ИССЛЕДОВАНИЕ ПРИМЕНИМОСТИ МЕТОДОВ ТЕМАТИЧЕСКОЙ КЛАССИФИКАЦИИ В ЗАДАЧЕ КЛАССИФИКАЦИИ ОТЗЫВОВ О КНИГАХ

Поляков П. Ю. (pavel@rco.ru),
Калинина М. В. (kalinina_m@rco.ru),
Плешко В. В. (volodia@rco.ru)

ООО «ЭР СИ О», Москва, Россия

В данной работе исследуются различные способы формирования обучающей выборки, методов извлечения классификационных признаков, а также методов построения классификаторов для решения задач классификации отзывов о книгах на 2 (положительный, отрицательный) и 3 (положительный, отрицательный и нейтральный) класса. Показано, что хороший результат можно получить путем применения к рассматриваемым задачам методов тематической классификации. Достигнутые показатели не уступают наилучшим результатам классификации Веб-сайтов и нормативно-правовых документов, полученным участниками семинара РОМИП.

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

RESEARCH ON APPLICABILITY OF THEMATIC CLASSIFICATION METHODS TO THE PROBLEM OF BOOK REVIEW CLASSIFICATION

Polyakov P. Yu. (pavel@rco.ru),
Kalinina M. V. (kalinina_m@rco.ru),
Pleshko V. V. (volodia@rco.ru)

RCO LLC, Moscow, Russian Federation

The paper examines the different approaches to forming the training set, methods for extracting classification features, as well as methods of constructing classifiers regarding the problem of book review sentiment analysis. The tasks were to divide book reviews into 2 groups (positive, negative) and into 3 groups (positive, negative, neutral). Several methods were tested in the solution of the two tasks. It was shown that good results could be obtained by using common document categorization methods. The obtained figures approach the best results of the Web-site and regulatory document classification track achieved by participants of ROMIP seminar. A method for enrichment of classification features within the linguistic approach using evaluative vocabulary dictionaries was proposed. It was established that this method gives a slight improvement in the results for the binary classification. We plan to explore in more detail the possibility of using expert-linguistic approaches to the construction of classification features.

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 соответствующей функции в коммерческих системах мониторинга социальных медиа. Тем не менее для русскоязычного контента до настоящего времени 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 2 (положительный, отрицательный) и 3 (положительный, отрицательный, нейтральный) класса в рамках новых дорожек РОМИП.

Постановка задачи

Участникам была предложена тестовая коллекция, представляющая собой набор отзывов пользователей рекомендательного портала Imhonet.ru 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».

Среди особенностей задачи следует отметить сильный дисбаланс тестовой коллекции в сторону положительных отзывов.

Формирование обучающей выборки

Двое экспертов независимо оценивали тестовую 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. В качестве обучающей выборки в разных прогонах брались как результат оценки одного эксперта, так и множество пересечений оценок обоих экспертов (отзывы, для которых оба эксперта выставили одинаковую оценку). Согласованность оценок экспертов при формировании обучающей выборки достигала 80 %.

Представление документа и извлечение терминов

Исследования проводились в рамках векторной модели представления документов, при которой документ описывается набором выделенных из текста терминов. В работе исследуется возможность обогащения классификационных признаков, полученных автоматическим (базовым) методом, который хорошо зарекомендовал себя в задаче тематической классификации документов [1–4], путем добавления к ним терминов, выделенных в рамках лингвистического подхода с использованием словарей оценочной лексики.

В базовом методе в качестве однословных терминов выделялись все слова документа за исключением служебных частей речи, числительных и дат.

Многословные термины выделялись при помощи алгоритма синтактико-семантического анализа [5] и представляли собой простые именные группы (напр. «глубокая мысль», «классика жанра»). Именные группы были усложнены включением в их структуру конструкций с предлогами в соответствии с моделями управления [6] (напр. «взгляд на мир», «книга для детей»).

Для повышения качества рубрицирования и обогащения набора классификационных признаков был применен лингвистический подход. Путем анализа имеющихся отзывов эксперт выделил атрибуты книги, на которые большинство пишущих обращали внимание. Таким образом, был получен список наиболее значимых для читателей вещей: язык, сюжет, герои, концовка, впечатления от прочтения, автор и т. д. Список данных атрибутов был расширен синонимами (*книга=книженция=опус=чтиво=произведение=сочинение* и т. д.; *конец=концовка=финал=развязка=хэппи-энд*; *герой=персонаж=характер* и т. д.) и гипонимами (*книга=роман=повесть=рассказ=детектив=пьеса=фэнтези=поэма* и т. д.; *автор=писатель=поэт*).

Далее были составлены словари оценочной лексики (прилагательные и глаголы), выражающей положительную, отрицательную или среднюю оценку. Примеры положительной оценки: *бесподобный, великолепный, яркий; запомниться, нравиться, потрясать*. Примеры отрицательной оценки: *бессодержательный, занудный, пошлый; устареть*. Примеры оценки «средне»: *неоднозначный, неровный, средненький, специфический*.

Для лингвистического анализа текста были использованы семантические шаблоны, описывающие возможные синтаксические связи в предложении между группами терминов из получившихся словарей [7]. Шаблон задает лексико-грамматические ограничения на искомую конфигурацию связей между словами в тексте, которые определяются синтаксическим анализатором. На Рисунке 1 приведен семантический шаблон для извлечения оценки книги, которая выражается прилагательным в конструкциях вида: *Книга оказалась достаточно интересной*; *Эти писатели стали культовыми еще в 60-е годы*.

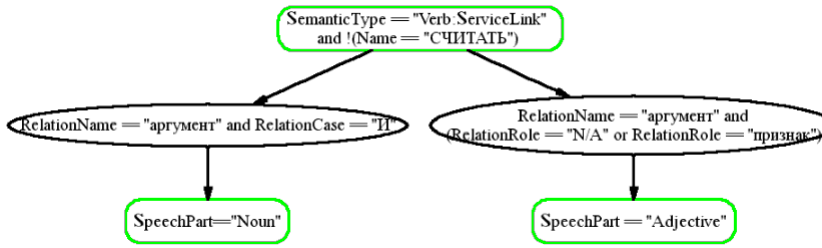

Рис. 1. Пример семантического шаблона для определения оценки книги

В вершинах указываются ограничения на части речи (*SpeechPart="Noun"* — существительное), конкретные слова *(!Name=="СЧИТАТЬ"* — запрещен глагол «считать»), или семантические разряды слов (*SemanticType="Verb:ServiceLink"* — глагол-связка). В эллипсах описываются ограничения на синтактико-семантические

связи между словами: тип связи (*RelationName=="аргумент"*), семантическая роль (*RelationRole=="признак"*), падеж (*RelationCase=="И"* — именительный). Окончательно такой шаблон параметризуется множеством конкретных слов из соответствующих словарей: множеством синонимов и гипонимов для книги и ее атрибутов параметризуется узел с ограничениями *SpeechPart="Noun"*, имеющий роль «Объект оценки» (*EstimatedObject*); множеством оценочных слов параметризуется узел с ограничениями *SpeechPart="Adjective"*, имеющий роль «Оценка» (*QualityAdjective*); узел с ограничением *SemanticType == "Verb:ServiceLink" and !(Name == "СЧИТАТЬ")* параметризуется множеством глаголов-связок за исключением глагола «считать». Выделенные таким образом пары «объект оценки + оценка» использовались в качестве терминов для автоматического определения тональности отзывов. Примеры положительных терминов: *глубокое произведение; книга зацепила; любить книгу; произведение интересное; проглотить книгу; сильная вещь; сюжет захватывает; хорошее чтение; финал неожиданный; язык легкий*. Примеры отрицательных терминов: *дочитать с трудом; испортить настроение; книга бессмысленная; неудачный перевод; скучная книга; сюжет не затягивает; стиль не понравился; читать по диагонали; язык уродливый*.

Методы классификации

В работе исследованы два подхода. В первом подходе для обучения классификатора использовались оценки самих пользователей. По ним строилась линейная регрессионная модель в реализации SVM-Light [8]. 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 [8]. 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 строит множество непересекающихся классов, то есть ставит в соответствие документу ровно в один из заданных классов. Использовалась реализация SVM-Multiclass [9].
- Линейный классификатор, который обучается независимо на классах положительных и отрицательных отзывов в реализации SVM-Light [8], 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 18 прогонов классификации на 2 класса и 24 прогона классификации на 3 класса. Прогоны варьировались:

- по способу формирования обучающей выборки: оценка первого эксперта [expert-1], второго [expert-2], согласованная оценка экспертов (обучающая выборка содержала только документы, которые эксперты оценили одинаково) [expert-and];
- по способу извлечения терминов: автоматический [base], смешанный (обогащение базового набора классификационных признаков в рамках лингвистического подхода с использованием словарей оценочной лексики) [hybrid];
- по методу классификации: регрессия [regression], линейный классификатор на независимые классы [one-per-class], на непересекающиеся классы [multiclass], линейный классификатор, обучающийся на классах положительных и отрицательных отзывов, но применяемый в задаче классификации на 3 класса [2-to-3-class].

Влияние различных подходов на качество классификации мы оценивали с помощью F1-меры [10], сильные и слабые требования к релевантности обозначены далее соответственно F-and и F-or.

Согласно Таблице 1, первый эксперт сформировал немного лучшую по качеству обучающую выборку, чем второй эксперт (лучшую в смысле качества обучения исследуемых методов), хотя различие между ними не превосходит нескольких процентов. Здесь и далее «average» и «maximum» обозначают в таблицах способ обобщения результатов по различным прогонам. В первом случае вычисляется средний результат, во втором случае берется максимальное значение. В частности, в Таблице 1 усреднение берется по различным методам и способам отбора терминов.

Таблица 1. Качество работы классификатора, построенного по разным обучающим выборкам, в задаче классификации на 2 и 3 класса

	average		maximum	
2-class	F-and	F-or	F-and	F-or
expert-1	0.691	0.721	0.723	0.747
expert-2	0.669	0.685	0.721	0.732
expert-and	0.690	0.706	0.723	0.724

	average		maximum	
3-class	F-and	F-or	F-and	F-or
expert-1	0.465	0.508	0.536	0.577
expert-2	0.419	0.449	0.484	0.516
expert-and	0.440	0.474	0.521	0.560

Данные, приведенные в Таблице 2, свидетельствуют о том, что привлечение «продвинутых» лингвистических признаков дает незначительное улучшение результата для задачи классификации на 2 класса и даже немного ухудшает результат в случае 3 классов. Небольшой прирост результата можно объяснить тем, что извлекаемые в базовом методе термины оказались, по сути, эквивалентны большинству созданных семантических шаблонов, за исключением

шаблонов, содержавших глаголы. Ухудшение результатов для задачи классификации на 3 класса, связано с некорректной группировкой терминов для класса нейтральных отзывов с использованием словарей оценочной лексики. Словарь оценочной лексики для нейтральных отзывов содержал только слова, соответствующие средней оценке, например, «средненький», «специфический», «сносный», «читабельный», «неровный». Но эта категория содержит согласно постановке задачи помимо нейтральных отзывов также документы, в которых книгу и хвалят, и ругают одновременно. В результате получилась низкая полнота покрытия языкового материала лингвистическими шаблонами документов данной рубрики.

Таблица 2. Качество работы классификатора в зависимости от способа извлечения терминов, описывающих документ, в задаче классификации на 2 и 3 класса

	average		maximum	
2-class	F-and	F-or	F-and	F-or
base	0.679	0.699	0.709	0.727
hybrid	0.688	0.709	0.723	0.747

	average		maximum	
3-class	F-and	F-or	F-and	F-or
base	0.445	0.481	0.536	0.577
hybrid	0.437	0.473	0.512	0.554

Согласно Таблице 3, в задаче классификации на 2 класса наилучший результат был достигнут при помощи метода one-per-class. Остальные методы незначительно ему уступают. В задаче классификации на 3 класса лучшим оказался метод regression. Ему немного уступает one-per-class. Следует отметить, что наименьший разброс результатов при варьировании способов отбора классификационных признаков и формирования обучающей выборки дает метод one-per-class, что свидетельствует о меньшей чувствительности данного метода к качеству входных данных по сравнению с другими методами.

Таблица 3. Качество работы классификатора в зависимости от метода обучения в задаче классификации на 2 и 3 класса

	average		maximum	
2-class	F-and	F-or	F-and	F-or
regression	0.674	**0.715**	0.701	0.727
one-per-class	**0.699**	**0.715**	**0.723**	**0.747**
multiclass	0.677	0.682	**0.723**	0.735
–	–	–	–	–

	average		maximum	
3-class	F-and	F-or	F-and	F-or
regression	**0.494**	**0.529**	**0.536**	**0.577**
one-per-class	0.487	0.525	0.512	0.554
multiclass	0.355	0.375	0.418	0.453
2-to-3-class	0.429	0.479	0.459	0.507

На Рисунках 2 и 3 показаны сводные результаты по всем участникам, участвовавшим в дорожках классификации отзывов о книгах на 2 и 3 класса, соответственно. Наши прогоны обозначены темной заливкой и упорядочены на рисунках по возрастанию меры F-or, прогоны других участников обозначены более светлой заливкой и упорядочены по убыванию F-or. Использование разметки экспертом 1 дало равномерно более высокий результат по сравнению

с разметкой эксперта 2 или их согласованной оценкой (возможно, это свидетельствует о том, что эксперт 1 имеет больше опыта). Поэтому наши прогоны на Рисунках 2 и 3 приводятся только для классификаторов, построенных по обучающей выборке первого эксперта. Расшифровка идентификаторов этих прогонов дана в Таблице 4.

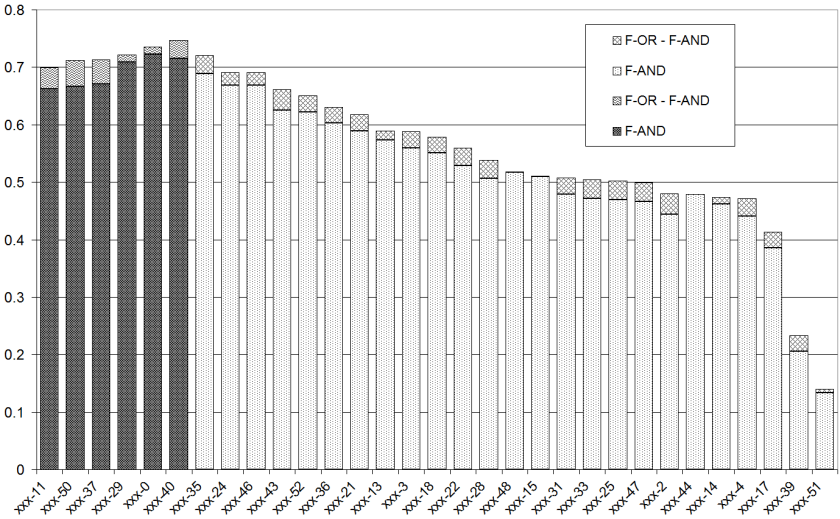

Рис. 2. Результаты оценки прогонов дорожки классификации на 2 класса

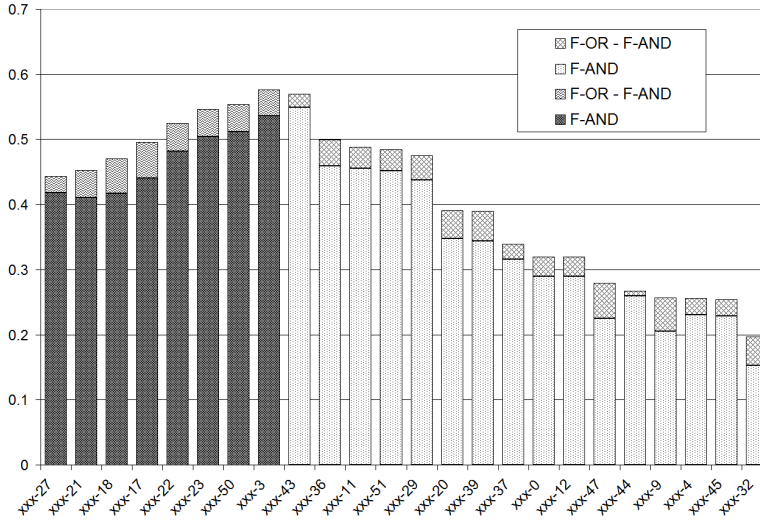

Рис. 3. Результаты оценки прогонов дорожки классификации на 3 класса

Таблица 4. Расшифровка идентификаторов наших прогонов. Номер строки определяет тип метода классификации, номер столбца — способ извлечения классификационных признаков

2-class	base	hybrid
regression	xxx-50	xxx-37
one-per-class	xxx-11	xxx-40
multiclass	xxx-29	xxx-0
–	–	–

3-class	base	hybrid
regression	xxx-3	xxx-23
one-per-class	xxx-22	xxx-50
multiclass	xxx-27	xxx-21
2-to-3-class	xxx-18	xxx-17

Заключение

Проведена апробация ряда методов решения задач классификации отзывов о книгах на 2 и 3 класса. Установлено, что использование методов, обычно применяемых для решения задачи тематической классификации, позволяет получить достаточно высокое качество, сопоставимое с наилучшими результатами дорожек РОМИП по классификации Веб-сайтов и нормативно-правовых документов. Предложен метод обогащения классификационных признаков в рамках лингвистического подхода с применением словарей оценочной лексики, который дает незначительное улучшение результата при классификации на 2 класса. В дальнейшем мы планируем более детально исследовать

возможность применения экспертно-лингвистических подходов для построения классификационных признаков.

Литература

1. *Плешко В. В., Ермаков А. Е., Голенков В. П., Поляков П. Ю.* RCO на РОМИП 2005 // Труды третьего российского семинара РОМИП'2005. (Ярославль, 6 октября 2005г.). — Санкт-Петербург: НИИ Химии СПбГУ — 2005 — с. 106–124.
2. *Поляков П. Ю., Плешко В. В.*, RCO на РОМИП 2006 // Труды четвертого российского семинара РОМИП'2006. (Суздаль, 19 октября 2006г.). — Санкт-Петербург: НУ ЦСИ — 2006 — с. 72–79.
3. *Плешко В. В., Поляков П. Ю.*, RCO на РОМИП 2008 // Труды РОМИП 2007–2008. (Дубна, 9 октября 2008г.). — Санкт-Петербург: НУ ЦСИ, 2008 — с. 96–107.
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. *Ермаков А. Е.* Значимость элементов текста в свете теории синтаксической парадигмы // Русский язык: исторические судьбы и современность. II Международный конгресс исследователей русского языка. Труды и материалы. — Москва: МГУ — 2004.
6. *Ермаков А. Е.* Эксплицирование элементов смысла текста средствами синтаксического анализа-синтеза // Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции Диалог'2003 (Протвино, 11–16 июня, 2003 г.). — Москва, Наука, 2003
7. *Ермаков А. Е.* Извлечение знаний из текста и их обработка: состояние и перспективы // Информационные технологии. — 2009. — N 7.
8. *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.
9. *Joachims T., Finley T., Yu Chun-Nam.* Cutting-Plane Training of Structural SVMs // Machine Learning Journal. — 2009, V.77, No.1, pp.27–59.
10. *Chetviorkin I., Braslavskiy P., Loukachevitch N.* Sentiment Analysis Track at ROMIP 2011.

References

1. *Pleshko V. V., Ermakov A. E., Golenkov V. P., Polyakov P. Yu*. RCO at RIRES 2005 [RCO na ROMIP 2005]. Trudy 3 rossijskogo seminara ROMIP'2005 [Proc. 3rd Russian Information Retrieval Seminar ROMIP'2005]. Yaroslavl, 2005. Saint Petersburg, 2005. pp. 106–124.
2. *Polyakov P. Yu., Pleshko V. V.* RCO at RIRES 2006 [RCO na ROMIP 2006]. Trudy 4 rossijskogo seminara ROMIP'2006 [Proc. 4th Russian Information Retrieval Seminar ROMIP'2006]. Suzdal, 2006. Saint Petersburg, 2006. pp. 72–79.
3. *Pleshko V. V., Polyakov P. Yu.* RCO at RIRES 2008 [RCO na ROMIP 2008]. Trudy ROMIP 2007–2008 [Proc. ROMIP'2007–2008]. Dubna, 2008. Saint Petersburg, 2008. pp. 96–107.
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. *Ermakov A. E.* The meaning of text elements from the point of view of syntactic paradigm theory [Znachimost' elementov teksta v svete teorii sintaksicheskoj paradigmy]. Ruskij jazyk: istoricheskie sud'by i sovremennost'. II Mezhdunarodnyj congress issledovatelej russkogo jazyka. Trudy I materialy [Russian Language: its Historical Destiny and Present State: The Second International Congress of Russian Language Researchers. Proceedings and materials]. Moscow, 2004.
6. *Ermakov A. E.* Explication of meaning of the text elements by means of syntactic analysis-synthesis [Eksplitsirovanie elementov smysla teksta sredstvami sintaksicheskogo analiza-sinteza]. Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2003" [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2003"]. Protvino, 2003. Moscow, 2003
7. *Ermakov A. E.* (2009) Extracting knowledge from text and processing: Status and Prospects [Izvlechenie znanij iz teksta i ih obrabotka: sostojanie i perspektivy]. Informatsionnye tehnologii [Information Technologies]. No.7.
8. *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.
9. *Joachims T., Finley T., Yu Chun-Nam* (2009). Cutting-Plane Training of Structural SVMs. Machine Learning Journal, V.77, No.1, pp.27–59.
10. *Chetviorkin I., Braslavskiy P., Loukachevitch N.* Sentiment Analysis Track at ROMIP 2011.