Оценка эмоциональной окраски постов социальной сети «ВКонтакте», включающих эмодзи, методами машинного и глубокого обучения

А. П. Быкова

Санкт-Петербургский государственный университет

st098553@student.spbu.ru

Аннотация

В данной работе исследуются возможности применения методов машинного и глубокого обучения к оценке эмоциональной окраски текста постов, содержащих эмо́дзи, из социальной сети «ВКонтакте». Описывается несбалансированный набор данных с текстом постов, размеченный по 15 классам, учитывающим эмоциональную и тональную составляющие в тексте. На полученном наборе данных проводятся эксперименты с использованием 6 методов классического машинного обучения, их ансамблей с мажоритарным и мягким голосованием и 3 нейросетевых методов. Лучший результат по метрикам качества классификации получился для модели BoW +VotingClassifier (soft) (мешок слов + ансамблевый метод с мягким голосованием) на лемматизированном тексте с пунктуацией и с эмо́дзи: F1-мера macro = 69.70%, F1-мера weighted = 82.06% и для рекуррентной нейросети GRU на 15 эпохах обучения: F1-мера macro = 48.77%, F1-мера weighted = 83.74%.

Ключевые слова: анализ эмоций, эмоциональная окраска текста, эмо́дзи, машинное обучение, нейронные сети

Библиографическая ссылка: Быкова А. П. Оценка эмоциональной окраски постов социальной сети «ВКонтакте», включающих эмо́дзи, методами машинного и глубокого обучения // Компьютерная лингвистика и вычислительные онтологии. Выпуск 7 (Труды XXVI Международной объединённой научной конференции «Интернет и современное общество», IMS-2023, Санкт-Петербург, 26–28 июня 2023 г. Сборник научных статей). — СПб: Университет ИТМО, 2024. С. 12–20. DOI: 10.17586/2541-9781-2024-7-12–20

1. Введение

Уже не одно десятилетие исследователи достаточно много внимания уделяют анализу тональности текста и речи. Результаты анализа эмоциональной окраски текстов имеют множество практических применений, например, в различных приложениях при работе с клиентами, в политологии при работе с политическими окрашенными текстами, здравоохранении. Изучается потенциал анализа эмоций для выявления и предотвращения различных форм онлайн-злоупотреблений, например, запугивания пользователей. Кроме того, растёт интерес к тому, как эмоции передаются в разных языках и культурах, и как это может повлиять на оценку эмоциональной окраски различной информации [1].

Оценка эмоциональной окраски текста может быть полезна во многих областях, например, для того, чтобы понять какое настроение выражено в тексте. Эта информация может использоваться для анализа мнений, анализа отзывов клиентов, мониторинга социальных сетей. Понимая эмоции, выраженные в тексте, организации могут лучше учитывать потребности и предпочтения своих клиентов. Понимание эмоций также можно использовать в личном общении, чтобы оценить настроение человека и отреагировать

соответствующим образом. В данном исследовании оценка эмоциональной окраски текста постов в социальной сети «ВКонтакте» проводится методами машинного и глубокого обучения. Для разметки постов использовалась автоматическая разметка на основании встречающихся в этих постах эмóдзи.

Эмóдзи — это цифровые изображения или значки, которые используются в текстовых сообщениях в различных социальных сетях, в том числе «ВКонтакте». Язык эмóдзи своего рода графический язык, где вместо слов используются сочетания картинок. Впервые эмóдзи появились в Японии и распространились по всему миру. В настоящее время использование эмóдзи достаточно популярно и доступно в самых разных стилях и дизайнах. Популярность эмóдзи обусловлена тем, что они могут передавать эмоции и добавлять контекст к текстовому общению. В некоторых случаях эмóдзи помогают преодолевать языковые барьеры и делают общение более доступным среди людей, которые владеют разными языками.

2. Подходы к анализу эмоциональной окраски текста

Анализ тональности текста — одно из направлений в компьютерной лингвистике, в рамках которого решается задача выявления мнения автора текста по поводу того, что обсуждается в тексте.

Тональность текста можно рассматривать как с точки зрения автора текста, так и с точки зрения того, кто читает и воспринимает этот текст. Поскольку в данном исследовании эмóдзи являются маркером для разметки, а эмóдзи проставляет сам автор текста, то в этом исследовании тональность и эмоциональная окраска текста рассматривается с точки зрения автора этого текста.

В целом, анализ эмоциональной окраски текста подразумевает собой применение методов, с помощью которых можно определить, к какому классу относится тот или иной текст. В основном используются алгоритмы на основе словарей и правил [2; 3] и методы на основе машинного обучения. Также существуют комбинированные методы, в которых словари оценочной лексики являются компонентом модели машинного обучения [4].

Для многих задач автоматической обработки текста необходимы специально размеченные текстовые данные, например, для автоматического распознавания в тексте иронии или сарказма [5].

Большую популярность в задачах анализа тональности приобрели методы машинного обучения. С начала 2000-х годов широко применяются классические методы машинного обучения, такие как логистическая регрессия, метод опорных векторов, наивный байесовский классификатор [6]. Также широко применяются классические нейронные сети, например, рекуррентные нейронные сети и свёрточные нейронные сети [7]. В 2019 году появились новые подходы к анализу текста на основе нейросетевой архитектуры трансформер, такие как модель BERT [8]. Использование архитектуры серии BERT для различных задач автоматической обработки текста привело к росту качества решений этих задач, в том числе и в задачах анализа тональности.

Первоначально модель BERT обучалась на многоязычных текстовых данных, затем в ряде исследований было выявлено, что дообучение BERT на данных конкретного естественного языка может дать лучшие результаты решения задач для этого языка. Так, например, в работе [9] исследователи описывают модель RuBERT, которая дообучена на модели BERT для русского языка.

3. Сбор и разметка данных

Набор данных создавался самостоятельно из постов социальной сети «ВКонтакте». Данные взяты из 100 наиболее популярных сообществ «ВКонтакте» на 5 февраля 2023 года.

Статистика по самым популярным сообществам взята с сайта «TOPPOST». Выбор постов из социальной сети «ВКонтакте» в качестве материала исследования обусловлен тем, что данная социальная сеть является популярной платформой, которой пользуются русскоязычные пользователи. В постах пользователи выражают собственное мнение и открыто взаимодействуют посредством различных реакций (лайки, комментарии, репосты). В социальной сети чаще происходит неформальное эмоционально окрашенное общение, поэтому текст постов можно использовать для оценки эмоциональной окраски текста.

API (Application Programming Interface) «ВКонтакте» представлен в открытом доступе, с помощью открытых методов был написан скрипт для скачивания текста постов.

В качестве маркеров для разметки текста использовались эмóдзи, которые встречаются в постах. Разметка на основе эмóдзи является ограничением данного исследования, поскольку такой разметки может быть недостаточно для точной классификации текста по эмоциям и тональности, особенно в том случае, если эмóдзи использовались авторами постов неоднозначно. Для пояснения значений эмóдзи и их классификации использовались карточки с описанием с сайта «Смайлики Эмóдзи».

Собран словарь эмóдзи, которые встречаются в скачанных постах, состоящий из 146 эмóдзи, входящие в тематическую группу Smileys & Emotion.

Полученные 146 эмóдзи распределены по классам. В этих классах учитывается эмоциональная составляющая и тональная составляющая, поскольку однозначно категоризировать эмоции достаточно сложно, например, для такой эмоции, как удивление, может присутствовать как положительная, так и отрицательная тональность (см. табл. 1).

Таблица 1. Классы эмоций и тональности

№	Эмоция	Настроение (тональность)
1	улыбка (smile)	позитивное/негативное (positive/negative)
2	нет эмоции (no_emotion)	нейтральное/скептическое (neutral/skeptical)
3	удовольствие	позитивное (positive)
4	нет эмоции (no_emotion)	позитивное/негативное (positive/negative)
5	грусть (sadness)	негативное (negative)
6	страх (feat)	негативное (negative)
7	стыд (shame)	негативное (negative)
8	гнев (anger)	негативное (negative)
9	отвращение (disgust)	негативное (negative)
10	удивление (surprise)	позитивное/негативное (positive/negative)
11	отвращение (disgust)	нейтральное/скептическое (neutral/skeptical)
12	удивление (surprise)	негативное (negative)
13	нет эмоции (no_emotion)	негативное (negative)
14	грусть (sadness)	позитивное/негативное (positive/negative)
15	испуг (fear)	позитивное/негативное (positive/negative)

Самое большое количество эмóдзи 26 из 146 относится к классу joy positive (удовольствие позитивное настроение).

Для оценки эмоциональной окраски проведена автоматическая разметка данных на основе использованных в тексте эмóдзи. Для обучения и оценки алгоритмов машинного обучения были выбраны посты с 1 эмóдзи и длиной поста не более 11 токенов вместе с эмóдзи, получилось 9220 постов. В дальнейшем планируется исследовать тексты с другими параметрами по количеству эмóдзи и количеству токенов в посте.

Среди выбранных данных больше всего постов с эмóдзи из класса smile positive/negative (улыбка позитивное/негативное настроение), полученный набор данных является несбалансированным. Для эффективной работы с данными необходима их предобработка. Для этого из текста постов удалены id пользователей и групп, текст переведён в нижний регистр. Полученные 9220 постов автоматически размечены по 15 выделенным классам.

4. Эксперименты с моделями машинного и глубокого обучения

Выбор метода для анализа эмоциональной окраски текста зависит от требований решаемой задачи и характера набора данных. Для того, чтобы узнать, какой метод больше всего подходит для данного исследования, был проведён ряд экспериментов.

Тестовая выборка данных составляла 20% из общего числа постов. Для оценки эмоциональной окраски размеченных постов использовались классические методы машинного обучения из пакета scikit-learn. Для предобработки постов использовался лемматизатор pymorphy2.

Эксперименты проводились для текста с пунктуацией и с эмо́дзи, для текста без пунктуации и без эмо́дзи, для лемматизированного с помощью pymorphy2 текста с пунктуацией и с эмо́дзи.

Использовались представления слов в виде мешка слов (Bag of Words) [10], предобученные плотные векторные представления слов для русского языка из библиотеки Navec и Word2Vec [11] для классических методов машинного обучения, таких как, наивный байесовский классификатор (GaussianNB), логистическая регрессия (Logistic Regression), метод опорных векторов (SVM), градиентный бустинг (Gradient Boosting), случайный лес (Random Forest), классификатор дерева решений (DecisionTreeClassifier). Также проведены эксперименты с использованием ансамблей классификаторов с мажоритарным и мягким голосованием с помощью VotingClassifier.

Для экспериментов использовались нейросетевые модели: одномерная свёрточная нейросеть CNN, рекуррентная нейросеть LSTM (Long Short-Term Memory) и рекуррентная нейросеть GRU (Gated Recurrent Units).

5. Результаты оценки эмоциональной окраски текста постов

Для оценки качества классификации использовались метрики: F1-мера по макроусреднению (macro) и F1-мера по взвешенному усреднению (weighted). Выбор данных метрик для оценки эмоциональной окраски текста постов обусловлен тем, что полученный набор данных не является сбалансированным.

По F1-мере лучший результат получился для модели BoW +VotingClassifier (soft) (мешок слов + ансамблевый метод с мягким голосованием) на лемматизированном тексте с пунктуацией и с эмо́дзи, F1-мера macro равна 69.70%, F1-мера weighted равна 82.06%. Полученные результаты представлены в таблице 2 (лучшие результаты выделены жирным шрифтом).

Таблица 2. Результаты оценки эмоциональной окраски текста для классических методов машинного обучения

Модель	F1 macro, %	F1 weighted, %
1	2	3
BoW + Logistic Regression (текст с пунктуацией и с эмо́дзи)	51.74	78.97
BoW + SVC (текст с пунктуацией и с эмо́дзи)	40.26	72.21
BoW + RandomForestClassifier (текст с пунктуацией и с эмо́дзи)	65.10	80.54
BoW + DecisionTreeClassifier (текст с пунктуацией и с эмо́дзи)	63.79	79.45
BoW + GaussianNB (текст с пунктуацией и с эмо́дзи)	28.49	62.17
BoW + GradientBoostingClassifier (текст с пунктуацией и с эмо́дзи)	65.27	81.48
BoW + VotingClassifier (hard) (текст с пунктуацией и с эмо́дзи)	64.64	81.34
BoW + VotingClassifier (soft) (текст с пунктуацией и с эмо́дзи)	67.99	82.02
Navec + Logistic Regression (текст с пунктуацией и с эмо́дзи)	11.31	49.51
Navec + SVC (текст с пунктуацией и с эмо́дзи)	6.78	50.25
Navec + RandomForestClassifier (текст с пунктуацией и с эмо́дзи)	11.02	51.75

Продолжение таблицы 2

1	2	3
Navec + DecisionTreeClassifier (текст с пунктуацией и с эмо́дзи)	10.21	46.40
Navec + GaussianNB (текст с пунктуацией и с эмо́дзи)	1.15	4.34
Navec + GradientBoostingClassifier (текст с пунктуацией и с эмо́дзи)	8.19	49.25
Navec + VotingClassifier (hard) (текст с пунктуацией и с эмо́дзи)	10.50	51.68
Navec + VotingClassifier (soft) (текст с пунктуацией и с эмо́дзи)	11.02	52.07
Word2Vec + Logistic Regression (текст с пунктуацией и с эмо́дзи)	5.17	48.99
Word2Vec + SVC (текст с пунктуацией и с эмо́дзи)	5.17	48.99
Word2Vec + RandomForestClassifier (текст с пунктуацией и с эмо́дзи)	23.64	62.53
Word2Vec + DecisionTreeClassifier (текст с пунктуацией и с эмо́дзи)	15.22	54.54
Word2Vec + GaussianNB (текст с пунктуацией и с эмо́дзи)	1.63	7.60
Word2Vec + GradientBoostingClassifier (текст с пунктуацией и с эмо́дзи)	15.55	57.24
Word2Vec + VotingClassifier (hard) (текст с пунктуацией и с эмо́дзи)	11.76	54.77
Word2Vec + VotingClassifier (soft) (текст с пунктуацией и с эмо́дзи)	13.21	57.14
BoW + Logistic Regression (текст без пунктуации и без эмо́дзи)	8.33	52.19
BoW + SVC (текст без пунктуации и без эмо́дзи)	6.99	50.60
BoW + RandomForestClassifier (текст без пунктуации и без эмо́дзи)	15.36	53.06
BoW + DecisionTreeClassifier (текст без пунктуации и без эмо́дзи)	13.42	50.57
BoW + GaussianNB (текст без пунктуации и без эмо́дзи)	10.86	39.11
BoW + GradientBoostingClassifier (текст без пунктуации и без эмо́дзи)	12.58	51.35
BoW + VotingClassifier (hard) (текст без пунктуации и без эмо́дзи)	14.38	52.34
BoW + VotingClassifier (soft) (текст без пунктуации и без эмо́дзи)	14.80	53.75
Navec + Logistic Regression (текст без пунктуации и без эмо́дзи)	12.28	50.47
Navec + SVC (текст без пунктуации и без эмо́дзи)	7.67	50.36
Navec + RandomForestClassifier (текст без пунктуации и без эмо́дзи)	10.57	51.80
Navec + DecisionTreeClassifier (текст без пунктуации и без эмо́дзи)	10.63	44.41
Navec + GaussianNB (текст без пунктуации и без эмо́дзи)	1.05	1.05
Navec + GradientBoostingClassifier (текст без пунктуации и без эмо́дзи)	9.12	49.62
Navec + VotingClassifier (hard) (текст без пунктуации и без эмо́дзи)	11.41	51.92
Navec + VotingClassifier (soft) (текст без пунктуации и без эмо́дзи)	11.89	52.67
Word2Vec + Logistic Regression (текст без пунктуации и без эмо́дзи)	5.17	48.99
Word2Vec + SVC (текст без пунктуации и без эмо́дзи)	5.17	48.99
Word2Vec + RandomForestClassifier (текст без пунктуации и без эмо́дзи)	9.71	51.74
Word2Vec + DecisionTreeClassifier (текст без пунктуации и без эмо́дзи)	9.83	45.48
Word2Vec + GaussianNB (текст без пунктуации и без эмо́дзи)	0.03	0.01
Word2Vec + GradientBoostingClassifier (текст без пунктуации и без эмо́дзи)	8.58	49.09
Word2Vec + VotingClassifier (hard) (текст без пунктуации и без эмо́дзи)	8.78	50.66
Word2Vec + VotingClassifier (soft) (текст без пунктуации и без	8.91	50.82
BoW + Logistic Regression (лемматизированный текст с пунктуацией и с эмо́дзи)	51.49	78.64

Продолжение таблицы 2

1	2	3
BoW + SVC (лемматизированный текст с пунктуацией и с эмо́дзи)	40.16	72.11
BoW + RandomForestClassifier (лемматизированный текст с пунктуацией и с эмо́дзи)	66.77	81.43
BoW + DecisionTreeClassifier (лемматизированный текст с пунктуацией и с эмо́дзи)	66.30	79.28
BoW + GaussianNB (лемматизированный текст с пунктуацией и с эмо́дзи)	27.26	60.74
BoW + GradientBoostingClassifier (лемматизированный текст с пунктуацией и с эмо́дзи)	67.01	81.74
BoW + VotingClassifier (hard) (лемматизированный текст с пунктуацией и с эмо́дзи)	63.35	81.32
BoW + VotingClassifier (soft) (лемматизированный текст с пунктуацией и с эмо́дзи)	**69.70**	**82.06**
Navec + Logistic Regression (лемматизированный текст с пунктуацией и с эмо́дзи)	11.39	49.15
Navec + SVC (лемматизированный текст с пунктуацией и с эмо́дзи)	6.59	50.04
Navec + RandomForestClassifier (лемматизированный текст с пунктуацией и с эмо́дзи)	11.14	51.98
Navec + DecisionTreeClassifier (лемматизированный текст с пунктуацией и с эмо́дзи)	9.89	46.18
Navec + GaussianNB (лемматизированный текст с пунктуацией и с эмо́дзи)	1.14	4.83
Navec + GradientBoostingClassifier (лемматизированный текст с пунктуацией и с эмо́дзи)	7.35	49.20
Navec + VotingClassifier (hard) (лемматизированный текст с пунктуацией и с эмо́дзи)	10.50	51.60
Navec + VotingClassifier (soft) (лемматизированный текст с пунктуацией и с эмо́дзи)	10.91	51.73
Word2Vec + Logistic Regression (лемматизированный текст с пунктуацией и с эмо́дзи)	5.17	48.99
Word2Vec + SVC (лемматизированный текст с пунктуацией и с эмо́дзи)	5.17	48.99
Word2Vec + RandomForestClassifier (лемматизированный текст с пунктуацией и с эмо́дзи)	25.85	64.19
Word2Vec + DecisionTreeClassifier (лемматизированный текст с пунктуацией и с эмо́дзи)	18.91	55.79
Word2Vec + GaussianNB (лемматизированный текст с пунктуацией и с эмо́дзи)	1.40	6.56
Word2Vec + GradientBoostingClassifier (лемматизированный текст с пунктуацией и с эмо́дзи)	18.77	56.97
Word2Vec +VotingClassifier (hard) (лемматизированный текст с пунктуацией и с эмо́дзи)	13.93	55.47
Word2Vec +VotingClassifier (soft) (лемматизированный текст с пунктуацией и с эмо́дзи)	18.09	59.61

Также для экспериментов использовались нейросетевые модели: одномерная свёрточная нейросеть CNN, рекуррентная нейросеть LSTM (Long Short-Term Memory) и рекуррентная нейросеть GRU (Gated Recurrent Units).

Лучший результат среди использованных нейросетевых моделей показала рекуррентная нейросеть GRU на 15 эпохах обучения: F1-мера macro равна 48.77%, F1-мера weighted равна 83.74%. Полученные результаты представлены в таблице 3 (лучшие результаты выделены

жирным шрифтом).

Таблица 3. Результаты оценки эмоциональной окраски текста для нейросетевых методов

Модель	F1 macro, %	F1 weighted, %
Одномерная свёрточная нейросеть (токенизатор Keras, optimizer='adam', epochs=5)	17.42	71.54
Рекуррентная нейросеть LSTM (токенизатор Keras, optimizer='adam', epochs=5)	11.85	62.66
Рекуррентная нейросеть GRU (токенизатор Keras, optimizer='adam', epochs=5)	28.29	78.42
Одномерная свёрточная нейросеть (токенизатор Keras, optimizer='adam', epochs=10)	40.20	81.86
Рекуррентная нейросеть LSTM (токенизатор Keras, optimizer='adam', epochs=10)	23.44	78.39
Рекуррентная нейросеть GRU (токенизатор Keras, optimizer='adam', epochs=10)	43.15	83.85
Одномерная свёрточная нейросеть (токенизатор Keras, optimizer='adam', epochs=15)	27.77	78.96
LSTM (токенизатор Keras, optimizer='adam', epochs=15)	29.34	79.81
Рекуррентная нейросеть GRU (токенизатор Keras, optimizer='adam', epochs=15)	**48.77**	**83.74**

Получили, что в случае макроусреднения, т. е. когда всем классам даётся одинаковый вес, независимо от их количества в наборе данных, лучший результат F1-мера macro = 69.70% для модели BoW +VotingClassifier (soft) (мешок слов + ансамблевый метод с мягким голосованием) на лемматизированном тексте с пунктуацией и с эмóдзи. В случае же взвешенного усреднения, т. е. когда вес классам даётся согласно количеству объектов в этих классах, лучший результат F1-мера weighted = 83.74% для модели рекуррентной нейросети GRU на 15 эпохах обучения.

В дальнейшем планируется сравнить полученные результаты по метрикам качества классификации F1-меры с результатами работы модели RuBERT, которая позволяет работать с текстами на русском языке и имеет качественные распределённые векторные вложения (embeddings) текстов.

6. Заключение

В данной статье представлена оценка эмоциональной окраски постов из социальной сети «ВКонтакте», описан процесс получения, обработки и использования полученного набора данных. Приводятся результаты экспериментов с использованием методов машинного и глубокого обучения с оценкой работы методов по метрикам качества классификации. По оценке качества классификации текста постов лучший результат по метрике F1-мера macro = 69.70% показала модель BoW +VotingClassifier (soft) (мешок слов + ансамблевый метод с мягким голосованием) на лемматизированном тексте с пунктуацией и с эмóдзи. Лучший результат по метрике качества классификации F1-мера weighted получен для модели рекуррентной нейросети GRU F1-мера weighted = 83.74%.

Поскольку эксперты не размечали полученные данные, а использовалась автоматическая разметка постов на основании, встречающихся в этих постах эмóдзи, в дальнейшем планируется провести экспертную оценку полученной автоматической разметки постов по выделенным классам.

Также планируется провести эксперименты на сбалансированных данных. В будущем можно продолжить исследование с использованием текстов с другими параметрами по количеству эмóдзи и токенов в тексте.

Литература

[1] Calvo R. A., D'Mello S. Affect detection: An interdisciplinary review of models, methods and their applications // IEEE Transactions on affective computing. 2010. Vol. 1 (1). P. 18–37.
[2] Кузнецова Е. С., Лукашевич Н. В., Четверкин И. И. Тестирование правил для системы анализа тональности // Компьютерная лингвистика и интеллектуальные технологии: по материалам международной конференции Диалог 2013. М.: Изд-во РГГУ, 2013. Вып. 12. Т. 2. С. 71–80.
[3] Loukachevitch N., Levchik A. Creating a general Russian sentiment lexicon // Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). 2016. P. 1171–1176.
[4] Kiritchenko S., Zhu X., Mohammad S. Sentiment analysis of short informal texts // Journal of Artificial Intelligence Research. 2014. Vol. 50. P. 723–762.
[5] Joshi A., Bhattacharyya P., Carman M. Automatic sarcasm detection: A survey // ACM Computing Surveys (CSUR). 2017. Vol. 50 (5). P. 1–22.
[6] Pang B., Lee L., Vaithyanathan S. Thumb up? Sentiment Classification using Machine Learning Techniques // Proceedings of Conference on Empirical Methods in Natural Language Processing EMNLP-2002. 2002. P. 79–86.
[7] Zhang L., Wang S., Liu B. Deep learning for sentiment analysis: A survey // Wiley Reviews: Data Mining and Knowledge Discovery. 2018. Vol. 8 (4).
[8] Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019. Vol. 1. P. 4171–4186.
[9] Куратов Ю., Архипов М. Адаптация глубоких двунаправленный многоязычных моделей на основе архитектуры Transformer для русского языка // Компьютерная лингвистика и интеллектуальные технологии: по материалам международной конференции Диалог 2019. М.: Изд-во РГГУ, 2019. Вып. 19 (25). С. 333–339.
[10] HaCohen-Kerner Y., Miller D., Yigal Y. The influence of preprocessing on text classification using a bag-of-words representation // PloS one. 2020. Vol. 15 (5).
[11] Mikolov, T. et al. Distributed representations of words and phrases and their compositionality // Proceedings of the 26th international conference on neural information processing systems. 2013. Vol. 2. P. 3111–3119.

Evaluation of Emotionality of Posts with Emojis in the VKontakte Social Network Using Machine and Deep Learning Methods

Anna P. Bykova

Saint Petersburg State University

In this paper possibilities of applying machine and deep learning methods to emotionality evaluation of posts text with emojis from the VKontakte social network are investigated. An unbalanced data set with posts text is described, the data set is annotated by 15 classes. In these classes emotional and tonal components of text are taken into account. Experiments are conducted on the obtained data set with using 6 classical machine learning methods, their ensembles with hard and soft voting, and 3 neural network methods. The best result by classification quality metrics was obtained for the Bag of words + VotingClassifier ensemble method with soft voting on lemmatized text with punctuation and emoji: F1-macro measure = 69.70%, F1-weighted measure = 82.06% and for the recurrent neural network GRU on 15 epochs of training: F1-measure macro = 48.77%, F1-measure weighted = 83.74%.

Keywords: emotion analysis, evaluation of emotionality in text, emoji, machine learning, neural networks

Reference for citation: Bykova A.P. Evaluation of Emotionality of Posts with Emojis in the VKontakte Social Network Using Machine and Deep Learning Methods // Computational Linguistics and Computational Ontologies. Vol. 7 (Proceedings of the XXVI International Joint Scientific Conference «Internet and Modern Society», IMS-2023, St. Petersburg, June 26–28, 2023). — St. Petersburg: ITMO University, 2024. P. 12–20. DOI: 10.17586/2541-9781-2024-7-12–20

Reference

[1] Calvo R. A., D'Mello S. Affect detection: An interdisciplinary review of models, methods and their applications // IEEE Transactions on affective computing. 2010. Vol. 1 (1). P. 18–37.

[2] Kuznecova E. S., Lukashevich N. V., Chetverkin I.I. Testirovanie pravil dlya sistemy analiza tonal'nosti // Komp'yuternaya lingvistika i intellektual'nye tekhnologii: po materialam mezhdunarodnoj konferencii Dialog 2013. M.: Izd-vo RGGU, 2013. Vyp. 12. T. 2. S. 71–80. (in Russian)[3] Loukachevitch N., Levchik A. Creating a general Russian sentiment lexicon // Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). 2016. P. 1171–1176.

[4] Kiritchenko S., Zhu X., Mohammad S. Sentiment analysis of short informal texts // Journal of Artificial Intelligence Research. 2014. Vol. 50. P. 723–762.

[5] Joshi A., Bhattacharyya P., Carman M. Automatic sarcasm detection: A survey // ACM Computing Surveys (CSUR). 2017. Vol. 50 (5). P. 1–22.

[6] Pang B., Lee L., Vaithyanathan S. Thumb up? Sentiment Classification using Machine Learning Techniques // Proceedings of Conference on Empirical Methods in Natural Language Processing EMNLP-2002. 2002. P. 79–86.

[7] Zhang L., Wang S., Liu B. Deep learning for sentiment analysis: A survey // Wiley Reviews: Data Mining and Knowledge Discovery. 2018. Vol. 8 (4).

[8] Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019. Vol. 1. P. 4171–4186.

[9] Kuratov Yu., Arhipov M. Adaptaciya glubokih dvunapravlennyj mnogoyazychnyh modelej na osnove arhitektury Transformer dlya russkogo yazyka // Komp'yuternaya lingvistika i intellektual'nye tekhnologii: po materialam mezhdunarodnoj konferencii Dialog 2019. M.: Izd-vo RGGU, 2019. Vyp. 19 (25). S. 333–339. (in Russian)

[10] HaCohen-Kerner Y., Miller D., Yigal Y. The influence of preprocessing on text classification using a bag-of-words representation // PloS one. 2020. Vol. 15 (5). DOI: 10.1371/journal.pone.0232525.

[11] Mikolov, T. et al. Distributed representations of words and phrases and their compositionality // Proceedings of the 26th international conference on neural information processing systems. 2013. Vol. 2. P. 3111–3119.