Machine Learning and the Detection of Anomalies in Wikipedia

Mentor Hamiti, Arsim Susuri and Agni Dika

***Abstract*—**This work analyses the current trend in applying machine learning in detection of anomalies, with the specific aim of analyzing anomalies in Wikipedia articles. Ever since it was created, in 2001, Wikipedia has grown with immense speed, enabling anyone the ability to edit articles, thus, establishing itself as one of the largest information sources on the Internet. Having become this popular, Wikipedia has become the source of an ever-increasing number of articles, created, modified and enhanced by different editors and, inadvertently, susceptible to various acts of vandalisms. This article aims to provide an overview of the initial research and developments in the field of machine learning applications in detecting anomalies in Wikipedia and future trends.

***Keywords*—**machine learning, Wikipedia, anomalies, vandalism, detection of anomalies.

I. Introduction

Ever since its inception, in 2001, Wikipedia has continuously grown to become one the largest information source on the Internet. One of its unique features is that it offers the ability to anyone to edit the articles. This popularity, in itself, means that, a number of articles can be read, edited, and enhanced by different editors and, inevitably, be subject to acts of vandalisms through illegitimate editing.

Vandalism means any type of editing which damages the reputation of an article or a user in Wikipedia. A list of typical vandalisms along with their chances of appearance, as shown in Fig. 1, was created as a result of empirical studies done by Priedhorsky et al. [1]. Typical examples include massive deletions, spam, partial deletions, offences and misinformation.

In order to deal with vandalism, Wikipedia relies on the following users:

- Wikipedia its users' ability and willingness to find (accidentally or deliberately) damaged articles
- Wikipedia administrators and
- Wikipedia users with additional privileges

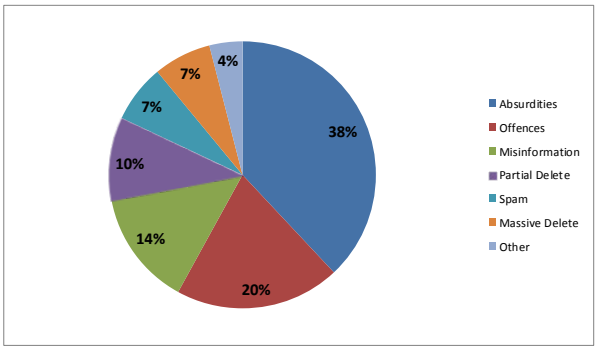

Fig. 1. Categories of vandalism based on empirical approach [1]

These users use special tools (e.g. Vandal Fighters) to monitor recent changes and modifications that enable retrieval of bad expressions or which are implemented by blacklisted users.

Wikipedia was subject to different statistical analysis from various authors. Viégas et al. [2] uses visualization tools to analyze the history of Wikipedia articles. When it comes to vandalism, authors were able to identify (manually) massive deletions as a jump in the history flow of a particular article page.

Since late 2006, some bots (computer programs designed to detect and revert vandalism), have appeared on Wikipedia. These tools are built on the primitive included in the Vandal Fighters. These use lists of common phrases, and consult databases containing blocked users or IP addresses in order to separate legitimate editing from vandalism.

One drawback of these approaches is emphasized that these world use static list of obscenities and grammatical rules which are difficult to maintain and easily "fooled". These detect only 30% of vandalisms committed.

Consequently, there is a need to improve the detection of this kind. One of the possible improvements is the application of machine learning.

The prior success implemented in interference detection, spam filtering for email, etc., is a good indicator for the opportunity that the machine learning shows in improvements in detecting anomalies in Wikipedia.

ISBN: 978-1-61804-318-4

II. Wikipedia Vandalism Detection

To define the vandalism detection task, we have to define some key concepts of MediaWiki (the wiki engine used by Wikipedia).

An article is composed of a sequence of revisions, commonly referred to as the article history. A revision is the state of an article at a given time in its history and is composed of the textual content and metadata describing the transition from the previous revision [3].

Revision metadata contains, among others, the user who performed the edit, a comment explaining the changes, a timestamp, etc. An edit is a tuple of two consecutive revisions and should be interpreted as the transition from a given revision to the next one. Wikipedia vandalism detection is a one-class classification task.

The goal is, given any edit, determine whether it is destructive or not. Through machine learning, anomalous contributions (edits) can be detected by inspecting Wikipedia edits. An edit $e = (r_-, r_+)$ is defined as a set of two consecutive revisions of an article which contains the original revision (r_-) and the new revision (r_+) once the changes have been submitted.

A revision r is a version of a Wikipedia article that, besides the article markup text, includes additional data (meta data) about the latest editing, such as the editor's user identification, his/her comment on the nature of the changes made, and a timestamp at which he/she edited the article.

Evaluating a vandalism detection system requires a corpus of pre-classified edits. Four different corpora have been reported in the literature:

1. Webis-WVC-07 - The Webis Wikipedia Vandalism Corpus 2007 (Webis-WVC-07) was the first Wikipedia vandalism corpus and consists of 940 human-annotated edits of which 301 are labelled as vandalism. It was compiled in 2007 and was first used by Potthast et al. [4]. English Wikipedia was the sole source for all edits.
2. PAN-WVC-10 - The PAN Wikipedia Vandalism Corpus 2010 (PAN-WVC-10), compiled in 2010 via Amazon's Mechanical Turk comprises 32439 edits from 28468 English Wikipedia articles of which 2394 have been annotated as vandalism.
   The dataset was created by 753 human annotators by casting 193022 votes, so that each edit has been annotated at least three times, whereas edits that were difficult to be annotated received more than three votes (Potthast [5]). The PAN-WVC-10 was first used in the 1st International Competition on Wikipedia Vandalism Detection (Potthast et al. [6]).
3. PAN-WVC-11 - The PAN Wikipedia Vandalism Corpus 2011 (PAN-WVC-11) from 2011 is an extension of the PAN-WVC-10. It was used in the 2nd International Competition on Wikipedia Vandalism Detection (Potthast and Holfeld [7]) and is the first multilingual vandalism detection corpus. The corpus comprises 29949 Wikipedia edits in total (9985 English edits with 1144 vandalism, 9990 German edits with 589 vandalism, and 9974 Spanish edits with 1081 vandalism annotations).
4. Wikipedia History Dump Wikipedia records all revisions of all articles and all other Wikipedia pages and releases them as XML or SQL dump files.

A. Wikipedia Bots

The vandalism problem on Wikipedia is probably as old as the encyclopedia itself. Kittur et al. [8] observe that the total number of vandalism edits is increasing over time. Although they report the total vandalism proportion to remain at the same level, increasing vandalism is a serious objective in the online encyclopedia.

To tackle this problem, the Wikipedia community resorts to manually protecting articles from being edited in case they are heavily vandalized.

Additionally, since 2006, vandalism detection bots are used, which automatically patrol for vandalism edits and partially revert them. Most often these bots use simple heuristic rules, word blacklists, and lists of blocked user IPs to identify vandalism edits (e.g. VoABot II or ClueBot).

The ClueBot NG bot which replaces ClueBot, uses machine learning approaches. It tries to enhance the heuristics-based techniques, which were difficult to maintain and easy to bypass. The bot uses a pre-classified edit dataset annotated by Wikipedia users to train an Artificial Neural Network.

AVBOT [9] is a bot created to automatically search for any vandalisms edits in Spanish articles of Wikipedia. So far, it has reverted more than 200,000 vandalism edits [10].

B. Approaches based on Machine Learning

Since 2008 Wikipedia vandalism detection based on machine learning approaches has become a field of increasing research interest. In Table 1 existing vandalism detection approaches from the literature are shown.

Potthast et al. [4] contributed the first machine learning vandalism detection approach using textual features as well as basic meta data features with a logistic regression classifier. Smets et al. [11] used a Naive Bayes classifier on a bag of words edit representation and were the first to use compression models to detect Wikipedia vandalism. Itakura and Clarke [12] used Dynamic Markov Compression to detect vandalism edits on Wikipedia.

Mola Velasco [13] extended the approach of Potthast et al. [4] by adding some additional textual features and multiple wordlist-based features. He was the winner of the 1st International Competition on Wikipedia Vandalism Detection (Potthast et al. [6]).

ISBN: 978-1-61804-318-4

TABLE I

VANDALISM DETECTION CLASSIFICATION OBTAINED FROM VARIOUS AUTHORS

Authors	Balanced Data	Classifier	Precision	Recall	PR-AUC	Corpora
Smets et al. [11]	x	Probablistic Sequence Modeling	0.3209	0.9171	-	Simplewiki
Smets et al. [11]	x	Naive Bayes	0.4181	0.5667	-	Simplewiki
Tran and Christen [20]	√	Gradient Tree Boosting	0.870	0.870	-	Historical Dump
Potthast et al. [4]	x	Logistic Regression	0.830	0.870	-	Webis-WVC-07
Velasco [3]	x	Random Forest	0.860	0.570	0.660	PAN-WVC-10
Adler et al. [15]	x	ADTree	0.370	0.770	0.490	PAN-WVC-10
Adler et al. [17]	x	Random Forest	-	-	0.820	PAN-WVC-10
West and Lee [18]	x	ADTree	0.370	0.770	0.490	PAN-WVC-10
Harpalani et al. [19]	x	LogitBoost	0.606	0.608	0.671	PAN-WVC-10
West and Lee [18]	x	ADTRee	-	-	0.820	PAN-WVC-11

West et al. [14] were among the first to present a vandalism detection approach solely based on spatial and temporal meta data, without the need to inspect article or revision texts.

Adler et al. [15], in a similar fashion, built a vandalism detection system on top of their WikiTrust reputation system (Adler and De Alfaro [16]). Adler et al. [17] combined natural language, spatial, temporal and reputation features used in their aforementioned works (Adler et al. [15], Mola Velasco [13], West et al. [14]). Besides Adler et al. [17], West and Lee [18] were the first to introduce ex post facto data as features, for whose calculation also future revisions have to be considered.

Their resulting multilingual vandalism detection system was the winner at the 2nd International Competition on Wikipedia Vandalism Detection (Potthast and Holfeld [7]).

Harpalani et al. [19] stated vandalism edits to share unique linguistic properties. Thus, they based their vandalism detection system on a stylometric analysis of vandalism edits by probabilistic context-free grammar models. They showed that this approach outperforms features based on shallow patterns, which match syntactic structures and text tokens.
Supporting the current trend of creating cross language vandalism classifiers, Tran and Christen [20] evaluated multiple classifiers based on a set of language independent features that were compiled from the hourly article view counts and Wikipedia's complete edit history.

C. Features of Anomalies

The literature provides an ever-growing set of features that are employed to model anomalous edits. After the first contributions to the Wikipedia vandalism detection task, most authors used a subset of existing features and added some new ones to their approaches.

Tables 2 provide an overview of textual data anomalies features that were used so far in the literature.

For the sake of simplicity, we use the following abbreviations to distinguish various authors: A17 (Adler et al. [17]), G14[1], J22 (Javanmardi et al. [22]), M3 (Mola Velasco [3]), P4 (Potthast et al. [4]), W18 (West and Lee [18]), and Wa21 (Wang and McKeown [21]).

Textual features are calculated by analyzing the new revision's markup text or rather both revisions' markup texts of an edit. Meta data features are compiled from the revision's meta data or are calculated by analyzing additional Wikipedia data, such as history dumps or article dumps.

While Mola Velasco [3] used three feature categories by considering textual, meta data and language features, his language features (wordlist-based features) could be categorized as textual features. Javanmardi et al. [22] split their features into four categories, namely textual, meta data, user and language model. The user category comprises user-related meta data features.

[1] https://github.com/webis-de/wikipedia-vandalism-detection

TABLE II

SOME TEXTUAL FEATURES USED BY VARIOUS AUTHORS, DESCRIBING ANOMALOUS EDITS IN WIKIPEDIA

Category	Feature	A17	G14	J22	M3	P4	W18	Wa21
Frequency	All words	√	√	√	√			
	Average term	√	√		√	√		
	Bad words	√	√	√	√			
	Biased words	√	√	√	√			
	Emoticons		√					
	Good/markup words	√	√	√	√			
	Sex words	√	√	√	√			
	Vulgarism		√	√	√	√	√	√
	Web slang							√
Impact	All words	√	√	√	√			
	Bad words	√	√	√	√			
	Emoticons		√					
	Good/markup words	√	√	√	√			
	Sex words	√	√	√	√			
	Vulgarism		√	√	√	√	√	
Ratio	Alphanumeric	√	√	√			√	
	Non-alphanumeric		√	√	√			
	Size	√	√	√	√	√		√
	Upper to all	√	√	√	√	√	√	
	Upper to lower	√	√		√			
Other	Blanking		√	√				
	Character diversity		√		√			
	Characters added or removed		√				√	
	Compressibility	√	√	√	√	√		
	Context relation					√		
	Digit ratio	√	√	√	√			
	External links added		√	√				
	Inserted wiki markup						√	
	Inserted words			√				
	Internal links added		√	√				
	Longest char sequence	√	√	√	√	√	√	
	Longest word	√	√	√	√	√	√	
	Punctuation misuse							√
	Removed words			√				
	Replacement similarity		√			√		
	Size increment	√	√	√	√		√	

ISBN: 978-1-61804-318-4

III. Conclusion

This brief review shows the overall progress in applying machine learning in detecting anomalies in Wikipedia. The problem of vandalism has grown over the years, along with the growth of popularity of Wikipedia. Applying machine learning as a tendency to automate detection of vandalisms is a great opportunity for maintaining and improving the credibility of Wikipedia, without compromising the ability of Various Wikipeida users to enhance articles online through editing.

Having in mind that in order to properly implement machine learning in detection of anomnalies there is a requirement to properly characterize anomalies. This is why implementation of specific features of anomalies in creating machine learning based anomaly detectors.

Based on our research, we can also conclude that, apart from English, German, French and Spanish, little or no progress is made in other language sections of Wikipedia, thus providing excellent grounds for future research. Furthermore, development of new language– independent methods to enhance detection of anomalies could improve the effect of machine learning approach on the credibility of Wikipedia.

References

[1] R. Priedhorsky, J. Chen, S. T. K. Lam, K. Panciera, L. Terveen, and J. Riedl. "Creating, destroying, and restoring value in Wikipedia," in proceedings of the international ACM conference on supporting GroupWork (GROUP), Sanibel Island, FL, pp 259-268, 2007.

[2] F. B. Viégas, M. Wattenberg, and K. Dave, "Studying cooperation and conflict between authors with history flow visualizations," in proceedings of the ACM Conference on human factors in computing systems (CHI), Vienna, Austria, pp 575-582, 2004.

[3] Santiago M. Mola-Velasco, "Wikipedia Vandalism Detection," - WWW 2011, Hyderabad, India. 2011.

[4] Martin Potthast, Benno Stein, and Robert Gerling, "Automatic vandalism detection in wkipedia," in advances in information retrieval, pp 663-668. Springer Berlin Heidelberg, 2008.

[5] Martin Potthast, "Crowdsourcing a wikipedia vandalism corpus," proceeding of the 33rd international ACM SIGIR conference on research and development in information retrieval - SIGIR '10, p 789, 2010.

[6] Martin Potthast, Benno Stein, and Teresa Holfeld, "Overview of the 1st International Competition on Wikipedia Vandalism Detection," in Martin Braschler, Donna Harman, and Emanuele Pianta, editors, Working Notes Papers of the CLEF 2010 Evaluation Labs, September 2010.

[7] Martin Potthast and Teresa Holfeld "Overview of the 2nd International Competition on Wikipedia Vandalism Detection", in Vivien Petras, Pamela Forner, and Paul D. Clough, editors, Notebook Papers of CLEF 11 Labs and Workshops, September 2011.

[8] Aniket Kittur, Bongwon Suh, Bryan A. Pendleton, and Ed H. Chi, "He says, she Says: conflict and coordination in Wikipedia;" in ACM Conference on Human Factors in Computing Systems, pp 453-462, 2007.

[9] Emilio-José Rodriguez-Posada, "AVBOT: Detecting and fixing Vandalism in Wikipedia" in proceedings of the CEPIS UPGRADE, vol. XII, issue no. 3, 2011.

[10] http://es.wikipedia.org/wiki/Especial:Contribuciones/AVBOT.

[11] Koen Smets, Bart Goethals, and Brigitte Verdonk, "Automatic vandalism detection in wikipedia: Towards a machine learning approach," in WikiAI '08: Proceedings of the AAAI Workshop on Wikipedia and Artificial Intelligence, 2008.

[12] Kelly Y. Itakura and Charles L. a. Clarke, "Using dynamic markov compression to detect vandalism in the Wikipedia," Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval - SIGIR '09, p 822, 2009.

[13] Mola Velasco Santiago Moisés Mola Velasco, "Wikipedia vandalism detection through machine learning: Feature review and new proposals - lab report for pan at clef 2010," in CLEF (Notebook Papers/LABs/Workshops), 2010.

[14] Andrew G. West, Sampath Kannan, and Insup Lee, "Detecting wikipedia vandalism via spatio-temporal analysis of revision metadata," in Proceedings of the Third European Workshop on System Security, EUROSEC '10, pp 22-28, New York, NY, USA, 2010.

[15] B. Thomas Adler, Luca De Alfaro, and Ian Pye, "Detecting wikipedia vandalism using wikitrust," notebook papers of CLEF, 2010.

[16] B. Thomas Adler and Luca De Alfaro, "A content-driven reputation system for the wikipedia," proceedings of the 16th international conference on World Wide Web WWW 07, 7(Generic):261, 2007.

[17] B. Thomas Adler, Luca De Alfaro, Santiago M. Mola-Velasco, Paolo Rosso, and Andrew G. West, "Wikipedia vandalism detection: Combining natural language, metadata, and reputation features," in proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing - Volume Part II, pp 277-288, Berlin, Heidelberg, 2011.

[18] Andrew G. West and Insup Lee, "Multilingual vandalism detection using language-independent & ex post facto evidence - notebook for pan at clef 2011. In CLEF (Notebook Papers/Labs/Workshop), 2011.

[19] Manoj Harpalani, Michael Hart, S Signh, Rob Johnson, and Yejin Choi, "Language of Vandalism: Improving Wikipedia Vandalism Detection via Stylometric Analysis," ACL (Short Papers), (2009):pp 83-88, 2011.

[20] Khoi-Nguyen Tran and Peter Christen, "Cross Language Prediction of Vandalism on Wikipedia Using Article Views and Revisions," Advances in Knowledge Discovery and Data Mining, pp 268-279, 2013.

[21] William Yang Wang and Kathleen R. McKeown, "Got You!: Automatic vandalism detection in Wikipedia with web-based shallow syntactic-semantic modelling," in Proceedings of the 23rd International Conference on Computational Linguistics, COLING '10, pp 1146-1154, Stroudsburg, PA, USA, 2010.

[22] Sara Javanmardi, David W. McDonald, and Cristina V. Lopes, "Vandalism detection in wikipedia: A high-performing, feature-rich model and its reduction through lasso," in Proceedings of the 7th International Symposium on Wikis and Open Collaboration, WikiSym '11, pp 82-90, New York, NY, USA, 2011.