Нейронные сети в приложении к задаче многозначной классификации постов пользователей в социальной сети

В. Д. Олисеенко
Санкт-Петербургский федеральный исследовательский центр РАН
vdo@dscs.pro

Т. В. Тулупьева
Санкт-Петербургский государственный университет;
Санкт-Петербургский федеральный исследовательский центр РАН;
Северо-Западный институт управления РАНХиГС
tvt@dscs.pro

***Аннотация*. В данном исследовании представлен результат автоматизации многозначной (англ. multi-label) классификации текстовых постов пользователей в социальных сетях с использованием нейронной сети, имеющей архитектуру долгой краткосрочной памяти (англ. long short-term memory). Полученная модель позволит автоматизировать часть процесса оценки степени выраженности психологических особенностей пользователей по их постам в социальных сетях, что в свою очередь является важным шагом для выработки рекомендаций к повышению их защищенности от социоинженерных атак.**

Ключевые слова: классификация постов; социальные сети; многозначная классификация; нейронные сети; социоинженерные атаки; машинное обучение; информационная безопасность; защита пользователя; профиль уязвимостей пользователя

I. Введение

По данным компании Positive Technologies за 2020 год [1] в 65 % атак на пользователей информационных систем использовалась элементы социальной инженерии. Именно поэтому вопрос повышения защищённости пользователей информационных систем от социоинженерных атак набирает всё большую актуальность. В качестве инструмента для оценки защищенности пользователей информационных систем может выступать профиль уязвимостей пользователя [2], который, в свою очередь, возможно построить в результате анализа информации, публикуемой пользователями в социальных сетях. Так в работах [3] была создана схема с критериями для выявления психологических особенностей пользователей по их текстовым постам на личной странице. Согласно данной схеме, все посты пользователей можно отнести к трем классам: информационные, эмоциональные и деятельные. Следует отметить, что один и тот же пост может быть информационным и эмоциональным или информационным и деятельным. А может быть отнесен сразу к трем классам. Данные классы делятся на непересекающиеся подклассы: информационные на формальные, событийные, личные, интеллектуально-рассудительные, ссылочные, кулинарные; эмоциональные на позитивные, негативные и поздравительные; деятельные на благотворительные, продающие, побудительные к действию.

Первые попытки автоматизации процесса данной классификации были предприняты в работе [4]. Однако, предложенный в ней подход предполагал использование множества бинарных классификаторов в формате «один против всех» для определения каждого из подклассов. Такой подход является наиболее простым в решении данной задачи, но в то же время может вызвать ситуацию, когда два и более классификатора ошибутся и выберут подклассы в одном классе, что противоречит изначальным критериям классификации. В качестве одного из решений данной проблемы можно использовать иерархическую (ансамблевую) классификацию, где классификатор первого уровня будет классифицировать по трём классам, а второго уровня уже находить подклассы в каждом, определенном ранее классе. Стоит отметить, что задача построения классификатора первого уровня осложнена тем, что пост пользователя может одновременно принадлежать нескольким классам. Таким образом, данная задача является задачей многозначной (англ. multi-label) классификации [5]. Целью данного исследования является автоматизация многозначной классификации текстовых постов пользователей в социальных сетях по разработанной ранее схеме с критериями. Теоретическая значимость заключается в возможности проработки и автоматизации подходов к оценке степени выраженности личностных особенностей пользователя и, опосредованно, уязвимостей пользователей. Практическая значимость заключается в дополнении функциональной составляющей существующего прототипа комплекса программ для анализа защищенности пользователей.

II. Релевантные работы

Одной из ключевых работ в систематизации теории многозначной классификации является работа [6], авторы которой провели комплексный обзор существующих методов, подходов алгоритмов и метрик, в том числе

Работа выполнена в рамках проекта по государственному заданию СПИИРАН № 0073-2019-0003, при финансовой поддержке РФФИ проект №20-07-00839.

применимых для текстов. В работах [5], [5] авторы выделяют два основных подхода к решению задач многозначной классификации: метод преобразования и метод адаптации. Метод преобразования разбивает задачу многозначной классификации на несколько задач бинарной классификации, а метод адаптации расширяет применения существующих методов бинарных и многоклассовых классификаций до многозначных. Одними из наиболее перспективных видятся методы адаптации на основе нейронных сетей [7], [8], которые позволяют находить и использовать скрытые связи между классами, повышая точность классификации. Наиболее близкой статьей по тематике является статья [9], где авторы используют многозначную классификацию вместе со словарями настроений для определения настроения пользователей в микроблогах. В данной работе будет рассмотрен подход на основе нейронной сети, имеющей архитектуру долгой краткосрочной памяти (long short-term memory) [10], [11].

III. Постановка задачи

Для решения задачи многозначной классификации текстовых постов пользователей в социальных сетях необходимо построить классификатор со следующим условием: пусть X – множество постов пользователей, $L = \{\lambda_1, \lambda_2, \lambda_3\}$ – множество возможных классов (меток), тогда $\alpha(X, Y)$ – классифицирующая функция, где $Y \subseteq L$ набор меток, иными словами $Y \in \{0,1\}^3$. Таким образом, классифицирующая функция $\alpha(X, Y)$ должна поставить в соответствие каждому посту пользователя три метки принимающие значения 0 или 1, которые определяют принадлежность к каждому из классов (рис. 1).

Информационный	Эмоциональный	Деятельный
Формальный	Позитивный	Благотворительный
Событийный	Негативный	Продающий
Личный	Поздравительный	Побудительный
Цитата/рассуждение		
Ссылочной		
Кулинарный		

Рис. 1. Критерии классификации

В качестве признака для классификации постов выступает текст, для этого его необходимо подготовить. В данной работе текст постов рассматривается на морфологическом уровне в векторной модели, т. е. на уровне отдельных слов в предложениях.

IV. Описание данных

Процесс подготовки текста выглядит следующим образом: перевод всех букв в нижний регистр, удаление цифр и неинформативных символов, удаление пунктуации, пробелов и стоп-слов, выделение слов в токены, приведение слов к канонической форме (лемме). Более подробное описание процесса содержится в разделе V.

Исходный набор данных был получен в [4] путем сбора текстов постов с личных страниц пользователей ВКонтакте (с получением согласия самих пользователей) и составляет более 2-х тысяч постов. Далее полученный набор был размечен группой экспертов. После предобработки распределение трех главных классов по постам пользователей (информационного, деятельного, эмоционального) выглядит следующим образом (рис. 2).

Рис. 2. Соотношение классов

Общее количество постов после предобработки составило 1568 штук. Также можно отметить сильный дисбаланс в классах, информационные посты преобладают в выборке, и в пересечениях классов. В итоговом наборе данных представлено 39 628 слов, самый длинный пост состоит из 1539 слов, самый короткий пост – из 2 слов.

V. Описание решения задачи

На первом этапе исследования необходимо подготовить текст. В качестве основного инструмента для подготовки текста и проведения исследования в целом используется язык Python 3 и ряд библиотек. Так, для удаления всех посторонних символов, которые не являются кириллицей, применяется стандартная библиотека для работы с регулярными выражениями re. Для того, чтобы нормализовать и лемматизировать текст используется библиотека Natasha [10], а для удаления стоп-слов – взят набор русскоязычных слов из библиотеки NLTK [13]. Процесс разделения предложений на слова-компоненты (токенизация) реализован через библиотеку Tensorflow [14], также как и построение модели нейронной сети.

На втором этапе строится и обучается нейронная сеть. Её архитектура имеет вид долгой краткосрочной памяти (рис. 3) со следующими параметрами обучения: оптимизатор – adam; функция потерь – бинарная перекрестная энтропия; метрика для оптимизации – точность (англ. accuracy); выходной слой содержит сигмоидную функцию активации.

Layer (type)	Output Shape	Param #
embedding_2 (Embedding)	(None, 250, 64)	320000
spatial_dropout1d_2 (Spatial	(None, 250, 64)	0
lstm_4 (LSTM)	(None, 250, 80)	46400
lstm_5 (LSTM)	(None, 20)	8080
dense_2 (Dense)	(None, 3)	63

Рис. 3. Архитектура нейронной сети

На вход в сеть подаётся токенизированный текст поста пользователя социальной сети, на выход оценка принадлежности к каждому классу в промежутке [0, 1]. Для проверки обобщающей способности алгоритма используется принцип скользящего контроля по четырем блокам (англ. 4fold). В качестве метрик для оценки полученной модели используются обычные и адаптированные метрики accuracy, precision, recall, f1-score, также AUC-ROC [6].

VI. Эксперимент

В таблице I представлен усредненный, на четырех тестовых блоках, результат метрик оценки качества работы классификатора. На пересечении первых трех строк и четырех столбцов рассчитаны классические метрики для каждого класса в отдельности. По полученным результатам можно сделать вывод, что по отдельности качество классификации довольно высокое, за исключением объектов из класса деятельный. Данный класс определяется чуть хуже других (метрика Recall), так как он самый малочисленный в рассматриваемом наборе данных (рис. 2). Обобщенная метрика точности определяет долю тех объектов, у которых по итогам классификации правильно поставлены все классы. Таких объектов не слишком много, так как в исходном обучающем наборе данных присутствует сильный дисбаланс в классах. Строки micro-average и macro-average отражают соответствующие метрики precision, recall, f1-score для всех классов одновременно.

ТАБЛИЦА I Precision, Recall, F1-score, Accuracy

	Precision	**Recall**	**F1-score**	**Accuracy**
Информационные	0,862	0,977	0,916	0,851
Эмоциональные	0,709	0,799	0,751	0.830
Деятельные	0,831	0,584	0,686	0.921
Micro average	0,820	0,888	0,853	
Macro average	0,801	0,787	0,784	
Обобщенная				0.665

На (рис. 4) построена ROC-кривая и посчитана площадь под ней. Из полученных значений площади также можно сделать вывод, что качество классификатора как в обобщенном варианте, так и в обычном довольно высоко.

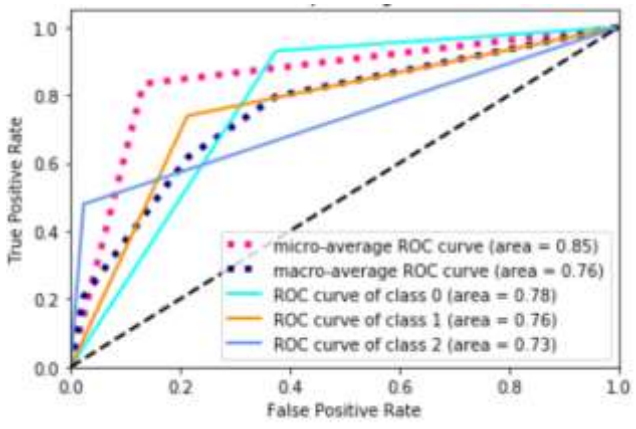

Рис. 4. ROC-кривая

VII. Заключение

В данной работе был представлен подход для автоматизации многозначной классификации текстовых постов пользователей социальных сетей на основе нейронной сети, имеющей архитектуру долгой краткосрочной памяти. Разработанный подход позволит частично автоматизировать процесс классификации постов пользователей для оценки выраженности их психологических особенностей, что является одной из ключевых задач в защите сотрудников компаний и предприятий [15]. Стоит также отметить, что полученные результаты являются лишь первой частью иерархической (ансамблевой) классификации и позволяют выделить только главные классы из схемы [3]. Возможные дальнейшие направления исследования заключаются в разработке второй части иерархической (ансамблевой) классификации, которая позволит в полученных классах выделить подклассы путём решения задачи многоклассовой классификации и также в существенном расширении набора данных для повышения точности разработанных подходов.

Список литературы

[1] Кибербезопасность 2020–2021 Тренды и прогнозы [Электронный ресурс]. URL: https://www.ptsecurity.com/upload/corporate/ru-ru/analytics/Cybersecurity_20-21.pdf

[2] Абрамов М.В. Автоматизация анализа социальных сетей для оценивания защищённости от социоинженерных атак // Автоматизация процессов управления. 2018. №1(51). С. 34–40.

[3] Тулупьева Т.В., Тафинцева А.С., Тулупьев А.Л. Подход к анализу отражения особенностей личности в цифровых следах // Вестн. психотерапии. 2016. № 60 (65). С. 124–137.

[4] Тулупьева Т.В., Суворова А.В., Азаров А.А., Тулупьев А.Л., Бордовская Н.В. Возможности и опыт применения компьютерных инструментов в анализе цифровых следов студентов пользователей социальной сети // Компьютерные инструменты в образовании. 2015. № 5. С. 3–13.

[5] Карпович С.Н. Многозначная классификация текстовых документов с использованием вероятностного тематического моделирования ml-PLSI // Труды СПИИРАН. 2016. № 47. С. 92–104.

[6] Gibaja E., Ventura S. A Tutorial on Multilabel Learning // ACM Comput. Surv. 2015. № 47 (3). Article 52. 38p. Doi: 10.1145/2716262

[7] Gargiulo F., Silvestri S., Ciampi M., De Pietro G. Deep neural network for hierarchical extreme multi-label text classification // Applied Soft Computing Journal. 2019. № 79. P. 125–138. Doi: 10.1016/j.asoc.2019.03.041.

[8] Chen W., Liu X., Guo D., Lu M. Multi-label text classification based on sequence model // Communications in Computer and Information Science. 2019. №1071. P. 201–210. Doi: 10.1007/978-981-32-9563-6_21.

[9] Liu S.M., Chen J.-H. A multi-label classification based approach for sentiment classification // Expert Systems with Applications. 2015. №42 (3). P. 1083-1093. Doi: 10.1016/j.eswa.2014.08.036.

[10] Shah D.K., Sanghvi M.A., Mehta R.P., Shah P.S., Singh A. Multilabel Toxic Comment Classification Using Supervised Machine Learning Algorithms // Lecture Notes in Networks and Systems. 2021. № 141. P. 23–32. Doi: 10.1007/978-981-15-7106-0_3.

[11] Liu G., Guo J. Bidirectional LSTM with attention mechanism and convolutional layer for text classification // Neurocomputing. 2019. № 337. P. 325-338. Doi: 10.1016/j.neucom.2019.01.078

[12] Библиотека для обработки русского языка языка Natasha [Электронный ресурс]. URL: https://github.com/natasha/natasha

[13] Библиотека для обработки естественного языка NLTK [Электронный ресурс]. URL: https://www.nltk.org/

[14] TensorFlow – открытая программная библиотека для машинного обучения [Электронный ресурс]. URL: https://www.tensorflow.org/

[15] Khlobystova A.O., Abramov M.V., Tulupyev A.L. Employees' Social Graph Analysis: A Model of Detection the Most Criticality Trajectories of the Social Engineering Attack's Spread // International Conference on Intelligent Information Technologies for Industry. Springer, Cham. 2019. С. 198–205. Doi: 10.1007/978-3-030-50097-9_20