Leveraging Wikipedia article evolution for promotional tone detection

Christine de Kock
Department of Computer Science
and Technology
University of Cambridge*
cd700@cam.ac.uk

Andreas Vlachos
Department of Computer Science
and Technology
University of Cambridge
av308@cam.ac.uk

Abstract

Detecting biased language is useful for a variety of applications, such as identifying hyperpartisan news sources or flagging one-sided rhetoric. In this work we introduce WikiEvolve, a dataset for document-level promotional tone detection in English. Unlike previously proposed datasets, it contains seven versions of the same article from Wikipedia, from different points in its revision history; one with promotional tone, and six without it. We adapt the gradient reversal layer framework to encode two article versions simultaneously, and thus leverage the training signal present in the multiple versions. In our experiments, our proposed adaptation of gradient reversal improves the accuracy of four different architectures on both in-domain and out-of-domain evaluation.

1 Introduction

Maintaining a neutral point of view is a desideratum in many communication channels, e.g. news articles, scientific writing, and encyclopaedias. Biased writing detection can help reduce the distribution of content which contains unfair representations of a topic. For this reason, datasets and methods have been developed to automate it.

A number of studies have approached biased writing detection in the context of news media (e.g. Fan et al. (2019); Chen et al. (2020); Färber et al. (2020)), primarily considering political stance and partisanship. However, biased writing also arises in other settings. In Wikipedia, the online encyclopaedia, it manifests itself in the form of promotional tone which violates the cornerstone "neutral point of view" policy of the platform. The latter allows users to flag articles with such policy violations by adding tags to the article mark-up, which are retained in its edit history. Leveraging this process, Recasens et al. (2013) and Aleksandrova et al. (2019) have released datasets of words and sentences which were altered in subsequent revisions, thus facilitating model development for word/sentence level bias detection.

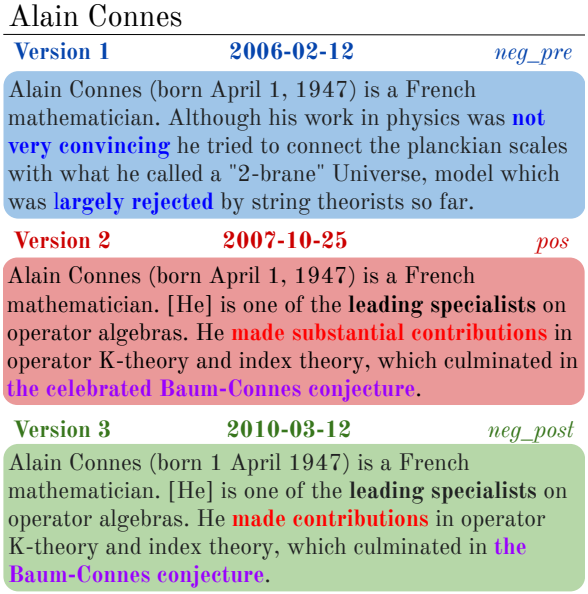

Figure 1: A sample from the dataset. The middle version was tagged as having a promotional tone problem. The first and third versions were sampled respectively before the tag was added and after it was removed.

In this work, we propose an alternative data collection methodology for document-level promotional tone detection, We sample multiple versions of the same article in Wikipedia and present WikiEvolve[1], a dataset of 68,498 labelled articles for this task. These articles are arranged into 13,887 sample sets, where each set contains multiple versions of the same article: one version tagged as having a promotional tone problem, and up to three versions respectively from before the tag was added and after it was removed. This is illustrated in Fig. 1; the second version was labelled as containing

*This work was initiated during an internship at the Wikimedia Foundation.

[1] github.com/christinedekock11/wiki-evolve

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 5601 - 5613

promotional tone (positive), whereas the first and third versions were considered negative.

In contrast with Recasens et al. (2013) and Aleksandrova et al. (2019), we choose to perform classification at the level of documents rather than sentences or words. Our motivation is that classifying a sentence out of context as biased is known to be difficult and prone to subjective judgements, while higher inter-annotator agreement is achieved at the document-level (Chen et al., 2020). Recasens et al. (2013) similarly found that identifying promotional tone at the word-level is challenging, with Mechanical Turk workers achieving 37% accuracy on this task. We hypothesise that there are article-level features which provide corroborating evidence to the intentions of the writer, which isolated sentences might not capture. We also see evidence of this in our own data – for instance, in Fig. 1, the mention of "leading specialist" in version 3 is dubious but justifiable; however, in version 2, it contributes to an overall assessment of biased writing.

To make better use of the training signal available in the multiple versions per article in WikiEvolve, we adapt gradient reversal (Ganin and Lempitsky, 2015). The latter entails adding an auxiliary task during training which shares the input encoder with the main task and is optimised concurrently, but its gradients are reversed during backpropagation. The model is therefore discouraged from learning features which are useful for the auxiliary task and assumed harmful for the main task. Our adaptation operates on pairs of samples rather than individual texts, and we define the auxiliary task as classifying whether two samples originated from the same article. The features we learn are therefore more likely to be informative of the tone, but not of the content.

In our experiments, gradient reversal improves the accuracy of all four architectures of increasing complexity. On a bag-of-words encoding followed by two neural network layers, the PR-AUC score improves from 0.60 to 0.64. Using a hierarchical attention network, performance is increased from 0.63 to 0.65. This illustrates that the additional structural information WikiEvolve provides can be utilised to improve performance on this task. To further assess the ability of gradient reversal to improve performance by encouraging models to learn features that do not rely on the topic or content, we also tested our models on out-of-domain data from the SemEval 2019 Shared Task on Hyperpartisan News Detection (Kiesel et al., 2019). Our results show that GRL training improves our accuracy on this dataset from 0.714 to 0.785.

2 Promotional tone on Wikipedia

A number of studies have utilised Wikipedia to develop labelled datasets for content-related issues, including promotional tone detection. Wikipedia has several favourable characteristics which enable this form of data collection. Firstly, articles evolve over time through different versions. Secondly, the chronological revision history of each article is preserved and open-sourced[2], meaning that the evolution of an article can be retrieved. Finally, the platform's decentralised quality control system allows users to tag articles that violate the platform's content policies, to warn readers of such issues and to attract the attention of editors to fix them. These tags are removed from the article once the problem is resolved, but they are preserved in the article's edit history. A more details on Wikipedia's policy violation tags see Anderka et al. (2012).

In this context, a revision of an article which contains a tag is considered a positive instance of that specific policy violation. Different methods have been proposed for sampling negatives. A popular approach is to find revisions of the same or other articles which do not contain the tag. However, this approach can introduce noise, as the absence of a policy violation tag from an article does not guarantee that the problem is not present. This characteristic of template-based Wikipedia datasets has been noted in previous work, e.g. Anderka et al. (2012); Bhosale et al. (2013); Orizu and He (2018). Another option is to look to other articles which are known to represent well-written content. For instance, Anderka et al. (2012) and Bhosale et al. (2013) select negatives from Wikipedia's list of featured and good articles. However, these articles are of a higher quality generally and therefore have other distinguishing characteristics, which may be misleading if the goal is to detect policy-violating content. Additionally, sampling negatives from different articles may introduce a topical bias.

3 Mining promotional articles

Our data extraction methodology consists of (i) finding articles tagged by a Wikipedia editor as having a promotional tone problem at some point

[2] A Creative Commons Attribution-Share-Alike License 3.0 applies.

in their edit history, (ii) selecting the revision where such a tag was added as a positive sample, and (iii) sampling negatives from revisions which did not contain the template.

Finding promotional tone tags To identify tags of interest, we refer to the Wikipedia category "articles with a promotional tone" (Wikipedia, 2021a) and identify the quality tags which most frequently occur in this articles of this category. These are "advert", "autobiography", "fanpov", "peacock" and "weasel". Each of these tags describes a different type of promotional tone issue, for which the definitions are contained in Appendix A. We then use regular expressions to collect all revisions which contain variations of these tags in the WikiText data lake (Wikipedia, 2021c).

Finding tag addition events Once incidences of promotional tone tags have been identified, we use the WikiHistory data lake (Wikipedia, 2021b) to find the full edit histories of these articles. For each article, we then identify the point in its edit history where a tag was added, and consider this version of the article as the positive sample. We exclude cases where the tag addition edit was reverted[3] by another editor. The article text at this timestamp is retrieved from the WikiText data lake.

Sampling negatives For each positive sample, we select negatives from the revision history of the same article. We consider as candidates all revisions which were not reverted, and which took place within 30 revisions (chronologically sorted) of the tag addition event. This is intended to ensure that the negative samples are of the same approximate stage of article development as the positive sample. We exclude the revision immediately before the tag addition event, as it is this version which prompted the tag to be added. Up to three revisions (depending on availability) are selected at random from these candidates, before and after the positive. We refer to such a set of samples as a **sample set**. The negatives sampled before the tag addition are denoted ***neg_pre***, and those from after are denoted ***neg_post***. The number of samples per tag and class are shown in Table 1. We split the data into train, test and validation sets with a ratio of 70-20-10. The datasets are stratified to contain samples from each tag type (set out in Table 1), and samples from the same sample set (i.e. revisions of the same article) are kept in the same split.

[3]From the platform guidelines: "On Wikipedia, reverting means undoing or otherwise negating the effects of one or more edits, which results in the page (or a part of it) being restored to a previous version."

Tag	# Positives	# Negatives
Autobiography	1578	9289
Advert	4361	25843
Fanpov	413	2446
Peacock	2859	16960
Weasel	906	5421
Total	**8 539**	**59 959**

Table 1: Number of samples per tag and class.

Although this work only considers promotional tone detection in English, the data collection methodology and training framework we propose could be extended to other languages on Wikipedia, as is done in Aleksandrova et al. (2019).

4 Data validation

As discussed in Sec. 2, Wikipedia tag-based datasets are known to contain a certain level of noise. To counteract this, we have implemented three measures: ensuring that the negatives are from the same stage of article development, sampling from different points in the same article's edit history, and sampling negatives before and after the positive. However, there is still a risk of including false negatives, i.e. articles not tagged as containing promotional tone even though they do. An example of such a case from our dataset is shown in Fig. 2. Despite containing non-neutral phrases such as "hit show", "made quite an impression", and "prove[d] herself to be intelligent", the *neg_pre* (first) sample is not tagged as containing biased language. It does however contain some information that reflects negatively on the subject ("for the wrong reasons"). This is removed in the positive (middle) sample, and more overtly biased descriptions are added ("quick wit, educational background, amazing looks", "bubbly personality and easy on the eye appearrance"). In the *neg_post* negative sample, the problematic phrasing is removed.

We perform manual validation of our dataset to estimate how frequently false negatives are included. We perform two types of validation: pairwise and independent prediction. For the former, the task is to rank two samples (i.e. revisions of the same article) as to which is more promotional. 40 articles, consisting of 20 positive-negative pairs in random order, are evaluated by two of the authors.

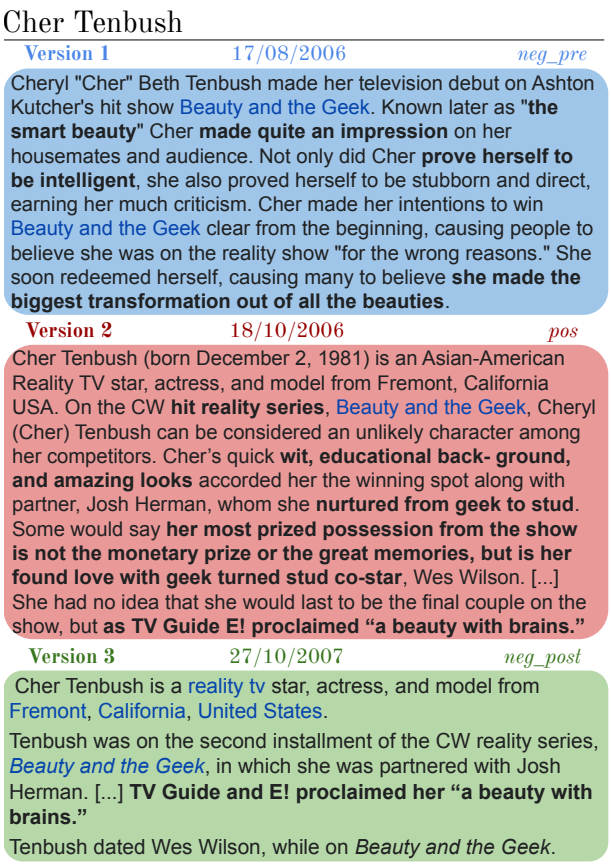

Figure 2: An example of a false negative obtained from the *neg_pre* version of a sample set.

The orderings of the annotators agreed with the assigned labels for respectively 16/20 and 14/20 pairs, with a Cohen's Kappa score of 0.79 indicating substantial agreement. This suggests that the collected data contains a trustworthy signal for comparing the extent to which two texts are promotional.

Since the task we are mainly interested in is text classification, rather than ranking, we also perform an evaluation on individual samples, annotating 30 samples of each type (positive, *neg_pre* and *neg_post*). The concurrence with the mined labels of the *neg_pre* and *neg_post* annotations are shown in Table 2. This task appears to be more challenging compared to the pairwise comparison, with both annotators achieving lower scores and a lower inter-annotator agreement Kappa score of 0.4805, indicating moderate agreement. A reason for this may be the subjective nature of the task, as illustrated by Chen et al. (2020).

Our evaluation indicates that the negative samples from before the tag was added contain more noise, compared to those sampled after it was removed. This can be attributed to the active removal of the tag by an editor in the version after

Annotator	***neg_pre***	***neg_post***
A1	$\frac{12}{30}$	$\frac{14}{30}$
A2	$\frac{14}{30}$	$\frac{22}{30}$
Total	$\frac{24}{60}$	$\frac{36}{60}$

Table 2: Agreement of two authors with mined labels of negative sample annotations.

the tag was added (*neg_post*), which indicates that the problem is resolved, while the lack of a policy violation tag in earlier versions (*neg_pre*) does not guarantee lack of promotional tone. However, ignoring the *neg_pre* samples altogether would expose the temporal bias mentioned in Sec. 1: if negatives are always sampled chronologically after positives and from a more developed version of the article, spurious correlations may be inferred.

Based on these insights, we have chosen to include the automatically mined *neg_pre* samples in training, but to create a separate set of manually validated *neg_pre* samples for evaluation. Thus, we randomly selected 100 *neg_pre* samples from the original test set and verified whether they represent a neutral writing style. 42 of the 100 samples were confirmed as true negatives. We balance these negatives with their corresponding positive samples, and refer to this dataset as *ValidNegPre*.

5 Gradient reversal training for promotional tone detection

Gradient reversal training (Ganin and Lempitsky, 2015) jointly optimises two classifiers which rely on a shared underlying encoder model: (i) a label predictor for the main task, which predicts class labels and is used during both training and test time, and (ii) a domain classifier, which predicts either the source or the target domain during training as the auxiliary task. The parameters of the encoder model are optimised to minimise the loss of the main task classifier while maximising the loss of the domain classifier. This is achieved through a gradient reversal layer, which leaves the input unchanged during forward propagation and reverses the gradient by multiplying it by a negative scalar during the backpropagation. This approach is motivated by theory on domain adaptation, which suggests that a good representation for cross-domain transfer is one for which an algorithm cannot learn to identify the domain of origin of the input observation (Ben-David et al., 2010).

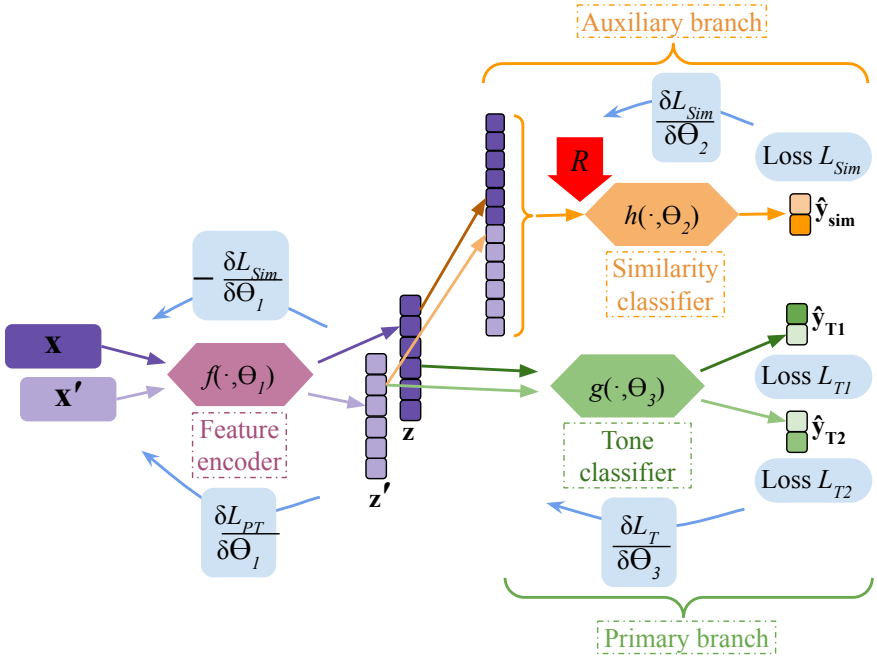

Figure 3: Model architecture for training with gradient reversal. x and x' represent two different samples which are encoded in parallel, by the same feature encoder model.

Our adaptation of this framework, shown in Fig. 3, differs from Ganin and Lempitsky (2015) in that it considers two text inputs concurrently (x and x'), as opposed to one. f represents a neural network encoder with parameters θ_1. f encodes the two texts independently, to produce z and z':

$$z = f(x;\theta_1); z' = f(x';\theta_1) \quad (1)$$

The network then splits into two branches. The primary (bottom) branch consists of a neural network model g, with parameters θ_3, which produces promotional tone predictions $\hat{y}_{T1}$ and $\hat{y}_{T2}$ for the two samples:

$$\hat{y}_{T1} = g(z;\theta_3); \hat{y}_{T2} = g(z';\theta_3) \quad (2)$$

The auxiliary branch concatenates the two input encodings as $[z, z']$, and then the similarity classifier h, a neural network parameterised by θ_2, provides a prediction $\hat{y}_{sim}$ of whether the two samples originate from the same Wikipedia article:

$$\hat{y}_{sim} = h([z, z'];\theta_2) \quad (3)$$

Our intention with this task is to encourage the encoder f to learn features that are topic agnostic. This should allow for better generalisation across datasets, as well as to avoid learning spurious correlations due to topical biases in the data.

The encoder and classifier models are trained simultaneously.

Given a set of training samples D = $[x_1, ..., x_N, y_1, ..., y_N]$, we construct M pairs with indices $P = (i, j) : i, j \in [1, ..., N]$. The process for generating these pairs is described in Sec. 6. Then, the loss is given by:

$$\begin{aligned} L(\theta_1, \theta_2, \theta_3, D, P) = \\ \frac{1}{M}\sum_{m=1}^{M}\Big(L_T(x_{P1}, x_{P2};\theta_1,\theta_2) \\ - \lambda L_{Sim}(x_{P1}, x_{P2};\theta_1,\theta_3)\Big), \end{aligned} \quad (4)$$

such that the loss with respect to the similarity label is maximised, while the loss with respect to the promotional tone label is minimised. λ is a scalar which controls the weight of the loss from the adversarial task, and $L_T = L_{T1} + L_{T2}$.

During testing, only the feature encoder and the main task branch are retained to perform the tone classification task:

$$\hat{y} = g(f(x;\theta_1),\theta_3). \quad (5)$$

Recall that the model encodes and predicts on the two input samples independently during training. We can therefore obtain predictions for individual test samples (rather than pairs), as is the more general case in other models and datasets for this task.

6 Experimental setup

6.1 Models

The two classifier models, g and h, are both MLP models. The feature encoder (indicated as f in Fig. 3) is responsible for producing an embedding of an article to be used in both the main and auxiliary task. We evaluate four options:

- **Bag-of-words (BoW + MLP)**: a bag-of-words representation of an article is propagated through a multilayer perceptron (MLP) to obtain an embedding,
- **Averaged embeddings (AvgEmb + MLP)**: GloVe embeddings (Pennington et al., 2014) for every word in the article are averaged, followed by an MLP model,
- **Hierarchical Attention Network (HAN)** (Yang et al., 2016): word embeddings are processed using an LSTM layer followed by an attention mechanism to build up sentence embeddings. Sentence embeddings are similarly combined to form an article embedding.
- **Longformer** (Beltagy et al., 2020): A transformer-based model, adapted for long-form documents. We finetune the pretrained "longformer-base-4096" model.

For the GRL models, we further experiment with the weights of the main versus auxiliary task λ on the validation set, finding that weighting the outputs equally yields the best results.

We compare the GRL training approach with the standard method of training the classifier with each feature extractor model. This is equivalent to training with the inference model in Equation 5; the auxiliary branch is removed and one sample is processed at a time. Implementation details are provided in App. B.

6.2 Metrics

For each model, we report two metrics:

- **PR-AUC**: The area under the precision-recall curve, which provides an aggregate measure of performance across all possible classification thresholds (Davis and Goadrich, 2006). Perfect performance is 1, and a random classifier would receive 0.
- **Accuracy**: The percentage of samples which are correctly classified, using a classification threshold based on Youden's J statistic (Fluss et al., 2005), which maximises the true positive rate and minimises the false negative rate.

6.3 Data

In order to train the main task we require samples with and without a promotional tone (i.e. positive and negative labels). To train the auxiliary we require both *matched* pairs (originating from the same sample set / article) and *unmatched* pairs (originating from different sample sets). Therefore, we include a number of different pairing configurations.

Firstly, given a training set consisting of K sample sets, we collect K positive-negative matched pairs. This means that we need to select one negative sample and one positive sample from each sample set. There are multiple negatives in every sample set, so we sample at random from all *neg_pre* and *neg_post* samples. There is only one positive per sample set, so this sample is used. We also collect K positive-negative unmatched pairs. We further include K matched and K unmatched negative-negative pairs. Finally, we include K positive-positive unmatched pairs. It is not possible to add positive-positive matched pairs, as there is only one positive per sample set.

Using this pair selection method, there are $7K$ samples for the tone classification task and $3.5K$ pairs for the similarity classification task, for a total of 48762 articles. For the baseline models, without GRL, only one sample is used at a time during training; thus, we retain the data generation method described above (to ensure the results are comparable), but ignore the pairings.

The training dataset is slightly unbalanced, with a ratio of 4:3 of negatives to positives for the similarity classification task and a ratio of 4:3 of unmatched to matched pairs. The numbers of samples per label and their origin are shown in Table 7 in App. C. Our validation and test sets consist of only positive-negative matched pairs, one from each sample set, and thus are fully balanced. As motivated in Sec. 4, for the main test set (denoted *FullTest*) negatives are only selected from the *neg_post* samples. For the *ValidNegPre* test set, all negatives are manually validated. The text preprocessing steps are described in App. B.

7 Results

The results from our evaluation on the *FullTest* are in Table 3. We observe that models trained with GRL consistently outperform models trained without it, on both the accuracy and PR-AUC metrics. All improvements, except for the Longformer, are

Model	PR-AUC	Accuracy
BoW + MLP	0.6019	0.5913
BoW + GRL	0.6409	0.6102
AvgEmb + MLP	0.6129	0.5848
AvgEmb + GRL	0.6415	0.6084
HAN	0.6271	0.5968
HAN + GRL	0.6459	0.6102
Longformer	0.6798	0.6392
Longformer + GRL	0.6984	0.6432

Table 3: Results using GRL training on $FullTest$.

	Test data	
Train data	*FullTest*	*ValidNegPre*
Incl. *neg_pre*	0.6984	0.6184
Excl. *neg_pre*	0.6962	0.5725

Table 4: Effect of ignoring *neg_pre* during training. PR-AUC scores are shown.

statistically significant at the $\alpha = 0.05$ level, using the permutation test to compare PR-AUC values. Larger gains are observed for the BoW+MLP and AvgEmb+MLP models, compared to the HAN and Longformer models. A possible explanation for this is that the simpler models rely only on word-level information, and thus more susceptible to topical biases which GRL mitigates.

These results support the motivation behind our data collection method and training framework: by incorporating our knowledge of how samples are related in our dataset and training, models are exposed to different versions of the same content (with and without promotional tone), and can therefore better learn features that are more effective for detecting promotional tone, compared to models that ignore this information.

7.1 Effect of *neg_pre* samples

Given our discussion in Sec. 4, we also evaluate the GRL approach on a separate, validated test set, which uses *neg_pre* rather than *neg_post* negatives (denoted *ValidNegPre*). In this evaluation, we are particularly interested in the effect of excluding *neg_pre* samples during training. The total number of samples in the training set remains the same as for experiments already reported, but all negatives are sampled from the *neg_post* samples for each sample set during training. We compare our best model (Longformer+GRL) under these conditions. For brevity, in Table 4 we only show the PR-AUC values, but the same trends hold for accuracy.

The original configuration is shown in the top left of the table; with training data including both *neg_post* and *neg_pre*, and testing on $FullTest$. We note that, using the same training data, the PR-AUC score is slightly lower on the *ValidNegPre* set (top right) compared to the *FullTest* set, indicating that these samples may be more difficult to classify correctly. The effect of excluding *neg_pre* samples during training is shown in the second row. The two training settings achieve similar performance on the *FullTest* test set, however, the performance on the *ValidNegPre* dataset is markedly lower when excluding *neg_pre*samples. This supports our motivation for including *neg_pre* samples during training, as described in Sec. 4, i.e. that not including them may lead to learning spurious correlations, such as temporal or article development biases. The *neg_pre* sampling adds useful information during training, despite including noise in the form of false negatives.

7.2 Ranked prediction

Since our main goal is to predict promotional tone for a given text, we did not optimise for ranked prediction; however, the pairwise accuracy is of interest since the GRL-based model is trained on pairs. This is similar to the pairwise human evaluation we performed in Sec. 4. For this we calculate the proportion of pairs for which the directionality of the predictions is correct. A score of 0.722 is achieved for the non-GRL Longformer model, compared to 0.741 for the GRL model. The fact that these values are higher than the accuracy values in Table 3 illustrates that there are samples which were incorrectly classified, but whose relative (pairwise) relationship was correctly predicted.

7.3 Error analysis

To better understand the differences in predictions made by models trained with GRL, we analyse more closely the test set and our predictions. There are 1318 samples on which both models are correct and 764 on which both are incorrect. There are 320 samples where the non-GRL model is correct while the GRL model is incorrect, and 356 samples in the reverse case. We are interested in the last two categories, where the two models disagree.

To better understand these classification categories, we evaluate the pointwise mutual information (PMI; Jurafsky and Martin, 2008) of each word

(w) with its classification status (c):

$$PMI(w, c) = log_2(\frac{P(w, c)}{P(w)P(c)}). \quad (6)$$

This gives us an indication of how much higher the probability of observing a word is to be in one of the categories, compared to the full test set.

The 50 words with the highest PMI, which were correctly classified as not promotional by our model but mislabeled as promotional by the non-GRL variant, are shown in Table 8 in Appendix D. Without GRL, these words were indicative of promotional tone; but with GRL, their use for promotional tone detection was reduced. Thus, these should be words that are misleading for the tone classifier, but helpful for the similarity classifier.

The list includes the terms "feminist", "feminism" and "female". This topical concentration may be caused by a bias in the training data, whereby there are more positive examples which contain these terms. Such an imbalance in the data may be related to the findings of Wagner et al. (2016), which explores the imbalance in representations of women versus men on Wikipedia. However, this is not the only topical bias we observe in the predictions of the non-GRL model; the terms "photograph", "photographer", and "graphics" are also in this list.

The PMI values for the opposite case where the non-GRL model is correct while the GRL model is incorrect is shown in Table 9. Here, too, we see some topical groupings; eg. "tumor", "physicians", "diagnosis". However, the PMI of these words are lower than that of the samples where the GRL model was correct (with a maximum PMI of 3.92 vs 2.76), meaning that the co-occurrence is on the whole lower.

7.4 Out-of-domain evaluation

We further evaluate our model on the test set from the "per-article" track of the SemEval 2019 Shared Task on Hyperpartisan News detection (Kiesel et al., 2019). Their dataset contains 314 positive (hyperpartisan) and 314 negative (not hyperpartisan) news articles. On this dataset, the Longformer+GRL model, trained on our training data, achieves a PR-AUC score of 0.759 (accuracy 0.785), compared to a PR-AUC of 0.736 (accuracy 0.714) when the GRL is omitted (statistically significant; P=0.043 on the signed rank test). The shared task received 42 entries and closed in June 2019. Compared against their leaderboard, our model would be ranked eighth, even though it was not trained on the provided training data.

Test set	No GRL	Content	Time
FullTest	0.6936	0.6984	0.6901
ValidNegPre	0.5769	0.6184	0.5948
SemEval	0.6942	0.785	0.7531

Table 5: The results on each of the test scenarios from Sec. 7, comparing models with no auxiliary task, the content-based task we proposed, and the time-based auxiliary task.

7.5 Time-based gradient reversal

A motivation for including the *neg_pre* samples is that they counteract the temporal bias introduced by only sampling *neg_post* samples. The gradient reversal layer also provides a debiasing mechanism, used to suppress topic-based biases in our proposed model. To observe the impact of the *neg_pre* sampling, we also evaluate models trained with a time-based auxiliary task. Specifically, we define the task as predicting which sample in an input pair is earlier in the revision history of an article. We use only *neg_post* samples, as they were found to be less noisy. Samples are generated from (*neg_post*, *positive*) pairs as well as (*neg_post*, *neg_post*) pairs, with the chronological ordering being swapped at random to give an equal probability of both outcomes. The results on each of the test scenarios from Sec. 7 are shown in Table 5, using the Longformer feature encoder and comparing to the results from the original formulation. We also compare to the same model trained without an auxiliary task. The Time-GRL model outperforms the model with no auxiliary task on the *ValidNegPre* and SemEval datasets, but the content-GRL model scores the highest on all three test sets. This indicates that using *neg_pre* samples to counter temporal biases and the auxiliary task to counter content biases achieves better performance on this task.

7.6 Comparing against sentence-level models

Previous work by Aleksandrova et al. (2019) explored a similar dataset creation strategy, using Wikipedia tags to identify sentences with a promotional tone. Our work focuses on document-level promotional tone detection, however, we also compare performance on our dataset using their models to verify whether document-level training captures more information than sentence-level training.

Model	Mean aggr.	Max aggr.
BoW+LogReg	0.55	0.54
LSTM+Attn	0.34	0.31

Table 6: Performance of models from Aleksandrova et al. (2019) on the *FullTest* dataset.

We replicate their best performing model and the reported test set score of F1 score of 0.62. To compare performance on our own document-level data, we obtain a prediction for each sentence in an article and apply two aggregation strategies: using the average prediction and the maximum score. We further implement an LSTM model with attention, which is similar to our HAN model without the hierarchical computation. Results are shown in Table 6. In both cases, the mean aggregation yields a slightly better score; however, models trained on our data, both with and without the GRL optimisation, achieve significantly higher scores, providing support that there is useful information contained in WikiEvolve for the task of document-level promotional tone detection.

Finally, it worth noting that the LSTM+Attn model performs worse than the BoW+LogReg model. The authors also report comparatively worse performance for a model using FastText (Bojanowski et al., 2017) embeddings.

8 Conclusion

In this work, we have proposed an alternative data collection method and dataset for promotional tone detection, which leverages the evolution of articles on the platform. To utilise the additional structure in our dataset, we extended the gradient reversal framework to train models which are more effective at detecting promotional tone. This was shown both on our own test set and on a test set from a different domain. We further provided insights on the effects of two negative sampling strategies on Wikipedia. These findings should be useful for researchers who use Wikipedia-based data more broadly, in addition to those who work on biased language detection.

Acknowledgements

Christine de Kock is supported by scholarships from Huawei and the Oppenheimer Memorial Trust. Andreas Vlachos is supported the EPSRC grant no. EP/T023414/1: Opening Up Minds. This work was initiated during an internship at the Wikimedia Foundation. We would like to thank the Foundation for granting us access to their data and resources, and in particular Diego Saez-Trumper for his support of the project.

References

Desislava Aleksandrova, François Lareau, and Pierre André Ménard. 2019. Multilingual sentence-level bias detection in wikipedia. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 42–51.

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, page 981–990, New York, NY, USA. Association for Computing Machinery.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. 2010. A theory of learning from different domains. *Machine learning*, 79(1):151–175.

Shruti Bhosale, Heath Vinicombe, and Raymond Mooney. 2013. Detecting promotional content in Wikipedia. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1851–1857, Seattle, Washington, USA. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

Wei-Fan Chen, Khalid Al Khatib, Benno Stein, and Henning Wachsmuth. 2020. Detecting media bias in news articles using Gaussian bias distributions. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4290–4300, Online. Association for Computational Linguistics.

Jesse Davis and Mark Goadrich. 2006. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240.

Lisa Fan, Marshall White, Eva Sharma, Ruisi Su, Prafulla Kumar Choubey, Ruihong Huang, and Lu Wang. 2019. In plain sight: Media bias through the lens of factual reporting. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6343–6349.

Michael Färber, Victoria Burkard, Adam Jatowt, and Sora Lim. 2020. A multidimensional dataset based on crowdsourcing for analyzing and detecting news bias. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, CIKM '20, page 3007–3014, New York, NY, USA. Association for Computing Machinery.

Ronen Fluss, David Faraggi, and Benjamin Reiser. 2005. Estimation of the youden index and its associated cutoff point. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 47(4):458–472.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR.

Daniel Jurafsky and James H Martin. 2008. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.

Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. 2019. SemEval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 829–839, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Udochukwu Orizu and Yulan He. 2018. Content-based conflict of interest detection on Wikipedia. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1650–1659.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Claudia Wagner, Eduardo Graells-Garrido, David Garcia, and Filippo Menczer. 2016. Women through the glass ceiling: gender asymmetries in wikipedia. *EPJ Data Science*, 5:1–24.

Wikipedia. 2021a. Category: Articles with a promotional tone. https://en.wikipedia.org/wiki/Category:Articles_with_a_promotional_tone. Accessed: 2021-11-15.

Wikipedia. 2021b. MediaWiki History. https://wikitech.wikimedia.org/wiki/Analytics/Data_Lake/Edits/MediaWiki_history. Accessed: 2021-11-15.

Wikipedia. 2021c. MediaWiki Wikitext history. https://wikitech.wikimedia.org/wiki/Analytics/Data_Lake/Content/Mediawiki_wikitext_history. Accessed: 2021-11-15.

Wikipedia. 2021d. Template: Advert. https://en.wikipedia.org/wiki/Template:Advert. Accessed: 2021-11-15.

Wikipedia. 2021e. Template: Autobiography. https://en.wikipedia.org/wiki/Template:Autobiography. Accessed: 2021-11-15.

Wikipedia. 2021f. Template: Fanpov. https://en.wikipedia.org/wiki/Template:Fanpov. Accessed: 2021-11-15.

Wikipedia. 2021g. Template: Peacock. https://en.wikipedia.org/wiki/Template:Peacock. Accessed: 2021-11-15.

Wikipedia. 2021h. Template: Weasel. https://en.wikipedia.org/wiki/Template:Weasel. Accessed: 2021-11-15.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489.

A Tag definitions

The definitions for the promotional tone tags used in this study are as follows, as per the relevant Wikipedia policy descriptions:

- **Fanpov**: Written from a fan's point of view, rather than a neutral point of view (Wikipedia, 2021f).
- **Peacock**: Contains wording that promotes the subject in a subjective manner without imparting real information (Wikipedia, 2021g).
- **Autobiography**: Article is extensively edited by the subject or by someone connected to the subject (Wikipedia, 2021e).
- **Advert**: Contains content that is written like an advertisement (Wikipedia, 2021d).

- **Weasel**: Vague phrasing that often accompanies biased or unverifiable information (Wikipedia, 2021h).

B Implementation details

Hyperparameters All models are implemented in Keras. For the BoW+MLP and AvgEmb+MLP encoder models, we use two-layer neural networks. The GRL label predictor and similarity predictor models also consist of two neural network layers. We experiment with learning rates in [0.0001,0.001,0.01], dropout (Srivastava et al., 2014) in [0.1,0.3,0.5] and hidden layer sizes in [64, 128, 256] on the validation set. The ReLU activation function is used. For the Longformer model, we finetune the `longformer-base-4096` snapshot using the Huggingface[4] package, using a learning rate of 5×10^{-6}.

Models were trained on a Nvidia Quadro RTX8000 GPU available at the authors' institution and training finished within less than 36 hours in all cases.

Text preprocessing The articles are preprocessed using the `mwparserfromhell` library, which extracts the article text from the marked-up wikicode. We remove the sections 'See also', 'External links' and 'References', as these mainly contain references to other sources rather than content. The resulting samples have a median length of 615 tokens. Samples longer than 1024 tokens are truncated.

C Data configuration for GRL training

The data configuration for GRL training is shown in Table 7.

D Error analysis

The top 50 words in terms of PMI for the two categories discussed in Sec. 7.3 are shown in Tables 8 and 9.

[4] https://huggingface.co/

Type	Similarity labels	Tone labels
Train (K=6 966)		
(pos,neg) pair from same article	{0:K}	{0:K, 1:K}
(pos,neg) pair from different articles	{1:K}	{0:K, 1:K}
(neg,neg) pair from same article	$\{0:\frac{K}{2}\}$	{0:K}
(neg,neg) pair from different articles	$\{1:\frac{K}{2}\}$	{0:K}
(pos,pos) pair from different articles	$\{1:\frac{K}{2}\}$	{1:K}
Total	**{0:1.5n,1:2n}**	**{0:4K, 1:3K}**
Validation (K=931)		
(pos,neg) pair from same article	{0:K}	{0:K,1:K}
Test (K=1 379)		
(pos,neg) pair from same article	{0:K}	{0:K,1:K}

Table 7: Train, validation and test set configuration. K refers to the number of sample sets in each data split. For the similarity classification task, 0 means same article, 1 is different. For the tone classification task, 1 is promotional, 0 is not.

Word	PMI
wow	3.916
continental	3.859
graham	3.497
fruit	3.470
aids	3.455
understood	3.105
feminist	3.067
poems	3.051
feminism	3.049
enemy	2.996
relate	2.926
gender	2.903
photographs	2.882
graphics	2.833
comparative	2.815
vancouver	2.807
bangalore	2.789
poets	2.744
translated	2.744
cry	2.709
abstract	2.699
implications	2.699
empowerment	2.686
tells	2.669
chapter	2.646
realize	2.640
strongly	2.595
junior	2.595
exist	2.590
playwright	2.590
poetry	2.577
choreographer	2.537
otherwise	2.477
contributing	2.450
berkeley	2.439
researcher	2.427
1945	2.425
photographer	2.401
relationships	2.372
pink	2.354
warren	2.351
singers	2.345
trends	2.318
writings	2.318
involving	2.303
animal	2.289
female	2.283
photography	2.274
evil	2.274
tiger	2.274

Table 8: Top 50 PMI words for samples which were correctly classified as not promotional by the GRL model, but incorrectly classified as promotional by the non-GRL model.

Word	PMI
giovanni	2.761
weil	2.591
tumor	2.485
provision	2.385
westminster	2.326
dee	2.298
protein	2.278
compilation	2.258
jung	2.248
physicians	2.217
malaysia	2.189
einstein	2.182
nervous	2.163
wolfgang	2.163
operatic	2.163
catalog	2.149
seats	2.144
kimmel	2.140
tumors	2.134
breast	2.134
injuries	2.129
dennis	2.120
sequences	2.112
cells	2.102
bafta	2.096
reduced	2.096
thomas	2.088
cbn	2.085
linda	2.085
stood	2.082
diagnosis	2.077
murphy	2.063
scene	2.053
causing	2.044
suffered	2.044
soprano	2.031
listing	2.022
migration	2.013
georgetown	2.007
madison	1.996
warming	1.989
verlag	1.970
mp	1.953
brain	1.936
postdoctoral	1.934
rates	1.930
experiment	1.929
researchers	1.923
disaster	1.917
replace	1.916

Table 9: Top 50 PMI words for samples which were correctly classified as not promotional by the non-GRL model, but incorrectly classified as promotional by the GRL model.