УДК 004.048 DOI: 10.26102/2310-6018/2020.30.3.014

Исследование задачи классификации публикаций социальных сетей на предмет выявления положительного отношения

М.А. Сазонов, С.В. Шекшуев

Академия ФСО России, Орел, Российская Федерация
Орловский государственный университет имени И.С. Тургенева, Орел, Российская Федерация

Резюме. В статье рассматривается актуальность решения класса задач, связанных с анализом публикационной активности пользователей социальных сетей. Приводится анализ существующих подходов к выявлению общественного мнения к публикациям в социальных сетях, в котором обосновывается превалирование методов, основанных на анализе тональности текстов. Приводятся недостатки указанных методов, снижающие эффективность процесса оценивания общественного мнения относительно публикационной активности пользователей социальных сетей. Выдвигается предположение о возможности использования метаданных сообщений без необходимости проведения процедуры анализа тональности текста для устранения указанной проблемы. Определяются первичные и производные показатели сообщений в социальных сетях, получаемые из совокупности метаданных. Рассматриваются подходы к решению задачи бинарной классификации на основе указанных показателей, как на базе статистических методов, так и с использованием методов машинного обучения. Делается предположение о приемлемой точности класса моделей на основе машинного обучения, обеспечивающих решение указанной задачи. Предлагается модель машинного обучения на основе случайного леса для решения задачи классификации положительного отношения к публикациям в социальных сетях, основанная на анализе первичных и производных показателей сообщений.

Ключевые слова: социальная сеть, данные, показатели социальных сетей, машинное обучение, случайный лес.

Для цитирования: Сазонов М.А., Шекшуев С.В. Исследование задачи классификации публикаций социальных сетей на предмет выявления положительного отношения. *Моделирование, оптимизация и информационные технологии.* 2020;8(3). Доступно по: https://moit.vivt.ru/wp-content/uploads/2020/08/SazonovShekshuev_3_20_1.pdf DOI: 10.26102/2310-6018/2020.30.3.014.

The researching of the social networks publications classification problem on the subject of positive attitude identification

M.A. Sazonov, S.V. Shekshuev

FSO Academy of Russia, Orel, Russian Federation
Oryol State University named after I.S. Turgenev, Orel, Russian Federation

Abstract. In article discusses the relevance of solving problems class publication activity analysis for users of social networks. An analysis of existing approaches identifying public opinion about publications in social networks is given, in which the prevalence is substantiated of methods based on the analysis of the texts sentiment. The disadvantages of these methods are given, which reduce the process of assessing public opinion regarding the publication activity of users of social networks efficiency. It is suggested that it is possible to use message metadata without the need a texts sentiment analysis procedure to eliminate this problem. The primary and derived indicators of messages in social networks are determined, obtained from the set of metadata. Approaches to solving the problem of

binary classification based on the indicated markers, both based on statistical methods and using machine learning methods, are considered. An assumption is made about the acceptable accuracy of a class of models based on machine learning that provide a solution to the specified problem. A machine learning model based on a random forest is proposed for solving the problem of classifying a positive attitude towards publications in social networks, based on the analysis of primary and derived indicators of messages.

Keywords: social network, data, social networks publications indicators, machine learning, random forest.

For citation: Sazonov M.A., Shekshuev S.V. The researching of the social networks publications classification problem on the subject of positive attitude identification. *Modeling, Optimization and Information Technology*. 2020;8(3). Available from: https://moit.vivt.ru/wp-content/uploads/2020/08/SazonovShekshuev_3_20_1.pdf DOI: 10.26102/2310-6018/2020.30.3.014 (In Russ).

Введение

Актуальность изучения общественного мнения сегодня велика. Общественное мнение представляет собой специфический способ существования и проявления массового сознания, посредством которого публично выражается духовное или духовно-практическое отношение большинства к актуальным для него фактам, событиям, явлениям и процессам действительности [1]. В современных условиях источником информации для изучения общественного мнения являются компьютерные социальные сети.

В настоящее время задача определения отношения к выбранному объекту, как частный случай задачи изучения общественного мнения, решается в основном на основе анализа тональности текстов публикаций, сообщений, комментариев [7]. Однако данный подход обладает рядом недостатков, самым существенным из которых авторы считают:

– наличие в тексте публикации сарказма, иронии, юмора, которые классификатор, основанный на анализе тональности, не может точно определить;

– наличии эмодзи (emoji – графический язык, где вместо букв используются картинки), значение которых не всегда явно;

– текст может отсутствует совсем, а вместо него находится мультимедийный контент.

Анализ задач, решаемых с помощью оценивания тональности текстов публикаций, показал, что в подавляющем большинстве случаев исследователей интересует тональность не одного конкретного текста сообщения, а множества сообщений, относящихся к одному объекту (публикации, лицу, факту и т. д.). Поиск решения задачи оценивания отношения к публикации, позволяющего нивелировать указанные выше недостатки, привел к идее исследовать не сам текст сообщения, а метаданные множества сообщений (как первичные, так и производные показатели) без анализа текста.

Фактически, подход реализует задачу многоклассовой классификации публикаций социальных сетей, где каждый класс публикации есть класс тональности. Для наглядной демонстрации этого метода, а также для упрощения вычислений в экспериментальной части задача многоклассовой классификации была сведена к задаче бинарной классификации.

Для реализации классификатора определим два класса публикаций социальной сети $Y = \{y_p, y_n\}$, где y_p – класс объектов социальной сети, к которым отношение некой социальной группы выражено положительно, а y_n – класс объектов социальной сети, к которым отношение некой социальной группы выражено не положительно, то есть либо

отрицательно, либо нейтрально, либо иным образом. Обозначим набор неких показателей социальной сети как множество:

$$X = \{x_0 ... x_m\} \quad (1)$$

где m – количество таких показателей. Предположим, что существует некий функционал F, такой, что

$$\forall x_m \in X, \exists x_m F(x_m) = y, y \in Y \quad (2)$$

Выражение (2) фактически представляет из себя постановку задачи бинарной классификации.

Производные показатели публикаций социальных сетей

На Рисунке представлен график изменения абсолютного количества комментариев к публикациям. Интервал времени – 10 минут. В качестве значений абсолютного количества комментариев в конкретный момент времени указано среднее значение абсолютного количества комментариев к публикации в этот же момент времени для каждой из подвыборок – подвыборки, содержащей экземпляры класса положительных y_p публикаций и подвыборки, содержащей экземпляры класса не положительных y_n публикаций.

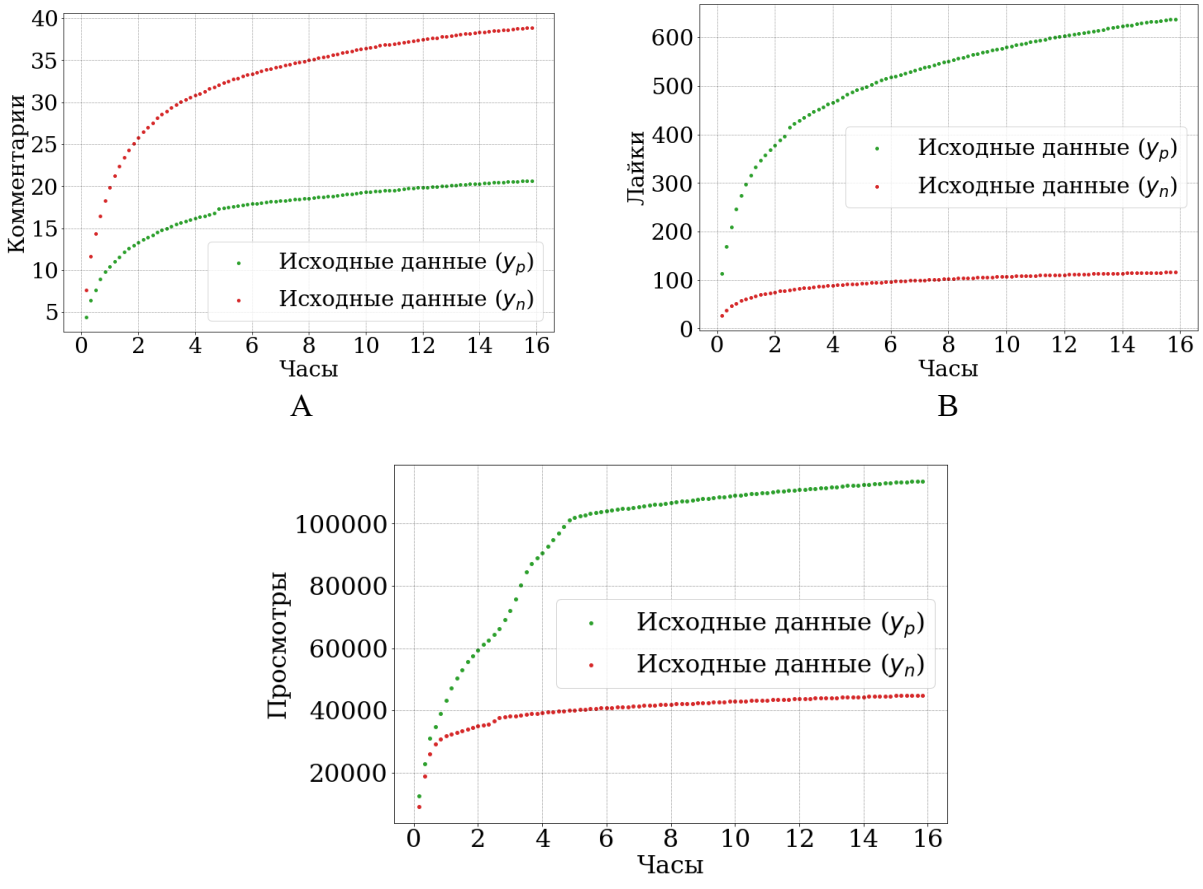

Рисунок – Графики распределения абсолютного количества комментариев (A), лайков (B) и просмотров (C) к публикации в конкретный момент времени

Figure - Graphs of the distribution of the absolute number of comments (A), likes (B) and views (C) for a publication at a particular point in time

Примечание. В дальнейшем на графиках в качестве значений оси ординат будет указано среднее значение по выборке соответствующего показателя. Такое решение принято в связи с тем, что нецелесообразно на графиках изображать значения каждого объекта выборки из-за ее большого объема, что повлечет за собой захламленность графиков. Среднее значение позволяет увидеть наглядную разницу между классами.

Анализ графиков показывает, что разница в количестве комментариев, а также во времени их появления зависит от характера публикации – посты положительной направленности пользователи комментируют меньше. Оценок «нравится» больше у положительных y_p публикаций, тогда как просмотров больше у не положительных y_n.

Зависимости, изображенные на графиках на рисунке 1, заданы табличными функциями (Таблица 1), анализ которых показал, что для каждой из них существует некая непрерывная функция $y = f(t)$ непрерывного аргумента *t*, причем явный вид этой функции неизвестен. Для нахождения приближенной функции $y = f(t)$, т. е. фактически для вычисления значения *y* для произвольного *t* существуют методы интерполяции и аппроксимации [4].

Таблица 1 – Пример табличной функции
Table 1 – Example of a table function

t_i	t_1	t_2	t_3	…	t_n
$f(t_i)$	y_1	y_2	y_3	…	y_n

С целью избегания больших погрешностей рекомендуется использовать кусочно-полиномиальную интерполяцию, когда весь отрезок [*a*, *b*] разбивают на частичные отрезки и на каждом частичном отрезке приближенно заменяют исходную функцию многочленом невысокой степени. Одним из способов интерполирования на всем отрезке является интерполирование с помощью сплайн-функций. Сплайн-функцией или сплайном называют кусочно-полиномиальную функцию, определенную на отрезке [*a*, *b*] и имеющую на этом отрезке некоторое число непрерывных производных [5]. В качестве сплайн-функции применены кубические сплайны, которые стали классическими интерполяционными функциями [2].

Если предположить, что длина вектора значений условного признака *y* равна $|y| = N + 1$, то кубический сплайн состоит из *N* звеньев, каждое из которых есть кубический полином, который включает четыре коэффициента и один признак $x_{i-1} \in X, 1 \le i \le N$ [2]:

$$S_3(x) = \sum_{i=1}^{N} (a_i + b_i(x - x_{i-1}) + c_i(x - x_{i-1})^2 + d_i(x - x_{i-1})^3), \quad (3)$$

где $x_{i-1} \le x \le x_i$.

В некоторых случаях (например, при наличии погрешности измерений или выбросов) лучше использовать аппроксимацию приближенной функцией, проходящей близко от табличных значений. Для выбранной функции определяются параметры, дающие наилучшее приближение, например, с использованием метода наименьших квадратов (МНК) [3]. Критерий наилучшего приближения базируется на

минимизации отклонений значений построенной функции $f(x)$ в узлах x_i от соответствующих чисел y_i [4].

Используя имеющиеся первичные показатели, а также их изменение во времени, можно образовать производные показатели социальной сети. Приблизив заданные табличные функции методами аппроксимации и интерполяции, в качестве производных показателей предлагается использовать параметры полученных функций. Это может быть степень полинома при полиномиальной аппроксимации, степень экспоненты, основание логарифма, коэффициенты сплайнов и прочее. Основой выбора параметров является то, что с помощью них функция, приближающая исходные зависимости, должна быть максимально описана.

Экспериментальная проверка гипотезы

Для проверки указанной ранее гипотезы разработан программный продукт, загружающий публикации и их первичные показатели, в течение заданного времени из заданных публичных сообществ социальной сети Вконтакте через заданные временные интервалы.

За несколько дней работы программного продукта из социальной сети Вконтакте загружено 1189 публикаций. Указанные публикации были отнесены к одному из двух классов $Y = \{y_p, y_n\}$ группой экспертов. В соответствии с мнением экспертов, количество положительных публикаций y_p составило 548, не положительных y_n – 641.

В качестве алгоритма машинного обучения для эксперимента был выбран алгоритм случайного леса (*Random Forest*).

Для оценки качества алгоритма классификации необходимо определить метрики качества. Для этого целесообразно ввести понятие матрицы исходов (Таблица 2) [6].

Таблица 2 – Матрица исходов
Table 2 – Outcome matrix

	$y = y_i$	$y \neq y_i$
$F(x_k) = y_i$	*True Positive (TP)*	*False Positive (FP)*
$F(x_k) \neq y_i$	*False Negative (FN)*	*True Negative (TN)*

Данная матрица состоит из следующих элементов:

1. верные срабатывания (TN – True Positive) – количество объектов из класса y_i которые алгоритм отнес к классу y_i;
2. ложные срабатывания (FP – False Positive) – количество объектов не из класса y_i, которые алгоритм ошибочно отнес к классу y_i;
3. ложный пропуск (FN – False Negative) – количество объектов из класса y_i, которые алгоритм не отнес к классу y_i;
4. верный пропуск (TN – True Negative) – количество объектов не из класса y_i, которые алгоритм не отнес к классу y_i.

В качестве критериев оценки качества классификации выбраны показатели доли правильных ответов (*accuracy*), точности (*precision*), полноты (*recall*), F-меры, площади под ROC-кривой и площади под кривой точности-полноты [6].

Доля правильных ответов *Accuracy* алгоритма есть отношение количества правильно классифицированных объектов к общему числу объектов:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}. \quad (4)$$

Показатель точности классификации *Precision* есть доля правильно классифицированных объектов. Точность вычисляется как отношение количества правильно отнесенных алгоритмом объектов к классу y_i к общему количеству объектов, отнесенных алгоритмом к классу y_i:

$$Precision = \frac{TP}{TP + FP}. \quad (5)$$

Показатель полноты классификации *Recall* – это доля правильно классифицированных алгоритмом объектов. Полнота вычисляется как отношение количества правильно отнесенных алгоритмом объектов к классу y_i к общему числу объектов в классе y_i:

$$Recall = \frac{TP}{TP + FN}. \quad (6)$$

F-мера является гармоническим средним полноты и точности, придает им одинаковый вес, вследствие чего F-мера будет падать одинаково при уменьшении как точности, так и полноты:

$$F = \frac{2 \times Precision \times Recall}{Precision + Recall}. \quad (7)$$

ROC-кривая представляет собой графическую характеристику качества бинарного классификатора – зависимость доли верных положительных классификаций от доли ложных положительных классификаций при варьировании уровня отсечения решающего правила [8]. Показателем, по которому можно оценить кривую, является площадь под ROC-кривой (AUC ROC – area under ROC curve), принимающий значения от 0,5 до 1, и чем ближе значение AUC ROC к 1, тем классификатор лучше. Если же значение площади под ROC-кривой близко к 0.5, имеется смысл полагать о том, что алгоритм классифицирует объекты случайным образом.

Еще один показатель оценки качества классификации – показатель площади под кривой точности-полноты (AUC PR – area under precision-recall curve). Высокое значение, близкое к 1.00, показывает, что классификатор работает точно и с высокой полнотой [9].

Таблица 3 – Результаты эксперимента
Table 3 – Experimental results

Алгоритм	accuracy	precision	recall	f1	roc_auc	pr_auc
Random forest	0,74	0,77	0,63	0,69	0,73	0,79

В Таблице 3 показаны средние значения показателей работы алгоритма, поскольку оценка проходила методом скользящего контроля, т. е. с помощью процедуры эмпирического оценивания обобщающей способности алгоритмов, которая заключается в том, что фиксируется некоторое множество разбиений исходной выборки на две подвыборки – обучающую и контрольную. Для каждого разбиения выполняется настройка алгоритма по обучающей подвыборке, а затем оценивается его средняя

ошибка на объектах контрольной подвыборки. Оценкой скользящего контроля называется средняя по всем разбиениям величина ошибки на контрольных подвыборках.

Заключение

Экспериментальные исследования показали, что при достаточном количестве информации в виде первичных параметров, для оценивания характера отношения пользователей аккаунтов к некоторому факту необязательно обращаться к существующим методам анализа тональности текста публикаций, так как возможно делать такие выводы, используя предложенные показатели социальной сети и методы их анализа. При этом выбранные критерии классификации будут иметь среднее значение в районе 0.7.

Этот вывод был получен, исходя из следующих сведений, полученных после проведения эксперимента:

1. Имеется корреляция между вектором признаков объекта $x = (f_1(x), ..., f_n(x))$ и его принадлежностью к одному из классов $Y = \{y_p, y_n\}$, и данная корреляция нелинейна;
2. Указанную зависимость возможно аппроксимировать ансамблевыми методами классификации, в основе которых лежат деревья решений, в данном случае – случайным лесом.

Эксперимент показал, что публикации из социальной сети Вконтакте, представленные в виде объекта $x = (f_1(x), ..., f_n(x))$ с ранее обоснованно выбранными показателями возможно использовать для их анализа и решения задачи бинарной классификации на предмет оценивания положительного отношения.

ЛИТЕРАТУРА

1. Франц В.А. Управление общественным мнением: учеб. Пособие. М-во образования и науки Рос. Федерации, Урал. федер. ун-т. Екатеринбург: Изд-во Урал. ун-та. 2016:135.
2. Беликова Г.И., Бровкина Е.А., Вагер Б.Г., Витковская Л.В., Матвеев Ю.Л. Численные методы. Учебное пособие. СПб., РГГМУ. 2019:174.
3. Лоусон Ч., Хенсон Р., Численное решение задач метода наименьших квадратов; Пер. с англ. М.: Наука. Гл. ред. физ.-мат. лит. 1986:232.
4. Фадеев М.А., Марков К.А. Численные методы: учебное пособие. ННГУ им. Н.И. Лобачевского. 2010.
5. Самарский А.А., Гулин А.В. Численные методы: учебное пособие для вузов. М.: Наука. Гл. ред. физ-мат. лит. 1989:432.
6. Davis J., Goadrich M. (2006). The Relationship Between Precision-Recall and ROC Curves. *Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA.*
7. Будыльский Д.В. Автоматизация мониторинга общественного мнения на основе интеллектуального анализа сообщений в социальных сетях: дис. … канд. техн. наук. Брянский гос. техн. университет, Брянск. 2015.
8. Гуськов С.Ю., Лёвин В.В. Интервальные доверительные оценки для показателей качества бинарных классификаторов ROC-кривых, AUC для случая малых выборок. *Инженерный журнал: наука и инновации.* 2015;3. URL: http://engjournal.ru/catalog/mesc/idme/1376.html.
9. Мюллер А., Гвидо С. Введение в машинное обучение с помощью Python. Москва, 2016-2017.

REFERENCES

1. Franc V.A. Upravlenie obshchestvennym mneniem: ucheb. Posobie. M-vo obrazovaniya i nauki Ros. Federacii, Ural. feder. un-t. Ekaterinburg: Izd-vo Ural. un-ta. 2016:135.
2. Belikova G.I., Brovkina E.A., Vager B.G., Vitkovskaya L.V., Matveev YU.L. CHislennye metody. Uchebnoe posobie. SPb., RGGMU. 2019:174.
3. Louson CH., Henson R., CHislennoe reshenie zadach metoda naimen'shih kvadratov; Per. s angl. M.: Nauka. Gl. red. fiz.-mat. lit. 1986:232.
4. Fadeev M.A., Markov K.A. CHislennye metody: uchebnoe posobie. NNGU im. N.I. Lobachevskogo. 2010.
5. Samarskij A.A., Gulin A.V. CHislennye metody: uchebnoe posobie dlya vuzov. M.: Nauka. Gl. red. fiz-mat. lit. 1989:432.
6. Davis J., Goadrich M. (2006). The Relationship Between Precision-Recall and ROC Curves. Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA.
7. Budyl'skij D.V. Avtomatizaciya monitoringa obshchestvennogo mneniya na osnove intellektual'nogo analiza soobshchenij v social'nyh setyah: dis. … kand. tekhn. nauk. Bryanskij gos. tekhn. universitet, Bryansk. 2015.
8. Gus'kov S.YU., Lyovin V.V. Interval'nye doveritel'nye ocenki dlya pokazatelej kachestva binarnyh klassifikatorov ROC-krivyh, AUC dlya sluchaya malyh vyborok. Inzhenernyj zhurnal: nauka i innovacii. 2015;3. URL: http://engjournal.ru/catalog/mesc/idme/1376.html.
9. Myuller A., Gvido S. Vvedenie v mashinnoe obuchenie s pomoshch'yu Python. Moskva, 2016-2017.

ИНФОРМАЦИЯ ОБ АВТОРЕ / INFORMATIONS ABOUT AUTHORS

Сазонов Михаил Анатольевич, сотрудник, Академия ФСО России, Орел, Российская Федерация.
e-mail: sma77@list.ru

Mikhail A. Sazonov, employee, FSO Academy of Russia, Oryol State University named after I.S. Turgenev, Orel, Russian Federation.

Шекшуев Сергей Васильевич, аспирант, Орловский государственный университет имени И.С. Тургенева, Орел, Российская Федерация.
e-mail: sergei.shekshuev@gmail.com

Sergey V. Shekshuev, аспирант, Oryol State University named after I.S. Turgenev Orel, Russian Federation.