OPI OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

Political discourse in Polish Internet – corpus of highly emotive Internet discussions

Antoni Sobkowicz[1], Katarzyna Baraniak[1]

[1] Ośrodek Przetwarzania Informacji

Państwowy Instytut Badawczy

4th Conference on CMC and Social Media Corpora for the Humanities

27–28 September 2016
Ljubljana

cmc-corpora.org

OPI
OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

Disclaimer

The original posts were gathered from an online forum (Onet.pl) and represent the fora participants' opinions. The names of Public figures and the nicknames used to describe them, and events were selected from the most active discussions for illustrative purposes of the algorithm capacity and do not represent the views of the authors and their organizations.

OPI
OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

Political Discussions in Polish Internet

The discussions in political forums in Poland contain large amounts of direct and indirect insults towards politicians.

Indirect insults are often expressed through the use of their nicknames, heavily associated with the politician's name, personal history, and public events they took part in.

OPI
OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

Our dataset

Our data comes from Onet.pl website, and was gathered for 9 months, from June 2015 to March 2016.

Onet.pl is one of the largest news sites in Poland, with tens of articles and thousands of comments written every day. Discussions are placed below the articles and are displayed in the tree format, allowing users to reply to each other.

OPI
OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

Our dataset

Data gathered from website was not heavily processed – out dataset contains:

- Article texts, publication date, tags (not used in this research)
- Comments with network structure preserved, with commenter usernames, posting dates, user scores

OPI
OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

Our dataset

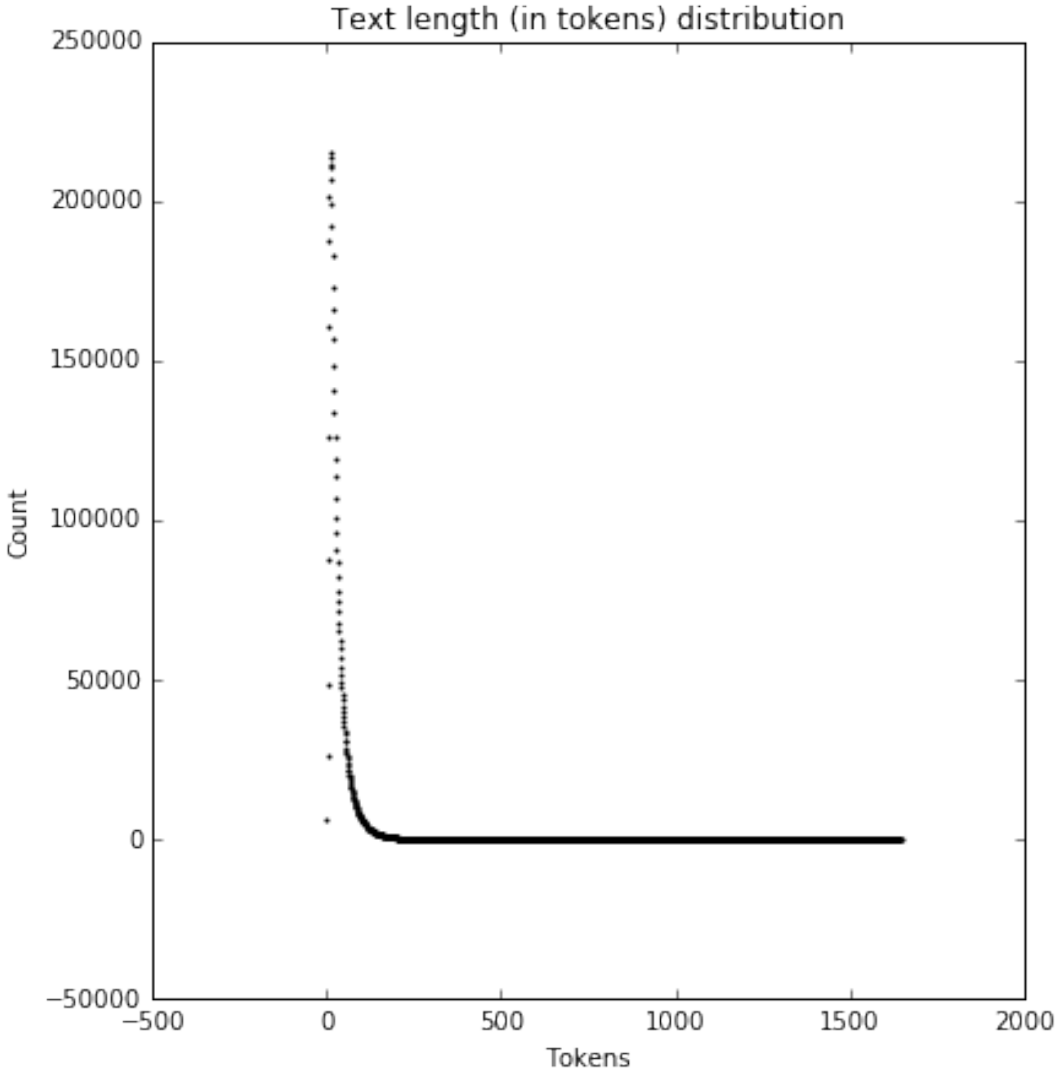
We gathered over 4.8 million comments, categorized as below

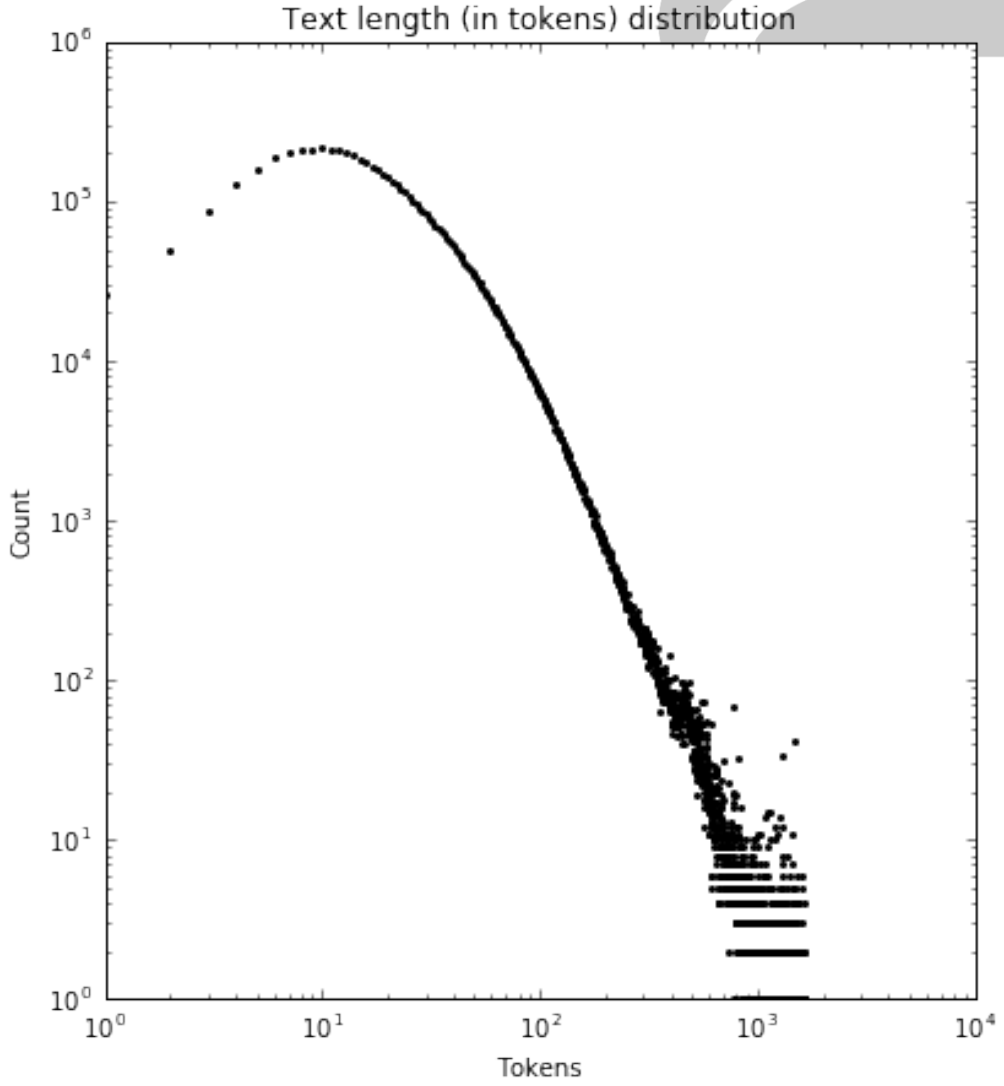
- **POL**, containing 4 829 076 comments from political categories from 2015, from Onet.pl, gathered in three distinct periods (Jun - Aug 2015, Sept - Dec 2015 and Jan - Mar 2016;

- **NONPOL**, containing 101 325 comments from non-political categories from Onet.pl, gathered between Feb - Mar 2016.

OPI
OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

Distributions – text length

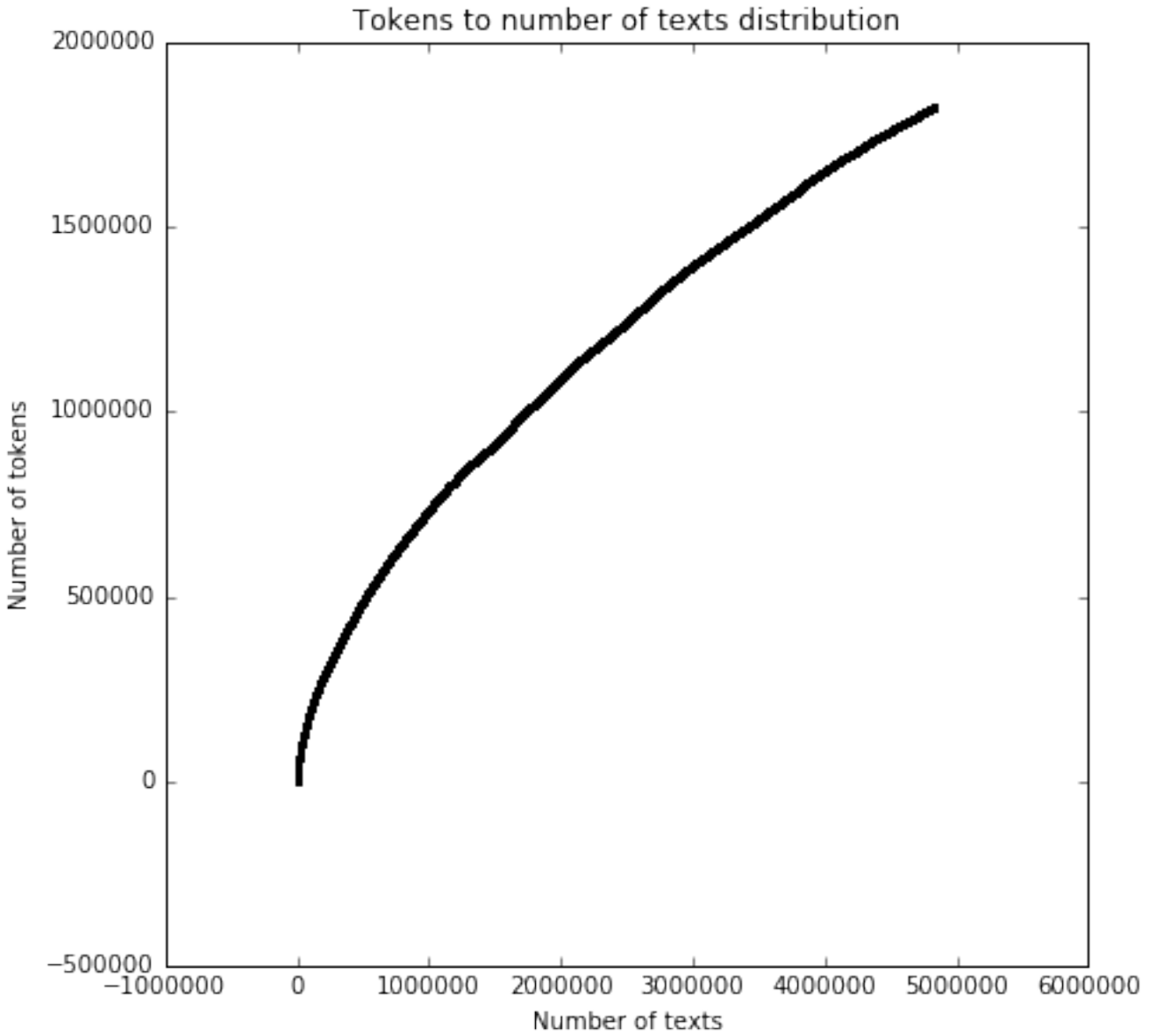




OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

Distributions

Unique tokens



OPI
OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

Goal

Our goal was to see if the combination of Machine Learning based algorithm for detecting emotional texts and network structure of discussions would allow us to find any special classes of users producing more emotion evoking content.

OPI
OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

Emotion evoking analysis

Compare:

"You should not listen to him, he doesn't know what he talks about." and *"Don't listen to that dickhead."*

Name calling (US examples, not Polish):
"crooked Hillary", "birther"

OPI
OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

Emotion evoking analysis

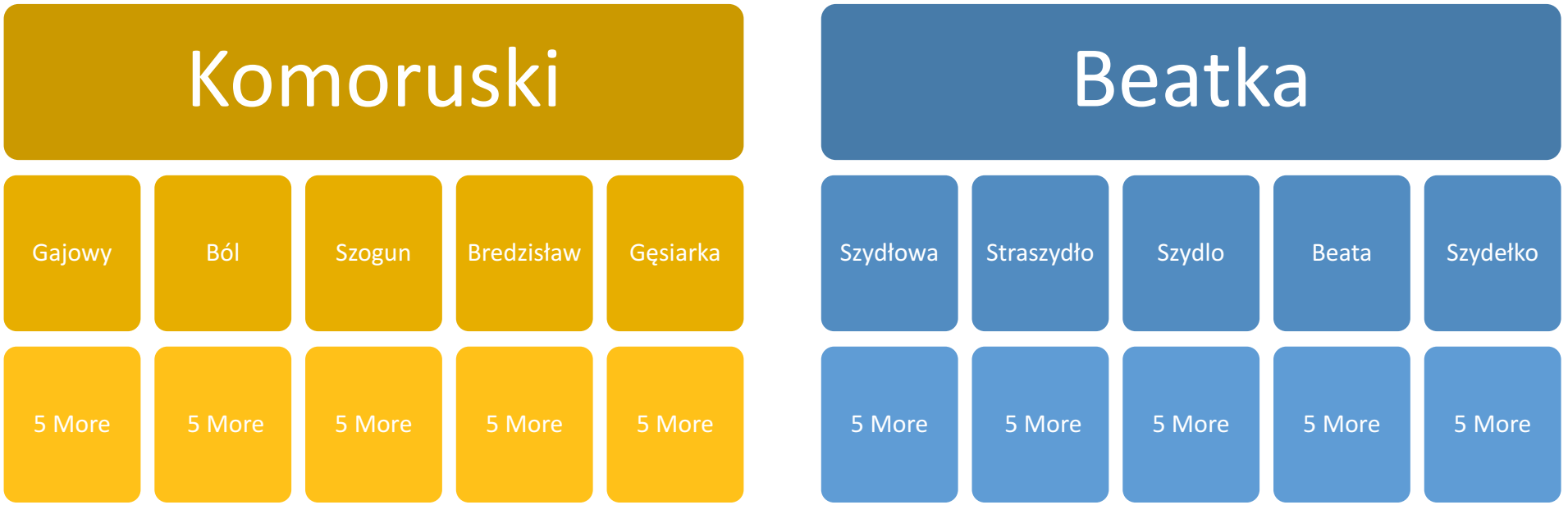
We preselected 5 nicknames, connected to politicians from the governing party and the opposition and used word2vec algorithm hoping it would allow us to find more nicknames automatically…

...and it went better than expected.

OPI
OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

Emotion evoking analysis

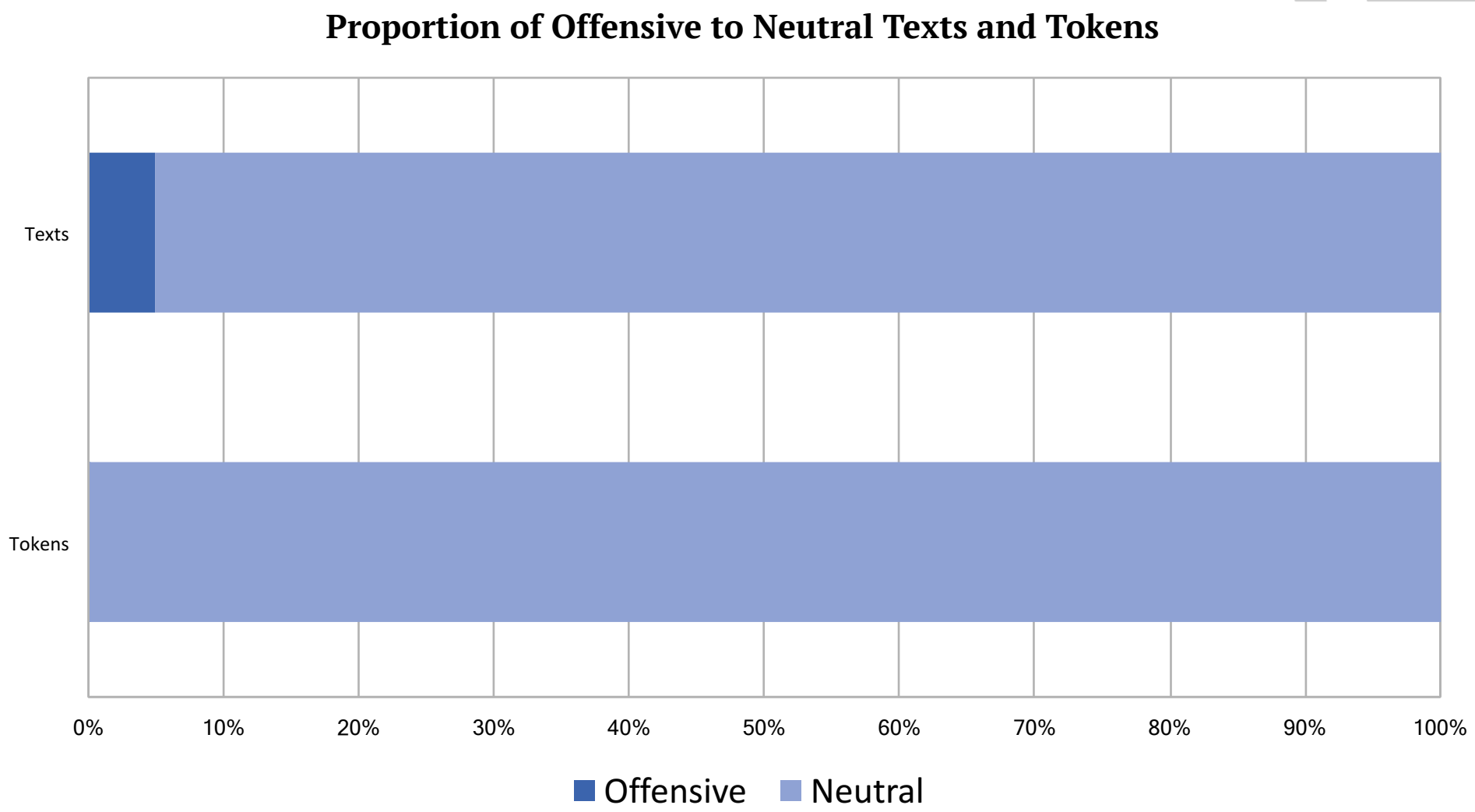
We got 125 “offensive” words – all connected to initial nicknames – out of 1,826,906 total different tokens found in texts.



Using these “offensive” words, we tagged texts – marking them as negative if one of selected tokens was found in text.

OPI
OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

Emotion evoking analysis



OPI
OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

Posting time patterns

Does the way you post can tell anything about what do you post?

OPI
OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

Posting time patterns

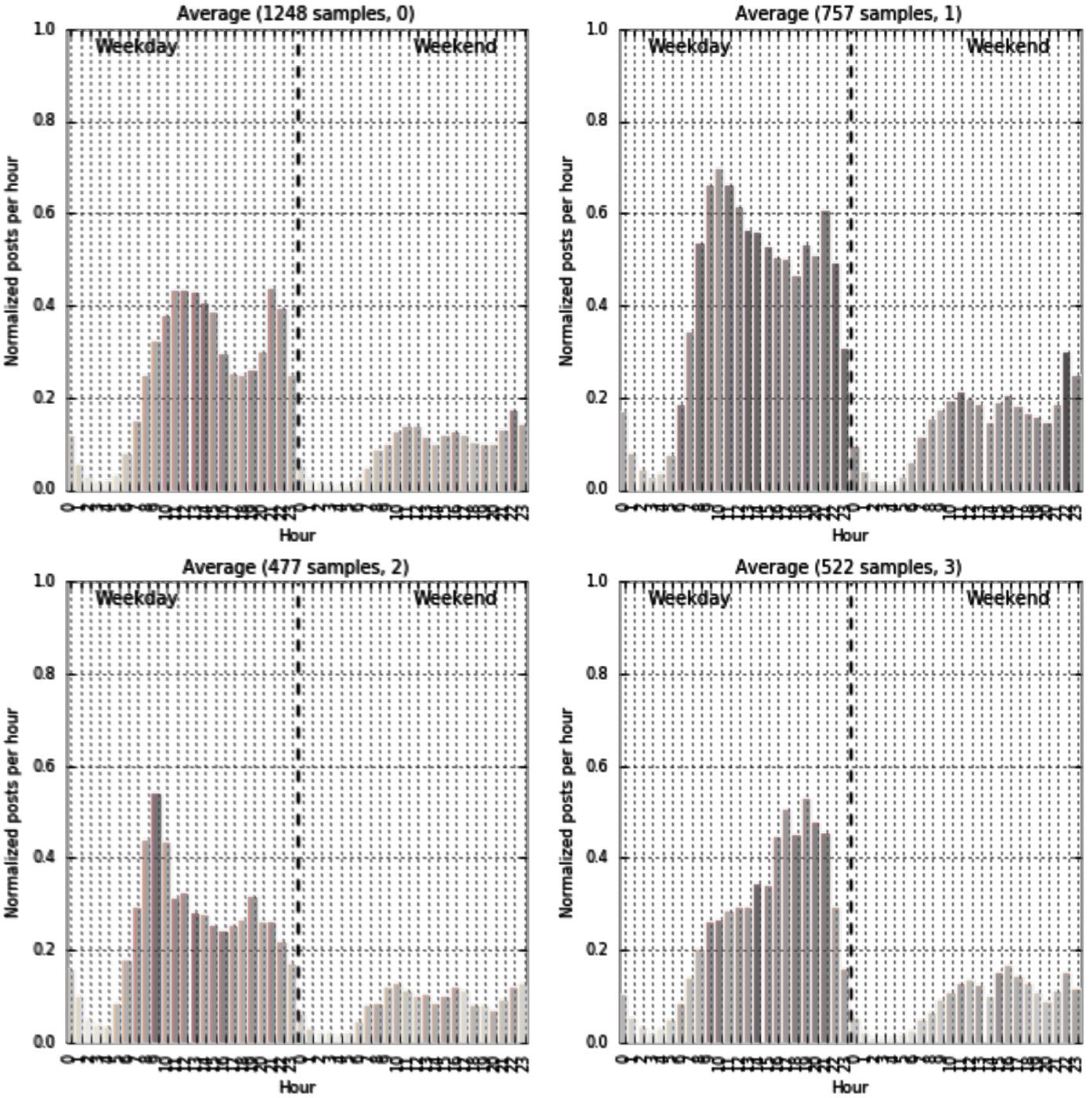
We analyzed posting patterns for users who had written more than 50 posts in the first 3-month time period.

We used 48-dimensional vector to represent how many posts a user has written during weekdays (first 24 dimensions for each hour) and similarly for the weekends.

The users who had similar posting patterns were gathered into four distinct groups by an Agglomerative Hierarchical Clustering algorithm using manually selected clustering seed.

OPI
OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

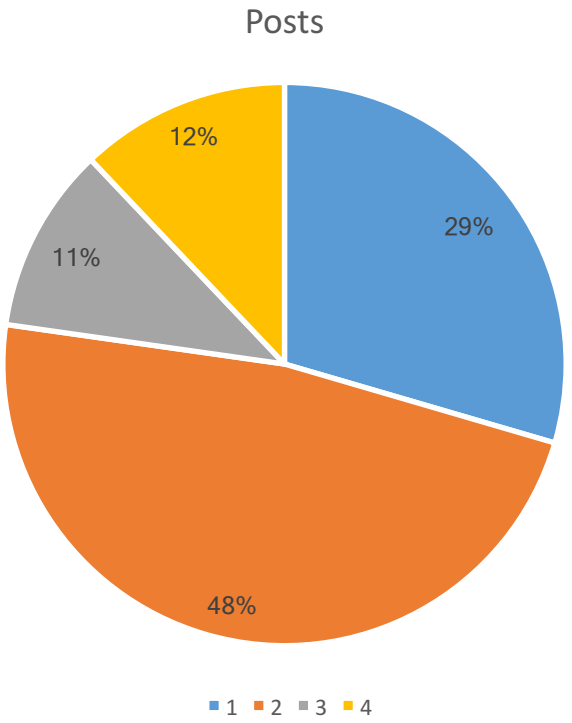
Grouping results

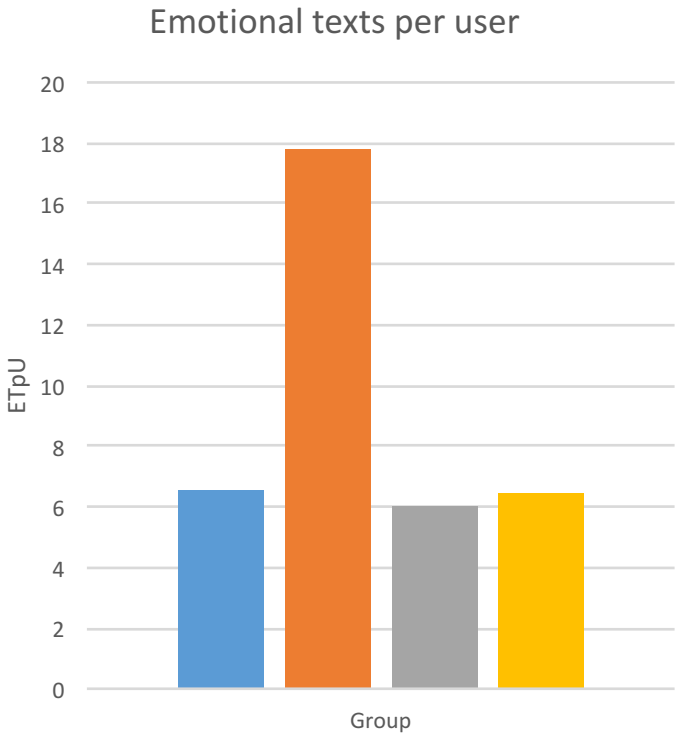


OPI
OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

Grouping results

Group	Users	Posts	ETpU
1	1248	157626	6,51923
2	757	254795	17,7873
3	477	57032	6,00839
4	522	64445	6,42529





OPI
OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

Posting time patterns

We are able to find users who produce most of emotion evoking content using only time patterns.

This process had manual components (optimizing the seed, selecting the number of groups)
– our next step was to find if the groups have any discriminating feature that would allow us to automate the selection process by selecting best clustering seed.

OPI
OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

Finding discriminating features for groups

We have decided to look at some network-related statistics for texts in each group, such as:

- Average in-degree in group
- Average response chain length
- Average out-degree
- Average number of responses
- Others…

OPI
OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

Finding discriminating features for groups

Many responses but no dialogue

Group	Avg. In-degree	σ	Avg. response chain length	σ
1	66,53	87,47	1,27	3,16
2	**155,11**	**229,7**	**0,98**	**2,69**
3	61,11	100,8	1,28	3,36
4	62,71	81,02	1,3	3,21

OPI
OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

Finding discriminating features for groups

Finally, we have two parameters that we use for automating seed selection process

- Average in-degree in group (number of responses)
- Average response chain length – length of dialogues

Most important: none of them is dependent on the text itself! Pure network properties.

OPI
OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

Automating clustering seed selection

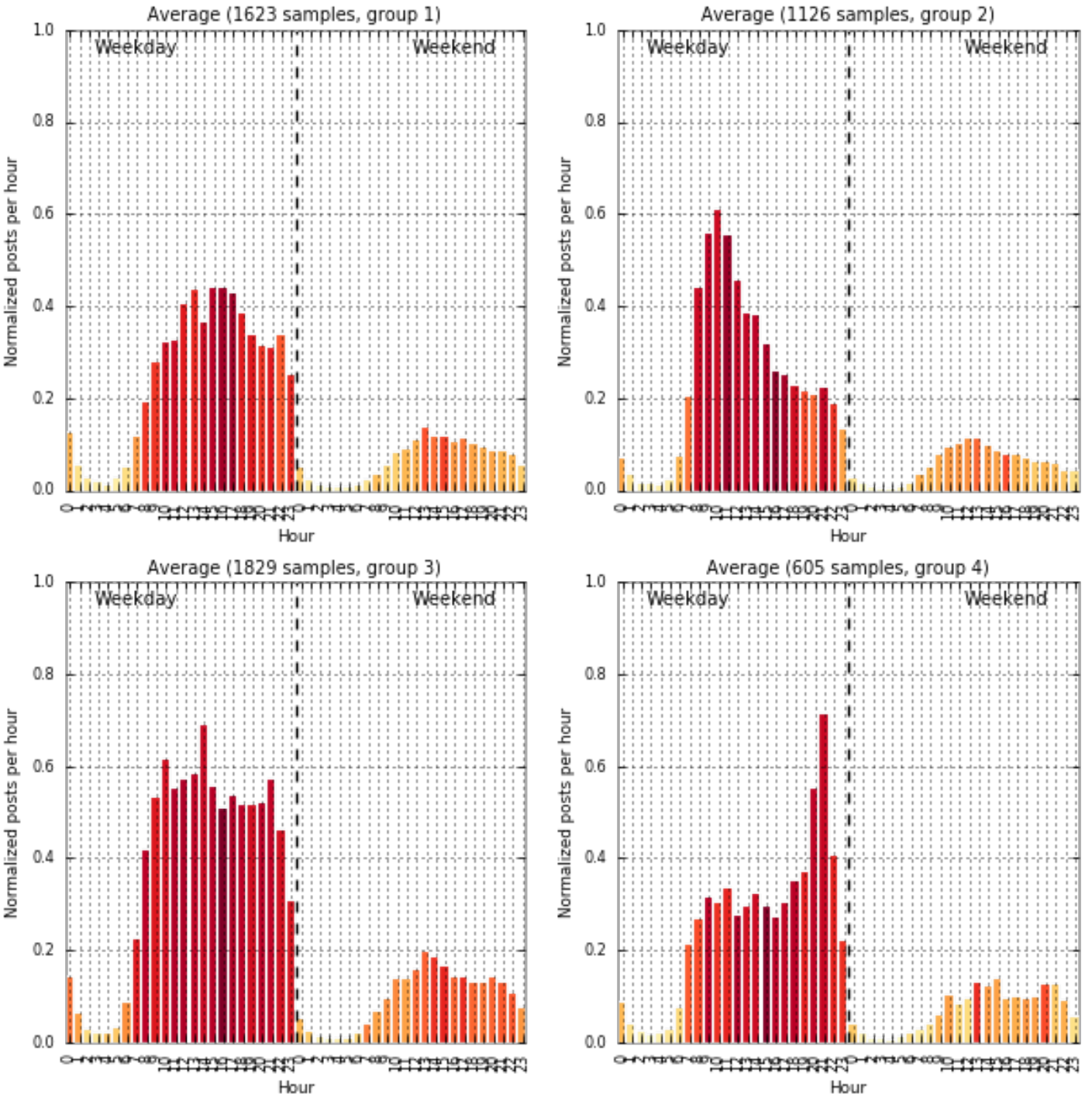
Having found out statistical properties that discriminate group, we must select such clustering seed which would make the difference between one of the groups from the rest will be as large as possible.

OPI

OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

Grouping results for automatic analysis

WHOLLY NEW DATASET

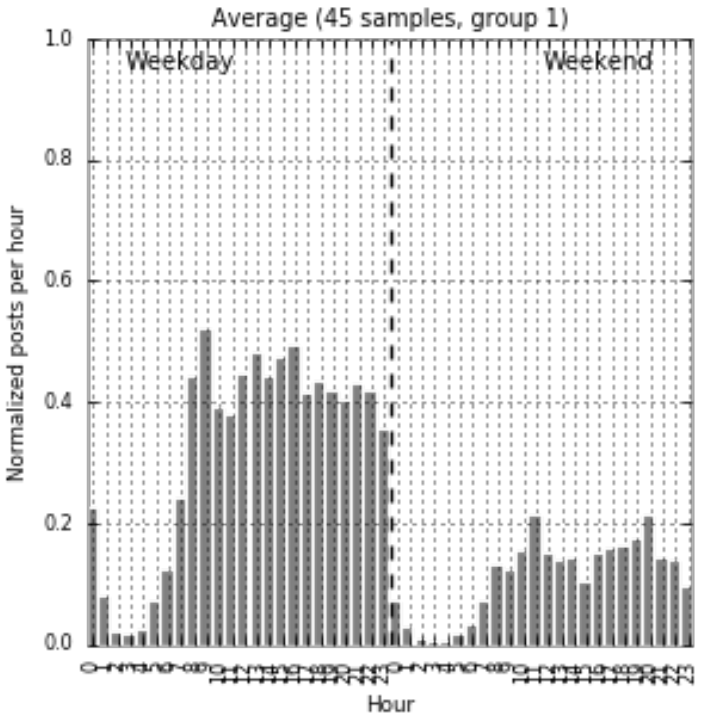


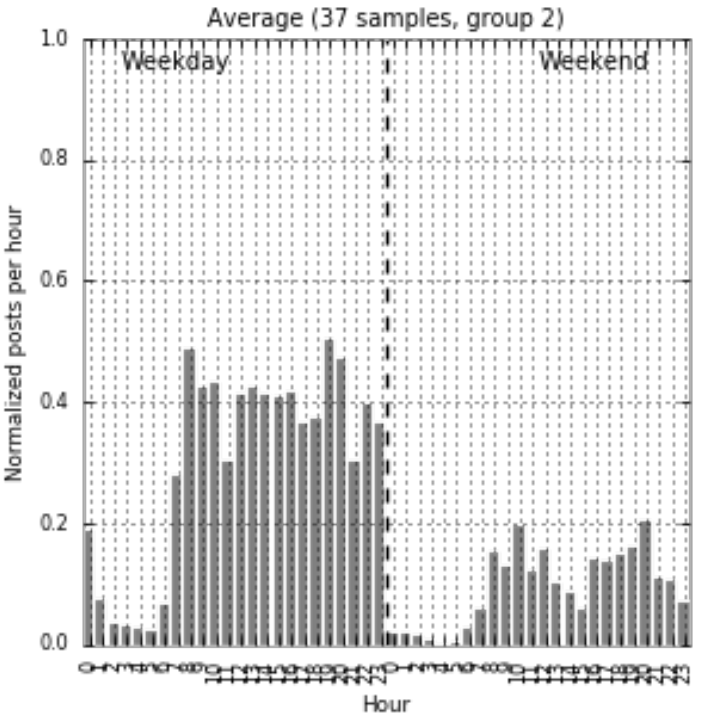
OPI

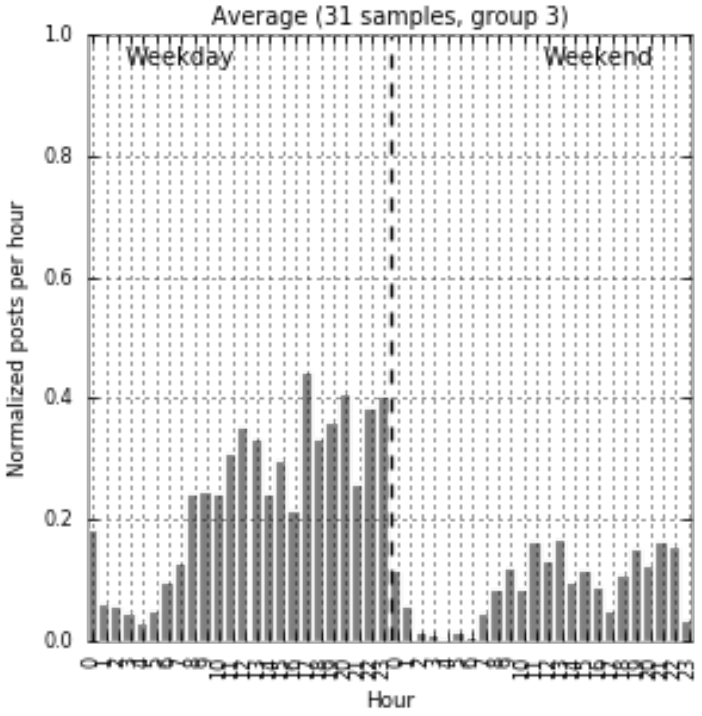
OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

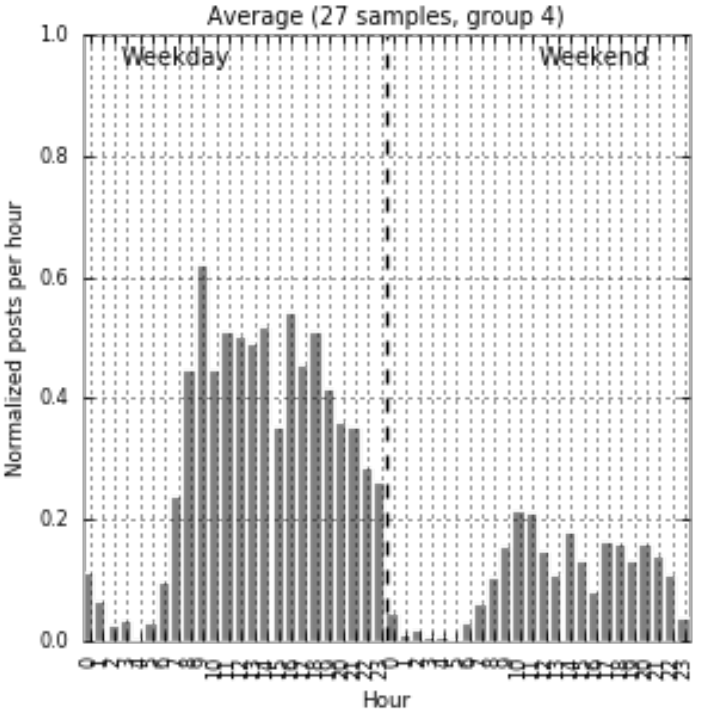
Grouping results for automatic analysis

SPORT NEWS









OPI
OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

Conclusions

We have found a way to find groups of users producing highly emotion evoking content only by using statistical properties of the way they post, not post content in highly polarized discussions.