OTSIKKOKONE
Headline Machine

Experimenting with natural language processing methods on Finnish online news media content

Sari Järvinen, Arttu Lämsä, Johannes Peltola, Tuomas Sormunen, Jaakko Tervonen

24/03/2020 VTT – beyond the obvious

Outline

- Introduction to Otsikkokone – Headline machine
- Metrics for measuring headline impact in online news media
- Theoretical model for Otsikkokone
- Experimental results
- Future potential
- Demonstration

OTSIKKOKONE – Headline machine

- OTSIKKOKONE project is funded by ”Media-alan tutkimussäätiö”
  - Aiming at improving the competitiveness of printed and electronic media
- In fall 2018 the research foundation organized a thematic call for projects:
  - New potential offered by personalisation and predictive analytics for content production and media business
- Our proposal: Tekoälyllä tehokkuutta otsikointiin – Effectiveness to headline creation by AI
  - 1.3.2019-29.2.2020
  - Co-operation with Kaleva

OTSIKKOKONE – Headline machine objectives

TARGET

- Develop a data-intensive solution for predicting news headline impact in content creation phase for different media channels

RESULT

- Build an intelligent tool for the news journalist to create an effective – not a clickbait - headline

News media transition from printed to digital

- The way we consume news media has changed notably
  - The interaction between the readers and media is moving online
  - The ways of working in news media have changed
- The role of headlines has changed and become more important
  - In printed news, the headline was supposed to deliver information on the content of the news article
  - In online media, the goal of the headline is to allure the reader to the article page
- Media business is evolving, and news media has to engage the readers to their portals and have them pay for the news service
- The competition on reader attention is not limited towards local news services anymore. For example in addition to news content, two million blog posts, 294 billion emails and 400 million tweets are published or sent daily globally.

Interviews of four journalists: what is an effective news headline?

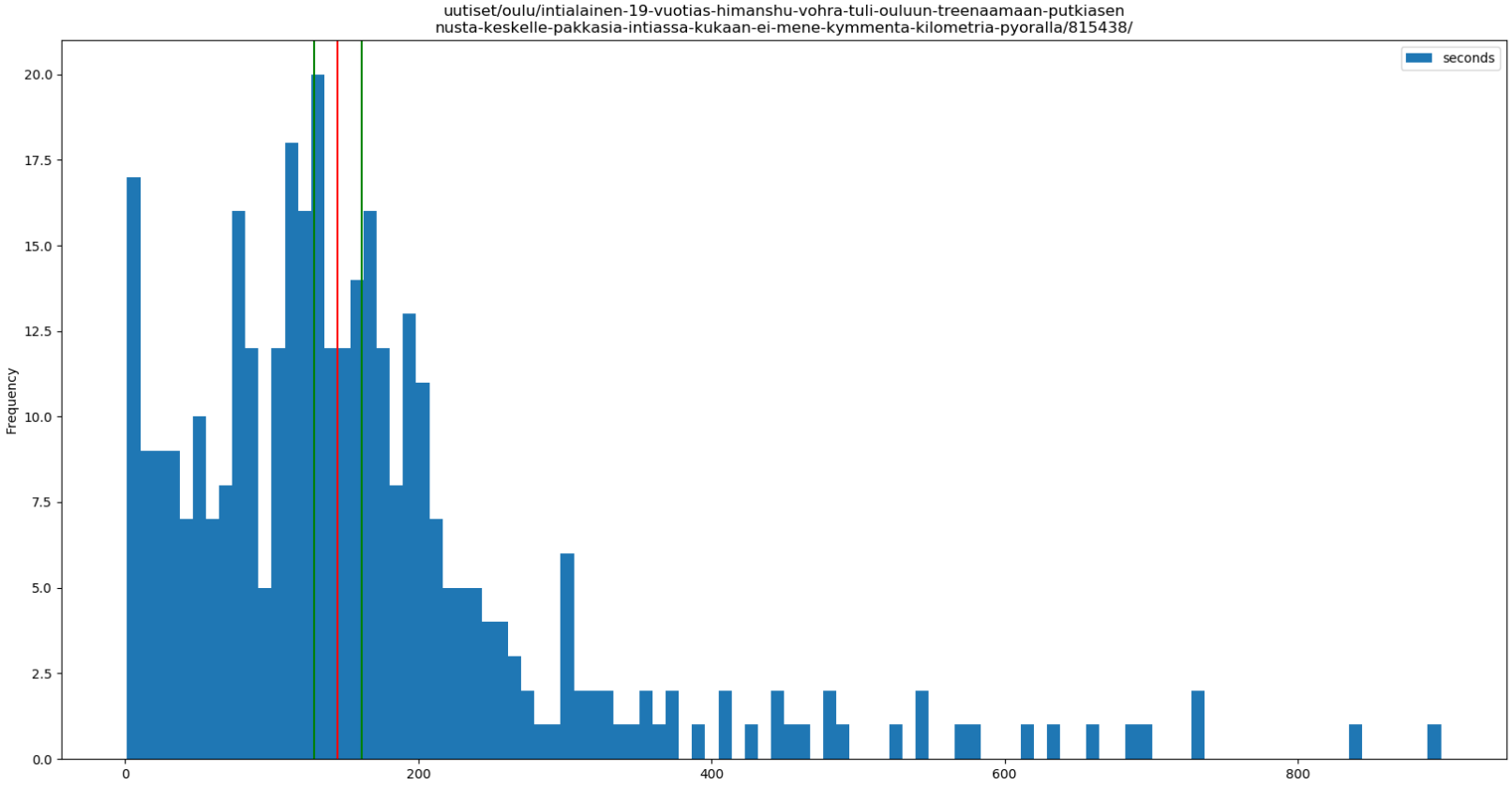
- All representatives of local news providers:
  - Targeting at engaging readers, who spend time on the site, return often and pay for the service
- Printed headline:
  - Visual image important, limited space for the headline
- Online headline:
  - Must allure readers to click and engage with the news service
  - Not so restricted as in printed media
  - Longer headlines, more informative
  - News media is avoiding “mystery”, clickbait headlines
- Journalist education does not train for online content creation

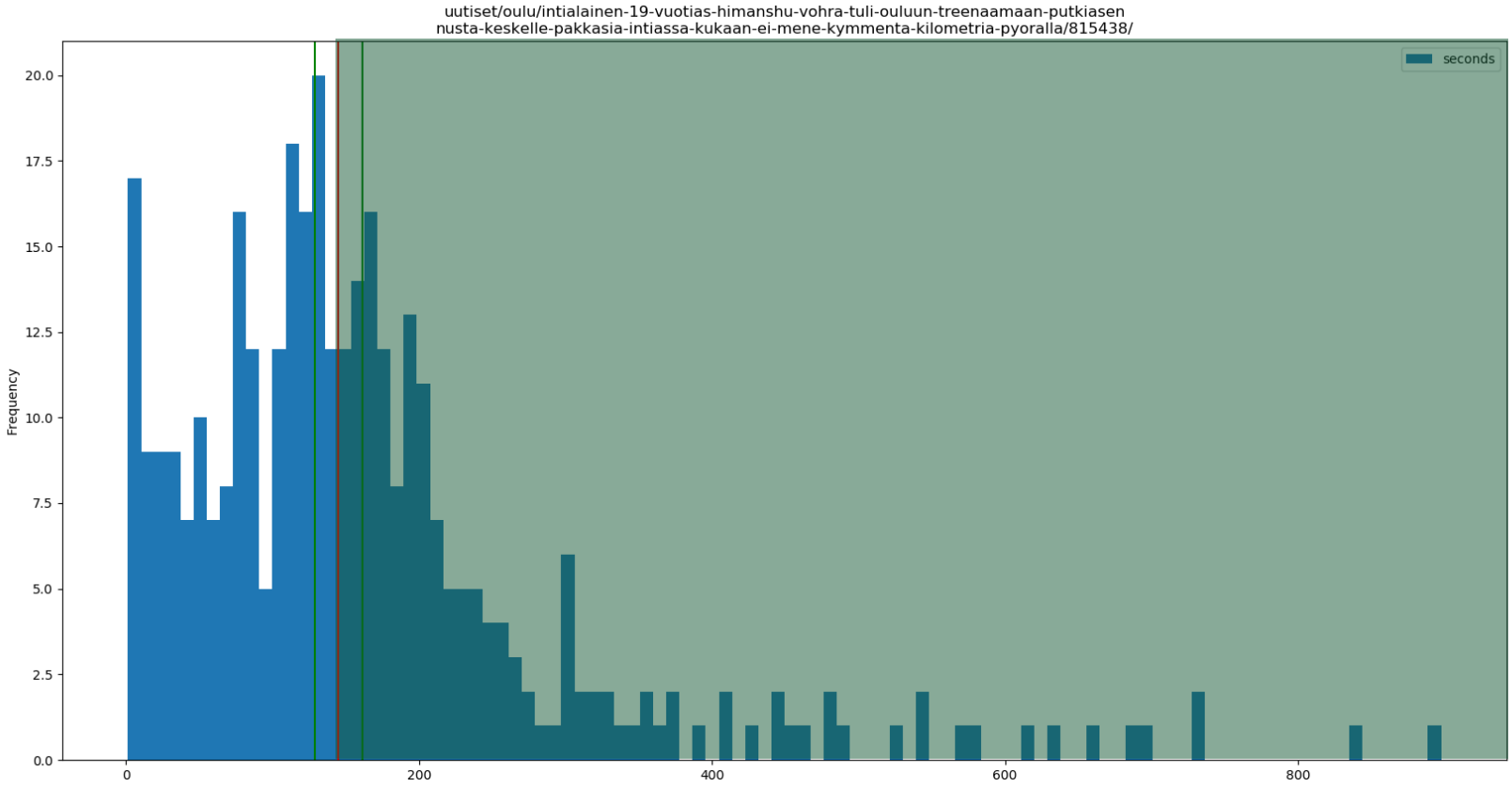
Metrics for measuring headline impact in online news media

Measuring the impact of a news headline

- Two key points:
  - Number of readers
  - Time spent on article, so-called *dwell time*
- Number of readers = Click-through rate (CTR)
- Dwell time is estimated from the click stream data:
  - Isolating and extracting unique users and their paths
  - Calculating the time between two consequent clicks → time spent on a single article (excluding the last click of a user)
- Reading time is dependent on article length
  - Literature review - average read time for Finnish: 161 ±18 words per minute [*] → estimation of how much of the article was read

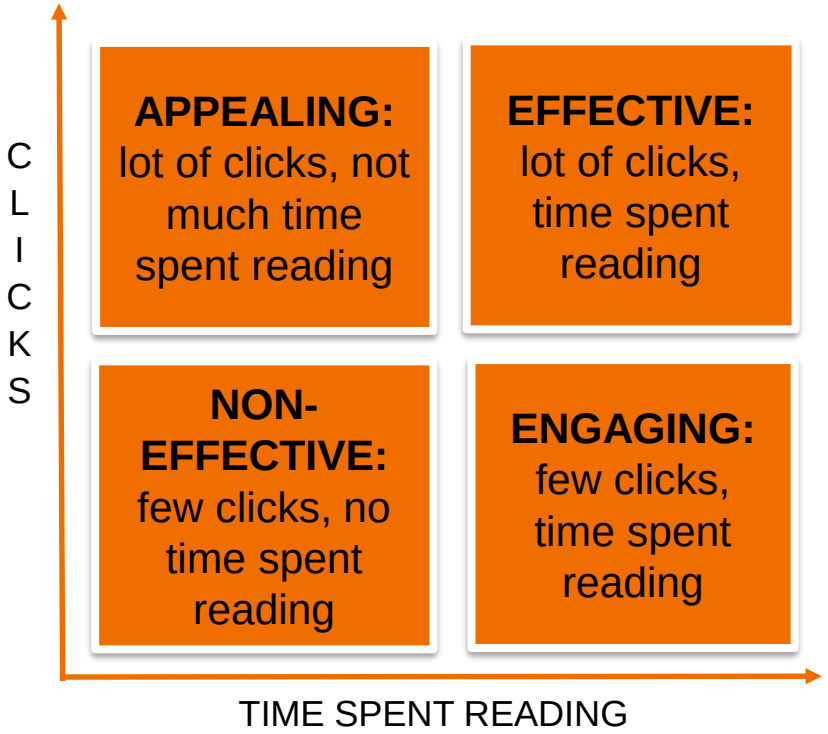
[*] “New standardised texts for assessing reading performance in four European languages”, 2006


uutiset/oulu/intialainen-19-vuotias-himanshu-vohra-tuli-ouluun-treenaamaan-putkiasen
nusta-keskelle-pakkasia-intiassa-kukaan-ei-mene-kymmenta-kilometria-pyoralla/815438/
seconds
Frequency
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
0
200
400
600
800


uutiset/oulu/intialainen-19-vuotias-himanshu-vohra-tuli-ouluun-treenaamaan-putkiasen
nusta-keskelle-pakkasia-intiassa-kukaan-ei-mene-kymmenta-kilometria-pyoralla/815438/
seconds
Frequency
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
0
200
400
600
800

Two-dimensional classification of articles

- Dimensions:
  - mean read percentage
  - number of clicks
- Four classes of article impact:
  - Non-effective (few clicks, low read-%)
  - Appealing (many clicks, low read-%)
  - Engaging (few clicks, high read-%)
  - Effective (many clicks, high read-%)

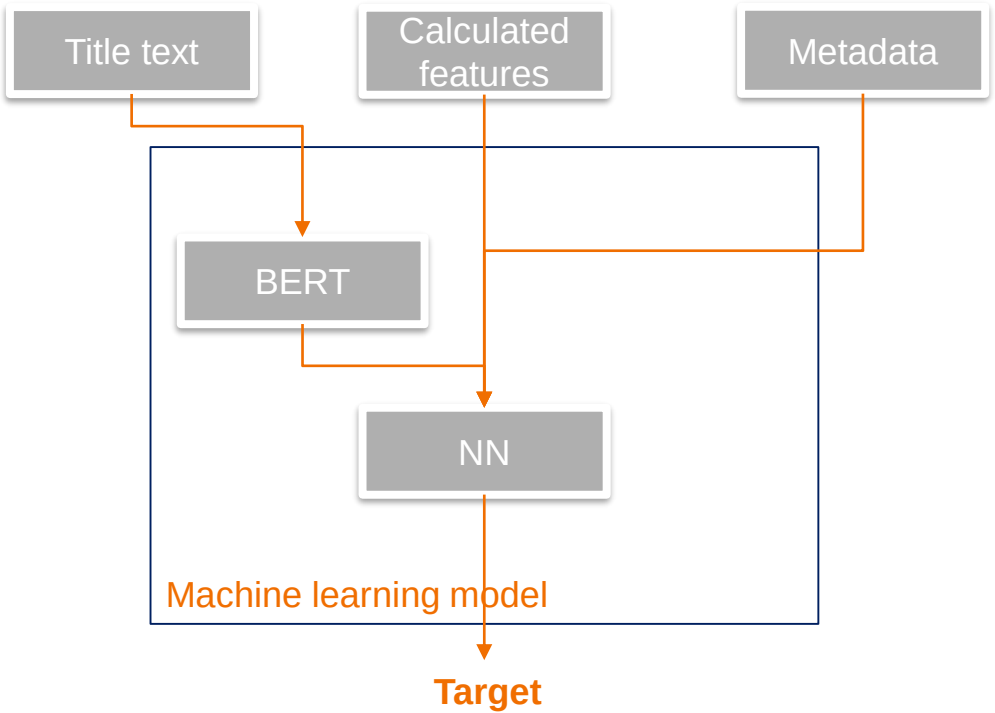


OTSIKKOKONE
the model

Model structure

Input data:

- Title text
- Calculated features
  - E.g. title length, named entity counts, distribution of word classes
  - Article text in vector format (Latent dirichlet allocation, LDA)
- Metadata
  - Section
  - Free/subscriber-only
  - Publish time



BERT (Bidirectional Encoder Representations from Transformers)

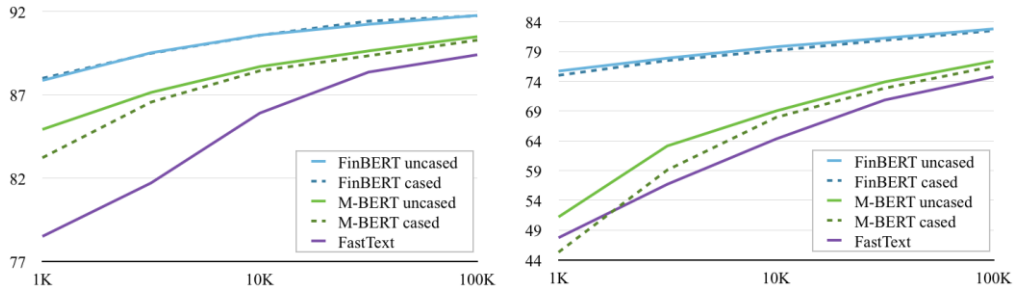
- Deep neural network model that can be used as a base in natural language processing (NLP) tasks
  - Unsupervised bidirectional language representation model (345M parameters)
  - Originally trained to predict missing words on texts → the model is able to understand the relations between words
- Enables the utilization of transfer-learning in NLP model development
  - Part of the original BERT model is used as it is but last layers of the model are retrained for a new task
    - Possible to train the model with less training material (compared to training the full model)
    - The pretrained features are utilized
    - Commonly used technique in machine vision
- Available for e.g. English and Chinese, also multilingual version exists

- *"BERT makes use of Transformer, an attention mechanism that learns contextual relations between words (or sub-words) in a text. In its vanilla form, Transformer includes two separate mechanisms — an encoder that reads the text input and a decoder that produces a prediction for the task. Since BERT's goal is to generate a language model, only the encoder mechanism is necessary."*
- *"When Google researchers presented a deep bidirectional Transformer model that addresses 11 NLP tasks and surpassed even human performance in the challenging area of question answering, it was seen as a game-changer in NLP/NLU."*

https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270

BERT (FinBERT)

- Finnish version available (FinBert)
  - Developed by Turku NLP Group (University of Turku)
  - Trained with news texts, online discussion and other material crawled from the internet
  - Outperforms the multilingual Bert → used in Otsikkokone



Virtanen, Antti, et al. "Multilingual is not enough: BERT for Finnish." *arXiv preprint arXiv:1912.07076* (2019).
https://medium.com/sciforce/googles-bert-changing-the-nlp-landscape-5f4a7bf65cc5
https://github.com/TurkuNLP/FinBERT

OTSIKKOKONE implementation - what did we do and how the results look like?

Dataset for Otsikkokone development

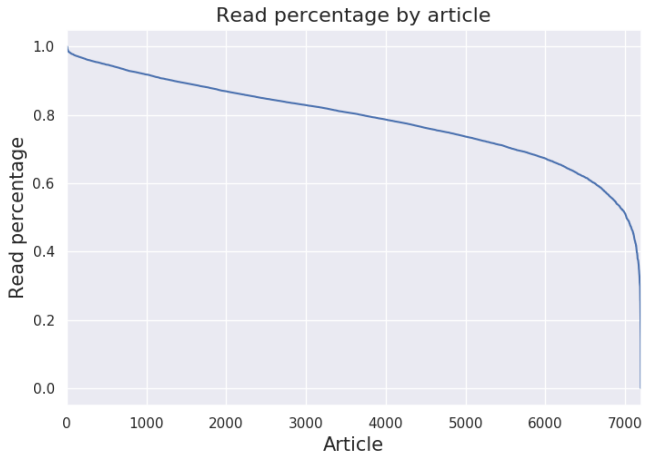
- Click-stream data stored as csv-files
- All the clicks to the news portal between 12/2018 –5/2019
- Each click contained records like timestamp, news title, news section, author, premium, userID etc

- Whole dataset ~108M clicks and ~205k unique article Ids
- Data used in development of Otsikkokone ~17M clicks and ~7k articles

Data preprocessing

- Removing irrelevant articles
  - Like photo galleries, comics, and corrections
- Scraping additional data
  - Publication times, article text, multimedia attached
- Obtaining read time values
- Removing duplicate articles

Data characteristics: CTR & read percentage



Defining target variable

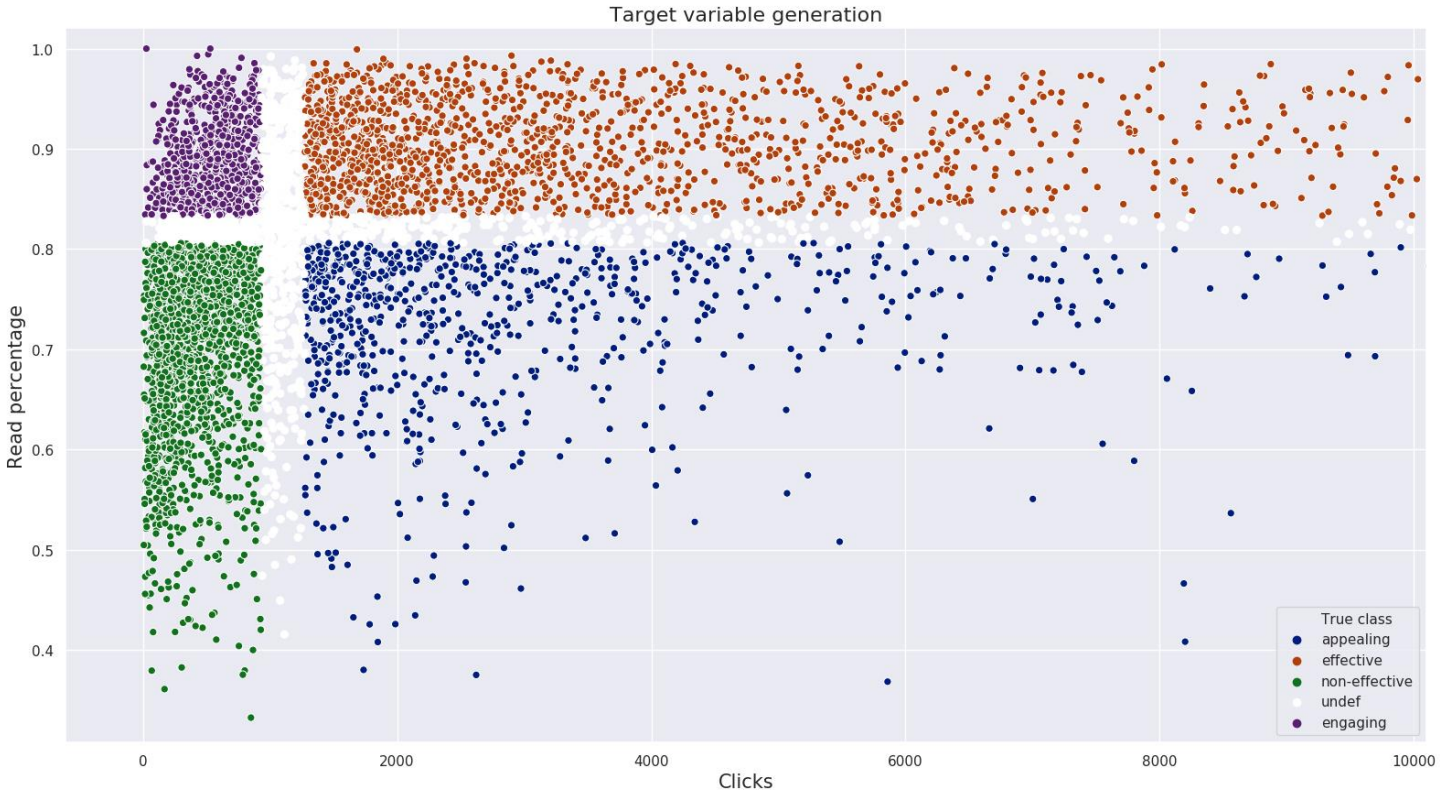
Starting point: CTR & read percentage as *continuous* variables

Target: Four-class variable combining both variables

Solution:

- Split both variables at median
- However, leave out 10% of articles around medians
  - At that zone, the true class is rather random

Defining target variable



Prediction tasks

1. Binary classification (CTR)
   - Best testing accuracy 77.0%

2. Binary classification (read percentage)
   - Best testing accuracy 71.8%

3. Binary classification (effective vs. rest)
   - Best testing accuracy 78.0%

4. Four class classification (CTR & read percentage)
   - Best testing accuracy 58.7% (baseline 25%)

Experiments for model optimization
Early tests conducted only for binary metric

1. Detect the best BERT model to use
   - FinBERT or multi-BERT
   - FinBERT some 6 %-points more accurate

2. Network structure
   - Number of hidden layers: 1 (2nd did not improve scores)
   - Number of neurons: 256 (scores did not improve if more were used)

Experiments for model optimization
Which features to include

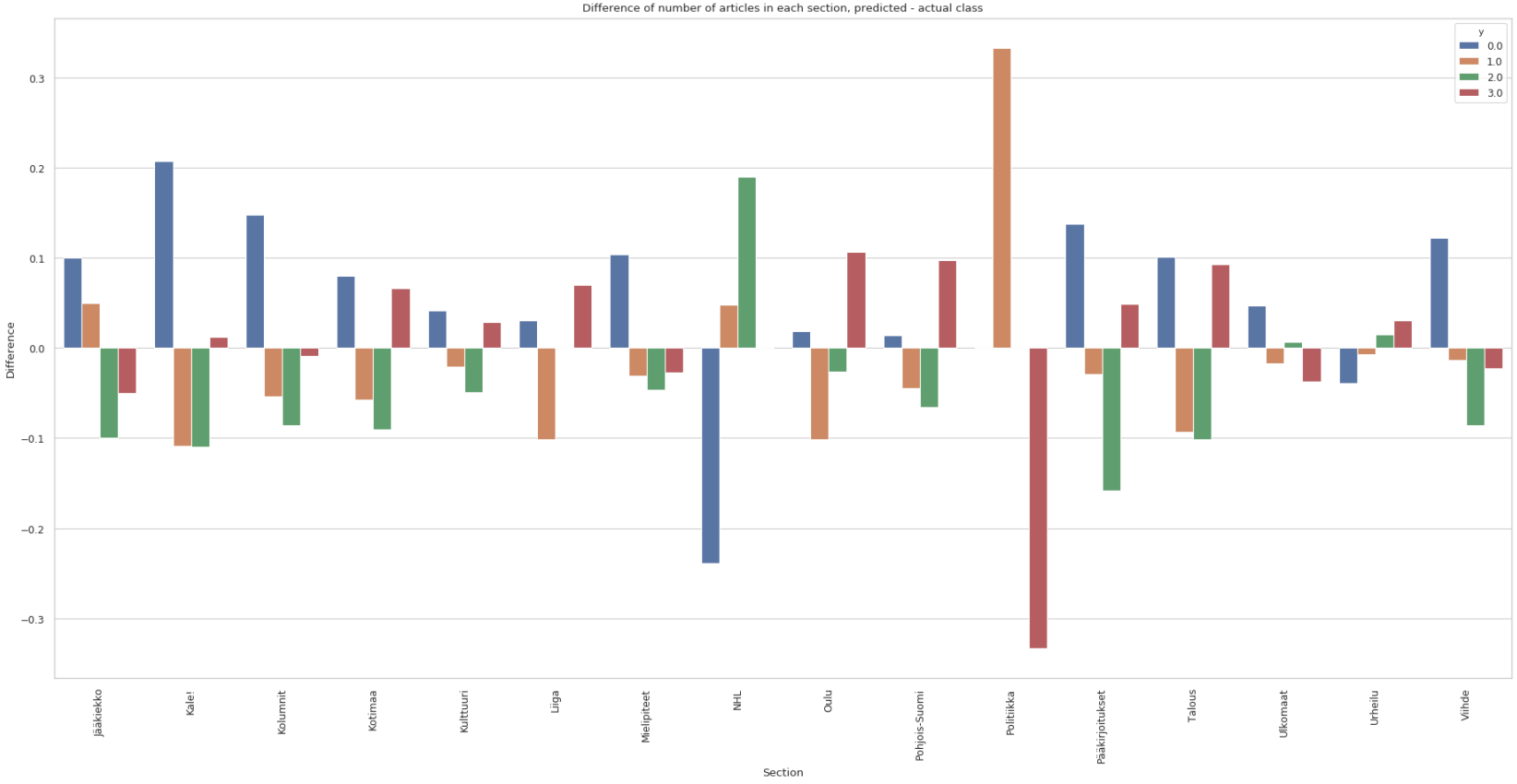
- Title features (length, word classes, punctuation...)
  - Currently included but no major differences
- Topic of article extracted with Latent Dirichlet Allocation
  - Excluded, no improvement
- Section
  - Improves scores by 2-3 %-points
  - However, huge impact on predicted class – demo version does not use
- Publication time
  - Excluded, no improvement
- Premiumity
  - Included a priori, affects how metric is calculated

Title features affecting the response

- If title contains:
  - The word **"Oulu"**, it is clicked more
  - A **quote**, it is clicked more but not read more
  - A **dash** (-), there is no effect
  - **Named entities**, the class "Effective" is less likely and classes "Appealing" and "Engaging" are more likely
  - **Quantities or money**, "Effective" is more likely

- The mean length of title does not vary much
  - 11.8, 14.3, 10.3, and 12.4 words in each class

How section affects the prediction?

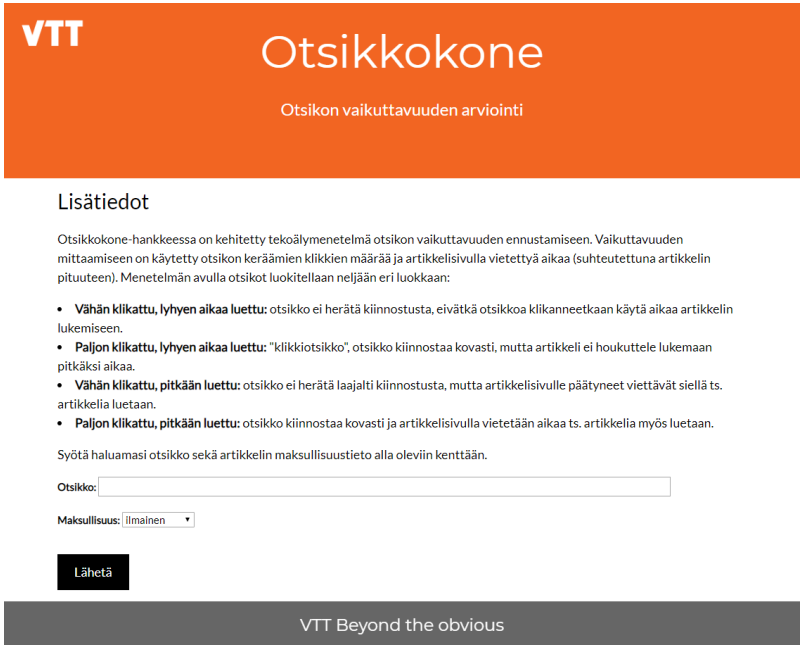


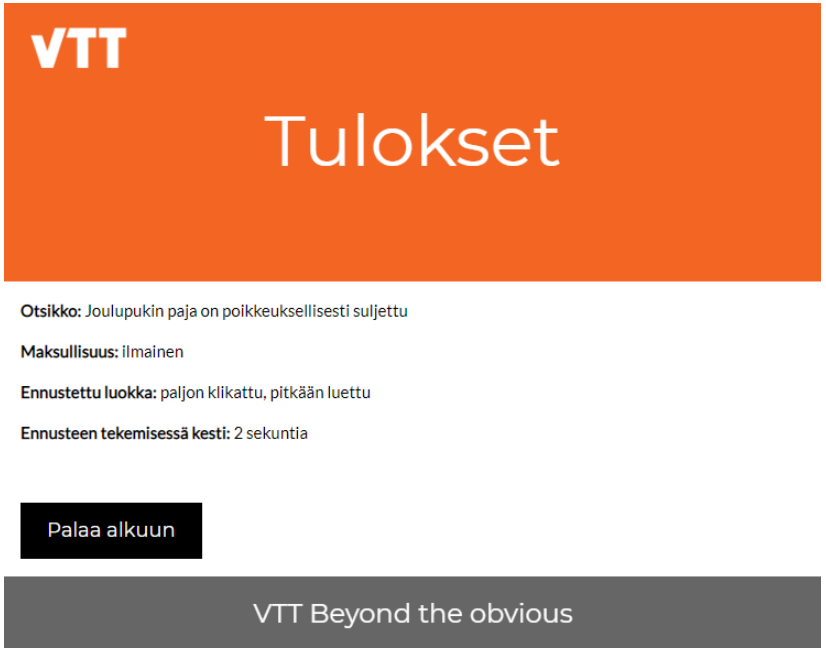
Evaluation against human experts

- 80 headlines from Kaleva's data randomly selected for web-based evaluation questionnaire
- Five journalists against Headline Machine

- Average accuracy of journalists: **25.75%**
- Accuracy of Headline Machine in the subset: **61.25%**

Browser-based Otsikkokone
Tool for predicting news headline impact





Additional use case: Section prediction from a news headline

- The FinBERT model was used to predict the section of a news article based solely upon the features of the headline text
- Test class: 857 articles, 15 different sections
- The accuracy for the whole test class was **69.6%**, with mean class-wise F1-score of 0.653
  - The most difficult section to classify was ”Eduskuntavaalit 2019”, with 30% accuracy of the total 30 headlines
  - The best classification accuracy was for the section ”Urheilu”, with 95.7% accuracy of the total 94 headlines

Section prediction - confusion matrix

	Eduskuntavaalit 2019	Kale!	Kolumnit	Kotimaa	Kulttuuri	Mielipiteet	Oulu	PPY	Pohjois-Suomi	Pääkirjoitukset	Talous	Ulkomaat	Urheilu	Vaalipuheet	Viihde
Eduskuntavaalit 2019	**9**	0	0	18	0	0	1	0	1	1	0	0	0	0	0
Kale!	0	**7**	0	0	1	0	1	0	0	0	0	0	0	0	0
Kolumnit	0	0	**7**	3	0	0	0	0	0	0	0	0	0	1	0
Kotimaa	5	0	0	**171**	2	4	12	0	2	1	4	10	2	0	5
Kulttuuri	0	0	0	3	**9**	1	4	0	0	0	0	1	1	0	8
Mielipiteet	0	0	2	1	0	**12**	3	0	0	1	0	0	0	2	0
Oulu	0	0	1	19	4	0	**109**	0	15	1	3	0	0	0	0
PPY	0	0	0	0	0	0	0	**6**	0	0	0	0	0	3	0
Pohjois-Suomi	0	0	0	13	1	2	19	0	**72**	0	1	0	1	0	0
Pääkirjoitukset	1	0	3	5	0	1	0	0	0	**4**	0	1	0	0	0
Talous	0	0	0	10	0	0	4	0	0	0	**8**	1	0	0	0
Ulkomaat	0	0	1	13	1	0	1	0	1	0	0	**86**	1	0	0
Urheilu	0	0	0	2	0	0	0	0	1	0	0	1	**90**	0	0
Vaalipuheet	0	0	0	0	0	1	0	2	0	0	0	0	0	**5**	0
Viihde	0	0	0	0	3	0	0	0	0	1	0	3	0	0	**20**

Otsikkokone open source package

- https://github.com/vttresearch/otsikkokone

Conclusion and future potential

Conclusion

- Definition of an effective – non clickbait - headline
- Metrics for measuring headline effectiveness in online news media
- Tools to analyze the web analytics data and news content to measure the effectiveness
- AI methods for predicting the online news headline impact
  - Accuracy of ~58%
- Evaluation compared to experts (journalists)
  - Accuracy of ~26%
- Otsikkokone test usage going live in Kaleva during the spring
- Draft of a conference paper, plans for a journal article

Future potential in media sector

- OTSIKKOKONE model improved with additional data sources (e.g. data on positioning in the portal)
- Different use cases for AI methods:
  - Content personalisation
  - Reader engagement, data on behaviour leading to purchase decision, optimization of the pay wall
  - Automatic analysis of comments and/or moderation
  - Sentiment analysis from textual content
  - Optimization of marketing and communications
  - Segmentation of text
  - Automatic language translation

beyond

the obvious

Sari Järvinen
sari.jarvinen@vtt.fi
+358 40 512 9662

@VTTFinland

www.vtt.fi