Predicting the Volume of Comments on Online News Stories (Abstract)*

Manos Tsagkias e.tsagkias@uva.nl
Wouter Weerkamp w.weerkamp@uva.nl
Maarten de Rijke mdr@science.uva.nl

ISLA, University of Amsterdam
Science Park 107, 1098 XG Amsterdam

ABSTRACT

On-line news agents provide commenting facilities for readers to express their views with regard to news stories. The number of user supplied comments on a news article may be indicative of its importance or impact. We report on exploratory work that predicts the comment volume of news articles prior to publication using five feature sets. We address the prediction task as a two stage classification task: a binary classification identifies articles with the potential to receive comments, and a second binary classification receives the output from the first step to label articles "low" or "high" comment volume. The results show solid performance for the former task, while performance degrades for the latter.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics

General Terms

Algorithms, Theory, Experimentation, Measurement

Keywords

Comment volume, prediction, feature engineering

1. INTRODUCTION

As we increasingly live our life online, in the form of blogs, discussion forums, comment facilities, etc., new types of data become available that can be mined for valuable knowledge. E.g., online chatter can be used to predict sales ranks of books [4]. Online news is an especially interesting data type for mining and analysis purposes. Much of what goes on in social media is a response to news events, as is evidenced by the large amount of news-related queries users submit to blog search engines [9]. Tracking news events and their impact as reflected in social media has become an important activity of media analysts [1]. We focus on online news articles plus the comments they generate, and attempt to predict news article comment volume prior to publication time.

*The full version of this paper appeared in *CIKM 2009*.

DIR'10, January 25, 2010, Nijmegen, the Netherlands.

One might raise the question why one should be interested in commenting behavior and the factors contributing to it. We envisage three types of application for predicting the volume of comments generated by news articles. First, *media and reputation analysis* is dependent on what users think of topics covered in the media. Predicting the comment volume might help in determining the desirability of an article (e.g., regarding the influence on one's reputation) or the timing of its publication (e.g., generate publicity and discussion during election time). Second, *pricing of news articles* by news agencies and *ad placement strategies* by news publishers could be made dependent on the expected comment volume; articles that are more likely to generate comments could be priced differently. Finally, news consumers could be served only news articles that are most likely to generate many comments; news sources can thus provide new services to their customers and can *save consumers' time* in identifying "important" articles.

Our aim in this paper is to predict comment volume of news articles prior to publication. To this end, we seek to answer the following two questions: (i) What are the dynamics of user generated comments on news articles? We look at article and comment statistics per source. (ii) Can we predict, prior to publication, whether a news story will receive any comments at all, and if so, whether it will receive few or many comments?

This work makes several contributions. First, it explores the dynamics of user generated comments in on-line Dutch media. Second, it introduces the problem of predicting the comment volume of a news article. Third, it provides a set of surface, cumulative, textual, semantic, and real-world features that can be used to predict the number of comments of a news story prior to publication. Fourth, it provides an evaluation of the introduced features. Fifth, an error analysis identifies possible causes for classification failure.

Section 2 contains related work; we explore news comments in Section 3; our feature sets are introduced in Section 4; predicting comment volume is done in Section 5; Section 6 contains discussion, error analyses, conclusions, and future work.

2. RELATED WORK

Different aspects of the comment space dynamics have been explored in the past. Schuth et al. [11] explore the news comments space of four on-line Dutch media, while Mishne and Glance [10] explored the weblog comment space. Kaltenbrunner et al. [6] measured community response time in terms of comment activity on Slashdot stories, and discovered regular temporal patterns on people's commenting behaviour. Recently, various prediction tasks and correlation studies have been considered in social media. Mishne and de Rijke [8] use textual features as well as temporal metadata of blog posts to predict the mood of the blogosphere. De Choudhury et al. [3] correlate blog dynamics with stock market activity,

and Gruhl et al. [4] perform a similar task with blogs/reviews and book sales. Szabó and Huberman [12] predict the popularity of a story or a video on Digg or YouTube, given an item's statistics over a certain time period after publication. Lerman et al. [7] forecast the public opinion of political candidates from objective news articles. Finally, Tsagkias et al. [13] predict podcast preference using surface features extracted from podcast RSS feeds.

To our knowledge, no prediction tasks have been published that concern the volume of comments generated by online news articles.

3. EXPLORING NEWS COMMENTS

Our data consists of the aggregated content from seven on-line news agents: *Algemeen Dagblad* (*AD*), *De Pers*, *Financieel Dagblad* (*FD*), *Spits*, *Telegraaf*, *Trouw*, *WaarMaarRaar* (*WMR*), and one collaborative news platform, *NUjij*. We have chosen to include sources that provide commenting facilities for news stories, but differ in coverage (regional/national), in political views, in subject (general/politics/arts/entertainment), and in type.

We turn to our first research question: What are the dynamics of user generated comments on news articles? Our data exploration reveals "big" and "small" news agents based on their respective number of published articles and received comments. User commenting behaviour is measured twofold: a) as the reaction time required for an article to receive a comment, and b) as the discussion timespan, for how long an article keeps receiving comments. Although we recorded variations of commenting behaviour between sources yet our findings are comparable to commenting behaviour in blogs [10]. These aspects of online news seem to be inherent characteristics of each source, possibly reflecting the credibility of the news organization, the interactive features they provide on their web sites, and their readers' demographics [2]. Our features attempt to capture the differences between the sources into account.

4. FEATURE ENGINEERING

We consider five groups of features: a) *surface*: captures feed metadata quality, b) *cumulative*: identifies impact of a news article by monitoring how many times an article's near-duplicate appears in our dataset, c) *textual*: captures which terms are correlated with most and least commented articles, d) *semantic*: similar to *textual* captures discriminative entities and locality, and e) *real-world*: outside temperature at time of publication.

5. PREDICTING COMMENT VOLUME

We now turn to the second research question: Can we predict, prior to publication, whether a news story will receive any comments at all? And if it receives comments, can we predict whether it receives few or many comments?

We address the prediction task as two consecutive classification tasks to compensate for the highly skewed datasets. First, we segregate articles with regard to their potential of receiving comments. A binary classification is performed with two classes: *with comments* vs. *without comments*. Second, we predict the comment volume level for the articles predicted to receive comments in the first step (positive class). This second classification is performed with two classes: *low volume* and *high volume*.

For the first classification step our results show high F1-scores for most sources but with low Kappa-statistic. Textual and semantic features perform the best among all feature sets. The run with all features combined did not lead to substantial improvements over the individual features.

For the second classification step F1-scores drop compared to previously. Textual and semantic features are again strong performers, although they exhibit high variance over the board. All features combined lead to better performance over the baseline. The lower Kappa-values of this run indicate more robust classification.

6. DISCUSSION AND OUTLOOK

We presented exploratory work on predicting the comment volume of news articles prior to publication. We have developed a set of surface, cumulative, textual, semantic, and real-world features and report on their individual and combined performance on two classification tasks: Classify articles according to whether they will (i) generate comments, and (ii) receive few or many comments. Our experiments show that predicting the volume of comments is more difficult than predicting whether an article will receive any comments at all. Textual and semantic features prove to be strong performers, and the combination of all features leads renders classification more robust. Our failure analysis indicates that the features used in this paper are not the only factors involved in the prediction process. Future work should therefore focus on extracting more feature sets (e.g., context and entity-relations), use different encodings for current features, optimize the number of textual and semantic features per source, and explore optimized feature sets.

Acknowledgments. This research was supported by the DuOMAn project (STE-09-12) carried out within the STEVIN programme and by the Netherlands Organisation for Scientific Research (NWO) under project numbers 017.001.190, 640.001.501, 640.002.501, 612.066.512, 612.061.814, 612.061.815, 640.004.802.

7. REFERENCES

[1] D. L. Altheide. *Qualitative Media Analysis (Qualitative Research Methods)*. Sage Pubn Inc, 1996.

[2] D. S. Chung. Interactive features of online newspapers: Identifying patterns and predicting use of engaged readers. *J. Computer-Mediated Communication*, 13(3):658–679, 2008.

[3] M. De Choudhury, H. Sundaram, A. John, and D. D. Seligmann. Can blog communication dynamics be correlated with stock market activity? In *HT '08*, pages 55–60. 2008.

[4] D. Gruhl, R. Guha, R. Kumar, J. Novak, and A. Tomkins. The predictive power of online chatter. In *KDD '05*. 2005.

[5] A. Kaltenbrunner, V. Gomez, and V. Lopez. Description and prediction of slashdot activity. In *LA-WEB '07*, pages 57–66. IEEE Computer Society, 2007.

[6] A. Kaltenbrunner et al. Homogeneous temporal activity patterns in a large online communication space. *CoRR*, abs/0708.1579, 2007.

[7] K. Lerman, A. Gilder, M. Dredze, and F. Pereira. Reading the markets: Forecasting public opinion of political candidates by news analysis. In *Coling 2008*, pages 473–480, 2008.

[8] G. Mishne and M. de Rijke. Capturing global mood levels using blog posts. In *AAAI-CAAW '06*, pages 145–152, 2006.

[9] G. Mishne and M. de Rijke. A study of blog search. In *ECIR '06*, pages 289–301, 2006.

[10] G. Mishne and N. Glance. Leave a reply: An analysis of weblog comments. In *WWE '06*, 2006.

[11] A. Schuth, M. Marx, and M. de Rijke. Extracting the discussion structure in comments on news-articles. In *WIDM '07*, pages 97–104. ACM, 2007.

[12] G. Szabó and B. A. Huberman. Predicting the popularity of online content. *CoRR*, abs/0811.0405, 2008.

[13] E. Tsagkias, M. Larson, and M. de Rijke. Exploiting surface features for the prediction of podcast preference. In *ECIR '09*, 2009.