ИССЛЕДОВАНИЕ ВОЗМОЖНОСТИ АНАЛИЗА ТОНАЛЬНОСТИ ТЕКСТА МЕТОДОМ АВТОМАТИЧЕСКОГО ТЕМАТИЧЕСКОГО КЛАССИФИЦИРОВАНИЯ ТЕКСТА НА ОСНОВЕ АЛГОРИТМОВ СТАТИСТИЧЕСКОГО АНАЛИЗА

Д.А. Горбушин, аспирант
Научный руководитель – Д.В. Гринченков, к.т.н, доцент
Южно-Российский государственный политехнический университет (НПИ)
имени М.И. Платова,
г. Новочеркасск
e-mail: gorinwww@gmail.com

Анализ тональности текста - класс математических методов выявления и изучения эмоциональной составляющей текста. Актуальность исследований на данную тематику обуславливается широким диапазоном применения алгоритмов анализа больших объемов неструктурированной текстовой информации в сети Интернет, позволяющих выявлять отношение потребителей к продукции, услугам, организациям и т.д. [1].

В данном исследовании была рассмотрена задача оценки возможности определения тональности текста методом автоматического тематического классифицирования текста на основе алгоритмов статистического анализа и морфологического анализа полученных результатов. Выбранные методы классификации имеют возможность работать с текстами большого объема, являются языконезависимыми и не используют правил построения текста, для описания которых требуется участие лингвистов [2].

Исследование проводилось в два этапа: на первом этапе тексты тематически классифицировались с помощью алгоритма латентно-семантического анализа; на втором был проведен морфологический анализ полученных результатов. Тематическая классификация текстов выполняется в несколько шагов:

1. Нормализация слов (термов) исходной коллекции текстов стеммером Портера. Выбор данного метода нормализации обусловлен сравнительным анализом существующих методов [3].
2. Составление частотной матрицы – матрицы терм-на-документы, каждым элементом которой является число повторений терма в каждом из документов.
3. Расчет весов термов и документов методом сингулярного разложения частотной матрицы (SVD) [4] и составление диаграммы семантического пространства коллекции текстов по тематическому признаку.

На втором этапе проводится морфологический анализ термов полученного семантического пространства и производится общая оценка тональности текстов.

В результате проведенного исследования была оценена возможность применения методов тематической классификации текстов на основе алгоритмов статистического анализа в качестве предварительного этапа анализа тональности текста. На эталонной тестовой коллекции метод показал 100% точность, на реальных текстах точность снижается вплоть до 33%, что обусловлено следующими факторами:

1. В исследовании рассматривалась одномерное эмотивное пространство – позитив/негатив, однако реальные коллекции состоят из нейтральных предложений более чем на 50%.
2. Необходим высокоточный алгоритм нормализации и морфологического анализа структурных единиц языка.

3. Определение тональности текстов возможно при условии появления в семантическом пространстве слов, характеризующих тональность текстов (в первую очередь прилагательные и связанные с ними наречия). Данная особенность сильно проявляется при работе с короткими или малосодержательными текстами и обусловлена главным принципом статистических алгоритмов тематической классификации - алгоритмы основаны на частотном анализе встречаемости слов в тексте и рассматривают текст как “мешок слов”, т.е. последовательность слов и их связи друг с другом не учитываются [4].

Таким образом, применение методов тематической классификации текстов на основе алгоритмов статистического анализа для решения задачи анализа тональности текста возможно при работе с объемными текстами качественного содержания. Низкая точность и языковая привязанность современных программных средств нормализации слов сводит к минимуму главную особенность представленного алгоритма - его языконезависимость, а также существенно снижает общую точность работы.

Список литературы

1. Д.А. Горбушин. Анализ методов автоматической классификации тональности текста // Научно-техническая конференция и выставка инновационных проектов, выполненных вузами и научными организациями ЮФО в рамках участия в реализации федеральных целевых программ и внепрограммных мероприятий, заказчиком которых является Минобрнауки России. – Новочеркасск : Лик, 2014. - С. 123-125.

2. Д.А. Горбушин, Д.В. Гринченков. Технологии машинного определения тематики текста как средство повышения эффективности информационного поиска // Проблемы баланса фундаментальности и профессиональной направленности физико-математической подготовки инженерных кадров : сб. науч. ст. по проблемам высшей школы. - Новочеркасск : ЮРГПУ, 2013. - С. 192-195.

3. Д.В. Гринченков, Д.А. Горбушин. Особенности применения алгоритмов стемминга при анализе тональности текста // Теория, методы проектирования, программно-техническая платформа корпоративных информационных систем. - Новочеркасск : ЮРГПУ, 2014. - С. 22-24.

4. Susan T. Dumais. Latent Semantic Analysis // Annual Review of Information Science and Technology. – 38: 2005. – С. 188.

5. Д.В. Машечкин, И.В. Петровский, М.И. Царёв. Методы вычисления релевантности фрагментов текста на основе тематических моделей в задаче автоматического аннотирования // Вычислительные методы и программирование, Т.14. – 2013. – С. 91–102.