Detecting Promotional Content in Wikipedia

Shruti Bhosale **Heath Vinicombe** **Raymond J. Mooney**
Department of Computer Science
The University of Texas at Austin
{shruti,vini,mooney}@cs.utexas.edu

Abstract

This paper presents an approach for detecting promotional content in Wikipedia. By incorporating stylometric features, including features based on n-gram and PCFG language models, we demonstrate improved accuracy at identifying promotional articles, compared to using only lexical information and meta-features.

1 Introduction

Wikipedia is a free, collaboratively edited encyclopedia. Since normally anyone can create and edit pages, some articles are written in a promotional tone, violating Wikipedia's policy requiring a neutral viewpoint. Currently, such articles are identified manually and tagged with an appropriate Cleanup message[1] by Wikipedia editors. Given the scale and rate of growth of Wikipedia, it is infeasible to manually identify all such articles. Hence, we present an approach to automatically detect promotional articles.

Related work in quality flaw detection in Wikipedia (Anderka et al., 2012) has relied on meta-features based on edit history, Wikipedia links, structural features and counts of words, sentences and paragraphs. However, we hypothesize that there are subtle differences in the linguistic style that distinguish promotional tone, which we attempt to capture using stylometric features, particularly deeper syntactic features. We model the style of promotional and normal articles using language models based on both n-grams and Probabilistic Context Free Grammars (PCFGs). We show that using such stylometric features improves over using only shallow lexical and meta-features.

2 Related Work

Anderka et al. (2012) developed a general model for detecting ten of Wikipedia's most frequent quality flaws. One of these flaw types, "Advert"[2], refers to articles written like advertisements. Their classifiers were trained using a set of lexical, structural, network and edit-history related features of Wikipedia articles. However, they used no features capturing syntactic structure, at a level deeper than Part-Of-Speech (POS) tags.

A related area is that of vandalism detection in Wikipedia. Several systems have been developed to detect vandalizing edits in Wikipedia. These fall into two major categories: those analyzing author information and edit metadata (Wilkinson and Huberman, 2007; Stein and Hess, 2007); and those using NLP techniques such as n-gram language models and PCFGs (Wang and McKeown, 2010; Harpalani et al., 2011). We combine relevant features from both these categories to train a classifier that distinguishes promotional content from normal Wikipedia articles.

3 Dataset Collection

We extracted a set of about 13,000 articles from English Wikipedia's category, "Category:All arti-

[1] http://en.wikipedia.org/wiki/Wikipedia:Template_messages/Cleanup

[2] "Advert" is the flaw-type of majority of the articles in the Category 'Articles with a promotional tone'.

Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1851–1857,
Seattle, Washington, USA, 18-21 October 2013.

Content Features
Number of characters
Number of words
Number of sentences
Average Word Length
Average, Minimum, Maximum Sentence Lengths, Ratio of Maximum to minimum sentence lengths
Ratio of long sentences (>48 words) to Short Sentences (<33 words)
Percentage of Sentences in the passive voice
Relative Frequencies of POS tags for pronouns, conjunctions, prepositions, auxiliary verbs, modal verbs, adjectives and adverbs
Percentage of sentences beginning with a pronoun, article, conjunction, preposition, adjective, adverb
Percentage of special phrases[3] such as peacock terms ('legendary', 'acclaimed', 'world-class'), weasel terms ('many scholars state', 'it is believed/regarded', 'many are of the opinion', 'most feel', 'experts declare', 'it is often reported') , editorializing terms ('without a doubt', 'of course', 'essentially')
Percentage of easy words, difficult words (Dale-Chall List), long words and stop words
Overall Sentiment Score based on SentiWordNet[4]

Table 1: Content Features of a Wikipedia Article

cles with a promotional tone" as a set of positive examples. We extracted a set of 26,000 untagged articles to form a noisy set of negative examples, which may contain some promotional articles that have not yet been tagged by Wikipedia editors. To counter this noise, we repeated the experiment using Wikipedia's Featured Articles and Good Articles (approx. 11,000) as a set of clean negative examples. We used 70% of the articles in each category to train language models for each of the three categories (promotional articles, featured/good articles, untagged articles), and used the remaining 30% to evaluate classifier performance using 10-fold cross-validation.

4 Features

4.1 Content and Meta Features of an Article

We used the content and meta features proposed by Anderka et al. (2012) as given in Tables 1-4. We also

[3] http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Words_to_watch

[4] This feature is not included in Anderka et al. (2012)

Structural Features
Number of Sections
Number of Images
Number of Categories
Number of Wikipedia Templates used
Number of References, Number of References per sentence and Number of references per section

Table 2: Structural Features of a Wikipedia Article

Wikipedia Network Features
Number of Internal Wikilinks (to other Wikipedia pages)
Number of External Links (to other websites)
Number of Backlinks (i.e. Number of wikilinks from other Wikipedia articles to an article)
Number of Language Links (i.e. Number of links to the same article in other languages)

Table 3: Network Features of a Wikipedia Article

added a new feature, "Overall Sentiment Score" for an article. This feature is the average of the sentiment scores assigned by SentiWordnet (Baccianella et al., 2010) to all positive and negative sentiment bearing words in an article. In total, this results in 58 basic document features.

4.2 N-Gram Language Models

Language models are commonly used to measure stylistic differences in language usage between authors. For this work, we employed them to model the difference in style of neutral vs. promotional Wikipedia articles. We trained trigram word language models and trigram character language models[5] with Witten-Bell smoothing to produce probabilistic models of both classes.

4.3 PCFG Language Models

Probabilistic Context Free Grammars (PCFG) capture the syntactic structure of language by modeling sentence generation using probabilistic CFG productions. We hypothesize that sentences in promotional articles and those in neutral articles tend to have different kinds of syntactic structures and therefore, we explored the utility of PCFG models for detecting this difference. Since we do not have ground-truth parse trees for sentences in our dataset,

[5] Modeling longer character sequences did not help.

Features based on PCFG models and n-gram Language models
Difference in the probabilities assigned to an article by the positive and the negative class *character trigram language models* (LM_char_trigram)
Difference in the probabilities assigned to an article by the positive and the negative class *word trigram language models* (LM_word_trigram)
Difference in the *mean* values of the probabilities assigned to sentences of an article by the positive and negative class *PCFG models* (PCFG_mean)
Difference in the *maximum* values of the probabilities assigned to sentences of an article by the positive and negative class *PCFG models* (PCFG_max)
Difference in the *minimum* values of the probabilities assigned to sentences of an article by the positive and negative class *PCFG models* (PCFG_min)
Difference in the *standard deviation* values of the probabilities of sentences of an article by the positive and negative class *PCFG models* (PCFG_std_deviation)

Table 5: Features of a Wikipedia Article based on PCFG models and n-gram Language models

Edit History Features
Age of the article
Days since last revision of the article
Number of edits to the article
Number of unique editors
Number of edits made by registered users and by anonymous IP addresses
Number of edits per editor
Percentage of edits by top 5% of the top contributors to the article

Table 4: Edit-History Features of a Wikipedia Article

we followed the method of (Raghavan et al., 2010; Harpalani et al., 2011), which uses the output of the Stanford parser to train PCFG models for stylistic analysis. We used the PCFG implementation of Klein and Manning (2003) to learn a PCFG model for each category.

4.4 Classification

The n-gram and PCFG language models were used to create a set of additional document features. We used the probability assigned by the language models to each sentence in a test document to calculate document-wide statistics such as the mean, maximum, and minimum probability and standard deviation in probabilities of the set of sentences in an article. The language-modeling features used are shown in Table 5.

Since we have a wide variety of features, we experimented with various ensemble learning techniques and found that LogitBoost performed best empirically. We used the Weka implementation of LogitBoost (Friedman et al., 2000) to train a classifier using various combinations of features. We used Decision Stumps as a base classifier and ran boosting for 500 iterations.

5 Experimental Evaluation

5.1 Methodology

We used 10-fold cross-validation to test the performance of our classifier using various combinations of features. We ran the classifier on the portion (30%) of the dataset not used for language modeling.[6] We measured overall classification accuracy as well as precision, recall, F-measure, and area under the ROC curve for all experiments. We tested performance in two settings (Anderka et al., 2012):

- *Pessimistic Setting*: The negative class consists of articles from the Untagged set. Since some of these could be manually undetected promotional articles, the accuracy measured in this setting is probably an under-estimate.
- *Optimistic Setting*: The negative class consists of articles from the Featured/Good set. These articles are at one end of the quality spectrum, making it relatively easier to distinguish them from promotional articles.

The true performance of the classifier is likely somewhere between that achieved in these two settings.

[6] We maintain an equal number of positive and negative test cases in both the settings.

Features	Pessimistic Setting				Optimistic Setting			
	P	R	F1	AUC	P	R	F1	AUC
Bag-of-words Baseline	0.823	0.820	0.821	0.893	0.931	0.931	0.931	0.979
PCFG	0.881	0.870	0.865	0.903	0.910	0.910	0.910	0.961
Character trigrams	0.889	0.887	0.888	0.952	0.858	0.843	0.841	0.877
Word trigrams	0.863	0.863	0.863	0.931	0.887	0.883	0.882	0.931
Character trigrams + Word trigrams	0.89	0.888	0.889	0.952	0.908	0.907	0.907	0.962
PCFG+Char. trigrams+Word trigrams	0.914	0.915	0.914	0.974	0.950	0.950	0.950	0.983
58 Content and Meta Features	0.866	0.867	0.867	0.938	0.986	0.986	0.986	0.996
All Features	**0.940**	**0.940**	**0.940**	**0.986**	**0.989**	**0.989**	**0.989**	**0.997**

Table 6: Performance (Precision(P), Recall(R), F1 score, AUC) of the classifier in the two settings

5.2 Results for Pessimistic Setting

From Table 6, we see that all features perform better than the bag-of-words baseline. We also see that character trigrams, one of the simplest features, gives the best individual performance. However, deeper syntactic features using PCFGs also performs quite well. Combining all of the language-modeling features (PCFG + character trigrams + Word trigrams) further improves performance. Compared to the 58 content and meta features utilized by Anderka et al., (2012) described in Section 4.1, the PCFG and character trigram features give much better performance, both individually and when combined. It is interesting to note that adding Anderka et al.'s features to the language-modeling ones gives a fairly small improvement in performance. This validates our hypothesis that promotional articles tend to have a distinct linguistic style that is captured well using language models.

5.3 Results for Optimistic Setting

In the Optimistic Setting, as shown in Table 6, the content and meta features give the best performance, which improves only slightly when combined with language-modeling features. The bag-of-words baseline performs better than all the language modeling features. This performance could be because there is a much clearer distinction between promotional articles and featured/good articles that can be captured by simple features alone. For example, featured/good articles are generally longer than usual Wikipedia articles and have more references.

5.4 Top Ranked Features and their Performance

To analyze the performance of different features, we determined the top ranked features using Information Gain. In the Pessimistic Setting, the top six features are all language-modeling features (character trigram model feature works best), followed by basic meta-features such as character count, word count, category count and sentence count. The new feature we introduced, "Overall Sentiment Score" is the 18th most informative feature in the pessimistic setting, indicating that the cumulative sentiment of a bag of words is not as discriminative as we would intuitively assume. Using the 10 top-ranked features, we get an F1 of 0.93, which is only slightly worse than that achieved using all features (F1 = 0.94).

In the Optimistic Setting, the top-ranked features are the number of references and the number of references per section. This is consistent with the observation that featured/good articles have very long and comprehensive lists of references, since Wikipedia's fundamental policy is to maintain verifiability by citing relevant sources. Features based on the n-gram and PCFG models also appear in the list of ten best features. Using only the top 10 features, gives an F1 of 0.988, which is almost as good as using all features (F1 = 0.989).

5.5 Optimistic and Pessimistic Settings

In the optimistic setting, there is a clear distinction between the positive (promotional) and negative (featured/good) classes. But there are only subtle differences between the positive and negative (untagged articles) classes in the pessimistic setting.

Best Features in Pessimistic Setting	Best Features in Optimistic Setting
LM_char_trigram	Number of References
LM_word_trigram	Number of References per Section
PCFG_min	LM_word_trigram
PCFG_max	Number of Words
PCFG_mean	PCFG_mean
PCFG_std_deviation	Number of Sentences
Number of Characters	LM_char_trigram
Number of Words	Number of Words
Number of Categories	Number of Characters
Number of Sentences	Number of Backlinks

Table 7: Top 10 Features (listed in order) in both Settings ranked using Information Gain

These two classes are superficially similar, in terms of length, reference count, section count etc. Stylometric features based on the trained language models are successful at detecting the subtle linguistic differences in the two types of articles. This is useful because the pessimistic setting is closer to the real-world setting of articles in Wikipedia.

5.6 Error Analysis

Since the pessimistic setting is close to the real setting of Wikipedia articles, it is useful to do an error analysis of the classifier's performance in this setting. There is an approximately equal proportion of false positives and false negatives.

A significant number of false positives seem to be cases of manually undetected promotional articles. This demonstrates the practical utility of our classifier. But there are also many false positives that seem to be truly unbiased. These articles appear to have been poorly written, without following Wikipedia's editing policies. Examples include use of very long lists of nouns, use of ambiguous terms like "many believe" and excessive use of superlatives. Other common characteristics of most of the false positives are presence of a considerable number of complex sentences with multiple subordinate clauses. These stylistic cues seem to be misleading the classifier.

A common thread underlying most of the false negatives is the fact that they are written in a narrative style or they have excessive details in terms of the content. Examples include narrating a detailed story of a fictional character in an unbiased manner or writing a minutely detailed account of the history of an organization. Another source of false negatives comes from biographical Wikipedia pages which are written in a resume style, listing all their qualifications and achievements. These cues could help one manually detect that the article, though not promotional in style, is probably written with the view of promoting the entity. As possible future work, we could incorporate features derived from language models for narrative style trained using an appropriate external corpus of narrative text. This might enable the classifier to detect some cases of unbiased promotional articles.

6 Conclusion

Our experiments and analysis show that stylometric features based on n-gram language models and deeper syntactic PCFG models work very well for detecting promotional articles in Wikipedia. After analyzing the errors that are made during classification, we realize that though promotional content is non-neutral in majority of the cases, there do exist promotional articles that are neutral in style. Adding additional features based on language models of narrative style could lead to a better model of Wikipedia's promotional content.

7 Acknowledgements

This research was supported in part by the DARPA DEFT program under AFRL grant FA8750-13-2-0026 and by MURI ARO grant W911NF-08-1-0242. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the view of DARPA, AFRL, ARO, or the US government.

References

Maik Anderka, Benno Stein, and Nedim Lipka. 2012. Predicting quality flaws in user-generated content: the case of Wikipedia. In *Proceedings of the 35th International ACM SIGIR Conference on Research and development in Information Retrieval*, SIGIR '12, pages 981–990, New York, NY, USA. ACM.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC'10), Valletta, Malta, May*.

Joachim Diederich, Jörg Kindermann, Edda Leopold, and Gerhard Paass. 2003. Authorship attribution with support vector machines. *Applied Intelligence*, 19(1-2):109–123.

Hugo J Escalante, Thamar Solorio, and M Montes-y Gómez. 2011. Local histograms of character n-grams for authorship attribution. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 288–298.

Rudolf Flesch. 1948. A new readability yardstick. *The Journal of Applied Psychology*, 32(3):221.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. 2000. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The Annals of Statistics*, 28(2):337–407.

Michael Gamon. 2004. Linguistic correlates of style: Authorship classification with deep linguistic analysis features. In *Proceedings of the 20th International Conference on Computational Linguistics*, page 611. Association for Computational Linguistics.

Manoj Harpalani, Michael Hart, Sandesh Singh, Rob Johnson, and Yejin Choi. 2011. Language of vandalism: Improving Wikipedia vandalism detection via stylometric analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers*, volume 2, pages 83–88.

Daniel Hasan Dalip, Marcos André Gonçalves, Marco Cristo, and Pável Calado. 2009. Automatic quality assessment of content created collaboratively by web communities: a case study of Wikipedia. In *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital libraries*, JCDL '09, pages 295–304, New York, NY, USA. ACM.

Michael Heilman, Kevyn Collins-Thompson, and Maxine Eskenazi. 2008. An analysis of statistical models and features for reading difficulty prediction. In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, pages 71–79. Association for Computational Linguistics.

Vlado Kešelj, Fuchun Peng, Nick Cercone, and Calvin Thomas. 2003. N-gram-based author profiles for authorship attribution. In *Proceedings of the Conference Pacific Association for Computational Linguistics, PACLING*, volume 3, pages 255–264.

Dan Klein and Christopher D Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 423–430. Association for Computational Linguistics.

Emily Pitler, Annie Louis, and Ani Nenkova. 2010. Automatic evaluation of linguistic quality in multi-document summarization. In *Proceedings of the 48th annual meeting of the Association for Computational Linguistics*, pages 544–554. Association for Computational Linguistics.

Sindhu Raghavan, Adriana Kovashka, and Raymond Mooney. 2010. Authorship attribution using probabilistic context-free grammars. In *Proceedings of the ACL 2010 Conference Short Papers*, ACLShort '10, pages 38–42, Stroudsburg, PA, USA. Association for Computational Linguistics.

Congzhou He Ramyaa and Khaled Rasheed. 2004. Using machine learning techniques for stylometry. In *Proceedings of International Conference on Machine Learning*.

Paul Rayson, Andrew Wilson, and Geoffrey Leech. 2001. Grammatical word class variation within the british national corpus sampler. *Language and Computers*, 36(1):295–306.

Klaus Stein and Claudia Hess. 2007. Does it matter who contributes: a study on featured articles in the German Wikipedia. In *Proceedings of the Eighteenth Conference on Hypertext and Hypermedia*, pages 171–174. ACM.

Kristina Toutanova and Christopher D Manning. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 13*, pages 63–70. Association for Computational Linguistics.

Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 173–180. Association for Computational Linguistics.

William Yang Wang and Kathleen R. McKeown. 2010. "Got you!": Automatic vandalism detection in Wikipedia with web-based shallow syntactic-semantic modeling. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 1146–1154, Stroudsburg, PA, USA. Association for Computational Linguistics.

Dennis M Wilkinson and Bernardo A Huberman. 2007. Assessing the value of coooperation in Wikipedia. *arXiv preprint cs/0702140*.