Министерство образования и науки Российской Федерации
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»

Кафедра дискретной математики и
информационных технологий

Анализ тональностей текста комментариев в социальных сетях

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

Студента 5 курса 521 группы
направления 09.03.01 «Информатика и вычислительная техника» факультета
компьютерных наук и информационных технологий
Киселева Романа Александровича

Научный руководитель
к. ф.-м.н., доцент _____ А.Н. Савин

Заведующий кафедрой
к. ф.-м.н., доцент _____ Л.Б. Тяпаев

Саратов 2017 год

Введение. Социальные сети с самого их появления притягивают пользователей возможностью делиться своими мыслями и впечатлениями где бы они ни находились. С каждым годом количество пользователей социальных сетей увеличивается, люди заводят все больше аккаунтов на разных ресурсах, чаще высказывают свое мнение по поводу товаров в различных магазинах, постоянно пишутся свежие обзоры на новинки гаджетов, на форумах обсуждаются различные события, происходящие в мире.

Таким образом в социальных сетях формируется огромнейшая база информации по различным аспектам. Знание мнения людей по поводу товаров или событий может дать огромное преимущество, к примеру, маркетологам, государственным организациям, крупным корпорациям и т.д.

По комментариям людей в социальных сетях можно производить анализ высказываний пользователей по тем или иным вопросам. Как пример – создатели проекта SportSense [1] разработали алгоритмы, способные определить уровень взволнованности спортивных болельщиков с помощью анализа их сообщений в Twitter, а это в свою очередь дает возможность отслеживать ключевые моменты проведения Национальной Футбольной Лиги США в реальном времени.

Целью дипломной работы является обзор классификации информации в интернете с точки зрения анализа тональности текста, изучение инструментов для обработки текста в Python, а также разработка программы для оценки эмоциональной окраски текста.

В главе 1 «Анализ эмоциональной окраски текста» содержится обзорная часть по классификации информации в интернете. В ней рассматриваются виды высказываний, информация о субъективности текста, а также уровнях, на которых производится определение эмоциональной окраски текста.

В главе 2 «Методы обучения» описаны методы обучения без учителя, то есть метод спонтанного обучения и обучение с учителем на примере наивного байесовского классификатора, метода опорных векторов и словарного метода.

Глава 3 «Разработка анализатора тональности текста» описывает создание программы для определения эмоциональной окраски текста, включающую в себя систему исправления опечаток и токенизацию текста в рамках его подготовки к определению тональности текста.

Существует большое количество задач, для которых необходим анализ эмоционально окрашенной лексики в текстах. Для их решения используются методы, называемые анализом тональности текста (Sentiment Analysis), или же анализом эмоциональной окраски текста. Сам по себе анализ тональности текста относится к задачам компьютерной лингвистики и является подзадачей получения и обработки информации.

1 Анализ эмоциональной окраски текста. Анализ эмоциональных тональностей текста – это автоматический анализ мнений и эмоционально окрашенной лексики, имеющиеся в тексте.

С точки зрения анализа эмоциональной окраски текста принято считать, что текстовая информация в интернете делится на две группы [2]:

1) Факты.
2) Мнения.

Факты с точки зрения анализа тональности текста не представляют интереса, поэтому не рассматриваются. Ключевым понятием являются мнения.

Мнения в свою очередь делятся на два типа:

1) Простое мнение.
2) Сравнение.

1.1 Мнение. Мнение содержит в себе изложение мысли автора о конкретном объекте. Оно может высказываться прямо, либо неявно. В обоих случаях мнение обычно имеет положительную или отрицательную окраску.

Выделяют три вида эмоциональной окраски мнений:

1) Позитивная.
2) Нейтральная.
3) Негативная.

1.2 Сравнение. Сравнение – это второй тип мнений. Его можно разделить на три категории [4]:

1) Сравнение аспектов объектов в пользу одного из них.
2) Приравнивание аспектов различных объектов.
3) Превосходство одного из объектов по отношению к другим.

1.3 Субъективность текста. Субъективность – термин, часто встречающийся в анализе эмоциональной окраски текста.

Согласно [2] предложение объективно, если оно выражает фактическую информацию, касающуюся объекта, и субъективно, если оно выражает чьи-либо личные предположения.

Предложения, которые содержат мнения, обычно субъективны, поэтому анализ исходного текста на субъективность информации, содержащейся в нем, зачастую является подзадачей по определению тональности текста.

1.4 Уровни определения текста. Определение полярности текста обычно рассматривается на двух уровнях:

1) На уровне документа.
2) На уровне предложения.

Задача на уровне документа состоит в определении полярности документа в целом. При этом текст документа может одновременно содержать предложения как с негативной, так и с позитивной эмоциональной окраской.

2 Методы обучения. Существуют два вида машинного обучения:

1) Обучение с учителем.
2) Обучение без учителя.

Методы обучения с учителем дают информацию о принадлежности объекта к какому-либо классу, на основании уже размеченного набора данных, полученных с помощью анализа тренировочных данных.

Обучение без учителя — один из способов машинного обучения, при котором испытуемая система спонтанно обучается выполнять поставленную задачу без вмешательства со стороны экспериментатора. Его так же можно назвать методом самообучения или спонтанным обучением.

2.1 Наивный байесовский классификатор. Наивный байесовский классификатор – [6] это простой классификатор, который работает с условными вероятностями, предполагая, что слова в предложении независимы.

Этот классификатор хорошо показывает себя в решении задачи классификации текстов [3]. Сперва необходимо выбрать закон, по которому, как предполагается, распределены данные. Затем по размеченным примерам вычисляются параметры этого распределения, которые в дальнейшем используются для разметки.

2.2 Метод опорных векторов. Метод опорных векторов (SVM) [7] работает по принципу разбиения пространства на подпространства, которые соответствуют классам.

В данном методе выбираются признаки, по которым приведенные примеры преобразуются в числовые векторы. В дальнейшем работа производится уже с готовыми векторами и пространством, в котором они располагаются.

При обучении задача метода – преобразование пространства таким образом, чтобы нашлись гиперплоскости, разделяющие примеры из разных классов обучающей выборки.

Предсказание делается по принципу попадания вектора данного примера в часть пространства найденных гиперплоскостей.

2.3 Словарные методы в определении окраски. Существуют методы, основанные на использовании словаря эмоционально окрашенных слов и словаря символов, обозначающих эмоции [5]. В словарных методах каждое слово обладает весом, характеризующим его эмоциональную окрашенность.

Основной проблемой словарных методов считается процесс составления словаря: чтобы получить метод, классифицирующий документ с высокой точностью, термины словаря должны иметь вес, адекватный предметной области документа. Например, слово «большой» по отношению к объему памяти жесткого диска является положительной характеристикой, но отрицательной по отношению к размеру мобильного телефона. Поэтому

словарные методы больше подходят для анализа узко специализированных текстов.

3 Разработка анализатора тональности текста. Полный код программы можно посмотреть на приложенном к дипломной работе диске.

Задача реализации анализатора эмоциональной окраски текста показана на рисунке 1. Она разбита на следующие подзадачи:

1) Создание языковой модели.
2) Исправление ошибок на основе составленной языковой модели.
3) Токенизация полученного текста.
4) Анализ подготовленного текста.

3.1 Входные данные. На вход программе подаются следующие данные:

1) Обучающий текст – большой файл для формирования языковой модели.
2) Словарь слов с оценкой эмоциональной окраски, таких как «good», «bad» и другие.
3) Словарь слов сравнения, например «better», «worse».
4) Список ключевых слов, оценку по которым необходимо выяснить, например «Applephone; iPhone». Если ключевые слова не были введены, производится общая оценка текста.
5) Исходный текст для анализа.

Импортированный исходный текст можно просмотреть, отредактировать вручную и, при желании, сохранить файл.

Разделение словарей позволяет наиболее точно выбрать оценку слов под конкретную тематику текста.

3.2 Система исправления ошибок и опечаток. Проверяемые тексты могут содержать как грамматические ошибки, так и опечатки. Неверное

исправление опечатки может в корни изменить оценку суждения, например слово «btter» может быть как «butter» (масло), так и «better» (лучше). Поэтому для решения задачи исправления ошибок было решено использовать стохастический подход, основанный на предсказании вероятности появления того или иного слова.

Для начала с помощью тренировочного текста на основе наивного байесовского классификатора создается языковая модель, в которой записана частота появления каждого слова в тренировочном тексте.

Далее сверяется наличие каждого слова исходного текста в языковой модели. Если слово найдено, то оно записывается в массив проверенных слов. Если такого слова нет, то создаются все возможные варианты этого слова с изменением одного и двух символо путем замены одного символа другим, удалением символа, транспозицией и вставкой нового символа.

Затем сгенерированные слова проверяются на наличие в языковой модели и выбирается наиболее вероятное.

$$\mathrm{Argmax}_B \, \mathrm{P}(A|B) \, \mathrm{P}(B),$$

где P(B) – вероятность появления слова B, задающаяся моделью языка, определяющуюся с помощью обучающего текста;

P($A|B$) – вероятность опечатки словом A вместо слова B;

Argmax_B – оператор перебора всех возможных B в поисках наиболее вероятного варианта.

При отсутствии всех сгенерированных слов в языковой модели в массив проверенных слов записывается исходное слово.

3.3 Токенизация текста. Токенизация - выделение в тексте слов, чисел, и иных токенов, в том числе, например, нахождение границ предложений.

Массив слов с исправлениями уже можно пытаться проанализировать, однако в нем все еще много стоп-слов, то есть слов, не передающих важную

информацию, таких как предлоги «a», «the», «is» и другие. Это мешает составлению корректных компактных n-грамм слов.

Компактность определяется следующим образом:

Пусть f – n-грамма из n слов, s– предложение, содержащее все слова из f (возможно расположенные не подряд). Если расстояние между любыми двумя словами, смежными в f, в предложении s составляет не более чем три слова, то f компактна в данном конкретном предложении.

3.4 Оценка подготовленного текста. После того как исходный текст проверен на наличие ошибок, можно произвести оценку эмоциональной окраски текста.

Для оценок в программе задается словарь эмоционально окрашенных слов, содержащий в себе как положительно, так и отрицательно окрашенные лексические единицы с их оценками, например:

- Good; 3
- Bad; -3
- Great; 5

Помимо этого, в программе задан список слов сравнения, таких как better, worse и подобные, задающихся так же с положительной, либо отрицательной оценкой, помогающие оценивать мнения со сравнениями. При необходимости рассмотрения специфической темы, пользователь может расширить этот список путем загрузки соответствующего теме словаря.

Так же в программе заданы слова усиления эмоциональной окраски, такие как very, slightly, overly и другие, которые будут добавлять, либо вычитать 1 балл (в зависимости от полярности слова, перед которым они стоят). Таким образом «Bad» будет иметь оценку -4, а «Very bad» будет иметь оценку -5.

Слова отрицания такие как «not» инвертируют знак усиления оценки, стоящего после них и если полученная оценка меньше нуля.

Если в предложении есть сравнительное слово, например «than», то рассматривается наличие ключевого слова после, либо перед словом сравнения в пределах компактности n-граммы, то есть трех ближайших слов. Так, например, покрывается результат предложения «HTC better than famous Samsung» с ключевым словом «Samsung».

Так же рассматриваются варианты нахождения в пределах трех слов наличие слов усиления, инвертирующих оценку слов и сравнительных слов.

Таким образом в программе были установлены правила для оценки языковых конструкций. На основе этих правил и словарей оценки производится анализ тональностей исходного текста.

Результатом анализа является сумма оценок предложений, поделенная на количество предложений с мнениями, а также текстовое значение:

1) Positive.
2) Negative.
3) Neutral.

Заключение. В ходе дипломной работы проведен обзор текстовых особенностей сообщений в социальных сетях в контексте разработки методов анализа их эмоциональной окраски и описаны методы анализа текста.

Разработана схема, позволяющая оценивать входной текст на предмет эмоциональной окраски. Реализован алгоритм исправления ошибок в каждом слове исходного текста, основанный на обучаемом предсказании вероятности слов.

Возможность задания специфических обучающих текстов для создания языковой модели, а так же задание эмоционально окрашенных слов, подходящих под конкретную тематику, позволяют анализатору оценивать тональность даже у узко специализированных текстов.

Таким образом все цели дипломной работы были успешно выполнены.

Список использованных источников

1 SportSense: Real-Time Detection of NFL Game Events from Twitter [Электронный ресурс] URL:

https://arxiv.org/abs/1205.3212 (дата обращения 15.03.2017 г. Заколовок с экрана. Язык английский)

2 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

3 Christopher D Manning и Hinrich Schütze. Foundations of statistical natural language processing. MIT press, 1999

4 Nitin Jindal and Bing Liu. Mining Comparative Sentences and Relations./ Proceedings of 21st National Conference on Artificial Intelligence, AAAI '06. Boston, MA. July 2006

5 Kerstin Denecke Using SentiWordNet for multilingual sentiment analysis/ IEEE 24th International Conference on Data Engineering Workshop. 2008 pp: 507-512

6 Kevin P. Murphy. Machine Learning: A Probabilistic Perspective (Adaptive Computation and Machine Learning series). The MIT Press, 2012. ISBN: 0262018020

7 Simon Tong и Daphne Koller. “Support vector machine active learning with applications to text classification”. B: The Journal of Machine Learning Research 2 (2002), с. 45—66