СЕМАНТИЧЕСКИЙ АНАЛИЗ СООБЩЕНИЙ В СОЦИАЛЬНЫХ СЕТЯХ

Кузьменков Юрий Сергеевич
магистрант, кафедра информационных технологий ГГТУ,
Беларусь, г. Гомель
E-mail: the_yuri@mail.ru

Мурашко Игорь Александрович
научный руководитель, доктор технических наук, профессор
Гомельского государственного технического университета им. П.О. Сухого,
Беларусь, г. Гомель

Основной задачей, поставленной в данной статье, является описание процесса реализации анализа контента в социальной сети. Решение задачи анализа контента можно разделить на следующие этапы:

1. Исследование особенностей сообщений в социальной сети.
2. Реализация метода определения полярности с учетом этих особенностей и тестирование его эффективности.
3. Реализация метода извлечения аспектов с учетом этих особенностей и тестирование его эффективности.

Исследование особенностей сообщений в Твиттер

Перейдем к решению первой задачи. В контексте задачи анализа эмоциональной окраски текста, основными особенностями сообщений в Твиттер являются:

- Малый размер сообщения – 140 символов.
- Наличие сленга, сокращений и грамматических ошибок.
- Наличие специальных символов.
- Использование ссылок на других пользователей и на внешние ресурсы.

Рисунок 1. Сообщения в социальной сети Twitter

В работах [1,2] эти особенности являются решающим фактором при разработке методов. Например, в работе [2] используется алгоритм определения полярности сообщения, основанный на словарном методе. В словарь входят символы и слова, обозначающие эмоции. Намеренные грамматические ошибки в словарных терминах, такие как повторение согласных букв, в словах, например «Победаааааа!», вместо «победа», увеличивают изначальный вес терминов.

Реализация алгоритма определения полярности сообщений

Для реализации эффективного алгоритма определения полярности сообщения было решено использовать методы машинного обучения, как наиболее эффективные в данных условиях.

Выбор меры эффективности алгоритмов

Традиционно эффективность задачи классификации текста формулируется в терминах точности и полноты. Интерпретируем эти термины для задачи определение полярности документов:

Таблица 1.

Результаты бинарной классификации

	Классифицированы как +	Классифицированы как -
Класс +	TP	FN
Класс -	FP	TN

Пусть в коллекции из N документов N_p документов имеют положительную эмоциональную окраску (принадлежат к классу +) и N_n документов – отрицательную (принадлежат классу -) .

В результате классификации этих документов, к классу + правильно отнесены TP документов, неправильно – FP, к классу - правильно отнесены TN документов, неправильно – FN. Тогда, относительно класса +:

Точность – отношение числа правильно отнесенных документов к классу + к числу всех документов, отнесенных к классу +:

$$Precision = \frac{TP}{TP+FP} \quad (1)$$

Полнота – отношение числа правильно отнесенных документов класса 1 к числу документов класса + в коллекции:

$$Recall = \frac{TP}{TP+FP} \quad (2)$$

Выбор признаков

Метод определения полярности, реализуемый в данной статье, не предполагает использования априорных предположений о том, какие слова или символы могут содержать сообщения, принадлежащие к какому-либо классу. Это значит, что все признаки априорно равнозначны.

Часто используемыми признаками при решении задачи определения полярности являются н-граммы слов. В данной работе под словом подразумевается любая последовательность букв алфавита, а под н-граммой порядка n - разделенная пробелами последовательность из n слов.

Например, сообщение «Сегодня шел дождь. ;)» содержит только следующие н-граммы слов первого и второго порядка: «Сегодня», «шел», «дождь», «Сегодня шел» и «шел дождь».

В ряде работ в качестве признаков используются части речи. Это объясняется тем, что мнение содержит субъективную лексику. Например, в работе [2] составляется словарь прилагательных и наречий, как терминов, выражающих эмоцию. По этой же причине было так же решено выбрать в качестве признаков н-граммы из частей речи.

Как было показано ранее, такие особенности сообщений в социальной сети, как сленг и эмотиконы сигнализируют. В качестве признаков был выбран набор часто употребляемых эмотиконов.

Выбор алгоритма классификации

В ряде исследований по определению полярности текста, высокую эффективность показали методы обучения с учителем. Эти методы использовались как в ранних работах по определению полярности документа в среднем, так и в современных работах, где анализируется предложения и короткие текстовые сообщения.

Для решения поставленной задачи были выбраны 2 алгоритма, показавшие себя наиболее эффективными при определении полярности коротких текстовых сообщений [2].

Метод опорных векторов

Метод опорных векторов относится к семейству линейных классификаторов. Целью линейной классификации является поиск гиперплоскости в пространстве признаков, разделяющей все объекты на два класса.

Основная идея метода опорных векторов состоит в поиске разделяющей гиперплоскости, максимально удаленной от ближайших к ней точек в пространстве признаков.

В случае линейно разделимой выборки поиск гиперплоскости можно записать в виде задачи оптимизации:

$$\frac{1}{2}\|\omega\|^2 \to \min_{\omega, b}$$
$$y_i(\omega^T x_i + b) \geq 1, j = 1,....,\mathrm{m} \quad (3)$$

Где $\frac{1}{\|\omega\|}$ - величина зазора между гиперплоскостью и ближайшими к ней

точками, как первого, так и второго класса. А $y_i(\omega^T x_i + b)$ - произведение значения класса точки и ее положения относительно гиперплоскости.

Для более общего случая линейно неразделимой выборки алгоритм может допускать ошибки на обучающих объектах. Новая задача оптимизации включает в себя требование минимизации ошибки:

$$\begin{aligned} &\frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{l} e_i \to \min_{\omega,b} \\ &y_i(\omega^T x_i + b) \geq 1 - e_i, i = 1,....,\mathrm{k} \\ &e_i \geq 0, \quad i = 1,....,\mathrm{k} \end{aligned} \tag{4}$$

Переменные e_i характеризуют величину ошибки на примере выборки из k элементов. Константа C позволяет находить компромисс между максимизацией величины зазора и минимизации суммарной ошибки на тренировочной выборке.

Наивный байесовский классификатор

Наивный байесовский классификатор — вероятностный классификатор, основанный на теореме Байеса и (наивном) предположении о статистической независимости случайных величин.

$$p(C \mid F_1,...,F_2) = \frac{p(C)p(F_1,...,F_2 \mid C)}{p(F_1,...,F_2)} \tag{5}$$

Основное достоинство данного классификатора заключается в низкой вычислительной сложности, а также в оптимальности, при условии действительной независимости признаков.

Выбор обучающей выборки

Для использования методов обучения с учителем требуется обучающая выборка. Обычно обучающее множество составляется из примеров той области, в которой будет применяться классификатор.

В качестве обучающей и проверочной выборки был составлен корпус, состоящий из 8000 предложений, для которых определена полярность. Часть этих предложений была извлечена из размеченного корпуса, предоставленного

авторами [2] для свободного доступа. Другая часть была получена с помощью онлайновой системы Sentiment140 анализа эмоциональной окраски сообщений социальной сети Твиттер.

Все примеры полученного обучающего множества получены из мнений об электронной технике, а именно о мобильных телефонах, планшетах, плеерах.

Тестирование эффективности

Для тестирования алгоритма определения полярности сообщений использовался метод кросс-валидации. Процедура кросс-валидации происходит следующим образом:

1. Фиксируется множество разбиений обучающего множества на тренировочное и контрольное подмножества.

2. Для каждого разбиения происходит обучение алгоритма на тренировочном множестве, затем тестирование на контрольном.

3. Результатом кросс-валидации алгоритма является среднее значение проведенных результатов тестирования на контрольном множестве.

В данной работе разбиение на множества производилось случайным образом. Попадание каждого предложения в одно из двух множеств равновероятно.

В таблице приведен результат тестирования алгоритмов, в качестве признаков выбраны н-граммы, в которые входят слова и эмотиконы:

Таблица 2.

Полярность сообщения: в качестве признаков выбраны н-граммы слов

Алгоритм	Юниграммы		Биграммы		совместно	
	Точность	**Полнота**	**Точность**	**Полнона**	**Точность**	**полнота**
Байес	0.73	0.7	0.72	0.68	0.72	0.71
Метод опорных векторов	0.81	0.74	0.76	0.67	0.81	0.72

В следующей таблице приведен результат тестирования, где в качестве признаков выбраны части речи и биграммы частей речи:

Таблица 3.

Полярность сообщения: в качестве признаков выбраны н-граммы частей речи

Алгоритм	Части речи		Биграммы ч.р.	
	Точность	Полнота	Точность	Полнона
Байес	0.6	0.55	0.54	0.48
Метод опорных векторов	0.65	0.57	0.56	0.53

Наиболее эффективной конфигурацией в терминах точности и полноты оказался метод опорных векторов, обученный на юниграммах слов и эмотиконов.

Выбор меры эффективности алгоритмов

Эффективность алгоритмов извлечения аспектов формулируется в терминах точности и полноты.

В контексте решаемой задачи эти метрики имеют следующий смысл. Алгоритм извлечения аспектов проверяет каждый термин документа на принадлежность множеству аспектов. Тогда точностью этого алгоритма называется отношение числа правильно определенных аспектов к числу всех терминов, отнесенных к классу аспектов, а полнотой – отношение числа правильно определенных аспектов к числу аспектов в документе.

Выбор алгоритма извлечения аспектов

Для решения этой задачи было решено реализовать алгоритм, основанный на методе распространения, описанный в [3].

Для улучшения эффективности алгоритма так же используются приведенные в обзоре статистические методы: кандидатами в термины могут быть н-граммы, содержащие только существительные и прилагательные и прошедшие через C-value фильтр и частотный фильтр.

Обучающая выборка

Метод распространения основан на извлечении терминов из связанных между собой предложений, но сообщения социальной сети Твиттер, как правило, независимы друг от друга. Поэтому в качестве тренировочного

множества используется множество полнотекстовых обзоров электронной техники, составленное вручную.

Процесс составления тренировочной выборки упрощает тот факт, что методы обучения с учителем не требуют создания выборки пар вида (пример, класс), как в случае обучения с учителем.

Тестирование эффективности

Для тестирования используются предложения, взятые из размеченного корпуса, предоставленного авторами работы [3].

Из корпуса взяты 1500 предложений из обзора двух цифровых фотоаппаратов и 1820 предложений из обзора Мп3 плеера. В каждом из этих предложений выделены аспекты. Результаты тестирования реализованного алгоритма показаны в таблице 4:

Таблица 4.

Извлечение аспектов

Домен	точность	полнота
фотоаппараты	0.65	0.76
Цифровые плееры	0.62	0.71

Проанализировав полученные результаты, можно сделать следующие выводы: для решения задачи определения полярности предложений и коротких сообщений эффективны как алгоритмы обучения с учителем, так и методы, основанные на словарях. Проблемой обучения с учителем является составление тренировочного корпуса с примерами из предметной области, в которой будет использоваться классификатор. Однако схожей проблемой обладают и словарные методы: веса терминов словаря, составленного для одной предметной области, могут оказаться неадекватными для другой.

Задача извлечения аспектов часто решается с помощью методов обучения без учителя и статистическими методами. Для увеличения эффективности этих методов используются лингвистические и частотные фильтры, позволяющие отсеивать слова, не имеющие отношения к аспектам.

Список литературы:

1. Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, K.P. Measuring User Influence in Twitter: The Million Follower Fallacy.// Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM), Washington, May 2010.
2. SportSense [HTML] (ec2.compute1.amazonaws.com/sportsense/)
3. Pang B. & Lee L. Opinion Mining and Sentiment Analysis // Foundations and Trends in Information Retrieval, v.2 n.1-2, January, 2008 - pp.1-135.