УДК 004.89 **Информационные технологии**

Батурин Михаил Михайлович, *студент-магистр, Калужский филиал ФГБОУ ВО «Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)»*

Федоров Виктор Олегович, *к.т.н., доцент, Калужский филиал ФГБОУ ВО «Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)»*

ОПРЕДЕЛЕНИЕ АВТОРСТВА КОРОТКИХ ТЕКСТОВ НА ОСНОВЕ ТРИПЛЕТНОЙ ФУНКЦИИ ПОТЕРЬ И НЕСКОЛЬКИХ ВИДОВ ВЛОЖЕНИЙ

Аннотация: Атрибуция авторства (АА), представляющая собой задачу поиска автора определённого текста, является важной и широко освещаемой проблемой. Методы глубокого обучения могут показывать высокую точность в задаче АА, тем не менее, большинство из этих методов не показывают высоких результатов на выборках данных с малой длиной текстов и большим количеством авторов. В этой статье предлагается новая модель, основанная на вложениях, изучающая представления характерных стилей письма при помощи нескольких видов вложений. Эксперименты, показывающие точность модели получены на реальном наборе данных коротких текстов. Результаты эксперимента показывают, что предложенная модель превосходит другие современные решения в задаче определения авторства текста.

Ключевые слова: Определение авторства текста, машинное обучение, вложения слов, POS теги, обработка естественного языка (NLP).

Annotation: Attribution of authorship (AA), which is the task of finding the author of a certain text, is an important and widely publicized problem. Deep learning methods can show high accuracy in the AA task, however, most of these methods do

not show high results on data samples with a small length of texts and a large number of authors. This article proposes a new embedding-based model that studies representations of characteristic writing styles using several types of embeddings. Experiments showing the accuracy of the model were obtained on a real data set of short texts. The results of the experiment show that the proposed model is superior to other modern solutions in the problem of determining the authorship of the text.

Keywords: Text authorship determination, machine learning, word embeddings, POS tags, natural language processing (NLP).

Введение

Задача поиска автора конкретного текста играет жизненно важную роль во многих приложениях. В частности, в контексте социальных сетей атрибуция авторства (AA) имеет решающее значение для решения проблемы распространения ложной информации и сбора доказательств в ходе судебных расследований. Несмотря на то, что задача AA была широко изучена, и многие ее функции были исследованы, традиционные методы, применяемые к большим документам, не очень хорошо работают с сообщениями в социальных сетях, поскольку они, как правило, короче и менее формальны.

Чтобы преодолеть проблему с короткими текстами, несколько сообщений в социальных сетях от одного и того же пользователя можно объединить в один длинный документ, однако такой подход неэффективен при AA для отдельного сообщения в социальной сети, что необходимо для некоторых приложений. Предложенная модель использует несколько представлений текста, чтобы преодолеть разреженность данных при определении стилей письма автора короткого текста.

Модуль встраивания постов

Пусть U — множество пользователей социальной сети, $U = \{u_1, u_2, \dots, u_N\}$, где N — общее количество пользователей платформы. У каждого пользователя u есть набор постов, $P_u = \{p_{u,1}, p_{u,2}, \dots, p_{u,M_u}\}$, где M_u— количество постов, принадлежащих пользователю u. Имея U, P_u и сообщение с запросом q, мы

стремимся предсказать наиболее вероятного автора q.

Классический подход к обучению классификатора для определения авторства коротких постов заключается в том, что для обучения используются размеченные посты всех пользователей. Предложенный в данной статье подход отличен от классического и заключается в изучении встраивания сообщений пользователя u с сообщениями M_u. Символ $e_{u,i}$ обозначает представление поста $p_{u,i}$. Полученные после обучения вложения постов $E_u = \{e_{u,1}, e_{u,2}, \dots, e_{u,M_u}\}$, затем объединяются для формирования встраивания пользовательского стиля s_u. Наконец, для каждого пользователя вычисляется u косинусное сходство между постом и запросом, встраивающим e^q и s_u, обозначаемое как $cos(e^q, s_u)$. Модель предсказывает, кто является наиболее вероятным автором сообщения с запросом q.

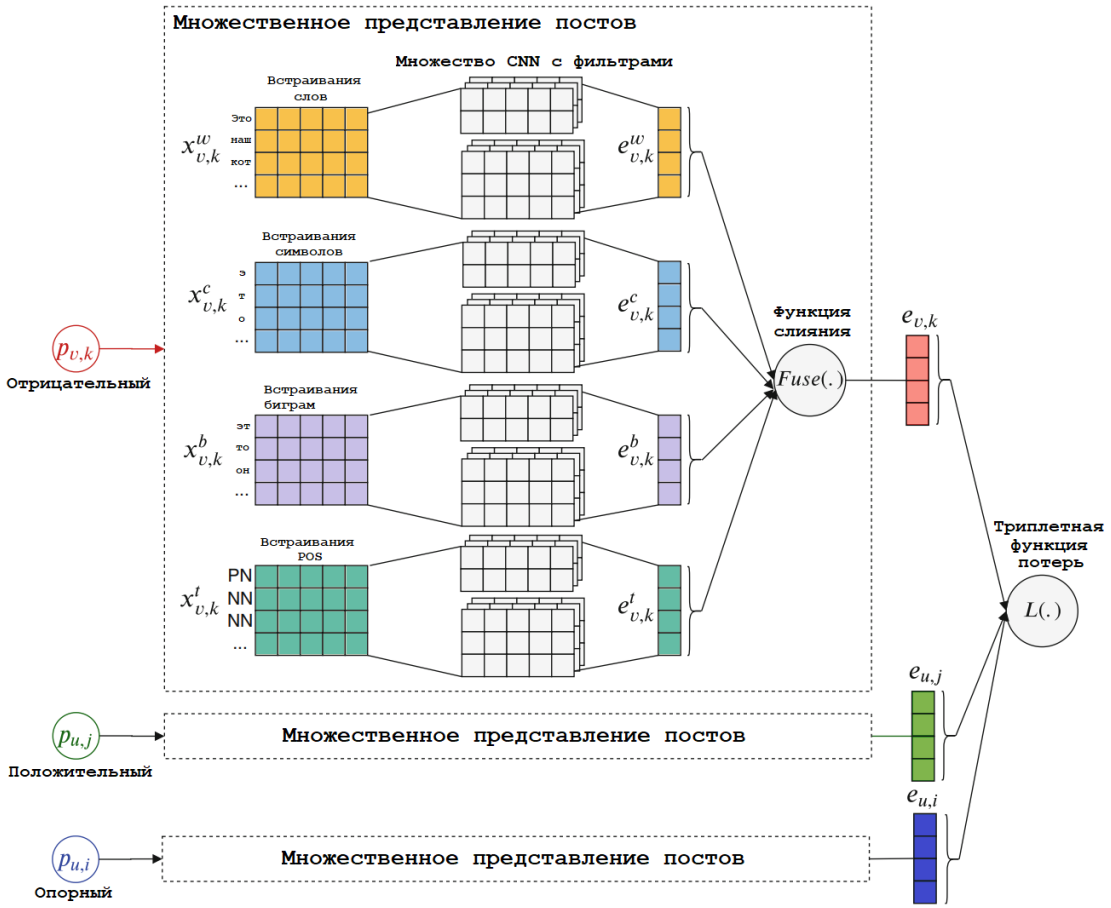

Рис. 1. Схема модуля встраивания постов

Рисунок 1 иллюстрирует общую архитектуру модуля встраивания поста. Модуль изучает представление встраивания, используя триплетную функцию

потерь. Каждый триплет формируется путем случайной выборки якорного поста $p_{u,i}$ и положительного поста $p_{u,j}$, принадлежащих одному и тому же пользователю u. Также случайным образом выбирается отрицательный пост $p_{v,k}$, принадлежащий другому пользователю v. В этой схеме сначала выполняется предварительная обработка данных, чтобы извлечь различные типы признаков для каждого поста среди триплетов. Эти вложения будут использоваться в качестве входных данных для набора сверточных нейронных сетей (CNN). Впоследствии функция слияния используется для вывода представления поста в виде его встраивания. Наконец, три встраивания, $e_{u,i}$, $e_{u,j}$ и $e_{v,k}$, оптимизируются с использованием триплетной функции потерь.

Извлечение вложений признаков

В предыдущих исследованиях [1, 2] широко изучались различные типы признаков, такие как слова, символы, n-граммы, синтаксическая и семантическая информация. В представленной работе результаты предыдущих исследований используются, чтобы извлечь вложения функций для представления поста $p_{u,i}, p_{u,j}$ или $p_{v,k}$. Для простоты используем $p_{v,k}$ в качестве примера.

Для четырех типов входных вложений $\{x_{v,k}^w, x_{v,k}^c, x_{v,k}^b, x_{v,k}^t\}$ получаем четыре типа выходных скрытых представлений через многоуровневый слой CNN, т. е. словесное скрытое представление $e_{v,k}^w$, скрытое представление символа $e_{v,k}^c$, скрытое представление биграммы $e_{v,k}^b$ и скрытое представление тега POS $e_{v,k}^t$. После получения скрытых представления различных типов вложений функций, применяется функция слияния (*Fuse*), объединяющая различные скрытые представления поста:

$$e_{v,k} = Fuse(e_{v,k}^w, e_{v,k}^c, e_{v,k}^b, e_{v,k}^t) \qquad (1)$$

Модуль агрегации стилей пользователя

Существует несколько способов комбинирования вложений сообщений пользователя v для формирования представления встраивания пользовательского стиля S_v. Используя агрегаторы типов *Mean, Max и Capsule* [3] вложения сообщений пользователей объединяются, чтобы сформировать

вложения пользовательского стиля, $S = \{s_1, s_2, \dots, s_N\}$, которые будут использоваться для выполнения операции определения авторства.

Эксперимент проводится на общедоступном наборе данных из социальной сети. Набор содержит посты от 7026 пользователей, каждый пользователь имеет около 1000 сообщений. Для оценки сравниваем предложенную модель с базовыми моделями: – следуя исследованиям [4, 1], мы обучаем модель CNN каждой из характеристик символов (CNN-S), биграмм (CNN-B) и слов (CNN-W). Модель LSTM: Long-Short Term Memory (LSTM) также была реализована в предыдущем работах [1].

Результаты экспериментов

Как и в предыдущих исследованиях [1, 5], интерес представляют возможности модели в задачах, когда проблема АА становится более сложной, т. е. когда количество пользователей увеличивается или, когда количество постов на автора уменьшается. В таблице 1 для этого эксперимента используется 100 постов от каждого пользователя для обучения и 20 постов от каждого пользователя для тестирования, а количество пользователей варьируется от 50 до всех пользователей. Стоит отметить, что хотя точность снижается с увеличением числа пользователей, модель показывает лучшие результаты, чем базовые модели.

Таблица 1.

Метод	**50**	**100**	**500**	**5000**	**Все**
CNN-S	60.9	47.1	35.6	23.1	18.7
CNN-B	59.3	46.7	36.5	24.6	20.1
CNN-W	50.0	37.1	28.4	17.7	16.4
LSTM-S	40.3	29.6	27.3	16.1	15.3
LSTM -B	45.1	34.5	25.2	16.5	15.5
LSTM -W	35.0	22.8	19.3	12.0	11.7
Предложенная модель	64.8	51.2	37.9	25.5	21.4

Заключение

В этой статье предложена новая модель, которая использует несколько типов вложений для изучения представления стиля письма пользователя для атрибуции авторства. Эффективность модели была оценена при помощи реального набора данных постов, и эксперименты показали, что модель превосходит базовые показатели в различных экспериментальных условиях.

Библиографический список:

1. Shrestha P., Sierra S., Gonzalez F., Montes M., Rosso P., Solorio T.: Convolutional neural networks for authorship attribution of short texts. // Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, 2017, Volume 2, Short Papers, pp. 669–674.

2. Schwartz, R., Tsur, O., Rappoport, A., Koppel, M. Authorship attribution of micro-messages. // Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013, pp. 1880–1891.

3. Sabour S., Frosst N., Hinton G.E. Dynamic routing between capsules. // Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17), 2017, pp. 3859–3869.

4. Батурин М.М., Белов Ю.С. Использование сверточных, рекуррентно-сверточных нейронных сетей и метода опорных векторов для определения авторства текста // Научные исследования в современном мире. Теория и практика. сборник избранных статей Всероссийской (национальной) научно-практической конференции. Санкт-Петербург, 2022. С. 47-49.

5. Ruder S., Ghaffari P., Breslin J.G. Character-level and multi-channel convolutional neural networks for large-scale authorship attribution. // Insight Centre for Data Analytics. National University of Ireland Galway, Technical Report, 2016.