Классификация групп в социальной сети «ВКонтакте» по возрастному признаку

Белов С. В., студент кафедры безопасных информационных технологий Университета ИТМО, syarhei.belov@gmail.com

Аннотация

В настоящее время для социальных сетей являются востребованными методы анализа тональности и модель автоматической возрастной классификации. Целью работы являлось исследование различных методов анализа тональности текстов для дальнейшего построения модели автоматической возрастной классификации информации в социальных медиа.

Актуальность исследования и построения данной модели объясняется вступившим в силу 1-го января 2012 года Федерального закона № 436-ФЗ. Принимая во внимание то, что все больше заметна тенденция к введению контроля информации в Интернете со стороны государственных структур, можно говорить о целесообразности построения подобной модели. В работе описаны исследования наивного байесовского классификатора, метода опорных векторов и искусственных нейронных сетей, результаты также описаны. При оценке тональности была получена наивысшая точность в 73,5 %.

Введение

В современное время в Интернете циркулирует огромное количество информации, и из года в год ее количество только увеличивается. Невольно встает вопрос: «Как мы можем оценить пригодность информации для какой-либо возрастной группы?». 1-го января 2012 года в России вступил в силу Федеральный закон от 29.12.2010 № 436-ФЗ «О защите детей от информации, причиняющей вред их здоровью и развитию» [1]. В этом законе описаны пять возрастных групп разграничения информации: 0+, 6+, 12+, 16+ и 18+. Также там описаны критерии, которыми необходимо пользоваться, при маркировании информации.

Обзор литературы

В [2] авторы описывают попытку математического моделирования проблемы автоматической классификации текстов. Однако, авторы данной работы сосредотачиваются не на приемлемости контента для какой-либо возрастной группы. Они пытаются определить, какой из возрастных групп

информация будет понята и воспринята. В этом случае авторы решают несколько отличную проблему. Они не фокусируются на специфических федеральных актах, не принимают во внимание правовые рамки.

В [3] авторы описывают близкую задачу. Они анализируют тексты песен и пытаются определить их возрастную приемлемость. Есть несколько различий между [3] и данным исследованием. В [3] авторы описывают семь возрастных групп и они исследуют тексты песен, как было сказано ранее, которые имеют свою специфику и особенности. Также, они анализируют тексты на английском языке, используя психолингвистическую базу данных (MRC Psycholinguistic Database) и проект GloVe Гарвардского университета. Данные методы неприменимы для русского языка, поскольку довольно сильно привязаны к английскому языку и его специфике.

Общая схема модели

Общая схема задачи заключается в присвоении метки c конкретному документу d (например, тексту, комментарию, посту и т. д.). Метка c получается из заранее определенного набора допустимых классов $C = \{c_1, c_2, ..., c_n\}$ [3]. В описываемой работе было определено лишь два класса: «18+» и «18-», поскольку данная задача является базовой для остальных. Базируясь на требованиях Федерального закона [1], была построена общая схема модели автоматической возрастной классификации информации (см. Рис. 1).

На входе модели подается дамп постов исследуемой группы в социальной сети. На выходе получается класс информации: «18+» или «18-».

Эксперименты (анализ тональности)

Первый эксперимент

Для проведения первого эксперимента, были выбраны 100 постов из различных групп в социальной сети «ВКонтакте». Данные группы содержат посты с различными шутками о жизни. Тональность была оценена при помощи пяти экспертов. Эти оценки были сравнены с машинными оценками, которые получены на выходе модели анализа тональности. Для сравнения была использована статистическая мода оценок экспертов.

Были изучены и сравнены точности четырех моделей:

- модель, основанная на наивном байесовском классификаторе (Naïve Bayes, NB);

- модель, основанная на методе опорных векторов (support vector machine, SVM);
- модель, основанная на методе опорных векторов со взвешиванием векторов при помощи Delta TF-IDF;
- модель, основанная на рекуррентной искусственной нейронной сети (artificial neural network, ANN).

Также, в первых трех случаях были использованы векторы с уни-, биграммами и их комбинацией. Результаты представлены в Таблице 1.

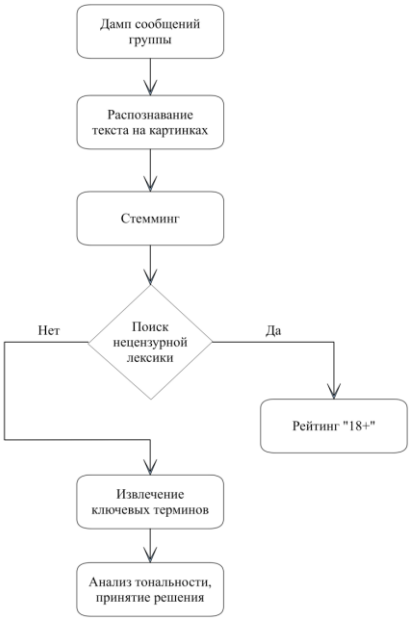

Рисунок 1: Общая схема модели автоматической возрастной классификации информации

	NB	**SVM**	**SVM + Delta TF-IDF**
униграммы	46 %	49 %	47 %
биграммы	52 %	49 %	49 %
комбинация	45 %	49 %	42 %

Таблица 1: Точность векторных моделей

Более хорошие результаты были получены при исследовании рекуррентной нейронной сети. Для обучения нейронной сети был использован размеченный корпус сообщений социальной сети «Twitter» [4]. Результаты представлены в Таблице 2.

Тип текста	Точность
Текст без каких-либо изменений	47 %
Удаление пунктуации	53 %
PyStemmer	50 %
Mystem	51 %
PyStemmer + удаление пунктуации	56 %
Mystem + удаление пунктуации	51 %

Таблица 2: Точность рекуррентной нейросети (первый эксперимент)

Второй эксперимент

После получения приблизительных оценок в первом эксперименте, был проведен второй эксперимент. В этом случае, тональность была оценена шестью экспертами, а также были использованы разные размеры обучающей выборки.

В этом эксперименте, были отобраны 300 постов из разных групп в социальной сети «ВКонтакте» вместо 100. 100 из этих 300 постов содержат различные шутки, сарказм или иронию. Полученная точность представлена в Таблице 3. Интуитивно ожидалось, что разделение на три класса тональности (положительный, негативный и нейтральный) более реалистично, чем на два (положительный и негативный). Также результаты представлены на графике (см. Рис. 2).

Размер обучающей выборки	2 класса		3 класса	
	С сарказмом	*Без сарказма*	*С сарказмом*	*Без сарказма*
25 000 постов	42,33 %	38,00 %	29,33 %	31,00 %
50 000 постов	65,33 %	70,00 %	40,33 %	38,00 %
100 000 постов	69,67 %	73,50 %	29,00 %	24,00 %
150 000 постов	67,33 %	69,50 %	29,33 %	22,50 %
200 000 постов	64,67 %	66,50 %	31,00 %	28,50 %

Таблица 3: Точность рекуррентной нейросети (второй эксперимент)

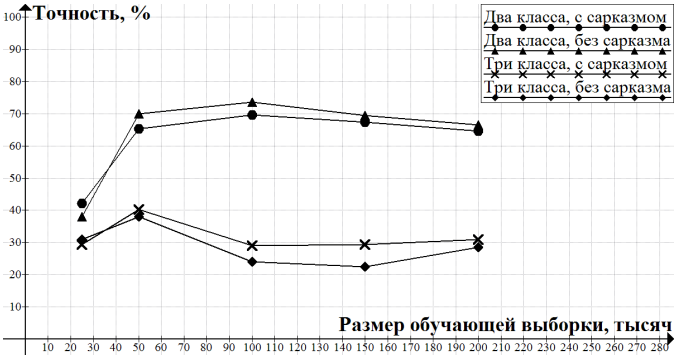

Рисунок 2: График зависимости точности от размера обучающей выборки

Заключение

На основании проведенных экспериментов и полученных данных, можно отемтить, что рекуррентная нейронная сеть лучше других моделей подходит для анализа тональности текста для текстов из «ВКонтакте».

Областями дальнейших исследований являются построение модели извлечения ключевых терминов из текста и создание библиотеки терминов, маркирующих категорию «18+». После этого будет возможно создать всю модель целиком.

Литература

1. О защите детей от информации, причиняющей вред их здоровью и развитию : федеральный закон Российской Федерации 29 дек. 2010 г. № 436-ФЗ (ред. от 29 июня 2015 г.) [Электронный ресурс] // КонсультантПлюс. – Режим доступа: http://base.consultant.ru/cons/cgi/online.cgi?req=doc;base=LAW;n=181927, своб. (дата обращения 15.04.2016).
2. Глазкова А. В., Захарова И. Г. Подход к моделированию задачи автоматической классификации текстов (на примере их отнесения к определенной возрастной аудитории) // Вестник ТюмГУ – 2014. – № 7. – С. 205‑211.
3. Maulidyani A., Manurung R. Automatic Identification of Age-Appropriate Ratings of Song Lyrics // Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 2: Short Papers – 2015. – № 2. – С. 583-587.
4. Рубцова Ю. В. Построение корпуса текстов для настройки тонового классификатора // Программные продукты и системы – 2015. – № 109. – С. 72‑78.