УДК 004.738

ПОСТРОЕНИЕ СИСТЕМЫ СЕНТИМЕНТНОГО АНАЛИЗА СООБЩЕНИЙ СОЦИАЛЬНЫХ СЕТЕЙ

О.О. Воронова; М.М. Соколов; С.А. Петров
Сумской государственный университет
e-mail: lvoronova1@gmail.com

Увеличение роли электронной коммуникации в современном социуме, открывает новое направление в исследованиях, таких как определение эмоциональной окраски текста и его классификация в зависимости от отношения субъекта к предмету оценки. Таким образом, становиться возможным разрабатывать информационные системы способные узнавать мнения различных социальных групп что может быть использовано в рекламе, о товарах, а также оценивать вероятность ожидаемых событий и соответствующее им изменение настроений в группах.

Классические проблемы подобного анализа связаны с определением релевантности текста по отношению к поставленому вопросу и выделением блоков предложений имеющих различную смысловую нагрузку и оценку по конкретным аспектам исследуемого вопроса [1].

Примером таких блоков могут послужить различные цитаты приводимые автором, а также блоки текста дающие оценку этим цитатам. Таким образом, необходимо отделить мнение выраженное в цитате от мнения автора текста. Еще одной важной проблемой является распознавание типичных словесных конструкций

отрицания / усиления / ослабления мнения [2]. Также, выделим проблему итоговой оценки результатов и определения критериев классификации текстов.

Одной из базовых идей сентиментного анализа является выявление эмоционально окрашенных слов в тексте. Данные ключевые слова, предварительно классифицированные, могут количественно отражать общий настрой текста.

В наиболее примитивной модели рассмотрим набор слов/словосочетаний разделенных на 2 группы положительного и отрицательного окраса соответственно. В упрощенной модели различные слова в одной группе имеют одинаковый вес для оценки. Дополним ее введя вес для каждого слова - величину из отрезка от -1,0 до 1,0. [-1.0; 0) – для негативно окрашенных слов и (0; 1.0] – для положительно. 0 ставится в соответствие словам ни разу не встретившихся в тексте.

В данной работе воспользуемся словарем составленным из 18,500 «взвешенных» слов. Теперь подсчитаем оценку текста по формуле:

$$C = \frac{\sum_i^n W_{ai}}{n} + \frac{\sum_j^m W_{bj}}{m},$$

где W_{ai} – вес слова из группы A,

W_{bj} – вес слова из группы B,

n – количество слов присутствующих в тексте из группы A,

m – количество слов присутствующих в тексте из группы B;

Чем выше полученная оценка – тем “положительнее” оценивается текст.

Данный алгоритм не учитывает сложные синтаксические структуры языка. Характерным для него будет большой процент ложно положительных оценок, в то время как ложно отрицательные будут встречаться значительно реже. Улучшить его можно с помощью применения семантических правил оценки частей предложения. Необходимо сформировать набор таких правил, которые учитывают сложные завуалированые конструкции языка. Один из наиболее прогрессивных на данный момент алгоритмов основывается на рекурсивном построение взвешенного дерева, отражающего структуру предложения [3].

Представленный в данной работе метод может использоваться для анализа коротких текстов, ограниченных как по количеству символов так и несложностью структуры предложений (сообщения в социальных сетях).

1. Pang B., Lee L. Opinion mining and sentiment analysis, Foundations and Trends in Information Retrieval 2(1-2), 2008, pp. 1–135.

2. Turney P. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews / P. Turney // Proceedings of the Association for Computational Linguistics (ACL). – 2002. – P. 417-424.

3. Socher R. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank / R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, C. Potts // Proceedings of EMNLP Conference. – 2013.