Are You Following the Right News-Outlet? A Machine Learning based approach to outlet prediction

Swati
swati@ijs.si
Jožef Stefan Institute
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia

Dunja Mladenić
dunja.mladenic@ijs.si
Jožef Stefan Institute
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia

ABSTRACT

In this work, we propose a benchmark task of outlet prediction and present a dataset of English news events tailored to the proposed task. Addressing this problem would not only allow readers to choose and respond to relevant and broader facets of events but also enable the outlets to examine and report on their work. We also propose a neural network based approach to recommend a list of probable outlets covering an event of interest. Evaluation results reveal that even in its simplest form, our model is capable of predicting the outlet significantly better than the existing rule based approaches. The proposed model will also serve as a baseline for evaluating approaches intended to address the task. Implementation scripts can be found at *https://github.com/Swati17293/outlet-prediction.*

KEYWORDS

News bias, Event Selection bias, News coverage, News Event Analysis, Recommendation System

1 INTRODUCTION

The advancement in the field of Natural Language Processing [9, 10, 5, 4] over the last decade, has made solutions to complex machine learning problems more convenient. The problems such as machine translation, text summarization, and segmentation are being solved much more efficiently than ever before. Consequently, it offered the researchers the opportunity to use these advanced techniques to solve problems in a variety of contexts such as news bias analysis. This analysis task is poised as the identification of the inherent bias present in the news production and its coverage process. It occurs when a news outlet publishes a news story selectively or incorrectly.

If the news is biased, then it can bias the thought process and decision making of the person listening, watching, and/or reading it [12]. It can have several direct or indirect implications whether political or social. For example, if the news shows only the positive or negative side of a political party; it has been observed to influence the public vote [2]. Not only politics but also the news about the disaster or spread of viral disease affects the belief system of the general public.

There are numerous events that happen continuously, and any form of bias can arise in numerous possible ways. It is not possible for any single outlet to capture every event. Thus, an outlet is forced to select a set of reporting events. Several factors, such as the geographical origin of the event, the involvement of an elite person or country, etc. influences such selection. Also the procedure requires rigorous monitoring of current affairs to determine the news value, and may result in event selection bias also known as gatekeeping bias.

However, no well-established automated method reveals to users the outlets that will cover the event of their interest. This drives the motivation of this study. The aim is to predict a list of outlets reporting on a given event. Addressing this problem would not only allow readers to choose and respond to relevant and broader facets of events but also enable the outlets to examine and report on their work. For instance, some outlets tend to publish events covered by well-established outlets. Instead of waiting for the news to be published, the proposed system will help them to get an insight into the degree of predictability of event selection by the major outlets.

1.1 contributions

We make the following contributions in this context:

- We propose a benchmark task of outlet prediction and present a dataset of English news events tailored to the proposed task.
- We provide a neural network model that can serve as a baseline for evaluating approaches intended to address the task.

The GitHub repository containing our code is available at ***https://github.com/Swati17293/outlet-prediction***.

1.2 Problem Statement

The problem is addressed as an outlet prediction task in which the bias is examined by comparing the learning ability of a classifier trained to predict the probability of event coverage by an outlet.

2 LITERATURE REVIEW

During the different stages of news production, various forms of news bias arise as described by Baker et al. [1]. The first stage begins with the selection of events also called gatekeeping, where an outlet selects or rejects an event for reporting. The selection process is driven by a number of factors, such as the geographical origin of the event, the involvement of an elite person or country, etc., and requires rigorous monitoring of current affairs to determine the news value. To our knowledge, only a few methods have been suggested that explicitly attempt to examine this bias.

Saez-Trumper et al. [11] attempted to identify bias in online news sources and social media groups surrounding them. They studied the disparity in the selection of events based on the quantity and exclusivity of stories published by 80 mainstream news

Information Society 2020, 5–9 October 2020, Ljubljana, Slovenia

outlets across the globe over a span of two weeks. From the review, it is found that there is a weak correlation between the quantity and exclusivity of news articles published by the outlets. It is also discovered that both the news and social media follow the same pattern of selection of events in similar geographical areas. However, media in the same region often choose the same events and publish similar-length posts.

Bourgeois et al. [3] used a matrix factorization method to extract latent factors that determine the selection of the event by an outlet. They combined the method with a BPR optimization scheme developed by Rendle et al.[8]. They used the events derived from the GDELT dataset and arranged the outlets in rows and their reported events in columns to form a matrix. Each cell value of the resulting matrix describes the selection/rejection of the event by the outlet.

For the bias analysis, they chose affiliation, ownership, and geographic proximity of the different outlets as the major factors. They suggest that each outlet follows its own latent preferences structure which facilitates the outlet to rank events. They also suggested that events should be selected such that the selected list should be diverse and should include a wide range of actively reported events. They thus adopted the method of Maximum Marginal Relevance which facilitates ranking based on the relevance and diversity of the events. It is discovered that event selection favors the most discussed topics rather than the unique ones.

F. Hamborg et al. [6] uses a matrix similar to the one created by Bourgeois et al.[3] Each cell in the matrix represent the most representative topic of the article reported by one country about the other. By spanning the matrix through outlets and topics in a region, the bias can be examined. They used a collection of 1.6 million articles from more than 100 countries over a two-month span from the Europe Media Monitor (EMM)[1] as their dataset.

Authors in [6] aggregates the related articles and then outsource the task of bias identification to the users, forcing them to determine the bias on their own. While the rest of the existing work analyzes the selection bias, it certainly does not present an automated approach suited to the outlet prediction task, unlike our work.

3 DATA DESCRIPTION

3.1 Raw Data Source

Event Registry[2] [7] monitors, collects, and provides news articles from news outlets around the world. It also aggregates them into clusters that are referred to as events. Each event is then annotated with several metadata such as unique id to track the event coverage, categories to which it may belong, geographical location, sentiment, etc. As a result, its large-scale temporal coverage can be used effectively to study the event selection process of news outlets.

[1]https://ec.europa.eu/knowledge4policy/
[2]https://eventregistry.org

3.2 Dataset

For our experiments, we first selected the top three news outlets based on Alexa Global Rankings [3]. We then used the Event Registry API to collect all news events reported in English between January 2019 and May 2020. We excluded events that were not covered by any of the selected outlets. We ended up with 51, 409 events for which we extracted basic information such as event id, title, summary, and source. Since the event coverage by these outlets is not uniform, which can be visualized in Figure 1, we used a stratified split to mimic this imbalance across the generated train-valid-test sets.

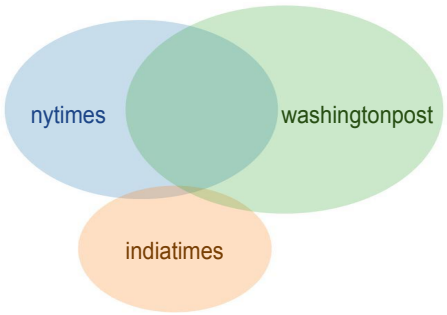

Figure 1: Distribution of event coverage by the outlets.

4 MATERIALS AND METHODS

4.1 Problem Modeling

For an event E and its associated pair (T, S), the task is to generate a list of outlets O expected to cover E. Here T is the event title and S is a short summary of the event as provided by the Event Registry. Mathematically, the task can be formulated as,

$$O = f(T, S, \alpha) \tag{1}$$

where, f is the outlet prediction function and α denotes the model parameters. O can have a well-thought-out variable length response generated from the list unique outlets O^l. For this work, $|O^l| = 3$.

4.2 Methodology

We extract feature vectors from T and S. We fuse them together to create a fused vector which is then passed through several layers to finally generate O. Figure 2 illustrates the entire prediction process. We further outline these tasks with more details in the following subsections.

4.2.1 Feature Extraction and Fusion. We used Google's *Universal Sentence Encoder* [4] *(USE)* to extract 128-dimensional feature vectors T' and S'. For feature fusion, we concatenated T' and S' and applied $tanh$ activation to generate F. We then used batch-normalization to increase the stability of the network and for regularization.

$$F = BN(tanh(T' \oplus S')) \tag{2}$$

In Eq 2, BN and $\oplus$ represents batch-normalization and concatenation respectively.

[3]https://www.alexa.com/topsites/category/Top/News/Newspapers
[4]https://tfhub.dev/google/universal-sentence-encoder/

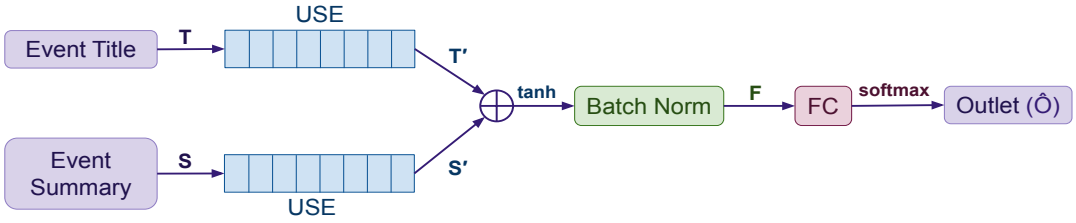

Figure 2: Outlet prediction process.

4.2.2 Outlet Prediction. We solve the problem using a multi-label classification model for which we create a separate outlet-index dictionary for outlets $D = \{o_1 : 1,\ o_2 : 2 \ldots\ o_n : n\}$, where n is the total number of unique outlets in O^l. To predict the list of outlets we pass F to the fully-connected layer (FC) having $softmax$ activation with n output neurons. Since an event can be covered by more than one outlet, we formulate the recursive prediction procedure as,

$$\hat{o} = \mathcal{P}(o_i|F, \hat{o}_{i-1}, b) = softmax(Fw_i + b_i) \tag{3}$$

$$= \frac{e^{Fw_i+b_i}}{\sum_{j=1}^{n} e^{Fw_j+b_j}} \tag{4}$$

where, $\hat{o}$ is the probability of selecting the i^{th} outlet (o_i) given F, bias (b), and the set of probabilities of previously predicted outlets ($\hat{o}_{i-1}$), and w is the weight. We use categorical cross entropy as the loss function as follows:

$$\mathcal{L}(o, \hat{o}) = -\sum_{j=1}^{n}\sum_{i=1}^{x}(o_{ij} * \log(\hat{o}_{ij})) \tag{5}$$

In Eq (5), for i^{th} outlet in the output sequence of length x, o_{ij} and $\hat{o}_{ij}$ denotes the actual and predicted probability of selecting the j^{th} outlet from D.

4.2.3 Hyper-parameters. We used Categorical accuracy[5] as the metrics to calculate the mean accuracy rate for multilabel classification problems across all the predictions. We consider a batch of size 128 and number of epocs as 100 for training. To optimize the weights during training we use Adam optimizer.

5 EXPERIMENTAL EVALUATION

5.1 Baselines

We use the following well-known and simplified methods as our baseline models.

- **Uniform:** Generate predictions randomly using a uniform distribution.
- **Stratified:** Generates predictions by respecting the class distribution of the training set.

5.2 Evaluation Metric

We aim to predict the list of outlets in this work. However, it is not necessary to predict the sequence in which outlets appear on this list. This is explained with an example given in Table 1. In other cases, a combination of correct and incorrect outlets may be predicted by the model.

We used the following metrics to evaluate the effectiveness of our model where, $\hat{o}$ is the predicted outlet, o is the true outlet, and N is the total number of instances.

[5]https://github.com/keras-team/keras/blob/master/keras/metrics.py

Table 1: Multiple correct predictions.

indiatimes nytimes washingtonpost
indiatimes washingtonpost nytimes

- **Subset Accuracy (a):** It measures the percentage of instances in which all of the outlets are correctly classified.

$$\text{Subset Accuracy } (a) = \frac{1}{N}\sum_{i=1}^{N}(\hat{o}_i - o_i) \tag{6}$$

- **Hamming Loss (ℓ):** It measures the fraction of the incorrectly predicted outlet to the total number of outlets. Since it is a loss function, its ideal value is 0.

$$\text{Hamming Loss } (\ell) = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{\hat{o}_i \cap o_i}{\hat{o}_i \cup o_i}\right| \tag{7}$$

5.3 Results and Analysis

Table 2 shows the comparison of our model with the baseline models in terms of subset accuracy and hamming loss.

Table 2: Comparison between the baseline models and our proposed model.

	Subset Accuracy	Hamming Loss
Uniform	0.140	0.526
Stratified	0.286	0.422
Ours	**0.546**	**0.275**

Quantitative analysis of the experimental results shows that, our model outperforms the Uniform and Stratified models by a margin of 0.41 and 0.26 points for subset accuracy and by 0.25 and 0.15 points for hamming loss respectively. The performance difference is clearly visible in Figure 3.

The intersection that we find among the different outlet pairs differs considerably as evident in Figure 1. This can be best seen by assessing the conditional probability of an event covered by an outlet given that it is covered by another outlet as listed in Table 3. For example, we can note that the $P(washington|nytimes) = 0.492$ which is quite high and indicates that *washingtonpost* tends to cover most of the events covered by *nytimes*. It is also interesting to note that *indiatimes* do not follow *washingtonpost* or *nytimes*, and vice versa.

6 CONCLUSIONS AND FUTURE WORK

It is important for a journalist to know which event is worthy enough to be published. Even readers would be interested to know

Table 3: Conditional probability of an event to be covered by an outlet, provided it is covered by another outlet.

P(x|y)	nytimes	indiatimes	washingtonpost
nytimes	1.000	0.067	0.364
indiatimes	0.034	1.000	0.023
washingtonpost	0.492	0.063	1.000

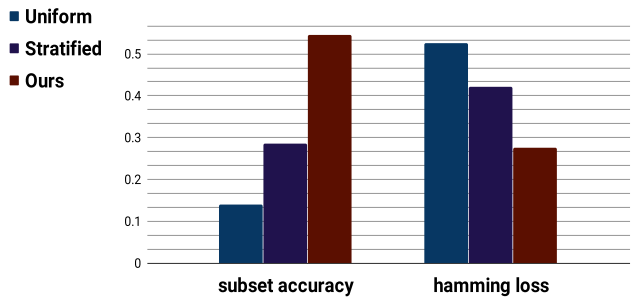

Figure 3: Comparison between the baseline models and our proposed model.

the outlets that are going to cover the event of their interest. Yet it is certainly not an automated approach, therefore in this work, we propose an approach to address the outlet prediction task given the event title and description. We also find that even in its simplest form, our model is capable of predicting the outlet. In the future, we intend to enhance our proposed model to better predict the outlets and to work in a cross-lingual setting. We plan to include a few more metadata provided by Event Registry (refer Section 3.1) along with Wikipedia concepts. We also plan to analyze the speed of reporting, time-span, and importance given to the events by the outlets. In addition, we will also be looking into how the outlets change their coverage style over time.

ACKNOWLEDGMENTS

This work was supported by the Slovenian Research Agency and the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 812997.

REFERENCES

[1] Brent H Baker, Tim Graham, and Steve Kaminsky. 1994. *How to identify, expose & correct liberal media bias.*

[2] Matthew Barnidge, Albert C Gunther, Jinha Kim, Yangsun Hong, Mallory Perryman, Swee Kiat Tay, and Sandra Knisely. 2020. Politically motivated selective exposure and perceived media bias, 82–103.

[3] Dylan Bourgeois, Jérémie Rappaz, and Karl Aberer. 2018. Selection bias in news coverage: learning it, fighting it. In *Companion Proceedings of the The Web Conference 2018*, 535–543.

[4] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems*, 13042–13054.

[5] Zihao Fu. 2019. An introduction of deep learning based word representation applied to natural language processing. In *2019 International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI)*. IEEE, 92–104.

[6] Felix Hamborg, Norman Meuschke, and Bela Gipp. 2018. Bias-aware news analysis using matrix-based news aggregation, 1–19.

[7] Gregor Leban, Blaz Fortuna, Janez Brank, and Marko Grobelnik. 2014. Event registry: learning about world events from news. In *Proceedings of the 23rd International Conference on World Wide Web*, 107–110.

[8] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. Bpr: bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence* (UAI '09). AUAI Press, Montreal, Quebec, Canada, 452–461. ISBN: 9780974903958.

[9] Sebastian Ruder. 2019. *Neural transfer learning for natural language processing*. PhD thesis. NUI Galway.

[10] Sebastian Ruder, Matthew E Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, 15–18.

[11] Diego Saez-Trumper, Carlos Castillo, and Mounia Lalmas. 2013. Social media news communities: gatekeeping, coverage, and statement bias. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, 1679–1684.

[12] Rune J Sørensen. 2019. The impact of state television on voter turnout. *British Journal of Political Science*, 257–278.