**УДК 004.9**

# Использование эмотиконов и эмодзи в методах анализа тональности текстов

И.А. Лизунова
(Университет ИТМО, г. Санкт-Петербург)
Научный руководитель – к. т. н., доцент, О.В. Махныткина
(Университет ИТМО, г. Санкт-Петербург)

Исследования выполнены за счет стартового финансирования университета ИТМО в рамках НИР № 618278 «Синтез эмоциональной речи на основе генеративных состязательных сетей».

В настоящее время в социальных медиа получило широкое распространение использование специальных пиктограмм и идеограмм, называемых эмотиконы и эмодзи, для выражения различных эмоций, чувств, настроений. В связи с этим в области анализа тональности текстов большой интерес представляет исследование и разработка алгоритмов, способных обрабатывать и интерпретировать текстовую информацию, содержащую эмотиконы и эмодзи. Целью данной работы является обзор существующих методов анализа тональности текстов, использующих данные средства выражения эмоций.

Эмотикон – пиктограмма, изображающая эмоцию, составленная из типографских знаков. В отличии от эмотикона, эмодзи – это идеограмма, представляющая собой не набор типографских символов, а готовое изображение с определенным смыслом. Однако некоторые эмодзи, изображающие лица, относят к эмотиконам.

Одним из наиболее распространенных методов анализа тональности текста, рассматривающих эмотиконы, является подход, основанный на создании словаря эмотиконов. Так в работе [1] создается словарь наиболее часто употребляемых эмотиконов, каждому из которых сопоставляется наименование соответствующей ему эмоции и бинарная оценка, обозначающая положительную или отрицательную эмоциональную окраску. Аналогичный метод применяется и в работе [2], где словарь создается на основе восьми различных общедоступных списков значений эмотиконов. В обеих работах итоговая эмоциональная оценка осуществляется на основе комбинации оценок, полученных в результате анализа текста и анализа эмотиконов. Также во многих работах, рассмотренных в [3] и [4], эмотиконы вместе с хештегами используются для создания размеченных по тональности корпусов преимущественно из социальной сети Twitter.

Однако следует отметить, что в большинстве существующих на данный момент работ для проведения анализа тональности рассматриваются только те эмотиконы, которые имеют ясные эмоциональные значения. При этом в сети Интернет в настоящий момент используются сотни эмотиконов и эмодзи, имеющих свои уникальные эмоциональные и семантические значения, причем у значительного их количества невозможно четко определить эмоциональную окраску. Кроме того, часто встречается несоответствие текста и использованного в нем эмотикона или эмодзи, из-за чего нейтральное сообщение может получить яркую эмоциональную окраску, а позитивное сообщение, содержащее негативный эмотикон, может являться саркастическим. Также пользователи прибегают к замене слов с помощью эмодзи, обозначающих действие или предмет, поэтому исключение данных символов при анализе текста приведет к изменению смысла сообщения. Таким образом, для эффективного анализа тональности текста необходимо рассматривать эмотиконы и эмодзи как контекстно зависимые единицы, способные изменять свою эмоциональную окраску в зависимости от окружающих их слов.

В связи с рассмотренными особенностями интерес представляют работы, в которых при анализе тональности с помощью эмотиконов учитывается контекст. В работе [4] рассматривается относительно большое число часто используемых эмотиконов, имеющих и не имеющих четких эмоциональных значений, для построения пространства эмотиконов, где каждый эмотикон служит одним измерением. Текстовое сообщение проецируется в это

пространство на основе сходства слов и эмотиконов. В полученном отображении сообщения с близкой эмоциональной окраской имеют близкие координаты. Работа [5] описывает метод машинного обучения без учителя. Он основан на извлечении из текста слов, несущих эмоциональную окраску, и последующем предсказании класса тональности слов с использованием эмотиконов и скрытых полярностей, где каждое слово рассматривается как документ, а каждый сопутствующий эмотикон – как слово в этом документе.

На данный момент недостаточно внимания уделяется обработке эмодзи, поскольку при создании корпусов для лексического анализа эмодзи как правило удаляются, так как они представляют собой только коды, относящиеся к изображениям. Сейчас в открытом доступе можно найти корпуса на английском и французском языках, содержащие эмодзи, однако на русском языке таких корпусов нет. Рассмотренные подходы к работе с эмотиконами также могут быть применены к работе с эмодзи, но с учетом некоторых особенностей, отличающих эмодзи от эмотиконов. Например, в отличии от эмотикона, эмодзи может представлять собой не только изображение эмоции, но и предмета.

На основании выполненного обзора методов можно сделать вывод, что наиболее перспективными представляются подходы, учитывающие контекст. Также существует необходимость в разработке методов для анализа тональности текстов, содержащих эмодзи. В дальнейшем в работе планируется разработка алгоритма анализа тональности текстов на основе комплексного применения подходов, использующих методы машинного обучения и словари тональностей с учетом особенностей работы с эмотиконами и эмодзи.

**Литература**


1. Wegrzyn-Wolska K., Bougueroua L., Yu H., Zhong J. Explore the Effects of Emoticons on Twitter Sentiment Analysis // CS & IT-CSCP 2016. –2016. –P. 65–77.
2. Hogenboom A., Bal D., Frasincar F., Bal M., de Jong F., Kaymak U. Exploiting Emoticons in Sentiment Analysis // Proceedings of the 28th Annual ACM Symposium on Applied Computing. –2013. –P.703–710.
3. Guibon G., Ochs M., Bellot P. From Emojis to Sentiment Analysis. // WACAI 2016. –2016.
4. Jiang F., Liu Y., Luan H., Zhang M., Ma S. Microblog Sentiment Analysis with Emoticon Space Model // Social Media Processing. –2014. –P. 76–87.
5. Wang F., Wu Y. Sentiment-Bearing New Words Mining: Exploiting Emoticons and Latent Polarities // Computational Linguistics and Intelligent Text Processing. –2015. –P. 166–179.