Министерство образования Республики Беларусь

Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК 004.855.5

Шинкевич
Наталья Николаевна

Предсказание популярности контента с использованием алгоритмов машинного обучения

АВТОРЕФЕРАТ
на соискание академической степени
магистра технических наук

по специальности 1-40 80 05 – Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

Научный руководитель
Егорова Н.Г.
кандидат технических наук,
доцент

Минск 2019

КРАТКОЕ ВВЕДЕНИЕ

Высокие скорости в принятии решений, многозадачность в управлении бюджетом и необходимость снижения рисков требуют современного подхода к организации и ведению бизнеса. Выходом в условиях усложняющейся внутренней и внешней среды предприятия становится комплексная автоматизация бизнес-процессов. Она позволяет гибко управлять ресурсами, позволяет менеджменту сконцентрироваться на ключевых направлениях работы компании.

Автоматизация является наиболее доступным, а иногда и единственным средством быстрого повышения эффективности предоставляемых услуг. В эпоху, когда аккаунты в Facebook, Twitter, Instagram и других социальных сетях есть как у большинства компаний, так и у 96% населения планеты, закономерным является то, что многие компании начали пользоваться автоматизированными средствами превращения информации социальных сетей в знания, имеющие практическую ценность.

В настоящее время масштабная предиктивная социальная аналитика доказала свою эффективность. За последнее десятилетие результаты научных исследований в сфере социальной аналитики не раз доказали свою потенциальную ценность, помогая создавать новые проекты и бизнес-решения на основе анализа поведения людей в интернете, анализа настроений, анализа влияния и взаимообусловленности различных событий и т. д.

Однако стоит отметить, что большинство таких решений еще не адаптированы к анализу данных в режиме реального времени, что, с учетом постоянно растущей динамики в экономической, социальной и других сферах, является по сути их главным недостатком.

Во многих системах предсказания в реальном времени имеют решающее значение, и задержки на часы (или даже минуты) могут сделать их практически бесполезными.

В качестве примеров: политические кампании могут очень быстро реагировать на противостояние мнений и интересов, которое распространяется на Facebook; различные так называемые «сети доставки контента», или CDN (Content Delivery Network) могут предварительно выбирать видео, фото или другой контент, который, как предсказывают, скоро станет вирусным, виральным; рекламные кампании в сети интернет могут быть корректировать

свою организацию бизнеса, чтобы уменьшить затраты и увеличить количество потенциальных потребителей.

Глобальная инфраструктура социальных сетей способна в настоящее время «понимать» человеческие чувства, делать прогнозы на их основе. В связи с тем, что появились новые технологические возможности для анализа огромного количества данных, появился такой социально-экономический феномен как Big Data. Однако наличие экстремально больших объемов самой разной информации создает значительные практические проблемы. В частности, масштабный анализ требует подавляющего количество вычислений, чтобы выявить действительно важные признаки среди шума. Под шумом следует понимать данные, которые в контексте решаемой задачи являются незначимыми, и только усложняют расчеты, не внося никакого вклада в общий анализ.

Следовательно, основная нетривиальная работа по пониманию социальных данных опиралась на пост-анализ, или анализ, использующий машинное обучение на исторических данных для выявления макромасштабных тенденций и закономерностей.

Среди данных, пригодных для анализа, я хочу выделить новости, так как на мой взгляд именно они являются неотъемлемой частью повседневной жизни каждого человека. Новостные агентства, такие как белорусские tut.by, onliner.by, иностранные Foxnews.com, mittelbayerische.de и другие ежедневно публикуют более тысячи материалов, но не все они становятся одинаково популярными.

Таким образом, прогноз популярности и является той стратегией, которой журналисты и редакторы должны придерживаться для более точного определения содержания, а также позиционирования своих статей на новостных ресурсах.

По результатам проведенного мной исследования можно выделить две проблемы, которые имеют решающее значение в решении проблемы прогнозирования популярности контента. Во-первых, в настоящее время существует огромное количество каналов, из которых люди получают информацию. Следовательно, есть несколько показателей популярности, например, количество просмотров статьи в группе VK, количество лайков или репостов на Facebook, или количество упоминаний определенной фразы или слова в поисковой системе. Во-вторых, популярность статьи может быть

определена в локальном или глобальном контексте. Мера локальной популярности контента прежде всего предназначена для использования в одном информационном агентстве или новостном ресурсе, в то время как мера глобальная показывает популярность одной конкретной статьи среди статей из других новостных агентств или ресурсов.

Прогнозирование локальной популярности материала является хорошей основой для разработки систем оценки и анализа контента, потому что в таком случае имеется более удобная возможность проверки результатов, чем в случае с глобальным прогнозированием. Нетривиальной и интересной задачей является интеграция диапазона измерений популярности через различные новостные каналы в глобальном контексте.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Цель и задачи исследования

Цель исследования в данной работе — осуществить анализ текстовой информации на примере публикаций новостных ресурсов с целью прогноза их популярности на различных платформах и ресурсах, а также создание инструмента, осуществляющего данный прогноз в режиме реального времени, с целью оптимизации публикуемого контента, а также расширения функционала новостных ресурсов, оптимизации рекламы и иных способов продвижения контента в сети интернет.

Для достижения поставленной цели необходимо решить следующие задачи:

1. Проанализировать предметную область, проконсультироваться с представителями новостных ресурсов и других каналов формирования общественного мнения;
2. Определить примерный список новостных ресурсов для анализа;
3. Выбрать конкретную прикладную бизнес-задачу, а именно прогнозирование популярности публикаций, и более детально исследовать эту область знаний;
4. Провести анализ и моделирование выбранной задачи;
5. Разработать систему, предназначенную для решения данной задачи. Провести эксперименты, сравнить результаты и сделать вывод о том, какие алгоритмы машинного обучения лучше всего справляются с решением поставленной задачи.

Объектом исследования выступают новостные публикации, контент, размещенный в сети интернет.

Предметом исследования — прогноз популярности контента новостных ресурсов сети интернет. В качестве данных были взяты статьи из открытых источников (для первоначального тестирования функциональности системы), а также собранные с новостных ресурсов данные, находящиеся в открытом доступе, а также данные, предоставленные заинтересованной в проекте третьей стороной. В качестве экспериментальных данных было взято небольшое количество публикаций из закрытых источников заинтересованных в проекте представителей иностранных новостных ресурсов.

Основная *гипотеза,* положенная в основу диссертационной работы, состоит в том, что популярность статьи, или виральность — это максимальное покрытие аудитории, существующей у новостного ресурса. Прогноз популярности данного контента в рамках той платформы, где он публикуется, состоит в том, чтобы определить, насколько контент может приблизиться к этому максимуму. Таким образом, довольно абстрактная задача прогноза популярности контента формализуется и впоследствии может быть разбита на отдельные подзадачи, такие как сбор данных об отдельной платформе, ее аудитории, анализ собранных данных, прогноз популярности в рамках платформы, проверка результатов и так далее.

Связь работы с приоритетными направлениями научных исследований и запросами реального сектора экономики

Работа выполнялась в соответствии научно-техническими заданиями и планами работ. Данные настоящего исследования были использованы напрямую для создания функционального прототипа системы для прогноза популярности контента для компании Exponenta Inc, разработки пользовательского интерфейса с возможностью ориентации на конкретного потребителя, а именно новостной ресурс, закрытого для остальных пользователей (аутентификация по временному API токену).

Исходя из заинтересованности в проекте крупных отечественных новостных ресурсов (tut.by, onliner, kyky.org), а также иностранных новостных ресурсов, в настоящее время уже обеспечена возможность тестирования прототипа в режиме реального времени, создана и функционирует система обратной связи, а также будет продолжено дальнейшее тестирование продукта с учетом требований и нужд потенциальных потребителей.

Личный вклад соискателя

Результаты, приведенные в диссертации, получены соискателем лично. Вклад научного руководителя Н. Г. Егоровой, заключается в формулировке целей и задач исследования.

Опубликованность результатов диссертации

Основные вопросы, рассматриваемые в данной диссертационной работе, подробно освещены в ряде публикаций, в том числе международного уровня, а именно: Третья Международная научно-практическая конференция «BIG DATA and Advanced Analytics: collection of materials of the third international scientific and practical conference» 2017 года (Минск, Беларусь, 2017), международная научная конференция «Информационные технологии и системы 2018 (ИТС 2018)». Полученные научные и практические результаты также докладывались на XV Белорусско-российской научно-технической конференции.

Структура и объем работы

Диссертация состоит из введения, общей характеристики работы, восьми глав, заключения, списка использованных источников, списка публикаций автора и трех приложений. В первой главе проанализирована предметная область. Во второй главе сделан обзор решений и систем, существующих в настоящее время и наиболее приближенных по своему функционалу к системе, разработанной в рамках данной диссертационной работы. В третьей главе дан обзор методов, используемых для прогноза популярности контента. В четвертой главе рассмотрено создание инфраструктуры для будущей системы прогнозирования, а именно базы данных. В пятой главе дан обзор вспомогательных методов и действий, направленных на обеспечение системы прогнозирования качественными данными для анализа. В шестой главе рассмотрены признаки, извлекаемые из текстовой информации в момент ее поступления. В седьмой главе описан анализ данных, этапы создания дополнительных характеристик и обучения моделей, а также сравнение их производительности. В восьмой (последней) главе проводится оценка полученных результатов.

Общий объем диссертации − 145 страниц. Работа содержит 22 формулы, 45 рисунков, 8 таблиц. Библиографический список включает 64 наименования.

ОСНОВНОЕ СОДЕРЖАНИЕ

Во введении рассмотрено современное положение и роль инструментов и алгоритмов машинного обучения в социально-экономической сфере, приведены примеры, объясняющие зависимость успеха большинства современных крупных компаний от их скорости реакции на различные события, имеющие место быть в социальной сфере, а также от масштабности использования автоматизированных средств превращения информации, полученной из социальных сетей, в знания, имеющие практическую ценность.

В первой главе сформулирована и конкретно поставлена задача, дан обзор ее предметной области, подробно освещены моменты, с которыми можно столкнуться при решении данной задачи, а именно: проблемы легального скрапинга сайтов, сбора информации из открытых источников.

Во второй главе дан обзор существующих решений, которые по своему функционалу наиболее приближены к разработанной в рамках данной диссертационной работы системе.

В третьей главе был проведен обзор и оценка методов прогнозирования популярности контента, приведена их классификация, рассмотрены области их применения. Также были введены и пояснены метрики качества, которые использовались для оценки производительности итоговой модели.

В четвертой главе приведена структура базы данных, необходимой для корректной работы системы, подробно описаны таблицы, их взаимосвязи.

В пятой главе описаны основные структурные компоненты системы, рассмотрены детали реализации двух режимов ее работы в условиях непрерывно поступающих данных, также объяснена необходимость наличия нескольких режимов работы. Рассмотрен этап очистки, или Data Cleaning, который осуществляется как для непрерывно поступающих, так и для единоразово поступивших данных для анализа.

В шестой главе рассмотрены признаки, извлекаемые из текстовой информации после того, как она прошла этапы проверки на содержательность и очистки соответственно. Пояснена необходимость создания новых дополнительных признаков на их основе.

В седьмой главе проведен обзор набора данных, используемого в качестве экспериментального при построении и тестировании системы. Рассмотрены аспекты выбора признаков для обучения моделей, также сделан

обзор методов конструирования новых признаков, а также результатов, которые были получены при применении всех вышеперечисленных методов на практике. Также в данной главе дан краткий обзор методов классификации, используемых для решения поставленной задачи, детально описаны проведенные эксперименты.

В последней (восьмой) главе проведена оценка методов машинного обучения, использованных при проведении экспериментов, даны результаты сравнительного анализа качества моделей, аргументирован выбор модели, основанной на Random Forest, в качестве основного алгоритма классификации для разработанной системы. Приведены предпосылки создания более усовершенствованной версии системы с учетом проведенных экспериментов и полученных результатов.

В заключении приведены основные достигнутые результаты и возможные будущие шаги, направленные на развитие области исследования.

ЗАКЛЮЧЕНИЕ

В работе была затронута тема прогнозирования популярности контента, а также возможности использования алгоритмов и методов машинного обучения к решению данной задачи. Для того, чтобы как можно более качественно решить поставленную задачу, были выполнены следующие шаги:

1. Проведен детальный анализ предметной области, а именно: осуществлено несколько встреч с представителями новостных ресурсов и других каналов формирования общественного мнения, проведено 4 опроса, из них 3 в режиме онлайн, 1 в виде рассылки анкет заинтересованным сторонам с целью выяснения существующих потребностей в разрабатываемой системе, организована встреча со специалистами в области обработки естественных языков (Natural Language Processing, NLP).
2. Определен конкретный список естественных языков, исходя из их общей распространенности в сети Интернет, и на новостных ресурсах в частности.
3. Определены новостные ресурсы, данные которых предполагалось использовать для построения системы.
4. Проведено всестороннее исследование выбранной области, а именно прогнозирования популярности контента.
5. Были исследованы популярные алгоритмы классификации, для обучения которых использовались собранные и обработанные специальном образом данные, проведен сравнительный анализ данных алгоритмов по различным критериям.
6. Предложена структура базы данных, необходимой для хранения и накопления данных.
7. Выбраны алгоритмы, на основании которых были построены модели и проведены эксперименты с их участием.
8. Сделан вывод о том, какая модель наилучшим образом решает поставленную задачу.
9. Разработана система, в основу которой была положена модель, превосходящая остальные по качеству.

Можно сформулировать следующий список достигнутых результатов:

- точность классификации на отложенной выборке составила 86.9%, метрика f-measure 0.81;
- время обучения модели составило 61 секунду;

- скорость формирования прогноза для конкретных данных составила
- 0.000014, что является, безусловно, отличным результатом в условиях необходимости быстро выдавать ответ пользователю, обрабатывать большие объемы данных; также это является качественно новым результатом по сравнению со скоростью, с которой человек принимает аналогичное решение.

Поставленная задача была достигнута с использованием методов, позволяющих анализировать текстовую информацию, извлекать из нее признаки при помощи существующих баз знаний, математических методов и алгоритмов. Несмотря на хорошее качество решения задачи, в настоящее время есть определенные методы, которые, возможно, позволят улучшить качество модели. К ним относятся использование рекуррентные нейронные сети (Recurrent Neural Network; RNN), сети долгой краткосрочной памяти (Long short-term memory; LSTM).

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

1-А. Использование вероятностного латентно-семантического анализа для построения вероятностных тематических моделей текстовых коллекций / М. И. Селюк, Н. Н. Шинкевич, М. В. Стержанов // Технические средства защиты информации : тезисы докладов XV Белорусско-российской науч.-техн. конф. (Минск, 6 июня 2017 г.). – Минск : БГУИР, 2017. – С. 101.

2-А. Оценка релевантности контента в режиме реального времени / Н. Н. Шинкевич // Информационные технологии и системы 2018 (ИТС 2018) = Information Technologies and Systems 2018 (ITS 2018) : материалы международной научной конференции, Минск, 25 октября 2018 г. / Белорусский государственный университет информатики и радиоэлектроники ; редкол. : Л. Ю. Шилин [и др.]. – Минск, 2018. – С. 138 - 139.

3-А. Функциональность системы получения и анализа текстовых данных / Шинкевич, Н. Н., Селюк, М. И., Рожков, Д. Н., Пресняцкий, В. Ю., Свито, А. И.М. В. Стержанов // BIG DATA and Advanced Analytics: collection of materials of the third international scientific and practical conference. (Minsk, Belarus, May 3 – 4, 2017) / editorial board : M. Batura [etc.]. – Minsk : BSUIR, 2017. – С. 242 - 245.