

Wikiganda: Identifying Propaganda Through Text Analysis

Rishi Chandy

Mentors: K. Mani Chandy and Virgil Griffith

Launched in 2001, Wikipedia is now one of the largest collaborative digital encyclopedias in the world, with article quality similar to that of Britannica. Users can edit Wikipedia articles directly and view changes instantly, providing an incentive for contribution. Consequently, millions around the world turn to Wikipedia as a first reference for information about practically anything that might cross their minds.

However, the growth of such publicly accessible information does not come without risks. With few restrictions on article editing, Wikipedia relies on its users to recognize and correct content that fails to adhere to its editing standards, such as conflicts of interest (COI) and biased points-of-view. As such, people seeking to portray history from a prejudiced standpoint or to run misinformation campaigns could potentially run rampant, unchecked by any independent authority, and taint the neutral viewpoint of an otherwise excellent encyclopedic entry.

In 2007, Virgil Griffith released Wikiscanner, an online tool that allows users to trace anonymous Wikipedia edits back to

the organization that owns the editor's IP address, thus exposing revisions made by that organization's employees. Wikiscanner became a sensation as several embarrassing edits from corporations became public and created minor public relations disasters.

Still, Wikiscanner independently is not able to determine if a Wikipedia revision contained propaganda; it merely identifies the source and relies on the user to judge the content. Here, we integrate Wikiscanner with recent research in opinion mining and sentiment analysis, which has provided new insight into the nature of emotion in text, especially in collaborative environments such as Wikipedia. This new tool, Wikiganda, uses text analysis and public data sources to pinpoint organizations that contribute possibly malicious propaganda to Wikipedia; Wikiganda aims to provide an effective platform to detect propaganda automatically. Wikiganda can be explored at <http://www.wikiwatcher.com>.

Categorizing Propaganda

Positive Propaganda, Negative Propaganda, and Vague Propaganda. Positive Propaganda occurs when an editor attempts to portray the subject of the article in a positive light by removing any negative criticism and adding excessive praise, while Negative Propaganda is the opposite. Vague Propaganda exhibits the properties of both Positive and Negative Propaganda.

In the example shown in Table 1, an editor from Exxon Mobil Corporation, according to IP2Location, a database of IP address owners, has added Positive Propaganda to the "Exxon Valdez Oil Spill" article. Casual users reading the article after this revision occurred would not see the text removed from the previous version, such as information about animal deaths, and therefore would assume that this was the only perspective on the subject. In addition, most users would not check the anonymous author's IP address and would fail to see the apparent conflict of interest, taking the biased information at face value.

The data set used in this analysis consisted of 45,315,618 edits by 11,106,420 unique anonymous contributors for 2,385,595 articles from the March 17, 2008 version of Wikipedia. Since the data set was so large and the web interface required propaganda statistics to be shown as users requested them, the algorithm was designed to be quick enough for user requests but applicable to every revision.

Wikiganda analyzes a revision by operating on the text added or deleted due to that revision, called the "diff". For each requested revision, Wikiganda retrieves the text for both that revision and the chronologically previous revision using the Wikipedia Application Programming Interface, which allows access to the raw text of Wikipedia revisions. Then, the changes between the two revisions are found using MediaWiki's native diff tool (the same tool that Wikipedia uses to display diffs). The revision is then classified as propaganda based on the metrics and the decision tree (Figure 1).

Text as of 21:38 December 29, 2004	Text as of 22:03 December 29, 2004 Edited by 192.67.48.156 (Exxon Mobil Corp.)
Thousands of animals perished immediately, the best estimates are: 250,000 sea birds ... and billions of salmon and herring eggs... In the long term, declines have been observed in various marine populations, including stunted growth and indirect mortality increases in pink salmon populations...	Peer-reviewed studies conducted by hundreds of scientists have confirmed that there has been no long-term severe impact to the Prince William Sound ecosystem. Thousands of species in Prince William Sound were never affected by the spill in the first place... As an example, six of the largest salmon harvests in history were recorded in the decade immediately following the spill...

Table 1: An example of Positive Propaganda on Wikipedia for the "Exxon Valdez Oil Spill" article.

Identifying Propaganda

Wikiganda identifies the polarity, or the positive or negative subjectivity, of a given Wikipedia revision using sentiment analysis to approximate the author's feelings. From this information, Wikiganda's propaganda classifier can place the revision into one of the propaganda types mentioned above.

The Wikimedia Foundation provides a complete revision history of every article, which lists the time, editor, and text of each edit. By analyzing this information, Wikiganda scores each revision on a propaganda scale based on several metrics. In addition, the intuitive web interface allows the public to find propaganda on Wikipedia, searching by article or conflicting organization.

Setup

Since one of the secondary goals of Wikiganda was to connect editor IP addresses to organizations and geographic locations, all of the anonymous edits were extracted from the March 17, 2008 Wikipedia database using a Python script. Then, each revision was labeled with the organization or ISP that owned the editor's IP address, according to the IP2Location database.

The web interface allows users to specify opposing "teams" of organizations, such as Microsoft and Yahoo VS. Google, so that Wikiganda would display all articles that one or more "players" from each team had modified. To facilitate this, the labeled revision history was aggregated so that articles were grouped by organization name, thus reducing the time needed to perform the intersect query.

Article Controversy Indicators

of Revisions on Talk Page (+)
of Revisions (+)
of Unique Editors (-)
Table 2: +/- show controversy correlation. "Talk Page" refers to the article's discussion page.

Revisions-level Metrics

Vandalism
Sentiment Detection
WikiTrust values
Conflict of Interest
Table 3: Revision-level Metrics used as indicators for propaganda in Wikipedia revisions.

The web interface allows users to specify opposing "teams" of organizations, such as The respectable Microsoft and Yahoo VS. Google

The Propaganda Metrics

Article-level Metric

Previous research has established indicators (Table 2) to identify controversial articles in Wikipedia. Wikiganda uses Article Controversy as a metric because controversial subjects attract propaganda from opposing interests.

After extracting 2,385,595 unique article names from the revision history, the article-level statistics were calculated for each article based on the revision history database. These statistics, which are shown in the user interface, are used to compute Article Controversy when scoring individual revisions for propaganda.

Revision-level Metrics

The revision-level metrics operate by analyzing the diff and computing a Propaganda Score from 1 to 10. Several user-created automated processes on Wikipedia, such as ClueBot, use heuristics to detect and correct vandalism. Using those heuristics, Wikiganda detects vandalism common in some types of propaganda, such as obscenities and large deletions.

Wikiganda uses a polarity classifier based on a lexicon of over 20,000 words built from the Positive, Negative, PosAff, and NegAff categories from the General Inquirer word list and the prior polarity word list from Wiebe. Every word that is changed in a revision is matched against this lexicon.

Previous sentiment classification research has focused on collections of writings that are flat, such as movie reviews. In contrast, Wikipedia diffs show the words changed from the previous revision. Because of this, classical sentiment classification techniques cannot be applied directly to the diffs. To place

the revision in one of the Propaganda Classes described above, Wikiganda computes the frequency of words added or deleted for each polarity, positive or negative. If the net positive change, or the difference between positive words added and positive words deleted, is greater than the net negative change, then the revision is labeled as Positive Propaganda. The reverse rule applies for Negative Propaganda. If the net positive change is equal to the net negative change, then it is considered Vague Propaganda.

Another important revision-level metric used in Wikiganda is WikiTrust. As part of the WikiTrust investigation into trust and content-driven author reputation in Wikipedia, researchers at the University of California Santa Cruz WikiLab have computed the trust values for revisions up to February 2007. The trust values indicate revision stability, or how long the changes of the revision lasted through the history of the article. Revisions that introduce flagrant propaganda can be unstable since users would recognize and correct them, so these trust values are a useful metric for revision-level propaganda.

The most blatant propaganda may originate from a conflict of interest, so Wikiganda also considers a Conflict of Interest Score. Daniel Erenrich has developed a system that flags revisions tainted by conflicts of interest. This depends on factors, such as the connection between the article and the organization that owns the editor's IP address. For a revision where a user from an Apple IP address edits Wikipedia article for the iPhone would be flagged as an interest. Based on these factors, the system computes a Conflict of Interest Score.

	Predicted Not Propaganda	Predicted Propaganda	Sum
Not true propaganda	93	39	132
True propaganda	25	43	68
Sum	118	82	200

Table 4: Confusion Matrix for Propaganda Classifier. Overall, Wikiganda is 68% accurate.

Evaluation

Using a manually labeled test set of 200 randomly selected revisions, RapidMiner was used to train a decision tree for Wikiganda to score propaganda. The algorithm is implemented in PHP so that Wikiganda can follow the decision tree with input from the web interface.

After stratified ten-fold cross-validation, Wikiganda's propaganda classifier is found to be 68% accurate, with 52.439% Precision (true positive divided by sum of true positive and false positive) and 63.235% Recall (true positive rate). Always choosing the majority class, or "not propaganda," would only give 66% accuracy, so Wikiganda is slightly more accurate.

By combining the Propaganda Classifier with the work of Virgil Griffith and other collaborators in Professor Chandy's Infospheres Laboratory, Wikiganda can approximately identify propaganda and conflicts of interest on Wikipedia.

Future Directions

These methods can be applied to other sources of user-generated content in addition to Wikipedia. In the future, this research could include a deeper analysis of the article revision history to

trace the evolution of specific misinformation. Also, it may be possible to automatically determine an organization's stance on issues based on its edits to related Wikipedia articles.

Further Reading

1. Adler, B., Chatterjee, K., de Alfaro, L., Faella, M., Pye, I., Raman, V. Assigning Trust to Wikipedia Content. WikiSym 2008: International Symposium on Wikis (2008) at <http://www.soe.ucsc.edu/~luca/papers/08/wikisym08-trust.html>
2. Erenrich, D. Wikiscanner: Automated Conflict of Interest Detection of Anonymous Wikipedia Edits. (2008).
3. Stone, P. J., Dunphy, D. C., Smith, M. S., Ogilvie, D. M. 1966. The General Inquirer: A Computer Approach to Content Analysis. MIT Press, Cambridge, MA.
4. Vuong, B. et al. On Ranking Controversies in Wikipedia: Models and Evaluation. Proceedings of the international conference on Web search and web data mining 171-182 (2008).
5. Wilson, T., Wiebe, J., Hoffmann, P. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, 347-354 (2005).

Acknowledgements

Rishi Chandy is a sophomore in Computer Science at Caltech. He would like to thank Professor K. Mani Chandy and Virgil Griffith for their valuable mentorship; B. Thomas Adler, Jason Benterou, Krishnendu Chatterjee, Luca de Alfaro, Ian Pye, and Vishwanath Raman for access to the WikiTrust data; and Daniel Erenrich for access to the Conflict of Interest data. He would also like to thank Charles Slamar and Kristen Vogen of J. Edward Richter Memorial Funds for their generous donation and sponsorship.

"It may be possible to automatically determine an organization's stance on minor issues based on its edits to related Wikipedia articles."

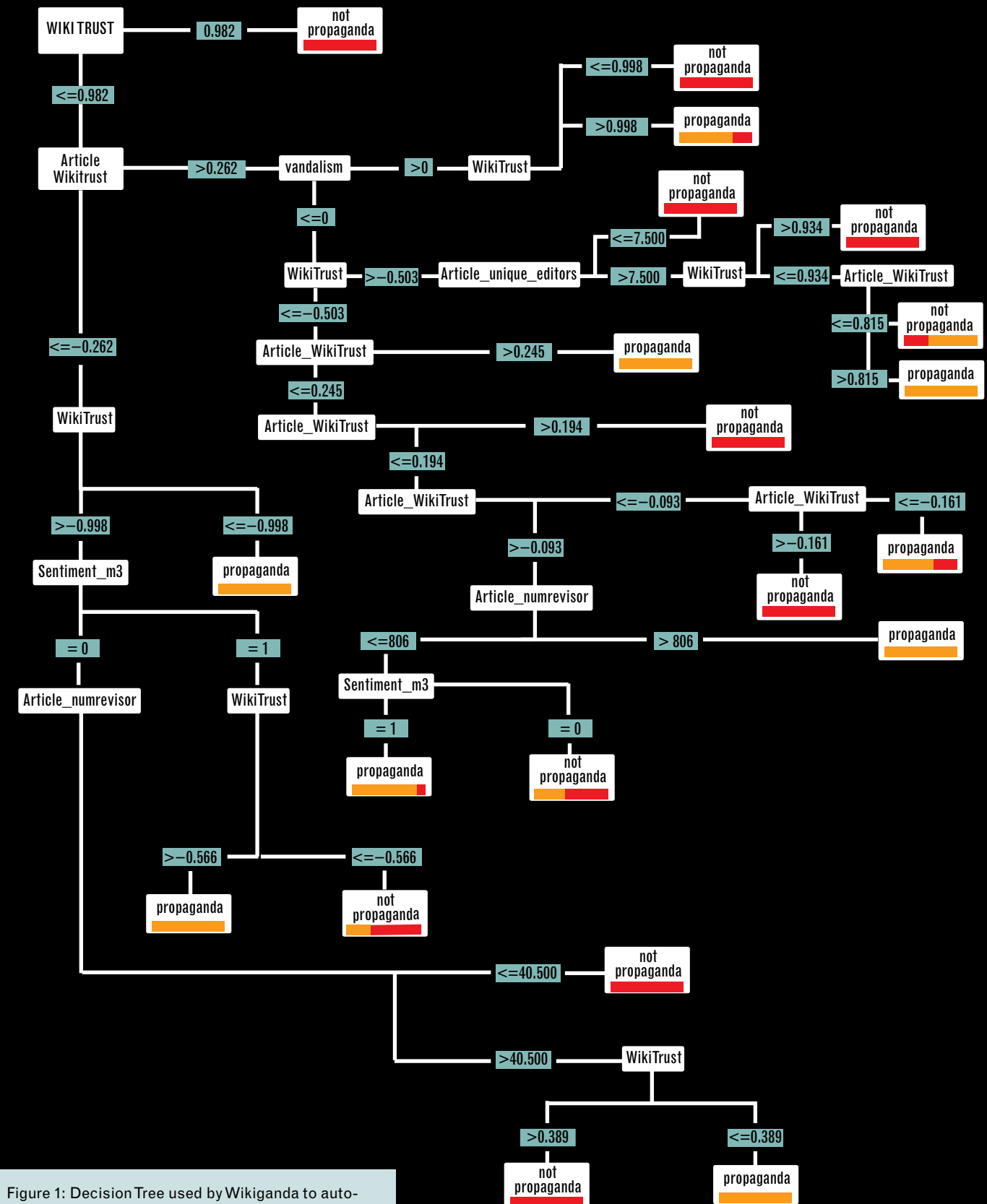


Figure 1: Decision Tree used by Wikiganda to automatically identify propaganda on Wikipedia. Starting at the top, Wikiganda follows the decision tree based on the indicator values to reach a final decision.