Определение характеристик городов, влияющих на тональность отзывов, на основе анализа социальной сети Twitter

Александр Зырянов, Никита Путинцев

Exposoft, Новосибирск, Россия
{alexander.zyryanov44,putintsevnikita}@gmail.com

Аннотация Статья посвящена анализу сообщений в социальной сети Twitter. В ходе работы устанавливается, какие характеристики городов России влияют на тональность сообщений, посвященных тому или иному городу, другими словами, от каких характеристик зависит отношение людей к городу.

Ключевые слова: тональность текста, FRiS, машинное обучение, кластеризация.

1 Введение

Сейчас многие населенные пункты в России теряют свое население в пользу более крупных и оживленных городов, о чем можно судить на основании данных Росстата[1] и Госкомстата[2]. Уезжают в основном молодые и перспективные люди, при этом обратный приток населения незначителен. Это ведет к уменьшению уровня производства местных предприятий вследствие недостатка кадров, к уменьшению качества образования в местных школах и университетах, к ухудшению экономического и социального состояния городов в целом. Эта проблема становится все более актуальной и ее решение – совсем нелегкая задача, требующая глубокого понимания причин, которые её формируют. В связи с этим возникает интерес попытаться выявить основные движущие факторы этой проблемы при помощи анализа текстовых сообщений в социальных сетях.

В социальных сетях люди охотно высказывают свое мнение по любому вопросу. Причем, в отличие от соцопросов, где люди часто отвечают неохотно, не задумываясь, так что их мнение искажено или не соответствует действительности, в сетях высказывания зачастую сформированы настоящими мыслями людей. Кроме того, социальные сети могут предоставить миллионы сообщений для обработки практически даром, тогда как для проведения опроса такого же объема потребуются значительные затраты времени и ресурсов. Извлекая из этих сообщений мнения о различных городах, мы можем выяснить насколько хорошо или плохо люди к ним относятся.

[1] http://www.gks.ru/free_doc/doc_2013/bul_dr/mun_obr2013.rar
[2] http://www.gks.ru/free_doc/new_site/perepis2010/croc/perepis_itogi1612.htm

2 Постановка задачи

В основе нашего исследования лежит предположение о том, что тональность сообщений об определенном городе и тональность сообщений, сделанных из этого города в различных социальных сетях, может зависеть от его социальных, экономических и географических характеристик. Предполагается исследовать влияние таких характеристик, как плотность населения, климат, средний уровень заработной платы, возрастной состав, половой состав, наличие крупных торговых центров, парков и зон отдыха.

Для решения поставленной задачи в первую очередь необходимо набрать базу сообщений, которые относятся к определенным городам России или были в них созданы. Для этого нужно, во-первых, найти источник информации и, во-вторых, отфильтровать нужные для исследования сообщения. Далее необходимо определить тональность собранных высказываний.

Следующим шагом необходимо набрать информацию о выбранных характеристиках городов России и привести ее к удобному для работы виду.

Наконец, планируется выделить те характеристики городов, которые оказывают наибольшее влияние на тональность сообщений. Опционально города планируется разбить на таксоны и определить, к какому типу городов люди относятся лучше всего.

3 Аналогичные работы

Идея использовать Twitter для анализа мнений людей по различным вопросам возникла довольно давно [2]. Существуют похожие исследования, в которых тональность сообщений используется для предсказания каких-либо событий [1]. Так же уже разработано и опробовано большое количество различных методов анализа тональности сообщений, как использующих словарь эмотивной лексики [4], так и обучающихся на выборке [3], [5] . Эти методы оказались достаточно эффективными и подходят для поставленной в данной работе задачи.

4 Предполагаемое решение

В качестве источника сообщений решено использовать Twitter, так как он имеет широкую и разнообразную аудиторию, и содержит огромное количество сообщений, которое растет с каждым днем. Кроме того данная сеть предоставляет API для работы с потоком новых сообщений и данные в качестве грантов[3]. Для отбора сообщений о городах используется, во-первых, словарь их полных и сокращенных названий в различных морфологических формах, во-вторых геолокация. Для фильтрации спама на обучающей

[3] `https://blog.twitter.com/2014/introducing-twitter-data-grants`

выборке тренируется наивный классификатор Байеса. Сообщения подвергаются предварительной обработке, которая включает нормализацию сообщений, осуществляемую при помощи Pymorphy[4], и исключение стоп-слов. Стоп-слова планируется убрать автоматически, используя индекс TF-IDF на всей коллекции собранных сообщений из Twitter[6],а так же находящиеся в открытом доступе словари. Так же планируется заменить все эмотиконы на специальные слова, соответствующие их тональности.

Для оценки тональности полученных сообщений выбраны наивный классификатор Байеса из-за его простоты и эффективности, а так же метод опорных векторов из-за его точности [5]

Отдельный интерес представляет использование алгоритма классификации FRiS Stolp [7]. Интерес обусловлен желанием проверить пригодность данного алгоритма для решения задач анализа текстов.

Информацию о городах планируется собрать в полуавтоматическом режиме, используя интернет ресурсы, в частности, Wikipedia. Для кластеризации городов будет использован алгоритм FRiS Tax [7].

Для определения наиболее значимых признаков предлагается использовать способность алгоритма Random Forest определять важность используемых признаков [8]. Достаточно просто обучить алгоритм на таблице объект-свойство всех городов с целевым признаком тональности, который высчитывается как сумма тональностей всех сообщений, относящихся к данному городу.

5 Заключение

В работе обозначена задача выявления характеристики городов, которые оказывают влияние на тональность сообщений в социальных сетях. Так же представлено предполагаемое решение этой задачи, основанное на анализе тональности сообщений социальной сети Twitter методами машинного обучения.

Список литературы

1. *Bollen, J., Maon, H., Zeng, H.* Twitter mood predicts the stock market // Journal of Computational Science. Март 2011. № 1(2). С. 1–8.
2. *Pak, A., Paroubek, P.* Twitter as a Corpus for Sentiment Analysis and Opinion Mining // Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010
3. *Kouloumpis, E., Wilson, T., Moore, J.* Twitter sentiment analysis: The good the bad and the omg! // The AAAI Press, 2011
4. *Клековкина, М. В., Котельников, Е. В.* Метод автоматической классификации текстов по тональности, основанный на словаре эмоциональной лексики /в сб. Труды XIV Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». С. 118–123. — Переславль-Залесский: изд-во «Университет города Переславль», 2012.

[4] `https://pythonhosted.org/pymorphy/`

5. *Клековкина, М. В., Котельников, Е. В.* Автоматический анализ текстов на основе методов машинного обучения // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной Международной конференции «Диалог» (Бекасово, 30 мая – 3 июня 2012 г.). – Вып. 11 (18). – М. : Изд-во РГГУ, 2012.
6. *Ramos, J.* Using TF-IDF to Determine Word Relevance in Document Queries // The First Instructional Conference on Machine Learnin, 2003.
7. *Borisova, I. A., Dyubanov, V. V., Kutnenko, O. A., Zagoruiko, N. G.* Use of the FRiS-Function for Taxonomy, Attribute Selection and Decision Rule Construction /в сб «Lecture Notes in Computer Science» С. 256–270. — Berlin: Springer Berlin Heidelberg, 2011.
8. *Breiman, L.* Random Forests // Machine Learning, 2001. Т. 45. № 1. С. 5–32.

Determining Which Cities' Features Affect the Opinions' Sentiments on Twitter

Alexander Zyryanov, Nikita Putintsev

Exposoft, Novosibirsk, Russia
{alexander.zyryanov44,putintsevnikita}@gmail.com

Abstract. The paper is devoted to analysis of messages in the Twitter social network. The present study is focused on which Russian cities' features do affect the opinions' sentiments expressed by people.

Keywords: text sentiment, FRiS, machine learning, clustering.