Detection of News Feeds Items Appropriate for Children

Tamara Polajnar[1], Richard Glassey, and Leif Azzopardi

School of Computing Science, University of Glasgow, Glasgow, UK
tamara.glasgow@gmail.com[1]

Abstract

Identifying child-appropriate web content is an important yet difficult classification task. This novel task is characterised by attempting to determine age/child appropriateness (which is not necessarily topic-based), despite the presence of unbalanced class sizes and the lack of quality training data with human judgements of appropriateness. Classification of feeds, a subset of web content, presents further challenges due to their temporal nature and short document format. In this paper, we discuss these challenges and present baseline results for this task through an empirical study that classifies incoming news stories as appropriate (or not) for children. We show that while the naïve Bayes approach produces a higher AUC it is vulnerable to the imbalanced data problem, and that support vector machine provides a more robust overall solution. Our research shows that classifying children's content is a non-trivial task that has greater complexities than standard text based classification. While the F-score values are consistent with other research examining age-appropriate text classification, we introduce a new problem with a new dataset.

Introduction

Children are embracing new technologies and increasingly using the web to find and engage with content for both educational and entertainment purposes [12, 13]. Whilst children should have equal rights to the content available online, and the barriers to access should be eliminated, the content they access must be appropriate, accessible and sensitive to their age, developmental stage, and level of understanding.

Much of the recent research into children's computer interaction has been concerned with the information seeking behaviour of children, and how website accessibility and design supports or hinders children [1, 3, 10]. This work has helped to produce design guidelines and has inspired the development of information access tools designed specifically for children [5, 6, 9]. Together, these efforts seek to improve access to the content by understanding the user needs of children and removing the barriers that adult user interfaces create for children; however, the issues of safety and content appropriateness have not been adequately addressed within these advances.

At present, the amount of content specifically designed and tailored for children is only a small fraction of all the content available online. For example, manual inspection shows that less than 5% of websites listed in DMOZ[1] are in the *kids and teens* category. Restricting children to this fraction of websites ensures that they do not encounter inappropriate content, but it also excludes them from accessing much more information online. Rather than manually creating safe but limited repositories that do not scale nor keep pace with the explosive growth of content, research efforts are focusing on developing methods that ensure unseen content is safe, appropriate and accessible.

Since it is not possible to manually check all available content, attempts to automatically classify appropriateness have been investigated. In [4], a webpage classifier was developed that tries to identify the pages in the correct categories of DMOZ. It was trained on features ranging from *topographic*, such as the amount of advertising on the page and the linking neighbourhood of the page, through to the *content-based*, such as the indicative words and the readability indices of the text in different parts of the page. On the other hand, related work has also been done on classifying text-only sources by reading age, as by [15].

As web pages are just one kind of online content, this paper explores the classification problem for news feeds: determining whether short snippets for individual feed entries are appropriate for children. The next section describes our motivation and the properties of the data. Then we describe the experiments conducted before concluding with the difficulties and challenges ahead for developing robust and reliable classifiers for children's or general feeds.

Problem Description

Feeds are an efficient means of receiving the latest information on news and current events. Most producers of content publish RSS/ATOM feeds in order to announce new content to users. The user-chosen feeds are processed by tools such as aggregators (e.g. Google Reader), which collect their content and present it to the users, without them having to visit websites. This is particularly useful in combination with mobile devices (which are also being used more by children [14]), especially since input options and limited screen size of mobile devices make traditional web browsing difficult.

While feeds themselves are omnipresent, child-specific feeds are not yet prevalent. Feed aggregators specifically designed for children [5] are well placed to improve safe access to new content. Because feeds are also published by social media and blogging sites, which actively interest young users, there is a real danger of exposure to inappropriate content. Feed aggregator tools also need to ensure that the content retrieved is safe for children. This is an important challenge, because whilst not all content is designed for children, children should have the right to to access all content that is appropriate for their them.

This problem is further compounded due to the nature of the data underlying the content. Feeds can have embedded metadata, but in general they are

[1] Open Directory Project – `http://www.dmoz.org/`

composed of a collection of entries, each of which has a title, link and a snippet (or a full article in some rare cases), which describes the content. Thus, a key technical challenge for such tools is to be able to classify feed entries for appropriateness, using only the limited information given by the title, link and snippet.

Whilst feeds are useful for any type of content, from any source possible, we chose here to focus upon news feed as a starting point to test the applicability of feed classification for appropriateness. This decision is supported by the availability of high-quality and accurate judgements of appropriateness for a news feed collection. However, this still presents significant challenges to overcome (i.e. that the same news stories written for children and for adults are generated from same events and that there is a large imbalance in the amount of content for children vs. adults.). Thus, unlike other feed classification problems, where topicality is the central problem that needs to be addressed, our problem is quite different and requires features beyond the term space, like the text readability features employed in [4, 15].

News Feed Scenario and Data

For this preliminary study, we selected two feeds from the British Broadcasting Company: (i) standard news feed (referred to as BBC), and (ii) children's news feed (CBBC). The feed data from these sources was collected from early March 2010 until March 2011, and consists of 51,833 entries from the BBC feed, and only 1,832 entries from the CBBC feed (i.e. 3.5%), making the dataset highly skewed. It should be noted that many of the entries written for children are derived from the same news events which generate the entries for the adult feed. An analysis of the incoming feeds revealed that the vocabulary used for children expanded at a much slower rate than for adults. Figure 1 shows that after 27,000 non stop-words there were approximately 2,000 more unique terms in the adult vocabulary than in the children's vocabulary, or approximately 40% more terms. Within the BBC feeds there were over 50,000 unique terms, while for the CBBC feeds there were around 6,000 unique terms, i.e. these terms only occurred in the BBC or CBBC feeds respectively.

Given this data, we would like to classify the entries from these feeds into stories appropriate for children and stories for adults. Before doing so, we performed a preliminary analysis of additional features that might be helpful in the content classification. We decided to examine the standard readability measures, such as the automated readability index (ARI)[2] [16], which are listed in Table 1.

We trained a naïve Bayes classifier to determine whether such features were discriminative of content for children. Using only these features resulted in an F-score of 0.15, indicating that they have some utility for classification. To determine which features were the best at distinguishing between classes, we used information gain (IG) to assess the measures. We found that ARI (IG = 0.03) was

[2] We calculated the measures using open source software `http://www.addedbytes.com/lab/readability-score/`.

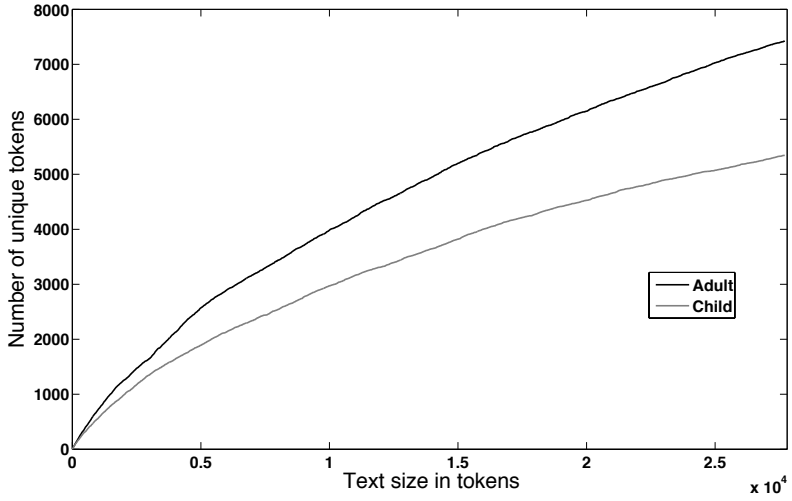

Fig. 1. Vocabulary Growth on children's and adult's news feeds.

Readability Measure	**IG**
Automated Readability Index	0.03
Flesch-Kincaid Reading Grade Level	0.014
Coleman-Liau Index	0.01
Flesch-Kincaid Reading Ease	0.003
Gunning-Fog Score	0
Average words per sentence	0
Average syllables per word	0

Table 1. Summary of the readability measures used and their associated information gain for the training data.

the most predictive measure, followed by the *Flesch-Kincaid reading grade level* (0.014), *Coleman-Liau Index* (0.01), and *Flesch-Kincaid Reading Ease* (0.003), while the *Gunning-Fog Score*, the *average words per sentence*, and the *average syllables per word*, were indistinguishable between the classes. The ARI measure differs from the other types of readability scoring because it looks at word length in characters instead of the number of syllables, which may be an indicating factor for this dataset. While these measures would provide more information over longer texts, they are still predictive for the 2 or 3 sentence snippets.

When we examined the unique words (tokens) that occurred in the feed entries, many of the words in the adult data occurred infrequently. By pruning away all words that occur in data less than 3 times, while keeping all the words in children's data, the feature space could be reduced to around 17,000 terms. Performing IG on these features, we found that short, informal words such as

pics and *footie* featured prominently. The word with highest information gain was *kids*, which is consistent with the findings in [4], and was followed closely by other words that indicate children's content, such as *newsround* (a CBBC-specific token), *cute*, *zoo*, or *animals*. However, the largest presence was indicated by the popular culture tokens, e.g., *justin* (for Justin Bieber) or *x* and *factor*. Adult-specific tokens that refer to areas of political unrest or military conflict like *iraq* and *government* occur much lower down the list.

From this brief analysis of shallow linguistic features, we identified some properties that are beneficial and others that may pose difficulties. The large amount of data and the skewed distribution of classes makes it suitable for the sparse classification offered by the SVM. While the class distribution property, along with the large number of terms may impede the naïve Bayes classifier. This can be addressed with feature pruning. However, this may reduce the generality of the classifiers trained i.e., when classifying new feed entries. Fortunately, the difference in writing style, with the use of short informal words as indicated by our analysis and by the ARI measure, should be helpful to distinguish the classes.

Experimental Setup

To perform the classification of the news feeds, we examined the data in two stages (training and testing). Firstly, we used the training data, which comprised of feeds from 2010 (approximately 9 months of data), to learn as much as possible about the characteristics of the problem and to test out classifiers through cross-validation. We then applied the best-performing strategies to the held out test data, which comprised of the feeds from 2011 (approximately 4 months of data).

Features

We extracted two types of features, one based on bag-of-word tokens from the feed entries, and the other based on the eight readability measures mentioned in the section above. We also used IG to filter these two sets of features. For the tokens, the dataset with 17,000 features will be referred to as, *long text* or LT. From this, we removed all the tokens that had low IG, reducing the feature set to about 3,000 (ST). For the readability measures, in the first instance we used 8 measures (*long readability* or LR), which were then filtered down to 4 (SR), by choosing the ones with IG larger than 0. The readability measures yield numbers from a continuous distribution, while the word-token features come from a discrete distribution. In order to make the readability features compatible with the naïve Bayes classifier, they are discretised by applying the floor function, and by zeroing negative values. We also examined feature weighting combinations using BM25 and TF-IDF, but this lead to a reduction in F-score by 0.02, leading us to employ the raw feature frequency counts instead.

Description

As a default baseline classifier, we used the multinomial naïve Bayes (NB) classifier [11]. In addition, we used an SVM [8], which is the state-of-the-art in large-scale text classification. Since we are considering two types of features, each with a different distribution, the non-parametric kernel-based method (i.e. SVM) may be more suitable.

To calibrate and validate our choice of parameters and features, the first set of experiments conducted on the training data, consisted of 10 randomised repetitions of 10-fold cross-validation (10x10CV). This provided 100 data points for significance testing, which was done using the t-test designed especially for 10x10CV experiments [2]. In the results section we take the word *significant* to indicate *statistically significant* results with $p < 0.05$. The SVM soft-margin parameter is tuned at each fold, using a separate, 2x2CV experiment conducted on the fold training data. This tested a range of suitable values (1, 2, 5, 10). The linear kernel provided the best results on the training data. The results are described using standard information retrieval measures precision (P), recall (R), F-score, and the area under the receiver operating characteristic curve (AUC). In order to make the results comparable to the related work, we show both the standard F_1 and the precision-skewed $F_{0.5}$ measures.

In the second set of experiments each of the classifiers was combined with the best settings from the first experiment, and the model trained on the full set of training data was tested on the unseen feed entries from 2011. For example, in the majority of the tuning experiment the chosen setting for the SVM soft-margin parameter was 1, so we used this parameter for testing. The test data contains 27,900 BBC feeds and 1,162 CBBC feeds resulting in a slightly different distribution of 4,5% positive samples. These documents are then represented by the tokens that were found in training data (all others are discarded) and the readability test values.

Results

Cross-validation (CV) experiments

Table 2 shows the results from the CV experiments. The baseline is provided by the NB classifier trained on the large feature space, LT. The NB classifier is trained to model both the class and feature distribution. The former distribution is sensitive to the large imbalance in the class sizes. As a result the NB classifier tends to produce results with higher recall and lower precision, and therefore a lower $F_{0.5}$ value. The latter distribution is sensitive to an overly large feature space; consequently, a statistically significant improvement in performance was gained by feature filtering. This improvement, is also significant in the AUC, which measures the quality of classification regardless of the class decision boundary. The AUC indicates that the majority of positive documents have a higher probability than negative documents, and that the balance is skewed

C	Data	F_1	$F_{0.5}$	P	R	AUC
NB	LT	49.53 ± 0.27	41.52 ± 0.27	37.48 ± 0.27	73.22 ± 0.33	94.38 ± 0.08
NB	ST	57.14 ± 0.27	47.82 ± 0.27	43.14 ± 0.27	**84.80 ± 0.32**	**97.11 ± 0.06**
NB	STSR	47.71 ± 0.25	37.85 ± 0.24	33.27 ± 0.23	84.59 ± 0.23	96.41 ± 0.07
SVM	LT	61.21 ± 0.30	61.26 ± 0.35	61.33 ± 0.40	61.27 ± 0.36	94.87 ± 0.08
SVM	LTLR	62.71 ± 0.33	63.52 ± 0.37	64.11 ± 0.43	61.52 ± 0.36	95.29 ± 0.06
SVM	ST	64.02 ± 0.38	65.81 ± 0.55	67.22 ± 0.74	61.73 ± 0.47	95.61 ± 0.10
SVM	STSR	**67.42 ± 0.28**	**72.69 ± 0.30**	**76.72 ± 0.36**	60.24 ± 0.34	96.49 ± 0.06

Table 2. Results of the cross-validation experiment on training data. All results are show with standard error values. Statistical significance of these values is discussed in the results section, although all results improve on the NB LT baseline.

towards recall due to a cutoff probability value that is lower than the natural distribution of positive and negative documents.

We conducted further experiments to chose the probability cutoff value that produces the best performance in the training data. Applying it to test data increased the precision, but reduced recall so that the overall F-score was equal to baseline SVM performance. Choosing the boundary that produces the highest $F_{0.5}$ score at each cross-validation significantly increases NB ST results to F = 63.46 and $F_{0.5}$ = 70.79. We will refer to this setup as NB ST+. The average optimised cut off value over the 10x10cv experiment is 0.9987. We can use this value on the test data to corroborate the improvement in performance.

The SVM LT performs significantly better than both of the NB LT and ST F-score results, but is not statistically significantly better then NB ST AUC score. The reduction in feature space (SVM ST) improves the SVM F-score and AUC results, significantly over NB LT and ST. However, they are not significantly different from the NB ST AUC or the SVM LT. The reason for this is that the SVM is not as sensitive to feature vector length as the NB classifier. The SVM is also better and predicting the boundary between the classes, and in general returns higher precision than recall on this data.

The distinct improvements in SVM performance come from the introduction of the readability measure features (LR and SR). The effect is a small reduction in recall combined with a large increase in precision, improving the overall quality of classification, as further evidenced by the increase in AUC. The STSR feature combination, produces a statistically significantly higher performance than the SVM with the ST features.

In order to use readability features with the NB classifier, we had to discretise values. Still, the underlying distribution is different from the word features and this results in a reduction in performance. By calculating the cutoff for each fold of the CV experiment we can engineer an improvement (NB STSR+ $F_{0.5}$ = 73.41 F = 66.48), and identify the mean value of 0.9999 to try on the test data.

Testing on unseen data

In the cross-validation experiments, we randomised the training data, this lead us to overestimate the performance of the classifier on temporal data. Concepts that occurred in 2010 may not be relevant in 2011. For example, the term *octopus* had high information gain. Referring to *Paul the Octopus*, the star of the 2010 World Cup, this highly discriminative term is no longer relevant in 2011. In Table 3 we can see that this leads to a 4-5% reduction in the AUC scores across all different models. Also it leads to a lesser reduction in models that employ readability features, because they are not dependent on trending topics.

Due to the increase in percentage of children's documents in the test data, the precision-recall balance is changed. This results in a seemingly smaller decrease in NB performance between Tables 1 and 2, as the precision increased. Using the estimated high cut-off values from the cross-validation experiments magnifies this effect. For example, the F_1-score of the ST features is reduced while the $F_{0.5}$ is increased due to the drastic shift in the precision-recall ratio. Similarly for the STSR features, the estimated cutoff value increases the precision by 44 percentage points, from P= 37 to P= 81, while the recall only drops by 38 from P= 64 to P=36. The estimated values provide a better precision-recall balance for the children's data; however, the tuned NB still does not outperform the SVM results.

In SVM models, there is a large reduction in recall. With the LT features the precision is also reduced, while with ST the precision increases. ST features still perform better overall, however, this may reduce over time because they contain less of the general tokens, and more named entities and current events.

Pattens can be observed by inspection of wrongly classified documents near the margin. In adult data, the documents that were misclassified dealt with child-appropriate topics such as football, celebrities, video reports, animals. In the children's data, news reports about astronomy, football, and animals, as well as, some reports of major political events, like the Libya protests and Japan earthquakes, were misclassified. When judging only from the snippets, the topics of these documents appear appropriate for either age group. The full articles, on the other hand, are written for different audiences, so an interesting challenge is to predict whether an article is appropriate based on the feed snippet.

C	Data	F_1	$F_{0.5}$	P	R	AUC
NB	LT	47.19	43.39	41.18	55.26	90.31
NB	ST	51.26	46.90	44.39	**60.64**	**92.29**
NB	ST+	39.04	54.23	73.22	26.61	92.29
SVM	LT	50.31	53.99	56.76	45.18	90.55
SVM	LTLR	51.50	56.95	61.28	44.41	91.30
SVM	ST	51.53	62.62	73.10	39.79	90.62
SVM	STSR	**52.10**	**64.43**	**76.50**	39.30	91.37

Table 3. Validation experiments on the test data.

Conclusion

In this paper we have presented an investigation into the problem of identifying news feed entries appropriate for children. While a minority of available feeds are explicitly written for children, much of the rest of the content may be appropriate for them. In this paper we presented a first look into an automatic method for seeking out and filtering out appropriate content of this type.

This task is characterised by short, time-ordered documents, whose distribution is overwhelmingly skewed towards the negative class, that is, the documents written for adults. As far as we are aware there is no prior work on this novel problem, but we can compare our results to related age-based classification experiments. For webpages [4] multiple types of features, which are available in webpages, but not in feeds, were used along with an SVM to achieve an $F_{0.5}$ score of 72 on the test data. While in classification of school materials [15], text-only features were used with an SVM to classify reading materials by primary school grade level, in order to identify appropriate reading materials for children of different ages. They achieved scores between $F_{0.5} = 41$ and $F_{0.5} = 76$ across different grade levels, the rage of values which might be explained by the difference in the number of training examples for each of the grades. The problem we have examined in this paper is more difficult than classification of web pages and reading materials, in that have very short documents with few linguistic clues and no topological data. Nonetheless, our results have been quite promising, where we obtained $F_{0.5} = 64$ on test data and $F_{0.5} = 73$. This is still quite low and hardly comparable with the performance on topic-based text classification problems which usually obtains F-score greater than 85 [7]. However, manual inspection of the false positive stories indicates that they mainly cover news events that are typically covered in child-specific feeds, and thus may not actually be inappropriate for children.

Since ensuring that the content is appropriate for children is a very sensitive area, i.e., false positives may have serious consequences, significantly more research is required in this area. We intend to build on this research and focus on a number of tasks: (i) to make the models more resilient to dissipation or introduction of concepts over time, (ii) to broaden the problem area and examine a greater variety of feeds and sources, (iii) to introduce human annotation to validate the labelling of test data (as stories which are recommended from feeds for adults may be suitable for children).

Acknowledgements

This research is supported by the PuppyIR project and is funded by the EC's FP7 2007-2013 under grant agreement no. 231507.

References

1. D. Bilal, S. Sarangthem, and I. Bachir. Toward a model of children's information seeking behavior in using digital libraries. In *IIiX '08*, pages 145–151, NY, USA, 2008. ACM.

2. R. R. Bouckaert and E. Frank. Evaluating the replicability of significance tests for comparing learning algorithms. In H. Dai, R. Srikant, and C. Zhang, editors, *Advances in Knowledge Discovery and Data Mining*, volume 3056 of *Lecture Notes in Computer Science*, pages 3–12. Springer Berlin / Heidelberg, 2004.
3. A. Druin, E. Foss, H. Hutchinson, E. Golub, and L. Hatley. Children's roles using keyword search interfaces at home. In *CHI '10*, pages 413–422, New York, NY, USA, 2010. ACM.
4. C. Eickhoff, P. Serdyukov, and A. P. de Vries. A combined topical/non-topical approach to identifying web sites for children. In *4th International Conference on Web Search and Data Mining (WSDM)*, Hong Kong, 2011. ACM, ACM.
5. R. Glassey, D. Elliott, T. Polajnar, and L. Azzopardi. Interaction-based information filtering for children. In *Proceeding of the third symposium on Information interaction in context*, IIiX '10, pages 329–334, NY, USA, 2010. ACM.
6. K. Gyllstrom and M.-F. Moens. A picture is worth a thousand search results: finding child-oriented multimedia results with collage. In *SIGIR*, pages 731–732, 2010.
7. T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the European Conference on Machine Learning*, pages 137–142. Springer, Chemnitz, Germany, 1998.
8. T. Joachims. *Advances in Kernel Methods - Support Vector Learning*, chapter Making large-Scale SVM Learning Practical. MIT-Press, Cambridge, Massachusetts, 1999.
9. A. Large, J. Beheshti, N. Tabatabaei, and V. Nesset. Developing a visual taxonomy: Children's views on aesthetics. *J. Am. Soc. Inf. Sci. Technol.*, 60(9):1808–1822, 2009.
10. A. Large, V. Nesset, and J. Beheshti. Children as information seekers: what researchers tell us. *New Review of Childrens Literature and Librarianship*, 14(2):121–140, 2008.
11. D. D. Lewis. Naive (Bayes) at forty: The independence assumption in information retrieval. In *ECML '98: Proceedings of the 10th European Conference on Machine Learning*, pages 4–15, London, UK, 1998. Springer-Verlag.
12. J. Nielsen. Usability of websites for children: Design guidelines for targeting users aged 3-12 years. Web publication, http://www.nngroup.com/reports/kids/, September 2010.
13. Ofcom. UK children's media literacy. http://stakeholders.ofcom.org.uk/market-data-research/media-literacy/medlitpub/medlitpubrss/ukchildrensml/, March 2010.
14. V. J. Rideout, U. G. Foeh, and D. F. Roberts. Report: Generation m2: Media in the lives of 8- to 18-year-olds. Technical report, Kaiser Family Foundation, 01 2010. http://www.kff.org/entmedia/upload/8010.pdf.
15. S. E. Schwarm and M. Ostendorf. Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 523–530, Ann Arbour, USA, 2005. ACL.
16. E. Smith and R. Senter. Automated readability index. Technical Report AMRL-TR-66-220, Cincinnati University, Ohio, 1967.