М.Д. Долгушин, Д.К. Исмакова, Ю.В. Бидуля, Г.Б. Барская

Тюменский государственный университет, г. Тюмень

УДК 004.912

РАЗРАБОТКА WEB-API КЛАССИФИКАЦИИ ТОКСИЧНЫХ ТЕКСТОВ

Аннотация. В статье представлен процесс разработки api для автоматизированного анализа русскоязычных комментариев в социальных сетях, и алгоритма классификации токсичных, ругательных комментариев. Полученное API будет использоваться в Web-сервисе поддержки деятельности русскоязычного модератора группы в социальных сетях.

Ключевые слова: API-приложение, модерация сообщений, обработка естественного языка, линейные алгоритмы классификации.

Введение

Спрос на услуги модератора социального сообщества с каждым годом все увеличивается [1], но системы поддержки его работы из-за сложности классификации комментариев практически отсутствуют. Однако, с популяризацией алгоритмов классификации текста проблема модерации комментариев становится решаемой. Потому задача разработки API-приложения для классификации русскоязычных комментариев социальных сетей с помощью алгоритмов классификации текста, является актуальной.

Существующие приложения для модерации сообщений

На данный момент приложений для автоматизированной модерации социальных сетей не существует, однако во многих социальных сетях внедряются простые фильтры [2] на основе списка запрещенных слов, составляемого модератором. Данные фильтры позволяют удалять сообщения с ругательствами или какими-то специальными запрещенными словами. Однако семантику сообщения такие фильтры не учитывают, что часто приводит к удалению вполне безобидных сообщений, либо пропуску комментариев с более тонким оскорблением.

Также проводились исследовательские работы [3,4] по классификации токсичных комментариев с помощью линейных классификаторов и нейронных сетей, показавшие хорошие результаты точности. Однако, большинство работ в данной сфере ведется относительно комментариев на английском языке, а не на русском, и используемый набор данных может быть ненадежен.

API классификации сообщений

Было решено разрабатывать API на языке Python, поскольку на нем реализовано множество модулей машинного обучения и он довольно прост в изучении.

Для разработки API классификации сообщений был собран и вручную размечен набор данных, состоящий из 5000 комментариев. Комментарии взяты из социальной сети ВКонтакте с помощью библиотеки vk_api. 1135 комментариев отмечены как токсичные, 3865 комментариев – не токсичные. Классы несбалансированы, а набор данных мал.

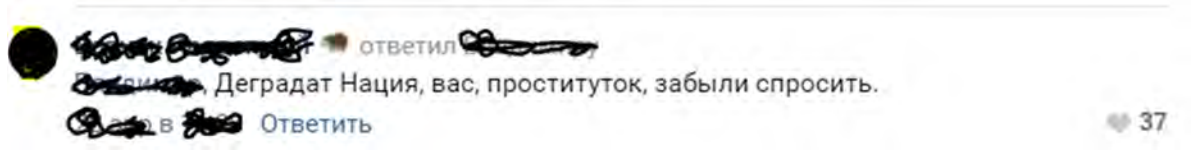

Рис. 1. Пример токсичного сообщения

На рис. 1 представлен пример токсичного сообщения из социальной сети ВКонтакте.

Для анализа набора данных были выделены дополнительные признаки, с помощью библиотек Gensim, Dostoevsky и простого фильтра ругательств. Выделены были следующие признаки: наличие указания личности, места, организации; негативный, позитивный, нейтральный эмоциональные окрасы; наличие ругательств.

	comment	toxic	PER	LOC	ORG	positive	negative	…	…	…	rude
2	[redacted] тут сложного?А) -10Б) 15/6Гуманитарий в треде кстати	1	1	0	0	0.022987367585…	0.01913403719663…	…	…	…	1
3	количество экспертов в интернетных спорах упало на 69%	0	0	0	0	0.023699468001…	0.50001001358032…	…	…	…	0
4	хммммм задумался над монетизацией	0	0	0	0	0.016924913972…	0.02443309314548…	…	…	…	0
5	Ну а что зато все мамины эксперты во всех областях сразу же исп…	1	0	0	0	0.023699468001…	0.31406053900718…	…	…	…	0
6	Бугагага) гендерные исследователи и политические аналитики тепер…	1	0	0	0	0.106700591742…	0.05501529201865…	…	…	…	0
7	А патау ясно.....	0	0	0	0	0.002641674596…	0.06755668669939…	…	…	…	0
8	Пора закрывать твиттер	0	0	0	0	0.009718479588…	0.03733688220381…	…	…	…	0
9	Они настолько [redacted] что не знают про ГДЗ?	1	0	0	1	0.001377025386…	0.30075559020042…	…	…	…	0
10	Позвонил в штаб навального а мне координатор шёпотом сказал чт…	0	0	0	0	0.048867784440…	0.16027602553367…	…	…	…	0

Рис. 2. Пример признаков для комментариев

На рис. 2 представлен пример выделенных признаков для комментариев, где:

- toxic – токсичность,
- PER – указание на личность,
- LOC – указание не место,
- ORG – указание на организацию,
- positive – позитивный эмоциональный окрас,
- negative – негативный эмоциональный окрас,
- rude – наличие ругательства.

Далее были получены результаты корреляции Пирсона для выделенных признаков с вручную размеченной токсичностью.

Таблица 1. Коэффициент корреляции Пирсона

Функция параметра с F("Toxic")	**Коэф. Корел. Пирсона**	**p**
F("PER")	-0.013487633351327665	0.3408563255644402
F("LOC")	-0.06327667493088972	7.721113062130543e-06
F("ORG")	-0.017656503724165926	0.6323986893140852
F("positive")	-0.04727098761866796	0.0008380399482287616
F("negative")	0.3613570789484974	9.22316478737137e-154
F("neutral")	-0.2673834873578516	1.9770095855740052e-82
F("rude")	0.4134216419158385	2.8822364997752216e-205

По представленным в таблице 2 коэффициентам заметна сильная корреляция параметра токсичности с ругательствами и негативным эмоциональным окрасом.

Тогда для классификации стала задача предобработки текстов комментариев. На этапе предобработки из комментариев были удалены символы пунктуации, стоп-слова, и все слова были приведены к начальной форме с помощью библиотеки PyMorphy2.

Далее для построения векторных моделей сообщения было решено использовать метрику TF для слов, поскольку ее использование более обоснованно на малых текстах комментариев, чем метрика TFIDF. Также использовалась векторная модель Word2Vec(min_count=5, window=10, size=150, negative=10, alpha=0.03, min_alpha=0.0007, sample=6e-5, sg=0), хорошо показавшая себя в классификации текстов [5]. Получаемые в модели вектора слов усреднялись для получения вектора сообщения.

Для классификации было решено использовать линейные классификаторы MNB(alpha=1), SVC(C=0.1), NBSVM(alpha=1, C=1) [6], поскольку на малом количестве записей их использование более обоснованно, чем использование нейронных сетей [7]. Выборка записей была предварительно разделена на тестовую и тренировочную в отношении 1/5.

Таблица 2. Результаты оценки классификаторов

Метод		**toxic**			**untoxic**			**Accuracy**
		Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**	
Count_Vec	SVC	0.819	0.438	0.571	0.780	0.954	0.858	0.787
	MNB	**0.878**	0.247	0.385	0.732	**0.984**	0.839	0.745
	NBSVM	0.825	0.433	0.568	0.779	0.956	0.859	0.787
Word2Vec	SVC	0.842	0.707	0.769	0.842	0.842	0.902	0.862
	MNB	0.778	**0.779**	**0.778**	**0.894**	0.894	0.894	0.857
	NBSVM	0.846	0.709	0.771	0.871	0.938	**0.903**	**0.864**

Из результатов таблицы 1 следует, что наибольшей точностью 0.864 (Accuracy) обладает набор Word2Vec векторизация и NBSVM классификация, хотя в некоторых частных случаях от данной классификации не отстает

Word2Vec векторизация с MNB классификацией. Потому для классификации был выбран набор Word2Vec + NBSVM.

Далее с применением фреймворка Flask было разработано API, позволяющее как возвращать различные данные по запросам к vk_api, так и определять токсичность с помощью обученного набора Word2Vec+NBSVM, а также определять дополнительные признаки сообщения: наличие имен собственных, эмоциональный окрас, наличие ругательств.

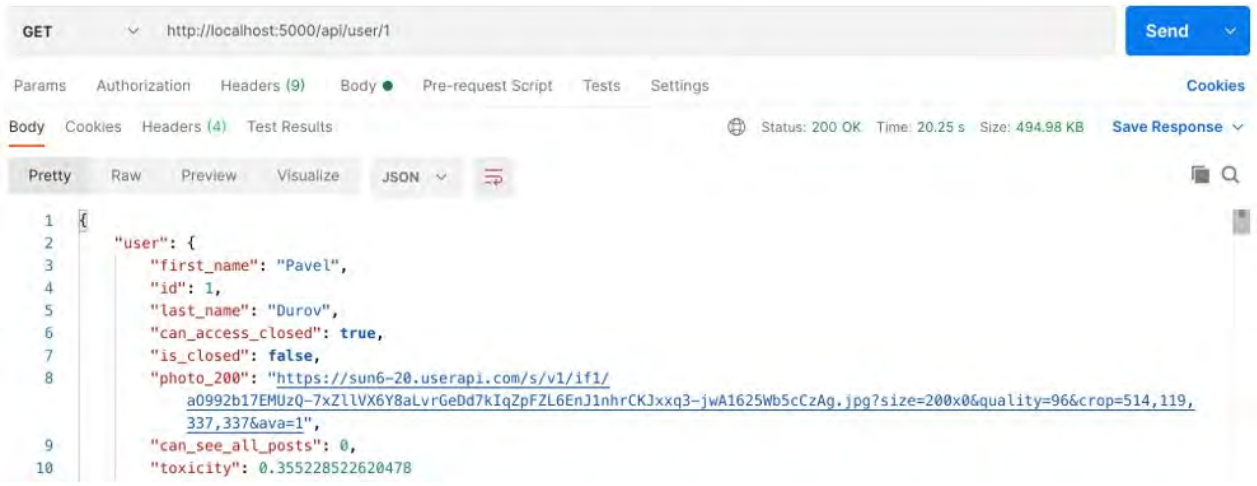

Рис. 3. Пример запроса на получение токсичности пользователя

На рис. 3 представлен пример результата запроса на получение средней токсичности 100 первых постов пользователя. Ограничение в 100 постов введено из-за запрета на большее количество запросов к VK единоразово.

Заключение

Был составлен, размечен и проанализирован набор данных по токсичным комментариям, состоящий из 5000 тысяч комментариев из социальной сети ВКонтакте. Далее был разработан алгоритм классификации токсичных комментариев на основе Word2Vec и NBSVM алгоритмов с точностью классификации 86.4%, а также обосновано его преимущество в сравнении с иными алгоритмами. Было разработано API, позволяющее: взаимодействовать через запросы с данными из социальной сети ВКонтакте, проводить

классификацию токсичности комментариев, а также получать эмоциональный окрас, наличие ругательств и наличие имен собственных в тексте.

Разработанное API предполагается использовать в полноценном WEB-сервисе поддержки деятельности модератора групп в социальных сетях. В будущем также полученный набор данных предполагается расширить и добавить определение разных видов токсичности сообщений, что может повысить точность определения токсичности и повысит полезность приложения в работе модератора соц. сети.

СПИСОК ЛИТЕРАТУРЫ

1. Калеушкина А. Модерация. Почему бизнесу важно обратить внимание на клиентскую поддержку в соцсетях уже сейчас. URL: https://vc.ru/social/109910-moderaciya-pochemu-biznesu-vazhno-obratit-vnimanie-na-klientskuyu-podderzhku-v-socsetyah-uzhe-seychas (Дата обращения: 05.05.2021).

2. Шаповалов Л. Как быстро фильтровать комментарии в соцсетях. URL: https://vc.ru/marketing/164843-kak-bystro-filtrovat-kommentarii-v-socsetyah (Дата обращения: 05.05.2021).

3. Клюев Л. Выявление и классификация токсичных комментариев. Лекция в Яндексе. URL: https://habr.com/ru/company/yandex/blog/414993/ (Дата обращения: 05.05.2021).

4. Сметанин С. Определение токсичных комментариев на русском языке. URL: https://habr.com/ru/company/mailru/blog/526268/ (Дата обращения: 05.05.2021).

5. Mauro Di Pietro. Text classificatoin with NLP: tf-idf vs Word2Vec vs BERT. URL: https://towardsdatascience.com/text-classification-with-nlp-tf-idf-vs-word2vec-vs-bert-41ff868d1794 (Дата обращения: 05.05.2021).

6. Wang Z. NBSVM | Kaggle. URL: https://www.kaggle.com/ziliwang/nbsvm (Дата обращения: 05.05.2021).

7. Wang S., Manning C.D. Baselines and Bigrams: Simple, Good Sentiment and Topic Classification. URL: https://nlp.stanford.edu/pubs/sidaw12_simple_sentiment.pdf (Дата обращения: 05.05.2021).