ТЕХНОЛОГИИ АНАЛИЗА ТОНАЛЬНОСТИ В СЕТЕВОЙ ОБРАЗОВАТЕЛЬНОЙ КОММУНИКАЦИИ

Концевой М. П., ст. преподаватель

г. Брест, БрГУ им. А.С. Пушкина

Тональность (сентимент, от англ. sentiment – чувство; настроение) – это эмоциональная составляющая коммуникации, выраженная на уровне лексемы (лексическая тональность) или коммуникативного фрагмента [1].

Тональность определяется несколькими факторами:

– субъект тональности (holder) – автор текста, автор цитаты, прямой или косвенной речи в тексте;

– собственно тональная оценка (orientation или polarity) – эмоциональное отношение субъекта к объекту: позитив / нейтрально / негатив;

– объект тональности (entity, feature) – сущность, насчет которой высказывается автор) или его свойства f (атрибуты, части объекта);

– момент времени (time), когда было оставлено мнение.

Определение тональности коммуникативного текста (письменного или устного) в психолингвистике рассматривается в контексте проблематики восприятия речи (как процесса извлечения смысла, находящегося за внешней формой речевых высказываний) и связано с решением следующих задач:

– описания речевых сообщений на основе изучения механизмов порождения и восприятия речи;

– изучения функций речевой деятельности в коммуникативном сообществе;

– исследования связи между речевыми сообщениями и характеристиками участников коммуникации (превращение намерений говорящего в сообщения, интерпретация их слушающим).

Определение тональности коммуникации предполагает выделение тех ее фрагментов (текста и речи), которые выражают эмоциональность субъекта (позитивную, нейтральную, негативную) по отношению к анализируемому в тексте объекту его эмоциональной оценки (объекту тональности). Объект тональности может определяться в предложениях текста как имя собственное или нарицательное или быть задан в целом для текста (с учетом его синонимических и анафорических употреблений).

Развитие сетевых образовательных сообществ предполагает рассмотрение Интернета не только как информационного поля, но и как социальной среды со своими собственными явлениями и социальными группами, которые необходимо изучать как важную неотъемлемую часть современного социума. Сетевые образовательные коммуникации отличаются в вербальном плане небывалым разнообразием, вариативностью, общедоступностью и множественностью способов влияния на психику человека, в том числе посредством продуцируемых ими текстов, которые становятся проводниками социального воздействия и коммуникации. Изучение и анализ текстов в сетевых образовательных коммуникациях открывает доступ к исследованию новых граней и уровней языка, психики современного человека и актуальной социальной проблематики. Анализ тональности находит свое практическое применение в педагогике и психологии, социологии образования и образовательном маркетинге (сбор данных из социальных сетей и блогов о взглядах и образовательных предпочтениях), медицине и лингвистике.

Большое количество размещенного в компьютерных сетях коммуникативного материала и его высокая динамичность делают невозможным обработку и анализ традиционными средствами и методами в «ручном режиме». Разработка компьютерного инструментария анализа текстов, в том числе сентимент-анализа, является необходимым условием научного исследования современного сетевого образовательного дискурса.

Основная прикладная цель компьютерного анализа тональности образовательной коммуникации – нахождение мнений в тексте и определение их свойств (opinion mining and sentiment analysis):

– авторства (кому принадлежит данное мнение);

– тематики (о какой области сообщается во мнении);

– тональности (позиция автора относительно упомянутой темы –обычно в одномерном эмотивном пространстве: «положительная» или «отрицательная»).

Несмотря на то, что тональность является лишь одной из характеристик мнения, именно задача классификации тональности наиболее часто ставится в наши дни.

Компьютерное определение тональности коммуникации реализуется преимущественно на основе подходов, в основании которых лежат либо правила, либо словари, либо машинное обучение.

Подходы, основанные на правилах, предполагают разработку набора правил, применяя которые система делает заключение о тональности текста. Такие подходы трудоемки, так как для хорошей работы системы необходимо составить большое количество правил, которые привязаны к определенной тематике. Данный подход является наиболее точным при наличии хорошей базы правил, но непригодным для психолингвистического исследования.

Подходы, основанные на словарях, предполагают создание и применение тональных словарей (affective lexicons), в которых каждой вокабуле приписывается значение тональности. Для анализа каждому слову в тексте присваивается его значение тональности из словаря (если оно присутствует в словаре), а затем вычисляется общая тональность всего текста. Вычисление тональности может быть реализовано на разных уровнях сложности: от нахождения простого среднего арифметического всех значений до применения машинного обучения классификатора на основе нейронной сети.

Для повышения эффективности извлечения из текста эмотивных мнений может быть использован словарь интенсификаторов (слов, увеличивающих или

уменьшающих вес тонального слова). Интенсификаторы обычно ранжируются по следующей шкале: высокий, средний и низкий. Лингвистическая база данных может содержать список слов, инвертирующих вес тонального слова, и слова, являющиеся ключевыми для выбранной предметной области, что позволяет избежать анализа той лексики, которая по своей сути является тональной, но не имеет отношения к анализируемой тематике [2].

Основной проблемой методов, основанных на словарях и правилах, считается трудоёмкость процесса составления словаря. Для того чтобы получить метод, классифицирующий документ с высокой точностью, термины словаря должны иметь вес, адекватный предметной области документа. Например, слово «огромный» по отношению к объему памяти жёсткого диска является положительной характеристикой, но отрицательной по отношению к размеру мобильного телефона. Поэтому данный метод требует значительных трудозатрат, так как для хорошей работы системы необходимо составить большое количество правил.

Машинное обучение (machine Learning) – системная совокупность инструментов извлечения знания из данных (data mining) на основе методов математической статистики, численных методов оптимизации, теории вероятностей, дискретного анализа. Машинное обучение разработано в проблемном поле искусственных нейросетей и реализует два основных способа обучения:

– с учителем (supervised learning) – для каждого прецедента задаётся пара «ситуация, требуемое решение»;

– без учителя (unsupervised learning) – для каждого прецедента задаётся только «ситуация», требуется сгруппировать объекты в кластеры, используя данные о попарном сходстве объектов, или понизить размерность данных.

Машинное обучение без учителя представляет собой перспективный, но наименее точный метод анализа тональности. Оно находит применение в автоматической кластеризации документов.

Машинное обучение с учителем является наиболее распространенным методом, используемым в исследованиях. Его суть состоит в том, чтобы обучить машинный классификатор на коллекции заранее тонально размеченных текстов, а затем использовать полученную модель для анализа новых документов.

Для создания системы анализа тональности на основе машинного обучения с учителем необходимо собрать коллекцию документов для обучения классификатора; каждый документ из обучающей коллекции нужно представить в виде вектора признаков; для каждого документа нужно указать «правильный ответ», т.е. тип тональности (положительная или отрицательная). По этим ответам и будет обучаться классификатор.

В исследованиях обычно рассматривается задача бинарной классификации тональности, т.е. классов всего два: «положительный» и «отрицательный». Качество результатов напрямую зависят от набора характеристик для составления вектора признаков. Наиболее распространенные способы представления документа в задачах психолингвистики основаны на модели «мешка слов» (bag-of-words), либо набора n-грамм (последовательности из n-элементов: слогов, слов, букв и др.). Обычно униграммы и биграммы дают лучшие результаты, чем n-граммы более высоких порядков (триграммы и выше), т.к. выборка обучения в большинстве случаев недостаточна большая для подсчета n-грамм высших порядков. Повысить качество анализа можно использованием дополнительных признаков (части речи, пунктуация, смайлики, междометия и т.д.).

Улучшить результаты анализа можно присвоением каждому признаку вектора некоторого «веса» (количественной характеристики, определяющей иерархическое положение признака относительно других признаков в некоторой системе оценки). Наиболее распространенным методом оценки веса признаков является TF-IDF (от англ. TF – term frequency, IDF – inverse document frequency) – статистическая мера, используемая для оценки важности слова в

контексте документа, являющегося частью коллекции документов или корпуса. В TF-IDF вес некоторого слова пропорционален количеству употребления этого слова в документе и обратно пропорционален частоте употребления слова в других документах коллекции.

Примером программного инструментария для компьютерного анализа текста является Sentistrength (http://sentistrength.wlv.ac.uk) – программа оценки силы положительных и отрицательных настроений в текстах, ориентированная на работу с краткими текстами.

SentiStrength работает на основе словаря эмоционально окрашенной лексики, все слова в котором закодированы от –5 до –1 для слов, выражающих отрицательные эмоции, и от 1 до 5 для слов, выражающих положительные эмоции. Словарь, который используется в англоязычной версии, включает в себя тысячи эмоционально окрашенных слов, а также так называемые слова-усилители «BoosterWords» (absolutely, definitely и др.), список идиоматических выражений, слова-отрицания «NegatingWords» (например, never, don’t), список вопросительных слов и сленговых выражений. SentiStrength также использует целый ряд признанных нестандартных написаний и других распространенных текстовых способов выражения чувств (сокращения, смайлы и т.п.).

Для определения положительных эмоций и отрицательных коннотаций SentiStrength имеет сравнимую степень точности в сравнении со стандартными алгоритмами «машинного обучения».

Предельная точность результатов сентимент-анализа для лучших программ не превышает 80%. Для англоязычных текстов удалось довести её на основе обучения рекурсивной нейросети до 85% с перспективой дальнейшего повышения точности в результате обучения [3]. Авторы создали модель NaSent (Neural Analysis of Sentiment), которую называют рекурсивной тензорной нейросетью (Recursive Neural Tensor Network). Она используется для обработки отдельных слов в каждой фразе, построения дерева взаимосвязей и проводит анализ эмоциональной окраски каждого слова, определяет влияние слов друг на

друга. Программа строит дерево с оценкой каждого слова, каждой фразы и всего текста целиком.

На демонстрационном онлайновом сервисе стэнфордского университета (http://nlp.stanford.edu:8080/sentiment/rntnDemo.html) можно изучить и опробовать работу Recursive Neural Tensor Network. Данный сервис одновременно является инструментом для обучения нейросети, и каждый пользователь может, предложив программе произвольный текст для анализа, скорректировать результат, исправив ошибки.

Литература

1. Гаспаров, Б. М. Язык, память, образ. Лингвистика языкового существования – М. : Новое лит. обозрение, 1996. – 352 с.
2. Бочкова А. Л. Лингвистическая база данных как основа системы автоматического извлечения мнений участников интернет-коммуникации // Карповские научные чтения: сб. науч. ст. 2014. URL: http://elib.bsu.by/handle/123456789/100518 (дата обращения: 22.03.2020).
3. Richard Socher, Alex Perelygin, Jean Y. Wu and etc. Recursive Deep Models for Semantic Compositionality Over a Sentiment URL: http://nlp.stanford.edu/~socherr/EMNLP2013_RNTN.pdf (дата обращения: 22.03.2020).