Градуировка тональной шкалы реакций на интернет-новости (на материале русского, английского и испанского языков)

Исполнитель: Савинцева Дарья Викторовна
Научный руководитель: к. ф. н., доц. Азарова И.В.

16 июня 2020 г.

1. *Титульный слайд*

   Добрый день, уважаемые члены комиссии! Вашему вниманию представлена выпускная квалификационная работа на тему *Градуировка тональной шкалы реакций на интернет-новости (на материале русского, английского и испанского языков)*.

2. *Цель и задачи работы*

   Целью работы стало создание тональной шкалы тем, представленных в пользовательских комментариях на актуальные новости, размещенных в социальных сетях.

   Для этого необходимо было выполнить следующие задачи:

   (a) подобрать и дополнить тональные лексиконы
   (b) выбрать тематическую модель;
   (c) создать корпусы комментариев в новостных группах социальных сетей;
   (d) определить алгоритм выделения тональных компонентов в темах;
   (e) представить шкалу для сопоставления тональных компонентов в текстах на разных языках.

3. *Анализ тональности и тематическое моделирование*

   Наше исследование касается двух направлений в обработке естественного языка:

   - Анализ тональности — направление, целью которого стоит извлечение из текста тонального компонента, которым может являться настроение, мнение, отношение, эмоции по отношению к какому-либо объекту или событию. Этот компонент обоснован наличием в семантической структуре слова эмоционально-экспрессивной составляющей, которая проявляется или в определенных контекстах, или заключена непосредственно в лексическом значении.

     Существует множество методов выделения тональных слов в тексте. Мы использовали словарный метод из-за относительной простоты и уже наличия материалов (тонального словаря).

   - Тематическое моделирование — инструмент статистического анализа текста, с помощью которого можно выделить темы в коллекции документов, и то, как эти темы связаны друг с другом.

     В нашей работе для реализации тематического моделирования задействован алгоритм латентного размещения Дирихле (LDA) из библиотеки gensim, которая написана специально для реализации тематических моделей и проста в использовании.

4. *Гипотеза*

   Основная гипотеза нашего исследования состоит в следующем: распределение тональных компонентов в темах комментариев может показать, насколько тонально окрашена та или иная тема в сравнении с остальными, а также выявить разницу в тональной окраске тем в разных языках. Таким образом становится возможным создать тональную шкалу тем, представленных в комментариях пользователей.

5. *Корпусы комментариев*

   Данные для анализа тональности и тематического моделирования на русском языке были собраны из социальных сетей Вконтакте и YouTube. Для отбора основных русскоязычных СМИ мы обратились к Базе данных СМИ Яндекса. Мы отобрали 47 новостных сообществ, которые:

- относятся к тематике «Главное»;
- относятся к типам «Газеты» (ежедневные, еженедельные), «Телеканалы», «Тематические сайты»;
- имеют свою группу Вконтакте с числом подписчиков примерно 100 тыс. человек;
- с открытыми комментариями.

Для отбора новостных каналов YouTube мы воспользовались интернет-ресурсом, предоставляющим статистику популярности новостных каналов разных стран, в частности, США и испаноговорящей Латинской Америки. Критерии для сбора были похожими на вышеперечисленные: новостная тематика канала, открытые комментарии под видео и большое количество подписчиков. Было отобрано по 15 каналов для США и стран Латинской Америки.

Далее мы произвели сбор комментариев из отобранных каналов и сообществ при помощи парсеров. Комментарии корпуса Вконтакте относятся к временному промежутку с 01.01.2020 по 16.04.2020. Комментарии корпусов YouTube относятся к временному промежутку с 01.01.2020 по 11.05.2020. Основные количественные показатели корпусов представлены на слайде. Корпусы российских и американских комментариев получились примерно одинаковыми по размеру, а корпус латиноамериканских — в разы меньше. После предобработки эта разница сокращается, но всё ещё видна по количеству слов в словарях.

6. *Тональный словарь для русского языка*

   Тональный лексикон для русского языка был создан на основе тонального словаря проекта *Linis-crowd*. Он был доработан и дополнен в ходе работы над групповым проектом по определению тональных компонентов в текстах новостей (в рамках учебной программы). Он состоит из тонально окрашенных лемм, тональность которых положительная $(+1)$ либо отрицательная (-1). Также в нём присутствуют нецензурные слова, тональность которых (-2). Всего в тональном лексиконе русского языка насчитывается 11,653 единицы. Справа на слайде представлен отрывок из лексикона.

7. *Тональный словарь для английского языка*

   В качестве тонального лексикона для английского языка использовался *Open Sentiment Lexicon*, созданный в 2004 году и с тех пор улучшаемый авторами. Он также состоит из тонально окрашенных слов, тональность которых положительная или отрицательная. Всего в тональном лексиконе английского языка насчитывается 6789 единиц. Справа на слайде представлен отрывок из лексикона.

8. *Тональный словарь для испанского языка*

   Для испанского языка были использованы лексиконы *iSOL* (*improved Spanish Opinion Lexicon*) и *Spanish Sentiment Lexicon*. Первый был получен исследователями на основе приведенного выше тонального лексикона на английском (переведен и отредактирован). Второй лексикон был создан на основе Spanish WordNet. Мы объединили эти тональные лексиконы, удалив возникшие в процессе дубликаты и привели к необходимому для анализа виду. Каждое слово в лексиконе принимает либо положительную, либо отрицательную тональность. Всего в тональном лексиконе испанского языка насчитывается 10,463 единицы. Справа на слайде представлен отрывок из лексикона.

9. *Выделенные темы в корпусах*

   Сначала мы построили пробные модели для каждого языка на 10, 15, 20, 30 тем, и посмотрели, какая больше соответствовала нашим критериям интерпретируемости тем и близости или удаленности тем при визуализации тем. На основании этого эксперимента мы установили, что 15 тем — это оптимальное количество тем для каждого корпуса.

   На слайде представлены темы для каждого корпуса комментариев. Они находятся в том порядке, в котором их выделила тематическая модель. Для российских комментариев характерно обсуждение коронавируса как в мировом, так и в бытовом масштабе; в темах американских комментариев видно, что практически все темы сконцентрированы вокруг происходящего внутри страны, более глобальные темы не затронуты; латиноамериканские комментарии представляются наиболее реакционными, насыщенными общественно-политическими вопросами местного и мирового масштаба. Также для каждого корпуса мы определили схожие между собой темы и уникальные, встречающиеся только в данном корпусе.

10. *Визуализация полученных тем*

    Затем мы составили и проанализировали тематические профили для каждого новостного сообщества. На слайде для примера представлено ранжирование тем обсуждений в комментариях в новостном сообществе Вконтакте *Медуза* и YouTube канала *HuffPost* по вероятности, с которой та или иная тема появится в комментариях сообщества. Можно сделать выводы, какие темы волнуют подписчиков той или иной группы в большей или меньшей степени и тем самым больше узнать аудиторию групп. Можно сказать, что внимание пользователей в комментариях новостных сообществ и каналов всех исследуемых групп сосредоточено больше на общественно-политических темах, нежели чем на личных и бытовых.

11. *Определение тональных компонентов в российском корпусе*

    Мы разметили получившиеся темы с помощью тональных словарей и определили количество тональных компонентов в каждой.

    Для русского языка получилось, что в бытовых темах (*повседневная жизнь*, *бытовые проблемы*, *жизнь в россии*, *работа*) в целом присутствует больше тональных компонентов, чем в общественно-политических. То есть темы, вероятность появления которых в комментариях меньше, имеют большую тональную окраску, чем те, которые с большей вероятностью присутствуют в новостных сообществах. Количество отрицательных тональных компонентов значительно превышает количество положительных. Более всего положительного отклика находят темы *повседневная жизнь*, *общественные вопросы* и *финансовые проблемы*, в то время как самыми отрицательно окрашенными темами являются *бытовые проблемы*, *жизнь в россии* и *работа*. У темы *бытовые проблемы* более всех отрицательных и менее всех положительных тональных компонентов: она воспринимается пользователями в наиболее негативном ключе в сравнении с остальными темами. Также примечательно, что тема *карантин* представлена как одна из наиболее насыщенных положительными компонентами и менее всех отрицательными.

12. *Определение тональных компонентов в американском корпусе*

    У американцев, как и у российских пользователей, большую эмоциональную реакцию вызывают темы скорее личные, бытовые, моральные, нежели чем общественно-политические (темы *лжи и страха*, *бога* и *здравоохранения* самые окрашенные). Эти темы так же, как и в российских комментариях, обсуждаются меньше, а эмоциональный отклик на них больше. Тема *коронавируса* представлена самой отрицательной и менее из всех положительной. Обсуждение и восприятие темы пандемии представляется из всех тем самым негативным, в то время как тема *здравоохранения* содержит наибольшее количество положительных компонентов — в противовес теме *коронавируса*. Для темы *америка, патриотизм* представлено примерно похожее распределение тональности. Такое распределение в теме американского патриотизма говорит о его характере скорее как об агрессивном и нападающем, чем как о защитном и оборонительном. Примечательно, что тема *бог* также насыщена не только положительными, но и большим количеством отрицательных компонентов.

13. *Определение тональных компонентов в латиноамериканском корпусе*

    В латиноамериканских комментариях менее вероятные темы в группах находят больший эмоциональный отклик среди пользователей, и наоборот: часто обсуждаемая тема имеет меньший отклик. Общая тональность темы *демократия* самая высокая среди представленных. Это ясно свидетельствует о том, что пользователей эта тема волнует более всего, и обсуждают они её в максимально негативном ключе. Сравнение положительных и отрицательных компонентов этой темы также это доказывает. Помимо этого, тема *воровства и терроризма* также сильно отрицательная в среднем, и находит такой же сильный (отрицательный) эмоциональный отклик пользователей, как и первая тема в списке. Тема *трамп*, посвящённая американскому президенту, не находит сильный эмоциональный отклик среди пользователей. Также присутствуют темы, насыщенные большим количеством как положительных, так и отрицательных тональных компонентов: *венесуэла и россия* и *проблемы людей*. Можно предположить, что эти темы анализируются пользователями с двух точек зрения, например, pro et contra отношений России и Венесуэлы. Менее всего окрашенная тема *правительство* также может дать предположение о том, что пользователи сконцентрированы скорее на желании каких-то изменений и обсуждении того, что можно сделать, чем на том, какое правление осуществляется в стране на данный момент.

14. *Средние значения тональности для каждой темы*

    На слайде приведены средние значения тональности для каждой темы каждого корпуса, отсортированные по убыванию окрашенности. У российских пользователей темы, касающиеся бытовых и повседневных вопросов, находят больше тонального отклика у пользователей, чем общественно-политические темы. В американском корпусе темы пандемии и патриотизма находят самый большой отрицательный эмоциональный отклик, чем более глобальные темы. В латиноамериканском корпусе наиболее тонально окрашенными в среднем являются темы, относящиеся скорее к насущным проблемам людей, нежели чем к глобальным.

15. *Тональная шкала тем*

    Из полученных тем мы определили темы, присутствующие в том или ином виде в каждом из корпусов комментариев, объединили их в тематики (общие обозначения для выделенных тем) и сравнили их средние тональности, чтобы посмотреть, насколько тонально окрашена та или иная тема в сравнении с аналогами в других языках. Тональная шкала тем, представленных в корпусах, может быть основана на сумме средних значений тональности для каждой темы. На слайде представлены подсчитанные суммы средних значений тональности, представленных в тематиках, а также средние значения этих сумм. Тематики отсортированы по второму столбцу, который является основой тональной шкалы.

По сумме средних значений тональностей *бытовые проблемы* и *общественные вопросы* находят у пользователей больший эмоциональный отклик, чем глобальные темы *политики* и *ситуации в мире*. Между этими темами располагается промежуточная тема *коронавируса* как явления, затрагивающего как жизнь каждого конкретного пользователя, так и страну и мир в целом. По среднему значению суммы средних тональностей картина меняется не сильно. Наиболее окрашенной тематикой в обсуждениях являются *бытовые проблемы*, далее темы *коронавируса*, *нужд человека* и *ситуации в стране*, затем более глобальные темы *президент*, *общественные вопросы*, *ситуация в мире* и *повседневная жизнь*.

Выводы о количестве и распределении тональных компонентов, сделанные для каждого языка в отдельности, экстраполируются и на общие выводы: темы, затрагивающие бытовые и повседневные проблемы жизни людей, более тонально окрашены и находят больший эмоциональный отклик, чем глобальные темы общества и политики. В бытовых темах всех трёх корпусов комментариев присутствует больше тональных компонентов, чем в общественно-политических.

16. *Выводы*

В завершение хочется отметить основные итоги проделанной работы:

(a) созданы корпусы комментариев к новостным группам на русском, английском и испанском языках на материале российских, американских и латиноамериканских СМИ;

(b) с помощью тематического моделирования выделены темы, присутствующие в комментариях, построены тематические профили сообществ;

(c) темы размечены при помощи тональных словарей, показаны распределения тональных компонентов в темах;

(d) определено среднее значение тональности для каждой темы, на основании которого построена тональная шкала тем для сходных тем из трёх корпусов.

Вывод исследования: внимание пользователей в комментариях новостных сообществ и каналов всех исследуемых групп сосредоточено больше на общественно-политических темах, нежели чем на личных и бытовых, но бытовые и общественные проблемы находят у пользователей социальных сетей больший эмоциональный отклик, чем темы политики и ситуации в мире.

Большое спасибо за внимание!