Специфика анализа текстовых данных из социальных медиа при оценке социального самочувствия жителей мегаполиса (на примере Санкт-Петербурга)

А.В. Чижик

***Аннотация*— Статья посвящена описанию экспериментов, связанных с попыткой выстраивания методологии анализа текстовых данных из социальных сетей с целью последующей оценки социального самочувствия жителей города. Основной задачей было выявить наиболее оптимальную модель векторизации коротких текстов (комментарии к постам) для дальнейшего использования в рамках анализа тональности. В статье приводятся результаты сравнения трех актуальных на данный момент подходов к созданию векторного представления: учет веса слова в документе (TF-IDF), использование дистрибутивной семантики при создании векторов слов (Word2Vec) и векторизация целых предложений (Laser). В статье описан дизайн исследования, приведены метрики качества, описаны данные, на которых проводились опыты. Далее приводятся промежуточные результаты последующего исследования текстовых данных в рамках анализа социального самочувствия: к текстам применяется тематическое моделирование, каждая тема измеряется по пятибалльной шкале эмоций. Для экспериментов использовались данные из социальной сети Вконтакте.**

***Ключевые слова*— векторизация текстов, анализ тональности, тематическое моделирование, социальные медиа, социальное самочувствие.**

I. Введение

Город является сущностью, которую можно достаточно легко описать в географических и административных категориях, однако также городская среда – это сложное социальное образование, которое имеет свою уникальную динамику и характер, продиктованный особенностями проживающих на этой территории людей. Таким образом, географически, город – это определенная территория, обычно характеризующаяся плотностью населения, типом зданий и системой инфраструктуры, истинная сущность которой просматривается через призму социокультурного анализа. Стоит отметить, что процесс притяжения населения к крупным городам продолжается, мегаполисы становятся все более плотно населенными и развитыми, а, значит, выявление факторов, которые влияют на динамику развития крупных городов, становится все более актуальным. Понимание динамики крупных городов и ее роли в формировании мировоззрения социальных групп является ключевым инструментом для создания более устойчивых, инклюзивных и прогрессивных городских сред.

Социальные отношения, структуры и динамика городской среды влияют на социальное самочувствие жителей, оказывая существенное влияние на психологическое и эмоциональное состояние индивидов, а, как следствие, на характер социальных связей и характеристики социальных групп, функционирующих внутри городской среды.

Социальное самочувствие населения мегаполисов можно рассматривать как интегральный показатель, который отражает общее состояние социального благополучия, удовлетворенности жизнью и социальных отношений в данной городской среде. Мегаполис (как и вообще город) в первую очередь является местом, где формируются и взаимодействуют различные социальные группы. В нем находятся школы и университеты, рабочие места, торговые центры, места культурного досуга и развлечений, что создает разнообразные социальные сети и структуры. Поэтому при классической оценке социального самочувствия жителей анализируется несколько важных аспектов жизни индивидов в городской среде: уровень удовлетворенности качеством жизни, уровень стресса и/или психологического благополучия, социальные связи, доступность здравоохранения, уровень безопасности, доступ к образованию, экономическое благополучие. В большинстве случаев для выявления этих показателей используются классические социологические методы (например, опросы). Однако в таком случае перед респондентом появляется вопрос-триггер, который может непроизвольно явится фактором, искажающим возможность оценки реального самочувствия индивида по исследуемым критериям (ответ индивида зависит от когнитивных процессов, которые побуждаются триггером). В то же время при взаимодействии человека с интернетом (социальными медиа, веб-сайтами, приложениями и другими

Статья получена 25 октября 2023.
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/).
А.В. Чижик, Центр технологий электронного правительства Института дизайна и урбанистики Университета ИТМО (afrancuzova@mail.ru)

цифровыми платформами) остается цифровой след в формате текстовых данных. Это означает, что информацию по важным индикаторам, характеризующим социальное самочувствие, можно попытаться получить, исследуя дискурс индивидов при их взаимодействии в онлайн-среде.

Собрав текстовые данные, можно получить информацию о тех темах, которые в большей степени интересуют индивидов (и социальные группы, с которыми человек себя ассоциирует), а также эмоциональный спектр, присутствующий у участников дискуссий. Таким образом, появляется прямая связь между анализом текстовых данных и детекцией общего эмоционального состояния и настроения отдельных социальных групп и общества в целом. В таком случае можно заключить, что анализ текстовых данных дает возможность оценить социальное настроение и связать отдельные эмоциональные шкалы с конкретными маркерами социального самочувствия (ЖКХ, здоровье, семья и т.п.).

Ниже будут представлены результаты опытов по созданию подхода к анализу текстовых данных применительно к заявленной задаче.

II. «Социальное настроение» vs «социальное самочувствие»

Изучая характерные особенности современного социокультурного пространства, определяющиеся социальной динамикой, важно отметить, что социальное самочувствие индивидов является движущей силой трансформации социального и культурного среза. Используя категорию самочувствия, исследователи получают доступ к системе оценивания изменений, которые происходят в общественном сознании и социальной структуре общества, что, в частности, позволяет детектировать зоны социальной напряженности и групповые представления о социокультурном пространстве, в котором индивиды осуществляют свои повседневные практики. Стоит, однако, сразу отметить, что сама по себе категория самочувствия статична (в силу того, что в ней заложена попытка фиксации текущего состояния) и не может являться репрезентативной единицей анализа динамики.

Термин «социальное самочувствие» используется в социологии с 70-х годов XX века. В ряде исследований, относящихся к американской социологической школе, можно встретить аналогичное ему понятие «субъективное благополучие». Близкое по значению понятие, «социальное настроение», вошло в категориальный аппарат философов и психологов в 60-х годах XX века. Настроение по своей идейной сути может описывать процесс социальной динамики за счет реактивности, которая заложена в сам термин (предполагающий прогностическую диагностику развития процесса), поэтому интересным представляется нахождение связующих нитей между социальным самочувствием и социальным настроением.

Чтобы раскрыть значение термина «социальное настроение» для начала необходимо выделить еще два важных термина: индивидуальное и групповое настроения. Эти феномены часто связаны друг с другом за счет циркулярной реакции, притягивающей индивидуальное настроение человека к полюсу, к которому стремится социальная группа, находящаяся в коммуникативном взаимодействии с индивидом в силу каких-то причин (например, такой группой может являться рабочий коллектив, жители одного района, родители детей, поступающих в один год в одном городе в ВУЗ, и т.д.). Иными словами, динамический стереотип восприятия является психофизической основой настроения индивида, а, значит, в настроении как целостной форме жизнеощущения человека находят свое выражение социальные процессы. Выходит, что индивидуальное настроение влияет на атмосферу в обществе. Однако в ряде исследований конца XIX и начала XX веков (В.М. Бехтерев, Л.И. Петражицкий, Н.М. Сеченов, П.А. Сорокин, Г. Тард, Г. Лебон, К. Юнг и Х. Ортега-и-Гасет) четко обосновывается факт того, что индивидуальное настроение является неустойчивым феноменом, имеющим тенденцию к быстрому угасанию. Поэтому важно понимать механизм соединения индивидуальных настроений в общее, выражающее мнение социальной группы, а также принимать тот факт, что групповое настроение – следствие суммы индивидуальных. Итак, объединение в групповое настроение дает возможность выражаемой эмоции существовать продолжительное время и, даже, с течением времени превращаться в общественное мнение, которое достаточно часто является прямым базисом социокультурной трансформации.

Следуя за традицией русской социологической школы, выделим категорию социального настроения как производную от группового [Ядов, 1995], которое В.А. Ядов предлагал разделять на политическое, массовое и социальное. Определяя социальное настроение, философ Б.Ф. Поршнев описывал его следующим образом [Поршнев, 1966]: «социальное настроение – это эмоциональные состояния, связанные с осуществлением или неосуществленностью. … Как правило, социальное настроение – это эмоциональное отношение к тому, что стоит на пути, кто мешает, или, напротив, кто помогает воплощению желаемого в жизнь». Таким образом, важной особенностью социального настроения является наличие субъекта и объекта настроения.

Понятие социального самочувствия как инструментарий для исследования социальной динамики появилось в середине 80-х годов XX века, формируясь на основе социального настроения. Предполагалось, что рассмотрев структуру жизнедеятельности субъекта в качестве внешних детерминант социального самочувствия, можно зафиксировать чувственное переживание социума на макро- или микроуровне. Социальное самочувствие является характеристикой устойчивого отношения людей к внешней, окружающей непосредственно исследуемую социальную группу, среде. Исследователи отмечают, что возникающие на фоне социального самочувствия мышление и поведение индивидов играет важную роль в формировании общественной атмосферы. Следовательно, в социальном самочувствии

выражается общая тональность общественных настроений социальной группы, оно формируется в процессе проводимого людьми сопоставления возможностей удовлетворения своих потребностей и реализации интересов с возможностями, которые видятся как базовые (по логике социального среза, к которому относится индивид или по культурным стереотипам общества в целом). Стоит отметить, что структура воздействующих на социальное самочувствие факторов сложна и включает в себя социальные явления различных уровней, поэтому полное представление об этом явлении может быть получено только посредством синергии методов с целью детекции совокупности всех воздействий.

Представители социально-конструктивистского направления П. Бергер и Т. Лукман в ходе своих исследований отмечали [Бергер, Лукман, 1995], что нормальное и отклоняющееся от нормы социальное самочувствие должно трактоваться как социальный конструкт, создаваемый сознанием самих людей. Ученые отмечают, что общество создается благодаря деятельности индивидов, которые обладают знанием в виде субъективных значений или коллективных представлений, при этом члены общества считают их реальными. Таким образом, социальный мир всегда переживает три ипостаси: экстернализация, объективизация и интернализация. Последняя фаза означает, что индивид присваивает «объективный» опыт, что трансформирует этот опыт в субъективную реальность. Эту же мысль доказывают и представители структуралистского конструктивизма (П. Бурдье, Э. Гидденс), которые основой социального самочувствия выдвигают размышление о себе «с оглядкой на других» [Бурдье, 1993; Бурдье, 2019; Гидденс, 2005]. Таким образом, социальное самочувствие является одновременно механизмом воспроизводства социальных практик и продуктом объективной социальной действительности.

Размышляя в прагматическом контексте над феноменом социального самочувствия, необходимо фокусироваться на факторном анализе социальных, экономических и политических аспектов жизнедеятельности общества. Поэтому социальное самочувствие традиционно выявляется путем опроса индивидов, в ходе которого выясняется их отношение к разным аспектам жизни внутри интересующей локации (страны, города, района): респонденту задаются вопросы, выявляющие степень реализации жизненных планов, удовлетворенности своим положением в обществе и т.п. Таким образом, в область фокусировки исследований попадает оценка эмоционально-оценочного отношения индивидов к конкретной социальной реальности в конкретный период времени и в конкретной локации. Такое понимание сущности социального самочувствия как индикатора сознания общества выявляет главную особенность: социальное самочувствие складывается из реакции на множество факторов, состоящих в отношении взаимодействия друг с другом, а значит работающих вместе как система.

Справедливо заметить, что прямые вопросы, задаваемые респондентам в рамках обычной практики анализа социального самочувствия, дают естественную погрешность, так как индивид в момент ответа сталкивается со стереотипами и прочими видами культурных смещений, зафиксированными в его сознании на уровне доминирующей культуры и, возможно, активно транслируемые социальной группой. Они являют в какой-то мере идеальное представление человека о таких категориях как «успешность», «благополучность» и «стабильность». Как было отмечено выше, достаточно важно проанализировать именно эмоциональный контекст социального самочувствия, так это подразумевает присутствие настоящего или даже будущего времени в выводах по итогу анализа самочувствия (в то время как обращения к идеальным оценкам индивида приводит к рефлексии над прошедшим временем, то есть моментом, который общество, по сути уже прошло). Таким образом, приступая к анализу динамично трансформирующегося социокультурного ландшафта необходимо в какой-то степени смешивать (объединять) понятия социального настроения и социального самочувствия. Таким образом, целесообразным представляется переключение на выявление социального настроения и его сопоставления с количественными данными, описывающими локацию функционирования индивидов (в рамках данного исследования рефлексия строится вокруг города как целостного объекта анализа, и его районов как единицу дифференциального анализа сущности явления).

В настоящее время высокий уровень вовлеченности индивидов в цифровые коммуникации предполагает, что каждый день в открытом доступе появляются текстовые данные, которые могут рассказать о передвижениях горожан, их мнении о городской инфраструктуре, об уровне агрессивности социальных групп. Концепция открытых данных, в свою очередь, предполагает наличие большого объема данных, описывающих разные стороны социального благополучия региона. Таким образом, опираясь на вычислительные модели эмоций, и анализируя взаимосвязь реальности (статистик) и эмоционального фона индивидов (анализ тональности текстов), становится возможным приблизиться к оценке социального самочувствия в обход стандартным социологическим методам и при этом учитывая социальное настроение как главный фактор его формирования.

III. Анализ текстов из социальных медиа для оценки социального самочувствия: подход и методы

Наиболее удобным местом сбора текстовых данных при исследовании социального самочувствия являются социальные сети. Во-первых, в них содержится большое количество публичных дискуссий зарегистрированных пользователей, а, значит, эмоциональные спектры можно смотреть в развитии и в том числе анализировать как индивидуальное настроение, так и настроение социальной группы. Во-вторых, достаточно часто в

социальных медиа фиксируется в той или иной форме географическая привязка к местности пользователя: в профиле можно найти город, с которым ассоциируется себя индивид, а сами дискуссии могут протекать в группах, связывающих представителей одной локации (города, района, улицы).

Однако дискурс социальных сетей является особой формой коммуникации, вследствие чего исследование таких текстов имеет ряд специфических особенностей. Главной особенностью дискурса являются краткость сообщений и наличие в письменной форме черт устной (разговорной) речи. Также необходимо учитывать, что в социальной сети общение строится вокруг постов, к которым доступна опция комментариев. На рис. 1 показана логика развития полемики в социальной сети.

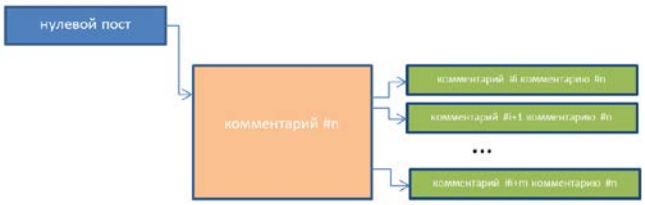

Рис. 1. Логика развития дискуссии под публичным постом в социальной сети ВКонтакте

Из рисунка видно, что пользователи могут писать как комментарии к посту, так и комментарии к n-му комментарию (и логически они будут связаны с ним, а не с постом как таковым). Это означает, что пост можно исследовать с точки зрение выделения темы, а комментарии под ним всегда насыщены информацией об эмоциях пользователей на предложенную тему. Оптимальным представляется проводить оценку настроений всей группы комментариев, что дает возможность оценить социальное настроение группы по отношению к теме.

Итак, общий подход к исследованию социального самочувствия может сводиться к следующей схеме (рис.2):

Рис. 2. Подход к исследованию социального самочувствия горожан

При этом необходимо понимать, что наибольшие проблемы могут возникнуть с анализом тональности, так как зафиксировать эмоцию в коротком сообщении – достаточно нетривиальная задача. Поэтому в рамках данного исследования был проведен анализ моделей векторизации текстов с оценкой успешности их использования в задаче оценки тональности.

A. Анализ тональности коротких текстов из социальных сетей

Итак, одним из важных nlp-методов, который актуален для исследования социального самочувствия, 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.

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].

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.

B. 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 (месседжер 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.

C. 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.

На первом этапе эксперимента было принято решение посмотреть способность анализируемых моделей эмбеддингов к разделению на кластеры (использовался алгоритм K-средних, k=2). Результаты разбиения на два кластера представлены на рис. 3. Для визуализации результатов использовался алгоритм понижения размерности PCA.

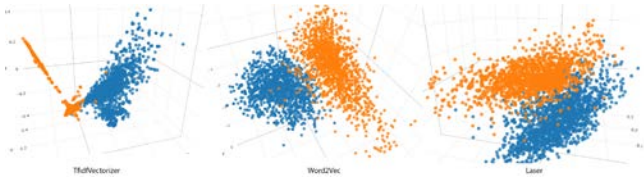

Рис. 3. Анализ разделимости классов с использованием 3D-сжатия векторного пространства с использованием алгоритма PCA

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. 4.

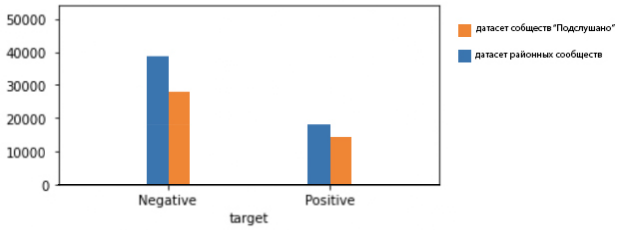

Рис. 4. Распределение размеченных классов в двух датасетах

После этого были выделены слова (и словосочетания), вносящие наибольший вклад в каждый из классов, и построены облака слов для обоих классов, что дало 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. 5 представлены результаты работы логистической регрессии при анализе датасета районных сообществ.

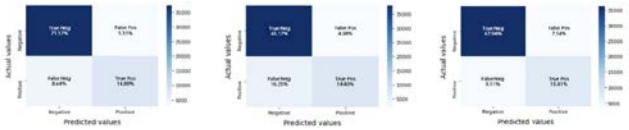

Рис. 5. Результаты работы модели логистической регрессии при отправленных в нее векторных представлениях, полученных тремя способами (слева направо: tf-idf, w2v, Laser)

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».

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.

D. Тематическое моделирование

В качестве базиса итогового анализа тональности текстов была использована модель П. Экмана. Ученый выделил пять основных базовых эмоций, которые считаются универсальными, так как переживаются всеми людьми независимо от их культурных или социальных особенностей. Эти эмоции задают мотивационные модели функционирования индивидов в городской среде, а также являются причинами демотивации и упадка социальной активности. Итак, базовыми П. Экман называл следующие эмоции:

1) Счастье (радость): эта эмоция характеризуется положительными чувствами, такими как удовлетворение, умиротворение и радость. Счастье может быть вызвано различными событиями, достижениями или приятными впечатлениями.
2) Грусть: выражает чувство потери, разочарования или печали. Она может быть вызвана разными событиями, включая утрату, разрыв отношений или неудачи.
3) Гнев: эмоция, связанная с негодованием, раздражением и даже яростью. Она может возникать при ощущении несправедливости, ущерба или нарушении личных границ.
4) Страх: возникает в ситуациях, когда человек ощущает угрозу или опасность. Это естественная реакция, которая помогает остерегаться потенциальных рисков и сохранять безопасность.
5) Удивление: эта эмоция связана с неожиданными событиями или открытиями, выражает интерес к необычному или новому.

Для анализа тональности мы использовали готовую модель rubert-tiny2 на основе набора данных CEDR [15]. На рис. 6 показаны результаты анализа тональности с учетом этих пяти эмоций.

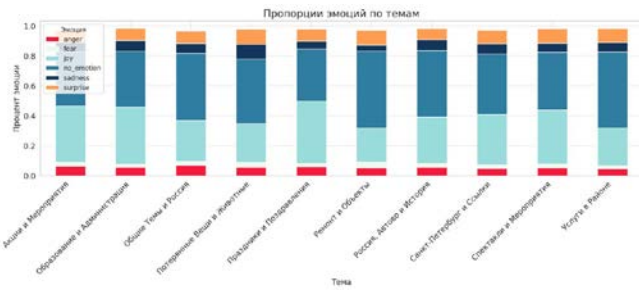

Рис. 6. Соотнесение тем дискуссий и эмоций горожан

Как видно из рисунка выделить спектр эмоций удалось. Далее предстоит работа по созданию более точного метода для выделения тем.

IV. ЗАКЛЮЧЕНИЕ

В данной статье сделана попытка обобщить результаты первых экспериментов по использованию анализа текстовых данных из социальных сетей для анализа социального самочувствия. Представляется важным использовать анализ тональности как метод, дающий информацию о социальном настроении населения. Серия экспериментов дала информацию о том, какую модель векторизации текстовых данных следует использовать для лучшей детекции тональности сообщений. Далее был проведен опыт по использованию модели П. Экмана для выявления эмоционального спектра комментаторов в социальных медиа. Также в статье обозначен подход, при котором тема поста может быть использована для соотнесения с конкретным маркером социального самочувствия, в таком случае выявленный спектр эмоций может быть связан с показателями, о которых в классическом подходе респондентам задают вопросы. Итак, в результате исследования была формализована логика подхода к латентному выявлению маркеров социального самочувствия. Далее планируется провести серию экспериментов для юстировки метода выделения тем: необходимо подобрать метод, при котором темы смогут распределяться, притягиваясь к заранее выделенным центроидам в виде социальных ролей (семья, работа, благосостояние и т.п.). На данный момент, как видно из итогового рисунка, темы выделялись просто по факту присутствия и мало подходят для репрезентативного сопоставления с индикаторами социального самочувствия. Однако проведенные опыты свидетельствуют о применимости подхода.

Благодарности

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/).

Библиография

[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. No. 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. No. 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. 2018. arXiv:1804.06659.

[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. No. 3. P. 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 // IMS. 2020. P. 277-287.
[15] Sboev A., Naumov A., Rybka R. Data-driven model for emotion detection in Russian texts //Procedia Computer Science. 2021. Vol. 190. P. 637-642.

Чижик Анна Владимировна, к. культ., старший научный сотрудник Центра технологий электронного правительства Института дизайна и урбанистики, Университет ИТМО (http://itmo.ru/), доцент кафедры информационных систем в искусстве и гуманитарных науках Санкт-Петербургского государственного университета, Санкт-Петербург, email: afrancuzova@mail.ru, elibrary.ru: authorid=708001, scopus.com: authorId=57222136821, ORCID: orcidID=0000-0002-4523-5167

Text analysis of Social Media comments for assessing the social well-being of metropolitan residents (St. Petersburg's example)

Anna V. Chizhik

Abstract— The paper is devoted to a description of experiments related to an attempt to build a methodology for analyzing text data from social networks with the aim of subsequently assessing the social well-being of city residents. The main task was to identify the most optimal vectorization model for short texts (comments on posts) for further use in sentiment analysis. The article presents the results of a comparison of three currently relevant approaches to creating vector embeddings: taking into account the weight of a word in a document (TF-IDF), using distributional semantics when creating word vectors (Word2Vec) and language-agnostic sentence embeddings (Laser). The article describes the design of the study, provides quality metrics, and describes the data on which the experiments were conducted. The following are intermediate results of a subsequent study of text data within the framework of the analysis of social well-being: topic modeling is applied to the texts, each topic is measured on a five-point scale of emotions. For the experiments, data from the social network Vkontakte was used.

Keywords— text vectorization, sentiment analysis, topic modeling, social media, social well-being.

REFERENCES

[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. No. 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. No. 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. 2018. arXiv:1804.06659.
[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. No. 3. P. 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 // IMS. 2020. P. 277-287.
[15] Sboev A., Naumov A., Rybka R. Data-driven model for emotion detection in Russian texts //Procedia Computer Science. 2021. Vol. 190. P. 637-642.

Anna V. Chizhik, Ph.D in Cultural Studies, Senior Researcher of E-Governance Center, Institute of Design and Urban Studies, ITMO University (http://itmo.ru/), Associate Professor, Department of Information Systems in Arts and Humanities, St. Petersburg State University, Saint-Petersburg, email: afrancuzova@mail.ru, elibrary.ru: authorid=708001, scopus.com: authorId=57222136821, ORCID: orcidID=0000-0002-4523-5167