Олеся Кольцова

Автоматизированные методы анализа больших массивов интернет-текстов

Сентябрь 2015

Рукопись для сборника «Онлайн исследования в России»

1. Введение

Автоматизированные методы анализа текстов становятся все более актуальными в связи с нарастающим потоком интернет-данных, представляющих интерес как для социальных наук, так и для бизнеса. Сейчас наблюдается взрывное развитие таких методов, и в то же время для большинства языков, включая русский, они находятся в стадии становления. Это ограничивает возможности их применения конечными пользователями – академическими исследователями и практическими аналитиками, в том числе и для анализа интернет-текстов. В виду своей сложности и незавершенности, успешно эти методы применяются, в основном, в крупных ИТ-компаниях – например, в поисковых сервисах – для их собственных нужд, и это знание остается закрытым.

Исходя из выше сказанного, существующие исследования в области автоматического анализа интернет-текстов можно классифицировать по нескольким основаниям. Во-первых, их можно разделить на интернет-центрированные и общество-центрированные. Первый тип - это исследования, в которых интернет-тексты являются самостоятельным объектом, а знание о них – конечной целью; например, это исследования тематики политических онлайн-дискуссий или тональности отзывов на товары. Второй тип – это исследования, в которых интернет-тексты служат средством для объяснения или прогнозирования явлений оффлайн-реальности – например, прогнозирование результатов выборов по текстам твитов или эпидемий гриппа по поисковым запросам. Оба типа пока крайне слабо представлены в России, за исключением группы работ по анализу тональности отзывов на товары.

Это подводит нас ко второму основанию классификации, по которому исследования можно разделить на «содержательные» и методологические, последние из которых доминируют. Исследователи в гораздо большей степени заняты разработкой методов, чем их применением и интерпретацией конкретных результатов. Методы же по понятным причинам разрабатываются в основном в сообществе специалистов в области программирования, математики и компьютерной лингвистики, а не социальных наук. Социальным исследователям зачастую сложно даже сформулировать запрос, не говоря уже о разработке и применении методов современного уровня, что и определяет некоторую оторванность методов от задач социальной науки. Этот разрыв намного меньше в коммерческой среде, где для его преодоления существуют более сильные стимулы и большие ресурсы.

Наконец, сами методы автоматизированного анализа текстов (которые не всегда специфичны именно для интернет-текстов) можно условно разделить на три больших группы:

1. Классификация, кластеризация и тематическое моделирование;
2. Сентимент-анализ и извлечение мнений;

3. Извлечение объектов и фактов.

Это не полная и не идеальная типология; методы могут комбинироваться, в зависимости от задач, и, более того, сами могут рассматриваться как задачи, требующие комбинации инструментов. Так, для оптимизации информационного поиска могут сочетаться кластеризация и извлечение объектов. Классификация может использоваться для задач сентимент-анализа и т.д. В этой статье мы сначала даем общую характеристику названных трех групп методов, а затем рассказываем об имеющихся российских исследованиях, в которых эти методы применяются для анализа интернет-текстов. Под российскими исследованиями понимаются работы с участием русскоязычных авторов, либо авторов, живущих в России, либо сделанные на российском или русскоязычном материале.

2. Автоматический анализ текстов для социальных наук

2.1. Подготовка текстов к анализу

Подготовка или препроцессинг текстов – с одной стороны, техническая задача, а с другой – требующая методологически выверенных и трудоемких процедур. Основная цель – сообщить компьютеру, из каких элементов состоит обрабатываемый текст для того, чтобы его потом можно было автоматически анализировать. Для большинства задач требуется, во-первых, расчленение текста на слова (что скорее представляет проблему в иероглифических языках, чем в русском). Во-вторых, обучение компьютера узнавать словоформы одного и того же слова, для чего проводится лемматизация или стемминг, и это уже отдельная большая проблема в случае флективных языков, включая русский. Чаще всего отдельные слова становятся признаками, которые потом подвергаются анализу. Но так бывает не всегда; поэтому – и это в-третьих - может потребоваться расчленение текста на последовательности слов (n-граммы), а также обучение компьютера узнаванию и вычленению частей речи, членов предложения, синтаксических конструкций, целых предложений и других элементов текста. Ни один исследователь не может осуществить препроцессинг «с нуля» и вынужден опираться на существующие инструменты: словари, программные продукты и т.д., в том числе лемматизаторы (например, MyStem (Mystem 2015) для русского языка) и парсеры (например, томита-парсер (томита-парсер 2015) для русского языка). Поэтому качество автоматического анализа текстов во многом зависит от развитости лингвистических инструментов для данного языка.

2.2. Классификация, кластеризация и тематическое моделирование

Под классификацией обычно понимают разделение элементов на заранее известные классы; в случае текстов, самым очевидным примером является разделение их по языкам. При автоматической классификации обычно применяются методы машинного обучения «с учителем»: машине подаются тексты, разделенные на классы людьми, и она «обучается» на них как на примерах. Анализируя частоту различных слов в поданных примерах, с помощью различных алгоритмов машина относит новые тексты к одному из классов: так, текст со словами “bonjour” и “merci” будет в нашем примере отнесен к классу «французский», поскольку в поданных примерах эти слова встречались в текстах

именно этого класса. К наиболее распространенным алгоритмам, используемым для классификации, относятся: метрические (например, метод ближайших соседей), вероятностные (например, наивный байесовский классификатор), линейные (например, машина опорных векторов (SVM), а также алгоритмы на основе регрессий и нейронных сетей) (обзор см. Aggarwal & Zhai 2012, 163-222; Воронцов 2008).

Однако зачастую состав классов не известен заранее: это особенно характерно для таких задач как определение тематического состава корпуса текстов. В таких случаях применяется кластерный анализ, не требующий «учителя» (такие алгоритмы часто называют обучением без учителя). Кластерный анализ в целом хорошо знаком социальным исследователям; кластеризация интернет-текстов отличается большими объемами – не только большим количеством объектов, но и большим количеством признаков: число уникальных слов в больших коллекциях обычно измеряется несколькими сотнями тысяч. Это требует особых, быстро работающих алгоритмов (обзор см. Aggarwal & Zhai 2012, 77-128) и подходящего программного обеспечения. Например, целое семейство алгоритмов для кластеризации данных больших размерностей (т.е. в частности текстов) было предложено Дж. Кариписом и воплощено в его программном обеспечении Cluto (Cluto 2007); алгоритмы включают специальные модификации иерархических и неиерархических моделей, включая графовые. Отдельную и очень важную задачу представляет интерпретация и присвоение имен кластерам. Кроме того, такие объекты как тексты зачастую воспринимаются как принадлежащие сразу к нескольким тематическим кластерам, а однозначное отнесение их только к одному кластеру обедняет представление о корпусе текстов.

Для решения этой проблемы используется либо нечеткая кластеризация, либо - все чаще – тематическое моделирование. Последнее - это группа вероятностных алгоритмов, сходных с факторным анализом (обзор см. Blei 2012; Steyvers & Griffiths 2007). Предполагается, что тексты являются порождением латентных переменных, называемых темами, чье распределение по словам и по текстам не известно и предстоит восстановить. В ходе работы такие алгоритмы, во-первых, приписывают каждому слову вероятность принадлежности к каждой теме, в результате чего получаются списки наиболее характерных для каждой темы слов, по которым можно судить о содержании тем, присущих корпусу текстов. Сумма вероятностей всех слов по теме служит индикатором значимости темы в корпусе. Во-вторых, вероятность принадлежности к каждой теме приписывается и каждому тексту; таким образом, выполняется задача нечеткой кластеризации текстов. Среди множества имплементаций тематического моделирования можно выделить TopicMiner как интерфейсный пакет, разработанный специально для русского языка (TopicMiner 2013).

Классификация, кластеризация и тематическое моделирование текстовых корпусов могут выполняться с использованием не только отдельных слов, но и других признаков, наиболее распространенными из которых являются биграммы (последовательности из двух слов). Это заметно улучшает качество результатов, но утяжеляет данные и усложняет работу. Поэтому исследователи пока редко идут дальше работы с отдельными словами; скорее, это дело будущего. Общей проблемой для всех перечисленных методов является вопрос выбора количества классов, кластеров или тем. Теоретически, его можно автоматизировать, задав параметры оптимизации, но вопросом остаются как раз эти

параметры, и, в частности, критерии качества работы данных алгоритмов. Наиболее известное свободное программное обеспечение, объединяющее все три подгруппы рассмотренных алгоритмов – Mallet (Mallet 2002).

2.3. Сентимент-анализ и извлечение мнений

Сентимент-анализ, или анализ тональности текста (обзор см. Pang & Lee 2008) проще всего помыслить на примере анализа отзывов на товары и услуги: его задача – автоматически разделить отзывы на негативные, то есть содержащие отрицательное отношение автора к объекту, и позитивные (положительное отношение). Существующие подходы к оценке тональности можно условно разделить на словарные (обзор см. Taboada et al 2011) и не использующие словарь. При словарном подходе алгоритм присваивает тексту оценку тональности на основании сравнения его словарного состава с тезаурусом, слова в котором вручную отнесены к позитивным или негативным. Преимущество такого подхода в простоте самого алгоритма, но качество его зависит от того, насколько удастся подобрать такие слова, которые будут приводить систему к правильному решению. Альтернативой является классификация на основе обучения с учителем: здесь, однако, потребуется вручную размечать тексты, которые будут служить образцами. В отношение отзывов это часто не проблема: отзывы оставляются на специальных сайтах, где текст сопровождается выставленной автором оценкой. Она и служит меткой тональности.

Совсем другая ситуация с записями в блогах и социальных сетях. Они не только не размечены авторами, но понятие «позитивности» и «негативности» в них не так однозначно. Во-первых, в отличие от отзывов, они могут содержать множественные объекты, и отношение к ним может быть разным. Во-вторых, структура мнений в таких текстах может быть вообще гораздо сложнее: так, во время катастроф в социальных сетях растет число записей, наполненных негативными словами («ужасно», «кошмар»), однако большинство этих текстов выражают не негативное отношение к событию или его участникам, а соболезнование пострадавшим. Отношение типа «соболезнование» не совсем укладывается в шкалу «негативное – позитивное» и требует иной типологии мнений. Для политических текстов также характерны пары отношений типа «согласие» - «несогласие» или «поддержка» - «оппонирование», все из которых могут быть выражены как с использованием эмоционально окрашенной лексики, так и без нее. Инструменты тональной оценки в целом очень чувствительны к предметной области – так, инструменты, настроенные на отзывы на автомобили, будут плохо работать даже на отзывах на фильмы. За пределами отзывов на товары и услуги качество сентимент-анализа вообще пока чрезвычайно низко. Пока можно наблюдать отдельные попытки классификации текстов в зависимости от мнений – например, Лин и Хауптманн. (Lin & Hauptmann 2006) предлагает метод автоматического разделения текстов СМИ на «произраильские» и «проарабские». Наиболее известным свободно доступным программным обеспечением для сентимент-анализа является SentiStrength (2012), однако оно пока не работает с русским языком.

К выявлению мнений примыкают задачи, связанные с поиском различных социальных проблем через пользовательский интернет-контент: нахождение экстремистских текстов

или речи ненависти, на основании расовой, этнической или религиозной принадлежности; выявление суицидальных настроений по текстам социальных сетей, выявление текстов, пропагандирующих употребление наркотиков или содержащих обращения к детям с целью их сексуального использования. Такие виды контента не совсем подходят под категорию «мнение», но методы их выявления очень схожи с сентимент-анализом и предполагают работу со словарями или машинное обучение. Также с сентимент-анализом их роднит значимость для качества результатов различных признаков текста, не сводимых к отдельным словам – синтаксических конструкций, словосочетаний, n-грамм, особенностей использования знаков препинания, заглавных букв и эмотиконов. Все эти признаки гораздо менее важны для качества тематического моделирования. Для обнаружения сложных признаков в алгоритмы вводятся правила: например, искать глагол желания 1 л. ед.ч + инфинитив глагола смерти («хочу умереть», «мечтаю сдохнуть»), возможно, на расстоянии 1-3 слова («мечтаю поскорее сдохнуть»). Использование правил часто сочетается со словарными подходами и иногда с машинным обучением.

2.4. Извлечение объектов, аспектов и фактов

Множественность объектов, на которые может быть направлено отношение – одна из проблем, решением которой занимаются методы извлечения объектов. Простейшим объектом являются персоны, которые могут быть обозначены разными способами, например: «Путин В.В.», «Владимир Путин», «президент РФ» или просто «он». Задача извлечения объекта – определить, какие слова и словосочетания указывают на один и тот же объект. После такого определения можно найти способы автоматического соотнесения выявляемых отношений или тем именно с интересующим объектом. Другие примеры объектов – организации или географические названия.

Можно выделить три основных подхода к извлечению объектов. Первый основан на онтологиях – специальных словарях, содержащих связи между элементами (например, о том, что президент РФ – должность Путина). Онтологии достаточно трудоемки в составлении и, как и все словари, чувствительны к предметной области. Другой подход – так же, как и в сентимент-анализе – машинное обучение, для которого, впрочем, требуются тексты не просто маркированные одной меткой, но детально размеченные. И третий подход предполагает ручное составление правил. Он, пожалуй, является основным для задачи извлечения фактов, то есть групп объектов и связей между ними. Примером такой задачи является автоматический поиск сообщений о терактах или о протестах. Такие тексты содержат схожие структуры, которые можно описать правилами: так, в описании теракта обычно участвуют глаголы или существительные, указывающие на взрыв; рядом с ними присутствуют объекты – устроители взрыва, не являющиеся военными, и объекты – места взрыва, как правило, публичные городские пространства вне зоны военных действий. Кроме того, помимо объектов, распространенной задачей является извлечение аспектов этих объектов – то есть параметров, по которым авторы текстов оценивают объекты в своих текстах. Это важно потому, что оценки одного и того же объекта по разным аспектам могут быть разными. Например, типичными аспектами отелей являются: местоположение, номерной фонд, питание, персонал и некоторые другие; однако аспекты других товарных категорий менее очевидны и требуют

извлечения, которое также обычно основано на методах машинного обучения. Извлечение объектов, аспектов и фактов (обзор см. Aggarwal & Zhai 2012, 11-42) – область, чрезвычайно развитая в коммерческом секторе, так как эти техники используются для оптимизации поисковых выдач и для маркетинговых исследований.

3. Российские исследования с применением автоматического анализа текстов

Перейдем к обзору российских исследований, использующих вышеописанные методы. Как уже говорилось, основные работы в области исследований интернета с использованием методов автоматического анализа текстов делаются в основном вне социальных наук. У этой группы работ нет единого «места» в научном сообществе. Часть работ докладывается на ежегодной конференции «Диалог» (Диалог 2015) и затем публикуется в ее сборнике «Компьютерная лингвистика и интеллектуальные технологии». Работы по интернет-данным получили здесь развитие благодаря многолетнему проекту РОМИП (РОМИП 2014)– Российскому семинару по оценке методов информационного поиска. Это некоммерческое партнерство десять лет проводило соревнования алгоритмов классификации, кластеризации, извлечения объектов, сентимент-анализа и извлечения мнений в рамках конференции «Диалог» на своих «дорожках» - тестовых коллекциях текстов, часть из которых извлекались из интернета. Это были как статичные сайты, так и новости и отзывы пользователей на товары. Еще одна существенная часть работ в этой области докладывается на конференциях Российской летней школы по информационному поиску (RuSSIR 2015) - здесь больше работ, связанных с интернетом, но они не публикуются (что, однако, позволяет увидеть здесь результаты коммерческих исследований). Часть работ можно встретить в трудах конференции АИСТ («Анализ изображений, сетей и текстов») (АИСТ 2015). Они носят в основном методологический характер, не представляя конечных эмпирических результатов, а предлагая способы их получения. Отдельные эмпирические работы разбросаны по разным изданиям и сообществам. В данном обзоре мы пытаемся сфокусироваться на этом последнем типе работ.

3.1. Классификация, кластеризация и тематическое моделирование в российских исследованиях

В группе методов классификации и кластеризации уже сейчас можно выделить некоторое количество ярких работ, сделанных математиками совместно с социальными исследователями. Так, Якушев и соавторы (2012) исследовали возможности моделирования наркотизации населения по данным социальных сетей. Для этого ими была использована база знаний, состоящая из словаря и правил, и на основе выраженности присутствия признаков из базы в текстах аккаунтов ВКонтакте, «хозяева» аккаунтов были классифицированы по степени вовлеченности в наркокультуру. Доли разных типов вовлеченности, определенные по материалам ВКонтакте, совпали с долями этих типов, смоделированными на оффлайновых данных. Статья не лишена ограничений: в ней не описана база знаний и нет никаких сведений об оффлайновой модели. Тем не менее, эта работа является наиболее продвинутой из ряда работ, использующих словарный подход; нередко социальные исследователи ограничиваются мониторингом

упоминаемости нескольких ключевых слов с использованием таких коммерческих систем мониторинга социальных медиа как IQBuzz или Wobot. Так, Биккулов (2013) исследует динамику обсуждения самоубийств в социальных сетях, изучая распределения ключевых слов во времени, по типам источников и пользователей на данных сервиса IQBuzz. Бершадская и Чугунов (2013) сходным образом, на основе того же сервиса анализируют динамику обсуждения в социальных сетях темы электронного правительства, а также проекта «Российская общественная инициатива».

Хотя метод классификации с машинным обучением широко используется для опознавания разных социально значимых видов текстов за рубежом, среди российских исследований удалось найти только исследования, направленные на решение коммерческих задач. Так, Гречников и соавторы (2009) тестируют алгоритм классификации TreeNet для задачи опознавания неестественных текстов (спама, сгенерированного ботами). Коршунов и соавторы (2013) применяют алгоритм классификации, известный как машина опорных векторов, для определения демографических характеристик пользователей Твиттера по текстам их записей. Метод применяется к шести европейским языкам, включая русский; он разбивает твиты пользователей на слова, биграммы и триграммы и приписывает пользователя к той демографической группе, для которой характерно наиболее сходное с данным пользователем сочетание этих признаков. Выборки составили по 500 пользователей (из них по 450 составили обучающую коллекцию) для каждого языка и для каждого исследуемого социально-демографического атрибута. Следует отметить, что это исследование может иметь применение не только в маркетинге, но и в социологии, где исследователям часто не достает демографических признаков изучаемых пользователей соцсетей. Метод демонстрирует большую точность, однако он не описан подробно, а используемые выборки довольно малы.

Тематическое моделирование было в последнее время использовано в нескольких социологических работах. Кольцова и Кольцов (Koltsova & Koltcov 2013) показали применимость тематического моделирования для исследования динамики тематики блогов на примере русскоязычного Живого Журнала, исследованного в «спокойный» и «предвыборный» периоды. Это первое в России социологическое исследование с применением тематического моделирования сравнивало две выборки постов популярных блоггеров, примерно по 25 тысяч постов, за август и декабрь 2011 года, и показало вытеснение темой выборов других социально-значимых тем во второй период по сравнению с первым, в то время как рекреационно-потребительские темы остались неизменными.

Алексеева и соавторы (2014) сравнивали тематику топовых (популярных) и случайных блоггеров Живого журнала и не обнаружили значимой разницы, которая заключалась лишь в количестве оставляемых постов и получаемых комментариев. Также, Кольцова и соавторы (Koltsova et al 2016) изучали принципы формирования сообществ совместного комментирования – групп пользователей, которые комментируют примерно одну и ту же группу постов. Анализ проводился на материале одной недели, включившей все посты топ-2000 блоггеров и все комментарии к ним. Авторы обнаружили, что лишь в некоторых сообществах посты близки лексически (а в большей степени они близки по авторству), а тематическое моделирование показало, что в сообществах, где комментируются

лексически близкие посты, можно вычленить, как правило, несколько доминирующих тем. Таким образом, нет убедительных свидетельств того, что люди завязывают совместные дискуссии о постах на основании их темы. Воскресенский и соавторы (Voskresenkiy et al 2015) изучили обсуждения в группах многоквартирных домов Петербурга во ВКонтакте и показали, что тематика открытых и закрытых групп отличается и связана также с членством их участников в других группах. В частности, открытые группы больше ассоциируются с членством в градозащитных группах и с правозащитной тематикой в противовес «бытовой»; тематическое сравнение открытых и закрытых групп - совершенно неизведанная область, способная многое сказать о влиянии настроек приватности на поведение пользователей. Горгадзе и Александров (Gorgadze & Alexandrov 2014) провели мэппинг групп «кавказцев» ВКонтакте с помощью сетевого анализа, куда вошло 887 активных групп, и тематического моделирования, куда 287 групп по трем южнокавказским этническим группам и псевдоэтнониму «кавказцы». Авторы выявили, что среди моноэтничных и полиэтничных групп последние являются посредниками; большинство групп не политизированы; религия и ислам, в частности, не являются центральной темой.

Все эти работы, однако, объединяет недостаток «наивного» использования алгоритма латентного размещения Дирихле, при котором его параметры (геперпараметы альфа и бета, количество тем, количество итераций) никак не подбираются и не оцениваются. Кроме того, алгоритм запускается один раз. Кольцов и соавторы (Koltsov et al 2014) показали, что даже при разных запусках с одними и теми же параметрами алгоритм дает радикально разные результаты, затрудняющие их социологическую интерпретацию, а о разнице результатов с разными параметрами остается только догадываться, так как эта тема практически не изучена. Николенко и соавторы (Nikolenko et al 2016) предлагают меру качества работы алгоритма и подтверждают ее успешность с помощью экспериментов на людях. Однако еще только предстоит разработать меры качества, оценивающие не отдельные темы, а целые решения, с разных точек зрения, и выработать рекомендации по подбору параметров для различных задач.

3.2. Сентимент-анализ и извлечение мнений в российских исследованиях

Сентимент-анализ широко представлен в российских работах, тестирующих новые методы анализа отзывов на книги, фильмы, компьютерные игры, мобильные телефоны, цифровые камеры - во многом благодаря конкурсу РОМИП. Эти работы не представляют никаких содержательных выводов по анализируемым товарным категориям, а те, которые представляют, крайне малочисленны. Например, Бойко (2014), предлагает не столько новый алгоритм анализа отзывов, сколько интегрирует уже имеющийся в схему конкретного маркетингового исследования проводит его на отзывах на банки. В качестве объектов используются четыре российских банка; в ходе исследования автоматически вычленяются аспекты для этих банков (персонал, кредит, депозит, карта, кассовое обслуживание) и рассчитываются рейтинги удовлетворенности каждым из аспектов в каждом из банков. Так, самые низкие оценки получают кредиты Альфа-банка, а самые высокие – депозиты его же и ВТБ24. Ценность этой работы – в попытке преодолеть разрыв между лингвистическими наработками в области методов и встраиванием их в программы конкретных исследований. Андреева и Никитина (2012) анализируют записи блогов об iPad, продукте Apple, однако их методика не совсем ясно описана. Вероятнее

всего, для анализа русскоязычной платформы F5 используется ручное кодирование сообщений, а при анализе англоязычных твитов применяется встроенный в Твиттер тональный анализатор. При неясности описания исследования выводы авторов представляются не совсем надежными.

Герданович и Герданович (2013) анализируют отношение к образованию в англоязычных твитах имеющих геотэги, с целью определить регионы, в которых отношение наиболее позитивное и которые, как предполагают авторы, таким образом могут быть наиболее вероятными потребителями белорусских образовательных услуг. Авторы классифицируют твиты на пять классов (от -2 до +2) с помощью готового метода, предложенного Брином (Breene), однако далее они затрудняются связать оценки с их причинами (объектами и аспектами). Построенные ими облака слов для положительных и отрицательных отзывов состоят в основном из общеупотребительных слов, что они ошибочно объясняют недостатками выборки (на самом деле в облаках представлены просто самые частотные для классов слова, а не слова с наибольшей дискриминационной силой). Тем не менее, интересной является попытка построения карт распределения сентимента.

Этлинг (Etling 2014) сравнивает отношение к украинским протестам 2013-2014 годов (вплоть до отставки Януковича) в англоязычных и русскоязычных интернет-источниках, выделяя среди последних российские и украинские. Данная работа демонстрирует возможности метода сентимент-анализа для исследования самых горячих политических тем. Автор использует данные и инструменты анализа компании Crimson Hexagon. Он показывает, что в целом англоязычные источники более негативны, чем русскоязычные, вопреки его ожиданиям, однако если выделить среди всех источников только блоги и социальные сети и разделить их на украинские, американо-британские и российские, то негативность будет нарастать от первых к последним. Следует отметить, что автор сначала характеризует выделяемые им классы текстов как поддерживающие и осуждающие протесты, затем как негативные и позитивные, не объясняя, как эти типы классов соотносятся и какие именно в итоге он выделяет. Кроме того, размеры размеченных обучающих коллекций (30-45 постов для каждой тональной категории на каждом языке) вызывают некоторые сомнения, несмотря на то, что автор ссылается на рекомендации Crimson Hexagon.

Следует отметить, что работы с применением сентимент-анализа для русскоязычного материала (а не предлагающие методы) – единичны, а работы с применением вычленения сущностей отсутствуют.

4. Заключение

Методы автоматического анализа текстов стремительно развиваются, и несмотря на то, что они еще слабо освоены в социальных науках, в будущем можно ожидать прорывов в социальных исследованиях именно на основе этих методов. Наиболее интересным в этой связи представляется сочетание разных методов. Так, после получения представления о тематической структуре корпуса текстов не менее важным оказывается вычленить сущности (людей, организации, события), характеризующих темы, аспекты этих сущностей, отношения авторов текстов в этим аспектам и их распределение во времени и пространстве. Коммерческие сервисы, предлагающие упрощенные варианты описанной системы аналитики, уже существуют, а когда такие системы разовьются в полную силу,

исследователи получат действительно мощные инструменты сбора и анализа текстов данных.

Использованные источники

1. Aggarwal C.C., Zhai C. A Survey of Text Classification Algorithms. In: C.Aggarwal. C. Zhai (eds) Mining Text Data, Springer 2012: 163-222.
2. Aggarwal C.C., Zhai C. A Survey of Text Clustering Algorithms. In: C.Aggarwal. C. Zhai (eds) Mining Text Data, Springer 2012: 77-128.
3. Aggarwal C.C., Zhai C. Information Extraction from Text. In: C.Aggarwal. C. Zhai (eds) Mining Text Data, Springer 2012: 11-42.
4. Blei D. Probabilistic topic models. *Communications of the ACM*, 55(4): 77-84, 2012.
5. Cluto 2007 http://glaros.dtc.umn.edu/gkhome/cluto/cluto/download (доступ осуществлен: 17.09.2015).
6. Etling, B. Russia, Ukraine, and the West: Social Media Sentiment in the Euromaidan Protests. The Berkman Center for Internet and Society at Harvard University, Research Publication No. 2014-13, September 25, 2014. http://cyber.law.harvard.edu/publications/2014/euromaidan (доступ осуществлен: 17.09.2015).
7. Gorgadze A., Alexandrov D. Virtual Transnational Movements in the Caucasus. Paper presented at *Social Media and Social Movements* conference, St.Petersburg, 2014. http://linisevents.hse.ru/data/2014/09/01/1313572271/Gorgadze%20Alexandrov%20SMSM%202014%20three%20pages.pdf (доступ осуществлен: 17.09.2015).
8. Koltsov S., Koltsova O., Nikolenko S. I. Latent Dirichlet Allocation: Stability and Applications to Studies of User-Generated content // *Proceedings of WebSci '14 ACM Web Science Conference, Bloomington, IN, USA — June 23 - 26, 2014*. NY : ACM, 2014. P. 161-165.
9. Koltsova O., Koltcov S., Nikolenko S. Communities of co-commenting in the Russian LiveJournal and their topical coherence. Internet Research, Vol 26, Iss. 3, 2016 (forthcoming).
10. Koltsova O., Koltcov S. Mapping the Public Agenda with Topic Modeling:The Case of the Russian LiveJournal // *Policy & Internet*. 2013. Vol. 5. No. 2. P. 207-227.
11. Lin W.-H., Hauptmann A. Are these documents written from different perspectives? A test of different perspectives based on statistical distribution divergence. In *Proceedings of the International Conference on Computational Linguistics (COLING)/Proceedings of the Association for Computational Linguistics (ACL)*, pages 1057–1064, Sydney, Australia, July 2006. Association for Computational Linguistics.
12. Mallet 2002. http://mallet.cs.umass.edu/ (доступ осуществлен: 17.09.2015).
13. Mystem 2015 https://tech.yandex.ru/mystem/ (доступ осуществлен: 17.09.2015).
14. Nikolenko S., Koltsov S., Koltsova O. Topic modelling for qualitative studies 2016 (forthcoming).
15. Pang, B., Lee, L. Opinion mining and sentiment analysis // Foundations and Trends in Information Retrieval. 2008, Vol. 2, No 1-2. P. 1–135.
16. RuSSIR 2015 http://romip.ru/russir2015/ (доступ осуществлен: 17.09.2015).
17. SentiStrength 2012. http://sentistrength.wlv.ac.uk/ (доступ осуществлен: 17.09.2015).
18. Steyvers M., Griffiths T. Probabilistic topic models. In: T. Landauer, D McNamara, S. Dennis, and W. Kintsch (eds), Latent Semantic Analysis: A Road to Meaning. Laurence Erlbaum, 2007.

19. Taboada, M., Brooke, J., Tofiloski, M., Stede, M. Lexicon-Based Methods for Sentiment Analysis // Computational Lingustics. 2011, Vol. 37, No. 2. P. 267-307.
20. TopicMiner 2013 http://linis.hse.ru/en/soft-linis (доступ осуществлен: 17.09.2015).
21. Voskresenskiy V., Sukharev K., Musabirov I., Alexandrov D. Online Communication in Apartment Buildings, in: *Lecture Notes in Computer Science* Vol. 8852: SocInfo 2014 International Workshops, Barcelona, Spain, November 11, 2014, Revised Selected Papers. Springer International Publishing, 2015.
22. АИСТ 2015 http://aistconf.org/ (доступ осуществлен: 17.09.2015).
23. Алексеева С., Кольцова О., Кольцов С. Общественное мнение онлайн: сравнение структуры и тематики постов «обычных» и «популярных» блогеров Живого Журнала // *Supplementary Proceedings of the 3rd International Conference on Analysis of Images, Social Networks and Texts (AIST 2014)*/ Ed. by D. I. Ignatov, M. Y. Khachay, A. Panchenko, N. Konstantinova, R. Yavorsky, D. Ustalov. Vol. 1197: Supplementary Proceedings of AIST 2014. CEUR-WS.org, 2014. С. 177-181.
24. Андреева А.Н., Никитина М.С. Сентимент-анализ брендов в Российской блогосфере как инструмент маркетинговых исследований // Бренд-менеджмент, 04 (65) 2012.
25. Бершадская Л.А., Чугунов А.В. Востребованность услуг электронного правительства: анализ дискуссий в социальных сетях // Интернет и современное общество: сборник научных статей. Труды XVI Всероссийской объединенной конференции «Интернет и современное общество» (IMS-2013), Санкт-Петербург, 9 – 11 октября 2013 г. – СПб: НИУ ИТМО, 2013. С. 67 – 71.
26. Биккулов А.С. Подростковые самоубийства в обсуждениях блогосферы // Интернет и современное общество: сборник научных статей. Труды XVI Всероссийской объединенной конференции «Интернет и современное общество» (IMS-2013), Санкт-Петербург, 9 – 11 октября 2013 г. – СПб.: НИУ ИТМО, 2013. С. 72 – 76.
27. Бойко М.В. Исследование удовлетворенности потребителей в банковской сфере на основе анализа текстовых отзывов // Вестник УГАТУ. 2014. Т. 18, № 5 (66). С. 139–145.
28. Воронцов К.В. Лекции по метрическим алгоритмам классификации. 2008. http://www.ccas.ru/voron/download/MetricAlgs.pdf (доступ осуществлен: 17.09.2015).
29. Гедранович, Б.А. Гедранович А.Б. Отношение к высшему образованию: сентимент-анализ данных микроблогов // Инновационные образовательные технологии. — 2013. — № 1 (33). — С. 46—54.
30. Гречников Е.А., Гусев Г.Г., Кустарев А.А., Райгородский А.М. Поиск неестественных текстов. Труды 11й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.
31. Диалог 2015 http://www.dialog-21.ru/ (доступ осуществлен: 17.09.2015).
32. Коршунов А., Белобородов И., Гомзин А., Чуприна К., Астраханцев Н., Недумов Я., Турдаков Д. Определение демографических атрибутов пользователей микроблогов // Труды Института системного программирования РАН, Том 25, 2013, С. 179-194.
33. РОМИП 2014 http://romip.ru/ (доступ осуществлен: 17.09.2015).
34. Томита-парсер 2015 https://tech.yandex.ru/tomita/ (доступ осуществлен: 17.09.2015).
35. Якушев А.В., Митягин С.А., Бухановский А.В. Имитационное моделирование наркотизации населения по данным мониторинга социальных сетей // Современные исследования социальных проблем (электронный научный журнал), № 2 (10) – 2012 – С 133-151.