The MediaBubble Dataset

A Crowdsourcing Dataset for Topic Detection Tasks for the Hungarian Language

László Grad-Gyenge, Linda Andersson
Creo Group, TU Vienna
Budapest,Vienna
laszlo.grad-gyenge@creo.hu, linda.andersson@tuwien.ac.at

Abstract

The paper presents the MediaBubble Dataset. Developing the dataset, our primary aim is to fill the gap of political topic detection dataset for Hungarian, a low density language. The dataset contains 1 000 political articles appeared in the Hungarian on-line media in political topics on the major news portals between 26.04.2017 and 29.04.2017. The dataset contains the topics and topic assignments created by 3 annotators. In addition, the dataset is initiated as a crowdsourcing dataset. It means that although the dataset is publicly available, in order to download it, a dedicated amount of annotations has to be conducted as a contribution to research.

Keywords: topic detection, crowdsourcing, dataset, Hungarian

1. Introduction

The importance of understanding political discourse on on-line platforms is becoming increasingly clear. There is several work in this direction done on high density languages such as German and English, but very few cover low density languages such as Hungarian. The purpose with the project MediaBubble [1] is to develop an adequate dataset for topics reflecting different political opinions of on-line news articles for the Hungarian language. Political preference on the same topic is referred to as *framing* (Card et al., 2015). Framing is related to the bias in a political discussions which emphasize or favor the speaker/writers opinion on a specific topic. Framing has been a central concept in political science and journalism for many decade (Goffman, 1974).

The primary goal of the project is to aid on-line news readers to eliminate / extend their filter bubble by recommending news articles on the same topic, but with a different political preference, i.e. frames, on the article the user is interested in. The frame for a news will have a different source frame aiming for a change in the perception of the issue among the reader (Scheufele, 1999). In (Fulgoni et al., 2016), they discover within the frame of *police violence* that the liberal press would use term as *uprising*, meanwhile the conservatives press would refer to the same event as *riot*.

In end-application of the MediaBubble project, we will utilize various semantic representation techniques on the articles appearing on-line. In order to train and evaluate different methods we first need to establish a dataset.

The Hungarian language belongs to the group of Uralic languages. Hungarian is an agglutinative language and unlike Germanic languages, does not follow a strict word order. In addition, to mention some of its properties, the language is rich on grammatical cases, lacks on grammatical gender, uses postpositions, involves specific plural markers, uses possessive suffixes and numeral expressions are singular. In order to develop semantic representation methods for such a language, specific techniques are to be involved. Our interest in creating this data set is two-folded. At first, we think that an adequate dataset for political mining of low density language is interesting by itself. At second, to be able to compare state-of-the-art mining algorithms involving distributional semantic methods on Uralic in comparison with Germanic language would contribute to the computational linguistic research community. Our contribution to the MediaBubble Dataset can be summarized as:

- collecting the news text from on-line portals with different political views,
- the initial annotations in the dataset,
- the maintenance of the underlying infrastructure,
- the user interface and work-flow logic to serve further annotation tasks.

The rest of the paper is organized as follows. Section 2. discusses related research conducted. Section 3. presents how the data was collected, pre-processed and its format. Section 4. introduces the user interface to prepare and to contribute to the dataset. Section 4.1. presents the format and the statistical properties of the dataset. Section 5. concludes the paper and gives insight into our future plans.

2. Related Work

In recent years, there has been a significant interest of mining social media for political preference. Tweets are by far the most popular, there have been several studies regarding tweets associated with elections in different countries e.g. Germany (Tumasjan et al., 2010), Ireland (Bakliwal et al., 2013) and U.S. presidential election (Williams and Gulati, 2008).

In (Tumasjan et al., 2010), the focus was on the federal election in order to investigate if the twitter flow could predict the outcome of the election. In (Bakliwal et al., 2013),

[1] The MediaBubble project has been carried out by Creo Group (creo.hu) in cooperation with Mérték Médiaelemző Műhely (mertek.eu). The project has been supported by the Google Digital News Initiative (digitalnewsinitiative.com)

they addressed sentiment analysis of the Irish General Election 2011, the goal was to classify tweets on a specific topic as positive/negative/neutral and also to see if it was possible to detect tweets as sarcastic i.e. if the literal sentiment was different to its actual sentiment. In (Williams and Gulati, 2008) they studied the effect of using social media platforms such as Facebook for vote sharing in the presidential primaries 2007.

Political tweets have also been used to discover party loyalty, in (Calzolari et al., 2016), they investigated if it was possible to discover if social and behavioural information available on Twitter would give sufficient data to train a classifier in order to identify *aisle-crossing politicians* i.e. those politicians who vote against their party. They collected 184,914 tweets from members of the U.S. Congress (both the House of Representatives and Senates) utilizing frames. Each tweet could be classified with one or more frames. They had a predefined list of 17 possible frames (e.g. Economic, Capacity & resources, Quality of Life, Culture identity, etc.

Frames have also been explored in public statements, congressional speeches, and news articles (Tsur et al., 2015; Baumer et al., 2015; Fulgoni et al., 2016). Tsur et al (Tsur et al., 2015), observed the language of framing in agenda setting campaigns. In (Baumer et al., 2015), they developed computational techniques in order to detect different framing on various political issues. Their work is based upon the fact that framing can have significant impact on the readers perception and therefore it is important to draw the reader's attention to the language of framing. They collected data from 15 political news feeds, and lay annotators have been used in order to reflect the frames among the general public. Fulgoni et al (Fulgoni et al., 2016), studied 17 different topics ranging from climate change to common core and from abortion to police violence. They observed divergence of themes between each partisans sides (conservative versus liberal), each sides would use different frames in order to appeal to their readers. For instance, in the abortion debate the conservative press use pro-life and the liberal press use anti-life.

We would like to contribute to research regarding language of framing for other languages than English such as Hungarian. It is of interest to study and to draw the readers' attention to the language of framing due to the pervasive influence of framing have on the reader perception on an issue.

3. Collecting the Dataset

The initial MediaBubble Dataset has been conducted on 1 000 news articles. The set of articles to be annotated is defined as the articles appeared on the major Hungarian news portals in the interval 26.04.2017 - 29.04.2017.

Table 1 summarizes the concrete portals involved in the initial annotation process of the dataset. The column denoted ”Name (Url)” contains the name and the URL of the portal. The column denoted ”Att” contains the political attitude of the particular portal according to the European political scale. The attitude is represented on a 1-5 scale. Value 1 represents the left attitude. Value 2 represents the moderate left attitude. Value 3 represents the centered political attitude. Value 4 represents the moderate right attitude. Value 5 represents the right attitude. The political attitude values are the subjective opinion of the author of this article and should not be considered as the ground truth. The column denoted ”Cnt” contains the number of articles involved in the annotation process from the specific news portal.

Name (Url)	Att	Cnt
24.hu (`24.hu`)	2	99
PestiSrácok (`pestisracok.hu`)	5	28
B1 BLOGCSALÁD (`b1.blog.hu`)	1	4
mandiner (`mandiner.hu`)	5	52
hvg.hu (`hvg.hu`)	2	99
Index (`index.hu`)	3	105
Kettős Mérce (`kettosmerce.blog.hu`)	2	2
Magyar Idők (`magyaridok.hu`)	5	102
Magyar Narancs (`magyarnarancs.hu`)	2	34
Magyar Nemzet (`mno.hu`)	4	102
Népszava (`nepszava.hu`)	2	134
ORIGO (`www.origo.hu`)	5	129
444 (`444.hu`)	1	62
888.hu (`888.hu`)	5	46
atlatszo.hu (`atlatszo.hu`)	1	1
Ténytár (`tenytar.hu`)	2	1

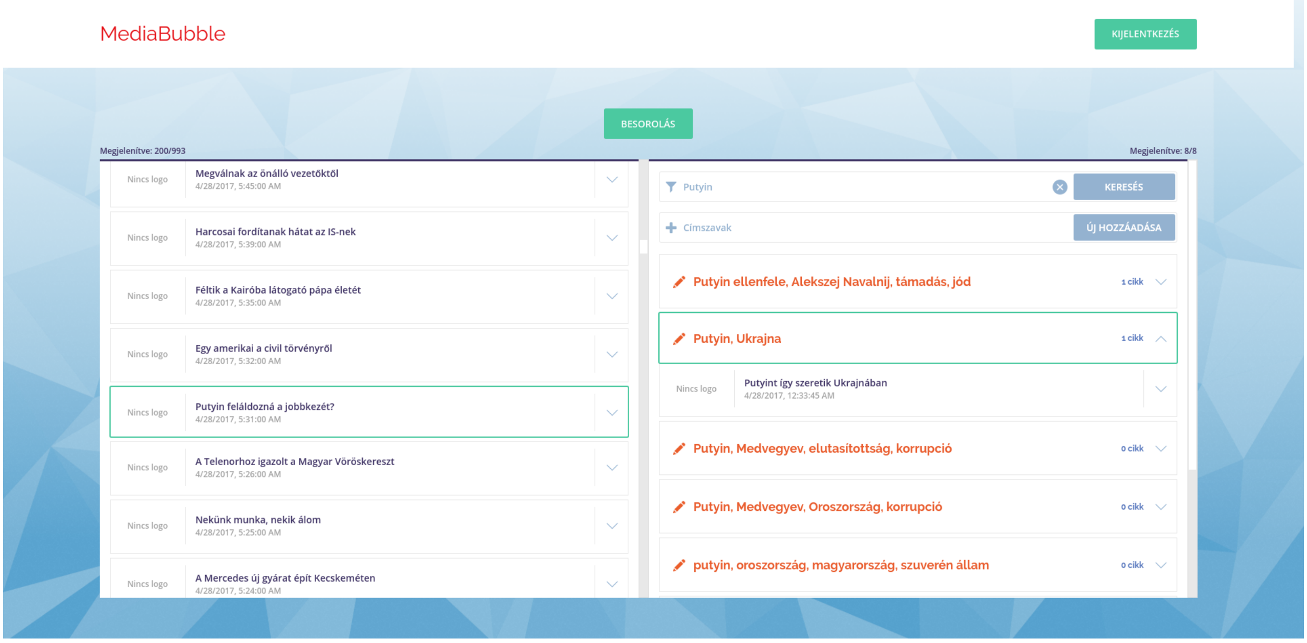
Table 1: List of news portals involved in the annotation process.

3.1. Format

The dataset can be downloaded as a compressed archive in tar.gz format. The files contained in the archive are in a tabular format and are the following.

- annotators.csv – The annotators involved in the project. The file contains an `id` and a `nick` column. The columns stand for the unique identifier of the annotator and for the nick name of the annotator, respectively.
- topics.csv – The topics defined by the annotators. The file contains an `id` and a `title` column. The columns stand for the unique identifier of the topic and for the title of the topic, respectively.
- articles.csv – The articles to be annotated. The file contains an `id` and a `title` column. The columns stand for the unqiue identifier of the article and for the title of the article, respectively.
- assignments.csv – The topic assignments of the articles. The file contains an `article`, a `topic` and an `annotator` column. The columns stand for the unique identifier of the article, the unique identifier of the topic and for the unique identifier of the topic assignment, respectively.

The concrete tabular file format is csv. The delimiter character is comma, the titles are quoted with double quotes. Double quotes in strings are escaped with repeated double quotes.



Megjelenítve: 200/993

Nincs logo Megválnak az önálló vezetőktől 4/28/2017, 5:45:00 AM

Nincs logo Harcosai fordítanak hátat az IS-nek 4/28/2017, 5:39:00 AM

Nincs logo Féltik a Kairóba látogató pápa életét 4/28/2017, 5:35:00 AM

Nincs logo Egy amerikai a civil törvényről 4/28/2017, 5:32:00 AM

Nincs logo Putyin feláldozná a jobbkezét? 4/28/2017, 5:31:00 AM

Nincs logo A Telenorhoz igazolt a Magyar Vöröskereszt 4/28/2017, 5:26:00 AM

Nincs logo Nekünk munka, nekik álom 4/28/2017, 5:25:00 AM

Nincs logo A Mercedes új gyárat épít Kecskeméten 4/28/2017, 5:24:00 AM

Megjelenítve: 8/8

Putyin

Címszavak

Putyin ellenfele, Alekszej Navalnij, támadás, jód 1 cikk

Putyin, Ukrajna 1 cikk

Nincs logo Putyint így szeretik Ukrajnában 4/28/2017, 12:33:45 AM

Putyin, Medvegyev, elutasítottság, korrupció 0 cikk

Putyin, Medvegyev, Oroszország, korrupció 0 cikk

putyin, oroszország, magyarország, szuverén állam 0 cikk

Figure 1: The user interface of the annotation tool.

4. The Annotation Process

The annotation process has been initiated involving 3 annotators. All of the annotators are highly educated and are outside of the computer science field. Before conducting the annotations, the goal and scope of the MediaBubble project has been described to them. The concrete meaning of the term ”topic” has not been described explicitly and typically has been referred to in an intuitive manner as our intention was not to influence the annotators into any specific direction.

Figure 1 presents the user interface developed specifically for the MediaBubble Dataset project. The annotation interface can be reached by conducting a login process. This is how the current annotator is identified. The user interface is clean and provides essential tools to support the annotation process. The main components of the user interface are the article list on the left and the topic list on the right. A counter is presented on the top of both lists to inform the annotators about the status of their annotation roadmap.

The article list contains the articles to be annotated. The list of the articles is fixed. Each article is presented with its title and the date of its appearance. In the case the title does not contain sufficient information, in order to get detailed information, the down arrow on the right of the title is to be clicked. The detailed view shows the abstract of the article and contains a link to display the article in its original location.

The topic list contains the topics the articles can be assigned to. The detailed view of a topic shows the assigned articles in a list below the title of the topic. The detailed view can be displayed by clicking the down arrow on the right of the title of the topic. Unlike the list of articles, the list of topics can be altered by the annotator. In a typical annotation work-flow, the annotator processes the articles sequentially. In the case no corresponding topic is available, the annotator can create a new topic by specifying its title. The title is used to help the annotators to identify the topic. The topics are shared among the annotators. In order to help the actual annotator to find a particular topic, a keyword-based search tool is created. By entering a search term, the annotator can filter the list of topics to the topics containing the search term.

An article can be selected by clicking on its title. The selection is indicated by an emphasizing border. The corresponding topic can be selected similarly, by clicking on its title. Having both items selected, the annotator has to click the assign button in order to finalize the topic assignment. In the case, the user clicks on an already assigned article, the title of the assign button changes to remove. The purpose of this button is to let the annotator undo mistaken assignments.

The novel contributions are to be conducted via the user interface described above.

4.1. The Statistical Properties of the established Dataset

In order to give a thorough overview of the dataset, the statistical properties of the dataset have been calculated from three different aspects as the empirical distribution of the cluster sizes, the empirical distribution of the agreement levels and the pairwise inter-annotator agreements of the annotators. The annotators are presented anonymously and are denoted with letter A, B and C.

Table 2 presents the count of topic sizes per annotator. The column ”Topic Size” denotes the size of the topic. Columns denoted as ”Annotator X” present the amount of topics of the corresponding size in the case of the specific annotator. The primary property of the dataset is that there is a specific amount of single articles assigned to a separate topic. On the other side, topics over size 10 are represented sparsely in this dataset. Topics containing only one article are validated. The purpose of these single articles is to provide control on false topic assignments of machine learning based methods.

Topic size	Annotator A	Annotator B	Annotator C
1	348	417	328
2	69	92	72
3	39	35	37
4	24	23	16
5	11	14	12
6	14	6	13
7	11	5	9
8	4	0	7
9	2	3	2
10	1	0	0
11	0	2	1
12	1	1	2
13	1	0	1
14	0	0	1
15	0	0	1

Table 2: Empirical distribution of the topic size per annotator.

In order to have an overview on the topic assignments of the articles, the annotator agreement level (AAL) is calculated for each article. The articles assigned to the same topic by all the annotators are denoted as having AAL 3. The articles having a majority vote meaning that two of the three annotators vote for the same topic. These articles are denoted as having AAL 2. Those articles assigned to three different topics by the annotators are marked as having AAL 1.
Having the measures calculated, the articles are separated into three different sets based on their AAL assignment. Table 3 presents the histogram of the articles regarding the AAL value. The column ”Annotator Agreement Level” contains the AAL value. The column ”Count of Articles” presents the amount of articles having the particular AAL value. Considering that a typical numerical experiment in the topic detection domain involves majority voting to determine the final / aggregated topic assignment of an article, the amount of articles having a final assignment is 736 which is 74% of the sample. This value could be looked upon as a kind of confidence level of the dataset, thus the dataset shows potential of be involved into further research projects.

Annotator Agreement Level	Count of Articles
3	222
2	514
1	264

Table 3: Count of articles per agreement level.

To analyze the topic assignments from the aspect of the annotators, Table 4 presents the pairwise inter-annotator agreement. As mentioned in the beginning of this section, 3 annotators are involved in the experiment. The column ”Annotator 1” and the column ”Annotator 2” contains the annotators. The column κ contains the inter-annotator agreement level of the two particular annotators. In order to present the results anonymously, the concrete annotators are denoted with A, B and C.

Annotator 1	Annotator 2	κ
A	B	0.466
A	C	0.414
B	C	0.423

Table 4: Pairwise inter-annotator agreement.

The Cohen’s kappa coefficients are in the interval $(0.4, 0.6]$ indicating a moderate agreement of the annotators.

5. Conclusion

In this paper, we have presented a first initiative to establish a political *framing* data set for the Hungarian language. The end-goal with this data set is to develop an application that gives the reader the possibility to read about topic of interest with different political aspects.
As our annotator resources are limited, the MediaBubble Dataset has been set up as a crowdsourcing dataset. The dataset is available for download for research purposes with the restriction that if someone would like to have access to the dataset, a specific amount of annotation has to be conducted as a contribution to the research community. Having the annotation task completed, the dataset is available for download. Details on downloading and contribution can be found on the homepage (Grad-Gyenge, 2017) of the dataset.
Our plans for the future can be described as the following. At first, we would like to emphasize the visibility of our initiative. We hope that crowdsourcing will be a potential technique to emphasize the size and the quality of the dataset. At second, we would like to involve the dataset into the development of novel semantic representation techniques especially for the Hungarian language.

Bakliwal, A., Foster, J., van der Puil, J., O’Brien, R., Tounsi, L., and Hughes, M. (2013). Sentiment analysis of political tweets: Towards an accurate classifier. In *Proceedings of the Workshop on Language Analysis in Social Media*, pages 49–58, Atlanta, Georgia, June. Association for Computational Linguistics.

Baumer, E., Elovic, E., Qin, Y., Polletta, F., and Gay, G. (2015). Testing and comparing computational approaches for identifying the language of framing in political news. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1472–1482.

Nicoletta Calzolari, et al., editors. (2016). *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*. ACL.

Card, D., Boydstun, A. E., Gross, J. H., Resnik, P., and Smith, N. A. (2015). The media frames corpus: Annotations of frames across issues. In *Proceedings of the*

53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 2: Short Papers, pages 438–444. The Association for Computer Linguistics.

Fulgoni, D., Carpenter, J., Ungar, L. H., and Preotiuc-Pietro, D. (2016). An empirical exploration of moral foundations theory in partisan news sources. In Nicoletta Calzolari, et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016.* European Language Resources Association (ELRA).

Goffman, E. (1974). *Frame analysis: An essay on the organization of experience.* Harvard University Press.

Grad-Gyenge, L. (2017). The MediaBubble Dataset. Available at: `http://laszlo.grad-gyenge.com/#!/mediabubble`.

Scheufele, D. (1999). Framing as a theory of media effects. *Journal of Communication*, 49(1):103–122.

Tsur, O., Calacci, D., and Lazer, D. (2015). A frame of mind: Using statistical models for detection of framing and agenda setting campaigns. In *ACL (1)*, pages 1629–1638.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. *Icwsm*, 10(1):178–185.

Williams, C. and Gulati, G. (2008). What is a social network worth? facebook and vote share in the 2008 presidential primaries. American Political Science Association.