

Predicting Headline Effectiveness in Online News Media using Transfer Learning with BERT

Jaakko Tervonen¹ ^a, Tuomas Sormunen¹ ^b, Arttu Lämsä¹,
Johannes Peltola¹, Heidi Kananen² and Sari Järvinen¹

¹*VTT Technical Research Centre of Finland, Kaitoväylä 1, Oulu, Finland*

²*Kaleva Media, Solistinkatu 4, Oulu, Finland*

Keywords: BERT, Headline Effectiveness, Journalism, Machine Learning, Natural Language Processing.


Abstract: The decision to read an article in online news media or social networks is often based on the headline, and thus writing effective headlines is an important but difficult task for the journalists and content creators. Even defining an effective headline is a challenge, since the objective is to avoid click-bait headlines and be sure that the article contents fulfill the expectations set by the headline. Once defined and measured, headline effectiveness can be used for content filtering or recommending articles with effective headlines. In this paper, a metric based on received clicks and reading time is proposed to classify news media content into four classes describing headline effectiveness. A deep neural network model using the Bidirectional Encoder Representations from Transformers (BERT) is employed to classify the headlines into the four classes, and its performance is compared to that of journalists. The proposed model achieves an accuracy of 59% on the four-class classification, and 72-78% on corresponding binary classification tasks. The model outperforms the journalists being almost twice as accurate on a random sample of headlines.


1 INTRODUCTION

During the last years, the ways to consume news articles have changed notably. The interaction between the readers and news media has moved to online channels, such as news portals and social networks. Due to this change, the headlines have a more important role in the news media. In printed news, the headline was supposed to briefly deliver information on the content of the news article. In online media, the goal of the headline is to attract the reader to the article page. As the news media is actively trying to engage the readers to their portals and widen the customer base willing to pay for the news service, the headline should not only allure the reader with false promises but also provide information on the actual content of the article.

Journalists are responsible for writing headlines but it is a difficult task to tell whether a certain headline is interesting to the readers while including correct information on article contents in the headline. Current practices are based on "trial and error" type of

approach, where the headline impact is monitored after publication and changes on the headline are made when considered necessary (Tandoc, 2014). The estimation of headline impact is done based on web analytics data, which is commonly used by the editors and journalists for evaluating the performance of the news sites and specific articles (Tandoc, 2015; Hanusch, 2017). Previous research studies concerning the effectiveness of a news headline have utilised simple univariate metrics such as click-through rate (Kuiken et al., 2017; Lai and Farbrot, 2014; Tenenboim and Cohen, 2015) or shares on a social media site (Szymanski et al., 2017), and for the news article itself, additionally, comments (Tenenboim and Cohen, 2015) and likes/recommendations (Sotirakou et al., 2018) have been used to gauge the impact. However, using these metrics arguably prevents capturing the exact behaviour of the news consumer in an article of a news portal. To account for this, some studies have implemented the use of viewport time (Lagun and Lalmas, 2016), i.e. what part of the article is seen on the screen at each moment in time, as a means to model reader behaviour. In addition to the viewport time, read speed and length as well as scroll

^a  <https://orcid.org/0000-0003-2236-0253>

^b  <https://orcid.org/0000-0001-7789-5867>

intervals have been used as a metric of headline impact (Lu et al., 2018).

Data analytics solutions are currently evolving from providing metrics and dashboard visualizations towards decision-making support tools able to provide actionable insight to their users. In journalism, this means for example automated content creation (Carlson, 2015) or tools supporting editors and journalists in their daily decisions (Petre, 2018). From consumer side, estimating and predicting headline effectiveness would help to filter content, or to get recommendations on articles with effective headlines.

In the present study, we establish a larger framework on how to define an effective headline to be used as a guideline for journalists and as a metric for machine learning prediction on online news popularity. We analyze a click-stream dataset from an online news media portal, and present a deep learning model to predict headline effectiveness. Further, we evaluate the practical value of the presented model by assessing whether it is useful for journalists. The performance is compared to expert evaluators scores, and we show that the proposed model outperforms the experts by a large margin.

2 RELATED WORK

Current machine learning solutions for predicting headline effectiveness can be categorized as considering pre- or post-publication prediction. As the former is more useful for practical usage, the focus in this study is on prediction before publishing the article.

Considering the previous machine learning solutions to pre-publication prediction, (Bandari et al., 2012) used regression models to predict whether the article received a low, medium, or high number of tweets. They found highest accuracy of 84% with a bagging method. However, they used articles from several news sites and they reported that the news source was the most important feature in the model. As the distribution of popularity across the different news sources varied, this suggests that the model actually learnt to distinguish popular news sites from mid-to unpopular ones. (Fernandes et al., 2015) predicted the number of Twitter shares. They extracted features related to both the headline and the article, and its publication time, and predicted whether the article received more or less than median amount of shares. The highest prediction performance was found with Random Forest, with accuracy of 67% and area under curve of 0.73. (Liu et al., 2017) considered publication time, author and news section as well as the grammatical construction of the headline and the arti-

cle to predict whether the article was popular or not. Popularity was based on number of clicks but it was unclear how the division between popular and unpopular articles was made. They found the highest area under curve of 0.825 with alternating decision tree.

Contrary to the aforementioned studies, (Lamprinidis et al., 2018) considered only features related to the headline. They predicted whether the article received more or less than a median amount of clicks. They compared two models: a baseline logistic regression trained on sequences of n characters and the TF-IDF scores of headline uni- and bigrams to a multi-task recurrent neural network trained on headline word embeddings (i.e. real number vector representations of the headline). They used part-of-speech tagging and news section prediction as auxiliary tasks for the recurrent network. Although the auxiliary tasks improved the prediction scores of the neural network, the network still did not perform better than logistic regression, both having highest accuracy of 67%. However, they did not consider using the extracted features as additional input to the neural network. The pretrained word embeddings were based on corpus consisting of the Danish Wikipedia and not news articles, and they did not comment on which model was used to train the word embeddings.

Recently, the Bidirectional Encoder Representations from Transformers (BERT) language model for extracting word embeddings has been demonstrated to achieve state-of-the-art performance in several natural language processing tasks (Devlin et al., 2019). BERT was used in a recent study where the quality of news headlines was defined in terms of number of clicks and dwell time, i.e. time spent on article page (Omidvar et al., 2020). They used a deep neural network to extract features from both the headline and the body text of the article, and predicted the probability of belonging to one of the four defined classes with a mean absolute error of 0.034.

To summarize, previous studies mainly used either the number of clicks or shares to define article popularity, not headline effectiveness per se, and they used a variety of features related to the article or the headline. Furthermore, previous studies lack practical validity since they are evaluated only in terms of numerical prediction accuracy but their actual usefulness as a tool for journalists is not considered.

In this study, we define headline effectiveness as two-dimensional through number of clicks and reading time. We use BERT word embeddings, compare the performance of multi-language BERT and BERT trained specifically for the Finnish language, and use both together with manually extracted features from the headline to predict its effectiveness.

Finally, we conduct an experiment with journalists to assess whether the model or the journalists can predict effectiveness better.

3 MATERIALS & METHODS

3.1 Estimating Headline Effectiveness

To describe headline effectiveness without resorting to simple univariate measures, it was defined in terms of click-through rate and time spent reading the article, which were considered to present the popularity of and engagement on the article. The effectiveness prediction task was formulated as a classification problem, since it is more prevalent in earlier studies than regression, and since it allows for more straightforward evaluation with the journalists. Thus, after obtaining the two values for each article, both variables were split at their median, giving rise to four classes (see Figure 1):

1. non-effective, few clicks and short time spent reading;
2. appealing, many clicks and short time spent reading;
3. engaging, few clicks and long time spent reading;
4. effective, many clicks and long time spent reading.

Since the correct class is rather random for articles whose click-through rate or reading time is close to the median, 5% of articles from both sides of the median for both dimensions were left out of the analysis.

As the studied dataset (see section 3.2) contained click-stream data, both metrics had to be calculated. Click-through rate was taken as the sum of clicks for each article. Reading time was estimated as the read percentage, i.e. time spent on the article page relative to the length of the article. It was estimated through the following procedure: 1) Calculate the word count for each article; 2) Isolate single users through user IDs, unique to each session; 3) Sort the clicks according to timestamp in ascending order; 4) Calculate the time between two consecutive clicks (except for the last click); 5) Evaluate the percentage of the article that has been read by the user. For step 5), literature values for the average read speed for Finnish language were considered; the value in an experiment with standardized texts approximated the mean read speed to be 161 words with standard deviation of 18 words per minute (Hahn, 2006). The minimum amount of time required to read a specific article was

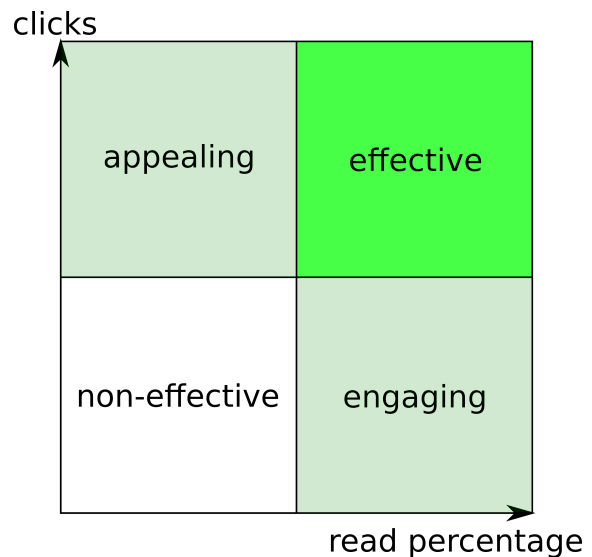


Figure 1: Division of headlines into non-effective, appealing, engaging, and effective. Each border denotes median value, and 5% of headlines were left out from both sides of the median.

estimated to be the article word count divided by the mean read speed value subtracted by two standard deviations (i.e. 125 words per minute). Users exceeding this time were evaluated to have read 100% of the article, whereas the read percentages of the users below this time value were estimated linearly from 0% up to 100%. Using these steps, the mean read percentage was obtained for each article.

3.2 Dataset

The used dataset contained click-stream data from a Finnish newspaper’s online portal, obtained between December 2018 and May 2019. Each click constituted of timestamp, the properties of the clicked article (its ID, URL, headline, section, publish time, and access policy), and an anonymous user ID.

In addition to news articles, the newspaper publishes other content like comics and photo galleries. The newspaper may also modify the contents of some articles after initial publication (e.g. updating the piece of news with additional information) without modifying the headline or article ID. As the read time estimate of the updated news would not be comparable (the same title had several different read time estimates based on different contents), duplicate headlines were removed. Because the main interest in the present work was to predict the headline effectiveness of news articles, all clicks targeting other content than news articles were removed, together with clicks targeting articles that were published prior to the defined data collection period.

After these restrictions, the dataset contained approximately 17 million clicks and 7198 articles, 6229 of them free and 969 subscription-only.

3.3 Feature Extraction

BERT provides contextual, bidirectional representations of words (Devlin et al., 2019). Effectively, the pretrained representation of a word is a real-valued feature vector. Two pretrained BERT models were considered, the cased multilanguage model by Google, multiBERT (Devlin et al., 2019), and the cased model specifically trained for Finnish language, FinBERT (Virtanen et al., 2019). FinBERT was trained with news articles and other material found online which can be considered to be linguistically similar to the dataset used in this study (see section 3.2). (Virtanen et al., 2019) showed that FinBERT outperforms multilingual versions of BERT in classification tasks of news articles and texts from discussion forum.

To complement the feature representation provided by BERT, features were extracted manually from the headline. These features contained the length of the headline (number of words and characters in the original and lemmatized headline, and number of sentences as provided by Natural Language Toolkit (Loper and Bird, 2002) and as separated by punctuation, mean length of words in the headline, punctuation (number of colons, semicolons, commas, dots, dashes, exclamation and question marks), whether the headline contained a quotation, whether the headline mentioned the name of the newspaper’s home city, and whether the headline started with a single string followed by a colon (e.g. ”Analysis:”). Named entities were recognized with DeepPavlov (Burtsev et al., 2018) and word classes were extracted with Turku Neural Parser Pipeline (Kanerva et al., 2018). Additionally, the access policy (free or subscription-only) was used as a metadata feature.

3.4 Model

The developed model is presented in Figure 2. The inputs of the model can be divided into three categories: the headline text, calculated features described in section 3.3, and metadata (access policy). The headline text acts as an input for the BERT model which then provides a transformed presentation of the text. The transformed headline presentation, along with the calculated features and the metadata features, are used as an input for the last layers in the neural network model. One layer combines the inputs and it is fol-

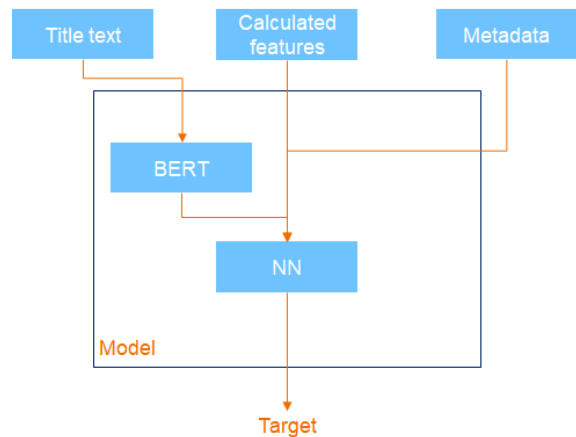


Figure 2: Structure of the developed model.

lowed by the output layer making the actual classification to the classes described in section 3.1. Hyperbolic tangent was used as an activation function in the hidden layer and the number of neurons used was set to 256. The model was implemented using Keras-BERT (HG, 2020).

4 EXPERIMENTS

4.1 Evaluation of the Proposed Model

Four prediction tasks were defined: binary classification on 1) click-through rate; 2) read percentage; 3) effective headline vs. the other three classes; and 4) four-class classification between all the groups.

For each task, multiBERT and FinBERT were used to extract features from the headline. It was expected that FinBERT provides features that perform better than the ones extracted with multiBERT. However, the multilanguage version was used for the sake of comparison and because the model might be of more interest to a wider audience than FinBERT.

Because the article’s access policy is visible to the reader before clicking the article, it likely affects the behavior of readers without subscription and therefore the access policy was included as a feature a priori. The model was first fitted with only the BERT feature representation and access policy as input features. To evaluate the significance of manually extracted features complementing the features calculated with BERT, the model for each task was then trained with BERT features, manually extracted features, and access policy as input features.

Since the access policy likely affects the behaviour of readers, the models were also trained separately using only the free articles. Similar inspection

Table 1: Results with free and subscription-only articles.

BERT Title features	click-through rate				read percentage				multiclass				effective vs. rest			
	acc	rec	prec	F1	acc	rec	prec	F1	acc	rec	prec	F1	acc	rec	prec	F1
Fin No	76.2	76.4	76.7	76.6	71.8	74.1	70.5	72.2	58.7	58.7	56.6	56.4	78.0	52.3	72.8	60.9
	77.0	77.4	77.4	77.4	71.8	73.6	70.7	72.1	57.5	57.5	56.1	56.4	77.3	51.2	71.3	59.6
Multi No	66.9	60.7	70.3	65.1	68.4	69.7	67.5	68.6	49.5	49.5	45.1	42.0	68.6	5.7	80.0	10.6
	66.7	66.5	67.6	67.0	67.0	66.1	66.9	66.5	50.2	50.2	48.5	46.6	69.1	14.9	61.8	24.1

Abbreviations: acc = accuracy, F1 = F1-score, Fin = FinBERT, Multi = MultiBERT, rec = recall, prec = precision

Table 2: Results using only the free articles.

BERT Title features	click-through rate				read percentage				multiclass				effective vs. rest			
	acc	rec	prec	F1	acc	rec	prec	F1	acc	rec	prec	F1	acc	rec	prec	F1
Fin No	77.0	78.2	75.6	76.9	72.1	69.3	72.4	70.8	56.9	56.9	56.2	52.9	79.7	54.5	74.4	62.9
	76.7	77.5	75.6	76.5	73.5	70.2	74.2	72.2	56.5	56.5	55.6	53.8	79.1	51.9	74.4	61.2
Multi No	66.2	60.5	67.2	63.7	66.6	69.0	64.9	66.9	47.1	47.1	30.9	37.2	69.6	5.1	80.0	9.6
	66.0	63.5	65.8	64.6	67.2	63.2	67.6	65.3	48.9	48.9	51.4	41.1	69.6	5.5	76.5	10.3

Abbreviations: acc = accuracy, F1 = F1-score, Fin = FinBERT, Multi = MultiBERT, rec = recall, prec = precision

was not done for the subscription-only articles since there were not enough of them for the model to provide comparable results.

Similarly to (Lamprinidis et al., 2018), the headlines were split into training, validation, and testing data using 70% for training the model and 15% for both validation and testing.

4.2 Evaluation with Journalists

As seen in section 2, earlier studies have evaluated the model only in terms of prediction accuracy. However, no matter how accurate the model is, it is not useful if experts are more accurate. Therefore, the practical value of the model was assessed by comparing its performance with that of journalists.

A survey was conducted to gather data on how five experts would place different headlines into the four classes defined. The experts were journalists working for the same news media from which the used data and headlines originated, thus having similar background knowledge on the headlines as was used to train the model. Eighty headlines from four different news sections (homeland, local news, sports, and economy, twenty from each section) were randomly sampled from the data. The sample was stratified so that the class distribution in the sample was the same as in the whole data. All the headlines selected were from free-to-view articles to make sure that access policy does not bias the experts' evaluations and that they focus only on the headline itself. These headlines were presented to the experts to place them in the four different classes, and the model was trained without using these headlines. The model used in this evaluation used FinBERT and the manually extracted features, and since all the headlines in the random sample were from free-to-view articles, the model was also

trained using only the free articles.

To get a more thorough view on how the proposed model performs on such a small, random subset, the experiment was repeated for ten thousand similar random samples (i.e. bootstrap samples). The model's performance was compared to random guessing and expert evaluators' scores.

5 RESULTS & DISCUSSION

5.1 Classification Results

The prediction results for each prediction task using all articles and using only the free articles, are reported in Tables 1 and 2, respectively. The baseline accuracy obtained by random guessing was 25% on the multiclass prediction, and 50% in binary prediction. The headline effectiveness could be predicted with up to 77% accuracy in terms of click-through rate, and up to 58.7% accuracy when using the multiclass metric. The performance when using only free articles, or using also the subscription-only articles, was similar, so the two types of articles can be used together in studying headline effectiveness.

As expected, FinBERT performed better than the multilanguage BERT in each of the prediction tasks, providing up to around ten percentage points higher accuracy in binary tasks, and up to around eight percentage points higher accuracy in the multiclass prediction. This is further evidence to complement the results reported in (Virtanen et al., 2019) that language-specific BERT model outperforms the multi-language model. Whereas the multilanguage model may serve as a baseline, language-specific models trained on a large corpus can detect more nu-

anced information and extract more useful features from the text, which leads to better performance in prediction tasks. Moreover, the multilanguage BERT performed especially poorly for the task of predicting an effective headline vs. rest of the classes in the sense that recall (and thus the F1-score) were notably lower than with FinBERT: the model classified nearly all articles to the ineffective (majority) class.

Using manual features to complement the features calculated with BERT did not make much difference. Regardless of the prediction task and performance metric, model performance was always within one or two percentage points from one another, for better or worse performance. Thus, it seems that the information provided by the manually extracted features is implicitly included in BERT’s feature representation. Indeed, manually extracted features consisted of features related to headline length, wording, punctuation, and named entities. Since all these elements are contained in the BERT input, BERT is able to convert all these aspects into its feature representation.

The necessity of BERT was evaluated in an ablation study, using just the manually extracted headline features as input to the neural network classifier without BERT output. The results for this experiment are shown in Table 3. In general, excluding BERT led to approximately 10-15 percentage point decrease in prediction performance, depending on the metric and prediction task. The most notable difference is in the effective vs. rest classification task, where just 2.1% of effective headlines were correctly predicted. Class imbalance may have affected this result (approximately one third of headlines were in the effective class) but when FinBERT was used, over 50% of effective headlines were correctly predicted with a higher accuracy despite the fact that it had exactly the same data splitting. Thus, BERT is a necessary component in the model.

Table 3: Prediction results without utilizing BERT.

	acc	rec	prec	F1
click-through rate	61.7	68.1	61.2	64.4
read percentage	62.8	66.5	61.5	63.9
multiclass	44.4	44.4	43.9	41.9
effective vs. rest	66.3	2.1	30.0	4.0

Abbreviations: acc = accuracy, F1 = F1-score, rec = recall, prec = precision

5.2 Journalist Evaluation Results

Results from the bootstrap simulations are displayed in Figure 3. The figure also visualizes the mean accuracies of both the expert evaluators and the neural network model in the random sampled distribution.

Across the ten thousand bootstrap samples, the proposed model’s accuracy ranged from 37.5% to 75% with a mean of 56%, and random guessing was significantly less accurate (range 8.8% – 43.8%, mean 25%). In the whole dataset, the testing accuracy with these model settings (FinBERT using manually extracted features and only free articles) was 56.5% which is similar to the average accuracy in the bootstrap samples. In the random sample that was presented to the experts, however, the proposed model scored an accuracy of 49.4% which is admittedly lower than on average. Since it is still only slightly more than one standard deviation from the mean, the lower performance is explained by the random selection of the evaluation headlines.

Even though the proposed model performed more poorly on this sample, it was still more accurate than the expert evaluators. The experts placed each headline in the correct class with an average accuracy of 26.1% (range 24.1% - 29.1%). The expert views on the headline effectiveness also varied greatly between different persons. The experts did not fully agree on the effectiveness of any headline and only in 3.8% of the cases four out of the five experts predicted the headline effectiveness correctly.

The classification of the expert evaluations in relation to model predictions is presented in Table 4. The scores presented in the table indicate that all experts were never correct for the headlines whose effectiveness the model predicted correctly or incorrectly. Further, for all the headlines whose effectiveness the model predicted correctly, at least one expert was correct only 42.5% of the time, but when the model was incorrect, at least one expert was correct on 22.5% of the headlines. Finally, all experts were incorrect for approximately 19% and 16% of the headlines that model was correct and incorrect about, respectively.

The results of the expert survey indicate that the proposed model is able to predict the headline effectiveness more accurately than the journalists that write the headlines. Based on these results, the model seems to be more capable of analytically processing large amounts of measured data leading to more accurate predictions of headline effectiveness compared to journalists. The experts estimate the effectiveness using their expert instinct and previous personal experiences, which might be the underlying reason for large variation in expert evaluations. Currently, if and when the headline effectiveness is measured, the measurement is based solely on the number of clicks.

Table 4: Comparison of model predictions in relation to expert predictions.

model vs. expert	all experts correct	one or more experts correct	all experts incorrect
model correct	0.0%	42.5%	18.8%
model incorrect	0.0%	22.5%	16.3%

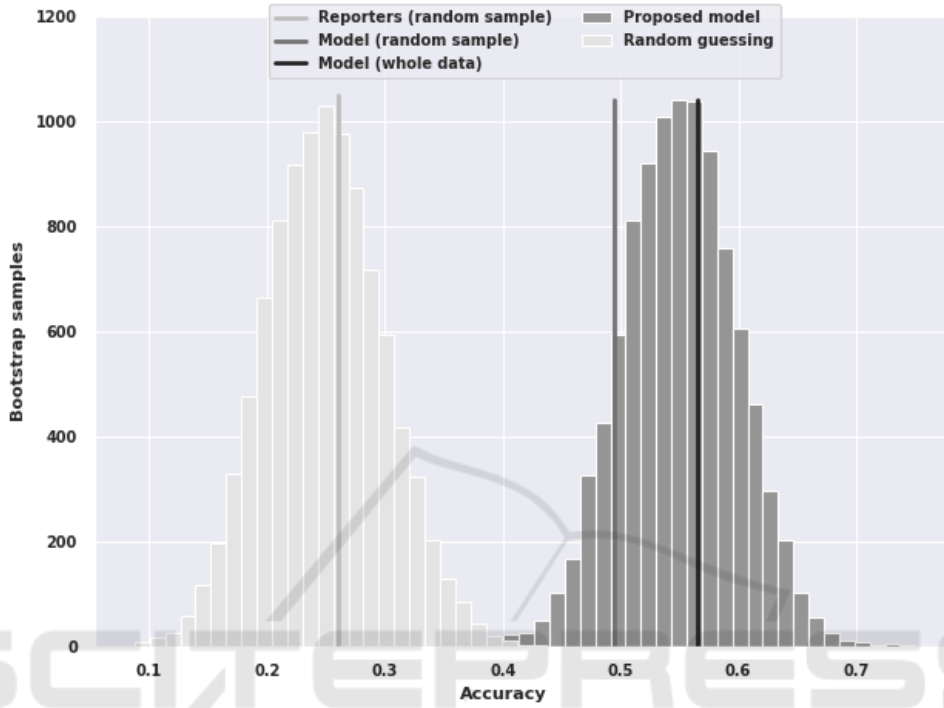


Figure 3: The accuracies obtained on bootstrap samples similar to the one used in the expert evaluation. The histogram on the left is the distribution of accuracy obtained with random guessing method and the one on the right with the proposed model for each of the samples. The leftmost vertical line is the expert accuracy and the middle vertical line is the model’s accuracy on the same sample as evaluated by the experts. The rightmost vertical line is the model’s testing accuracy on the whole dataset.

5.3 Comparison of Results with Previous Works

The obtained model performance is also comparable to or exceeds the performance reported in previous studies: (Bandari et al., 2012) found an accuracy of 84% on three-class classification to low, medium, or high number of tweets, (Fernandes et al., 2015) got an accuracy of 67% on binary classification of number of shares on Twitter, and (Lamprinidis et al., 2018) reported the highest accuracy of 67% on binary classification of number of clicks. For comparison purposes, the logistic regression model that performed the best in (Lamprinidis et al., 2018) was trained to perform the same classification tasks as BERT-based approaches. The title texts were first transformed into numerical format with TF-IDF by using 2-6 character n-grams and then the classifier was trained. The classification results are presented in Table 5. In almost all of the measured metrics the logistic regression

based approach is not able reach the performance of FinBERT but outperforms multilanguage BERT.

5.4 Limitations and Future Work

This study presents and evaluates the first proof-of-concept version implementation of a tool for supporting news editors and journalists in their work. The tool predicts the effectiveness of the headline more accurately than journalists but there are numerous pos-

Table 5: Prediction results with a logistic regression model used in (Lamprinidis et al., 2018).

	acc	rec	prec	F1
click-through rate	72.4	72.4	72.4	72.4
read percentage	66.3	66.3	66.4	66.3
multiclass	53.4	53.4	51.5	48.2
effective vs. rest	73.9	73.9	75.2	69.3

Abbreviations: acc = accuracy, F1 = F1-score, rec = recall, prec = precision

sibilities to improve both the accuracy and usability of such a tool.

The current implementation uses just the article headline and its access policy as inputs, as opposed to several existing studies. This selection was done since the headline is the medium the journalists use to convey the topic of the article to the readers and to allure them to read the article. However, the headline should not make false promises and article text should fulfill the expectations set by the headline. Thus, writing the headline is a difficult task and the goal was to build a tool to help predict whether the headline itself is effective or not. The prediction could be improved with additional inputs, such as the body text or pictures.

The current implementation relies on relatively simple measures of article popularity and engagement. Popularity was measured with the number of clicks received, and engagement with reading time, relative to the article length and average reading speed. Since the user behavior in online news media varies and some people may simply browse the headlines on the front page, skim through the article, read only the introduction or view the pictures and captions, measuring popularity and engagement could be improved with more advanced web analytics functionalities in the future. Information on article presence and location on the landing page of the news portal, scrolling patterns and interaction with the article could be used as inputs, or they could be utilized to determine popularity and engagement more accurately.

If taken to use in news desk, the tool should be integrated into the news editing workflow. It could also provide added functionalities such as hints on how to improve the headline or even suggest headlines based on the article content. The feasibility of our theory on how to measure headline effectiveness should also be evaluated, i.e. whether or not the effective headlines actually lead to an increase in reader engagement or number of subscriptions.

6 CONCLUSIONS

This work proposed a metric based on click-through rate and read percentage to estimate headline effectiveness in online news media, a model using BERT word embeddings to predict the effectiveness of the given headline under the new metric, and a comparison of the model's performance against expert evaluators. We also carried out a simulation procedure to estimate the model's performance for small random samples. The results indicated that a BERT model specifically trained for Finnish language out-

performed a multilanguage BERT model in predicting headline effectiveness, and that manually extracted features from the headline could not improve the performance. It was found that the model performed significantly better than the experts in evaluating the headline effectiveness in a four-class classification task. However, more extensive data sources describing user behavior on the news site might help in providing more accurate predictions, and integrating the prediction functionality into a larger set of AI-driven tools would provide support for news journalists in their day-to-day work.

CODE AVAILABILITY

The model implementation with example data is available at <https://github.com/vttresearch/otsikkokone>.

ACKNOWLEDGEMENTS

This work was financially supported by Media Industry Research Foundation of Finland and VTT. The authors would like to thank Kaleva Media for sharing their data for model development and the anonymous journalists who took part in the expert evaluation.

REFERENCES

- Bandari, R., Asur, S., and Huberman, B. A. (2012). The Pulse of News in Social Media: Forecasting Popularity. *Sixth International AAAI Conference on Weblogs and Social Media*.
- Burtsev, M., Seliverstov, A., Airapetyan, R., Arkhipov, M., Baymurzina, D., Bushkov, N., Gureenkova, O., Khakhulin, T., Kuratov, Y., Kuznetsov, D., Litinsky, A., Logacheva, V., Lymar, A., Malykh, V., Petrov, M., Polulyakh, V., Pugachev, L., Sorokin, A., Vikhрева, M., and Zaynutdinov, M. (2018). DeepPavlov: Open-Source Library for Dialogue Systems. In *Proceedings of ACL 2018, System Demonstrations*, pages 122–127, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Carlson, M. (2015). The Robotic Reporter. *Digital Journalism*, 3(3):416–431.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186,

- Stroudsburg, PA, USA. Association for Computational Linguistics.
- Fernandes, K., Vinagre, P., and Cortez, P. (2015). A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News. In Pereira, F., Machado, P., Costa, E., and Cardoso, A., editors, *Lecture Notes in Computer Science (including sub-series Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9273 of *Lecture Notes in Computer Science*, pages 535–546. Springer International Publishing, Cham.
- Hahn, G. A. (2006). New standardised texts for assessing reading performance in four European languages. *British Journal of Ophthalmology*, 90(4):480–484.
- Hanusch, F. (2017). Web analytics and the functional differentiation of journalism cultures: individual, organizational and platform-specific influences on newswork. *Information, Communication & Society*, 20(10):1571–1586.
- HG, Z. (2020). keras-bert. <https://github.com/CyberZHG/keras-bert>. [Online; accessed 19-April-2021].
- Kanerva, J., Ginter, F., Miekka, N., Leino, A., and Salakoski, T. (2018). Turku Neural Parser Pipeline: An End-to-End System for the CoNLL 2018 Shared Task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 133–142, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Kuiken, J., Schuth, A., Spitters, M., and Marx, M. (2017). Effective Headlines of Newspaper Articles in a Digital Environment. *Digital Journalism*, 5(10):1300–1314.
- Lagun, D. and Lalmas, M. (2016). Understanding User Attention and Engagement in Online News Reading. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pages 113–122, New York, NY, USA. ACM.
- Lai, L. and Farbroth, A. (2014). What makes you click? The effect of question headlines on readership in computer-mediated communication. *Social Influence*, 9(4):289–299.
- Lamprinidis, S., Hardt, D., and Hovy, D. (2018). Predicting News Headline Popularity with Syntactic and Semantic Knowledge Using Multi-Task Learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 659–664, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Liu, C., Wang, W., Zhang, Y., Dong, Y., He, F., and Wu, C. (2017). Predicting the Popularity of Online News Based on Multivariate Analysis. In *2017 IEEE International Conference on Computer and Information Technology (CIT)*, pages 9–15. IEEE.
- Loper, E. and Bird, S. (2002). NLTK: The Natural Language Toolkit. In *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics*, volume 1, pages 63–70, Morristown, NJ, USA. Association for Computational Linguistics.
- Lu, H., Zhang, M., and Ma, S. (2018). Between Clicks and Satisfaction: Study on Multi-Phase User Preferences and Satisfaction for Online News Reading. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 435–444. ACM.
- Omidvar, A., Pourmodheji, H., An, A., and Edall, G. (2020). Learning to Determine the Quality of News Headlines. In *Proceedings of the 12th International Conference on Agents and Artificial Intelligence*, pages 401–409. SCITEPRESS - Science and Technology Publications.
- Petre, C. (2018). Engineering Consent: How the Design and Marketing of Newsroom Analytics Tools Rationalize Journalists’ Labor. *Digital Journalism*, 6(4):509–527.
- Sotirakou, C., Germanakos, P., Holzinger, A., and Mourlas, C. (2018). Feedback Matters! Predicting the Appreciation of Online Articles A Data-Driven Approach. In *Machine Learning and Knowledge Extraction, CD-MAKE 2018*, volume 11015, pages 147–159. Springer International Publishing.
- Szymanski, T., Orellana-Rodriguez, C., and Keane, M. T. (2017). Helping News Editors Write Better Headlines: A Recommender to Improve the Keyword Contents & Shareability of News Headlines. *arXiv preprint arXiv:1705.09656*.
- Tandoc, E. C. (2014). Journalism is twerking? How web analytics is changing the process of gatekeeping. *New Media & Society*, 16(4):559–575.
- Tandoc, E. C. (2015). Why Web Analytics Click. *Journalism Studies*, 16(6):782–799.
- Tenenboim, O. and Cohen, A. A. (2015). What prompts users to click and comment: A longitudinal study of online news. *Journalism: Theory, Practice & Criticism*, 16(2):198–217.
- Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., Ginter, F., and Pyysalo, S. (2019). Multilingual is not enough: BERT for Finnish. *arXiv preprint arXiv:1912.07076*.