МЕТОД АВТОМАТИЧЕСКОЙ ФИЛЬТРАЦИИ РЕКЛАМНЫХ СООБЩЕНИЙ ДЛЯ СИСТЕМЫ СЕНТИМЕНТ-АНАЛИЗА (на примере китайского языка)

А. А. Улахович

IHS Global
Минск, Беларусь
e-mail: a.ulahovich@gmail.com

В данной работе речь идет о фильтрации сообщений рекламного характера из общего потока сообщений, опубликованных на китайских специальных медиаресурсах и поступающих на вход построенной системы их сентимент-анализа. Рассмотрены особенности рекламы в китайских социальных сетях и вариант решения вышеуказанной проблемы, реализованный для известной системы автоматизации инженерии знаний и решения инновационных задач «Goldfire». В основу указанного решения положен так называемый «rule-based» метод, использующий специально составленные словари и множество лингвистических правил.

Ключевые слова: сентимент-анализ; фильтрация рекламы; китайские социальные сети.

METHOD OF AUTOMATIC FILTRATION OF ADVERTISING MESSAGES IN SENTIMENT ANALYSIS SYSTEM (in terms of chinese language)

A. A. Ulakhovich

IHS Global
Minsk, Belarus

This article is about the problem of filtration of advertising messages from the data stream, entering the sentiment-analysis system, created for processing of messages, taken from Chinese social media resources. Article presents special aspects of advertisement in Chinese social media and the solution of abovementioned problem, implemented in «Goldfire» searching platform – the system of engineering knowledge automatization and innovation problems solving. The method involves using of «rule-based» system, based on special dictionaries and set of linguistic rules.

Keywords: sentiment-analysis; advertisement filtration; Chinese social media.

ВВЕДЕНИЕ

В последние годы все большую роль в жизни современного человека играют социальные сети: наряду с частной перепиской люди публично делятся своими мнениями по самым разнообразным вопросам, оставляют комментарии и отзывы. Данная ин-

формация является очень важной, в том числе и для различных компаний: она позволяет знакомиться с мнением потребителей об их товарах и услугах, выявлять предпочтения и пожелания своих потенциальных клиентов, слабые места продуктов и т. п. Именно поэтому возникла потребность в технологиях, позволяющих распознавать необходимые данные в открытых источниках и определенным образом их обрабатывать.

В качестве одной из технологий указанного выше типа может быть успешно использован анализ тональности текста – класс методов контент-анализа в компьютерной лингвистике, предназначенный для автоматизированного выявления в текстах эмоционально окрашенной лексики и эмоциональной оценки авторов (мнений) по отношению к объектам, о которых идет речь [7] (далее в статье упоминается под названием «сентимент-анализ»). Существует несколько разных, но достаточно эффективных методов сентимент-анализа: методы, основанные на правилах и словарях [3], методы машинного обучения с учителем и без [7], методы, основанные на построении теоретико-графовых моделей [5]. В данной работе речь пойдет о первом из указанных типов методов. Эти методы строятся на поиске в тексте эмоционально окрашенных лексических единиц, входящих в специальный словарь. Наличие того или иного слова сигнализирует о том, что в предложении содержится оценочное мнение, которое может быть интерпретировано с помощью набора специальных правил. Например, в некотором предложении анализируемого текста найдено слово «качественный», которое входит в заранее построенный словарь слов с положительной окрашенностью; предложение, в котором встретилось данное слово, поступает далее в модуль обработки на предмет его соответствия условию некоторого паттерна из заданного их множества. Если такое соответствие достигнуто, то на выходе модуля выдается оценочное мнение (о разнообразных товарах, услугах, брендах, торговых марках и т. п.), формальное описание которого может быть представлено в виде кортежа: S = (Object, Value, Opinion), где Object – это объект оценочного мнения, Value – значение оценочного мнения (положительное или отрицательное), Opinion – часть предложения, в которой данное мнение выражается. Пример работы системы:

Вход: «Я обожаю свой iPhone, после него другими телефонами пользоваться совсем не хочется».

Выход: «iPhone» – нравится – «Я обожаю свой iPhone», где «iPhone» выступает в качестве объекта оценки, «нравится» является значением положительной оценки мнения, «Я обожаю свой iPhone» – частью предложения, в котором выражается данное мнение.

Вместе с тем серьезной проблемой при использовании такой технологии может стать большой объем текстов с рекламным содержанием, который не отражает мнения реальных потребителей. Попадая в систему автоматизированного сентимент-анализа данных, такие сообщения снижают точность и достоверность полученных результатов. Поэтому возникает необходимость создания специальной системы их предварительной фильтрации, что в конечном счете существенно уменьшит трудоемкость решения всей задачи и повысит ее качественные характеристики.

В настоящее время существует множество фильтров, работа которых направлена на обнаружение и удаление рекламных сообщений и спама [2, 4], однако чаще всего когда речь идет о спаме, имеются в виду сообщения электронной почты рекламного характера. В данном случае речь идет о системе фильтрации, которая создана специально для обнаружения рекламных сообщений в социальных сетях (в частности, на портале Weibo – китайском аналоге популярной сети Twitter).

Существует два основных подхода к обнаружению рекламного текста – анализ метаданных («никнейм» автора, хештеги, количество «ретвитов» и т. д.) и анализ непосредственного содержания сообщения. В нашем случае, ввиду особенностей формата данных, поступающих на вход в систему, анализирующую тональность текста, работа с метаданными представляется невозможной, поэтому система фильтрации строится на поиске в тексте специальных ключевых слов, сигнализирующих о том, что с большой долей вероятности анализируемый текст имеет рекламный характер. Слова такого рода также представлены в виде специальных лингвистических правил (паттернов, шаблонов).

ОСОБЕННОСТИ РЕКЛАМНЫХ СООБЩЕНИЙ В КИТАЙСКОЙ СОЦИАЛЬНОЙ СЕТИ «ВЭЙБО»

Sina Weibo (кит. 新浪微博) – китайский сервис микроблогов, запущенный компанией Sina Corp. в 2009 г. Сервис представляет собой своеобразный гибрид социальных сетей Twitter и Facebook и является одним из самых популярных сайтов в Китае. По некоторым данным, в 2016 г. количество пользователей сайта достигло 600 миллионов человек, при этом более 260 миллионов являются его активными пользователями [9].

Благодаря своей структуре и содержанию, портал – прекрасный источник данных для разработки и использования систем сентимент-анализа сообщений на китайском языке: по аналогии с сервисом Twitter пользователи ведут личные микроблоги, в которых делятся своим мнением о разнообразных товарах, выражают свои предпочтения и пожелания. Помимо рядовых «юзеров» сервис популярен среди различных компаний, торговых представительств, магазинов, ресторанов и частных продавцов. Их сообщения, особенно рекламного характера, зачастую также содержат эмоциональную лексику и оценочные суждения («товар превосходного качества», «высокий уровень обслуживания» и т. д.), из-за чего могут быть выделены системой и использованы при проведении сентимент-анализа того или иного объекта. Это искажает реальную картину, нивелируя саму суть работы системы, направленную на анализ мнений реальных пользователей. Поэтому была поставлена задача распознавания такого рода сообщений в общем массиве данных и их фильтрации (удаления) на стадии предварительной обработки входа. Для этого был построен специальный модуль, основанный на блоке лингвистических правил, обеспечивающих распознавание спама. Сложность задачи состояла также в невозможности использования метаданных: даже если «твит» сделан с официальной страницы той или иной компании, в названии страницы (никнейме) часто отсутствуют маркеры того, что это страница официального лица (нет таких слов, как «магазин», «компания», «официальный» и т. д.), поэтому работать приходилось с непосредственным текстом сообщения. Кроме того, в китайской интернет-среде рекламные сообщения, а также сообщения с официальных страниц брендов и компаний часто носят разговорный, «личный» характер, используется множество уменьшительно-ласкательных и разговорных слов и выражений, большое количество эмотиконов, знаков восклицания, повторений – элементов, свойственных в большей степени частным сообщениям пользователей, что также усложняет процесс распознавания рекламных текстов.

Несмотря на вышеуказанные трудности в конечном счете удалось выделить целый ряд маркеров, которые с большой долей вероятности сигнализируют о рекламном

характере сообщения. В соответствии с их смысловым значением можно выделить несколько групп таких маркеров:

– сылки, адреса сайтов, фразы «пройдите по ссылке», «кликните»;

– наличие контактных данных: телефонов, номеров сервисов мгновенного обмена сообщениями, адресов и т. д., призывы «связаться», «зафолловить», «добавить в адресную книгу»;

– информация о проведении акций формата «сделай ретвит, оставь комментарий – участвуй в розыгрыше призов»; слова и фразы «дерзай», «спеши поучаствовать», «не упусти свой шанс»;

– информация о ценах, скидках, акциях: «специальное предложение», «специальная цена», «подарочная карта», «распродажа»;

– упоминание о реальных и интернет-магазинах: «адрес магазина», «купи онлайн», «ссылка на страницу магазина», «приглашаем посетить».

При разработке паттернов пришлось также учитывать особенности китайского языка, которые в значительной степени затрудняли работу:

– слабо развитая морфология: в китайском языке предложения «я бесплатно получил» и «получи бесплатно» могут выглядеть абсолютно одинаково (**免**费获得); особенно это характерно для текстов сообщений в социальных сетях, где пользователи зачастую используют обороты разговорной речи, опускают личные местоимения, не ставят знаки препинания, не пользуются другими грамматическими показателями, обязательными для «правильных» печатных текстов;

– сильно развитая (в противовес слабой морфологии) синонимия: для того чтобы выделить все слова, обозначающие одно понятие, эксперту необходимо проделать огромную работу, при этом всегда существует риск, что какие-либо слова и выражения останутся неохваченными;

– замена одних иероглифов в слове на другие, близкие по звучанию (как намеренная, так и в результате опечатки).

ПРИНЦИПЫ РАБОТЫ СИСТЕМЫ ФИЛЬТРАЦИИ РЕКЛАМНЫХ СООБЩЕНИЙ

Когда мы говорим о «фильтрации рекламных сообщений», речь идет об удалении целых текстов, как правило, небольшого объема. Тексты удаляются при определенных условиях, выполненных для одного или нескольких фрагментов этих сообщений.

Всего существует три типа правил, регулирующих процесс принятия решений об удалении того или иного текста.

1. Правила «*killMsg(Cond1)*» – удалить сообщение, если оно удовлетворяет условию (шаблону) *Cond1*; шаблоны этих правил оперируют конкретными лексическими единицами, присутствие которых в анализируемых сообщениях с очень высокой степенью вероятности свидетельствет от том, что они носят рекламный характер. Например, шаблон типа «"**特价**为X» (специальная цена составляет *X*).

2. Правила «*killMsg(Cond1^Cond2)*» – удалить сообщение, если оно удовлетворяет одновременно условию *Cond1* и условию *Cond2*. Причем фрагменты сообщения, удовлетворяющего этим условиям, часто принадлежат различным его предложениям. Показательным примером для срабатывания таких правил являются сообщения со ссылками. Например, наличие в тексте лексических единиц «скидка», «распродажа»,

«покупка онлайн» и т. д. еще не гарантирует рекламного характера таких сообщений, но в совокупности с наличием ссылки уже дает возможность принимать решение об удалении всего текста. Пример: «在线购买相机» (купить фотоаппарат в интернет-магазине) + < ссылка >.

3. Правила «*killMsg(Cond1* ^$\overline{Cond2}$*)*» – удалить сообщение, если оно удовлетворяет условию *Cond1* и одновременно не удовлетворяет условию *Cond2*, которое ориентировано на наличие в тексте сообщений личных местоимений. Например, если перед фрагментом сообщения, удовлетворяющим шаблону «бесплатно получить подарок», не будет личного местоимения в единственном числе, то сообщение будет удалено. Пример: существует шаблон "免费获得礼物" (дословно – «бесплатно получить подарок»). При наличии перед шаблоном личного местоимения в единственном числе предложение будет означать «я бесплатно получил подарок»; система примет решение оставить данное сообщение и отправить его на дальнейшую обработку. При отсутствии личного местоимения предложение приобретет значение «бесплатно получи подарок» и будет удалено.

Что касается количества правил, оно зависит от степени их обобщения на этапе формализации. В нашем случае их 44; все они разделены на группы по семантическому признаку, что существенно повышает эффективность работы экспертов на этапах разработки и совершенствования системы фильтрации. К плюсам представленного метода можно отнести простоту принципа построения системы, возможность учитывать контекст и быстро вносить требуемые изменения и дополнения, тем самым – наращивать качественные показатели решения задачи, полностью контролировать работу системы. К недостаткам – относительно высокую трудоемкость построения самих шаблонов, поэтому, как показывают исследования, вполне перспективным может оказаться гибридный метод, сочетающий в себе данный с элементами машинного обучения.

БИБЛИОГРАФИЧЕСКИЕ ССЫЛКИ

1. Клековкина М., Котельников Е. Метод автоматической классификации текстов по тональности, основанный на словаре эмоциональной лексики // RCDL-2012. Переславль-Залесский, 2012.
2. Ларионова А., Хорев П. Метод фильтрации спама на основе искусственной нейронной сети // Интернет-журн. «Науковедение». 2016. Т. 8, № 3.
3. Пазельская А., Соловьев А. Метод определения эмоций в текстах на русском языке // The international conference on computational linguistics and intellectual technologies Dialogue 2011. М., 2011. С. 510–512.
4. Семенова М., Семенов В. Метод автоматической фильтрации при борьбе со «спамом» // Изв. вузов. Сер.: Приборостроение. 2009. Т. 52. № 9.
5. Усталов Д. Извлечения терминов из русскоязычных текстов при помощи графовых моделей. Екатеринбург : УРФУ, 2012.
6. Hu M., Liu B. Mining and summarizing customer reviews // KDD, Seattle. 2004. P. 168–177.
7. Pang B., Lee L. Opinion mining and sentiment analysis // Foundations and trends in information retrieval. 2008. № 2. P. 1–135.
8. Peng Q., Zhong M. Detecting Spam Review through Sentiment Analysis // J. of software. 2014. Т. 9, № 8.
9. By the numbers: 58 amazing Weibo statistics (May 2016): [Electronic resource]. URL: <http://expandedramblings.com/index.php/weibo-user-statistics/ >