Плетнева Анастасия Дмитриевна

Санкт-Петербургский государственный университет (СПбГУ),
Санкт-Петербург, Россия
adpletneva@perm.ru
Научный руководитель – М. В. Хохлова, канд. филол. наук

ИССЛЕДОВАНИЕ ТОНАЛЬНОСТИ В ЗАДАЧЕ АВТОМАТИЧЕСКОГО ОПРЕДЕЛЕНИЯ ТИПА АВТОРА

Ключевые слова: анализ тональности, корпусы текстов, блоги, язык веба, авторское профилирование.

В статье описывается эксперимент по изучению тональности текстов интернет-блогов стажеров, находящихся на международных волонтерских стажировках. Был составлен корпус текстов блогов, создан частотный список слов и проведено исследование тональности слов и предложений текста, а также эмотиконов и других средств выражения эмоциональной оценки авторов по отношению к стажировкам.

Pletneva Anastasiia

Saint Petersburg State University (SPbSU),
St. Petersburg, Russia

THE STUDY OF SENTIMENTS APPLIED TO THE AUTOMATIC DETECTION OF AN AUTHOR'S TYPE

Keywords: sentiment analysis, text corpora, blogs, web language, author profiling.

The paper describes an experiment on studying the sentiments of the weblogs written by the volunteer interns. A corpus of texts has been compiled and a frequency list of lexemes has been created. The paper also provides an analysis of the sentiment of words and sentences as well as emoticons and other means of an expressive attitude presented in the texts.

В настоящее время все больше людей используют социальные сети для выражения своего мнения и своих эмоций для написания текстов, называемых блогами. При этом возникает такое удивительное явление, как язык Интернета, который обнаруживает характеристики разговорного языка и новые явления, которые отсутствовали ранее при письменном общении: хэштеги, смайлики, ненормативная пунктуация и др.

Отмечается воздействие компьютеров и глобальной сети на русский язык с двух сторон: во-первых, происходит одновременное усложнение одних и упрощение других средств сравнительно с аналогичными в русском языке, не подвергшимися воздействию глобальной сети, а во-вторых, видна конкуренция норм письменного и устного языков. В целом же, можно констатировать тот факт, что язык Интернета пока остается недостаточно изученным в современной лингвистике из-за своего активного развития.

В качестве материала для исследования были рассмотрены тексты блогов стажеров, которые участвовали в программах международных обменов от организации AIESEC. Стажеры выбирали волонтерскую программу, которая длилась 6–8 недель, по одному из семнадцати направлений, соответствующих целям устойчивого развития ООН. Авторы, которые вели записи о волонтерских стажировках, в большинстве своем являлись студентами 2–3 курсов бакалавриата, возраст составлял от 19 до 23 лет, обычно у них отсутствовал профессиональный опыт. Цель исследования состоит в изучении и сравнении тональности текстов интернет-блогов, что может быть востребовано при автоматической обработке блогов и определении типа автора, написавшего блог.

Анализ тональности (или сентимент-анализ) связан с эмоциональной окраской текста. Он подразумевает выявление отношения автора к определенной теме или предмету. Автоматическое определение тональности текста подразумевает под собой определение определенных фрагментов текста, несущих в себе позитивную или негативную оценку по отношению к какому-либо объекту. При этом одной из главных задач при обработке и анализе естественного языка сегодня является определение и верификация авторства текстов.

Авторство основывается на классификации текстов на основе лингвистических характеристик авторов. Кроме самого определения авторства, при котором учитывается стиль авторов, изучается социальный аспект, т. е. то, как используется язык. Это помогает при идентификации аспектов профилирования, таких как возраст, пол или образование. В данный момент фокус в задаче авторского профилирования все больше смещается на анализ блогов пользователей социальных сетей, использующих язык Интернета, и на определение того, как он отражает личностные характеристики человека. Однако все чаще информация, которую определенный человек указывает о себе, оказывается неверной, например, пол или возраст. Именно поэтому очень важно описать демографический и психологический портрет пользователя на основе их текстов.

Не существует единого мнения, какой же набор характеристик с наибольшей точностью указывает на авторство. Что касается английского языка, в работах применялись различные критерии. Например, некоторые исследователи [Koppel, Argamon, Shimoni, 2003, p. 401–412] связывали использование языка с такой демографической характеристикой как пол человека. Изучалось также влияние возраста и гендерной принадлежности на стиль текстов блогов [Schler et al., 2006, p. 199–205.]. Ученые доказали, что языковые особенности авторов блогов коррелируют с возрастом, что отражается, например, в использовании предлогов и детерминантов.

Для русского языка задача автоматического определения типа автора остается актуальной, так как было проведено немного исследований, направленных на авторское профилирование русскоязычных текстов. На материале русского языка впервые были использованы методы распознавания образов в задаче атрибуции анонимных текстов с учетом индивидуальных особенностей авторов [Марусенко, 1990, с. 164]. Данный способ авторского профилирования дал высокие результаты при обработке историко-литературных текстов [Синелева, 2001, с. 22]. В большинстве случаев метод позволяет четко классифицировать тексты в зависимости от стилистических характеристик произведений.

Наше исследование посвящено анализу корпусов текстов блогов волонтерских стажировок. Был создан корпус, который состоит из 61 текста различных авторов общим объемом около 370#000 словоупотреблений. Тексты были написаны авторами, которые в 2014–2017 годах побывали на волонтерских стажировках в 12 разных странах. Таким образом, задачами текущего исследования являются:

1. Построение частотного списка слов волонтерских стажировок.
2. Сравнение частотного списка слов по волонтерским стажировкам по эмоциональной окраске с тональным словарем краудсорсингового веб-ресурса Linis Crowd.
3. Определение тональности отдельных предложений с помощью программы SentiStrength.
4. Исследование эмотиконов и других средств выражения экспрессивности.

Результаты сравнения частотного списка слов с тональным словарем Linis Crowd показывают, что слов с позитивной окраской наибольшее количество (59%), что согласуется с первоначальной гипотезой об их превалировании в текстах. Слов с маргинальными оцен-

ками (–2 и 2) обнаружено примерно одинаковое количество (4%), но слов с оценкой 1 больше, чем слов с оценкой –1 (55% vs. 37%).

Чтобы проверить результаты, которые показывают, что положительно окрашенных слов больше, чем отрицательно, и, следовательно, общая тональность текста скорее позитивная, чем негативная, с помощью программы SentiStrength была проанализирована тональность предложений текста.

Результаты работы программы показывают, что в текстах преобладают нейтральные предложения (56%). Больший интерес представляет то, что количество положительных предложений более чем в два раза превышает количество отрицательных предложений, что подтверждает гипотезу о положительной окрашенности текстов блогов в целом. Ниже в качестве примера приведены предложения с 1) положительной ([4;–1]) и 2) отрицательной ([1;–3]) оценками (авторские орфография и пунктуация сохранены во всех примерах):

1) *В итоге, я очень рада, что она у меня наконец то есть!*

2) *Пока мы ехали в школу пошел сильный дождь и это, я скажу вам, не шутки.*

Оставшиеся средства выражения экспрессивности были разделены на две части: выражающие положительные и отрицательные эмоции. Рассмотрены следующие средства: эмотиконы, смайлики, удлинения слов, ненормативная пунктуация, использование слов и выражений на других языках, кроме русского, слова, написанные буквами верхнего регистра.

Эмотикон — это пиктограмма, которая служит для выражения эмоции в языке Интернета, выражающаяся в сочетании типографических знаков (например, *:)*, *;)* и др.).

В ходе исследования было выяснено, что эмотиконов, отражающих положительные эмоции в 4 раза больше тех, что выражают отрицательные:

Я накупила столько украшений, еды и прочего:))))))))))))))))))))))))))))))))

При этом отметим, что количество первых в тексте велико — 44%, а вместе они покрывают 57% из общего числа средств, связанных с эмоциональностью текста.

В отличие от эмотиконов, смайлик — это графическое изображение человеческого лица, отображающего определенную эмоцию. Также как и в случае с эмотиконами, они делятся на три группы: передающие положительные, отрицательные эмоции и нейтральные смайлики (т. е. те, которые не выражают определенной эмоции).

Количество смайликов в текстах невелико (всего 5.5%), тогда как при первичной обработке данных нами было высказано предположение, что именно они играют большую роль при передаче различных эмоций в тексте. В примере ниже смайлики заменены специальными тэгами:

[*EMOJI*] [*EMOJI*]*Откуда такая возможность*[*EMOJI*] [*EMOJI*]

Оставшиеся средства выражения эмоций — удлинения слов, пунктуация, использование иностранных слов и слов, написанных верхним регистром, — было решено рассматривать вместе.

Наибольший процент здесь занимает использование ненормативной пунктуации (15%). Стоит отметить, что при этом внутренняя организация предложения и, соответственно, пунктуация внутри него не нарушается, несмотря на ненормативность в конце предложений:

И тут такая новость в группе #aiesec_msc (*скриншот прикрепляю*), *от которой все мысли сбежались в кучу……ГОА!!!!……МЕРОПРИЯТИЕ !!!!! …..МЕЖДУНАРОДНЫЙ ОПЫТ РАБОТЫ!!!!!……ТАНЦЫ!!!!….это же правда #стажировкамечты !!!!*

Далее идет использование иностранных слов (12%), которые могут использоваться как частично в структуре предложения (заменять определенные русские слова и выражения), так и все предложение может быть написано на иностранном языке. Данное явление объясняется либо написанием целых кусков текста на не русском (чаще всего, английском) языке, либо языковыми лакунами в русском языке, вследствие чего авторы используют иностранные слова для их покрытия:

Вчера мы были на Средиземном море! Finally!

Верхний регистр и удлинения слов также нельзя причислить к наиболее используемым средствам выражения эмоций в тексте — их всего 2 и 5% соответственно:

В начале нам предстояло подняться на гору, затем спуститься с нее к пляжу, это было ооооооооооочень ОЧЕНЬ тяжело!

Таким образом, были выявлено, что тональность, передающаяся различными средствами выражения экспрессивности, в частности смайликами, эмотиконами, эмоционально окрашенными словами, может использоваться при автоматических методах определения типа авторов текстов. Мы планируем продолжить исследование в данном направлении, используя дополненный ряд лингвистиче-

ских характеристик, например, типичные синтаксические структуры словосочетаний и предложений, морфологические и семантические признаки, а также в плане создания отдельных программ для определения авторства с применением описанных признаков.

ЛИТЕРАТУРА

Марусенко, 1990 — *Марусенко М. А.* Атрибуция анонимных и псевдонимных литературных произведений методами распознавания образов. Л., 1990. С. 164.

Синелева, 2001 — *Синелева А. В.* Атрибуция «Романа с кокаином»: лингвостатистическое исследование. СПб., 2001. С. 22.

Koppel et al. 2003 — *Koppel M., Argamon S., Shimoni A.* Automatically categorizing written texts by author gender // Literary and Linguistic Computing. Vol. 17 (2003). P. 401–412.

Schler et al., 2006 — *Schler J., Koppel M., Argamon S., Pennebaker J.* Effects of age and gender on blogging // In AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs. AAAI, 2006. P. 199–205.