A MODEL FEATURES

The GBDT model was trained using 3 types of features - article features, user features, and user-article features. The article features are:

- *article tags:* multi-label encoding with a dimension of 3,000 items;
- *article authors:* multi-label encoding with a dimension of 1,000 items;
- *article section:* the section under which the article was published on the website, represented as a one-hot encoding with a dimension of 1,500 items;
- *is background story:* binary variable, whether or not the article contains a background story, as specified by the authors;
- *is free article:* binary variable, whether or not the article is behind the paywall;
- *contains BNR:* binary variable, whether or not the has a BNR radio program associated with it;
- *published weekday:* day of the week when the article was published on the website, represented as an integer;
- *published day:* day of the month when the article was published on the website, represented as an integer;
- *published month:* month of the year when the article was published on the website, represented as an integer;
- *published time:* time of the day when the article was published, represented as an integer $\in [0, 3]$ (0 for before 6AM, 1 for 6 to 9AM, 2 for 9AM to 6PM, 3 for after 6PM);
- *published timestamp:* UNIX timestamp when the article was published;
- *word embeddings average:* average word embeddings of all the words in the article; we use pre-trained fastText Dutch language vectors[3]with a dimension of 300;
- *word embeddings sum:* sum of word embeddings of all the words in the article;
- *word embeddings variance:* variance of word embeddings average;
- *word embeddings L2:* L2 norm of word embeddings;
- *number of sentences* in the article;
- *average sentence length:* measured in words;
- *variance of average sentence length*;
- *number of words* in the article;
- *average word length:* measured in characters;
- *variance of average word length*;
- *number of paragraphs* in the article;
- *average paragraph length:* measured in words;
- *variance of average paragraph length*;
- *number of characters*;
- *"Also Read" items*: number of articles linked under "Also Read";
- *sentiment*: article polarity and subjectivity scores;
- *Hapax Legomena*: percentage of words that occur only once in the article;
- *Hapax Dislegomena*: percentage of words that occur only twice in the article;
- *complexity:* Fleisch reading ease score of the article;
- *number of images* in the article;
- *number of embedded Tweets* in the article.

The user features are:

- *followed tags:* multi-label encoding with a dimension of 3,000 items;
- *average word embeddings:* averaged across all articles the user has read;
- *word embeddings sum:* summed across all articles the user has read;
- *word embeddings variance:* variance of word embeddings averaged across all articles the user has read;
- *word embeddings L2 norm:* aggregated across all articles the user has read;
- *average sentiment*: averaged across all articles the user has read;
- *average word length:* across all articles the user has read;
- *word length variance:* variance of average word length across all articles the user has read;
- *average words in a read article*;
- *average character length of a read article*;
- *average number of paragraphs in a read article*;
- *average number of images in a read article;*
- *average Hapax Legomena in a read article*;
- *average Hapax Dislegomena in a read article.*

The user-article features are:

- *author overlap:* given an article and a user, the number of article authors that feature in the top 10 of the user's most read authors;
- *tag overlap:* given an article and a user, the number of article tags that feature in the top 10 of the user's most read tags;
- *cosine similarity of average word embeddings:* between the article average word embeddings and the user average word embeddings;
- *cosine similarity of word embeddings sum*;
- *cosine similarity of word embeddings variance*;
- *difference in word embeddings L2 norm:* between the article word embeddings L2 norm and the user word embeddings L2 norm;
- *difference in number of words:* between the article number of words and the user average number of words;

[3] https://fasttext.cc/docs/en/crawl-vectors.html

- *difference in character length:* between the article character length and the user average character length;
- *difference in average word length:* between the article average word length and the user average word length;
- *difference in variance of average word length:* between the article variance of average word length and the user variance of average word length;
- *difference in number of paragraphs:* between the article number of paragraphs and the user average number of paragraphs;
- *difference in number of images:* between the article number of images and the user average number of images;
- *difference in Hapax Legomena:* between the article Hapax Legomena and the user average Hapax Legomena;
- *difference in Hapax Dislegomena:* between the article Hapax Dislegomena and the user average Hapax Dislegomena.

B MODEL HYPERPARAMETERS

The Gradient Boosted Decision Trees (GBDT) model employed throughout the paper uses the following hyperparameters:

- learning rate: 0.01;
- number of gradient-boosted trees: 1500;
- maximum tree depth for base learners: 15;
- minimum loss reduction required to make a further partition on a leaf node of the tree: 0;
- subsample ratio of the training instance: 0.7;
- subsample ratio of columns when constructing each tree: 0.8;
- learning objective: logistic regression.