ВЕБ-ПРИЛОЖЕНИЕ ДЛЯ ОЦЕНКИ КОЛИЧЕСТВА ПРОСМОТРОВ ИНТЕРНЕТ-ПУБЛИКАЦИЙ

Пучило Т.Н.[1], Щегрикович Д.В.[2] Email: Puchylo636@scientifictext.ru

[1]Пучило Татьяна Николаевна – магистрант;
[2]Щегрикович Дмитрий Васильевич – кандидат физико-математических наук, доцент,
кафедра интеллектуальных систем,
Белорусский государственный университет, г. Минск, Республика Беларусь

Аннотация: *целью данной работы является разработка веб-приложения для предсказания количества просмотров интернет-публикаций на основе содержащейся в них текстовой информации. В работе для достижения поставленной цели решаются такие задачи, как сравнительный анализ методов машинного обучения и разработка с использованием методов обработки естественного языка функций для извлечения признаков из текстов публикаций. Описываются классификация извлекаемых 120 признаков и преимущества использования выбранного метода машинного обучения «Случайный лес».*
Ключевые слова: *обработка естественного языка, корпусная лингвистика, машинное обучение, бинарная классификация, извлечение признаков, веб-приложение.*

WEB APPLICATION FOR EVALUATION OF NUMBER OF ONLINE PUBLICATIONS VIEWS

Puchylo T.N.[1], Shchegrikovich D.V.[2]

[1]Puchylo Tatsiana Nikolaevna – Graduate Student;
[2]Shchegrikovich Dmitry Vasilevich – Candidate of Physico-Mathematical Sciences, Associate Professor,
INTELLIGENT SYSTEMS DEPARTMENT,
BELARUSIAN STATE UNIVERSITY, MINSK, REPUBLIC OF BELARUS

Abstract: *the goal is to develop web application for prediction of viewing count of news online publications based on the text of these publications. This has been done by using natural language processing for features extraction functions creating and machine learning algorithms comparison. Upon examination of these algorithms, it becomes clear to use 120 features and Random forest method to estimate of viewing count of news online publications. Web application was created using microframework "Flask" on Python.*
Keywords: *natural language processing, corpus linguistics, machine learning, binary classification, feature extraction, web application.*

УДК 004.8

Введение

Новостные ресурсы, составляющие значительную часть от общего объёма всех ресурсов в сети Интернет, ежедневно публикуют более тысячи публикаций. Продвижение таких публикаций в сети Интернет – сложная динамически развивающаяся отрасль. Данный процесс требует денежных затрат, поэтому авторам таких публикаций выгодно понимать возможности их популяризации. Создание публикаций, способных распространяться самостоятельно, также является ресурсоемким процессом. Всё это обуславливает необходимость создания инструмента, способного оценить будущую популярность новостных публикаций, тем самым позволив редакторам оптимизировать их содержание, а в итоге и вовсе уменьшить затраты на рекламный бюджет. Прогноз популярности новостных публикаций может быть интересен не только редакторам и журналистам, но и содержателям новостных ресурсов с целью расширения функционала, и рекламодателям для более качественного планирования рекламного бюджета и его распределения по рекламным каналам.

Веб-страница, на которой размещается текстовый материал статьи, кроме текста, содержит информацию других типов, а также сама обладает характеристиками, влияющими на восприятие пользователем информации: веб-дизайн, опыт взаимодействия. Интерес для данной работы представляет прогнозирование популярности публикаций на основе текстовой информации, содержащейся в ней.

В существующих работах часть исследователей акцентирует внимание на содержимое публикуемого материала, часть – на структуре ресурса, на котором он публикуется. Общим недостатком для всех этих работ является то, что оценка будущей популярности публикации осуществляется с использованием данных, собранных непосредственно после публикации материала [1-3].

Исходя из имеющегося интереса к оптимизации текстов новостных публикаций до их публикации на ресурсах, а также предпочтения совершать это в режиме реального времени, для мгновенного редактирования текстов, целью данной работы является разработка веб-приложения, способного

предсказывать количество просмотров интернет-публикаций до выхода материала на основе содержащейся в нем текстовой информации.

Извлечение признаков из текстовой информации

Текстовая информация в контексте данной работы является информацией, представленной на естественном языке – языке, используемом для общения людей. Для анализа такого рода информации используется направление «автоматическая обработка естественного языка», которое изучает проблемы компьютерного анализа и синтеза естественного языка [4-6]. Признаки для описания текста новостной публикации, выделяемые с помощью методов обработки естественного языка, были классифицированы следующим образом.

Метапризнаки. Метапризнаки содержат подробную информацию о каждой публикации, доступную любому пользователю на веб-странице: название, краткое описание статьи, автор, дата и время публикации. Так как в рамках одного ресурса веб-страницы имеют общую структуру HTML-кода, получение метапризнаков было реализовано с использованием регулярных выражений.

Графематические признаки. В процессе графематического анализа текста происходит определение элементов грамматической структуры: от количества слов в предложениях до количества абзацев в тексте [5]. Часть графематических признаков (количество абзацев, цифр, заголовков и использований тегов «b», «strong» и других элементов форматирования текста), по аналогии с метапризнаками, были извлечены из HTML-кода веб-страниц с использованием шаблонов регулярных выражений. Для вычисления количественных характеристик предложений и слов использовались тривиальные функции, работающие с числом пробелов и пунктуационных знаков. Вычисляемыми графематическими признаками стали количество слов, предложений, символов, букв, цифр, абзацев, использований элементов форматирования, заголовков, а также статистические моменты и такие характеристики, как минимальное и максимальное значения, всех описанных величин.

Морфологические признаки. На этапе морфологического анализа происходит определение принадлежности каждой словоформы к определенному слову, имеющемуся в размеченном словаре; классов слов; разбиение словоформ на части слов; грамматических признаков для каждой словоформы. Например, для существительных такими признаками являются род, число, падеж и склонение. Морфологическими признаками, характеризующими текст, были выделены не только признаки, относящиеся к морфологическим свойствам слов, но и признаки, которые можно было вычислить на основе данных о словах и словоформах [7]: количество уникальных неизменяемых частей слов; лексическое разнообразие – количество слов, включая всевозможные словоформы каждого слова, и количество уникальных слов (например, в высказывании «хорошее вино - это вино, которое вам нравится» присутствует 7 слов и 6 уникальных слов, так как слово «вино» имеет 2 вхождения в данное высказывание); количество слогов в словах; скорость роста словаря – количество уникальных слов на определённое число слов (например, значение скорости «0,25» означает в среднем 25 уникальных слов на каждые 100 слов). Для оценки признаков, характеризующих исключительно количественный состав слов, использовалась частотная матрица неизменяемых основ слов [8]. В каждой ячейке данной матрицы указано, какое число данной основы слова содержит каждый текстовый документ.

Семантические признаки. Семантический анализ позволяет связать количественный состав слов текста с их качественными характеристиками, определить смысловую составляющую слов. Методы семантического анализа позволили определить количественный состав слов определённых категорий в тексте. К семантическим признакам были отнесены: количество стоп-слов, имён собственных, присутствующих в тексте: места, страны, города, организации, персоны, должности, даты и др. Извлечение данных признаков реализовано с использованием функции «NER» – распознавания именованных объектов [9], которая позволяет создавать аннотации для каждого текста, содержащие информацию о словах и классах, к которому они относятся: места, страны, города, организации, персоны, должности, даты и др. Функция работает на основе нейронной сети, обученной на выборке из словарей англоязычных слов и классов, к которым данные слова относятся.

Сентиментальные признаки. Признаки, описывающие тональность текста, извлекались на основе предположения, что значение тональности предложения складывается из тональности отдельных слов, входящих в него; а текста – из среднего значения тональности отдельных предложений. Вычисление тональности слов происходит на основе «словаря эмоциональностей», предложенного в лаборатории штата Небраска, который по умолчанию настроен на художественные тексты, поскольку термины, на которых он был обучен, были извлечены из коллекции из 165000 закодированных вручную текстов, взятых из корпуса современных романов [10]. Сентиментальными признаками, используемыми в работе, стали: тональность текста, количество слов различного типа тональности (гнев, шутка, грусть, страх, предвкушение, удивление, доверие), количество негативных и позитивных слов; их статистические моменты; максимальные и минимальные значения.

Пунктуационные признаки. Пунктуационный анализ позволяет находить пунктуационные знаки в предложении или тексте. Самым значимым фактом является то, что пунктуационный анализ упрощает

семантический и сентиментальный анализ, так как большинство знаков препинания сигнализирует об определенной части речи или об эмоциональном окрасе фрагмента текста. Используемыми пунктуационными признаками в работе стали: количество точек, запятых, восклицательных и вопросительных знаков в тексте.

Исходная выборка. Исходя из результатов проведенного литературного обзора, был определен оптимальный временной отрезок для оценки количества просмотров новостных публикаций, продолжительностью в одну неделю. Для создания системы оценки количества просмотров новостных публикаций на основе вышеописанных признаков, был выбран интернет-портал «Forbes.com»: число новостных публикаций, ежедневно появляющихся на данном ресурсе, – более 200; 47 миллионов уникальных посетителей ежемесячно; 6,7 миллионов просмотров публикаций за месяц [11].

Исходная выборка должна была включать в себя текста публикаций и данные о количестве их просмотров за неделю. Данная задача решалась с помощью разработки алгоритма, способного сохранять веб-страницы и выгружать их HTML-код. В результате, исходная выборка представляла собой совокупность текстов публикаций, данных о количестве их просмотров за неделю и 120 признаков, характеризующих каждую из них. Исходя из свойств полученных данных, решение задачи оценки количества просмотров публикаций была сведена к решению задачи бинарной классификации. Два класса были определены следующим образом: «Не популярная публикация» (от 0 до 2500 просмотров за неделю) и «Популярная публикация» (более 2500 просмотров за неделю). Исходная выборка состояла из 364 статей (185 публикаций в которой относилось к первому классу, 179 – ко второму).

Процесс обработки информации, содержащейся на веб-страницах, для извлечения признаков проиллюстрирован на рисунке 1. Из каждой веб-страницы необходимо было получить текст публикации, её название и краткое описание (блок 2). Извлеченный HTML-код из веб-страниц позволил получать данные компоненты публикаций с помощью регулярных выражений. В блоке 3 условно показаны признаки, извлеченные из данных. Такие признаки, как количество заголовков, количество ссылок, изображений и применений тегов, были извлечены с помощью регулярных выражений. Остальные признаки – с помощью функций, описанных выше. В итоге, из каждой публикации было извлечено 120 признаков. Предобработка признаков включила очистку от признаков с высоким значением корреляции (например, «Количество букв в тексте» и «Количество слов в тексте», «Количество символов в названии» и «Количество букв в названии»); удаление признаков, имеющих одинаковое значение для всех данных (например, «Максимальная частота уникальных слов в названии публикации» для каждой публикации соответствовало единице); числовое кодирование категориальных признаков и стандартизацию.

Рис. 1. Процесс обработки HTML-кода веб-страницы

Веб-приложение для оценки количества просмотров новостных интернет-публикаций

Сравнительный анализ методов машинного обучения. Для выбора метода решения задачи бинарной классификации, был проведён сравнительный анализ [12, 13]. Качество каждой модели оценено посредством 3 повторов 10-кратной перекрестной проверки по таким параметрам, как точность (количество правильно классифицированных объектов из тестовой выборки), время обучение и минимальное количество настраиваемых параметров. Результаты данного анализа представлены в таблице 1.

Таблица 1. Сравнительный анализ методов машинного обучения для решения задачи бинарной классификации

Метод	Точность	Настраиваемые параметры	Время обучения (мс)
Деревья решений	0.5764 (мин:0.37; макс:0.77)	3	10 000
Бэггинг над ДР	0.579 (мин:0.4: макс:0.72)	1	40 540
Случайный лес	**0.633** (мин: 0.3928; макс: 0.793)	1	172 890
Байесова обобщенная линейная модель	0.5667 (мин: 0.4; макс: 0.793)	0	10 000
Метод опорных векторов	**0.633** (мин: 0.466; макс: 0.822)	2	15 440
Метод k ближайших соседей	0.5592 (мин: 0.55252; макс:0.75)	3	2 845
Логистическая регрессия	0.5112 (мин: 0.322; макс: 0.7)	2	32 000
Стохаостический градиентный бустинг	**0.633** (мин: 0.46; макс:0.862)	4	34 190

Лучшие значения точности относятся к логическим моделям и ансамблям моделей (ДР, случайный лес, стохаостический градиентный бустинг). Предпочтение было отдано методу «Случайный лес», исходя из того, что им были продемонстрированы одни из лучших характеристик в сравнительном анализе, а также дополнительных преимуществ [14], таких, как: способность эффективно обрабатывать данные с большим числом признаков; нечувствительность к масштабированию значений признаков; возможность содержания исходной выборкой признаков, измеренных в разных шкалах, что недопустимо для многих классификаторов, и что существенно упрощает процесс предобработки данных; возможность оценить значимость отдельных признаков в модели.

Процент правильно классифицированных объектов из тестовой выборки считался метрикой для оценки качества модели. Зависимость точности классифицирующей модели от количества деревьев решений в ансамбле и от количества признаков, используемых при создании очередного дерева решений, представлена на рисунке 2.

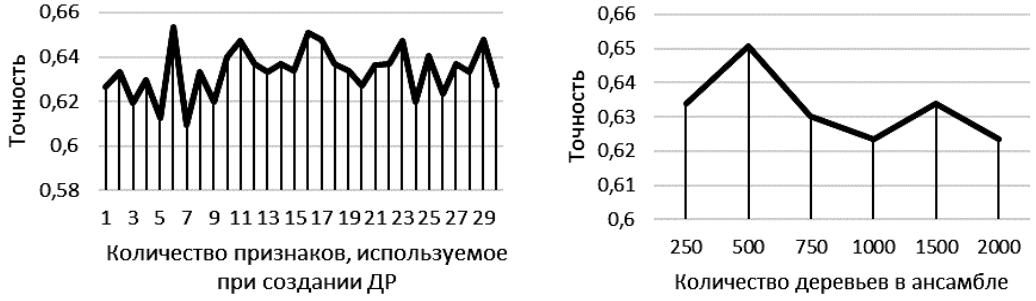

Рис. 2. Зависимость точности классификации от количества деревьев в ансамбле и от количества признаков, используемых при создании очередного дерева решений

Значениями параметров настраиваемого алгоритма для решения задачи были выбраны 500 решающих деревьев в ансамбле и 6 признаков, используемых при обучении очередного дерева решений.

Результатами обучения модели стали:

- Точность на обучающей выборке – 0,65.

Результаты перекрестной проверки на 5 разбиениях:

- Модель верно классифицирует 66,8% объектов из тестовой выборки.
- Точность для класса «Популярная публикация» = 67,7 %
- Точность для класса «Не популярная публикация» = 65,9 %

Так же метод «Случайный лес» позволяет оценить значимость признаков, на которых обучается модель. Оценка значимости позволяет исключить бесполезные признаки и сделать выводы о том, какие из признаков в большей степени влияют на популярность публикаций. Наиболее значимыми признаками стали читаемость текста по Диксу-Стайверу, среднее число слогов в тексте, сложность текста, количество чисел в тексте и количество символов в описании. Самыми незначимыми признаками оказались количество заголовков третьего уровня, количество вхождения форматированного текста (тег «b»), а так же среднее количество букв в словах в тексте. Расчет значимости признаков для оценки популярности публикаций позволяет оптимизировать время работы программы, исключая расчет признаков, которые не вносят никакой вклад в точность работы модели.

Разработка веб-приложения. Оптимальное использование данного инструмента подразумевает систему, позволяющую в режиме реального времени загружать и оптимизировать текст потенциальной публикации непосредственно перед публикацией. Для решения данной задачи было разработано веб-приложение с использованием микрофреймворка «Flask», предназначенного для создания веб-приложений на языке Python.

Веб-приложение, обрабатывающее текст потенциальной публикации и выдающее на выходе один из двух классов, к которым относится данная публикация, представляет собой классическое клиент-серверное приложение. Название потенциальной публикации, краткое описание и её текст, а также такие характеристики, как количество изображений, которые планируется использовать в публикации, количество ссылок, заголовков, абзацев, вносятся пользователем в поля формы на главной странице, после чего эта форма отправляется на сервер, где происходит обработка названия, текста и описания публикации описанными в первой части данной статьи функциями по извлечению признаков. Далее извлеченные признаки, а также признаки, введенные пользователем вручную, компонуются и обрабатываются обученным методом классификации. Результат работы классифицирующей модели передается в ответ клиенту, и он отображается на странице в браузере как «Популярная статья» или «Не популярная статья» соответственно.

Заключение

В работе проанализированы существующие методы оценки количества просмотров интернет-публикаций и выявлен их основной недостаток – оценка популярности публикации непосредственно после публикации на ресурсе. Описано формирование исходной выборки из публикаций, а также извлечение данных о количестве просмотров данных публикаций за неделю. Разработаны функции для извлечения признаков на основе методов обработки естественного, применены к исходным данным. В результате сравнительного анализа методов машинного обучения выбран метод «Случайный лес», характеризующийся высокой точностью, возможностью обрабатывать как непрерывные, так и дискретные признаки, а также возможностью оценки значимости отдельных признаков. Проведена оценка точности классификации метода «Случайный лес» на исходной выборке из 364 новостных публикаций, достигнута точность 0,65. Произведено тестирование метода посредством перекрестной проверки на 5 разбиениях и оценка значимых признаков, в результате верно классифицировано 67% объектов из тестовой выборки. Разработано программное решение для пользователей в виде веб-приложения.

Планируются следующие усовершенствования разработанной системы: добавление новых признаков, характеризующих текст на естественном языке; решение задачи многоклассовой классификации и регрессии; реализация дополнительного функционала в веб-приложении для создания рекомендаций по оптимизации текста потенциальной публикации после её анализа.

Список литературы / References

1. *Szabo G., Huberman B.A.* Predicting the popularity of online content // Communications of the ACM 53(8), 2010. 80–88 p.
2. *Deza Arturo, Parikh Devi.* Understanding Image Virality // Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference, 2015. 1818–1826 p.
3. *Lee Jong Gun, Sue Moon, Kav´e Salamatian.* An Approach to Model and Predict the Popularity of Online Contents with Explanatory Factors // IEEE, WIC, ACM International Conferences on Web Intelligence and Intelligent Agent Technology. Toronto,. Canada, 2010. 623–630 p.

4. *Herbrich Ralf, Graepel Thore.* Handbook of natural language processing. Second edition // Microsoft Research Ltd. Cambridge. UK // Chapman and Hall // CRC; 2 edition February 22, 2010. 704 p.
5. *Kao Anne and Stephen R.* Natural Language Processing and Text Mining // Poteet (Eds). UK, 2007. 272 p.
6. *Manning Christopher D., Schiitze Hinrich*. Foundations of Statistical Natural Language Processing // The MIT Press. Cambridge. Massachusetts. London, England. 704 p.
7. Package 'koRpus' for R // The Comprehensive R Archive Network. [Электронный ресурс]. Режим доступа: https://cran.r-project.org/web/packages/koRpus/index.html/ (дата обращения: 19.02.2017).
8. *Stefan TH. Gries.* Quantitative corpus linguistics with R: A Practical Introduction // Routledge; 1 ed. – February 22, 2009 – 260 p.
9. *Rahul Sharnagat.* Named Entity Recognition: A Literature Survey // Indian Language Technology (CFILT), June 30, 2014 – 27 p.
10. Package 'suyzhet' for R // The Comprehensive R Archive Network. [Электронный ресурс]. Режим доступа: https://cran.r-project.org/web/packages/syuzhet/syuzhet.pdf/ (дата обращения: 11.02.2017).
11. Recourse of global media, branding and Technology Company, with a focus on news and information about business, investing, technology, entrepreneurship, leadership and affluent lifestyle. [Электронный ресурс]. Режим доступа: http://www.forbesmedia.com./ (дата обращения: 11.02.2017).
12. *Guha R., Manjunath Shreya, Palepu Kartheek.* Comparative analysis of machine learning techniques for detecting insurance claims fraud // Wipro limited, Doddakannelli, Bangalore. 560 035, India. 19 p.
13. *Alesheykh R.*. Comparative Analysis of Machine Learning Algorithms with Optimization Purposes // Department of Information Technology. Payame Noor University. P.O. BOX. 19395-3697. Tehran. Iran, 12 p.
14. *Flah P.* Machine Learning: The Art and Science of Algorithms That Make Sense of Data // Cambridge University Press. 1 ed., November 12, 2012. 409 p.