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.wikiscanner.com.
In addition, the intuitive web interface allows the public to find each revision on a propaganda scale based on several metrics. By analyzing this information, Wikiganda scores the history of every article, which lists the time, editor, and text of each edit. This information can be used to approximate the author's feelings. From this information, Wikiganda identifies the polarity, or the positive or negative subtext, of a given Wikipedia revision using sentiment analysis. 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 GisheBot, use heuristics to detect and correct vandalism. Using these heuristics, Wikiganda detects vandalism common in some types of propaganda. 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 GisheBot, use heuristics to detect and correct vandalism. Using these heuristics, Wikiganda detects vandalism common in some types of propaganda.

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

Table 2: +/- Controversy correlation. Talk Page refers to the article’s discussion page.

### The Propaganda Metrics

#### Article-level Metrics

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

After extracting 2,385,599 unique article names from the revision history, the article-level statistics were calculated for each article based on the revision history database. Those 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 GisheBot, use heuristics to detect and correct vandalism. Using these heuristics, Wikiganda detects vandalism common in some types of propaganda.

### Sentiment Detection

*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).*

### Vandalism

*Vandalism can be identified using two metrics: the number of revisions for an article and the number of unique editors who contributed to those revisions. If the number of revisions is significantly higher than the number of unique editors, it suggests vandalism.*

### Conflict of Interest

*To identify conflicts of interest, Wikiganda uses the IP address of the editor and maps it to a financial organization or a geographic location using the IP2Location database. For example, if an editor from a university IP address edits the “Exxon Valdez Oil Spill” article, a trust value is assigned to that editor. If an editor from an Apple IP address edits the “Exxon Valdez Oil Spill” article, a trust value is assigned to that editor. The system computes a Conflict of Interest Score for each revision.*

### 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.*

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

*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.*

*Another important revision-level metric used in Wikiganda is 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).*
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

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 Adie, Jason Bentecu, Krithnudu 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 Blame and Kristin Venö 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.