Московский государственный университет имени М. В. Ломоносова
факультет вычислительной математики и кибернетики
кафедра математических методов прогнозирования

Курсовая работа

Выявление пропаганды в текстах с помощью моделей архитектуры BERT.

Detection of propaganda in texts using BERT architecture.

Выполнил: студент 317 группы
Воробьев Сергей Юрьевич

Научный руководитель: д.ф.-м.н.
Воронцов Константин Вячеславович

Москва 2022

Аннотация

В наши дни очень актуальной является задача автоматического выявления пропаганды и манипуляции в новостных и не только текстах. Я рассматриваю способ выделения конкретных фрагментов текста, содержащих пропаганду, с помощью модели `BERT`. В частности, будет показано, что модель `spanBERT` лучше справляется с этой задачей, а также, что нет ощутимой разницы между `BERT-base` и `BERT-large`.

Содержание

1 Введение

1.1 Актуальность проблемы

Задача выявления пропаганды в тексте стала очень актуальной в наши дни. Из-за существенного упрощения доступа к интернету, развития социальных сетей и медиа-платформ, любой человек может высказывать мнение, охватывая очень большую аудиторию. В связи с этим может возникать соблазн манипулировать мнением большого числа людей и продвигать выгодную автору повестку. Информация распространяется очень быстро и это сильно усложняет борьбу с пропагандой и манипуляцией стандартными средствами. В связи с этим возникает задача автоматического определения пропаганды в тексте. В обзоре Estela Saquete et al., (2020)[1] очень подробно описана проблема борьбы с постправдой, в том числе манипуляцией и пропагандой.

Анализу текстов посвящена целая область машинного обучения – NLP. Она активно развивается последние несколько лет и в ней уже достигнуты значительные результаты. Для нашей задачи хорошо подойдут нейросетевые модели, основанные на механизме внимания[2]. Задача выявления пропаганды может ставиться по-разному. В самом классическом случае нужно классифицировать новостную статью, абзац или предложение. Такая задача обычно решается получением векторного представления всего текста (абзаца или предложения) и его дальнейшая классификация. Можно поставить более сложную задачу классификации на уровне слов. В этом случае задача усложняется, поскольку, по одному слову сложно судить о какой-либо манипуляции или пропаганде и приходится обращать внимание на контекст этого слова. Именно такую задачу я рассмотрел в своей работе.

1.2 Существующие решения

Задача выявления пропаганды на уровне текстов и предложений является обычной задачей классификации и может решаться базовыми средствами, например логистической регрессией с TF-IDF. Можно использовать нейронные сети или даже ансамбли сложных и простых моделей. В статье Pankaj Gupta et al., (2019)[3] используется ансамбль разных моделей и эмбеддингов для задачи классификации предложений. Задача классификации отдельных токенов решается реже, поскольку является не такой очевидной. Я буду опираться на статью Da San Martino et al., (2019)[4]. В ней предложена модель, которая использует дополнительно эмбеддинг предложения для классификации каждого токена.

В своей работе я решил изучить базовую задачу классификации токенов с помощью `BERT`. Посмотрел, как гиперпараметры влияют на обучение, также сравнил модели с разным числом весов и попробовал `spanBERT`, который изначально обучался моделировать целые фрагменты текста, а не отдельные токены.

2 Данные

Я взял данные из соревнования NLP4IF на конференции EMNLP-IJCNLP 2019 [4]. Это набор новостных статей, для которых размечены фрагменты, содержащие пропаганду, а также, указан соответствующий класс. Всего в данных присутствуют 18 классов. Все они представлены ниже:

1. Loaded language. Использование слов/фраз с сильным эмоциональным подтекстом (положительным или отрицательным) для воздействия на аудиторию. Например: «. . . a lone lawmaker's childish shouting.»
2. Name calling or labeling. Обозначение объекта пропаганды как чего-то, что целевая аудитория боится, ненавидит, считает нежелательным или иным образом любит или восхваляет. Например: «Republican congressweasels», «Bush the Lesser.»
3. Repetition. Повторение одного и того же сообщения снова и снова, чтобы аудитория в конечном итоге приняла его.
4. Exaggeration or minimization. Представлять что-то в чрезмерной манере: делать вещи больше, лучше, хуже. Например: «the best of the best», «quality guaranteed»

5. Doubt. Сомнение в компетенции/достоверности кого-либо или чего-либо. Например: «Is he ready to be the Mayor?»
6. Appeal to fear/prejudice. Стремление заручиться поддержкой идеи, вызывая у населения тревогу и/или панику в отношении альтернативы, возможно, на основе предвзятых суждений. Например: «stop those refugees; they are terrorists.»
7. Flag-waving. Игра на сильных национальных чувствах (или по отношению к группе, например, расе, полу, политическим предпочтениям) для оправдания или продвижения действия или идеи. Например: «entering this war will make us have a better future in our country.»
8. Causal oversimplification. Предполагать одну причину, когда существует несколько причин, стоящих за проблемой. Перекладывание вины на одного человека или группу людей без изучения сложности вопроса. Например: «If France had not declared war on Germany, World War II would have never happened.»
9. Slogans. Короткая и эффектная фраза, которая может включать ярлыки и стереотипы. Лозунги, как правило, действуют как эмоциональные призывы. Например: «Make America great again!»
10. Appeal to authority. Заявление о том, что утверждение верно просто потому, что его поддерживает некий авторитет или эксперт по данному вопросу, без каких-либо других подтверждающих доказательств.
11. Black-and-white fallacy, dictatorship. Представление двух вариантов как единственно возможных, хотя на самом деле существует больше альтернатив. Например: «There is no alternative to war.»
12. Thought-terminating cliche. Слова или фразы, препятствующие осмысленному обсуждению данной темы. Как правило, это короткие общие предложения, предлагающие простые ответы на сложные вопросы или отвлекающие внимание от других направлений мысли. Например: «it is what it is», «you cannot judge it without experiencing it», «it's common sense», «nothing is permanent except change», «better late than never» и так далее.
13. Whataboutism. Дискредитация позиции оппонента, обвиняя его в лицемерии, не опровергая напрямую его аргумент. Например, упоминание о событии, дискредитирующем оппонента: «What about…?»
14. Reductio ad Hitlerum. Убеждение аудитории не одобрять действие или идею, предполагая, что эта идея популярна среди групп, которые целевая аудитория презирает. Это может относиться к любому человеку или понятию с негативной коннотацией. Например: «Only one kind of person can think this way: a communist.»
15. Red herring. Внесение не относящегося к обсуждаемому вопросу материала, чтобы отвлечь внимание всех от обсуждаемых моментов.
16. Bandwagon. Попытка убедить целевую аудиторию присоединиться и принять желаемый курс действий, потому что «все остальные делают то же самое». Например: «Would you vote for Clinton as president? 57% say yes.»
17. Obfuscation, intentional vagueness, confusion. Намеренное использование неясных слов, чтобы у аудитории была своя интерпретация.
18. Straw man. Когда предложение оппонента заменяется аналогичным, которое затем опровергается вместо исходного

Тексты новостных статей разбиваются на предложения, и в итоге мы получаем 14808 предложений в тренировочной выборке и 2022 в валидационной.

3 Постановка задачи

3.1 Метрика

Наша задача заключается в правильной классификации фрагментов текста. Очевидно, что стандартные метрики здесь не подойдут, так как классы могут накладываться друг на друга, а классифкация может выделять фрагмент не целиком, но при этом оставаться качественной. Авторы статьи предложили следующую метрику. Пусть документ d представлен последовательностью токенов. Тогда фрагмент с пропагандой можно обозначить $t = [t_i \dots t_j]$. Документ включает некоторое множество T таких фрагментов. Наш алгоритм порождает множество $s = [s_m \dots s_n]$ для документа d. Функция $l(x) \in \{1, \dots, 18\}$

ставит в соответствие каждому $s \in S$ один из 18 классов.

Введем следующую функцию:

$$C(s,t,h) = \frac{|(s \cap t)|}{h}\delta(l(s), l(t)) \tag{1}$$

которая поможет нам оценивать пересекающиеся области. Здесь h – нормализующий множитель, $\delta(a,b) = 1$, если $a = b$, иначе 0.
Далее определим понятные нам точность и полноту:

$$P(S,T) = \frac{1}{|S|}\sum_{s \in S t \in T} C(s,t,|s|), \tag{2}$$

$$R(S,T) = \frac{1}{|T|}\sum_{s \in S t \in T} C(s,t,|t|), \tag{3}$$

Теперь мы можем посчитать F_1-меру.

3.2 Задача

В нашем случае нас интересуют точность и полнота одновременно, поэтому критерием качества возьмем F_1-меру, их среднее гармоническое. Поставим задачу выявления и классификации фрагментов текста. При этом, если токен отмечен как пропаганда, но класс предсказан неверно, то мы не будем засчитывать такой ответ как правильный. Теперь у нас есть метрика качества и данные – будем считать задачу поставленной.

4 Модель

Для классификации отдельных токенов я буду пользоваться моделью BERT[5]. Эта модель хорошо подходит под нашу задачу, поскольку позволяет получать контекстно-зависимые эмбеддинги слов благодаря механизму attention[6]. Рассмотрим схематично архитектуру модели BERT[7]:

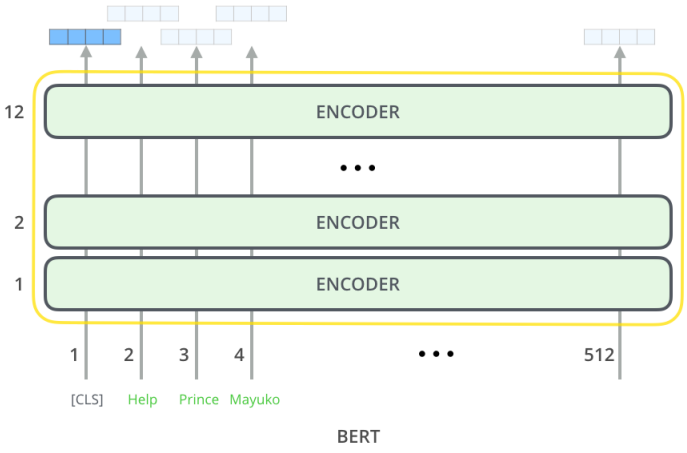

Рис. 1: Архитектура BERT.

На схеме видно, что BERT получает эмбеддинги для каждого слова, и они все одинакового размера. Соответственно, мы можем решать задачу классификации для каждого токена с помощью обычного линейного слоя. Веса классификатора будут одинаковыми для всех представлений. Также, в качестве альтернативы выходным эмбеддингам, можно взять представление с нескольких слоев и произвести классификацию на их основе, как предлагается в [7]. Дополнительно, можно рассмотреть другие модели на основе BERT. В экспериментах я покажу результаты работы spanBERT[8].

5 Эксперименты

5.1 Переобучение

Начнем с базового эксперимента. Посмотрим, как ведет себя F_1-мера на тренировочной и валидационной выборках. Иными словами, попытаемся понять, страдает ли наша сеть от переобучения.

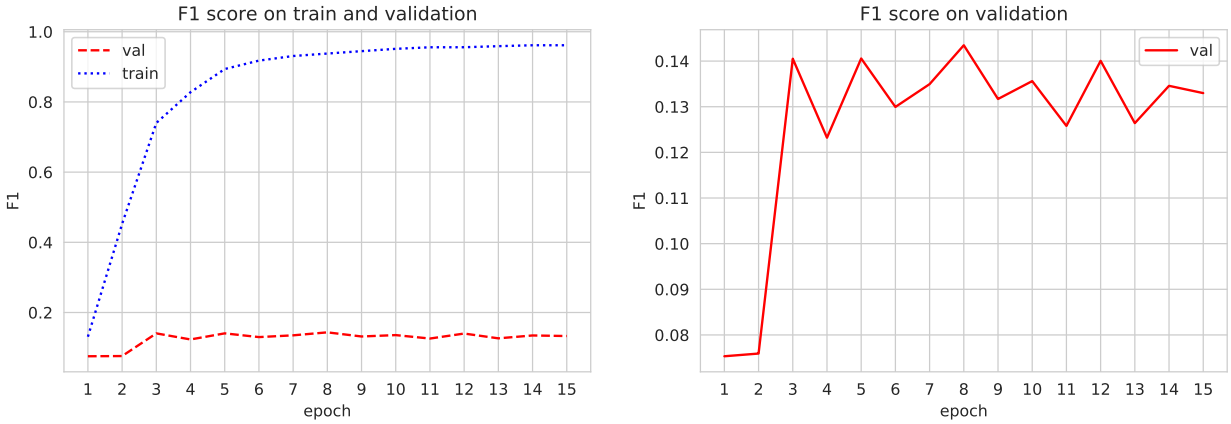

Рис. 2: Качество на train и val.

На Рис. 2 видно, что модель очень хорошо подстраивается под обучающую выборку. При этом, у нас не происходит существенного падения качества на валидационной выборке, поэтому, мы можем считать, что подстраивание под обучающую выборку не мешает нам получать хорошую модель. Обратим внимание, что 10 эпох вполне хватает для дообучения на наших данных.

5.2 Размер батча

Далее посмотрим, как размер батча влияет на качество.

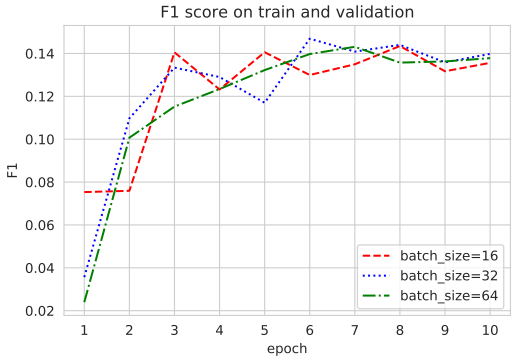

Рис. 3: Сравнение размера батча.

Судя по графику, существенной разницы в качестве нет. Стоит отметить, что увеличение размера батча замедляет обучение, но в итоге результаты выравниваются.

5.3 Темп обучения

Далее посмотрим, как темп обучения влияет на качество.

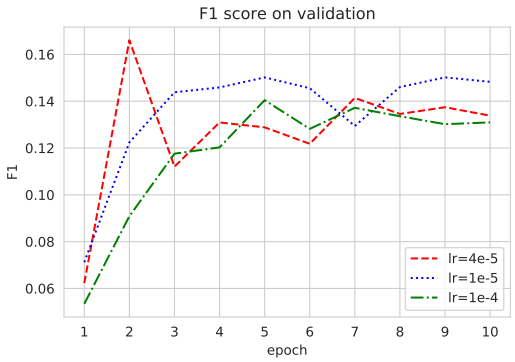

Рис. 4: Сравнение темпа обучения.

Судя по графику, более низкий темп обучения в среднем справляется немного лучше.

5.4 BERT-base vs BERT-large

У модели `BERT` есть две разновидности. В «большом» варианте существенно больше весов (340 млн против 110)[5]. Сравним `BERT-base` и `BERT-large` между собой. Поскольку модель `BERT-large` содержит гораздо больше параметров, можно предположить, что она позволит получать более качественную классификацию.

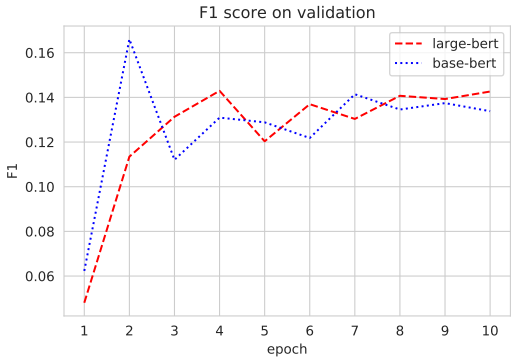

Рис. 5: Сравнение bert-base и bert-large

Однако, на Рис. 5 мы не видим существенной разницы. Отсюда можно сделать предположение, что базовой модели хватает параметров, чтобы получать сопоставимые по информативности эмбеддинги для нашей задачи.

5.5 spanBERT

`spanBERT` представляет из себя обычную модель `BERT`, которая обучалась по другому принципу. Вместо маскирования отдельных слов, при обучении `spanBERT` маскируются целые куски текста, что позволяет лучше моделировать участки в тексте. Это должно помочь в нашей задаче.

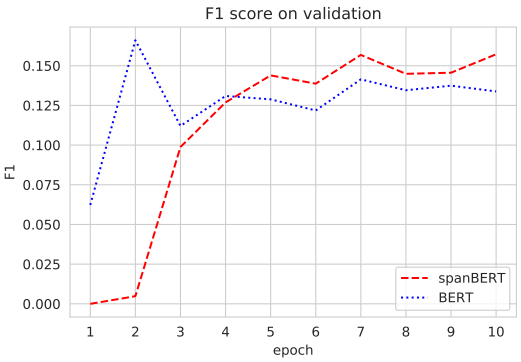

Рис. 6: Сравнение bert и span-bert

Действительно, мы видим, что качество у `spanBERT` немного превосходит качество обычного `BERT`. Из этого мы можем сделать вывод о том, что процедура обучения модели играет важную роль в ее способности решать задачу.

5.6 Визуализация работы модели

Теперь визуализируем предсказания нашей модели. Зеленым я выделил те слова, для которых модель верно предсказала класс, оранжевым те, для которых модель неверно предсказала класс, но поняла, что это является манипуляцией и красным те, на которых модель полностью ошиблась.

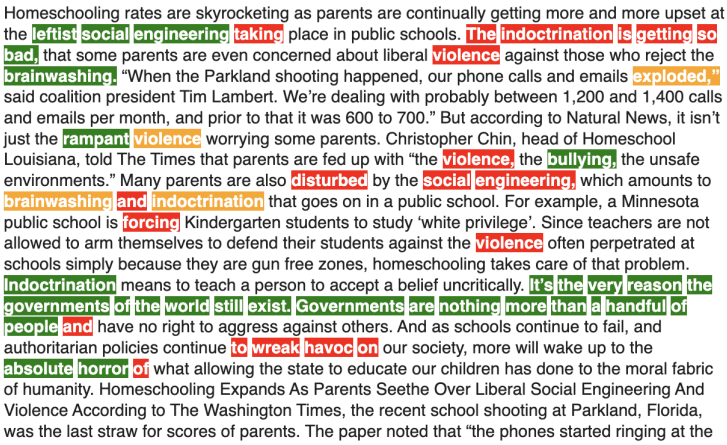

Рис. 7: Результат работы модели.

Разберем пару примеров на Рис. 7. В первом фрагменте модель почему-то отнесла к классу **Name calling or labeling** глагол «taking», хотя грамматически это неверно. Во втором фрагменте она выделила всю фразу в класс **Loaded language**, хотя правильным тут является только слов «bad». Также модель не смогла понять, что слово «violence» повторяется в тексте, что тоже является манипуляцией.

Рассмотрим еще несколько примеров. Ниже представлен текст, на котором модель довольно плохо справилась с задачей:

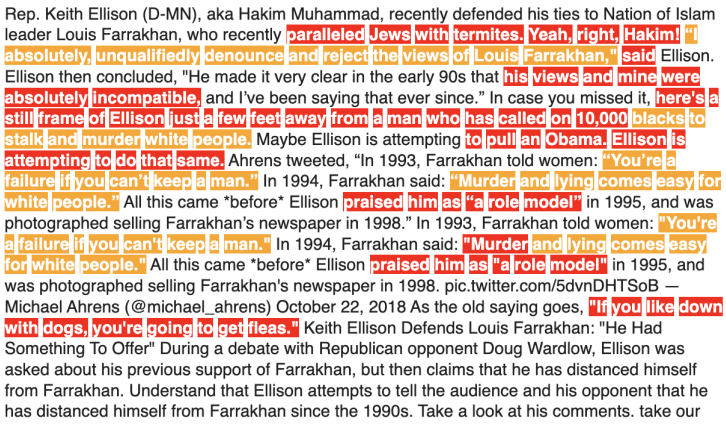

Рис. 8: Результат работы модели.

В целом модель около половины токенов верно отнесла к пропаганде, однако не смогла предсказать точно их класс. Так, например, фразу «Farrakhan said: Murder and lying comes easy for white people.», модель ошибочно отнесла к классу **Appeal to fear/prejudice**, хотя верный ответ тут **Exaggeration or minimization**. Похожая ошибка в предложении «"I absolutely, unqualifiedly denounce and reject the views of Louis Farrakhan,"said Ellison.» Вместо **Loaded language** алгоритм предсказал **Appeal to authority**. Это говорит о том, что алгоритм все еще не очень хорошо понимает текст и в этом направлении нужно работать.

6 Выводы

В данной работе я рассмотрел задачу выявления и классификации фрагментов текста, содержащих пропаганду. В частности, было показано, что:

- Модель хорошо запоминает тренировочною выборку, но это не влияет на качество на валидационной.
- Гиперпараметры могут влиять на качество обучения, поэтому лучше исследовать их для каждой конкретной задачи.
- Количество параметров в модели не сильно влияет на качество. Это может быть связано с тем, что модели хватает параметров для получения хороших представлений.
- Способ обучения модели сильно влияет на качество. `spanBERT` в среднем показал качество выше, так как обучался на похожей задаче.

Список литературы

[1] Estela Saquete, David Tomás, Paloma Moreda, Patricio Martínez-Barco, and Manuel Palomar. Fighting post-truth using natural language processing: A review and open challenges. *Expert Systems with Applications*, 141:112943, 2020.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[3] Pankaj Gupta, Khushbu Saxena, Usama Yaseen, Thomas Runkler, and Hinrich Schutze. Neural architectures for fine-grained propaganda detection in news. *arXiv:1909.06162v1*, 2019.

[4] Giovanni Da San Martino, Seunghak Yu, Alberto Barron-Cedeno, Rostislav Petrov, and Preslav Nakov. Fine-grained analysis of propaganda in news articles. *In Proceedings of the 2019 Con- ference on Empirical Methods in Natural Language Processing and the 9th International Joint Confer- ence on Natural Language Processing, EMNLP- IJCNLP 2019, Hong Kong, China.*, 2019.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. 2017.

[7] Jay Alammar. The illustrated transformer. https://jalammar.github.io/illustrated-transformer/.

[8] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. SpanBERT: Improving pre-training by representing and predicting spans. *arXiv preprint arXiv:1907.10529*, 2019.