The Appeal of Politics on Online Readers

Elena Hensinger, Ilias Flaounas, Nello Cristianini
Intelligent Systems Laboratory, University of Bristol, UK

Abstract

We analysed the choices of online readers of newspapers in order to model their preferences, by using automated methods operating on a very large scale. We were able to obtain models which are predictive of users' choices, and which we applied to explore the relationships between audience preferences and topics of news articles. We found that for 12 of 14 modelled audiences, the presence of "Public Affairs" content, such as "Politics", reduced the appeal of an article.

The models, describing the appeal of a given article to each audience, are formed by linear functions of word frequencies, and are obtained by comparing articles that became "Most Popular" on a given day in a given outlet with articles that did not. We make use of 2,432,148 such article pairs, collected over a period of over 1.5 years.

Those models are shown to be predictive of user choices, and in the next step, they are used to compare both the audiences and the contents of various news outlets. First, we visualise the information contained in the models themselves – via word clouds. Next, we use a dataset of half a million articles from one year of time, and we compute for each article its appeal score for each modelled audience. Next, we determine an article's topic affiliation and compare it to its appeals. For an average audience, we find significantly less interest in "Public Affairs" topics, such as "Politics" and "Business", than in "Non-Public Affairs" topics such as "Sport" or "Crime".

Introduction

The availability of news in digital format allows large-scale automated analysis of reading preferences of internet users without using questionnaires, polls or log file analysis of clicks by individual users. Instead, we acquired data about user preferences by collecting specific sets of articles advertised by online news outlets through news feeds: the "Most Read", "Most Clicked" and "Most Viewed". This was compared to another feed, the "Top Stories", which corresponds roughly to the main webpage of an outlet. In other words, we relied on the only available

information about click-through rates of news articles that is released by news outlets, from which we extracted preference data. The drawback of this approach is that this information is not available for all outlets and there is not a fine-grained user segmentation. The advantage is that this data is naturally present "in the wild" and therefore is very plentiful. In our previous work we explored such datasets with different techniques to model user preferences in terms of prediction performance and applications (Hensinger, Flaounas and Cristianini 2010; 2011; 2012).

The key idea of the modelling was to derive some simple functions of the word content of articles that could capture the appeal of any given article for a specific audience. The underlying assumption is that an article will become more popular if it is more appealing. The modelling functions would be – not unusual in text mining –linear functions of word frequencies. In other words, we asked the following question: "Can we define a linear function of word frequencies in a given article that correlates well with the appeal that this article exercises on a user?"

We know well that there are other factors apart from textual content that could affect readers' attention, such as presence of images, size of font, or position of an article in a webpage. For this study, we ignored all this additional information, focusing on the textual content of articles only. Furthermore, we only made use of the text contained in the titles and short descriptions provided to readers, as this is typically the same limited data that readers assess when making their decision on a news webpage. Could it be that there is still some information contained in such a simplified setting? It turns out that the answer is "yes", if we use sufficiently large datasets and an appropriate experimental design.

The problem of inferring the parameters of the required linear function is solved by comparing pairs of articles, which had appeared on the same webpage in the same day, only one of which had become popular. We used a Machine Learning algorithm to identify the parameter settings that conferred a higher appeal to the more popular article in as many document pairs as possible. The resulting linear models – one for each outlet we analysed – were then proven to have a significant predictive power on the choices of readers on a separate dataset (not used for the parameter selection), as well as having other interesting properties.

Once we had built working models for "article appeals" we could run a series of comparisons between audiences and topics. Among other findings, we observed that – perhaps unsurprisingly – a general reader prefers "Non-Public Affairs" over "Public Affairs". Based on these findings, various questions can be followed-up, including that of explaining why news editors give such emphasis to "Public Affairs" articles, when audiences seem to click away from them.

Data Collection and Analysis

We use two datasets, one for modelling user preferences, and the second for understanding the relationship between those preferences and news topics. The first dataset is comprised of articles from 14 oneline news outlets, for 20 months between the 1st December 2009 and 31st July 2011, gathered from news feeds (Flaounas et al. 2011). The outlets were the newspapers “Los Angeles Times”, “The New York Times”, “The Seattle Times” and “The Wall Street Journal”, the broadcasters “BBC News”, “CBS News”, “CNN.com” and “KSBW”, the magazines “Forbes” and “TIME”, the news website “News.com.au”, the media organisation “NPR”, the newswire “Reuters” and the news aggregator “Yahoo! News”. For each of those outlets, we use two sets of articles: the first one corresponding to articles published on the main webpage of the outlet, which we call “Top Stories”; the second set consisting of the most read articles for every single day, in the following called “Most Popular”.

To create preference data pairs, we use the “Most Popular” articles to separate the “Top Stories” ones into two groups: those articles that audiences preferred to read, i.e. which were in “Top Stories” and also in “Most Popular”, and those, which had the same initial condition, but were not preferred by the audiences, i.e. they were in “Top Stories” but not in “Most Popular”. Based on these sets, we acquire 2,432,148 preference pairs.

Each article was represented by its title and description, as provided via the news feeds, reflecting the short textual item overviews on a typical news webpage. Typical text pre-processing techniques of stop word removal and stemming (Porter 1980) were carried out, after which articles were stored in a standard format for automatic processing (Salton, Wong and Yang 1975).

The preference data pairs from a given outlet contain the reading preferences of this outlet's entire audience group. In the following Section, we will show that it is possible to predict the choices of readers better than random, by automatically fitting linear models to such preference data, even when using just the words in title and description as data features.

Appeal Models

For each outlet, we used the preference pairs to train the Machine Learning algorithm Ranking SVM (Joachims 2002) (10-fold cross validation, training on 18 and testing on 2 months), obtaining models that predict user choices on news significantly better than random. Their pairwise prediction performances are presented in Figure 1. Each model has the functionality to decide on a preferred orientation for a pair of input articles, based on which article has a higher appeal

score. Due to the modelling approach, models have the additional capability to receive as input individual articles and compute their individual appeal scores, which we exploit in further work.

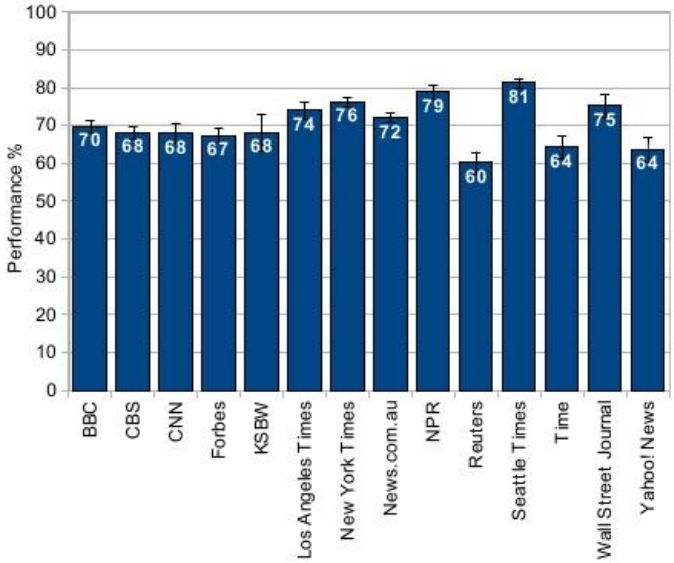

Figure 1: Mean performances and standard errors for 14 models of audience group preferences, with confidence intervals set to 95% on estimation of the mean.

As a first investigation, we "look inside" two example models, by presenting the highest and lowest weighted terms, i.e. those terms that contribute most to increase or decrease an article's appeal, if the respective words occur in the article. In Figure 2, we show those terms for the audience of the outlet "News.com.au", an Australian newspaper belonging to the "News Limited" corporation. Notice the many references to people, events and location of this outlet. In contrast, Figure 3 displays the same information, but for the outlet "Forbes". This outlet focuses on business topics, and it is known for its lists of billionaires and their possessions[1].

1 http://en.wikipedia.org/wiki/Forbes (July 2012)

Figure 2: The word cloud shows in pink (black) the words that increase (decrease) the appeal of a text for readers of the website "News.com.au". The size of a word in the image reflects the magnitude of this effect. Notice the references to people and to the location of the outlet in Australia.

Figure 3: The word cloud shows in pink (black) the words that increase (decrease) the appeal of a text to the readers of "Forbes.com". The size of a word in the image reflects the magnitude of this effect. Notice the many superlatives, referring to wealth rankings published by that outlet.

Topic classification

We make use of various topic categorisation tools to detect the strength of an article's affiliation with a topic (Flaounas 2011; Hensinger, Flaounas and Cristianini 2011). These tools compute a topic score for an input article which is compared to a topic-individual threshold in order to decide whether an article is of a topic or not. In the presented case study, we utilise the topic scores before thresholding.

We divide the topics into two macro categories, which are comprised of the average scores for their contained topics, namely *Elections*, *Inflation and Prices*, *Markets*, *Business*, *Politics*, and *Petroleum* for "Public Affairs", and *Crime*, *Disasters*, *Fashion*, *Art*, *Environmental issues*, *Religion*, *Science*, *Sports*, *Travel*, and *Weather* for "Non-Public Affairs".

At this stage, we use a second large dataset: 579,805 articles published in the "Top Stories" feeds of 37 different outlets, for one year of time between the 1st June 2010 and the 31st May 2011. This time, we use an article's title, description and full article text in order to determine more precisely its topic affiliation. We also compute for each article and model the appeal score, and we average the resulting 14 appeal scores to approximate a general audience and the general appeal, i.e. how a universal audience perceives this article.

Overall, we end up with three values per article: the average appeal score over all 14 models, and the two topic category scores for "Public Affairs" and "Non-Public Affairs", respectively. Next, we compute correlation coefficients over all three scores for all 579,805 articles, grouped by the 37 outlets the articles were published in. Overall, we could assign 63% of articles to be of any category. This allows to compare outlets in terms of their published articles' themes, as presented in Figure 4: the appeal for a general audience is depicted along the x-axis and ranks the outlets accordingly. The y-axis shows the results of the ratio of outlets' "Non-Public" to "Public" Affairs news. It is only for two outlets, "The Wall Street Journal" and the "Reuters", for which more of their published articles are about "Public" than about "Non-Public" affairs.

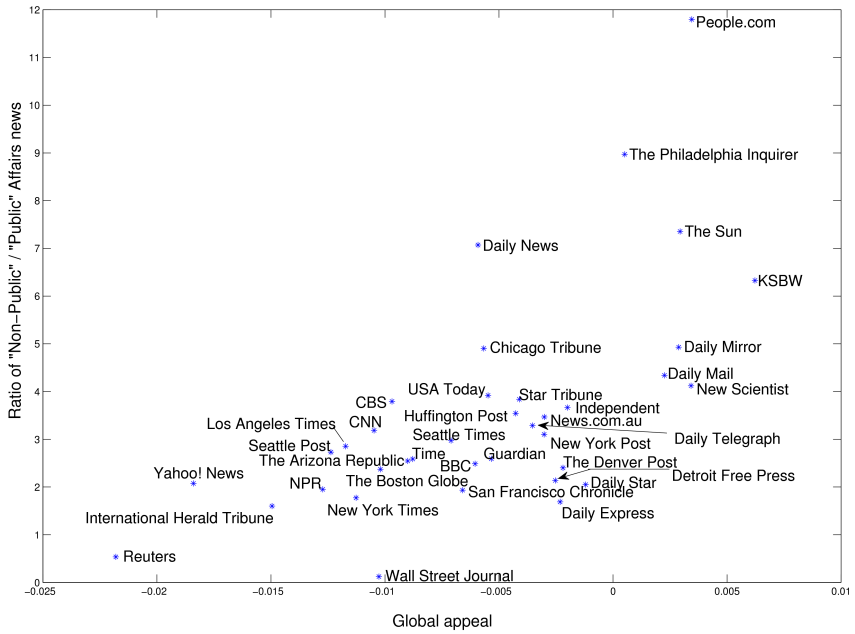

Figure 4: News outlets from around the globe are embedded in the two-dimensional plane, spanned by their appeal to a general audience (x-axis), and by their ratio of published "Non-Public" to "Public" affairs news (y-axis). Outlets in the lower part of the image tend to cover more "Public Affairs" news.

Results

In the first experiment, we analyse how topic correlates with appeal to readers, and we do this for the different 37 outlets separately. For each outlet's articles for one year of time, we compute the correlation coefficients between the general audience appeals and the articles' "Public" and "Non-Public" Affairs scores.

For each of the investigated outlets, we obtain the result that the general appeal of its published articles is significantly anti-correlated with them being about "Public Affairs", and positively correlated with them being about "Non-Public Affairs". All measured correlation coefficients are presented in Figure 5. The correlation of general appeal and "Non-Public Affairs" scores is on average 0.28, and can reach up to 0.43 for articles from "Denver Post" (with p-values < 0.001). On the contrary, general appeal and "Public Affairs" scores are anti-correlated with an average of -0.31, with minimal value -0.43 for articles published in "Daily News" (with p-values < 0.001). This presents a universal pattern of what general audiences find interesting to read about.

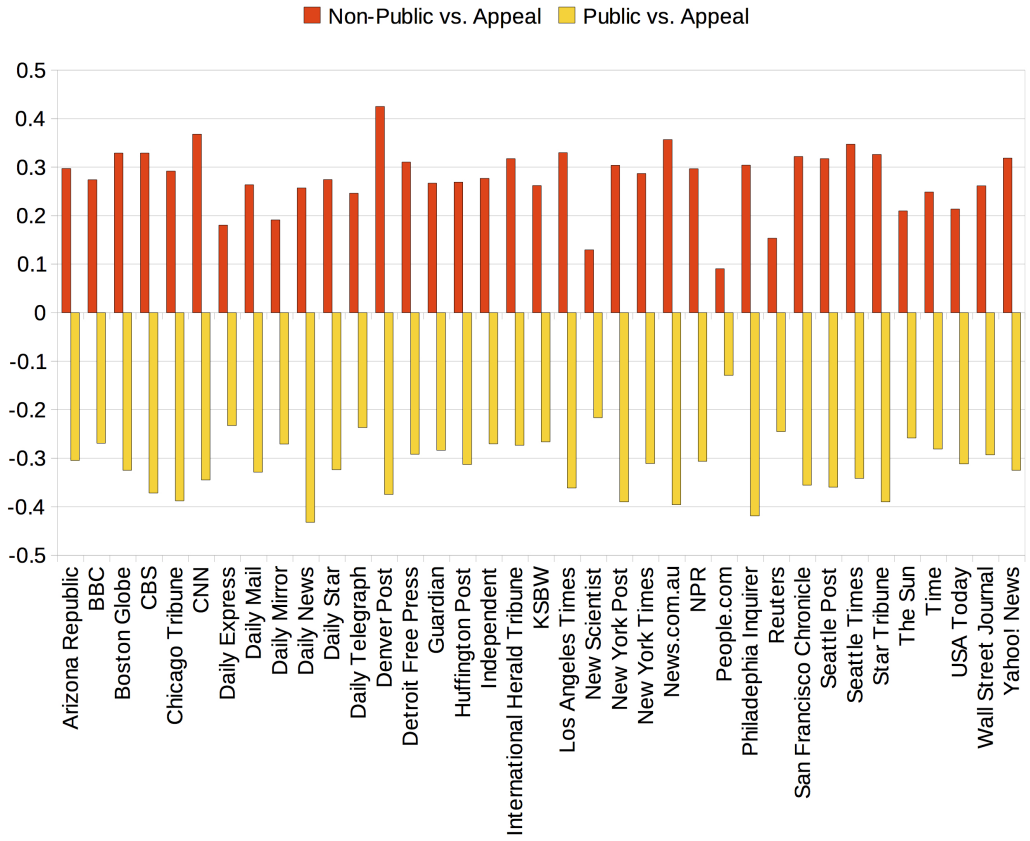

Figure 5: Correlation coefficients between general appeal of ``Top Stories'' articles and their scores for ``Public'' and ``Non-Public'' Affairs topics, grouped by the 37 outlets of articles' publication.

The second experiment focuses on the different models and what topics their modelled audiences perceive as appealing. We therefore used all 579,805 articles, independently of their publishing outlet, and we computed their appeal scores with each of the 14 preference models individually. As before, we also utilised the “Public Affairs” and the “Non-Public Affairs” score for each of those articles. Next we computed, model by model, the correlation coefficients between the appeal scores this audience model assigns to a vast amount of data from various sources, and the topic category of those articles. We find a significant correlation (p-value<0.001) for all but one model of article appeal and its topic being about “Non-Public Affairs”. The correlations range from 0.04 for “Los Angeles Times” to 0.32 for “Yahoo! News”. The only model with non-significant results was “Reuters”.

As for appeal and “Non-Public Affairs”, we get significant results for all 14 models: for 12 of those, there exists an anti-correlation, i.e. the more the article

is about topics such as “Politics” and “Business”, the less appealing it is perceived for the audience. It is only for two outlets that the opposite can be said: for “The Wall Street Journal” with a correlation coefficient of 0.03 and the “Reuters” with 0.09. These results are displayed in Figure 6, plotting the models in the space of significant correlation coefficients between appeal and “Non-Public Affairs” along the x-axis, and appeal and “Public Affairs” along the y-axis. The model of “Reuters” audience is omitted due to non-significant results along the x-dimension.

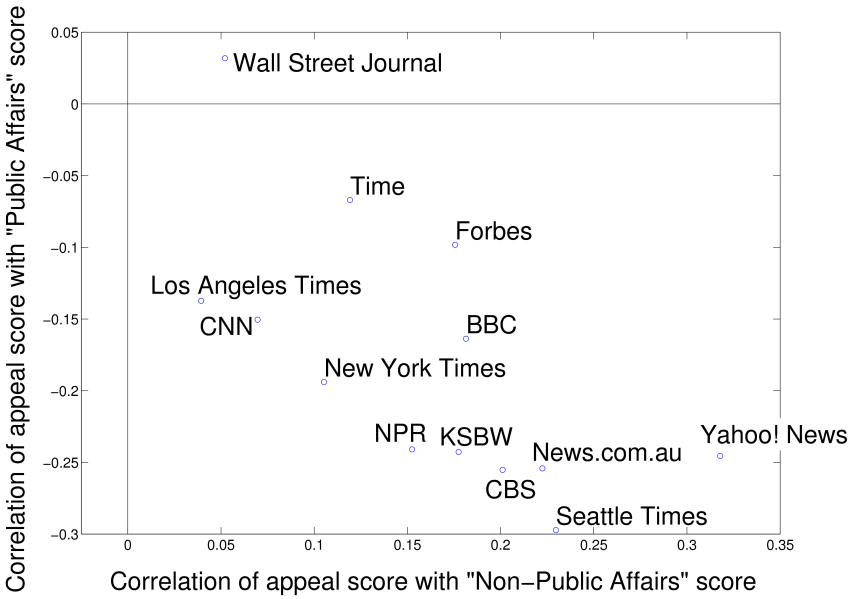

Figure 6: Models in space of significant correlation coefficients of the appeal scores they assign to 579,805 articles, and those articles' ``Public'' and ``Non-Public'' themes. Only for two outlets, the ``Wall Street Journal'' and the ``Reuters'', a significant correlation between articles' appeal scores and their ``Public Affairs'' scores can be observed. For all other models, this relationship is an anti-correlation, i.e. these topics reduce the perceived appeal. The ``Reuters'' model is excluded from visualisation due to non-significant results along the x-axis.

The presented results were achieved with Machine Learning techniques both for learning user preferences and for topic assignments of news articles. The presented studies demonstrate how computational approaches can be employed in the setting of big data and media analysis.

Conclusions

The comparison of a “popular” article with an “un-popular” one reveals significant differences, when performed on 2.5 million such pairs. Even by just using word frequencies as data features, it is possible to predict which of two articles is the popular one with an average probability of more than 70% percent. While this could be improved, it is sufficient to enable an analysis of user preferences towards articles of different topic groups.

This is part of a more general trend which demonstrates how the use of vast amounts of data, naturally available “in the wild” and therefore inexpensive, can lead to insights about the preferences of news readers, if combined with appropriate Machine Learning or data mining algorithms. By comparing millions of news items in the right way, we were able to model “What people prefer to read", and to present some explanations to the question of “What influences those preferences". Such analyses can be helpful for journalists and editors on the one hand, and for political and media scientists on the other hand, who have been exploring questions of what becomes news (Harcup and O'Neill 2001) or how media bias, in terms of liberal ideological views, can be measured (Groseclose and Milyo 2005).

The data we used for modelling has its limitations: it contains only text, and misses other factors that might affect readers' interests, such as accompanying pictures or videos. Furthermore, the data conveys preference information for entire audience groups, leading to a rather coarse-grained segmentation of users by their choice of outlet. While such an approach is not uncommon and is known in marketing as “behavioural segmentation" (Assael and Roscoe 1976), it is likely that modelling could be further improved if finer-grained user data could be acquired.

As data features for modelling, we use very limited information: articles' titles and descriptions only. While this mimics the real-life situation users experience at news webpages, it results in an average number of features of less than 30 per article. Given all these characteristics and challenges in our data, it is remarkable that it is still possible to reliably predict news preferences of audiences with an average performance of 70.6%.

Finer-grained user models could allow for further investigations and insights into why audiences like some articles more than others. Avenues for further work include incorporation of more information on users and news, for instance demographic data, geographic proximity of users to news, the presence of celebrities in the text, or the reporting of scandals.

Acknowledgements

I. Flaounas and N. Cristianini are supported by the CompLACS project (European Community's Seventh Framework Programme - grant agreement No. 270327); All authors are supported by Pascal2 Network of Excellence.

References

Assael, H., and A.M Roscoe Jr.. 1976. “Approaches to Market Segmentation Analysis.” *The Journal of Marketing* 40, 4: 67-76

Flaounas, I., O. Ali, M. Turchi, T. Snowsill, F. Nicart, T. De Bie and N. Cristianini. 2011. “NOAM: News Outlets Analysis and Monitoring System.” *Proceedings of the 2011 International Conference on Management of Data (SIGMOD '11)*:1275-1278

Flaounas, I. 2011. “Pattern Analysis of News Media Content.” *PhD Thesis* University of Bristol

Groseclose, T. and J. Milyo. 2005. “A Measure of Media Bias.” *The Quarterly Journal of Economics.* 120(4): 1191-1237

Harcup, T. and D. O'Neill. 2001. “What is News? Galtung and Ruge revisited.” *Journalism Studies*: 2(2): 261-280

Hensinger, E., I. Flaounas and N. Cristianini. 2010. “Learning the Preferences of News Readers with SVM and Lasso Ranking.” *Proceedings of Artificial Intelligence Applications and Innovations - 6th IFIP WG 12.5 International Conference (AIAI)*: 179-186

Hensinger, E., I. Flaounas and N. Cristianini. 2011. “Learning Readers' News Preferences with Support Vector Machines.” *Proceedings of Adaptive and Natural Computing Algorithms - 10th International Conference (ICANNGA)*: 322-331

Hensinger, E., I. Flaounas and N. Cristianini. 2012. “What Makes Us Click? - Modelling and Predicting the Appeal of News Articles .” *Proceedings of the 1st International Conference on Pattern Recognition Applications and Methods, Volume 2 (ICPRAM)*: 41-50

Joachims, T. 2002. “Optimizing search engines using clickthrough data.” *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*: 133-142

Porter, M.F. 1980. “An Algorithm for Suffix Stripping.” *Program* 14: 130-137

Salton, G., A. Wong and C.S. Yang. 1975. “A vector space model for automatic indexing.” *Communications of the ACM*. 18(11), November: 613-620