ИСПОЛЬЗОВАНИЕ СЕМАНТИЧЕСКИХ КАТЕГОРИЙ В ЗАДАЧЕ КЛАССИФИКАЦИИ ОТЗЫВОВ О КНИГАХ

USING SEMANTIC CATEGORIES IN APPLICATION TO BOOK REVIEWS SENTIMENT ANALYSIS

Фролов А.В. (*anton_frolov@rco.ru*), Поляков П.Ю (*pavel@rco.ru*), Плешко В.В. (*vp@rco.ru*)

ООО «ЭР СИ О», Москва, Россия

В данной работе исследуется метод использования семантических категорий фактов в качестве классификационных признаков для решения задач классификации отзывов о книгах на 2 (положительный, отрицательный) и 3 (положительный, отрицательный и нейтральный) класса. Кроме того, проанализированы основные ошибки и подводные камни, которые могут встречаться в задачах подобного рода.

Ключевые слова: анализ мнений, определение тональности, автоматическая классификация, машинное обучение, извлечение классификационных признаков, метод опорных векторов, регрессия

Frolov A.V. (*anton_frolov@rco.ru*), Polyakov P.Yu. (*pavel@rco.ru*), Pleshko V.V. (*vp@rco.ru*)

RCO LLC, Moscow, Russian Federation

This paper studies use of fact semantic categories in application to book reviews sentiment analysis. The tasks were to divide book reviews into 2 classes (positive, negative) and into 3 classes (positive, negative, neutral). Moreover, main machine learning pitfalls concerning opinion mining were classified and analyzed.

Key words: opinion mining, sentiment analysis, document categorization, machine learning, classification feature extraction, support vector machine, regression, two-class classifier, multi-class classifier

Введение

Задача автоматической классификации отзывов о товарах является на сегодняшний день весьма востребованной, о чем свидетельствует появление коммерческих сервисов мониторинга социальных сетей и блогов (например, [1]). Тем не менее, для русскоязычного контента долгое время отсутствовали общедоступные размеченные корпуса, на которых разработчики могли бы провести оценку качества своих методов. Данный пробел были призваны восполнить новые дорожки семинара РОМИП, в рамках которых участникам предлагалось решить задачу классификации отзывов о книгах, фильмах и фотокамерах.

В настоящей работе исследуются методы решения задачи классификации отзывов о книгах на 2 (положительный, отрицательный) и 3 (положительный, отрицательный, нейтральный/средний) класса в рамках дорожек РОМИП 2012 [2].

Постановка задачи

Участникам была предложена коллекция для обучения, представляющая собой набор отзывов пользователей блогов на книги различных жанров (всего 24 160 отзывов). Каждый отзыв имел пользовательскую оценку от 1 до 10 баллов. Из имеющихся дорожек нами были выбраны две: дорожка по классификации отзывов пользователей на 2 класса и дорожка по классификации отзывов пользователей на 3 класса. В первом случае требовалось разделить отзывы на положительные и отрицательные. Во втором случае требовалось разделить отзывы на 3 класса: "положительный", "средний" (в отзыве указываются достаточно значимые положительные и отрицательные стороны оцениваемой книги) и "отрицательный".

Объединение фактов в классы

Было решено усовершенствовать лингвистический подход, представленный в [3] и продемонстрировавший хорошие результаты на прошлогодней дорожке. Для этого, был проведен анализ прошлых результатов и проверена гипотеза о том, что обучащая колекция слишком мала, чтобы обеспечить достаточные частоты отдельных фактов. Одним из возможных способов исправления этого недостатка является применение семантических фильтров, которые позволяют объединять несколько фактов в один класс.

Напомним, что извлечение фактов производится с помощью семантических шаблонов. Семантический шаблон представляет собой орграф с ограничениями на вершины, которые могут включать ограничения на часть речи, имя, семантический тип, синтаксические связи и тому подобные атрибуты (см. рис. 1).

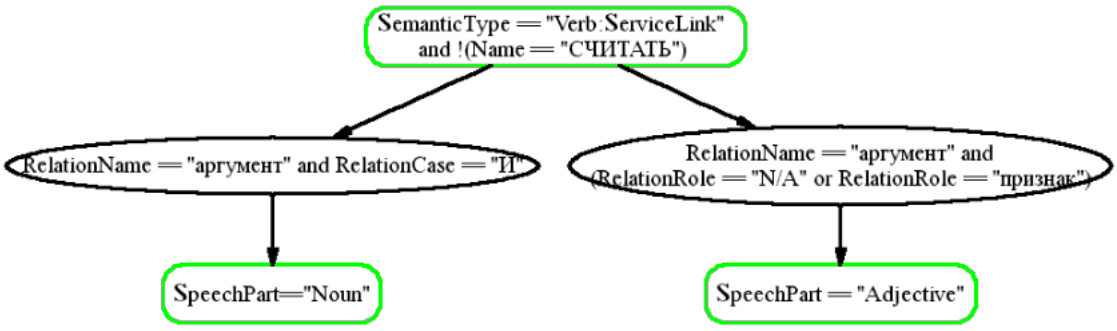

Рис. 1. *Пример семантического шаблона для определения оценки книги.*

Кроме того, факты можно обобщать с использованием специальных словарей именуемых фильтрами, используя наборы синонимов для положительных, отрицательных и нейтральных оценок. Недостатком такого подхода является необходимость ручного отбора лексики для фильтров: трудоемкая задача, требующая привлечения эксперта-лингвиста. С другой стороны, возможно отобрать лишь базовую лексику усилиями человека и дополнить её уже автоматически. Именно этот способ и был выбран основным.

Для реализации были взяты фильтры, использовавшиеся в прошлогодней дорожке, и пополнены новой лексикой, которую система находила самостоятельно. Более подробно процесс извлечения фактов и возможности применения шаблонов и фильтров описан в [3].

Новые лексика формировалась следующим образом:

На обучающей выборке делался прогон системы, настроенный на извлечение фактов. Полученные факты принимались как единственные признаки документов и на их основе проводилась классификация. При выборе классификатора предпочтение было отдано байесовскому классификатору с функцией Пуассона в качестве функции плотности распределения вероятностей слов в тексте [4]. Далее, система брала по отдельности профили каждого из классов и использовала заполненные слоты фактов из этих профилей для отбора кандидатов новой лексики. После чего, полученные списки фильтровались по частотному порогу и сливались с уже имеющимися. Для улучшения качества использовался список оценочной лексики опубликованный организаторами РОМИП [5]. Если слот факта содержался в нем, то его вес увеличивался в 10 раз.

Таблица 1: Пример фильтра

Subject	**Quality Verb**	**Quality Emotion**	**Quality Adjective**
КОНЕЦ КНИГИ *КОНЦОВКА* *ФИНАЛ* *РАЗВЯЗКА* *ХЭППИЕНД* *ХЭППИ ЭНД* *ХЭППИ* *ХЭППИ-ЭНД*	*УБИТЬ* *ИДТИ* *РАСТЯГИВАТЬСЯ* *СДЕЛАТЬ* *НЕ ПОНРАВИТЬСЯ*	*ЖДАТЬ* *УБИТЬ* *НЕ ЖДАТЬ* *ИДТИ* *РАСТЯГИВАТЬСЯ* *НЕ ПОНРАВИТЬСЯ* *НАЗВАТЬ*	*УМОПОМРАЧИТЕЛЬНЫЙ* *ОПТИМИСТИЧНЫЙ* *ДУРАЦКИЙ* *ЗАКРЫТЫЙ* *УТОМИТЕЛЬНЕЙШИЙ* *СКУЧНЫЙ* *ПЕЧАЛЬНЫЙ*

В таблице 1 приведен пример автоматически заполненного фильтра, позволяющий объединять несколько фактов в один класс: “негативный отзыв о конце произведения”. В данном случае, факты с четырьмя слотами (субъект, глагол, эмоция, прилагательное) отождествляются, если содержимое соответствующих слотов содержится в одном и том же фильтре.

Несмотря на редкие ошибки (в примере – слово “оптимистичный” оказалось в фильтре для негативного класса), большая часть лексики адекватна. Далее, качество отбора терминов можно улучшить повысив репрезентабельность и размер обучающей выборки.

Таким образом были получены классы для выявления оценки книги, персонажей произведения, языка,сюжета и автора.

Методы классификации

Для формирования качественной обучающей выборки, двое наших экспертов независимо оценивали коллекцию и проставляли оценки: негативный отзыв, позитивный отзыв, отзыв содержит как положительные, так и отрицательные характеристики. Каждый эксперт оценил порядка 4000 отзывов, большая часть из которых была отнесена к положительным. Согласованность оценок экспертов при формировании обучающей выборки достигала $r \sim 0.8$, где r - корреляционный коэффициент Пирсона.

В эксперименте использовалось два подхода. В первом подходе для обучения классификатора использовались оценки самих пользователей. По ним строилась линейная регрессионная модель в реализации SVM-Light [6]. Затем по этой модели вычислялись веса документов из обучающей выборки и подбирались пороги отнесения документа к заданным классам таким образом, чтобы получить наилучшее соответствие между получаемым разбиением и разметкой экспертов (максимизировалась F-мера).

Во втором подходе классификатор строился только на основе обучающей выборки, сформированной экспертами. Рассмотрены следующие методы классификации:

- Линейный классификатор, в котором обучение производится для каждого класса независимо от других классов, в реализации SVM-Light [6]. Если в процессе обработки тестовой выборки один документ попадал в несколько классов, то мы принудительно относили его к одному классу, в котором этот документ имел самый большой вес.

- Линейный классификатор, который обучается независимо на классах положительных и отрицательных отзывов в реализации SVM-Light [6], а используется в задаче классификации на 3 класса. К классу нейтральных отзывов мы относили документы, которые классификатор приписывал одновременно и классу положительных, и классу отрицательных отзывов.

Результаты

В работе проанализированы результаты оценки 4 прогонов классификации на 2 класса и 4 прогона классификации на 3 класса. Прогоны варьировались по типу классификатора:

- SVM: метод опорных векторов с разделением классов по принципу “один против всех”

- Regression: линейная регрессионная модель

и по набору признаков, используемых для описания документов:

- Base: признаками являются леммы (отдельные слова) и темы (словосочетания)

- Hybrid: в добавок к признакам из Base добавляются ещё классы фактов.

Влияние различных подходов на качество классификации мы оценивали с помощью F1-меры [7]. Также, для удобства в таблицах приведены значения полноты, точности и аккуратности.

Таблица 2: результаты прогонов для 2 классов

	P-macro	R-macro	F-macro	Accuracy
Base SVM	0.676425	0.620273	0.647133	0.86046
Hybrid SVM	0.577041	0.552521	0.564515	0.82945
Base Regression	0.627363	0.627363	0.627363	0.82945
Hybrid Regression	0.605004	0.634454	0.619379	0.79845

Данные, приведенные в Таблице 2, говорят о том, что лучше всего с задачей справился классификатор, не использующий продвинутые лингвистические признаки. Некоторое объяснение этому факту приводится в следующем разделе статьи. Также, заметно, что использование регрессионной модели даёт лучшие результаты, нежели классический SVM в случае гибридной модели и худшие в случае базовой модели.

Таблица 3: результаты прогонов для конкретных классов (2 класса)

	P-pos	R-pos	F-pos	P-neg	R-neg	F-neg
Base SVM	0.898305	0.946428	0.921739	0.454545	0.294117	0.357143
Hybrid SVM	0.881355	0.928571	0.904348	0.272727	0.176471	0.214286
Base Regression	0.901785	0.901786	0.901786	0.352941	0.352941	0.352941
Hybrid Regression	0.90566	0.857143	0.880734	0.304348	0.411765	0.35

Если взглянуть на Таблицу 3, то становится ясно, что основную сложность для классификатора составили негативные рецензии. Сложность задачи классификации негативных отзывов может быть объяснена следующими факторами:

1. В тестовой коллекции, как и в обучающей, оказалось на порядок больше положительных отзывов: 17 против 112.
2. Размер тестовой выборки оказался очень невелик: всего 129 документов. Вместе с фактором 1 это привело к достаточно большому влиянию статистических погрешностей на итоговый результат.

3. Значительная часть (8 из 17) отрицательных отзывов не содержала резко негативной оценки рецензируемого объекта и при разделении на три класса была отнесена к нейтральным.

Стоит отметить, что тестовая коллекция размечалась всего одним экспертом, что дополнительно снижает объективность итогового результата.

При классификации на три класса картина существенно меняется: регрессионная модель показывает существенное отставание от SVM.

Таблица 4: результаты прогонов для 3 классов

	P-macro	R-macro	F-macro	Accuracy
Base SVM	0.544343	0.554074	0.549165	0.697674
Hybrid SVM	0.450879	0.467037	0.458816	0.666666
Base Regression	0.354825	0.333703	0.343940	0.542636
Hybrid Regression	0.354826	0.333704	0.343941	0.542636

Таблица 5: результаты прогонов для нейтрального класса (3 класса)

	P-neu	R-neu	F-neu
Base SVM	0.891566	0.74	0.808743
Hybrid SVM	0.870588	0.74	0.8
Base Regression	0.857142	0.72	0.782608
Hybrid Regression	0.864864	0.64	0.735632

Таблица 6: результаты прогонов для негативного и позитивного классов (3 класса)

	P-pos	R-pos	F-pos	P-neg	R-neg	F-neg
Base SVM	0.341463	0.7	0.459016	0.4	0.222222	0.285714
Hybrid SVM	0.282051	0.55	0.372881	0.2	0.111111	0.1428571
Base Regression	0.147058	0.25	0.185185	0.090909	0.111111	0.1
Hybrid Regression	0.116279	0.25	0.158730	0.083333	0.111111	0.095238

Анализ результатов

Значительное влияние на результаты оказали особенности тестовой коллекции, а именно: сильный дисбаланс в сторону нейтральных (положительных, при разбиении на 2 класса) отзывов и её небольшой размер (вдвое меньше прошлогодней). Кроме того, негативные отзывы оказались достаточно сильно смещены: около половины из них касались одного и того же произведения: “Ангелы и демоны” Дэна Брауна.

Согласованность эксперта РОМИП и нашего составила $r = 0.78$.

Как можно видеть из таблиц результатов, основную сложность для классификатора представили негативные отзывы. Нами были проанализированы и категоризированы случаи, которые были некорректно оценены системой. Условно их можно разделить на следующие категории:

1. Автор большую часть рецензии пересказывает содержание книги. В этом случае, в тексте может содержаться достаточное количество шумовой лексики, чтобы классификатор выбрал неверное решение.

2. Автор перечисляет положительные стороны произведения, но итоговую оценку даёт негативную. Пример: “Сюжет есть. И интрига присутствует. А вот то, как разворачиваются действия - не вдохновляет ни коим образом.” В итоге, положительная лексика перевешивает негативную за счет количества. Особенно уязвим к такому роду ошибок метод с извлечением классов фактов, поскольку статистики по фактам будет собрано существенно меньше нежели по леммам.

3. Автор ссылается на положительные отзывы других людей, но собственную оценку даёт негативную. Пример: “С сожалением сообщаю: не для моих мозгов. Говорят, книга очень хорошая. Промолчу.”

4. В представленной системе вместо классических списков стоп-слов использовался фильтр по семантическим категориям. Так, например, отфильтровывались любые числительные или служебные слова. Этот метод продемонстрировал хороший результат на рецензиях с портала Imho-net. Однако, из-за того, что вместо обычных рецензий для оценки были представлены записи из блогов, множество шумовых слов существенно изменилось.

Следует отметить, что использовалось разбиение “один против всех” с отнесением документа к классу, с наибольшим весом. В результате, многие из ошибочно проклассифицированных документов имели отрицательный вес для обоих классов. В задаче же классификации на три класса система успешно отнесла такие рецензии к нейтральным, улучшив показатели.

Кроме того, п.4 усугубляется тем, что тип рецензий значительно отличается от прошлогоднего. Если в прошлый раз тестовая выборка производилась из записей ресурса Imhonet.ru, сейчас классифицируемыми объектами были записи из персональных блогов. Действительно, рецензии из блогов меньше фокусируются на объекте оценки и имеют тенденцию содержать большое количество шумовой лексики, которую невозможно отсеивать семантическими фильтрами.

Таблица 7: Сравнение результатов на прошлогодней дорожке

	Expert 1 F-macro	Expert 2 F-macro
New hybrid SVM	0.503129181	0.500560892
Old hybrid SVM	0.467705308	0.484938518
Base regression	0.4903	0.4998

Как видно из результатов, классификатор с возможностью выделения фактов справился с задачей хуже чем базовый. У нас возникла гипотеза, что это произошло из-за смещенности новой коллекции. Для её проверки, мы провели тесты на прошлогодней дорожке. Оказалось, что новый классификатор показал улучшение в задаче классификации на три класса по сравнению с прогоном hybrid и даже нежели метод регрессии из [3], который продемонстрировал наилучший результат среди всех систем, участвовавших в прошлогодней дорожке. Таким образом, можно сделать вывод, что при наличии несмещенной коллекции новый классификатор работает лучше.

Возможные улучшения

Для решения упомянутых в предыдущем разделе проблем следует изменить набор признаков, характеризующих документ.

Во-первых, для правильной оценки больших рецензий необходимо уметь корректно выделять резюмирующую оценку. Действительно, большую часть таких текстов занимает пересказ сюжета или же отвлеченные рассуждения. В то время как реальная оценка делается либо в первых нескольких предложениях, либо в последних.

Во-вторых, при классификации рецензий, желательно выделять объект рецензии. Дело в том, что в одном тексте может “рецензироваться” и сравниваться между собой множество книг. Пример: “Сегодня я читал X и мне не понравилось. Гораздо хуже замечательной книги Y, которую я читал вчера”. Если объект не указан при постановке задачи, то система должна уметь выделять его сама.

В-третьих, при разделении отзывов на два класса необходимо отделять мнение автора рецензии от характеристик из внешних источников (“говорят книга хорошая, но мне не очень понравилась”). Очевидно, что в этом случае мнение автора должно иметь больший вес. В задаче классификации на три класса это уже не так критично и, в целом, различным источникам можно присваивать схожие веса.

Заключение

Проведена апробация ряда методов решения задач классификации отзывов о книгах на 2 и 3 класса и усовершенствован метод обогащения классификационных признаков в рамках лингвистического подхода с применением словарей оценочной лексики и машинного пополнения фильтров. Кроме того, были проанализированы и

категоризированы основные ошибки допускаемые классификатором и выяснено направление для дальнейшей работы.

Литература

[1] Sentiment 140 // http://www.sentiment140.com
[2] *Четверкин И.И., Лукашевич Н.В.* Тестирование систем анализа тональности на семинаре РОМИП-2012
[3] *Поляков П. Ю., Калинина М. В., Плешко В. В.,* Исследование применимости методов тематической классификации в задаче классификации отзывов о книгах // Труды Диалог'12 (Наро-Фоминск, 30 мая – 3 июня 2012г.)
[4] *Плешко В.В., Поляков П.Ю., Ермаков А.Е.* RCO на РОМИП 2009 // Труды РОМИП 2009. (Петрозаводск, 2009г.). - Санкт-Петербург: НУ ЦСИ, 2009 - с. 122-134
[5] *Chetviorkin I., Loukachevitch N.* Extraction of Russian Sentiment Lexicon for Product Meta-Domain // In Proceedings of COLING 2012
[6] *Joachims T.* Making large-scale support vector machine learning practical // Advances in Kernel Methods: Support Vector Machines / B.Scholkopf, C. Burges, A. Smola (eds.) - MIT Press: Cambridge, MA" – 1998
[7] Chetviorkin I., Braslavskiy P., Loukachevitch N. Sentiment Analysis Track at ROMIP 2011

References

[1] Sentiment 140 // http://www.sentiment140.com
[2] *Chetviorkin I., Loukachevitch N.* Sentiment analysis track at ROMIP‘12
[3] *Polyakov P. Yu., Kalinina M. V., Pleshko V. V.,* Research of applicability of thematic classification to the problem of book review classification. Dialog ’12. Naro-Fominsk, 2012
[4] *Pleshko V.V., Polyakov P.Yu, Ermakov A.E.* RCO at RIRES 2009 [RCO na ROMIP 2009]. Trudy ROMIP 2009 [Proc. ROMIP 2009]. Petrozavodsk, 2009. Saint Petersburg, 2009. pp. 122-134
[5] *Chetviorkin I., Loukachevitch N.* Extraction of Russian Sentiment Lexicon for Product Meta-Domain // In Proceedings of COLING 2012
[6] *Joachims T.* Making large-scale support vector machine learning practical. Advances in Kernel Methods: Support Vector Machines / B.Scholkopf, C. Burges, A. Smola (eds.) - MIT Press: Cambridge, MA" – 1998
[7] *Chetviorkin I., Braslavskiy P., Loukachevitch N.* Sentiment Analysis Track at ROMIP 2011