УДК 004.02

АНАЛИЗ ЭМОТИВНОСТИ ЦИФРОВЫХ ТЕКСТОВ

Н.С. Куличок
Магистрант кафедры ПОИТ БГУИР

А.И. Парамонов
Доцент кафедры программного обеспечения информационных технологий факультета компьютерных систем и сетей БГУИР, кандидат технических наук, доцент

Белорусский государственный университет информатики и радиоэлектроники,
Республика Беларусь.

Н. С. Куличок
Окончил Белорусский государственный университет информатики и радиоэлектроники. Магистрант кафедры программного обеспечения информационных технологий факультета компьютерных систем и сетей БГУИР.

А. И. Парамонов
Доцент кафедры программного обеспечения информационных технологий факультета компьютерных систем и сетей БГУИР, кандидат технических наук, доцент.

Аннотация. Работа посвящена проблеме анализа цифровых текстов на естественном языке с целью выявления их эмотивной характеристики. В статье дается понятие эмотивности и ее составляющих. Рассмотрены существующие методы анализа тональности текстов как основы эмотивного фона текста. Обозначены дальнейшие шаги решения задачи.

Ключевые слова: эмоции в тексте, анализ тональности, машинное обучение, эмотивность.

Введение.

Постоянный рост объемов текстовой информации в цифровом виде приводит к возникновению потребности в новых инструментах для ее анализа. На сегодня достаточно эффективно решаются задачи классификации текстов и автоматического синтаксического анализа. Известны подходы по семантическому анализу текстов. На современном этапе развития автоматической обработки текстов актуальной и востребованной становится задача эмотивной оценки текстов. Анализ эмотивности является достаточно сложной задачей [1], что связано с трудностями при выделении нужной эмоциональной лексики в текстах, а также с определением самого эмотивного пространства, количества и состава его измерений.

Ежедневно множество данных генерируется пользователями соцсетей, новостными порталами, различными блогами. Весь этот контент несет в себе огромное количество информации, которую можно и даже нужно использовать на эмоциональную окраску. Анализ эмотивности требуется для мониторинговых, аналитических и сигнальных систем, для систем документооборота и рекламных платформ, таргетированных по тематике веб-страниц и многих других. Область применения анализа эмотивности текстов обширна. Эмотивность рассматривается как свойство языка выражать психологические (эмоциональные) состояния и переживания автора текста. Эмотивность это отображение термина «эмоциональность» на категории естественного языка. Эмотивность – это выражение эмоциональности на лингвистическом уровне с помощью разнообразных языковых и речевых средств, которые представлены на каждом уровне любого естественного языка и образуют его эмотивный код. Каждый отдельный текст имеет свой эмотивный код. Эмотивный код

текста рассматривается как система сигналов эмотивности текста, отражающих общее настроение документа и эмоциональное отношение автора к описываемым в тексте действиям и объектам.Существует множество характеристик эмотивности, которые можно анализировать, в числе которых: эмотивный фон, эмотивная тональность, эмотивная окраска, эмотивная направленность, эмотивная модальность, эмотивные интенции текста [2].

Анализ тональности как основа эмотивного содержания текста.

Наибольшего успеха исследования эмотивности текста достигли при анализе его тональности Анализ тональности можно рассматривать как метод количественного описания качественных данных, с присвоением оценок настроения. Целью является нахождение мнений в тексте по отношению к объектам, речь о которых идёт в тексте, и определение их свойств [3]. Анализ тональности играет важную роль в принятии решений и в системе рекомендаций. Однако современные большие объёмы цифровой информации не позволяют человеку оценивать весь поток текстов самостоятельно. Анализ тональности упрощает эту задачу, поскольку описывает полярность текста, так что пользователь может напрямую узнать, является ли данный текст (или набор документов) положительным или отрицательным, не просматривая его. Большинство современных систем используют бинарную оценку – «положительный сентимент» или «отрицательный сентимент», однако некоторые системы позволяют выделять силу тональности.

Для определения настроения в анализе тональности используются три термина: объект, о котором дается мнение, особенности этого объекта, а также владелец мнения об объекте [4]. Анализ тональности выполняет задачу классификации в три этапа: на уровне документа, уровне предложения и уровне характеристик. Уровень классификации документа используется там, где задача состоит в том, чтобы найти общую полярность темы независимо от того, кто придерживается мнения. Классификация по уровням предложений предполагает, что каждое предложение придерживается единого мнения. На уровне характеристик (или аспектов) выполняется анализ различных характеристик объекта. Анализ тональности включает в себя предварительную обработку данных, выбор характеристик и классификацию, а затем определение полярности данных. Предварительная обработка данных включает в себя токенизацию, удаление стоп-слов, выделение корней, лемматизацию и прочие процедуры. Токенизация – задача разбиения последовательности слов на отдельные слова, называемые токенами. Стоп-слова – слова, которые не придерживаются какого-либо мнения, поэтому их полезно удалить (например, is, am, are, in, to и т. п.). Стемминг – задача преобразования альтернативных форм слова в его базовую форму (например, helping в help).

Методы анализа тональности делятся на 2 большие группы: основанные на машинном обучении и основанные на правилах и словарях.В методах машинного обучения используется маркированный набор данных, где полярность предложения уже определена. Из этого набора данных извлекается признак, который помогает классифицировать полярность неизвестного входного предложения. В свою очередь методы машинного обучения разделены на обучение с учителем и обучение без учителя. Машинное обучение с учителем используется в случае, когда для обучения модели доступны помеченные данные [5]. Реализация этого метода предполагает выполнение двух этапов: первый шаг – обучение модели, второй – решение задачи прогнозирования. Во время обучения набор данных с его метками подается в алгоритм классификации, который дает модель в качестве выходных данных. После этого тестовые данные вводятся в модель для прогнозирования категории. Сегодня известны различные алгоритмы классификации для машинного обучения с учителем: наивный байесовский классификатор, байесовская сеть, метод опорных векторов, нейронные сети и другие. Метод машинного обучения без учителя используется, когда сбор помеченных данных затруднён. Зачастую собрать немаркированные данные легче, чем маркированные. При таком подходе предложения и документы классифицируются на основе списков ключевых слов каждой категории, которые также должны быть заранее подготовлены [6]. Методы на основе правил и словарей включают две соответствующие подгруппы: методы на основе правил и методы на основе словарей. В основе методов на основе правил лежит идея, что система состоит из набора правил, применяя которые делается заключение о тональности текста. В данном случае для достижения хорошего результата необходимо составить большое количество правил. Зачастую правила привязаны к определенной предметной области и при её изменении требуется

заново составлять правила. Тем не менее, этот подход является наиболее точным при наличии достаточно полной базы правил. Методы на основе словарей используют так называемые тональные словари, которые представляют из себя списки слов с ранжированием (с указанным для каждого слова значением его тональности). Для достижения хорошего результата нужно вычислить значения двух оценок: положительной составляющей текста и отрицательной. Положительная составляющая текста вычисляется как сумма тональностей всех положительных терминов, которые присутствуют в тексте. Аналогичным образом рассчитывается значение отрицательной составляющей текста. Итоговая оценка тональности всего исследуемого текста рассчитывается как отношение этих двух составляющих. Естественно, точность результатов напрямую зависит от размера и «качества» словаря (корректности указанных значений тональности терминов). Следует отметить, что для текстов на русском языке на сегодня практически не существует экспертно-размеченных словарей тональности. Построение подобных словарей целая отдельная научная задача.

Заключение.

Устранение обозначенных вопросов существующих методов анализа тональности позволит эффективнее обрабатывать большие объёмы данных, а также более точно определять эмотивность текстов не только в одномерных эмотивных пространствах типа позитив-негатив, но и в куда более сложных размерностях. Основные пути решения задачи связаны с комбинацией существующих подходов и формализацией методик психолингвистики для последующего построения модели эмотивной составляющей текстов на естественном языке.

Список литературы

[1] Анна Пазельская, Алексей Соловьев. Метод определения эмоций в текстах на русском языке // The international conference on computational linguistics and intellectual technologies «Dialogue 2011» : конференция. – Москва, 2011. – С. 510–522.

[2] Ленько Галина Николаевна Уровни анализа текстовой эмотивности (на примере текстов художественного стиля) // Вестник ЛГУ им. А.С. Пушкина. 2014. №2. – 10 с.

[3] Гербик, А. И. Анализ тональности текста / А. И. Гербик, П. Е. Дорошкевич, А. И. Свито // Компьютерные системы и сети: материалы 53-й научной конференции аспирантов, магистрантов и студентов (Минск, 2–6 мая 2017 г.). – Минск: БГУИР, 2017. – С. 167 – 168.

[4] Алгоритмы анализа тональности текста / Н. С. Иванин и другие // BIG DATA and Advanced Analytics: collection of materials of the third international scientific and practical conference. (Minsk, Belarus, May 3 – 4, 2017) / editorial board : M. Batura [etc.]. – Minsk : BSUIR, 2017. – С. 150 - 154.

[5] Гербик, А. И. Определение тональности текста методом обучения с учителем / А. И. Гербик, Е. А. Макович, М. В. Аксамит // Технические средства защиты информации : тезисы докладов XV Белорусско-российской науч.-техн. конф. (Минск, 6 июня 2017 г.). – Минск : БГУИР, 2017. – С. 30.

[6] Гербик, А. И. Определение тональности текста методом обучения без учителя / А. И. Гербик, М. В. Аксамит, Е. А. Макович // Технические средства защиты информации : тезисы докладов XV Белорусско-российской науч.-техн. конф. (Минск, 6 июня 2017 г.). – Минск : БГУИР, 2017. – С. 29.

ANALYSIS OF DIGITAL TEXTS EMOTIVENESS

N.S. KULICHOK
Master's degree Student at Software for Information Technologies Department of BSUIR

A.I. PARAMONOV,
PhD (Candidate of Computer Sciences), Associate Professor at Software for Information Technologies Department of BSUIR, Associate Professor

Belarusian State University Of Informatics And Radioelectronics, Republic of Belarus
E-mail: nikita.kulicok@gmail.com

Abstract. The work deals with the problem of analyzing digital texts in natural language in order to identify their emotive characteristics. The article gives the concept of emotiveness and its components. The existing methods of analyzing the sentiment of texts as the basis of the emotive background of the text are considered. Further steps for solving the problem are indicated.

Keywords: emotive text, sentiment analysis, machine learning, emotiveness