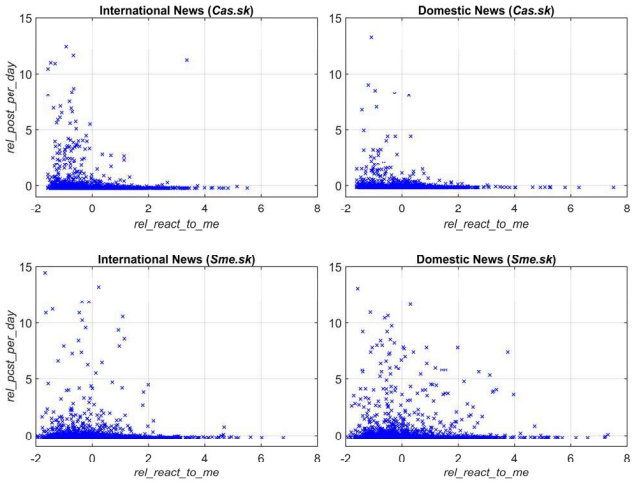


Figure 1. Scatter plots of our four datasets for the feature pair (*rel_react_to_me*, *rel_post_per_day*)

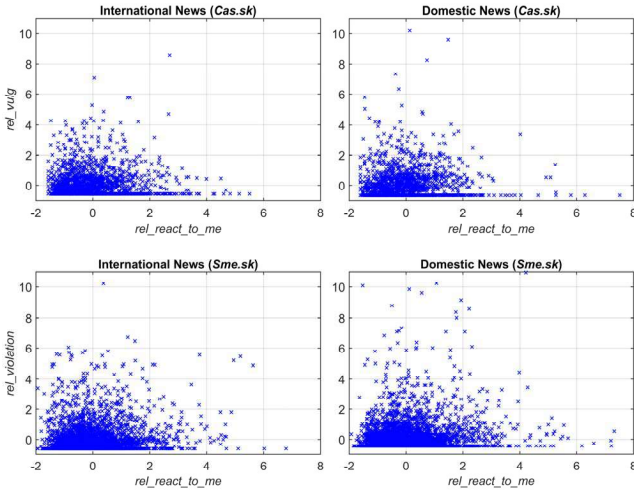
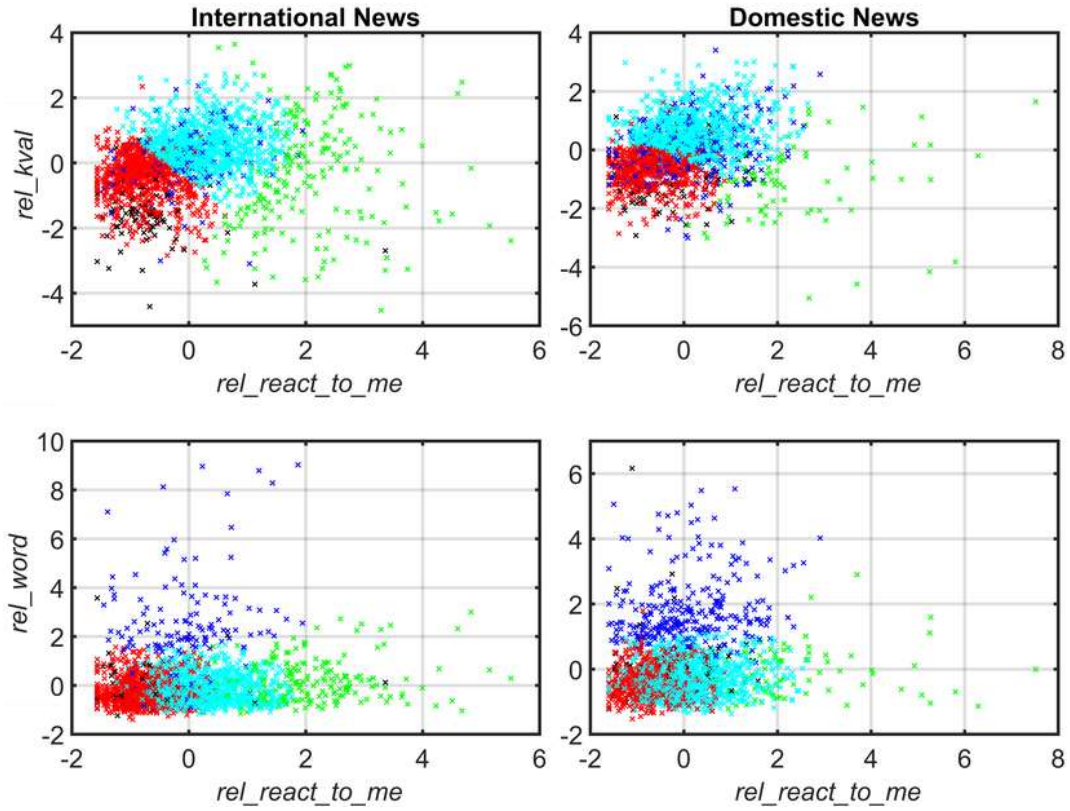


Figure 2. The top row shows scatter plots for *Cas.sk* and the feature pair (*rel_react_to_me*, *rel_vulg*), while the bottom one for *Sme.sk* and the approximated feature pair (*rel_react_to_me*, *rel_violation*)



independent. Of course, each *Cas.sk* forum member can freely discuss both the domestic and the international news, but the discussed topics should be vastly different. The fact that the EM method recognized the same number of clusters in both datasets (without having it imposed by us) is also very encouraging. Moreover, it could do so without any guidance regarding the target attribute, which makes the result doubly significant. (The presence or absence of a target attribute typically has a large influence on unsupervised machine learning results.) We can therefore conclude that these clusters represent not just these two particular datasets, but to some extent generalize the sample density on the whole domain *Cas.sk*.

Broadly speaking, the *red* cluster in Fig. 3 comprises the participants with less than average karma (*rel_kval*) and response elicitation factor (*rel_react_to_me*). The *cyan* cluster is its complement, because its members tend to score above average in both respects. The members of the *green* cluster are widely scattered, but their unifying characteristic is high response elicitation factor. The dark *blue* cluster represents the authors of longish posts. (The same color-coding of clusters was used in Figures 4 to 6.)

Regarding cluster positions in the bottom row of Fig. 3 (*rel_word* vs. *rel_react_to_me*), a similar spatial structure is present on both charts, with three clusters (*red*, *cyan* and *green*) perched next to each other on the abscissa and the fourth (*blue*) hovering above them. Adjacent clusters also tend to overlap to some extent. This structure is present in both datasets and the relative cluster sizes are very similar too, as shown in Tab. IV. (This table also lists the absolute cluster sizes in terms of sample counts.)

Despite strong similarities we also noted a few minor differences:

1. The *green* clusters visually appear to have different sample densities. But their relative sizes are quite similar (4,3% vs 2,8%), so this visual difference is likely caused by the smaller size of the Domestic Dataset which seems to have affected its *green* cluster in particular.

2. The angles of the hyperplane separating the *cyan* from the *red* cluster are different in the top row of Fig. 3 (*rel_kval* vs. *rel_react_to_me*). However, this hyperplane passes through a high-density area where even a small difference in sample density can strongly affect its angle.

3. Going purely by Tab. IV, the *blue* clusters appear to have significantly different relative sizes. However, the charts in the bottom row of Fig. 3 (*rel_word* vs. *rel_react_to_me*) reveal that this is caused mainly by their lower boundary being positioned a bit higher on the left chart (international news) than on the right (domestic news). With increasing size of the datasets we would expect the relative size of the *blue* clusters to stabilize somewhere between 8% and 12%.

TABLE IV: ABSOLUTE AND RELATIVE CLUSTER SIZES

Cluster Number	Cluster Color	International News from *Cas.sk*		Domestic News from *Cas.sk*	
		Sample Count	Relative Size [%]	Sample Count	Relative Size [%]
1	*cyan*	1179	53.57	918	48.70
2	*red*	701	31.85	531	28.17
3	*blue*	142	6.45	275	14.59
4	*green*	94	4.27	53	2.81
5	*black*	85	3.86	108	5.73

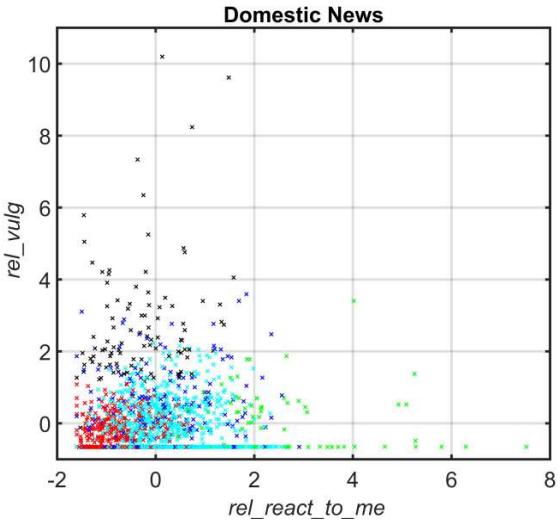


Figure 4. The black cluster in the Domestic News dataset represents the "foul-mouthed" participants

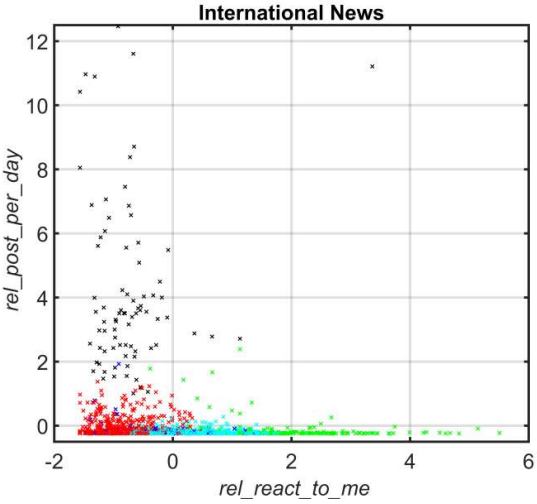


Figure 5. The black cluster in the International News dataset represents the frequent posters

4. In the charts in Fig. 3 there is also a *black* cluster, but it is not clearly visible. However, as Fig. 4 and Fig. 5 testify, this is just a matter of perspective: in other two-dimensional projections it becomes easily separable. It also turns out that in each dataset it represents a different group of participants: in the domestic news it is the "foul-mouthed" ones (i.e. those characterized by higher values of *rel_vulg* feature in Fig. 4); while in the international news it is the frequent posters (i.e. those with higher values of *rel_post_per_day* in Fig. 5). Both groups are relatively small with about 100 samples or 6% of dataset population. We believe this difference emerged because the counterpart of each group in the other dataset (i.e. the foul-mouthed participants in the international news and the frequent posters in the domestic news) was less clearly delineated from the rest than its more successful competitor. At present we cannot say whether the two datasets are really different in this respect, because 100 samples are not enough for a reliable assertion. Further exploration with more data will be required.

At this point in our experiments we decided to verify the stability of our clustering results. In order to test it, we provided a new, sixth attribute to our clusterer (*rel_hoax*, i.e. the average number of hoax links per post) and checked how it influenced the cluster structure and boundaries. This attribute can help us to indentify users prone to spreading hoaxes, but we were also interested in any relationship it might exhibit with respect to the user's karma or posting frequency. When we repeated the clustering exercise with this attribute added, our clusterer again identified five clusters in the Domestic News

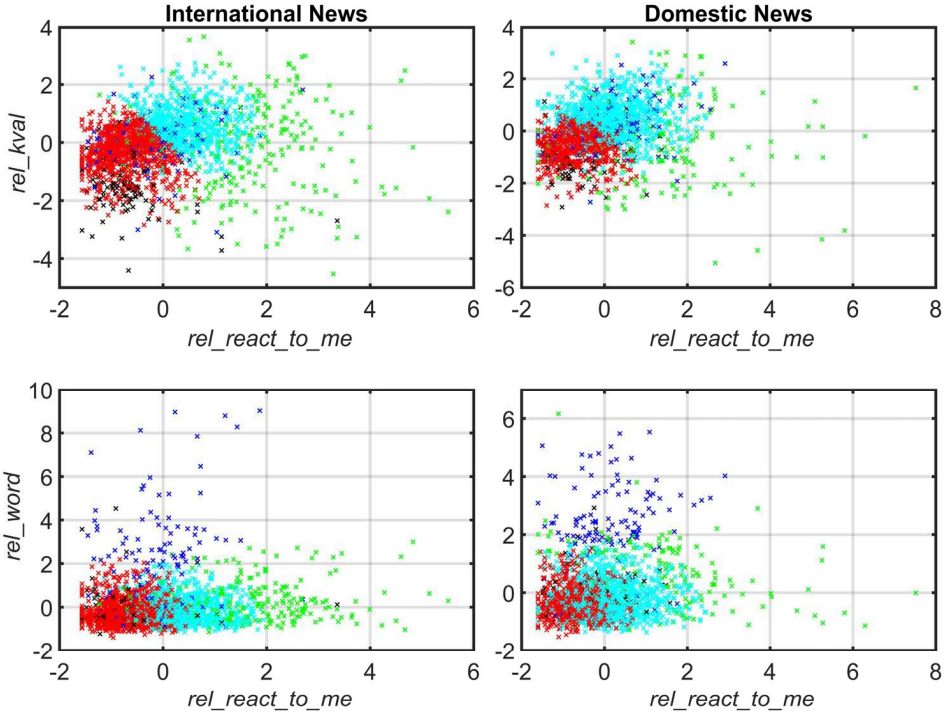


Figure 6. Visualisation of five clusters for international news (left) and domestic news (right) on *Cas.sk* after adding the sixth feature (rel_hoax). In this case the number of clusters for international news was imposed by us

TABLE V.
SELECTED STATISTICAL CHARACTERISTICS OF THE ATTRIBUTE REL_HOAX IN THE ANALYSED DATASETS

rel_hoax characteristics: Dataset name:	mean	standard deviation	Kurtosis
Cas.sk – Domestic News	0.0001	0.011	419.28
Cas.sk – International News	0.0040	0.038	324.36

dataset, but only three in the International News one.

By subsequent investigation we determined that this difference was probably caused by the varying character of the *rel_hoax* attribute in the two datasets: its mean and standard deviation were nearly four times higher for the international news than for the domestic news (see Table V). While the standardization effectively suppresses these differences, there are others that persist. For example, there is also a significant difference in *Kurtosis*, which is not affected by the standardization. We therefore consider it likely that such deeper statistical differences affect the clustering and its results. Nevertheless, we felt it would be worth seeing the same number of clusters in both datasets in this new six-dimensional feature space. We therefore imposed five clusters on our EM clusterer for the international news as well. The resulting clusters were again of similar relative sizes, shapes and positions, as can be seen in Fig. 6. The differences with respect to the clusters identified earlier in the five-dimensional feature space (see Fig. 3) are hardly noticeable. We therefore consider these five clusters as sufficiently stable and well-defined for our future research.

CONCLUSION AND FUTURE WORK

In this paper we compared the structure of online discussions in two major Slovak national newspapers (*Sme.sk* and *Cas.sk*) in two different categories, domestic news and international news. We also carried out a more focused clustering analysis of both categories for *Cas.sk*. The structure of online discussions was expressed in terms of participant profiles derived from publicly available data on both domains. Our base participant profiles included content-oriented features, such as the number of expletives or URL links in the participants' posts, as well as communication-oriented ones, such as how many distinct people ever responded to them. For our analysis, we derived circa 18 relative attributes from the base ones, out of which we selected five or six most effective ones for clustering.

For each category on *Cas.sk* we identified five relatively stable clusters, four of which exhibited significant similarity between the two categories. The fifth, smallest cluster turned out to be category-specific. We therefore conclude that at least the four common clusters generalize to some extent the structure of online discussions on the domain *Cas.sk*. Each cluster represents a group of discussion participants with similar characteristics and could be interpreted as a kind of participant type or role.

In the future we would like to analyze data from more discussion forums in order to see to what extent the identified clustering structure would generalize across them. Our features are deliberately defined in a language-independent way so as to be directly usable in other languages besides Slovak. Our future ambitions include more advanced forms of content analysis, such as sentiment analysis, user stance identification, topic recognition and text segmentation. We believe these can help us to better understand and more effectively respond to various social and socio-psychological phenomena in online social media, such as the spread of hoaxes and conspiracy theories or the diverse forms of anti-social behavior.

ACKNOWLEDGMENT

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

REFERENCES

[1] Mojzis, J. (2016). Visualization, Navigation and Relationship Discovery in Graphs. Information Sciences and Technologies, 8(2), 45.

[2] Morrison, Donn, et al. "Evolutionary Clustering and Analysis of User Behaviour in Online Forums." ICWSM. 2012.

[3] Kumar, Srijan, Justin Cheng, and Jure Leskovec. "Antisocial Behavior on the Web: Characterization and Detection." Proceedings of the 26th International Conference on World Wide Web Companion. International World Wide Web Conferences Steering Committee, 2017.

[4] Hardaker, Claire. "Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions." Journal of Politeness Research, 6 (2). ISSN 16125681(2010): 215-242.

[5] Cheng, Justin, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. "Antisocial Behavior in Online Discussion Communities." ICWSM. 2015.

[6] Cheng, Justin, et al. "Anyone Can Become a Troll." American Scientist 105.3 (2017): 152.

[7] Wood, M. J., & Douglas, K. M. (2013). "What about building 7?" A social psychological study of online discussion of 9/11 conspiracy theories. Frontiers in Psychology, 4, 409.

[8] Tan, C., Niculae, V., Danescu-Niculescu-Mizil, C., & Lee, L. (2016, April). Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In Proceedings of the 25th international conference on world wide web (pp. 613-624). International World Wide Web Conferences Steering Committee.

[9] Mojžiš, J., & Budinská, I. (2017, October). Analyzing news articles from the side of discussion threads. In Intelligent Engineering Systems (INES), 2017 IEEE 21st International Conference on (pp. 000297-000300). IEEE.

[10] Batsakis, S., Petrakis, E. G., & Milios, E. (2009). Improving the performance of focused web crawlers. Data & Knowledge Engineering, 68(10), 1001-1013.

[11] Diligenti, M., Coetzee, F., Lawrence, S., Giles, C. L., & Gori, M. (2000, September). Focused Crawling Using Context Graphs. In VLDB (pp. 527-534).

[12] Du, Y., Xu, Y., & Wang, M. (2017). A Novel Cooperation and Competition Strategy Among Multi-Agent Crawlers. Computing and Informatics, 35(5), 1050-1078.