УДК 004.912

Ю.В. Рубцова

РАЗРАБОТКА И ИССЛЕДОВАНИЕ ПРЕДМЕТНО НЕЗАВИСИМОГО КЛАССИФИКАТОРА ТЕКСТОВ ПО ТОНАЛЬНОСИ

Рубцова Ю.В. **Разработка и исследование предметно независимого классификатора текстов по тональности.**

Аннотация. В статье представляется метод построения классификатора для классификации текстов по тональности на два и на три класса (положительные и негативные; положительные, нейтральные и негативные тексты). Представляются результаты экспериментов, показывающие высокую точность работы метода при классификации на два класса независимо от предметной области к которой принадлежит текст. Эффективность представленного метода подтверждается экспериментами на текстовой коллекции блогов с разметкой по оценочной тональности семинара РОМИП-2012. Для оценки используются метрики: precision, recall, accuracy и F-меры. Значение F-меры для предлагаемого метода при классификации на 2 класса составляет 93%. Помимо блоговой коллекции РОМИП-2012, используются коллекция новостей и коллекция текстов социальных сетей.

Ключевые слова: анализ тональности текстов, машинное обучение, классификация текстов, автоматическая классификация, извлечение классификационных признаков.

Rubtsova Y.V. **Research and Development of Domain Independent Sentiment Classifier.**

Abstract. The paper presents the method of constructing a sentiment classifier on two and three classes (positive and negative, positive, neutral and negative texts). It is also presents the results of experiments. Results show the high accuracy of the proposed method on texts, which are not belong to any pre specified domains. The effectiveness of the presented method is confirmed by experiments' results on the text collection of blogs from ROMIP 2012 seminar. It was used following metrics for classifier evaluation: precision, recall, accuracy and F-measure. The value of F-measure of the proposed method for classification into 2 classes is up to 93%. In addition to blog collection ROMIP 2012 for experiments were used a collection of news and a collection of short-texts from social networks.

Keywords: sentiment analysis, machine learning, text classification and categorization, feature extraction.

1. Введение. Практически каждый второй россиянин старше 18 лет является активным пользователем Интернета, а более 80% из них имеют хотя бы один аккаунт в социальных сетях. Социальные сети становятся объектом пристального внимания социологов, психологов, маркетологов и PR-специалистов. Огромный объем информации, появляющийся в социальных сетях, приходится регулярно обрабатывать и классифицировать для того, чтобы иметь возможность решать поставленные перед специалистами задачи. Одна из таких задач – это поиск отзывов и упоминаний и классификация найденных текстов по тональности. Практическая ценность анализа тональности включает, но не ограничивается следующими примерами:

– при запуске нового продукта, компании могут быстро узнать, как покупатели оценивают этот продукт, нужно ли что-то исправить в продукте или рекламных материалах;

– правительство сможет отследить реакцию населения на новый закон, уточнения, заявление;

– организаторы мероприятий (напр. конференций) могут собрать отзывы участников в социальных сетях и оценить как прошло мероприятие, понравилось оно или нет;

– IT компании могут разработать эффективную систему поддержки пользователей, учитывая вопросы и отзывы пользователей;

– проведение маркетинговых исследований: изучение потребительских предпочтений, измерение степени удовлетворения потребностей потребителей, определение эффективности распространения продуктов или услуг;

– финансовые рынки. В работе [1] говорится, что существует множество новостей, статей в блогах и сообщений в твиттере о каждом акционерном обществе. Система автоматического анализа тональности может использовать эти источники для извлечения отзывов, что может стать основой для автоматической торговой системы.

В задаче классификации текстов по тональности можно выделить несколько подзадач:

– Классификация на 2 класса: положительный и отрицательный;

– Классификация на 3 класса: положительный, нейтральный и отрицательный;

– Классификация на 5 классов;

– Классификация на более, чем 5 классов, например, классификация по 10-бальной шкале.

Задачу классификации отзывов на два класса достаточно успешно решают как с помощью словарей и правил, так и с помощью машинного обучения. При классификации узкотематических текстов на два класса, точность классификаторов, основанных на униграммах, превышает 82% [2]. При решении задачи классификации более, чем на два класса, точность классификации существенно снижается – это связанно с субъективным восприятием информации: то, что один человек считает «позитивным», другой может отнести к «нейтральному» или даже «склонному к негативному». В статье [3] авторы показывают, что при разделении текстов на большое количество классов, даже человек показывает низкую точность классификации – точность снижается до 55%.

В данной работе предлагается новый метод автоматического извлечения и взвешивания признаков на основе размеченной коллекции коротких сообщений платформы Твиттер. Коллекция была подготовлена

автором статьи для решения задачи тренировки классификатора текстов по тональности на 2 и на 3 класса. Особое внимание в статье уделяется описанию коллекции, из которой извлекаются списки слов – униграммы. Для каждого слова проводится морфологический анализ и рассчитывается набор статистических характеристик.

2. Подготовка коллекций для извлечения униграмм. Машинное обучение с учителем показывает наиболее точные результаты при классификации текстов по сравнению с другими методами, такими как машинное обучение без учителя и машинное обучение, основанное на правилах и словарях. Для обучения классификатора в качестве вектора признаков, на которые раскладывается документ, используют униграммы, биграммы или -граммы. В задачах обработки текста на естественном языке популярно [2-5] представление документов в виде N-грамм, где N-граммы — последовательности слов или символов длины n. Для n=1 такая последовательность состоит из одного слова и называется униграммой, для n=2 такая последовательность называется биграммой, и т.д. В большинстве предыдущих работ [2-5] для извлечения N-грамм использовались тренировочные коллекции текстов, принадлежащих к определенной предметной области: рецензии на фильмы; отзывы на товары; коллекция новостей, собранная с одного сайта. В этих случаях, на тестовых коллекциях, классификаторы показывали высокую точность как для моделей, основанных на униграммах [2], так и на N-граммах [6]. Однако, подобные классификаторы не пригодны для классификации текстов по тональности, принадлежащих к другим предметным областям. Качественные характеристики оценки одной предметной области могут не выражать тонального отношения или вовсе не встречаться в другой предметной области. Так, например, негативная характеристика «скучный» (сюжет) выражающая отношение к фильму, может не встречаться в предметной области «цифровые фотоаппараты». Поэтому, несмотря на наличие размеченных по тональности узкотематических коллекций текстов на русском языке, они не подходят для построения универсального словаря униграмм для задачи классификации текстов по тональности, который будет показывать удовлетворительные результаты независимо от предметной области.

В статье выдвигается и успешно подтверждается гипотеза, что для извлечения n-грамм можно использовать достаточно представительный, однородный, сбалансированный корпус коротких сообщений, собранных на основе платформы Твиттер. Термин «достаточно представительный корпус» означает, что добавление новых сообщений к коллекции повлечет за собой добавление сравнительно небольшого числа новых терминов – это позволит

извлечь достаточно полный набор -грамм для классификации текстов, принадлежащих разным предметным областям.

В качестве платформы для сбора корпуса был выбран Твиттер потому что:

1.Во-первых, пользователи твитера часто выражают субъективное и эмоционально окрашенное мнение о чем-либо;

2.Для выражения эмоций, пользователи используют живой, разговорный язык, который может содержать сленг и ненормативную лексику, усиливающие тональность сообщений;

3.При написании сообщений, пользователи могут допускать широко распространённые ошибки, которые исправляются редакторами новостных изданий, но которые нужно учитывать при классификации по тональности текстов из Интернета (например, блогов или сайтов отзывов на товары).

С помощью Streaming API twitter [7] была собрана коллекция текстов, состоящая из около 15 миллионов коротких сообщений, на основе которой с помощью метода [8] и предложенной автором фильтрации [9] был сформирован сбалансированный корпус, состоящий из следующих коллекций:

- коллекция положительных сообщений 114 991 записей;
- коллекция негативных сообщений 111 923 записей;
- коллекция нейтральных сообщений 107 990 записей.

Чтобы удостовериться в однородности исследуемых коллекций воспользуемся законом Ципфа [10], который описывает распределение частот слов в естественном языке, если все слова достаточно длинного текста упорядочить по убыванию частоты их использования, то частота n-го слова в таком списке окажется приблизительно обратно пропорциональной его порядковому номеру n. Например второе по используемости слово встречается примерно в два раза реже, чем первое, третье — в три раза реже, чем первое, и т. д. Проверим предположение, что высокая частота встречаемости слов, таких как “и”, “я”, “в” в разных коллекциях (позитивной и негативной) должна быть сопоставимо одинакова. В то же время более редкие слова вроде “ассоциация”, “фактор” должны встречаться сопоставимо реже независимо от принадлежности коллекции к положительному или негативному классу [8]. На рисунке 1 для каждой из коллекции (позитивной и негативной) показано распределение нормированной частоты встречаемости слов в зависимости от количества твитов, в котором это слово встречается. Количество твитов, в которых встречается слово, отложено на оси x. Видно, что графики близки друг к другу на однородных наборах данных.

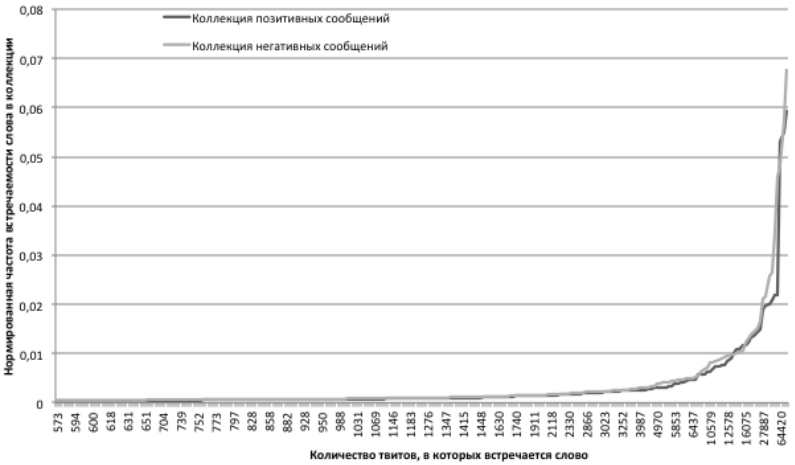

Рис 1. График распределения плотности слов в зависимости от количества твитов, в котором слово встречается

Для того чтобы удостовериться в полноте корпуса объединяем все три коллекции в одну, после чего производим вычисление количества уникальных терминов в зависимости от размера коллекции. На рисунке 2 показано, что при небольшом количестве твитов, добавление к коллекции новых сообщений влечет за собой увеличение числа уникальных терминов. Но, после достижения 340 000 уникальных терминов, добавление новых твитов к коллекции не влечет за собой значительного увеличения уникальных терминов.

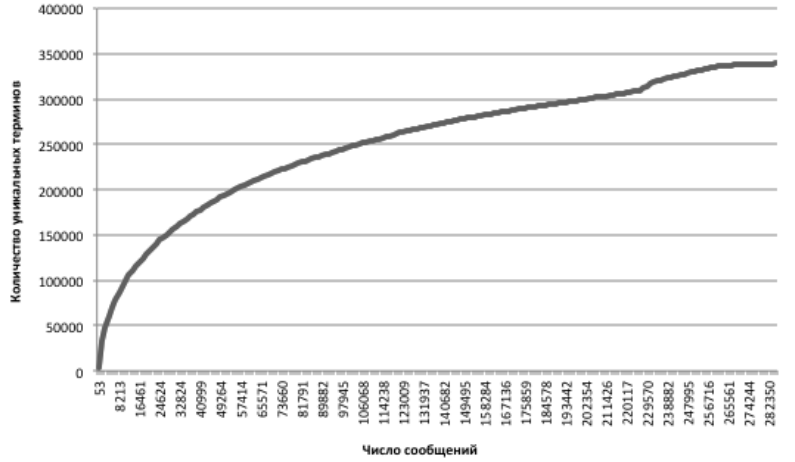

Рис 2. График числа уникальных терминов в зависимости от числа текстов в коллекции

Количество уникальных словоформ в коллекциях распределено следующим образом:

– Положительная коллекция – 150 720 уникальных словоформ;

– Нейтральная коллекция – 105 239 уникальных словоформ;

– Негативная коллекция – 191 677 уникальных словоформ.

Следующим шагом работы является выделение признаков для сокращения размерности векторного пространства и построение взвешенного словаря униграмм на основе размеченных коллекций.

3. Выделение признаков для задачи текстовой классификации по тональности. В формальном виде задача текстовой классификации представляется следующим образом [11]. Пусть существует описание документа $d \in X$, где X — векторное пространство документов, и фиксированный набор классов $C = \{c_1, c_2, \ldots, c_m\}$. Из обучающей выборки (множества документов с заранее известными классами) $D = \{\langle d, c\rangle \mid \langle d, c\rangle \in X \times C\}$ с помощью метода обучения F необходимо получить классифицирующую функции $F(D) = \gamma$, которая отображает документы в классы $\gamma : X \to C$. В решаемой задаче определения тональности множество C состоит из двух или трех элементов (положительный, отрицательный или положительный, нейтральный, отрицательный).

Все документы из обучающей и тестовой выборки представляют собой k-мерные векторы признаков. Таким образом документ определяется в виде вектора $d = (w_1, w_2, \ldots, w_{|V|})$, где V — множество всех уникальных униграмм из обучающей выборки, w_i — вес i-й униграммы.

Для взвешивания униграмм в данной работе используются следующие пять весовых схем:

1. TF-IDF, которая вычисляется по формуле 1:

$$tf.idf = tf \times \log \frac{T}{T(t_i)}, \tag{1}$$

здесь и далее, tf – это частота встречаемости термина в коллекции (положительных или отрицательных твитов). T – общее число сообщений в коллекциях положительных и отрицательных, а $T(t_i)$ – число сообщений в положительной и отрицательной коллекциях, содержащих термин.

2. TF-RF, которая рассчитывается по следующей формуле:

$$tf.rf = tf \times \log(2 + \frac{a}{\max(1, c)}), \tag{2}$$

где a – количество сообщений (положительной) коллекции, содержащие термин, c – количество сообщений (отрицательной) коллекции, содержащие взвешиваемый термин. В работе [12] показано, что методы, основанные на мере RF, показывают лучшие результаты при вычислении веса слова с учетом принадлежности слова к разным классам, чем методы, основанные на мере TF-IDF.

3. Prob-Based [18], которая вычисляется по формуле 3:

$$prob - based = tf \times \log(1 + \frac{a}{c} \times \frac{a}{b}), \quad (3)$$

где a и c аналогично формуле 2, b – число сообщений (положительной) коллекции, которые не содержат взвешиваемый термин.

4. Формула 4 предназначена для вычисления TF-ICF [19]:

$$tf.icf = tf \times \log(1 + \frac{|C|}{cf(t_i)}), \quad (4)$$

где C – это число категорий, cf – число категорий, в которых встречается взвешиваемый термин.

5. ICF-Based [19] вычисляется по формуле 5:

$$icf - based = tf \times \log(2 + \frac{a}{\max(1, c)} \times \frac{|C|}{cf(t_i)}), \quad (5)$$

где C и cf аналогично формуле 4, a и c аналогично формуле 2.

В сообщениях, полученных из Твиттера, часто присутствует «шум» (очень редкие или наоборот очень частые слова, которые встречаются одинаково часто во всех классах), поэтому, перед взвешиванием униграмм, для уменьшения размерности вектора признаков и улучшения качества и скорости работы классификатора, произведена предобработка коллекций:

– Предлоги были отфильтрованы;

– Удалены спецсимволы (стикеры, изображение эмоций), которые пользователи используют при публикации записей с мобильных телефонов;

– Так же были отфильтрованы знаки препинания, такие как: запятая, точка с запятой, двоеточие, тире, точка. Восклицательные и вопросительные знаки были оставлены;

– Удалены имена собственные;

– Удалены значимые события (например, олимпиада);

– Все гиперссылки на другие документы были заменены паттерном «Link»;

– Все заглавные буквы приведены к строчным;

– Большое количество точек (более трех подряд идущих точек) были заменены паттерном «…».

В результате, получен словарь, который состоит из 17 143 взвешенных пятью способами униграмм.

С целью сокращения размерности вектора признаков, был проведен морфологический анализ слов коллекции и составлен второй словарь. Морфологический анализ был произведен с помощью TreeTagger для русского языка [13]. TreeTagger – это вероятностный инструмент для разметки текстов, разрешающий морфосинтаксические неоднозначности русского языка. В качестве униграмм во втором словаре выступают части речи. Как и в случае первого словаря, все униграммы были взвешены согласно пяти весовым схемам. Количество униграмм в морфологическом словаре всего 810. Чтобы не возникало путаницы, использование термина «словарь униграмм» далее по тексту относится к словарю униграмм построенному на извлеченных словоформах, использование термина «словарь морфологических униграмм» – к словарю униграмм построенному на информации о частях речи. На рисунке 3 схематично представлен процесс создания словарей униграмм.

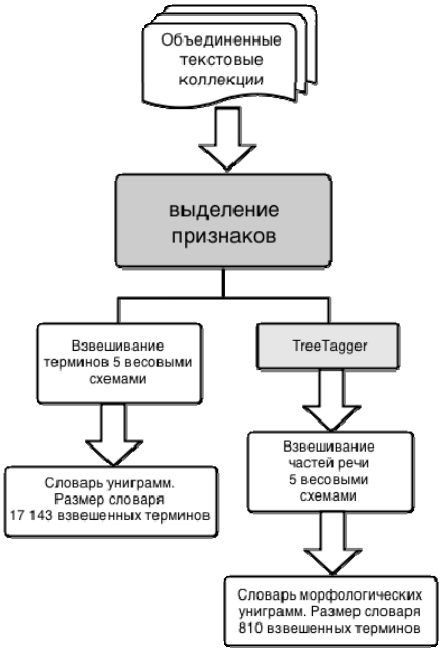

Рис 3. Схема создания словарей униграмм

4. Описание тестовых коллекций. Автоматическая классификация отзывов на два и три класса проводилась на трех коллекциях.

4.1. Коллекция коротких сообщений. Коллекция подробно описана в разделе 2.

4.2. Коллекция новостей. Новостные коллекции были собраны на новостных web-сайтах. Разметка корпуса на положительные, нейтральные и негативные коллекции производилась асессорами вручную. Отличие новостной коллекции от коллекции коротких текстов в том, что новости менее эмоциональны, лексика новостей более нейтральна и не изобилует жаргонизмами, сокращениями и ненормативной лексикой. Как правило, новостные тексты не содержат орфографических ошибок и символов, обозначающих эмоции на письме (смайликов) – тексты более строгие. Новостные тексты существенно длиннее 140 символов. Корпус новостных текстов состоит из следующих коллекций:

– коллекция положительных документов состоит из 46 339 новостей;

– коллекция негативных документов состоит из 46 337 новостей;

– коллекция нейтральных документов состоит из 46 340 новости.

4.3. Коллекция текстов из блогов с разметкой по оценочной тональности и объектам ROMIP 2012. Число текстов коллекции 874. Предлагаемый набор текстов из блогов, участвовал в тестировании дорожек в рамках РОМИП-2011 [14]. В случае классификации на три класса имеем 534 положительных, 236 нейтральных и 103 отрицательных сообщений (одно сообщение исключено из-за ошибок кодировки). В случае классификации на два класса: 749 положительных текстов и 124 негативных сообщений. Разметка коллекции на два и три класса производилась асессорами вручную. В коллекции содержатся тексты блогов пользователей с отзывами на цифровые фотоаппараты, рецензии на книги и на фильмы. Так же, как и в коллекции новостных текстов, длина текстов из блогов не ограничена. Следует отметить сильный дисбаланс тестовой коллекции в сторону положительных текстов.

Три выбранные для эксперимента коллекции принадлежат разным предметным областям, они достаточно разнообразны по содержанию и стилистике, имеют разную длину сообщения. Причем, если в коллекции коротких текстов есть лимит по длине, то в других коллекциях, текст может быть как очень длинным, так и сопоставимым по длине с текстом коротких Твиттер сообщений.

5. Алгоритм классификации. Согласно исследованиям [15, 16], наилучшие результаты в решении задачи автоматической текстовой классификации в целом и отзывов в частности, демонстрирует метод опорных векторов (Support Vector Machines – SVM). Так как вектор признаков имеет достаточно большую размерность, поэтому, в данной работе, для реализации алгоритма SVM была использована библиотека LIBLINEAR [17]. Библиотека LIBLINEAR – это реализация алгоритма SVM с линейным ядром. Все параметры алгоритма были оставлены в соответствии со значениями по умолчанию.

6. Оценка качества. В качестве оценки качества результатов обучения и работы классификатора выбраны четыре общепринятые характеристики: accuracy, precision, recall и F–мера [20].

По результату работы обученного классификатора на тестовой выборке для каждого класса вычисляются следующие значения:

– *TP* – количество истинно-положительных результатов.

– *TN* – количество истинно-отрицательных результатов.

– *FP* – количество ложно-положительных результатов.

– *FN* – количество ложно-отрицательных результатов.

Precision (точность) – это доля объектов классифицированы как X, которые действительно принадлежат классу X (формула 6).

$$Precision = \frac{TP}{TP + FP}, \tag{6}$$

Recall (полнота) – это доля всех объектов класса X, что классифицируется по алгоритму как принадлежащие классу X, вычисляется по формуле 7.

$$Recall = \frac{TP}{TP + FN}, \tag{7}$$

F-мера – это гармоническое среднее между Precision и Recall, формула 8.

$$F\text{-}measure = \frac{2 \times \Pr ecision \times \mathrm{Re} call}{\Pr ecision + \mathrm{Re} call}, \tag{8}$$

Accuracy (аккуратность) – это доля правильно классифицированных объектов среди всех объектов, обработанных с помощью алгоритма классификации (формула 9):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{9}$$

Для того, чтобы избежать проблемы переобучения, когда модель слишком хорошо работает на тестовых примерах, но плохо на реальных данных, для проверки модели используется метод перекрестной проверки (cross-validation). Вся обучающая выборка делится на k-частей, затем $\forall\ i \in \{1, 2, \ldots, k\}$ алгоритм классификации обучается на всей обучающей выборке, кроме части i, после чего, тестируется на i-й части. Результатом работы считается среднее арифметическое по всем проходам. Для данной работы значение k принимаем равным 5.

7. Эксперименты и результаты. Задача работы состоит в исследовании возможности и оценке качества применения словаря униграмм, извлеченных из эмоционально-окрашенных текстов платформы Твиттер для классификации текстов по тональности, извлеченных их других площадок и принадлежащих к заранее не известным предметным областям. С целью сравнения точности классификатора и выбора лучшего подхода, в зависимости от выбранного метода определения веса термина в коллекции для извлеченного набора униграмм, было проведено несколько экспериментов на трех разных по структуре наборах данных.

7.1. Корпус коротких текстов. Первый эксперимент был проведен для корпуса коротких текстов. Результаты работы классификатора на двух классах: положительный и отрицательный, представлены в таблице 1.

Таблица 1. Результаты классификации коротких текстов на два класса в зависимости от выбранной весовой схемы

	TF-IDF	TF-ICF	TF-RF	ICF-Based	Prob-Based
Accuracy	97.3805%	96.5346%	98.4741%	*98.4953%*	98.3026%
Precision	97.3402%	97.2926%	98.0017%	*98.0157%*	97.3489%
Recall	97.4928%	95.8256%	99.0067%	99.0345%	*99.3549%*
F-Measure	0.9742	0.9655	0.9850	*0.9852*	0.9834

Результат работы классификатора на трех классах: положительный, нейтральный, отрицательный, представлен в таблице 2. Меры TF-RF, ICF-Based и Prob-based предназначены для бинарной классификации, поэтому они не участвуют в классификации на три класса. Лучший результат выделен курсивом.

Таблица 2. Результаты классификации коротких текстов на три класса в зависимости от выбранной весовой схемы

	TF-IDF	TF-ICF
Accuracy	*95.5981%*	95.06645%
Precision	*95.8093%*	95.3112%
Recall	*95.5205%*	94.9847%
F-Measure	*0.9566*	0.9515

Так как эксперимент ставился на тех же коллекциях, которые использовались для извлечения униграмм, был получен ожидаемо хороший результат как при классификации на два класса, так и при классификации на три класса.

7.2 Корпус новостей. Следующий эксперимент был проведен для коллекции новостей, размеченных в ручную на три класса: положительные, нейтральные и отрицательные. Для представления текстов в векторном виде использовался вышеописанный словарь униграмм с пятью весовыми схемами. Работа классификатора, как и в первом эксперименте, проводилась перекрестной проверкой с шагом 5. Результаты работы классификатора на двух классах: положительный и отрицательный, представлены в таблице 3. Результат работы классификатора на трех классах: положительный, нейтральный, отрицательный, представлен в таблице 4.

Таблица 3. Результаты классификации текстов новостей на два класса в зависимости от выбранной весовой схемы

	TF-IDF	TF-ICF	TF-RF	ICF-Based	Prob-Based
Accuracy	89.4892%	79.2082%	*95.0095%*	95.0084%	91.1153%
Precision	92.3608%	80.0591%	94.5777%	*94.589%*	89.7698%
Recall	86.1003%	77.7941%	*95.4941%*	95.479%	92.8074%
F-Measure	0.8912	0.7891	*0.9503*	*0.9503*	0.9126

Таблица 4. Результаты классификации текстов новостей на три класса в зависимости от выбранной весовой схемы

	TF-IDF	TF-ICF
Accuracy	*69.8619%*	58.1397%
Precision	*70.9246%*	61.2780%
Recall	*69.8624%*	58.1403%
F-Measure	*0.7039*	0.5967

Эксперимент показал хороший результат при классификации новостей на два класса с использованием схемы взвешивания униграмм TF-RF. Более того, алгоритм достаточно точно классифицировал новостные тексты на три класса, показатель F-measure превышает 0.7 при взвешивании униграмм с помощью схемы TF-IDF. Следовательно, можно сделать вывод, что использование взвешенного словаря униграмм, выделенного из корпуса коротких эмоциональных текстов можно использовать для классификации текстов новостей по тональности.

7.3. Коллекция текстов из блогов ROMIP 2012. Третий эксперимент был поставлен на коллекции текстов из блогов. Результаты работы классификатора на двух классах: положительный и отрицательный, представлены в таблице 5.

Таблица 5. Результаты классификации текстов блогов на два класса в зависимости от выбранной весовой схемы

	TF-IDF	TF-ICF	TF-RF	ICF-Based	Prob-Based
Accuracy	79.3814%	78.1214%	*87.6289%*	*87.6289%*	86.0252%
Precision	89.5688%	89.0756%	90.0125%	90.0125%	*90.9804%*
Recall	85.9813%	84.9132%	*96.2617%*	*96.2617%*	92.9239%
F-Measure	0.8774	0.8694	*0.9303*	*0.9303*	0.9194

При классификации текстов блогов были получены одинаковые значения для весовых схем TF-RF и ICF-Based, это означает, что для коллекции блогов, в словаре униграмм нет слов, которые встречаются только в положительном классе или только в отрицательном. Следующий эксперимент был проведен для той же самой коллекции, но классификация проводилась на три класса, результаты представлены в таблице 6.

Таблица 6. Результаты классификации текстов блогов на три класса в зависимости от выбранной весовой схемы

	TF-IDF	TF-ICF
Accuracy	53.9773%	*57.9545%*
Precision	*56.1341%*	55.8903%
Recall	53.1164%	*53.579%*
F-Measure	0.5458	*0.5471*

При классификации длинных текстов на три класса, обе меры показали неудовлетворительные результаты согласно F-мере.

7.4. Классификация текстов с использованием морфологического словаря униграмм. Следующий эксперимент был поставлен с использованием морфологического словаря униграмм. Эксперимент показал, что сокращение размерности вектора признаков сильно сказывается на качестве классификации. Это связано с тем, что практически для 20% слов часть речи не определяется автоматически с помощью инструмента [13], эти слова, как правило, написаны с ошибками, использован сленг или повторение гласных букв внутри слова.

Результаты классификации коротких сообщений с использованием морфологического словаря униграмм на два и три класса представлены в таблицах 7 и 8 соответственно. Результаты

классификации новостных текстов на два и три класса представлены в таблицах 9 и 10 соответственно. Для коллекции блогов не проводилось экспериментов по классификации текстов по тональности с использованием морфологического словаря униграмм из-за сравнительно низких результатов, полученных для первых двух коллекций.

Таблица 7. Результаты классификации текстов коротких сообщений на два класса с использованием морфологического словаря униграмм

	TF-IDF	TF-ICF	TF-RF	ICF-Based	Prob-Based
Accuracy	54.3832%	50.5929%	*61.588%*	61.5814%	61.5395%
Precision	53.8769%	51.5864%	*61.574%*	61.5673%	61.487%
Recall	69.153%	40.1662%	64.3068%	64.3025%	*64.4461%*
F-Measure	0.605666	0.451656	0.6291	0.629	*0.6293*

Таблица 8. Результаты классификации текстов коротких сообщений на три класса с использованием морфологического словаря униграмм

	TF-IDF	TF-ICF
Accuracy	50.7952%	*54.0306%*
Precision	*51.819%*	49.0239%
Recall	52.0946%	*52.5921%*
F-Measure	*0.5196*	0.5074

Таблица 9. Результаты классификации новостных текстов на два класса с использованием морфологического словаря униграмм

	TF-IDF	TF-ICF	TF-RF	ICF-Based	Prob-Based
Accuracy	62.2532%	80.4198%	*81.4557%*	81.4449%	80.9593%
Precision	57.0629%	*84.657%*	77.7725%	77.7682%	77.2874%
Recall	*98.9964%*	74.3067%	88.0868%	88.0652%	87.6875%
F-Measure	0.724	0.7914	*0.826*	*0.826*	0.8216

Таблица 10. Результаты классификации новостных текстов на три класса с использованием морфологического словаря униграмм

	TF-IDF	TF-ICF
Accuracy	*48.3684%*	46.0746%
Precision	53.4827%	*60.1676%*
Recall	*48.641%*	46.56%
F-Measure	0.50947	*0.52496*

Заключение. Как показали эксперименты, извлечение однословных униграмм из достаточно представительной однородной коллекции коротких эмоционально окрашенных сообщений площадки Твиттер может успешно применяться для создания вектора признаков для SVM-классификатора при решении задачи автоматической

классификации текстов по тональности, независимо от того, к какой предметной области принадлежит текст. Полученные результаты классификации на два класса сопоставимы, а в некоторых случаях даже превосходят результаты классификаторов, спроектированных для классификации одной заранее определенной предметной области. Однако, при классификации на три класса, точность классификатора заметно ухудшается на длинных текстах (новости и тексты блогов).

В результате эксперимента определены весовые схемы, дающие наибольшую точность при классификации текстов по тональности на два и на три класса. Для классификации коротких текста на два класса лучший результат, согласно метрике F-measure, показывает схема ICF-Based, а для неограниченных по длине текстов – схема TF-RF.

Эксперименты показали, что использование морфологической разметки не дает сопоставимые результаты точности классификации в сравнении с использованием словаря униграмм. В дальнейшем планируется объединить словари и поставить эксперимент на объединенном словаре униграмм и морфологических униграмм. Также к перспективам исследования относится поиск возможностей для сокращения размерности вектора признаков.

Литература

1. *Feldman R.* Techniques and Applications for Sentiment Analysis // Communications of the ACM. 2013. vol. 56. no. 4. pp. 82–89.
2. *Pang B., Lee L.* Thumbs up? Sentiment classification using machine learning techniques // Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP). Philadelphia: ACL. 2002. pp. 79–86.
3. *Pang B., Lee L.* Seeing stars: exploiting class relationships for sentiment categorization with respect of rating scales // Proc. of ACL, 43rd Meeting of the Association for Computational Linguistics. Ann Arbor: ACM. 2005. pp. 115–124.
4. *Bespalov D., Bai B., Qi Y., Shokoufandeh A.* Sentiment classification based on supervised latent n-gram analysis // In Proceedings of the 20th ACM international conference on Information and knowledge management (CIKM '11). ACM. New York. NY. USA. 2011. pp. 375–382.
5. *Nguyen D.Q., Nguyen D.Q., Vu T., Pham S.B.* Sentiment classification on polarity reviews: an empirical study using rating-based features // In: 5th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis. Baltimore, Md. 2014. pp. 128–135
6. *Cui H., Mittal V., Datar M.* Comparative experiments on sentiment classification for online product reviews // Proceedings of the 21st national conference on Artificial intelligence. AAAI Press. 2006. vol. 2. pp. 1265–1270.
7. The Streaming APIs. URL: https://dev.twitter.com/docs/streaming-apis (дата обращения: 28.10.2014).
8. *Reed J.W., Jiao Y., Potok T.E., Klump B.A., Elmore M.T., Hurson A.R.* TF-ICF: A New Term Weighting Scheme for Clustering Dynamic Data Streams // In: Proc. Machine Learning and Applications (ICMLA '06). 2006. pp. 258–263.
9. *Рубцова Ю.В.* Метод построения и анализа корпуса коротких текстов для задачи классификации отзывов // Электронные библиотеки: перспективные методы и

технологии, электронные коллекции: Труды XV Всероссийской научной конференции RCDL'2013. Ярославль. 2013. С. 269–275.
10. *Kechedzhy K. E., Usatenko O.V., Yampol'skii V. A*. Rank distributions of words in additive many-step Markov chains and the Zipf law Arxiv LANL // Phys. Rev. E. 2005. vol. 72. pp. 046138(1)–046138(6).
11. *Manning D., Raghavan P., Schutze H.* Introduction to Information Retrieval // Cambridge University Press. 2008.
12. *Lan M., Tan C.L., Su J., Lu Y.* Supervised and Traditional Term Weighting Methods for Automatic Text Categorization // IEEE Transactions on Pattern Analysis and Machine Intelligence. vol. 31. no. 4. 2009. pp. 721–735.
13. *Schmid H.* Probabilistic part-of-speech tagging using decision trees // In Proc. of the International Conference on New Methods in Language Processing. 1994. pp. 44–49.
14. Коллекция текстов из блогов с разметкой по оценочной тональности и объектам. URL: http://romip.ru/ru/collections/sentiment-blog-collection-2012.html. (дата обращения: 28.10.2014).
15. *Joachims T.* Text categorization with support vector machines: Learning with many relevant features // In Proc. of the European Conference on Machine Learning (ECML 1998). 1998. pp. 137–142.
16. *Sebastiani F.* Machine Learning in Automated Text Categorization. ACM Computing Surveys. vol. 34. no. 1. March 2002. pp. 1–47.
17. *Fan R.-E., Chang K.-W., Hsieh C.-J., Wang X.-R., Lin C.-J.* LIBLINEAR: a Library for Large Linear Classification // J. of Machine Learning Research. 2008. vol. 9. pp. 1871–1874.
18. *Liu Y., Loh H.T., Sun A.* Imbalanced text classification: A term weighting approach // Expert systems with Applications. 2009. vol. 36. no. 1. pp. 690-701.
19. *Wang D., Zhang H.* Inverse-category-frequency based supervised term weighting scheme for text categorization // arXiv preprint arXiv: 1012.2609. 2010.
20. *Olson D. L., Dursun D.* Advanced Data Mining Techniques (1st edition) // Springer. 2008. 138 p.

References

1. Feldman R. Techniques and Applications for Sentiment Analysis. Communications of the ACM. 2013. vol. 56. no. 4. pp. 82–89.
2. Pang B., Lee L. Thumbs up? Sentiment classification using machine learning techniques. Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP). Philadelphia: ACL. 2002. pp. 79–86.
3. Pang B., Lee L. Seeing stars: exploiting class relationships for sentiment categorization with respect of rating scales. Proc. of ACL, 43rd Meeting of the Association for Computational Linguistics. Ann Arbor: ACM. 2005. pp. 115–124.
4. Bespalov D., Bai B., Qi Y., Shokoufandeh A. Sentiment classification based on supervised latent n-gram analysis. Proceedings of the 20th ACM international conference on Information and knowledge management (CIKM '11). ACM. New York. NY. USA. 2011. pp. 375–382
5. Nguyen D.Q., Nguyen D.Q., Vu T., Pham S.B. Sentiment classification on polarity reviews: an empirical study using rating-based features. In: 5th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis. Baltimore, Md. 2014. pp. 128–135
6. Cui H., Mittal V., Datar M. Comparative experiments on sentiment classification for online product reviews. Proceedings of the 21st national conference on Artificial intelligence. AAAI Press. 2006. vol. 2. pp. 1265–1270
7. The Streaming APIs. Available at: https://dev.twitter.com/docs/streaming-apis (accessed 28.10.2014).

8. Reed J.W., Jiao Y., Potok T.E., Klump B.A., Elmore M.T., Hurson A.R. TF-ICF: A New Term Weighting Scheme for Clustering Dynamic Data Streams. In: Proc. Machine Learning and Applications (ICMLA '06). 2006. pp. 258–263.
9. Rubtsova Y.V. [A Method for development and analysis of short text corpus for the review classification task]. *Trudy XV Vserossiiskoy naychnoy konferencii RCDL'2013* [In proc. of The XVth All-Russian Scientific Conference RCDL'2013]. Jaroslavl'. 2013. pp. 269–275. (In Russ.)
10. Kechedzhy K. E., Usatenko O.V., Yampol'skii V. A. Rank distributions of words in additive many-step Markov chains and the Zipf law Arxiv LANL. Phys. Rev. E. 2005. vol. 72. pp. 046138(1)–046138(6).
11. Manning D., Raghavan P., Schutze H. Introduction to Information Retrieval. Cambridge University Press. 2008.
12. Lan M., Tan C.L., Su J., Lu Y. Supervised and Traditional Term Weighting Methods for Automatic Text Categorization. IEEE Transactions on Pattern Analysis and Machine Intelligence. vol. 31. no. 4. 2009. pp. 721–735.
13. Schmid H. Probabilistic part-of-speech tagging using decision trees. In Proc. of the International Conference on New Methods in Language Processing. 1994. pp. 44–49.
14. Kollekcija tekstov iz blogov s razmetkoj po ocenochnoj tonal'nosti i ob'ektam [A collection of marked by sentiment texts from blogs]. Available at: http://romip.ru/ru/collections/sentiment-blog-collection-2012.html. (accessed 28.10.2014). (In Russ.)
15. Joachims T. Text categorization with support vector machines: Learning with many relevant features. In Proc. of the European Conference on Machine Learning (ECML 1998). 1998. pp. 137–142.
16. Sebastiani F. Machine Learning in Automated Text Categorization. ACM Computing Surveys. vol. 34. no. 1. March 2002. pp. 1–47.
17. Fan R.-E., Chang K.-W., Hsieh C.-J., Wang X.-R., Lin C.-J. LIBLINEAR: a Library for Large Linear Classification. J. of Machine Learning Research. 2008. vol. 9. pp. 1871–1874.
18. Liu Y., Loh H.T., Sun A. Imbalanced text classification: A term weighting approach. Expert systems with Applications. 2009. vol. 36. no. 1. pp. 690-701.
19. Wang D., Zhang H. Inverse-category-frequency based supervised term weighting scheme for text categorization. arXiv preprint arXiv: 1012.2609. 2010.
21. Olson D. L., Dursun D. Advanced Data Mining Techniques (1st edition). Springer. 2008. 138 p.

Рубцова Юлия Владимировна — аспирант, Институт систем информатики им. А.П. Ершова СО РАН. Область научных интересов: мат. лингвистика, корпусная лингвистика, классификация текстов по тональности. Число научных публикаций — 5. yu.rubtsova@gmail.com; 630105 г. Новосибирск, ул. Линейная 47/2; р.т. +79059516757.

Rubtsova Yuliya Vladimirovna — Ph.D. student, A.P. Ershov Institute of Informatics Systems. Scientific interests: sentiment analysis, mathematical linguistics, machine learning. The number of publications — 5. yu.rubtsova@gmail.com; 630105 Novosibirsk, Lineynaya street 47/2; office phone +79059516757.

РЕФЕРАТ

Рубцова Ю.В. **Разработка и исследование предметно независимого классификатора текстов по тональности.**

В статье представляется метод построения классификатора для классификации текстов по тональности на два и на три класса (положительные и негативные; положительные, нейтральные и негативные тексты). Во введении обосновывается постановка задачи, приводятся примеры практического применения.

Второй раздел статьи посвящен описанию коллецкий из которых будут выделяться признаки для классификатора. Показывается, что коллекции достаточно представительные и однородные и они могут быть использованы для извлечения униграмм.

В третьем разделе описывается алгоритм извлечения признаков для задачи текстовой классификации по тональности. Описана процедура фильтрация коллекций и введены 5 весовых схем для взвешивания униграмм.

В следующем разделе приведены характеристики текстовых коллекций, которые были использованы для экспериментов. Эксперименты ставились на трех типах коллекций:

– короткие сообщения микроблогов;

– новостная коллекция;

– коллекция отзывов, тексты собраны из блогов.

В качестве алгоритма классификации был использован метод опорных векторов, этому посвящен шестой раздел.

Далее были введены и описаны оценки качества классификатора. В качестве оценки качества результатов обучения и работы классификатора выбраны четыре общепринятые характеристики: accuracy, precision, recall и F–мера.

В разделе 7 представлены результаты экспериментов. Всего было поставлено 10 экспериментов на 3-х тестовых коллекциях.

В результате экспериментов определены весовые схемы, дающие наибольшую точность при классификации текстов по тональности на два и на три класса. Для классификации коротких текста на два класса лучший результат, согласно метрике F-measure, показывает схема ICF-Based, а для неограниченных по длине текстов – схема TF-RF.

SUMMARY

Rubtsova Y.V. **Research and Development of Domain Independent Sentiment Classifier.**

The paper presents a method of constructing a classifier for text classification in tone on the two and three classes (both positive and negative, positive, neutral and negative texts). In the introduction the problem statement, and examples of practical application.

The second section is devoted to describe collections which will be used for feature extraction for the classifier. It was shown that the collections are representative and homogeneous enough tobe used for extracting unigrams.

The third section describes the feature extraction algorithm for the task of text sentiment classification. The procedure of filtering collections and description of 5 weight schemes for weighting unigramm also explained.

The following section describes the characteristics of text collections, which were used for the experiments. Experiments were performed on three types of collections:

– Short microblogging messages;
– A collection of news;
– A collection of reviews, the texts collected from blogs.

As a classification that algorithm was used is support vector machine, it is dedicated to the sixth section.

Next was introduced and described quality assessment for the classifier. As the assessment of the quality four common measures were selected: accuracy, precision, recall and F-measure.

Section 7 presents the results of experiments. It was delivered 10 experiments on 3 test collection in total.

The experiments determined weighting schemes which are the most accurate in the text sentiment classification into two or three classes. For the short text classification into two classes the best result, according to the F-measure, shows scheme ICF-Based, and for the long texts - scheme TF-RF.