ПРЕДВАРИТЕЛЬНАЯ ОБРАБОТКА В АНАЛИЗЕ ТОНАЛЬНОСТИ РЕЦЕНЗИЙ

Д.К. Волчик
Кафедра вычислительной техники, Национальный исследовательский университет
«Московский энергетический институт»
Смоленск, Российская Федерация
E-mail: diana-lex@bk.ru

В рамках данной работы была разработана система анализа тональности текстовых документов на основе Наивного байесовского классификатора. В качестве исходных данных были собраны рецензии кинофильмов с интернет-ресурса kinopoisk.ru. Для исследования влияния предварительной обработки текстовых данных на точность были разработаны две модификации системы, с каждой из которых был проведен ряд тестов. В статье представлены результаты исследования и сделаны выводы.

Введение

Анализ тональности – это область компьютерной лингвистки, которая занимается автоматизированным выявлением и анализом эмоций, чувств автора по отношению к объектам, речь о которых идет в текстовых документах [1]. Целью такого анализа является классификация такого отношения автора текста.

В рамках данной работы были разработаны программные средства анализа тональности рецензий кинофильмов, полярность мнений которых и требовалось определить. Выбрана двухуровневая система оценки: позитивная – негативная.

I. Описание системы

Анализ тональности – достаточно новая, перспективная для изучения и развития, область компьютерной лингвистики. В настоящее время существует не так много систем, способных определять тональность текстовых документов, и большинство из них ориентированы на иностранные языки. Разработанная система работает с русскоязычными текстами, имеет простой алгоритм и хорошую точность работы. Кроме того, была реализована система, позволяющая выделять данные из сети Internet вместе с необходимыми признаками для работы классификатора [2]; проведен анализ этих данных - сравнение разных подходов к предварительной обработке текста.

Процесс создания реализованной системы можно в общем виде представить как последовательность следующих действий:

1. сбор коллекции документов для обучения классификатора;
2. выделение для каждого документа из обучающей коллекции признака, по которому будет обучаться классификатор;
3. обучение классификатора;
4. использование полученной модели.

Для решения задач обработки естественного языка, в частности анализа тональности, необходимо наличие хорошей базы (корпуса) текстовых данных (в данном случае – рецензии кинофильмов), с помощью которой можно осуществлять построение и проверку математических моделей. Был собран корпус из 11159 рецензий вместе с оценками [2], который затем был разбит на следующие наборы:

- тренировочный набор «training» – на этом наборе происходит обучение классификатора;
- контрольный набор «validation» – на этих данных тестируется обученный классификатор, а полученные результаты контролируются за счет сравнения полученных оценок с достоверными;
- тестовый набор «test» – набор используется при тестировании системы.

В данной работе был использован классификатор Байеса (от англ. Naïve Bayes Classifier – Наивный Байесовский классификатор) – это простой вероятностный классификатор, который основывается на применении теоремы Байеса о сильных (наивных) независимых предположениях (присутствие или отсутствие какого-либо признака не связано с наличием или отсутствием любого другого признака) [3,4]. Метод представления входных для него данных - «мешок слов». Под «мешком слов» понимаются словари – это набор слов, выделенных из рецензий соответствующего типа.

II. Обоснование актуальности

Результаты анализа тональности могут зависеть от многих факторов в силу особенностей языка и обработки текстовых данных. Например, при выделении слов из рецензии учитывается регистр букв, а, следовательно, одно и то же слово, начинающееся на строчную и прописную буквы, будет выделяться как два разных слова. Если в рецензиях избавиться от верхнего регистра, то изменятся размеры словарей, и значения вероятностей при вычислении будут другими, что повлияет на результат работы классификатора. Кроме того, как и в любых текстовых данных, рецензии содержат слова без эмоцио-

нальной нагрузки (так называемые, стоп-слова). Например, к таким словам относятся союзы, частицы, местоимения, а также большинство наиболее часто встречающихся слов.

Отсюда вытекает необходимость исследования влияния этих факторов на точность результатов анализа тональности.

III. Результаты

Были разработаны две модификации метода анализа тональности: с учетом и без учета регистра при одних и тех же входных данных. И для каждой из модификаций были проведены тесты на контрольном (validation) и тестирующем (test) наборах для разной величины списка стоп-слов. Полученные результаты приведены в таблице (см. табл. 1).

Из полученных результатов можно сделать вывод, что точность системы хорошая - 75%. Кроме того, наличие регистра в исходных данных оказывает некоторое влияние на количество ошибочно определенных рецензий (1,5 – 2%). Но вне зависимости от того, учитывается ли регистр или нет, система показала, что если из исходных данных «выбрасывать» стоп-слова, то точность работы системы уменьшается.

Исходя из полученных значений, можно отметить следующую зависимость: чем большее количество стоп-слов исключается из словарей, тем точность работы становится меньше (см рис. 1).

Данная работа посвящена исследованию влияния таких факторов, как наличие регистра и списка наиболее часто встречающихся слов в исходных данных на точность работы системы анализа тональности рецензий кинофильмов. В работе использован Наивный Байесовский классификатор, который, несмотря на простоту, дает довольно хорошие результаты: точность вычислений для контрольного набора данных с учетом регистра составила 72%, а без учета регистра – 75%. При исключении наиболее часто встречающихся слов из словарей точность работы системы уменьшается.

1. Course of Natural Language Processing [Electronic resource] / D. Jurafsky. – Stanford University, 2008. – Date of access: www.coursera.org
2. Волчик, Д.К. Сбор текстовых данных из сети INTERNET в задачах обработки естественного языка // Сборник трудов десятой международной научно-технической конференции студентов и аспирантов «Информационные технологии, энергетика и экономика». В 3 т. Т 2. – Смоленск: "Универсум 2013. - С.108-111.
3. Волчик, Д.К. Анализ тональности русскоязычных отзывов с помощью наивного байесовского классификатора // Сборник трудов 11-ой международной научно-технической конференции студентов и аспирантов «Информационные технологии, энергетика и экономика». В 3 т. Т.2 - Смоленск, 2014.- С.220-223
4. Сегеран Т. Программируем коллективный разум. – Пер. с англ. – Спб:Символ-Плюс, 2008. – 369 с., ил.

Таблица 1. Точность работы системы

Длина списка стоп-слов	Validation без учета регистра, %	Test без учета регистра, %	Validation с учетом регистра, %	Test с учетом регистра, %
0	75,03	74,14	73,59	72,53
10	74,49	73,79	73,06	72,13
30	74,04	73,74	72,38	71,46
100	73,73	72,89	71,94	70,92
1000	68,04	66,03	66,11	64,69

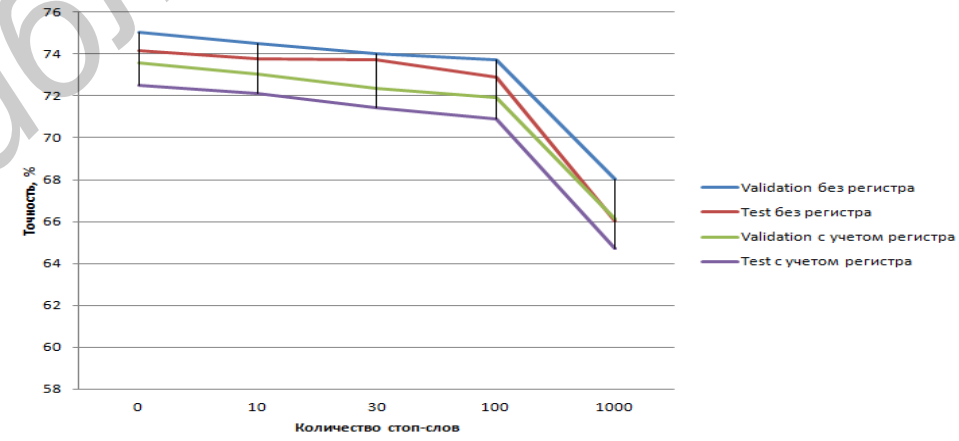

Рис. 1 – Точность работы модификаций системы