Концепция контроля достоверности информации в профессиональной социальной сети с применением сверточной нейронной сети

Ю. В. Катенко
Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина);
ООО «БалтИнфоКом»
katenkoyuliya@gmail.com

С. А. Петренко
Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)
s.petrenko@rambler.ru

***Аннотация*. В статье описан подход к оценке достоверности информации, размещенной в социальной сети. Достоверность информации рассматривается с точки зрения ее правдивости, истинности. Предлагается оценивать достоверность информации, приведенной в записи социальной сети, с применением алгоритмов классификации. Для анализа текстов записей предложено использовать сверточные нейронные сети. Также в статье описан алгоритм построения и использования инструмента для оценки достоверности, а также возможные варианты его применения.**

Ключевые слова: достоверность информации; социальные сети; классификация текстов; сверточные нейронные сети

I. История вопроса

За последние годы социальные сети стали одним из основных источников новостей и другой информации для множества людей. Они позволяют увидеть освещение происходящих в мире событий с разных точек зрения, быстро и удобно получать любую тематическую информацию. Вместе с этим социальные сети все чаще используются для распространения разнообразных слухов, мошеннической и прочей недостоверной информации, а также политической пропаганды. Информация может искажаться как злоумышленником, преследующим собственные корыстные цели, так и обычными пользователями, неумышленно распространяющими недостоверную информацию посредством размещения на своих страницах данных, полученных из непроверенных источников. Все это может оказывать влияние на социальную, политическую, экономическую стабильность.

По данным ВЦИОМ, 62% россиян хотя бы раз в неделю пользуются социальными сетями. [1] Однако не все понимают необходимость критической оценки информации, представленной в социальных сетях, поэтому зачастую записи, содержащие недостоверные данные, могут оказывать на пользователей негативное влияние. Примером такого влияния может быть распространение во время чрезвычайных происшествий слухов, вызывающих у людей страх (например, слухи о взрывах наземного транспорта после теракта в Петербургском метрополитене в 2017 году). Также в качестве примера можно рассматривать такое явление как отказ от вакцинации, которое ВОЗ назвала одной из главных глобальных угроз здоровью, — одним из главных каналов распространения недостоверной информации на эту тему являются именно социальные сети. [2]

Следовательно, в сложившейся ситуации, когда количество преднамеренных и непреднамеренных угроз нарушения достоверности информации возрастает, и при этом в большинстве популярных социальных сетей нет доступных обычному пользователю встроенных механизмов оценки, возникает необходимость в создании системы контроля достоверности информации. В процессе работы система должна анализировать высказывания, содержащиеся в текстах записей социальной сети, и в случае обнаружения недостоверной информации сигнализировать об этом. Кроме того, эта система может позволить следить за распространением недостоверной информации и оценивать ее влияние на пользователей.

В литературе предлагаются различные определения достоверности информации.

Свойства достоверности информации — это свойства, характеризующиеся возможностями воздействующей информации отражать в сообщениях, содержащихся в ней, реально существующие объекты с необходимой точностью. [3]

Достоверность информации — объективное отражение информацией реальных процессов. [4]

Достоверность информации — свойство информации быть правильно воспринятой, вероятность отсутствия ошибок; степень соответствия данных, хранимых в памяти ЭВМ или документах, реальному состоянию отображаемых ими объектов предметной области. [5]

В данной работе достоверность информации рассматривается с точки зрения ее правдивости, истинности, непротиворечивости, соответствия реальному положению вещей.

Достоверность информации измеряется доверительной вероятностью необходимой точности (вероятностью того,

что отображаемые в сообщениях, содержащихся в информации, значения параметров отличаются от истинных значений этих параметров в пределах необходимой точности). [3]

II. Критический анализ известных методов

В работе В.В. Зубец и И.В. Ильиной «Оценка достоверности сетевой информации» предлагается оценивать информацию, размещенную в Интернете, путем вычисления суммы баллов по четырем критериям: степени идентификации владельца сайта, мотивации в предоставлении достоверной информации, актуальности информации и качеству выходных данных. Этот подход прост в использовании, однако он не универсален и подходит в большей степени для оценки научной информации. [6]

Работа С.М. Ивановой «Оценка достоверности информации, найденной в сети Интернет» посвящена обучению организации процесса поиска информации в Интернете. Автор предлагает определять достоверность информации с учетом ее полноты, целостности и истинности. Наиболее сложной является задача определения истинности информации, и в качестве решения в статье предлагается оценивать достоверность источника в целом.

Источник может считаться достоверным, если он принадлежит научной, образовательной или правительственной организации. В противном случае в качестве способа оценки истинности информации может быть использован анализ ссылок на используемые источники данных. Однако зачастую сайты не содержат ссылок на используемые источники, и в этом случае необходимо оценить, насколько достоверной является представленная на сайте информация на известную для пользователя тему. Основываясь на этом, пользователь в дальнейшем может делать выводы о тематически близкой информации, достоверность которой ему необходимо оценить. Для решения этой задачи предлагается применять контроллер Мамдани аппарата нечеткой логики. Недостаток этого подхода заключается в том, что у пользователя не всегда может быть достаточно сведений, чтобы оценить, насколько истинной является какая-либо информация по искомой тематике, и, следовательно, в таком случае он не сможет судить о достоверности информации в источнике в целом. Однако этот подход интересен тем, что он применим и для анализа данных из социальных сетей — вместо общей достоверности информации на сайте таким образом можно определять достоверность информации на странице пользователя или в сообществе. [7]

Общим недостатком приведенных подходов является невозможность полной автоматизации процесса оценки достоверности информации, пользователь вовлекается в этот процесс, что, во-первых, требует затрат времени, во-вторых, наличия у пользователя навыков, необходимых для анализа. Также необходимо учитывать, что люди в целом склонны к подтверждению собственной точки зрения, что делает оценку достоверности информации предвзятой. Для автоматизации оценки достоверности информации в социальной сети могут быть использованы методы машинного обучения. В таком случае задача анализа достоверности информации сводится к задаче классификации записей социальной сети. Все записи могут быть разделены, к примеру, на следующие классы:

- достоверная информация;
- скорее достоверная информация;
- информация с неопределенной достоверностью;
- скорее недостоверная информация;
- недостоверная информация.

Существенным ограничением для применения алгоритмов классификации является необходимость создания обучающей выборки, включающей большое количество объектов, для которых вручную определен класс. В работе Leon Derczynski и др. «SemEval-2017 Task 8: RumourEval: Determining rumour veracity and support for rumours» предлагается использовать краудсорсинг для определения достоверности высказываний, приведенных в записи социальной сети, и отношения к записи прокомментировавших ее пользователей. [8] Краудсорсингом (от английского “crowd” — «толпа» и “sourcing” — «использование ресурсов») называется процесс передачи каких-либо функций или операций некоторому кругу лиц за пределы компании, решение поставленных задач силами добровольцев, используя информационные технологии. Как правило, краудсорсинг — это неоплачиваемая деятельность, однако, существуют сервисы, позволяющий привлечь к решению задач большое количество людей за определенную плату. [9]

Другой важной задачей является выбор набора признаков, которые будут использованы для построения классификатора. Во-первых, признаки содержатся в самом тексте записи в социальной сети, во-вторых, из социальных сетей может быть извлечена дополнительная информация, которую можно применить при анализе.

Социальная сеть может быть представлена в виде графа, содержащего следующие классы узлов:

- пользователь;
- сообщество;
- запись;
- комментарий;
- оценка (одобрение или неодобрение).

Следовательно, в качестве дополнительных признаков можно использовать сведения об авторе записи и его связях с другими пользователями, реакцию пользователей на запись, а также данные, которые содержатся в записи — изображения, ссылки, хэштеги, именованные сущности.

Для оценки достоверности текста записи социальной сети могут быть использованы различные алгоритмы

классификации текстов, в том числе могут применяться нейронные сети.

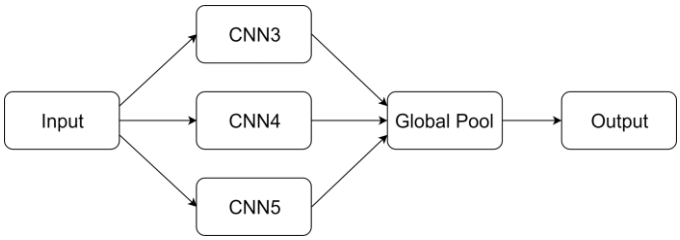

Рис. 1. Схема классификатора, включающего слои свертки

Большой интерес представляет подход, описанный Yoon Kim в работе «Convolutional Neural Networks for Sentence Classification». [10] Предлагается использовать для анализа текстов классификатор, в котором используются слои свертки. Его схема представлена на рисунке 1. Input и Output на схеме — это входной и выходной слои; CNN3, CNN4 и CNN5 — сверточные слои, каждый из которых включает набор ядер свертки размером 1×3, 1×4 и 1×5 соответственно; Global Pool — слой, предназначенный для объединения выходных значений с каждого слоя свертки в один вектор.

В процессе обучения нейронной сети с описанной архитектурой с помощью сверточных слоев будут выделены определенные последовательности, включающие три, четыре и пять слов, наиболее характерных для текстов определяемой категории.

III. Замысел решения

Использование предложенного метода включает три этапа.

Этап 1. Подготовка обучающей выборки, необходимой для обучения моделей классификации текстов записей и записей в целом. Включает следующие шаги:

Шаг 1. Получение из социальной сети большого массива записей, комментариев, оценок и связанных пользователей, подготовка этих данных к ручному анализу.

Шаг 2. Разметка данных усилиями множества пользователей.

Шаг 3. Оценка качества полученного набора данных.

Этап 2. Обучение моделей классификации — полученные на этом этапе модели будут применяться непосредственно для оценки достоверности информации, представленной в записи. Данный этап включает следующие шаги:

Шаг 1. Обучение модели классификации текстов записей. Данные шаг включает:

- Предобработку текстов — разбиение на токены, удаление стоп-слов, приведение токенов к начальной форме — стемминг или лемматизация, приведение наборов токенов к векторному виду — например, с помощью моделей word2vec, описанных в работе Tomas Mikolov и др. «Efficient Estimation of Word Representations in Vector Space». [11]
- Обучение сверточной нейронной сети, архитектура которой описана ранее.

Шаг 2. Обучение модели классификации записей социальной сети. Этот шаг включает:

- Выбор признаков исследуемого объекта — в набор признаков будет входить класс текста, а также некоторые дополнительные признаки, зависящие от пользователя, комментариев и т. д.
- Обучение модели. Возможно построение классификаторов с применением различных алгоритмов и последующий выбор того, который показал наилучшие результаты.

Этап 3. Применение полученных моделей для оценки достоверности. Включает следующие шаги:

Шаг 1. Получение из социальной сети необходимой записи, ее автора, комментариев, оценок и связанных пользователей.

Шаг 2. Предобработка текста записи — процесс предобработки должен соответствовать тому, который был применен при создании классификатора текстов.

Шаг 3. Классификация текста записи с помощью сверточной нейронной сети.

Шаг 4. Извлечение дополнительных признаков из данных.

Шаг 5. Классификация записи в социальной сети с учетом класса текста записи и дополнительных признаков.

Шаг 6. Добавление к записи метки, которая отображает класс достоверности.

IV. Заключение

Контроль достоверности информации в социальных сетях в настоящее время является весьма актуальной задачей, поскольку все больше людей становится активными их пользователями, и при этом количество угроз постоянно возрастает.

Сегодня ответственность за проверку того, насколько правдива представленная в социальной сети информация, ложится на плечи каждого пользователя, заинтересованного в получении достоверной информации. Однако ручной анализ, при котором необходимо оценить репутацию ресурса и автора и совершить перекрестную проверку данных в различных источниках, отнимает много времени и сил, кроме того, он требует от пользователя беспристрастности.

Реализация предложенного подхода позволит упростить эту проверку. Разрабатываемый метод должен быть достаточно универсальным для того, чтобы его можно было применять в любой социальной сети. Возможно два варианта его использования.

Во-первых, в качестве одного из внутренних механизмов социальной сети — в таком случае, все высказывания, представленные в текстах записей должны анализироваться автоматически. В случае, если высказывание с определенной вероятностью является недостоверным, к записи должна быть добавлена предупреждающая метка. При таком варианте применения метода информация о достоверности записей будет постепенно накапливаться, что, возможно, повысит точность анализа новых записей.

Во-вторых, в качестве внешнего механизма проверки достоверности, который может быть реализован, например, в виде бота в социальной сети или на стороннем веб-сайте. Многие популярные социальные сети, в том числе ВКонтакте, Одноклассники и Твиттер, предоставляют открытые API, с помощью которых можно получить данные, необходимые для анализа. В этом варианте использования проверка достоверности информации находится в ответственности пользователя, однако ему не нужно заниматься поиском информации, а достаточно будет указать ссылку на запись.

В перспективе планируется детальная разработка предложенного в данной работе метода, а также его программная реализация.

Список литературы

[1] Каждому возрасту — свои сети // ВЦИОМ URL: https://wciom.ru/index.php?id=236&uid=116691 (дата обращения: 13.03.2019).

[2] Ten threats to global health in 2019 // World Health Organization URL: https://www.who.int/emergencies/ten-threats-to-global-health-in-2019 (дата обращения: 20.03.2019).

[3] ГОСТ Р 43.0.4-2009 «Информационное обеспечение техники и операторской деятельности. Информация в технической деятельности. Общие положения». М: Стандартинформ, 2018.

[4] Большой экономический словарь. 19000 терминов / Под ред. А.Н. Азрилияна. М.: Институт новой экономики, 1997. 864 с.

[5] Термины и определения в области информационной безопасности / Комов С.А. и др. М.: АС-Траст, 2009. 291 с.

[6] Зубец В.В., Ильина И.В. Оценка достоверности сетевой информации // Вестник Тамбовского университета. Серия: Естественные и технические науки. 2011. №1. URL: https://cyberleninka.ru/article/n/otsenka-dostovernosti-setevoy-informatsii (дата обращения: 20.03.2019).

[7] Иванова С.М. Оценка достоверности информации, найденной в сети Интернет // Преподаватель XXI век. 2015. №4. URL: https://cyberleninka.ru/article/n/otsenka-dostovernosti-informatsii-naydennoy-v-seti-internet (дата обращения: 18.03.2019).

[8] Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, Arkaitz Zubiaga. SemEval-2017 Task 8: RumourEval: Determining rumour veracity and support for rumours. URL: https://arxiv.org/pdf/1704.05972.pdf (дата обращения: 18.03.2019).

[9] Сивакс А.Н. Краудсорсинг как способ оптимизации функционирования предприятий // Интернет-журнал Науковедение. 2015. №1 (26). URL: https://cyberleninka.ru/article/n/kraudsorsing-kak-sposob-optimizatsii-funktsionirovaniya-predpriyatiy (дата обращения: 20.03.2018)

[10] Yoon Kim. Convolutional Neural Networks for Sentence Classification. URL: https://arxiv.org/pdf/1408.5882.pdf (дата обращения: 18.03.2018)

[11] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. URL: https://arxiv.org/pdf/1301.3781.pdf (дата обращения: 18.03.2018)