Детекция манипуляций в новостном потоке

A Preprint

Мелихов Дмитрий Александрович
Факультет вычислительной математики и кибернетики
МГУ им. Ломоносова
melikhov.dmitry.a@gmail.com

Воронцов Константин Вячеславович
Факультет вычислительной математики и кибернетики
МГУ им. Ломоносова
vokov@forecsys.ru

Abstract

В работе решается задача выявления манипуляций в новостном потоке. В новостных статьях выделяются манипулятивные фрагменты и помечается тип манипуляции. Для этой задачи представляется новый набор данных на основе новостных статей из открытых сточников, где выделяются фрагменты указывающие на продвигаемые ценности. В работе рассматриваются модели на основе больших лингвистических моделей, которые выявляют фрагменты и тегируют. Для выявления фрагментов и тегирования рассматриваются постановки задачи: классификация токенов на принадлежность к фрагментуу, классификация пар токенов начало-конец. Предлагаются новые критерии качества модели, которые учитывают длину фрагментов.

Keywords span identification · text tagging · manipulation detection

1 Введение

Современные средства массовой информации генерируют огромный поток данных на социально-политические темы. При этом они охватывают огромное число читателей и во многом формируют у них определённый набор ценностей Mutz and Goldman [2010]. Возникает потребность автоматически обрабатывать новостной поток для выявления манипуляций Martino et al. [2020], суммаризации текста Kemahduta et al. [2021].

Манипуляцией в тексте назвается воздействие на читателя с целью сформировать определённое отношение к цели (мишени манипуляции). Среди манипуляций можно выделить: эмоциональное воздействие, предоставление недостоверной информации, ложные причинноследственные связи.

К обработке новостного потока можно подходить с точки задачи классификации или регрессии. В 2007 году появилось соревнование SemEval-2007 Task 14 Strapparava and Mihalcea [2007], где основная задача - выявление эмоциональной нагрузки заголовков статей, которая сформулирована как задача регрессии - каждой эмоции сопостовляется число от 0 до 100. В 2022 году предложили датасет на основе статей с Rappler для выявления эмоций читателя K. et al. [2022]. Существует постановка задачи, в которой новости нужно классифицировать по политической идеологии, например определить кто написал статью: левый, правый или центрист. В статье Baly et al. [2020a] предлагается датасет на основе данных с сайта AllSides[1]. Также в данной статье была найдена важная проблема - смещение по источнику новости (media bias). Моделям проще выучить стиль написания статей разными СМИ

[1] https://www.allsides.com/media-bias/media-bias-rating-methods

и предсказывать политическую идеологию по ним, вместо того, чтобы опираться на утверждения в тексте.

Для классификации раньше использовались алгоритмы, основанные на праилах (rule based) и классические подходы. Они были популярны в соревновании SemEval-2007 Task 14 Strapparava and Mihalcea [2007]. С развитием нейросетей, в том числе больших языковых моделей, качество решений улучшилось. В 2019 году была предложена модель BERTDevlin et al. [2019]. Данная модель стала популярной для задачи классификации текстов. В статье K. et al. [2022], где предлагался новый датасет, предлагалась модель Bi-LSTM с attention усреднением эмбеддингов. В статье по предвзятости новостей Baly et al. [2020a] для классификации рассматривались LSTM и BERT.

Другой подход - выделение фрагментов (span identification)Papay et al. [2020], Toshniwal et al. [2020]. Выделение фрагментов используется для выделения ошибок в текстеChen et al. [2020], построения синтаксической структуры текстаYeung and Lee [2015], суммаризации Ma et al. [2018], La Quatra et al. [2019], Liu et al. [2021], анализа цитирования научных статейLa Quatra et al. [2019], построения графа знанийCheng et al. [2020], выделение именованных сущностейLi et al. [2019], Rojas et al. [2022]. В задаче детекции манипуляций выделяются фрагменты, указывающие на манипуляцию. Для модерации платформ популярна задача выделения оскорбительных фрагментов. В 2021 был проведено соревнование SemEval-2021 Task 5Pavlopoulos et al. [2021], где требовалось выделить оскорбительные фрагменты текста. Также проводилось соревнование на платформе codalab [2], результаты описаны в статье Ravikiran et al. [2022]. В некторых постановках задач нужно сопоставлять фрагментам теги, указывающие на тип манипуляции. В 2020 году было проведено соревнование SemEval-2020 Task 11Martino et al. [2020], где требуется выделять манипулятивные фрагменты и классифицировать их на 14 классов.

Для нахождения фрагментов в задаче суммаризации использовалось SVM, логистическая регрессия, решающие деревья Ma et al. [2018], синтаксические деревья Yeung and Lee [2015]. С развитием нейросетей стали популярны подходы с трансформерами BERTXu et al. [2023], RoBERTaRavikiran et al. [2022], Jurkiewicz et al. [2020], свёрточные нейронные сети Dewantara et al. [2020], GPT-2 Nouri [2022]. Данные модели можно ансамблировать и получать результат лучше, что можно заметить в обзоре результатов соревнования SemEval-2020 Task 11Martino et al. [2020]. Сравнение трансформерных моделей можно увидеть в статьеToshniwal et al. [2020]. Также для данной задачи предобучен SpanBERT Joshi et al. [2020]. Для выделения пересекающихся фрагментов можно использовать графовые нейронные сети с BERTZaratiana et al. [2022]. Также для задачи с вложенными фрагментами можно использовать multiple LSTM+CRFRojas et al. [2022].

2 Постановка задачи

В данной работе рассматривается новый русскоязычный датасет, составленный лабораторией "Машинного обучения и семантического анализа" института искусственного интеллекта. Для данного датасета рассматривается задачи выделения фрагментов, выделения связи между фрагментами и тегирования.

Задача выделения фрагментов - задача классификации слов:

$$CE(S, p(S \mid \theta')) \to \min_{\theta'}$$

Где $S \in \{B, I, O\}^N$ - BIO разметка текста.

Тегирование - multilabel классификация:

$$BCE(Y, p(Y \mid \theta'')) \to \min_{\theta''}$$

$Y_{i,j} = 1$, если фрагмент f_i имеет тег T_j.

3 Метод решения

Для решения задачи упрощается структура разметки и решается сразу несколько задач:

1. Нахождение фрагментов в тексте.
2. Объедининение фрагментов в элементы разметки. Истинные фрагменты известны.

[2]https://competitions.codalab.org/competitions/36395

3. Тегирование фрагментов. Истинные фрагменты известны.
4. Тегирование элементов разметки. Истинные фрагменты и элемены разметки известны.

Задачи решаются в одном итерационном процессе. На каждой итерации для каждой задачи последовательно делаются градиентные шаги. Подобные подходы применяются для доменной адаптации (Baly et al. [2020b]).

Для вычисления эмбеддингов слов применяются лингвистические модели на основе BERT для русского языка от SberDevices(Zmitrovich et al. [2023]) и DeepPavlov(Kuratov and Arkhipov [2019]). Для выявления связей между фрагментов используются text2graph модели Guo et al. [2020]. Для тегирования используются эмбеддинги фрагментов.

Находим слова, которые являются частью фрагмента.

$$X = \{x_n\}_{n=1}^{N} - \text{текст} \qquad X \in \mathbb{N}^N$$

$$S \in \{0,1,2\}^N - \text{разметка фрагментов (BIO разметка)}$$

$$E = BERT(X) \in \mathbb{R}^{N \times d} - \text{эмбеддинги слов}$$

$$\hat{S}_{i,j} = \text{softmax}(EW^F) \qquad W^F \in \mathbb{R}^{N \times 3} - \text{обучаемый параметр}$$

1) Оптимизационная задача для выделения фрагментов:

$$CE(S, \hat{S}) \to \min_{BERT, W^F}$$

От структуры с элементами разметки перейдем к графу. Матрица смежности такого графа: $A_{i,j} = 1$, если i, j фрагменты находятся в одном элементе разметки. Допустим, известны фрагменты $\phi_i = \{E_j\}_{j \in I(i)}$. Введём преобразование, которое по эмбеддинам слов строит эмбеддинг фрагмента:

$$F^e : \mathbb{R}^{N \times d} \to \mathbb{R}^d \qquad N - \text{произвольное}$$

$$f_i = F^e(\phi_i) - \text{эмбеддинг i-го фрагмента}$$

Введём близость фрагментов:

$$s : \mathbb{R}^{2 \times d} \to [0,1] \qquad \hat{A}_{i,j} = s(f_i, f_j)$$

$s(f_i, f_j) = \sigma(\alpha f_i^T (W + W^T) f_j)$
Где α, W - обучаемые параметры

2) Оптимизационная задача для объединения фрагментов в элемент разметки:

$$BCE(A, \hat{A}) \to \min_{BERT, F^e, s}$$

Задача тегирования – задача multilabel classification. $T-$ множество тегов. Для известных фрагментов находим эмбеддинги ($f \in \mathbb{R}^{|f| \times d}$) и для них проводим классификацию:

$$\hat{Y} = \sigma(fW^{ft}) \qquad W^{ft} \in [0,1]^{d \times |T|} - \text{обучаемый параметр}$$

$$Y \in \{0,1\}^{|f| \times |T|} - \text{ground truth}$$

$Y_{i,j} = 1$, если фрагмент f_i имеет тег T_j. 3) Оптимизационная задача для тегирования фрагментов:

$$BCE(Y, \hat{Y}) \to \min_{BERT, F^e, W^{ft}}$$

Для аналогичного решения достаточно найти эмбеддинг элемента разметки.
Эмбеддинг элемента разметки можно найти по эмбеддингам фрагментов:

$$e = E^e(elem) - \text{эмбеддинг элемента разметки}$$

Где $elem = \{f\}$ - множество эмбеддингов фрагментов, находящихся в элементе разметки.

$$\hat{Z} = \sigma(eW^{et}) \qquad W^{et} \in [0,1]^{d \times |T|} - \text{обучаемый параметр}$$

$$Z \in \{0,1\}^{|e| \times |T|} - \text{ground truth}$$

$Z_{i,j} = 1$, если элемент разметки e_i имеет тег T_j.

4) Оптимизационная задача для тегирования элементов разметки:

$$BCE(Z, \hat{Z}) \to \min_{BERT, F^e, E^e, W^{et}}$$

4 Данные

Данные состоят из текстов новостных статей статей. Для каждого текста выделяются фрагменты, по которым можно определить каких ценностей придерживается автор. Соответсвующие ценности указываются в тегах. Фрагменты объединяются в элементы разметки. Фрагменты в элементах разметки связаны семантически, связь указывается в тегах (рис. 1). Всего 784 документов, 2746 разметок, 7620 элементов разметки, 9617 фрагментов.

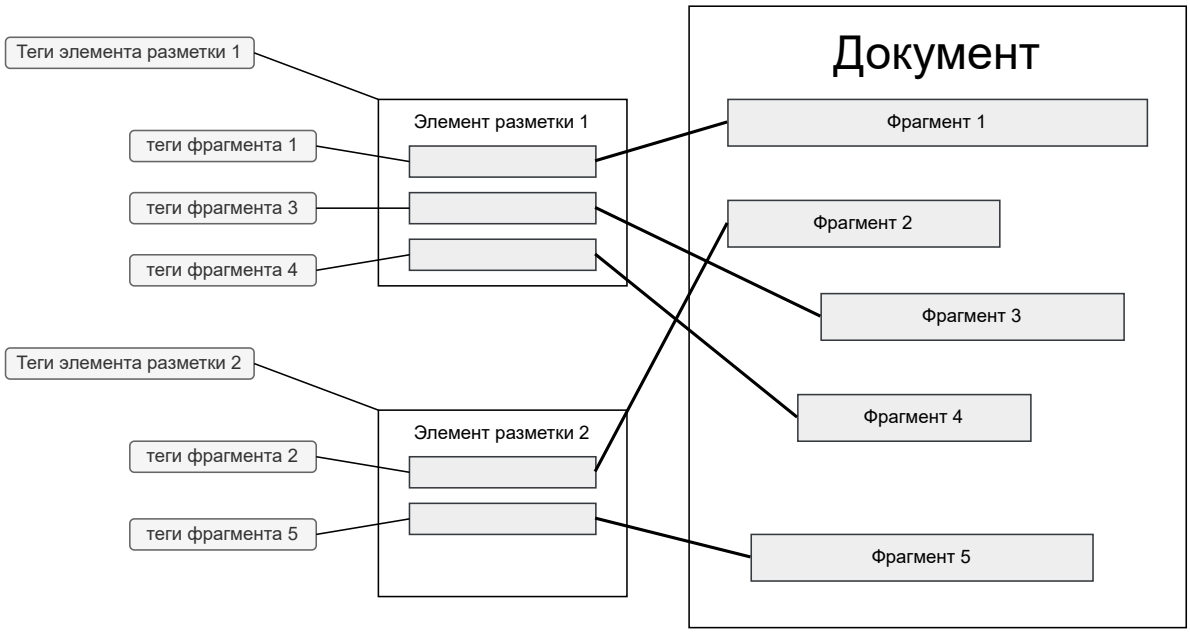

Рис. 1: Общая схема разметки документа

5 Критерии качества модели

Для оценки качества модели для выделения тегированных фрагментов во многих работах используется $F_1 score$ с микро усреднением. По всей выборке для каждого тега считается матрица ошибок, усредняются по тегам и считается среднее гармоническое полученных точности и полноты:

$$TP_t = \sum_{y,\hat{y}} \sum_{i=0}^{|X|} [y_{t,i} = 1][\hat{y}_{t,i} = 1] \qquad FP_t = \sum_{y,\hat{y}} \sum_{i=0}^{|X|} [y_{t,i} = 0][\hat{y}_{t,i} = 1]$$

$$FN_t = \sum_{y,\hat{y}} \sum_{i=0}^{|X|} [y_{t,i} = 1][\hat{y}_{t,i} = 0]$$

$$TP = \frac{1}{|T|} \sum_{t \in T} TP_t \qquad FP = \frac{1}{|T|} \sum_{t \in T} FP_t \qquad FN = \frac{1}{|T|} \sum_{t \in T} FN_t$$

$$F_1 = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}$$

Где пара $(y, \hat{y})$ – истинная разметка и предсказание модели.

$y_{t,i} = 1 \Rightarrow$ i-е слово принадлежит фрагменту, помеченному тегом t.

6 Эксперименты

7 Нахождение фрагментов

Решается только задача поиска фрагментов. Для оценки качества используется $F_1 score$ с микро усреднением, где в качестве тегов используется {0, 1, 2} (BIO разметка). Результаты обучения представлены на графиках (рис. 2, 3, 4). Параметры обучения: $lr = 2 \cdot 10^{-5}$, cosine scheduler, $batch\_size = 4$.

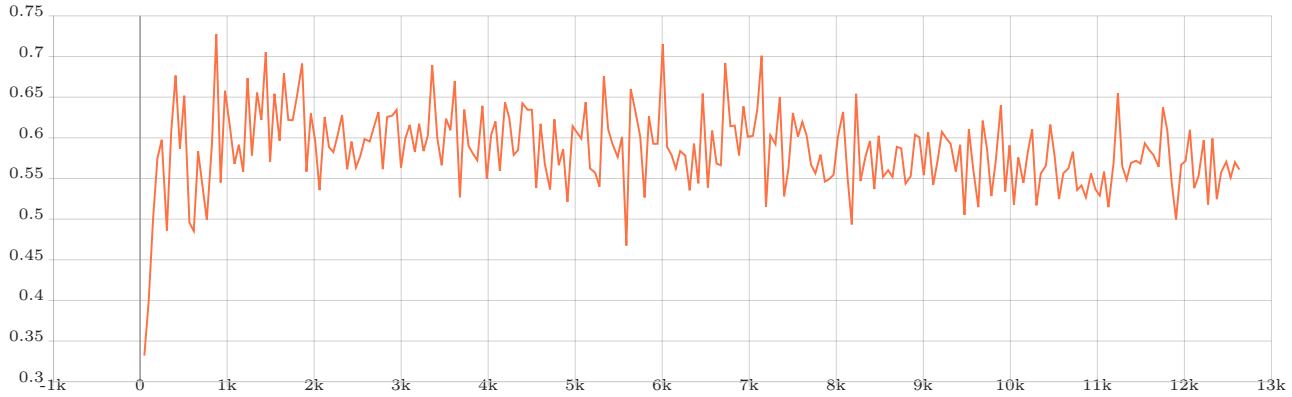

Рис. 2: Величина precision в зависимости от итерации

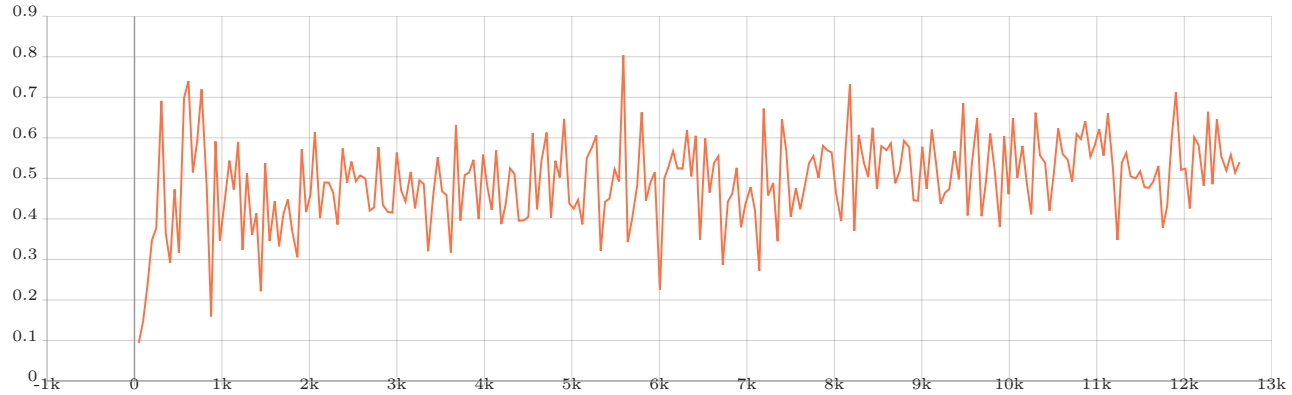

Рис. 3: Величина recall в зависимости от итерации

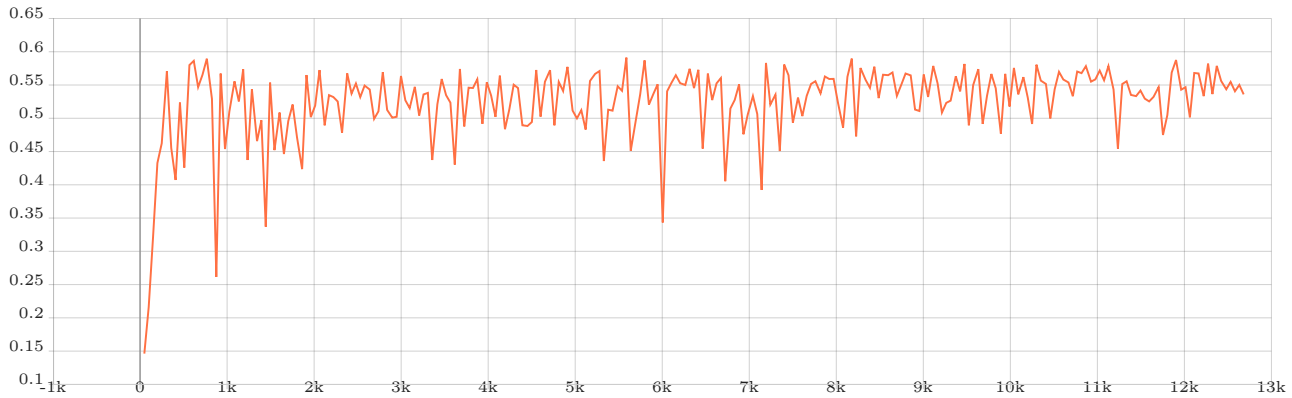

Рис. 4: Величина $F_1 score$ в зависимости от итерации

Для константного предсказания (весь текст выделяется как фрагмент) $precision = 0.346$, $F_1 score = 0.514$. В результате обучения *precision* выходит на уровень 0.55-0.6, а *recall* – 0.5-0.6, в результате $F_1 score$ – 0.55.

Список литературы

Diana Mutz and Seth Goldman. Mass media. 2010.

G Martino, Alberto Barrón-Cedeno, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. Semeval-2020 task 11: Detection of propaganda techniques in news articles. arXiv preprint arXiv:2009.02696, 2020.

Sigi Kemahduta, Sari Widya Sihwi, and Wisnu Widiarto. Automatic text summarization with categorization on online news about indonesian public figures using fuzzy logic method. In 2021 International Conference on Artificial Intelligence and Big Data Analytics, pages 10–15. IEEE, 2021.

Carlo Strapparava and Rada Mihalcea. Semeval-2007 task 14: Affective text. In Proceedings of the fourth international workshop on semantic evaluations (SemEval-2007), pages 70–74, 2007.

Anoop K., Deepak P., Savitha Sam Abraham, Lajish V. L., and Manjary P. Gangan. Readers' affect: predicting and understanding readers' emotions with deep learning. Journal of Big Data, 9(1):82, Jun 2022. ISSN 2196-1115. doi:10.1186/s40537-022-00614-2. URL `https://doi.org/10.1186/s40537-022-00614-2`.

Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. We can detect your bias: Predicting the political ideology of news articles. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 4982–4991, Online, November 2020a. Association for Computational Linguistics. doi:10.18653/v1/2020.emnlp-main.404. URL `https://aclanthology.org/2020.emnlp-main.404`.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

Sean Papay, Roman Klinger, and Sebastian Padó. Dissecting span identification tasks with performance prediction, 2020.

Shubham Toshniwal, Haoyue Shi, Bowen Shi, Lingyu Gao, Karen Livescu, and Kevin Gimpel. A cross-task analysis of text span representations, 2020.

Mengyun Chen, Tao Ge, Xingxing Zhang, Furu Wei, and Ming Zhou. Improving the efficiency of grammatical error correction with erroneous span detection and correction. arXiv preprint arXiv:2010.03260, 2020.

Chak Yan Yeung and John SY Lee. Automatic detection of sentence fragments. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 599–603, 2015.

Shutian Ma, Jin Xu, and Chengzhi Zhang. Automatic identification of cited text spans: a multi-classifier approach over imbalanced dataset. Scientometrics, 116(2):1303–1330, Aug 2018. ISSN 1588-2861. doi:10.1007/s11192-018-2754-2. URL `https://doi.org/10.1007/s11192-018-2754-2`.

Moreno La Quatra, Luca Cagliero, Elena Baralis, et al. Poli2sum@ cl-scisumm-19: Identify, classify, and summarize cited text spans by means of ensembles of supervised models. BIRNDL@ SIGIR, 2414:233–246, 2019.

Shifeng Liu, Yifang Sun, Bing Li, Wei Wang, Florence T Bourgeois, and Adam G Dunn. Sent2span: span detection for pico extraction in the biomedical text without span annotations. arXiv preprint arXiv:2109.02254, 2021.

Liying Cheng, Lidong Bing, Qian Yu, Wei Lu, and Luo Si. Ape: Argument pair extraction from peer review and rebuttal via multi-task learning. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 7000–7011, 2020.

Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. A unified mrc framework for named entity recognition. arXiv preprint arXiv:1910.11476, 2019.

Matías Rojas, Felipe Bravo-Marquez, and Jocelyn Dunstan. Simple yet powerful: An overlooked architecture for nested named entity recognition. In Proceedings of the 29th International Conference on Computational Linguistics, pages 2108–2117, 2022.

John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. Semeval-2021 task 5: Toxic spans detection. In Proceedings of the 15th international workshop on semantic evaluation (SemEval-2021), pages 59–69, 2021.

Manikandan Ravikiran, Bharathi Raja Chakravarthi, Anand Kumar Madasamy, Sangeetha Sivanesan, Ratnavel Rajalakshmi, Sajeetha Thavareesan, Rahul Ponnusamy, and Shankar Mahadevan. Findings of the shared task on offensive span identification from code-mixed tamil-english comments, 2022.

Weiwen Xu, Xin Li, Yang Deng, Wai Lam, and Lidong Bing. Peerda: Data augmentation via modeling peer relation for span identification tasks, 2023.

Dawid Jurkiewicz, Łukasz Borchmann, Izabela Kosmala, and Filip Graliński. Applicaai at semeval-2020 task 11: On roberta-crf, span cls and whether self-training helps them. arXiv preprint arXiv:2005.07934, 2020.

Dimas Sony Dewantara, Indra Budi, and Muhammad Okky Ibrohim. 3218IR at SemEval-2020 task 11: Conv1D and word embedding in propaganda span identification at news articles. In Proceedings of the Fourteenth Workshop on Semantic Evaluation, pages 1716–1721, Barcelona (online), December 2020. International Committee for Computational Linguistics. doi:10.18653/v1/2020.semeval-1.225. URL `https://aclanthology.org/2020.semeval-1.225`.

Nasim Nouri. Data augmentation with dual training for offensive span detection. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2569–2575, 2022.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. Spanbert: Improving pre-training by representing and predicting spans. Transactions of the association for computational linguistics, 8:64–77, 2020.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. Gnner: Reducing overlapping in span-based ner using graph neural networks. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pages 97–103, 2022.

Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. We can detect your bias: Predicting the political ideology of news articles. arXiv preprint arXiv:2010.05338, 2020b.

Dmitry Zmitrovich, Alexander Abramov, Andrey Kalmykov, Maria Tikhonova, Ekaterina Taktasheva, Danil Astafurov, Mark Baushenko, Artem Snegirev, Tatiana Shavrina, Sergey Markov, Vladislav Mikhailov, and Alena Fenogenova. A family of pretrained transformer language models for russian, 2023.

Yuri Kuratov and Mikhail Arkhipov. Adaptation of deep bidirectional multilingual transformers for russian language, 2019.

Qipeng Guo, Zhijing Jin, Xipeng Qiu, Weinan Zhang, David Wipf, and Zheng Zhang. Cyclegt: Unsupervised graph-to-text and text-to-graph generation via cycle training, 2020.