Сравнение моделей векторизации текстов для задачи анализа тональности коротких сообщений из социальных сетей

А. В. Чижик

Университет ИТМО

chizhik@itmo.ru

Аннотация

Анализ тональности текстов является одной из актуальных задач, которая способна выявлять важные факторы, влияющие на вектор социального настроения общества. При использовании для решения этой задачи методов машинного обучения требуется преобразовать текст в его векторное представление. Существует ряд методов векторизации текстов. В данной статье сравниваются три актуальных на данный момент подхода к созданию векторного представления: учет веса слова в документе (TF-IDF), использование дистрибутивной семантики при создании векторов слов (Word2Vec) и векторизация целых предложений (Laser). Сравнивая эти три модели векторизации текстов для задачи анализа тональности коротких сообщений из социальных сетей, можно сказать, что каждая из них имеет свои преимущества и недостатки. В статье описан дизайн исследования, приведены метрики качества, описаны данные, на которых проводились опыты.

Ключевые слова: векторизация текстов, анализ тональности, социальные медиа

Библиографическая ссылка: Чижик В. А. Сравнение моделей векторизации текстов для задачи анализа тональности коротких сообщений из социальных сетей // Компьютерная лингвистика и вычислительные онтологии. Выпуск 7 (Труды XXVI Международной объединённой научной конференции «Интернет и современное общество», IMS-2023, Санкт-Петербург, 26–28 июня 2023 г. Сборник научных статей). — СПб: Университет ИТМО, 2024. С. 81–89. DOI: 10.17586/2541-9781-2024-7-81–89

1. Введение

Информационные технологии, проникая во все сферы жизни, изменяют способы взаимодействия между людьми, обеспечивая доступ к новым источникам информации и создавая новые формы коммуникации. Благодаря доступности интернета, развитию социальных сетей, мессенджеров и других цифровых инструментов коммуникации в текущий момент наблюдается резкое увеличение количества цифровых взаимодействий между людьми. Для индивидов присутствие виртуальной коммуникативной среды обеспечивает расширение круга общения и позволяет находить единомышленников и достигать общих целей вне пространственных, временных и социальных ограничений. С точки зрения социокультурной динамики это означает увеличение интерактивности общественного пространства через более активное участие людей в различных социальных и политических процессах. Таким образом, цифровые технологии и цифровые взаимодействия являются мощным инструментом создания новых связей между социальными группами и отдельными индивидами. При этом онлайн-среда представляет собой многочисленные горизонтальные связи, образующие масштабный неориентированный граф, именно эта структура и обеспечивает преодоление социальных барьеров и приводит к пересечению индивидов из различных слоев общества.

Коммуникация в виртуальной реальности имеет ряд важных характеристик, две из которых стратегически важны в рамках актуализации потенциала этой коммуникативной среды как объекта исследований. Во-первых, большинство цифровых взаимодействий происходит в форме обмена текстовыми сообщениями, что означает наличие зафиксированного в удобном формате цифрового следа; во-вторых, часто коммуникация происходит в публичной зоне (посты и комментарии к ним, чаты) и имеет ряд метаданных (например, дата поста, социально-демографические характеристики автора), а, значит, доступна для структурированного сбора данных и дальнейшего изучения компьютерными методами.

Таким образом, возрастающее количество публичных текстовых данных превратили анализ текстов (natural language processing, nlp) в актуальный метод для анализа социальной динамики общества. Основным объектом исследований стал контент из социальных и новых медиа. Стоит отметить, что nlp-методы применяются в различных областях, в том числе активно используются в сфере маркетинга для оптимизации пути клиента, в медицине для обеспечения дистанционного взаимодействия «клиника-пациент», но, главное, они позволяют анализировать поведение социальных групп на микро- и макроуровнях, привязывая его к временной шкале и событийным фактам. Так выявляются и объясняются скрытые закономерности, приводящие общество в движение, а также предсказываются вероятности наступления явления (например, общественных волнений или, наоборот, апатии социальной группы или общества в целом).

Одним из важных nlp-методов, который актуален для исследования медиапространства, является анализ тональности текста (sentiment analysis). Он дает возможность понять отношения, мнения и эмоции, лежащие в основе онлайн-текста. Формализуя понятие, можно дать следующее определение: анализ тональности — это класс методов контент-анализа в компьютерной лингвистике, основная задача которого заключается в классификации текста по его настроению. Обобщая тональность текстов, можно вычислять индекс субъективного благополучия, прогнозировать результаты выборов или экономических показателей, оценивать реакцию на события или новости.

По сути, тон текста помогает понять эмоциональное состояние автора и определить его отношение к поднятой теме. Так как любой публичный пост подразумевает наличие серии комментариев на него, то становится доступным целый ряд научных рефлексий: исследование общей реакции социальной группы на тему, анализ реакции активных акторов на проблему, детекция реакции пассивных акторов коммуникации на лидеров мнений.

2. Постановка проблемы

В простых случаях задача анализа тональности сводится к бинарной классификации текстов на две категории: позитивные и негативные (в ряде случаев также включают категорию «нейтральный текст»). Однако подобное разделение на 2-3 класса не всегда репрезентативно для выявления глобальных социальных закономерностей, и задача переформатируется в мультиклассовую классификацию, когда необходимо более четко определить эмоциональные состояния индивидов. В таком случае дополнительная фаза исследования отводится под разработку актуальной шкалы, способной связать в единую логику используемые для анализа данные и выявляемые закономерности. В таких целях может использоваться численная шкала или категории типа «страх», «злость», «печаль», «счастье». В результате могут быть вынесены суждения об индексе социального благополучия или векторе социального настроения. Такие классы легко связываются с количественными данными (например, в задачи социального картирования, где необходимо визуально показать взаимосвязь эмоций жителей страны, города или района и ряда количественных данных, отражающих различные характеристики жизни в этой

локации). Глобально анализ тональности текстов можно разделить на три направления методов:

1. Подходы на основе правил (rule-based). В них используются размеченные словари эмоций, для русского языка крупнейшими являются RuSentiLex и LINIS Crowd [1, 2], которые имеют информацию о привязке слов к категориям «позитивно» и «негативно», то есть не дают четких характеристик эмоций в отличие от англоязычных SenticNet, SentiWordNet и SentiWords [3, 4, 5]. Так же эта группа методов предполагает вручную созданные наборы правил классификации. Очевидным минусом подхода является низкая способность к обобщению (невозможно масштабировать для анализа текстов, не имеющих предсказуемой конкретной тематики).
2. Подходы на основе машинного обучения, которые подразумевают автоматическое извлечение признаков из текста, что позволяет анализировать тексты, относящиеся к разным тематикам, в едином конвейере. Часто используемыми в рамках анализа тональности моделями машинного обучения являются логистическая регрессия, дерево решений и метод опорных векторов. Последние несколько лет помимо классических алгоритмов машинного обучения для решения этой задачи применяются свёрточные (CNN) и рекуррентные (RNN) нейросети [6, 7]. Группа этих методов показывает хорошие результаты с точки зрения метрик качества (точность от 70% в зависимости от конкретной задачи) и масштабируемости (применимость для текстов разных типов и дискурсов).
3. Гибридные подходы, которые объединяют в себе первые два (примером может служить ALDONAr [8]).

Из перечисленных групп методов с точки зрения применимости для анализа процессов социальной динамики выделяются подходы на основе машинного обучения. Возможность их применения первично строится на необходимости переформатирования текста в числовые векторы, так как алгоритмы машинного обучения подразумевают манипуляции в математическом пространстве. К тому же идея заключается в том, что векторы (embedding), представленные в геометрическом пространстве, могут быть описаны через расстояние до соседей, что дает информацию об их взаимосвязях. Эмбеддинги слов могут быть созданы различными методами векторизации: самая простая из них — «мешок слов» (bag of words), также часто используется tf-idf векторизация, которая учитывает важность слова в документе, а не только частоту его появления. В более сложных системах для генерирования эмбеддингов слов применяются модели дистрибутивной семантики, например, Word2Vec, GloVe и FastText [9, 10, 11]. Существуют подходы к векторизации, позволяющие создать эмбеддинги предложений или параграфов (а не слов), к этой логике векторизации относятся, например, модели ELMo, BERT и LASER [12, 13, 14].

Стоит отметить, что, несмотря на частое появление в методологиях исследований компонента анализа тональности, метод не имеет четких рамок и устоявшихся правил использования: тональность, содержащуюся в тексте, можно анализировать на уровне бинарной классификации, или детализировать на несколько классов (часто используют пятибалльную шкалу). В зависимости от того, какая модель векторизации будет использована, примененный в дальнейшем алгоритм машинного обучения для задачи классификации тональности текста сработает точнее или наоборот более ординарно. Таким образом, исследование применимости моделей векторизации к конкретным типам текстов является актуальной исследовательской проблемой.

В рамках данного исследования была поставлена задача анализа успешности моделей векторизации применительно к коротким текстам из социальных сетей. Было решено сфокусироваться на бинарной классификации, так как основной вопрос: какая техника создания векторного представления точнее фиксирует особенности коротких текстов на русском языке (разговорного формата) с точки зрения возможности далее ml-моделью уловить негативные и позитивные тональности.

3. Данные

В настоящее время интерес представляют два социальных медиа: ВКонтакте (консервативная по формату социальная сеть, состоящая из пабликов и групп с наличием публичных постов и комментариев к ним) и Телеграмм (мессенджер с большим количеством публичных чатов, где обсуждение тем может развиваться параллельно, без наличия побуждающего нулевого поста). С точки зрения возможностей привязывать анализируемые текстовые данные к реальности (например, к геоданным) полезнее оказывается информация, полученная из социальной сети ВКонтакте. Поэтому для тестирования моделей векторизации было собрано два набора данных именно из этой сети: 1) датасет постов и комментариев к ним из публичных районных сообществ города Санкт-Петербурга (18 групп, выкачивались данные за 2019-2020 гг.); 2) датасет, составленный из контента пабликов «Подслушано» (4 группы, выкачивались данные за 2022 год). Средняя длина комментариев в первом наборе данных — 21 слово, а постов — 41 слово; во втором датасете средняя длина постов — 11 слов, комментариев к ним — 15 слов. Полярность тематик собранных датасетов — намеренная стратегия, так как дискурсы текстов и длина сообщений — важные характеристики, влияющие на подбор метода векторизации. Идея заключалась в том, чтобы проверить, будет ли какое-то заметное различие в метриках качества для двух датасетов. Общий объем данных — 319 335 записей. Собранные текстовые данные были поэтапно предобработаны по следующей схеме: 1) разбиение текстов реплик на токены; 2) удаление спецсимволов, эмодзи, ссылок и знаков пунктуации; 3) удаление стоп-слов; 4) нормализация токенов. На выходе из такого пайплайна препроцессинга был получен предобработанный текст, готовый к вероятностной векторизации. Также в рамках очистки датасетов от данных, не вносящих концептуальный вклад в эксперимент, посты (начало дискуссии по теме), не содержащие более пяти комментариев, были удалены. Это было сделано исходя из того, что для проводимого эксперимента была важна оценка тональности поста и серии комментариев к нему с точки зрения направленности социального настроения, таким образом, нейтрально окрашенные темы (например, обсуждение потерянных ключей или графика работы какого-то учреждения) не представляли для данного исследования интереса. После этого этапа предобработки данных были получены обновленные датасеты, общим объемом 204 107 строк.

4. Описание эксперимента

В качестве базовой модели векторизации была выбрана TF-IDF (учитывались биграммы). Также были обучены: модель Word2Vec (size = 100, sg = 1, min_count = 1, window = 5, учитывались биграммы) и базирующаяся на библиотеке глубокого обучения PyTorch модель LASER (использовалась предобученная модель для русского языка из библиотеки laserembeddings). Таким образом, эксперимент заключался в сравнении успешности трех основных подходов к созданию векторных представлений текстов.

Прежде чем обратиться к обучению классификатора, было решено исследовать данные методом кластеризации, чтобы удостовериться, что в текстах действительно есть закономерности, которые с математической точки зрения заметны. Кластеризация данных является классической задачей восстановления распределения данных: это дает понимание того, как объекты распределены в пространстве признаков, какие наиболее характерные значения у них есть, где объектов мало, а где они лежат плотным облаком. Таким образом, на первом этапе анализа данных мы использовали эмбеддинги текстов, полученные тремя способами, и было принято решение посмотреть способность анализируемых моделей эмбеддингов к разделению на кластеры. Кластеризация была проведена с использованием алгоритма K-средних (k=2), так как он позволяет задать количество искомых кластеров. Результаты разбиения на два кластера представлены на рис. 1. Для визуализации кластеризации использовался алгоритм понижения размерности PCA.

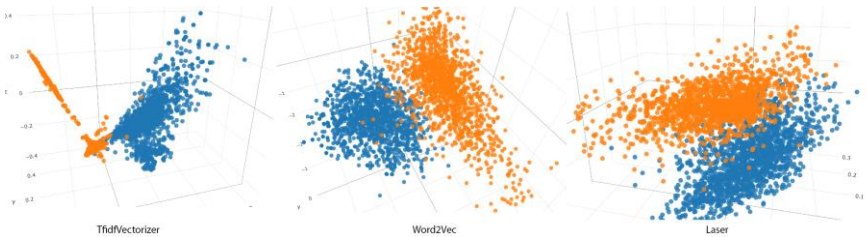

Рис. 1. Анализ разделимости классов с использованием 3D-сжатия векторного пространства с использованием алгоритма PCA

Графики показывают, что векторизация методом TF-IDF дает неплохие результаты: кластеры визуально выглядят сепарированными друг от друга. Векторизация с помощью Word2Vec и Laser гораздо хуже фиксирует особенности текстов, позволяющие их сепарировать как легко вчитывающиеся кластеры, по крайней мере при k=2. Стоит отметить, что кластерный анализ не дает точного представления о конкретных особенностях текстов: признаки, по которым алгоритм делит данные на группы, остаются не интерпретируемыми. Однако, этот опыт показывает, насколько векторное представление в принципе фиксирует полярность кластеров (по какому-либо признаку). Заметим, что неожиданным стала низкая репрезентативность Word2Vec-векторизации, так как этот метод обычно хорошо улавливает синонимичность слов, что, как следствие, помогает близкие по значению тексты отнести к одному кластеру.

На втором этапе эксперимента часть собранных данных была размечена вручную на два класса (негативный и позитивный).

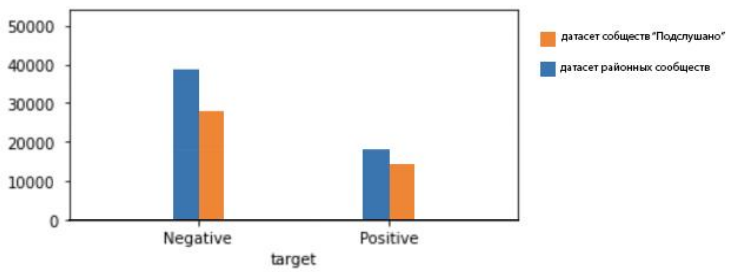

Рис. 2. Распределение размеченных классов в двух датасетах

После этого были выделены слова (и словосочетания), вносящие наибольший вклад в каждый из классов, и построены облака слов для обоих классов (рис. 3).

Рис. 3. Датасет районных сообществ: облако слов, вносящих наибольший вклад в «негативный» класс (слева); облако слов, вносящих наибольший вклад в «позитивный» класс (справа)

Такое визуальное представление классов дает возможность выдвинуть важную гипотезу: «негативные» тексты гораздо важнее правильно детектировать нежели «позитивные», так как они явно более содержательны с точки зрения возможностей

дальнейшего анализа контекстов. То есть, выявив «негативный» класс далее отдельно с ним можно проводить дополнительные исследования: тематическое моделирование, выделение именованных сущностей, анализ тональности уже с мультиклассовым разделением на эмоции. Соответственно, для оценки качества работы модели классификатора можно использовать матрицу ошибок. Она дает информацию о процентном содержании истинно-положительного, истинно-отрицательного, ложно-положительного и ложно-отрицательного решений классификатора. Таким образом, отдельно от общей производительности модели, становится возможным проверить, в скольких случаях был спрогнозирован «негативный» класс, и это оказалось правдой.

Далее размеченный набор данных был разделен на обучающую и тестовую выборку (пропорция 70% и 30%). В качестве алгоритма классификации была выбрана логистическая регрессия. На рис. 4 представлены результаты работы логистической регрессии при анализе датасета районных сообществ.

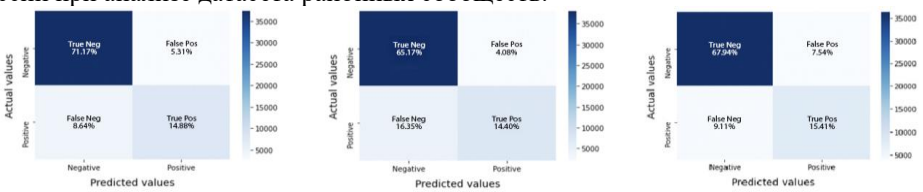

Рис. 4. Результаты работы модели логистической регрессии при отправленных в нее векторных представлениях, полученных тремя способами (слева направо: tf-idf, w2v, Laser)

Как видно из результатов, все три метода векторизации сработали достаточно хорошо. Однако различия касаются степени ошибок первого (ложно-положительное решение) и второго (ложно-отрицательное решение) рода. По приведенным матрицам видно, что лучше всего «негативный» класс детектируется логистической регрессией, работающей на векторном представлении tf-idf. Удивительным фактом является то, что модель векторизации Laser сработала достаточно хорошо, это значит, что для анализа тональности коротких текстов актуальным подходом может быть векторизация целых предложений.

Эксперименты над вторым датасетом дали схожие результаты, при этом стоит отметить, что Laser показал лучший результат относительно tf-idf, а модель Word2Vec осталась на третьем месте (56,17% истинно-отрицательных решений, что немного хуже, чем это было на данных из районных сообществ). Таким образом, появляется гипотеза, что чем менее текст насыщен контекстов (и присутствует только эмоция), тем хуже Word2Vec способствует улавливанию нюансов, важных для классификации по тону сообщения. И в то же время Laser, вероятно, показывает наиболее успешные результаты в рамках векторизации коротких текстов, чем меньше в них присутствует категория «содержание».

5. Заключение

При сравнении моделей векторизации TF-IDF показала лучшую способность улавливать необходимые особенности в коротких текстах. Модель логистической регрессии с использованием данного векторного представления показала хорошую итоговую производительность (F1_score=0,81), к тому же именно эта векторизация позволяет при анализе тональности точнее детектировать «негативный» класс, который, как было показано выше, является более интересным с точки зрения дальнейших поисков закономерностей при анализе социального настроения. Однако стоит отметить, что Word2vec предоставляет дополнительные инструменты анализа текстов (благодаря учету квази-синонимичности) и, соответственно, при определенной постановке задачи может быть полезным. Касательно целесообразности использования Laser, стоит дополнительно отметить, что модель требует больших ресурсов оперативной памяти (и сама векторизация

занимает достаточно длительное время), однако само векторное представление показало неплохие результаты при использовании в классификаторе.

В дальнейшем планируется доразметить собранные наборы данных, составить из них датасет, содержащий уравновешенное количество примеров из обоих классов, и затем повторить опыт с теми же настройками, что описаны в данном эксперименте (так как некоторая туманность в оценке «позитивного» класса на данный момент остается).

Исследование выполнено при поддержке Российского научного фонда и Санкт-Петербургского научного фонда, грант № 23-28-10069 «Прогнозирование социального самочувствия с целью оптимизации функционирования экосистемы городских цифровых сервисов Санкт-Петербурга» (https://rscf.ru/project/23-28-10069/).

Литература

[1] Loukachevitch N., Levchik A. Creating a general Russian sentiment lexicon // Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). — 2016. — P. 1171–1176.

[2] Koltsova O. Y., Alexeeva S., Kolcov S. An opinion word lexicon and a training dataset for Russian sentiment analysis of social media // Computational Linguistics and Intellectual Technologies: Materials of DIALOGUE. — 2016. — Vol. 2016. — P. 277–287.

[3] Cambria E., Poria S., Bajpai R., Schuller B. SenticNet 4: A semantic resource for sentiment analysis based on conceptual primitives // Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics. — 2016. — P. 2666–2677.

[4] Baccianella S. et al. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining // Lrec. — 2010. — Vol. 10. — № 2010. — P. 2200–2204.

[5] Gatti L., Guerini M., Turchi M. SentiWords: Deriving a high precision and high coverage lexicon for sentiment analysis // IEEE Transactions on Affective Computing. — 2015. — Vol. 7. — №. 4. — P. 409–421.

[6] Baziotis C. et al. Ntua-slp at semeval-2018 task 3: Tracking ironic tweets using ensembles of word and character level attentive rnns // arXiv preprint arXiv:1804.06659. — 2018.

[7] Baziotis C., Pelekis N., Doulkeridis C. Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis // Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017). — 2017. — P. 747–754.

[8] Meškelė D., Frasincar F. ALDONAr: A hybrid solution for sentence-level aspect-based sentiment analysis using a lexicalized domain ontology and a regularized neural attention model // Information Processing & Management. — 2020. — Vol. 57. — №. 3. — Art. 102211.

[9] Mikolov T., Sutskever I., Chen K., Corrado G., Dean J. Distributed representations of words and phrases and their compositionality // Advances in neural information processing systems. — 2013. — Vol. 26. — P. 3111–3119.

[10] Pennington J., Socher R., Manning C. D. Glove: Global vectors for word representation // Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). — 2014. — P. 1532–1543.

[11] Joulin A., Grave E., Bojanowski P., Mikolov T. Bag of tricks for efficient text classification // Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. — 2017. — Vol. 2. — P. 427–431.

[12] Lee K., Filannino M., Uzuner Ö. An Empirical Test of GRUs and Deep Contextualized Word Representations on De-Identification // MedInfo. — 2019. — P. 218–222.

[13] Devlin J., Chang M.-W., Lee K., Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. — 2018. — Vol. 1. — P. 4171–4186.

[14] Chizhik A., Zherebtsova Y. Challenges of Building an Intelligent Chatbot // International Conference «Internet and Modern Society» (IMS-2020). CEUR Proceedings. — 2021. Vol. 2813. — P. 277–287.

Comparison of Text Vectorization Models for the Sentiment Analysis of Short Messages from Social Media

Anna V. Chizhik

ITMO University

Sentiment analysis is one of the urgent tasks that can identify important factors which affect the vector of the social mood. When using machine learning methods to solve this problem, it is required to convert the text into its vector representation. There are a number of text vectorization methods. This paper compares three currently relevant approaches to creating a vector representation: taking into account the weight of a word in a document (TF-IDF), using distributive semantics when creating word embeddings (Word2Vec), and sentence embedding models (Laser). Comparing these three text vectorization models for the task of analyzing the sentiment of short messages from social networks, it is obvious that each of them has its own advantages and disadvantages. The paper describes the design of the study, provides quality metrics, describes the data on which the experiments were conducted.

Keywords: text vectorization, embeddings, sentiment analysis, social media

Reference for citation: Chizhik A.V. Comparison of Text Vectorization Models for the Sentiment Analysis of Short Messages from Social Media // Computational Linguistics and Computational Ontologies. Vol. 7 (Proceedings of the XXVI International Joint Scientific Conference «Internet and Modern Society», IMS-2023, St. Petersburg, June 26–28, 2023). — St. Petersburg: ITMO University, 2024. P. 81–89. DOI: 10.17586/2541-9781-2024-7-81–89

Reference

[1] Loukachevitch N., Levchik A. Creating a general Russian sentiment lexicon // Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). — 2016. — P. 1171–1176.

[2] Koltsova O. Y., Alexeeva S., Kolcov S. An opinion word lexicon and a training dataset for Russian sentiment analysis of social media // Computational Linguistics and Intellectual Technologies: Materials of DIALOGUE. — 2016. — Vol. 2016. — P. 277–287.

[3] Cambria E., Poria S., Bajpai R., Schuller B. SenticNet 4: A semantic resource for sentiment analysis based on conceptual primitives // Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics. — 2016. — P. 2666–2677.

[4] Baccianella S. et al. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining // Lrec. — 2010. — Vol. 10. — № 2010. — P. 2200–2204.

[5] Gatti L., Guerini M., Turchi M. SentiWords: Deriving a high precision and high coverage lexicon for sentiment analysis // IEEE Transactions on Affective Computing. — 2015. — Vol. 7. — №. 4. — P. 409–421.

[6] Baziotis C. et al. Ntua-slp at semeval-2018 task 3: Tracking ironic tweets using ensembles of word and character level attentive rnns // arXiv preprint arXiv:1804.06659. — 2018.

[7] Baziotis C., Pelekis N., Doulkeridis C. Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis // Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017). — 2017. — P. 747–754.
[8] Meškelė D., Frasincar F. ALDONAr: A hybrid solution for sentence-level aspect-based sentiment analysis using a lexicalized domain ontology and a regularized neural attention model // Information Processing & Management. — 2020. — Vol. 57. — №. 3. — Art. 102211.
[9] Mikolov T., Sutskever I., Chen K., Corrado G., Dean J. Distributed representations of words and phrases and their compositionality // Advances in neural information processing systems. — 2013. — Vol. 26. — P. 3111–3119.
[10] Pennington J., Socher R., Manning C. D. Glove: Global vectors for word representation // Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). — 2014. — P. 1532–1543.
[11] Joulin A., Grave E., Bojanowski P., Mikolov T. Bag of tricks for efficient text classification // Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. — 2017. — Vol. 2. — P. 427–431.
[12] Lee K., Filannino M., Uzuner Ö. An Empirical Test of GRUs and Deep Contextualized Word Representations on De-Identification // MedInfo. — 2019. — P. 218–222.
[13] Devlin J., Chang M.-W., Lee K., Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. — 2018. — Vol. 1. — P. 4171–4186.
[14] Chizhik A., Zherebtsova Y. Challenges of Building an Intelligent Chatbot // International Conference «Internet and Modern Society» (IMS-2020). CEUR Proceedings. — 2021. — Vol. 2813. — P. 277–287.