Computer Science • 16 (2) 2015 http://dx.doi.org/10.7494/csci.2015.16.2.185

Michał Jankowski-Lorek
Kazimierz Zieliński

DOCUMENT CONTROVERSY CLASSIFICATION BASED ON THE WIKIPEDIA CATEGORY STRUCTURE

Abstract *Dispute and controversy are parts of our culture and cannot be omitted on the Internet (where it becomes more anonymous). There have been many studies on controversy, especially on social networks such as Wikipedia. This free on-line encyclopedia has become a very popular data source among many researchers studying behavior or natural language processing. This paper presents using the category structure of Wikipedia to determine the controversy of a single article. This is the first part of the proposed system for classification of topic controversy score for any given text.*

Keywords Wikipedia, controversy, classification

Citation Computer Science 16 (2) 2015: 185–198

1. Introduction

Over the years, the Internet has been rapidly evolving and has integrated into every part of our lives. Starting from work, where it gives us a means of fast and reliable communication, through entertainment and social portals, and up to learning and searching for information. Fewer and fewer people are using hard-copy versions of encyclopedias; instead, they use digital encyclopedias or simple search engines to find their required information.

Therefore, a very important question arises concerning the quality and credibility of the information presented on the Internet. Many studies have focused on its credibility and detecting highly non-credible information or pages. The bounded aspect of credibility is a controversy which is an integral part of any text posted on the Internet. Even subjects that are well supported by facts may still occasionally be questioned. Controversy may help us to encourage others to follow a topic and introduce new evidence, but it can also be destructive to a theory or generate antagonism among authors.

In this paper, we concentrate on detecting controversy that should be considered as yet another, third state of text credibility (besides credible and non-credible). Controversy may occur due to opinions, interpretations, and points of view among the authors or readers. Besides single statements, posts, or articles, there are some potentially controversial topics.

Our controversy-detection system uses the structure of categories in Wikipedia. Wikipedia has a set of content policies and conduct policies defining – among others – how to prevent controversy and how to handle controversial topics[1]. The list of known controversies is maintained manually by admins[2]. Our system follows the definition of controversy from Wikipedia by using it "as it is".

The main goal is to verify if we can use scores of articles aggregated by controversial topics to determine the controversy of a single article. This is the first part of a proposed new method for detection of controversy for any given text based on its topic.

The next section provides a background of controversy and determining topics based on Wikipedia. Then, we present an overview of a proposed system for controversy detection – its main parts and requirements. Next, in section 4, we describe and validate controversy aggregation from articles to categories. Finally, section 5 concludes the system proposition and current results.

[1] http://en.wikipedia.org/wiki/Wikipedia:NPOV,
http://en.wikipedia.org/wiki/Wikipedia:CONS,
http://en.wikipedia.org/wiki/Wikipedia:DISPUTE

[2] http://en.wikipedia.org/wiki/Wikipedia:CONT

2. Related work

Wikipedia has received a lot of attention from researchers, especially in the field of social behavior and content quality. In 2004, Viegas et al. [20] published the first paper about the problem of cooperation and conflict between editors on Wikipedia. Authors created a tool for the visualization of conflicts between co-editors. Another study about conflicts was performed by Buriol et al. [4], who focused on reverts between co-authors. In contrast to studies about conflicts, Borzymek et al. [2], Turek et al. [18], and Wierzbicki et al. [23] studied collaboration and teamwork based on the Wikipedia social network.

Wikipedia articles are stored as consequent revisions, allowing users to revert to a previous version at any time. Everyone can revise previous versions of the article and, therefore, can revert it to their preferred version. Repeating mutual reverts of single articles is called an edit war and usually indicates an argument about the content or controversy of the topic. Edit wars were studied by Sumi et al. [16] and [17]. Furthermore, Yasseri et al. [24] presented that edit wars finish in consensus or lead to permanent controversy on the specified topic.

Another way of detecting controversy on Wikipedia is to use available meta information about editors. Studies carried out by Kittur et al. [10] and Voung et al. [22]. Recently, Rad and Barbosa [14] have compared several controversy-detection models on Wikipedia. This work constituted a base for our latest research *Predicting Controversy of Wikipedia Articles Using the Article Feedback Tool* [7].

Wikipedia has previously been successfully used as a data source for semantic-information retrieval to improve the results from search engines and to create a categorization of texts.

Milne and Witten [12] measured semantic relatedness by using the hyperlink structure of Wikipedia articles. They used tf-idf link counts weighted by the probability of each link to compute the relatedness of each article. Behanam et al. [5] created a multi-tree for each entity in the Wikipedia category network, combined them, and then used a multi-tree similarity algorithm to compute the similarity of entities. Recently, Han [6] proposed a method of measuring semantic similarity that uses Wikipedia as an ontology and spreading-activation strategy.

Besides semantic similarity, Wikipedia category graph (WCG) was used in research to improve an ad hoc document retrieval (Kaptain et al. [8]), identifying document category (Schønhofen [15]), acquiring knowledge (Nastase and Strube [13]). Medelyan et al. [11] published an extensive overview of research that mines Wikipedia. In 2006, J. Voss [21] called the category structure a collaborative thesaurus. The structured form of Wikipedia categories allowed for the automated learning of ontology (Yu et al. [25]). Kittur et al. [9] used WCG to detect contentious topics in Wikipedia using annotated data. Recently, Biuk-Aghai et al. [1] made an attempt to visualize human collaboration in Wikipedia. They visualized WCG subtrees by transforming them into simple trees.

3. Design of a Web Content Controversy Detection System

Our research focuses on building a system that will determine the controversy of a given text document (web page, article, post, etc.) based on its similarity to existing articles in Wikipedia. The proposed similarity measure is cosine similarity between tf-idf vectors representing the documents. Figure 1 illustrates an overview of the proposed system. The first step is to find articles similar to a rated document in Wikipedia. After getting a subset of Wikipedia articles, the relevance of each to the categories in which they belong will be calculated. The final step is to apply a controversy model that will calculate the overall controversy degree for the text.

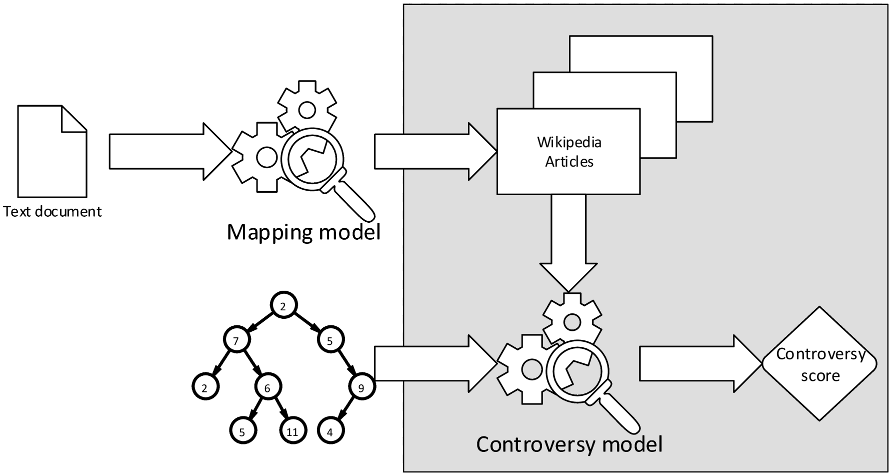

Figure 1. Functional overview of the proposed system.

In this paper, we are concentrated on the part of the system highlighted in Figure 1. We are demonstrating that it is possible to use aggregated scores of controversy to get a level of controversy for each topic, and consequently provide the controversy of a new article based on those topics.

3.1. Graph of topics controversy

The Wikipedia category system is a directed acyclic graph, with the articles and categories as nodes and directed edges indicating a parent relationship. Each category or article can be a child of many other categories and have many subcategories. The main category is *Content* and has 7 subcategories. There are over 1.5 million subcategories.

To calculate a controversy graph, we want to use the articles that are below the current node of the category structure. By starting with categories at the lowest level (containing only articles), we will calculate the controversy of each of the nodes. During this process, we will also prune the structure and remove all of the nodes that do not have an appropriate number of subcategories and articles (as to be derived in further studies). This part of the proposed system will be addressed in future studies.

Once the controversy graph is calculated, the controversy of the text will be dependent on the set of articles and their belonging to the categories.

3.2. Article mapping tool

The second part of the system is the mapper, which will allow us to map the analyzed text to a set of Wikipedia articles. A classic approach for determining semantical similarity and text classification is to treat both texts as a bag of words. This approach does not work for classification of all topics at once.

In this system, we propose using a semantic-similarity algorithm for only titles and first paragraphs of the articles. Similar work was carried out by Vandamme and Turck [19]. They proposed a way of searching related articles by stemming all words in the query and comparing n-grams to the stemmed-article titles and categories.

For most of the articles, the essence of the topic is contained in the first sentence or paragraph. This mapping system will be used only to determine the category. That is why its accuracy regarding the level of article is not necessary. For each article, we will use the relevance score in a further calculation of the overall controversy score. This score will be combined with the categories to which the article belongs as well as the controversy level of the subcategories. Based on this information, the overall controversy score will be calculated for each category.

4. Controversy score aggregation

As described in section 3.1, to build the proposed system, we will need to calculate a graph of topic controversy. To do this, we need to compute and aggregate controversy scores of all articles in Wikipedia.

4.1. Wikipedia article controversy classifier

In our previous research on predicting controversy on Wikipedia based on the Article Feedback Tool (AFT) [7], we prepared a classifier for determining the controversy of a single article. We created the learning dataset with 438 records. We selected 219 controversial articles from the official Wikipedia controversial-article list based on the number of AFT evaluations. We also selected 219 non-controversial articles based on a similarity of the length of text.

For each article, we retrieved meta information described in previous research by Kittur et al. [10]. We used the seven most important features based on the number of revisions, unique editors, anonymous edits in articles, and discussion pages for those articles. We also added a new features derived from AFT evaluations of articles: the frequency of different AFT ratings and the total number of votes. Beside those two types of features, we computed the emotion polarity of utterances from discussion pages.

Then we used a random forest machine learning algorithm [3] implementation in R to train classification models for three subsets of features (AFT, Kittur, Sentiment). Figure 2 presents the ROC curve for trained models, and Table 1 summarizes the performance of all of the models mentioned.

The algorithm proposed by Kittur et al. [10] delivers the best overall results, and it can be calculated for every article directly from its editing history. We chose this model to prepare a dataset for further work.

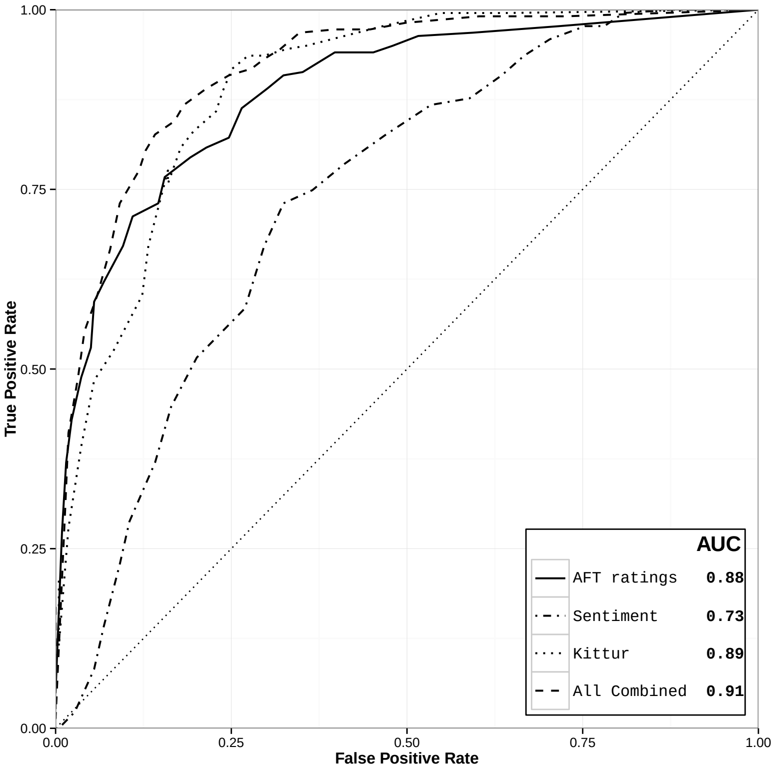

Figure 2. ROC of article controversy classifiers.

Table 1
Performance comparison in AFT dataset.

Measure	AFT	EP	Kittur	MR	Combined
F-measure	80.37%	69.05%	81.48%	73.27%	84.10%
Precision Controversial	79.96%	67.66%	79.77%	79.03%	83.14%
Precision Non-controversial	80.78%	70.78%	83.42%	69.73%	85.11%
ROC-AUC	0.88	0.73	0.89	0.76	0.91

4.2. Dataset

The current version of Wikipedia consists of more than 4.6M articles. We computed all of the required features and used the model described in the previous section to determine the controversy score and confidence level of this score for each of the articles. This dataset is available for others for further studies and can be downloaded from datahub.io website[3].

Only 0.5% (23,103) of the articles were classified as controversial. This can be considered a plausible result, as on the official list of controversial articles, there are 963 positions, and in the discussion pages, we can find 2,153 articles with a possible controversial note added (controversial article template). The mentioned list is not updated often and definitively does not contain many controversial topics and articles.

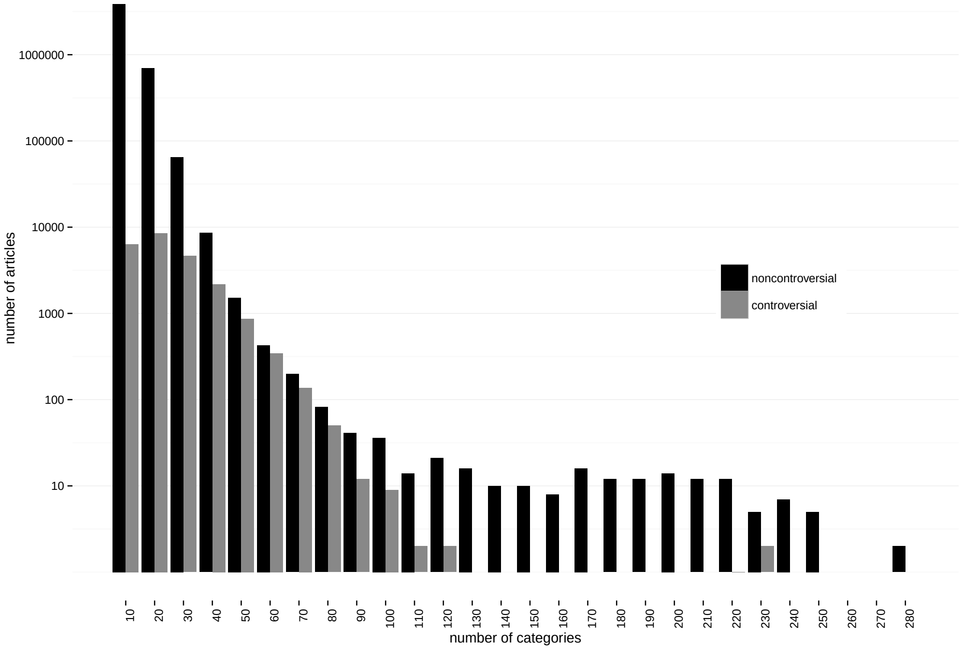

Figure 3. Histogram of the number of categories per article for non-controversial and controversial articles.

For each article, we retrieved all of the categories to which they are assigned. One article can be in many categories and on different levels of the Wikipedia category graph. Figure 3 presents a histogram of the number of categories per article for non-controversial and controversial classes. As we can see for both classes, most articles

[3] `http://datahub.io/dataset/controversy-of-wikipedia-articles`

have fewer than 20 categories assigned; however, there are articles with more than 200 categories.

There are 354,940 categories with only one or two pages assigned. These categories are discarded in further studies, as there is no need for any aggregation on such a small number of articles per category.

4.3. Controversy of Wikipedia categories

We have a good model for determining if a single article is controversial or not; but for further studies, we need to aggregate the micro scores and be able to determine the controversy of each node in the Wikipedia category graph (allowing us to detect new articles that are potentially controversial.

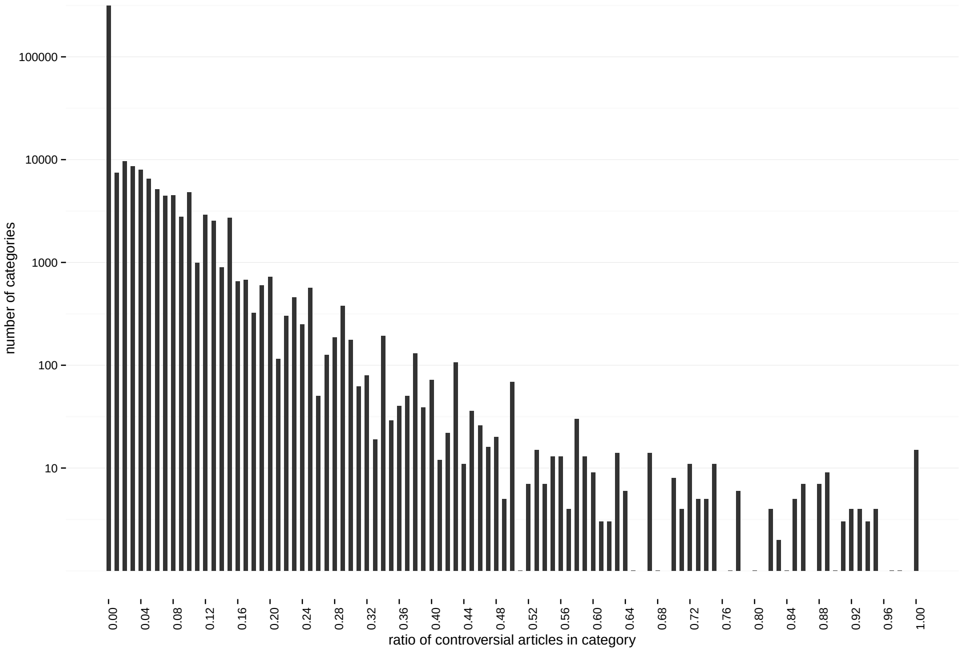

Figure 4. Histogram of the ratio of controversial articles in categories.

In Figure 4, we can see the number of categories with a given percentage of controversial articles. The distribution looks exponential for ratios smaller than 0.5. Above ratio 0.5 (i.e., where the number of controversial articles prevail over non-controversial ones), distribution is irregular, with a notable peak of 15 articles with a ratio equal to 1.0. This peak corresponds to categories which contain only controversial articles. Only 346 categories have 50% or more controversial articles. These categories can be treated as controversial.

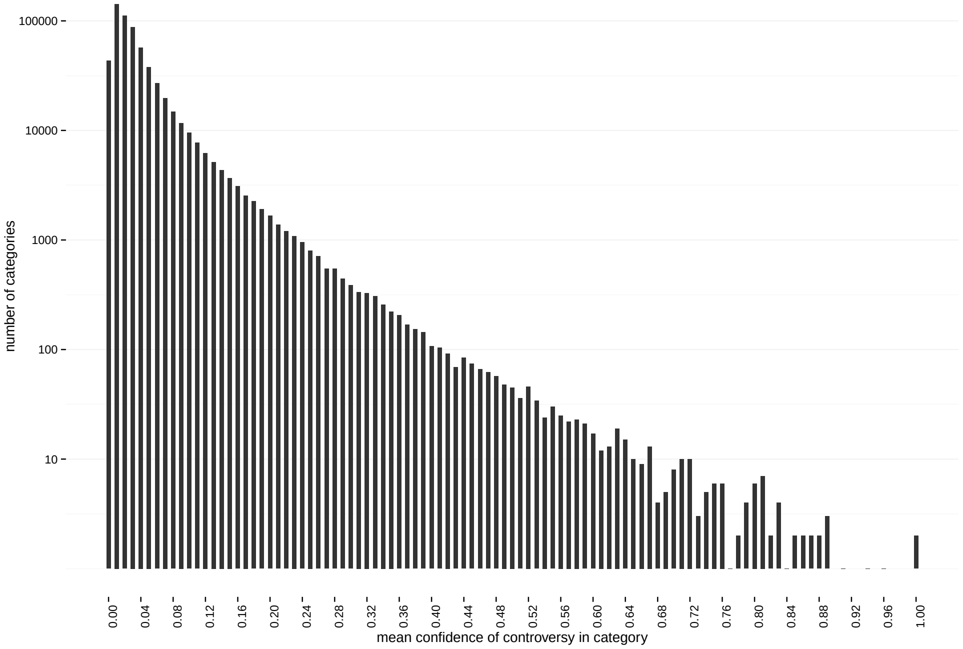

Figure 5. Histogram of mean confidence level of controversy in categories.

Figure 5 presents a histogram of mean confidence level of controversy in categories. Distribution is similar to ratio; but this time, we have 514 categories that have a 0.5 or higher mean of confidence level.

In both cases, the majority of categories are non-controversial; thus, all of Wikipedia should be treated as not controversial.

We tested our way of detection of controversy categories by segregating the article score in two approaches. The first one is based on the manual checking of a selected list of categories, while the second is based on cross validation.

4.3.1. Empiric validation

Based on the fact that some topics are generally known as controversial all over the world (for example: politics, religion, racism, etc), we manually tested the list of controversial and non-controversial categories.

Table 2 contains the top 20 controversial categories based on the mean of confidence of articles. The top of the list is occupied by "G8 nations", which seems to be the most controversial content-related category in English Wikipedia. All of the top 20 categories are about politics or religion.

Table 2

Top 20 controversial categories.

rank	category	mean confidence	# of articles
1	G8 nations	0.9978125	8
2	G7 nations	0.9975	7
3	G20 nations	0.952368421	19
4	Hindustani-speaking countries and territories	0.935833333	3
5	NUTS 1 statistical regions of the United Kingdom	0.905	3
6	People banned from entering China	0.888333333	3
7	Federal constitutional republics	0.8875	9
8	Slavic countries and territories	0.883076923	13
9	People of the American Enlightenment	0.87375	6
10	Near Eastern countries	0.873333333	9
11	Wars involving Qatar	0.861666667	3
12	Member states of the Union for the Mediterranean	0.860064103	39
13	Member states of NATO	0.855535714	28
14	Member states of the South Asian Association for Regional Cooperation	0.85375	8
15	Middle Eastern countries	0.842916667	18
16	Mormonism	0.841666667	3
17	Northeast Asian countries	0.8325	7
18	Member states of the Council of Europe	0.828928571	14
19	Democratic-Republican Party Presidents of the United States	0.828125	4
20	Western Asian countries	0.827368421	19

In the top 300 categories, 240 were manually verified as controversial or belonging to topics well-known as controversial. In a randomly-selected 100 categories from the list of all-controversial categories based on mean confidence, 82% were verified as controversial. On the randomly-selected list of 100 non-controversial categories, we found only 5% of controversial topics (false negatives).

This validation confirms that using the mean level of confidence of controversy for articles can be used to find controversial categories.

4.3.2. Cross validation

For a second validation, we decided to randomly split our dataset into training and test subsets with a ratio of 0.7. The training-article subset was used to calculate the controversy levels of categories based on the ratio of controversial to non-controversial articles as well as mean controversy confidence level. Then, we used these categories to calculate which articles in the test subset are controversial, based on the average level of confidence of all categories to which they were assigned. In testing the dataset, there were 1,399,303 articles, and 6,926 were originally classified as controversial.

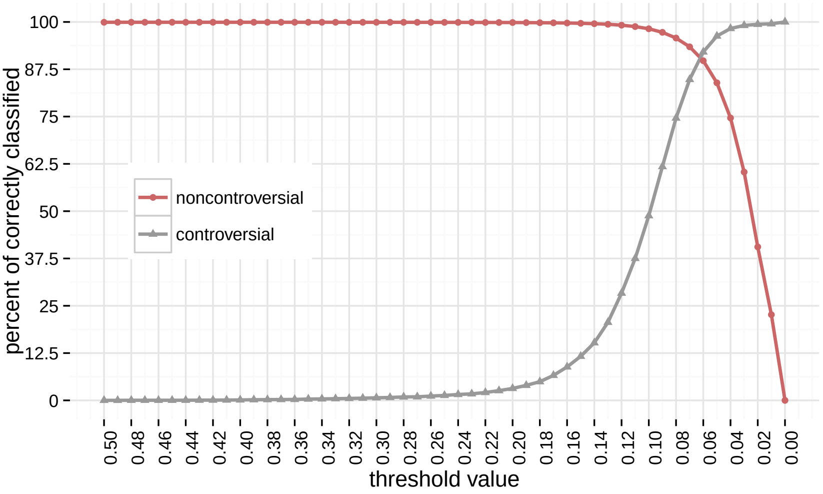

Figure 6. Percentage of correctly-classified controversial and non-controversial articles in correspondence to the chosen threshold value.

Figure 6 presents the percentage of correctly-classified controversial and non-controversial articles in correspondence with a chosen threshold value. An article was assigned as controversial if the average confidence controversy level of its categories exceeded the threshold.

As we can see, the percentage of correctly-detected controversial articles rapidly increases after lowering the threshold value below 0.2 and levels up at 0.06, while the percentage of correctly-classified non-controversial articles is steady until 0.12, but then starts to decrease slightly, dropping rapidly after the 0.05 threshold value. Based on this, we chose 0.06 as the best value for the threshold. At this point, we can detect 92% of all controversial articles with only 10.3% of false positives.

The second validation also confirmed that using the mean level of confidence of controversy for articles can be used to find controversial categories, although we need to choose an appropriate level as a threshold value to be able to correctly detect new controversial articles.

5. Conclusions and future work

Controversy is an integral part of all materials on the Internet. Therefore, there should be a way to predict and warn users about the possible controversy of the content. This is important because users use the Internet for searching information about all aspects of their lives. For example, in a search of health advice, one can easily find the so-called alternative medicine sites. Before following such advice, the person shall be made aware of the difference between academic and alternative approaches and decide which one to trust. The proposed solution should satisfy this need by using combined methods and two-step classification.

In this paper, we present that aggregation at the micro level (articles) can be used to detect controversial categories. Furthermore, this approach can be used to determine potential controversy of a single new article, even if it has yet to contain any editing history.

Future studies should focus on pruning the Wikipedia category structure and determining a method of aggregation of controversy for higher-level categories. Further, searching for certain linguistic patterns typical of dissent or disputes may help identify controversies in Internet texts. Another way to identifying controversy shall be by adding knowledge of the world with the use of an Internet search engine. If one can find contradictory claims and statements relating to the same topic, then the number of such contradictory claims and statements may serve as an indication of controversy, and this relationship shall be researched. The relationship between contradiction and controversy is still untouched and worth researching by the use of linguistic methods.

Acknowledgements

This work is supported by Polish National Science Center grant 2012/05/B/ST6/03364.

References

[1] Biuk-Aghai R.P., Pang C.I., Si Y.W.: Visualizing Large-scale Human Collaboration in Wikipedia. *Future Generation Computet Systems*, vol. 31, pp. 120–133, doi:10.1016/j.future.2013.04.001, 2014, `http://dx.doi.org/10.1016/j.future.2013.04.001`.

[2] Borzymek P., Sydow M., Wierzbicki A.: Enriching Trust Prediction Model in Social Network with User Rating Similarity. In: *Computational Aspects of Social Networks, 2009. CASON '09. International Conference on*, pp. 40–47, 2009, doi: 10.1109/CASoN.2009.30.

[3] Breiman L.: Random Forests. *Machine Learning*, vol. 45(1), pp. 5–32, `http://dx.doi.org/10.1023/A%3A1010933404324`.

[4] Buriol L., Castillo C., Donato D., Leonardi S., Millozzi S.: *Temporal Evolution of the Wikigraph.* IEEE CS Press., Hong Kong, 2006.

[5] Hajian B., White T.: Measuring Semantic Similarity using a Multi-Tree Model. In: *CEUR Workshop Proceedings*, vol. 756, Sun SITE CE, Aachen, Germany.

[6] Han M. S.: Semantic Information Retrieval based on Wikipedia Taxonomy. *International Journal of Computer Applications Technology and Research*, vol. 2(1), pp. 77–80, 2013.

[7] Jankowski-Lorek M., Nielek R., Wierzbicki A., Zielinski K.: Predicting Controversy of Wikipedia Articles Using the Article Feedback Tool. In: *Proceed-*

ings of the 2014 International Conference on Social Computing, SocialCom '14, pp. 22:1–22:7, ACM, New York, NY, USA, 2014, `http://doi.acm.org/10.1145/2639968.2640074`.

[8] Kaptein R., Koolen M., Kamps J.: Using Wikipedia Categories for Ad Hoc Search. In: *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pp. 824–825, ACM, New York, NY, USA, 2009, `http://doi.acm.org/10.1145/1571941.1572147`.

[9] Kittur A., Chi E. H., Suh B.: What's in Wikipedia?: Mapping Topics and Conflict Using Socially Annotated Category Structure. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pp. 1509–1512, ACM, New York, NY, USA, 2009, `http://doi.acm.org/10.1145/1518701.1518930`.

[10] Kittur A., Suh B., Pendleton B. A., Chi E. H.: He Says, She Says: Conflict and Coordination in Wikipedia. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, pp. 453–462, ACM, New York, NY, USA, 2007, `http://doi.acm.org/10.1145/1240624.1240698`.

[11] Medelyan O., Milne D., Legg C., Witten I. H.: Mining Meaning from Wikipedia. *International Journal of Human-Computer Studies*, vol. 67(9), pp. 716–754, 2009, `http://dx.doi.org/10.1016/j.ijhcs.2009.05.004`.

[12] Milne D., Witten I. H.: An Effective, Low-Cost Measure of Semantic Relatedness Obtained from Wikipedia Links. In: *Proceedings of AAAI 2008*, 2008.

[13] Nastase V., Strube M.: Decoding Wikipedia Categories for Knowledge Acquisition. In: *Proceedings of the 23rd National Conference on Artificial Intelligence*, vol. 2, AAAI'08, pp. 1219–1224, AAAI Press, 2008, `http://dl.acm.org/citation.cfm?id=1620163.1620262`.

[14] Rad H. S., Barbosa D.: Identifying Controversial Articles in Wikipedia: A Comparative Study. In: *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*, WikiSym '12, pp. 7:1–7:10, ACM, New York, NY, USA, 2012, `http://doi.acm.org/10.1145/2462932.2462942`.

[15] Schonhofen P.: Identifying Document Topics Using the Wikipedia Category Network. In: *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, WI'06, pp. 456–462, IEEE Computer Society, Washington, DC, USA, 2006, `http://dx.doi.org/10.1109/WI.2006.92`.

[16] Sumi R., Yasseri T., Rung A., Kornai A., Kertész J.: Characterization and prediction of Wikipedia edit wars. In: *Proceedings of the ACM WebSci 11*, 2011.

[17] Sumi R., Yasseri T., Rung A., Kornai A., Kertesz J.: Edit Wars in Wikipedia. In: *IEEE third international conference on social computing (socialcom)*, pp. 724–727, 2011, doi:10.1109/PASSAT/SocialCom.2011.47.

[18] Turek P., Wierzbicki A., Nielek R., Hupa A., Datta A.: Learning About the Quality of Teamwork from Wikiteams. In: *Social Computing (SocialCom), 2010 IEEE Second International Conference on*, pp. 17–24, 2010, doi: 10.1109/SocialCom.2010.13.

[19] Vandamme S., De Turck F.: Algorithms for Recollection of Search Terms Based on the Wikipedia Category Structure, 2014, http://dx.doi.org/10.1155/2014/454868.

[20] Viegas F. B., Wattenberg M., Dave K.: Studying cooperation and conflict between authors with history flow visualizations. In: *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 575–582, 2004, http://dl.acm.org.proxy.lib.umich.edu/citation.cfm?id=985765.

[21] Voss J.: Collaborative thesaurus tagging the Wikipedia way. *CoRR*, vol. abs/cs/0604036, 2006, http://arxiv.org/abs/cs/0604036.

[22] Vuong B. Q., Lim E. P., Sun A., Le M. T., Lauw H. W., Chang K.: On Ranking Controversies in Wikipedia: Models and Evaluation. In: *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, pp. 171–182, ACM, New York, NY, USA, 2008, http://doi.acm.org/10.1145/1341531.1341556.

[23] Wierzbicki A., Turek P., Nielek R.: Learning About Team Collaboration from Wikipedia Edit History. In: *Proceedings of the 6th International Symposium on Wikis and Open Collaboration*, WikiSym '10, pp. 27:1–27:2, ACM, New York, NY, USA, 2010, http://doi.acm.org/10.1145/1832772.1832806.

[24] Yasseri T., Sumi R., Rung A., Kornai A., Kertész J.: Dynamics of conflicts in Wikipedia. *PloS one*, vol. 7(6), p. e38869, 2012.

[25] Yu J., Thom J.A., Tam A.: Ontology Evaluation Using Wikipedia Categories for Browsing. In: *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, pp. 223–232, ACM, New York, NY, USA, 2007, http://doi.acm.org/10.1145/1321440.1321474.

Affiliations

Michał Jankowski-Lorek
Polish-Japanese Academy of Information Technology, Warsaw, Poland,
m.jankowski@pja.edu.pl

Kazimierz Zieliński
Polish-Japanese Academy of Information Technology, Warsaw, Poland,
kazimierz.zielinski@pja.edu.pl

Received: 19.01.2015
Revised: 02.04.2015
Accepted: 03.05.2015