Аннотация—С расширением обсуждаемых тем, освещаемых в социальных сетях, читателям становится все труднее извлекать фактическую информацию из новостных статей, содержащих большое количество мнений. Эта задача также широко известна как задача обнаружения субъективности — отличить субъективные предложения от беспристрастных, или объективных. В этой статье представлено создание инструмента для выявления субъективности в новостных статьях с нуля. Во-первых, мы собираем современный набор данных новостных статей, аннотированных для субъективности работниками краудсорсинга. Затем мы настраиваем модель BERT, достигая точности, близкой к лучшим достигнутым результатам, в задаче обнаружения субъективности. Мы также используем продвинутые методы дообучения, такие как one cycle policy. Наконец, мы создаем удобный инструмент командной строки для автоматического аннотирования новостных статей с применением дообученной модели. Польза от этой работы выходит за рамки работы с новостями, поскольку точное распознавание мнений и фактов имеет важное значение для таких областей, как поиск информации и анализ тональности.

Ключевые слова — обработка естественного языка, обнаружение субъективности, краудсорсинг, сбор аннотированных данных, обработка новостей

I. Введение

В данной статье рассматривается задача отличия фактов от мнений в новостных статьях. Эта задача также известна как задача обнаружения субъективности. Это важная подзадача анализа тональности, поскольку модели анализа тональности работают лучше, когда работают с корпусами, содержащими исключительно оценочные суждения.

В настоящее время многие действующие информационные агентства, в том числе BBC, The Guardian и другие, ежедневно публикуют тонны материалов. Количество горячо обсуждаемых тем сегодня также растет: среди недавних — экологические проблемы, COVID-19, гонка поисковых систем между Microsoft и Google и другие. С таким количеством обсуждаемых новостей обычному читателю становится трудно извлечь факты, чтобы иметь непредвзятый взгляд на происходящие события.

В этой статье рассматривается создание инструмента для различения фактических и субъективных предложений в новостных статьях. Инструмент позволяет пользователю загружать текст новостной статьи и аннотировать каждое предложение меткой факта или мнения.

Мы достигаем поставленной цели пошагово. Сначала мы создаём набор предложений из современных новостных статей. Этот набор данных аннотирован людьми на субъективность. Wiebe, J. et al. создали аналогичный набор данных под названием MPQA в [1]. Их исследование является основным аналогом нашей работы по созданию современного набора новостных данных. Аннотации предложений для этого набора данных были созданы в 1999 году в [2] и сейчас являются устаревшими. Во-вторых, мы дообучаем модель BERT для обнаружения субъективности. Pang, B. et al. рассмотрели различные методы дообучения модели BERT в [3]. Мы переиспользуем и проверяем результаты их исследования — они достигли точности около 90%, что верифицируется в данной статье. В-третьих, мы создаём новый инструмент, который по ссылке на новостную статью извлекает ее текст, размечает его на предмет субъективности и выдаёт размеченную статью. В настоящее время у данного инструмента нет аналогов; ближайший похожий инструмент — Grammarly, который может помочь обнаружить слишком оценочные предложения и адаптировать текст под настроенные критерии.

Структура работы выглядит следующим образом. Раздел 2 охватывает научные статьи, связанные с нашим исследованием, то есть работы, охватывающие создание как наборов данных, так и моделей. В разделе 3 мы описываем нашу методологию для сбора и человеческой разметки набора данных. В разделе 4 мы подробно описываем средства оценки и полученные результаты. Раздел 5 содержит выводы статьи.

II. Анализ литературы

Достаточное количество исследований было проведено в области обнаружения субъективности. Существует несколько различных наборов данных, предназначенных для этой задачи в нескольких областях. В [4] был представлен набор данных SUBJ. Он состоит из 5000 субъективных и 5000 объективных предложений, взятых из 1346 веб-страниц, аннотированных людьми. Предложения маркируются с учетом субъективности (субъективное/объективное) и полярности (положительное/отрицательное). В [5] создается набор данных, содержащий утверждения из Википедии с тегами POV (point of view). Утверждения с тегом POV — это утверждения, которые, как сообщается в статье, нарушают принцип NPOV (neutral point of view) Википедии. Эти утверждения затем аннотируются людьми, и в конечном итоге набор данных состоит из 1843 предвзятых утверждений, 3109 нейтральных, 1843 нейтральных из избранных статей в Википедии и 1843 нейтральных из избранных статей с таким же распределением предвзятых утверждений. В [1] был представлен набор данных MPQA (Multi-Perspective Question Answering). Он содержит предложения из 535 испанских новостных статей, переведенных на английский язык. Всего в наборе данных около 9700 предложений, и около 55% из них субъективны. Все эти наборы данных имеют отношение к нашей задаче, но только MPQA удовлетворяет нашей области, то есть новостным статьям. Однако набор данных MPQA был создан около 20 лет назад. Мы считаем, что за 20 лет стиль новостных текстов в информационных агентств изменился; кроме того, были введены новые темы, такие как COVID-19. Было создано несколько новых агентств, а некоторые старые обанкротились. По всем этим причинам мы решили

создать более современный набор данных, который в целом будет похож на MPQA, но будет состоять из современных данных.

Большой объём работ был также проделан на тему аннотации текстов. В процессе разработки набора данных MPQA авторами Wiebe, J., et al. были выпущены статьи [6], [7] и [8]. Во многом они использовались как вдохновители для нашей инструкции по разметке данных, в том числе, некоторые примеры были взяты из этих статей. При создании набора данных SUBJ, в статье [4], за изначальные данные были взяты отзывы с сайтов RottenTomatoes (отмечены как полностью субъективные) и IMDb (отмечены как полностью объективные), и взяты близости предложений из неразмеченного набора данных с размеченными. Затем был использован алгоритм min-cut-max-flow, использующий потоки, для разметки оставшейся части набора данных. В статье [5], самой новой из тех, что представляют новый набор данных, был использован краудсорсинг (сервис Amazon Mechanical Turk) — аннотаторам предлагались предложения и достаточно простая инструкция; необходимо было выбрать один из трёх вариантов ответов. В настоящее время краудсорсинг стал наиболее популярной техникой разметки наборов данных, что подтверждается тем, что статья [5] первая из нами рассмотренных стала использовать эту технику.

Помимо создания наборов данных, большой объем работы был посвящен обучению моделей для выявления субъективности. В [9] показан временной прогресс таких моделей. Первые методы использовали примитивные функции, например, обнаружение ключевых слов. Затем исследователи перешли к онтологическим моделям, то есть к определению набора онтологий, которые определяют отношения между разными классами слов и проецируют их в векторное пространство. Следующими были статистические методы, похожие на классическое машинное обучение — модели, обученные на аннотированном наборе данных. Одним из примеров такой модели является Passive-Aggressive Classifier, упомянутый в [10]. Авторы этого исследования утверждают, что их метод достиг около 85% F1-score на кросс-валидации. Это показывает, что ранние подходы к задаче были уже достаточно успешными. Следующими моделями были так называемые LDM (Latent Dirichlet Models), которые использовали частоту слов для вычисления апостериорного распределения. Все вышеперечисленные методы были синтаксическими, т. е. пытались использовать синтаксис, а не значения слов, в роли признаков. Следующие методы, наоборот, используют семантические значения. Начнем с того, что существуют семантические модели предложений, которые пытаются смоделировать вероятность слова с учетом слов, стоящих перед ним. Затем есть деревья синтаксического анализа, методы, которые пытаются моделировать вероятности предложений, разлагая представления на матричные умножения и используя рекуррентные нейронные сети (RNN) поверх этого. Затем существуют сверточные модели (CNN) для обнаружения субъективности, где CNN служат основной моделью (backbone), а затем ответ извлекается с помощью RNN или других методов. Один из таких методов описан в [11] — они используют байесовскую сетевую машину экстремального обучения (BNELM) поверх CNN для достижения точности 89% на TASS 2015 — наборе данных, содержащем твиты на испанском языке для обнаружения субъективности. Наконец, новейшими моделями обнаружения субъективности являются модели, основанные на архитектуре трансформеров, такие как BERT. Одним из таких исследований является [3], где авторы достигают точности от 84% до 95% на разных наборах данных. Они также пробуют различные методы дообучения для повышения качества и обнаруживают, что такие методы, как One Cycle Policy и Multitask-Learning являются наиболее полезными. Мы собираемся повторить этот эксперимент в нашей работе.

Как упоминалось ранее, в настоящее время нет аналогов нашей работы именно для обнаружения субъективности. Существуют средства для написания новостей, такие как Headline Analyzer, Ahrefs и Grammarly, но нет инструментов для автоматического анализа новостей.

III. Сбор и разметка данных

A. Выбор исходного набора данных

Перед разметкой необходимо подобрать неразмеченный набор данных. Нашими главными критериями для подбора такого набора данных были:

- Достаточный размер: одно из главных требований к нашему размеченному набору данных было наличие минимум 5000 субъективных и 5000 объективных предложений, поэтому необходимо было, чтобы неразмеченный набор данных был достаточно большим.
- Наличие вариативности в новостных источниках: наш инструмент в итоге должен получиться достаточно универсальным, так что с его помощью можно было бы размечать как, например, новости про спорт, так и новости про экологию.
- Возможность расширения: по возможности, способ сбора набора данных должен быть достаточно простым, чтобы можно было собрать самые новые статьи.
- Простота использования: наша работа нацелена более на аннотацию, чем на сбор данных, поэтому нужен набор данных, который будет легко использовать без дополнительной обработки.

Среди вариантов наборов данных, предложенных на сайте HuggingFace[1], были рассмотрены следующие:

[1]https://huggingface.co/

- news_commentary[2] — набор данных с переводом большого количества статей между различными языками. Не был выбран, так как данные изначально были собраны не для классификации, и понадобилась бы дополнительная обработка, чтобы вывести из него нужные статьи на английском языке.
- multi_news[3] — набор данных для суммаризации, в котором для суммаризации даются статьи из разных источников. Не был выбран по той же причине: этот набор данных был собран для другой задачи, и была необходима предобработка, чтобы достать статьи в нужном нам формате.
- argilla[4] — набор данных, содержащий статьи, изначально собранные для классификации. Содержит около 20000 статей, что подходит нам, но не был выбран, так как нет возможности собрать его заново из самых новых статей.
- cc_news[5] — набор данных, состоящий из статей, собранных с помощью утилиты news-please[6], которую легко использовать, и потенциально можно было бы использовать также для итогового инструмента командной строки. Этот набор данных прост в использовании, легко расширяется, а также содержит сотни тысяч статей, собранных из различных новостных порталов, поэтому он был выбран в качестве неразмеченного набора данных для нашей задачи.

B. Подготовка исходного набора данных к аннотации

Выбранный набор данных содержал несколько сотен тысяч статей, что было слишком много для аннотации. Необходимо было отобрать столько самых подходящих статей, чтобы суммарно в них было около 20000 предложений. Также необходимо было разделить эти статьи на предложения, чтобы подготовить данные к разметке. Для этого были проделаны следующие шаги:

- Из набора данных были убраны все статьи, содержащие менее пяти предложений — мы признали эти статьи выбросами из общего распределения.
- Статьи были дедуплицированы — набор данных мог содержать одинаковые статьи из одного и того же источника, но разных веб-сайтов, например, reuters.co.uk и reuters.com.
- В оставшемся наборе данных были оставлены только источники, содержащие от 100 статей. Это было сделано для удаления слишком редких источников — вероятных выбросов из общего распределения.
- Были выбраны самые новые статьи, так, чтобы суммарное число предложений было около 20000. В итоге было оставлено 1024 статьи из июля 2018 года, содержащие в сумме 19953 предложения. Мы посчитали, что июль 2018 года — это достаточно новые статьи, и можно просто взять их, и таким образом, сбор новейших данных и разработка инструмента, использующего их, остаётся для будущей работы.

C. Написание инструкции

Значительное количество времени было выделено на написание качественной инструкции. Для этого был предпринят следующий алгоритм действий. Изначально была составлена инструкция на основе [8], адаптированная под нашу конкретную задачу. Пользователям предлагалось выбрать либо опцию "Not Applicable предназначенную для предложений, не содержащих никакие утверждения и, следовательно, не подходящих для разметки, либо степень субъективности по дискретной шкале от 1 до 5, также известной как шкала Ликерта. Дальше происходили три итерации улучшения инструкции, состоящие в том, что автор и третье лицо размечали три заранее выбранных статьи и сравнивали результаты. После одной из итераций состоялась консультация с студентом Школы лингвистики НИУ ВШЭ, после которой инструкция была улучшена. После трёх итераций удалось получить корелляцию 0.62 и F1-score 0.41 между авторскими ответами и ответами третьего лица. Так как особенность задачи состоит в субъективности разметки, такие результаты были признаны достаточно хорошими, чтобы запускать разметку полного набора данных.

D. Разметка с помощью краудсорсинга

Для более масштабной разметки был использован сервис Toloka AI[7]. В данном сервисе есть два важных понятия: проект и пул. Проект — это сущность, содержащая несколько пулов. На уровне проекта задаётся инструкция и некоторые правила контроля качества. Пул — это набор данных, который непосредственно размечают работники краудсорсинга. Бывают разные виды пулов, например, тренировочный пул создан для того, чтобы на примерах объяснить работникам инструкцию; этот пул размечается бесплатно. Экзаменационный пул необходим, чтобы отобрать работников с достаточным умением, и отсеять мошенников и роботов. Далее, существует общий вид пула, в котором работники уже непосредственно размечают рабочие данные. В нашем проекте были все эти три вида пулов. Для тренировочного пула было специально написано несколько примеров под каждую метку; также были написаны пояснения к этим примерам. Для экзаменационного пула были использованы те размеченные автором данные, которые использовались при улучшении инструкции. Остальные данные были размечены с помощью общего пула. Параметр перекрытия

[2]https://huggingface.co/datasets/news_commentary
[3]https://huggingface.co/datasets/multi_news
[4]https://huggingface.co/datasets/argilla/news-summary
[5]https://huggingface.co/datasets/cc_news
[6]https://github.com/fhamborg/news-please
[7]https://toloka.ai/

(то есть, сколько аннотаторов должны разметить одно предложение) был выбран равным 3. В качестве правил контроля качества использовались ограничения по числу размеченных наборов задач на человека за день, ограничение по пропуску наборов задач подряд, а также отложенная приёмка (то есть, перед оплатой денег разметка должна была быть проверена вручную, и аннотации могли быть отклонены). Изначально было размечено всего 5 статей, чтобы удостовериться в правильности настройки. После того, как настройки были улучшены, а разметка этих пяти статей выглядела адекватно, были запущены в разметку остальные статьи.

E. Пост-обработка разметки

В качестве способа агрегации ответов аннотаторов использовался алгоритм WAWA[8]. Это итеративный алгоритм, который в начале выдаёт каждому работнику вес 1, потом вычисляет для каждого предложения голос взвешенного большинства и перераспределяет веса между работниками в зависимости от того, как близко они к этому большинству на каждом примере. После агрегации голосов аннотаторов для каждого работника был посчитан F1-score, и задачи, размеченные аннотаторами, получившими F1-score меньше 0.35, были переразмечены. Таким образом, нам удалось достичь достаточного уровня согласия между разными аннотаторами.

IV. Результаты

Мы собрали современный набор данных, размеченных на задачу обнаружения субъективности, с помощью разметки работниками краудсорсинга. Мы получили достаточно большой уровень согласия аннотаторов — F1-score каждого аннотатора против агрегированных меток составляет не менее 0.35. В итоговом наборе данных присутствует около 1500 меток "Not Applicable около 14000 меток "Objective"(оценки 1 и 2 из проекта разметки), около 500 меток "Neutral"(оценка 3 из проекта разметки) и около 4000 меток "Subjective"(оценки 4 и 5 из проекта разметки), что практически удовлетворяет поставленным целям.

V. Заключение

В этой статье описана работа по созданию нового инструмента для распознавания мнений и фактов в новостных статьях. На данный момент был собран набор данных, необходимый для эффективного дообучения модели. Мы уверены, что этот проект значительно повлияет на анализ новостей, и инструмент будет достаточно удобным для публичного распространения.

[8]https://toloka.ai/docs/crowd-kit/reference/crowdkit.aggregation.classification.wawa.Wawa/

Список литературы

[1] J. Wiebe and E. Riloff, “Creating subjective and objective sentence classifiers from unannotated text,” in Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing, 2005, pp. 486–497.

[2] J. Wiebe, R. Bruce, and T. O’Hara, “Development and use of a gold standard data set for subjectivity classifications,” in Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL 1999), 1999.

[3] H. Huo and M. Iwaihara, “Utilizing bert pretrained models with various fine-tune methods for subjectivity detection,” in Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data. Springer, 2020, pp. 270–284.

[4] B. Pang and L. Lee, “A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts,” in Proceedings of ACL 2004, 2004.

[5] C. Hube and B. Fetahu, “Neural based statement classification for biased language.” in Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining. ACM, 2019, pp. 195–203.

[6] J. Wiebe, “Instructions for annotating opinions in newspaper articles,” University of Pittsburgh, Tech. Rep., 2002.

[7] J. Wiebe, T. Wilson, and C. Cardie, “Annotating expressions of opinions and emotions in language,” Language resources and evaluation, vol. 39, pp. 165–210, 2005.

[8] T. Wilson and J. Wiebe, “Annotating attributions and private states,” in CorpusAnno ’05: Proceedings of the Workshop on Frontiers in Corpus Annotations II, 2005.

[9] I. Chaturvedi, E. Cambria, R. Welsch, and F. Herrera, “Distinguishing between facts and opinions for sentiment analysis: survey and challenges,” Information Fusion, vol. 44, pp. 65–77, 2018.

[10] A. Stepinski and V. Mittal, “A fact/opinion classifier for news articles,” in Proceedings of the ACM Special Interest Group on Information Retrieval (SIGIR). ACM, 2007, pp. 807–808.

[11] I. Chaturvedi, E. Ragusa, P. Gastaldo, R. Zunino, and E. Cambria, “Bayesian network based extreme learning machine for subjectivity detection,” Journal of The Franklin Institute, vol. 355, pp. 1780–1797, 2018.