Clickbait Detection based on Comparison of Headlines to Articles

Daria Kim (dvkim@edu.hse.ru)

National Research University Higher School of Economics
Moscow, Russia

Due to the popularity of social networks and personalized search algorithms, it becomes more and more difficult for people to evaluate media quality and resist disinformation. In order to facilitate this task, new methods for automated fact checking are required. The first step to solving the problem of automatic news verification is automated clickbait filtering. This paper presents a dataset of 11 thousand news for clickbait classification and provides machine learning classification baseline. In addition, it was proved that the use of comparative characteristics of headings to texts for training helps to improve models' performances.

Key words: *clickbait detection, recurrent neural networks*

1 Introduction

The growing spread of personalised search algorithms and social media newsfeeds may cause filter bubbles – a state of intellectual isolation which prevents users from receiving information they weren't previously looking for. Positive for pastime reading, filter bubbles often lead to lack of knowledge about real diversity of views on substantial and controversial topics, resulting in vulnerability to misinformation. This can explain the recent media prevalence of the term «fake news»[1]: the absence of diversity in opinions makes online communities prone to propaganda. If the news verification task became too hard for average user to solve, automated fact-checking systems are in demand.

According to Shu et al. [1] machine-based fact-checking systems are now far from being widely implemented. Expert-based validation methods are still in use, and most of approaches listed in [1] review aim to assist human experts, extracting facts from article's text, highlighting possibly misleading sentences and clickbait headlines etc. The latter is said to be the distinctive feature of internet hoaxes: clickbait is a headline which is intended to attract the user to click the link and read the article.

The other technique of filtering misinformation and provocation in media is stance detection [2]. This approach relies on online comments left to post or news article. Authors of SemEval [3], a benchmark dataset addressing this task, define stance as «a subject's reaction to a claim made by a primary actor», therefore the problem of stance detection for fact verification can be rephrased as the following question: «Do the users trust the news article?»

The recent study [4] implementing stance detection model to news dataset has proved that adding comparative characteristics of text and article results in increase of deep learning model performance.

Our paper proposes news dataset for clickbait detection in Russian and baseline for clickbait classification, proving that adding text comparison features improves the accuracy and recall of the models.

[1] A made-up story with an intention to deceive

2 Related Work

2.1 FNC-1

The motivation for exploring stance detection techniques can be found in works published by Fake News Challenge organizers[2] [5] and participants. FNC-1, a machine learning competition held in 2017, aimed to address the problem of fake news detection on the Internet. Authors of FNC-1 task and dataset believed that results achieved by participants may serve in stance detection applications such as clickbait detection.

The organizers of the FNC-1 provided a baseline[3] of gradient boosting classifier trained on hand-crafted features: word overlap count and polarity score. Participants had suggested wide range of ML models solving the news classification task — from linear algorithms such as logistic regression or SVM to MLP and voting model combining gradient-boosting decision trees and deep neural model into one predictor, which resulted in the highest competition score beating organizers benchmark by 28 percent.

Several papers addressing FNC-1 problem were published after the main competition track was closed. Borges, Martins and Calado [4] obtained both title-text and title-lead[4] similarity rates from data to train the model. Exploring correlations between the title and the lead helped to improve predictions accuracy. Our model, discussed in section is inspired by preprocessing and training pipeline suggested in this work.

2.2 RuStance [6]

So far, automatic stance prediction, as well as clickbait/fake news classification has been investigated for a limited range of languages: English, Catalan and Spanish. At first, a multi-purpose dataset for stance detection related applications was obtained by [6] who published their data collection named RuStance.

RuStance contains the set of 958 tweets and comments responding to news articles published by Meduza and RT in late 2017, each labeled as either «support», «deny», «query» or «comment» opposed to replied/cited article. The best performance of 0.83 F1-score and 0.92 accuracy was achieved using the gradient boosting algorithm on word vector representations. However, RuStance authors and developers admit that the size of the dataset is not enough to train deep learning models, demonstrated the best results on stance detection task according to recent publications [2].

3 Data

The original corpus for further analysis and training was collected during 2018 web scraping sessions. Figure 2 describes the list of the news media sources used in retrieval and processing.

Initial news dump contained over 1 million items dated back from mid-2004 to late 2018. To retain relevance of both the data and the models trained on it, all the articles published before January 2018 were ignored.

Clickbait articles were labeled manually, while most of «trustworthy» items were labeled such because of the sources they were obtained from: mostly news agencies, e.g. Novaya Gazeta, RIA and Iterfax. News agencies have to follow article naming rules, controlling headline style and lexical characteristics.

[2] http://www.fakenewschallenge.org/

[3] https://github.com/FakeNewsChallenge/fnc-1-baseline

[4] Using *lead* – generally two first sentences of an article – to briefly report on key facts is a conventional journalist practice.

Figure 1: Russian news dataset statistics

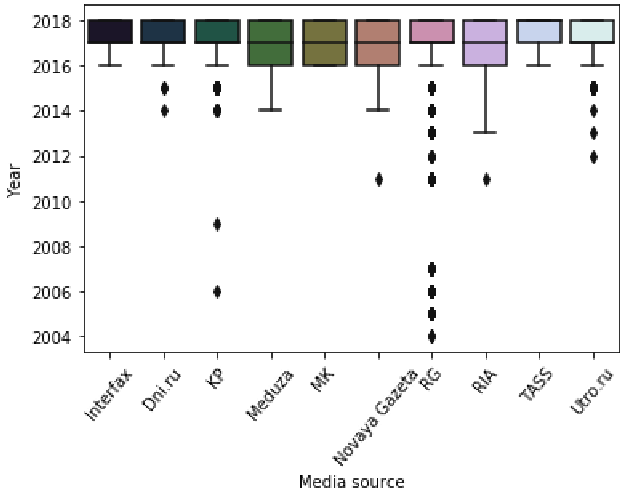

Figure 2: Sources distribution in dataset

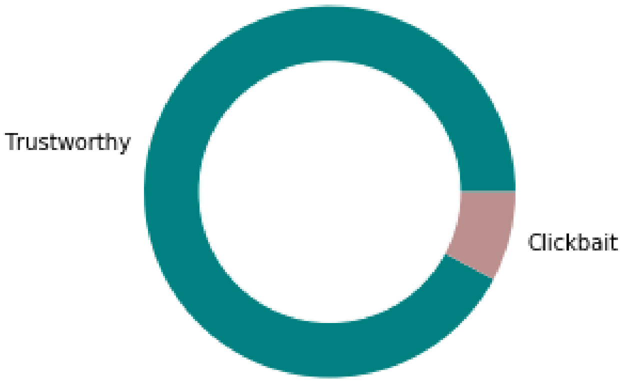

Figure 3: Classes ratio

Demography of clickbait articles annotators: 60% women, 40% men aged between 18 and 30, finished or continuing higher education. If more than a half of 20 annotators marked the article as ambiguous, it was discarded because of inconsistency with the task and unreliability of human annotators decision.

During the experiment with manual labeling of smaller corpus with equal presence of both classes, the following results emerged:

- 45% of headlines were labeled as clickbait by more than 90% of annotators
- 46% of headlines were considred trustworthy by more than 90% of annotators
- 9% were labeled «ambiguous» by more than 90% of annotators or considered clickbait and trustworthy by the same number of annotators

The dataset remained after removing duplicates and unfinished entries contains 11.5K articles labeled either as «Clickbait» or «Journalism».

3.1 Preprocessing and Feature Extraction

Texts were normalized with *Mystem* [7] and cleaned from stopwords, hyperlinks, punctuation etc. All entries of information agency signature such as «Москва. 28 апреля. INTERFAX.RU» were also removed to prevent model overfitting to the pattern not related to clickbait/non-clickbait definition itself.

Hand-crafted features for Boosting and Linear Models

The methods used to extract headline-article comparison features were proposed by FNC-1 authors in their baseline model and were further used by all of the participating teams.

ROUGE similarity score and cosine distance between *word2vec* representations were also used to enhance model performance. Table 1 in Experiments section represents the results of model training and testing.

Additional preprocessing for LSTM

To prepare data for hierarchical LSTM model input, article bodies were splitted into separate sentences using *razdel* [8] sentenizer. First two sentences from each article were extracted as a lead.

4 LSTM model

The neural network architecture inspired by model proposed in [4] exploits a hierarchical approach [9] [10] for modeling headlines and bodies of news articles.

In this approach, two stacked layers of Long Short-Term Memory units followed by summerization|attention mechanism are used to obtain the matrix of encoded sentences of the text. Matrix of encoded sentences, in turn, is also processed through the same bi-directional recurrent layer to form a vector representation of the whole text. Summarization mechanism is applied to the output of bi-directional LSTM to output a single vector.

In contrast to [4] our hierarchical LSTM model is fed only with tokenized texts, headlines and leads, not accompanied by any hand-crafted features like string distance metrics used by FNC-1 participants. Headline, lead sentences and text of articles are encoded separately and then matched the following way:

1. the vector of headline is concatenated with, element-wise multiplied and subtracted by the vector of body

2. the vector of headline is concatenated with, element-wise multiplied and subtracted by the vector of lead

Then results of the previous steps are concatenated and fed forward to dropout, dense and sigmoid activation layers.

Figure 4 describes proposed Hierarchical LSTM model.

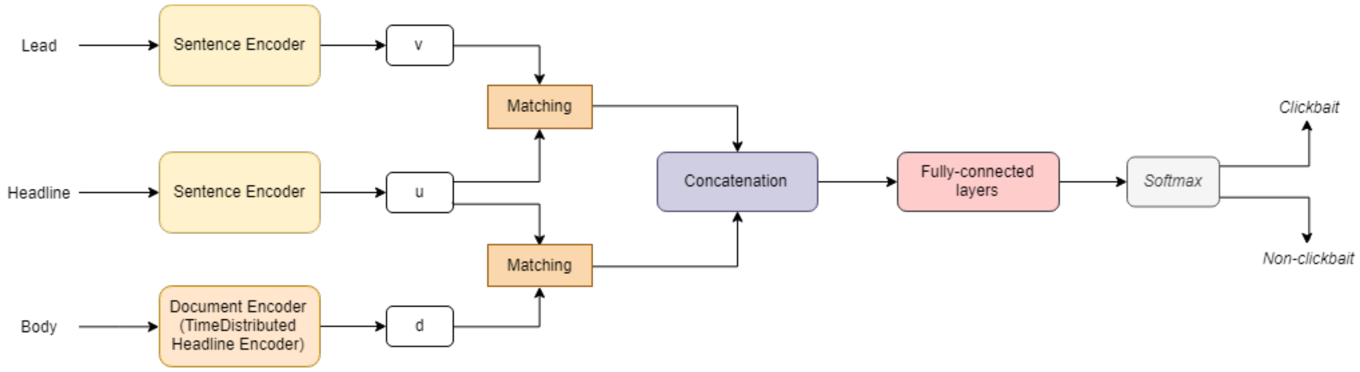

Figure 4: LSTM model scheme

5 Experiments and Results

The key advance of suggested model is the use of *concatenate* and *multiply* layers for comparison of text encodings obtained by LSTM. To test whether our approach provides an increase in classification quality, several different models were evaluated on the same data. Stochastic gradient boosting model is basically the one used by FNC-1 authors to construct competition baseline, and the logistic regression model was suggested and evaluated on FNC-1 dataset by [11]. Both gradient boosting and logistic regression models were trained on three types of input features vectors:

- only TF-IDF representation of the whole article, ignoring the headline-body structure
- only headline-body distance metrics: word overlap score, BLEU, ROUGE and cosine distance
- combination of previous two feature sets

In addition to linear models and ensembles we tested the simple recurrent network model, which ignores article’s headline-body structure, receiving the whole text as an input.s

Table 1 describes models’ hyperparameters and performance. It’s important to note that the higher the recall was, the better the model performs – as the number of articles labeled 0, or «Journalism» was almost 7 times higher than the number of articles considered «Clickbait».

The experiments lead us to several conclusions:

1. Deep learning model results require shorter and simpler preprocessing and feature engineering pipeline, but are equal or better than ones obtained via ensemble model
2. Recall achieved of hierLSTM model is equal or better than results achieved on ensemble and linear models. This error score is more substantial than precision, which decreased on LSTM, as ’trustworthy’ labels of articles may be false or ambiguous, while presented ’clickbait’ is free of labeling mistakes.

The code for data preprocessing and models training can be found at GitHub.[5]

№	Classifier	N-gram range	Features	Tokenizer	Parameters	Accuracy	Recall	Precision
1	Logistic Regression	1-2	TF-IDF	Mystem	’alpha’: 0.01 ’l1_ratio’: 0.15	0.86	0	0
2	Logistic Regression	1-2	N-gram cooccurrence score Overlap distance Cosine distance BLEU	Mystem	’alpha’: 0.1 ’l1_ratio’: 0.15	0.86	0.03	0.52
3	Logistic Regression	1-2	TF-IDF N-gram cooccurrence score Overlap distance BLEU Cosine distance	Mystem	’alpha’: 0.01 ’l1_ratio’: 0.55	0.87	0.9	0.35
4	Gradient Boosting	1-2	TF-IDF	Mystem	’n_estimators’: 400 ’subsample’: 0.8	0.96	0.92	0.98
5	Gradient Boosting	1-2	N-gram cooccurrence score Overlap distance BLEU Cosine distance	Mystem	’n_estimators’: 100 ’subsample’: 0.8	0.9	0.51	0.73
6	Gradient Boosting	1-2	TF-IDF N-gram cooccurrence score Overlap distance BLEU Cosine distance	Mystem	’n_estimators’: 400 ’subsample’: 0.8	0.97	0.96	0.99
7	Non-hierarchical LSTM	–	Tokenized article	Mystem Keras Tokenizer	100 LSTM units	0.97	0.99	0.81
8	Hierarchical LSTM	–	Tokenized text Tokenized headline Tokenized lead	Razdel Mystem Keras Tokenizer	50 LSTM units	0.96	0.992	0.75

Table 1: Models, parameters and performance score

6 Conclusion and future work

The experiment shows that comparison of vector representations of texts for clickbait detection provides an increase in quality both in cases of ensemble and recurrent network model. Let us try to explain its effectiveness in simple terms.

[5] https://github.com/kimaril/clickbait-classifier

When reading we move from one word or sentence to another sequentially, following the narrative. Bidirectional LSTM layer models this process to some extent, which enables model to learn the «context» vector representation. The following layers of network serve to extract the most important components from this representation and to measure similarity between vectors of headline, lead and body of the article.

References

[1] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36, 2017.

[2] Ray Oshikawa, Jing Qian, and William Yang Wang. *CoRR*, abs/1811.00770, 2018.

[3] Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. Semeval-2017 task 8: Rumoureval: Determining rumour veracity and support for rumours. *arXiv preprint arXiv:1704.05972*, 2017.

[4] Luís Borges, Bruno Martins, and Pável Calado. Combining similarity features and deep representation learning for stance detection in the context of checking fake news. *ACM Journal of Data Information and Quality*, 2019.

[5] Andreas Hanselowski, Avinesh PVS, Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M. Meyer, and Iryna Gurevych. A retrospective analysis of the fake news challenge stance-detection task. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1859–1874, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

[6] Nikita Lozhnikov, Leon Derczynski, and Manuel Mazzara. Stance prediction for russian: Data and analysis. *CoRR*, abs/1809.01574, 2018.

[7] nlpub/pymystem3: A python wrapper of the yandex mystem 3.1 morphological analyzer (http://api.yandex.ru/mystem). `https://github.com/nlpub/pymystem3`.

[8] natasha/razdel: Rule-based tokenizer, sentenizer for russian language. https://github.com/natasha/razdel.

[9] Frederic Morin and Yoshua Bengio. Hierarchical probabilistic neural network language model. In *Aistats*, volume 5, pages 246–252. Citeseer, 2005.

[10] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 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, 2016.

[11] Razan Masood and Ahmet Aker. The fake news challenge: Stance detection using traditional machine learning approaches. 09 2018.