Вестник УГАТУ
Уфа : УГАТУ, 2012 Т. 16, № 6 (51). С. 91–99

УДК 004.912

Н. И. Юсупова, Д. Р. Богданова, М. В. Бойко

АЛГОРИТМИЧЕСКОЕ И ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ ДЛЯ АНАЛИЗА ТОНАЛЬНОСТИ ТЕКСТОВЫХ СООБЩЕНИЙ С ИСПОЛЬЗОВАНИЕМ МАШИННОГО ОБУЧЕНИЯ

В статье рассматривается задача классификации тональности текстовых сообщений с использованием методов машинного обучения, в частности наивного байесовского классификатора. Для повышения точности классификации тональности рассматривается применение алгоритмов бустинга и бэггинга. *Анализ текста*; *анализ тональности*; *сентимент анализ*; *машинное обучение*

ВВЕДЕНИЕ

Современный этап развития человечества характеризуется бурным ростом количества информации. Одной из наиболее распространенных форм хранения информации являются тексты на естественном языке. Текстовая форма информации естественна для человека и легко им воспринимается. Развитие информационных технологий сопровождается интенсивным ростом числа веб-сайтов, которое на сегодняшний день составляет более 285 млн, и, как следствие, ростом объема текстовых данных. Огромное количество информации скапливается в многочисленных текстовых базах, хранящихся в персональных компьютерах, в локальных и глобальных сетях. Рядовому пользователю становится все сложнее работать с гигантскими объемами данных. Чтение объемных текстов, ручной поиск и анализ нужной информации в гигантских массивах текстовых данных малоэффективны. Для решения данной проблемы и автоматизации процессов получило развитие направление обработки естественного языка (natural language processing), решающее задачи информационного поиска (information retrieval), машинного перевода (machine translation), извлечения информации (information extraction), анализа тональности текста (sentiment analysis) и др.

Статья посвящена вопросам анализа тональности текстовых сообщений (sentiment analysis) с использованием машинного обучения. Анализ тональности в тексте является одним из направлений в анализе естественно-языковых текстов. Тональностью называется эмоциональная оценка, которая выражена в тексте. Она может иметь одномерное эмотивное пространство (два класса) или многомерное (несколько классов). Перспективность анализа тональности текста заключается в том, что на основе текстовой информации он позволяет оценить успешность рекламной кампании, политических и экономических реформ; выявить отношение прессы и СМИ к определенной персоне, к организации, к событию; определить, как относятся потребители к определенной продукции, к услугам, к организации. Так, в работе [1] авторами рассматривается применение анализа тональности к исследованию мнений потребителей.

Несмотря на перспективы данного направления, пока оно не столь активно применяется в системах обработки текстовой информации. Причинами являются трудности выделения эмоциональной лексики в текстах, несовершенство существующих текстовых анализаторов, зависимость от предметной области. Поэтому совершенствование и разработка новых методов

Контактная информация: 8-987-255-15-05

Результаты, приведенные в статье, являются частью научных исследований по тематическому плану НИР по заданию Министерства образования и науки РФ «Исследование интеллектуальных технологий поддержки принятия решений и управления для сложных социально-экономических объектов» и научно-исследовательской работы по теме «Исследования и разработка интеллектуальных технологий поддержки принятия решений и управления на основе инженерии знаний». Исследования поддержаны грантом РФФИ №09-07-00408-а «Распределенная интеллектуальная система поддержки принятия решений при выполнении проектов фундаментальных исследований сложных систем» и грантом Президента Российской Федерации № НШ-65497.2010.9 «Теоретические и методические основы разработки информационных систем, а также их применения в промышленности и в социально-экономической среде с учетом тенденции развития информационных технологий».

анализа тональности на основе машинного обучения является актуальной задачей.

В статье приводятся результаты исследования применения классификатора тональности текстов на русском языке с использованием машинного обучения. В качестве тестовых данных использовались отзывы клиентов российских банков.

1. ПОДХОДЫ К АНАЛИЗУ ТОНАЛЬНОСТИ ТЕКСТОВЫХ СООБЩЕНИЙ

На сегодняшний день существует три подхода к анализу тональности текстовых сообщений:

1. Анализ тональности текста по заранее составленным тональным словарям с применением лингвистического анализа. Тональные словари состоят из таких элементов как слова, словосочетания, паттерны, каждый из которых имеет свою эмоциональную окраску. Тональность текста определяется по совокупности найденной эмотивной лексики и оценивается в зависимости от количества позитива и негатива.

2. Анализ тональности текста методами машинного обучения. Текст представляется в векторной форме. По имеющейся обучающей выборке производится обучение классификатора. После этого производится классификация тональности текста.

3. Комбинация первого и второго подходов.

Первый подход достаточно трудоемок из-за необходимости составления тональных словарей, получения списка тональных паттернов и разработки лингвистических анализаторов, но отличается большей гибкостью. Преимуществом данного подхода является то, что он позволяет увидеть эмоциональную лексику на уровне предложении.

В работе [2] авторами представлен алгоритм анализа тональности текста, основанный на тональных словарях, состоящий из нескольких этапов: морфологического анализа текста, разметки по словарным спискам тональной лексики, синтаксического анализа и непосредственно определения тональности. Работу алгоритма можно оценить на веб-сайте [3].

В работе [4] авторами разработан следующий алгоритм оценки тональности текста, который включает распознавание объекта тональности, синтаксический анализ текста, выделение и классификация пропозиций, которые выражают тональность, оценка общей тональности на основе тональности всех пропозиций.

За рубежом так же ведется активный поиск и совершенствование анализа тональности текста на основе тональных словарей и лингвистического анализа. Одно из таких исследований представлено в работе [5]. В ней описан анализатор, который состоит из: 1) извлечения специальной терминологии из текста; 2) определения тональности; 3) ассоциативного анализа отношений. Анализатор использует два лингвистических аппарата: тональный словарь и базу тональных шаблонов.

Подход, основанный на использовании машинного обучения, предполагает наличие предварительно размеченного обучающего набора данных. Целью обучения в случае анализа тональности является получение необходимых и достаточных правил, с помощью которых можно произвести классификацию по тональности новых текстовых сообщений, сходных с теми, которые составляли обучающую выборку. Недостатком алгоритмов анализа тональности на основе машинного обучения является зависимость от качества и количества обучающих данных, к тому же подход не позволяет провести глубокий анализ текста, выявить объект и субъект тональности.

Методы машинного обучения для решения задачи классификации тональности текстовых сообщений активно развиваются за рубежом. В российской научной практике пока не известны случаи успешного применения машинного обучения для анализа тональности, поэтому рассмотрим некоторые работы зарубежных авторов.

Большой вклад в развитие анализа тональности текстовых сообщений внесли исследователи из Корнельского университета Б. Пэнг и Л. Ли. В 2008 году они выпустили книгу «Opinion Mining and Sentiment Analysis» [6], посвященную современным методам и подходам к анализу тональности в текстовых сообщениях. В их работе [7] рассматривается классификация тональности с использованием машинного обучения и показывается, что такой подход превосходит простые техники, основанные на составлении словарей часто употребляемых позитивных и негативных слов. В дальнейшей работе [8] авторы описывают алгоритм, который позволяет классифицировать тональность, используя только субъективные предложения. Объективные предложения, как правило, не имеют тональной окраски, но создают шум в данных.

В работе [9] авторами рассматривается проблема того, что при классификации тональности из обучающих данных извлекается очень большое число термов. Авторы описывают способы для отбора наиболее информативных термов и оценки их тональности.

Для устранения недостатков рассмотренных выше подходов применяется их комбинирование. Так, в работе [10] метод основан на извлекаемых лексических правилах, при этом обучение с участием человека и машинное обучение комбинируются в один алгоритм классификации тональности.

В другой работе [11] исследователи из корпорации Майкрософт предлагают пути сокращения времени на составление тональных словарей. Результат достигается за счет совместного использования автоматического извлечения информативных шаблонов и машинного обучения.

Комбинированный подход является перспективным направлением, так как объединяет достоинства двух первых подходов. Здесь важной задачей для исследования является определение способа их взаимодействия.

2. ПРЕДЛАГАЕМЫЙ МЕТОД

Предлагаемый метод основан на подходе анализа тональности текстовых сообщений с использованием машинного обучения. В качестве алгоритма машинного обучения выбран наивный байесовский классификатор. Для повышения точности классификации рассматриваются мета-алгоритмы машинного обучения – бустинг и бэггинг.

Математически задача классификации тональности может быть представлена следующим образом. Имеется два класса – класс положительных сообщений c_1 и класс негативных сообщений c_2:

$$C = \{c_1, c_2\}, \quad (1)$$

имеется множество сообщений:

$$D = \{d_1, d_2, ..., d_n\}, \quad (2)$$

и неизвестная целевая функция:

$$F : C \times D \to \{0,1\}, \quad (3)$$

необходимо построить классификатор F', максимально близкий к целевой функции F. Для решения данной задачи с помощью машинного обучения имеется множество размеченных сообщений.

$$K \subset C \times D^l. \quad (4)$$

Признаковое пространство в задаче распознавания тональности может быть представлено с помощью векторной модели. Каждое текстовое сообщение рассматривается в виде набора слов (bag of words). Это представление текстового сообщения в качестве точки в многомерном пространстве. Близко лежащие друг к другу точки соответствуют семантически схожим сообщениям. В данной модели игнорируется последовательность слов. Так, например, «хорошая книга» и «книга хорошая» одно и тоже. Таким образом, сообщение представляет собой «мешок» со словами.

Для решения задачи классификации тональности в работе используется метод машинного обучения на основе наивного байесовского классификатора. Основные его преимущества – простота реализации и низкие вычислительные затраты при обучении и классификации. Основной его недостаток заключается в предположении независимости признаков, что в реальных ситуациях редко выполняется.

Пусть каждое сообщение d принимает значения из словаря V и описывается некоторым набором слов $\{w_1, w_2, \ldots, w_n\}$. Имеется множество классов $C = \{c_1, c_2\}$, состоящее из класса позитивных сообщений и класса негативных сообщений. Необходимо найти наиболее вероятное значение соответствующего класса данного набора слов:

$$c_{NB} = \arg\max_{c_j \in C} p(d = c_j \mid w_1, w_2, ..., w_n). \quad (5)$$

Известно, что вероятность условного события может быть найдена по теореме Байеса:

$$p(d = c_j \mid w_1, w_2, ..., w_n) = \\ = \frac{p(w_1, w_2, ..., w_n \mid d = c_j) \cdot p(d = c_j)}{p(w_1, w_2, ..., w_n)} \quad (6)$$

Тогда выражение (5) примет вид:

$$c_{NB} = \arg\max_{c_j \in C} \frac{p(w_1, w_2, ..., w_n \mid d = c_j) \cdot p(d = c_j)}{p(w_1, w_2, ..., w_n)} \quad (7)$$

Из выражения (7) нам интересен только числитель дроби, так как ее знаменатель не зависит от класса. Таким образом, знаменатель является константой и может быть сокращен. Предположив условную независимость признаков, получим выражение, по которому производится классификация:

$$c_{NB} = \arg\max_{c \in C} p(w_1, w_2, ..., w_n \mid d = c) \cdot p(d = c). \quad (8)$$

Наивный байесовский классификатор работает при следующих допущениях:

- слова и словосочетания в сообщении независимы между собой;
- не учитывается последовательность слов;
- не учитывается длина сообщения.

Существуют два способа реализации наивного байесовского классификатора – многомерный или другое название – модель Бернулли (*multivariate, Bernoulli model*) и мультиномиальный (*multinomial*). Различие заключается в том, что в модели Бернулли в качестве признака рассматривается присутствие слова в тексте. В мультиномиальной модели рассматривается число появлений какого-либо слова в тексте. В табл. 1 представлен пример векторной записи текста.

Таблица 1

	Векторная запись
Модель Бернулли	[0, 0, 1, 1, 0, 1, 1, 1, 0, 0, 0]
Мультиномиальная модель	[0, 0, 2, 1, 0, 3, 1, 2, 0, 0, 0]

Модель Бернулли

Рассмотрим алгоритм классификации тональности в модели Бернулли [12]. В модели Бернулли сообщение описывается вектором, состоящим из атрибутов, принимающих значения 0 либо 1. Таким образом, рассматривается только присутствие или отсутствие слова в сообщении, а сколько раз оно повторяется в сообщении неважно.

Пусть дан словарь $V = \{w_t\}_{t=1}^{|V|}$. Тогда сообщение d_i описывается вектором длины $|V|$, состоящим из битов b_{it}. Если слово w_t встречается в сообщении d_i, то $b_{it} = 1$, если отсутствует, то $b_{it} = 0$. Тогда правдоподобие принадлежности сообщения d_i к классу c_j можно посчитать по формуле:

$$p(d_i \mid c_j) = \prod_{t=1}^{|V|} (b_{it} \cdot p(w_t \mid c_j) + \\ + (1 - b_{it}) \cdot (1 - p(w_t \mid c_j))). \quad (9)$$

Для обучения классификатора нужно найти вероятности $p(w_t \mid c_j)$. Пусть имеется обучающий набор сообщений $D = \{d\}_{i=1}^{|D|}$, которые уже распределены по классам c_j, тогда можно подсчитать оценки вероятностей того, что то или иное слово встречается в том или ином классе:

$$p(w_t \mid c_j) = \frac{1 + \sum_{i=1}^{|D|} b_{it} \cdot p(c_j \mid d_i)}{2 + \sum_{i=1}^{|D|} p(c_j \mid d_i)}. \quad (10)$$

Априорные вероятности классов можно подсчитать по формуле:

$$p(c_j) = \frac{\sum_{i=1}^{|D|} p(c_j \mid d_i)}{|D|}. \quad (11)$$

Тогда классификация будет осуществляться по формуле:

$$c_{NB} = \arg\max_{c_j \in C} p(c_j) \cdot p(d_i \mid c_j) = \\ = \arg\max_{c_j \in C} [\log(\sum_{i=1}^{|D|} p(c_j \mid d_i)) + \sum_{t=1}^{|V|} \log[b_{it} \times \\ \times p(w_t \mid c_j) + (1 - b_{it})(1 - p(w_t \mid c_j))]]. \quad (12)$$

Из формулы (10) следует, что некоторые вероятности будут иметь нулевые значения, так как какие-то слова могут присутствовать в одном классе обучающих данных и отсутствовать в другом. Сложности с нулевыми вероятностями возникают, когда они перемножаются в формуле (12). При этом все выражение обнуляется, и происходит потеря информации. Чтобы избежать получения нулевых вероятностей, применяется *add-one* или Лапласовское сглаживание, которое заключается в добавлении единицы к числителю:

$$p(w_t \mid c_j) = \frac{1 + \sum_{i=1}^{|D|} b_{it} \cdot p(c_j \mid d_i)}{2 + \sum_{i=1}^{|D|} p(c_j \mid d_i)}. \quad (13)$$

Алгоритм классификации тональности в текстовом сообщении с использованием модели Бернулли представлен на рис. 1–2. Он состоит из обучающей части и классифицирующей части. В обучающей части входными параметрами являются множество обучающих сообщений и множество классов. Здесь создается словарь термов V, находятся вероятности $p(c_j)$ и вероятности $p(w_t \mid c_j)$, настраивается пороговое значение h для минимизации ошибки классификации. Выходными данными является полностью обученный классификатор с настроенными параметрами. Классифицирующая часть применяется для сообщений, тональность которых нужно определить.

Мультиномиальная модель

В мультиномиальной модели [12] сообщение представляет собой последовательность случайного выбора одного слова из словаря. В данной модели учитывается количество повторений каждого слова в одном сообщении, но не учитываются слова, которых нет в сообщении.

Пусть дан словарь $V = \{w_t\}_{t=1}^{|V|}$. Тогда сообщение d_i описывается вектором длины $|V|$, состоящим из слов, каждое из которых вынуто из словаря с вероятностью $p(w_t \mid c_j)$. Тогда правдоподобие принадлежности сообщения d_i к классу c_i считается по формуле:

$$p(d_i \mid c_j) = p(|d_i|) \cdot |d_i|! \cdot \prod_{t=1}^{|V|} \frac{1}{K_{it}!} p(w_t \mid c_j)^{K_{it}}, \quad (14)$$

где K_{it} – количество вхождений слова w_t в сообщение d_i.

Вход: множество сообщений $D = \{d_1, d_2, \dots, d_n\}$, множество классов $C = \{c_1, c_2\}$

1. Извлечь все слова из D в словарь V
2. Для каждого $c_j \in C$ делать
   3. Подсчитать число сообщений N^c в классе
   4. Вычислить вероятность $p(c_j) = N^c/N$
   5. Для каждого $w_t \in V$ делать
      6. Подсчитать число сообщений $N_{w_t}^c$, содержащих слово w_t
      7. Вычислить вероятность $p(w_t|c_j) = N_{w_t}^c + 1)/(N^c + 2)$
8. Найти пороговое значение h, минимизирующее ошибку классификации

Выход: $V, p(c_j), p(w_t|c_j), h$

Рис. 1. Алгоритм обучения классификатора

Вход: сообщений $d, V, p(c_j), p(w_t|c_j), h$

1. Извлечь все термы из d в словарь V_d
2. Для каждого $c_j \in C$ делать
   3. $score[c_j] = \ln p\,(c_j)$
   4. Для каждого $w_t \in V$ делать
      5. Если $w_t \in V_d$, тогда $score[c_j] \mathrel{+}= \ln p\,(w_t|c_j)$
      6. Иначе $score[c_j] \mathrel{+}= \ln(1 - p\,(w_t|c_j)$
7. Если $score[c_1] > h \cdot score[c_2]$, тогда $d \in c_1$, иначе $d \in c_2$

Выход: тональность сообщения d

Рис. 2. Алгоритм работы классификатора

Для обучения классификатора нужно тоже найти вероятности $p(w_t \mid c_j)$. Пусть имеется обучающий набор сообщений $D = \{d\}_{i=1}^{|D|}$, которые уже распределены по классам c_i, и мы знаем число вхождений слов в сообщения K_{it}. Тогда можно подсчитать оценки вероятностей того, что то или иное слово встречается в том или ином классе. В данном случае также применяется сглаживании *add-one*:

$$p(w_t \mid c_j) = \frac{1 + \sum_{i=1}^{|D|} K_{it} \cdot p(c_j \mid d_i)}{|V| + \sum_{s=1}^{|V|} \sum_{i=1}^{|D|} K_{is} \cdot p(c_j \mid d_i)}. \quad (15)$$

Априорные вероятности классов можно подсчитать по формуле:

$$p(c_j) = \frac{\sum_{i=1}^{|D|} p(c_j \mid d_i)}{|D|}. \quad (16)$$

Тогда классификация будет осуществляться по формуле:

$$\begin{aligned} c_{NB} &= \arg\max_j p(c_j) \cdot p(d_i|c_j) = \\ &= \arg\max_j [\log(\sum_{i=1}^{|D|} p(c_j \mid d_i)) + \\ &+ \sum_{t=1}^{|V|} K_{it} \cdot \log p(w_t \mid c_j)]. \end{aligned} \quad (17)$$

Разработанный алгоритм классификации тональности в текстовом сообщении с помощью Multinomial Naïve Bayes model представлен на рис. 3–4.

Вход: множество сообщений $D = \{d_1, d_2, \dots, d_n\}$, множество классов $C = \{c_1, c_2\}$

1. Извлечь все слова из D в словарь V
2. Для каждого $c_j \in C$ делать
   3. Подсчитать число сообщений N^c в классе
   4. Вычислить вероятность $p(c_j) = N^c/N$
   5. Для каждого $w_t \in V$ делать
      6. Подсчитать число употреблений $M_{w_t}^c$ слова w_t
      7. Вычислить вероятность $p(w_t|c_j) = (M_{w_t}^c + 1)/\sum_{t=1}^{|V|}(M_{w_t}^c + 1)$
8. Найти пороговое значение h, минимизирующее ошибку классификации

Выход: $V, p(c_j), p(w_t|c_j), h$

Рис. 3. Алгоритм обучения классификатора

Вход: сообщений $d, V, p(c_j), p(w_t|c_j), h$

1. Извлечь все термы из d в словарь V_d
2. Для каждого $c_j \in C$ делать
   3. $score[c_j] = \ln p\,(c_j)$
   4. Для каждого $w_t \in V$ делать
      5. Если $w_t \in V_d$, тогда $score[c_j] \mathrel{+}= \ln p\,(w_t|c_j)$
6. Если $score[c_1] > h \cdot score[c_2]$, тогда $d \in c_1$, иначе $d \in c_2$

Выход: тональность сообщения d

Рис. 4. Алгоритм работы классификатора

Он состоит из обучающей части и непосредственно классификации сообщения. В обучающей части создается словарь термов *V*, находятся вероятности $p(c_j)$ и вероятности $p(w_t \mid c_j)$, настраивается пороговое значение h для минимизации ошибки классификации. В классификационной части производится распознавание тональности сообщений.

Алгоритм бустинга

С целью повышения точности классификации тональности были рассмотрены алгоритмы построения ансамблей классификаторов – бустинг и бэггинг.

Бустинг является мета-алгоритмом машинного обучения ансамблей классификаторов, позволяющий повысить точность классификации, предложенный Й. Фройндом и Р. Шапиро [15]. Бустинг является адаптивным алгоритмом в том смысле, что каждый следующий классификатор строится по объектам, неверно классифицированным предыдущими классификаторами. Алгоритм бустинга применяется для линейных классификаторов.

Алгоритм одной из модификаций бустинга – AdaBoost – представлен на рис. 5. Алгоритм начинается с инициализации начальных весов $w_t(i)$. Далее строится классификатор h_t с минимальной взвешенной ошибкой ε_t. Затем находятся значения α_t и обновляются веса $W_{t+1}(i)$. Начинается следующая итерация. И так до тех пор, пока все *T* классификаторов не будут обучены.

Анализ алгоритма бустинга показал, что он хорошо работает с классами, которые имеют сложную нелинейную границу. На рис. 6 представлено визуальное распределение позитивных и негативных сообщений в зависимости от вероятности их принадлежности к тому или иному классу. По оси *X* откладывается вероятность принадлежности сообщения к положительному классу. По оси *Y* – вероятность принадлежности сообщения к негативному классу. Из рисунка видно, что, в случае классификации тональности с помощью наивного байесовского классификатора, классы имеют строгую линейную разделимость. Построение дополнительных классификаторов не снижает ошибку классификации. Поэтому из-за неприменимости мета-алгоритма бустинга в дальнейшем использовался мета-алгоритм бэггинга.

Вход: $(d_1, y_1), (d_2, y_2), \dots, (d_m, y_m)$, где $d_i \in D$ - множество размеченных сообщений;
$y_i \in Y = \{+1, -1\}$, $y_i = +1$, если позитивная и $y_i = -1$, если негативная тональность текста;
T- число классификаторов в ансамбле

1. Инициализировать веса $W_1(i) = \frac{1}{m}, i = \overline{1, m}$
2. Для $t = 1$ *до* T делать
   3. Построить классификатор $h_t: D \to \{+1, -1\}$ с минимальной взвешенной ошибкой $h_t = argmin_{h_j \in H} \varepsilon_j$, $\varepsilon_j = \sum_{i=1}^{m} W_t(i) \cdot [y_i \neq h_j(d_i)]$
   4. Если $\varepsilon_t \geq 0{,}5$, то СТОП
   5. Найти $\alpha_t = \frac{1}{2} ln \frac{1-\varepsilon_t}{\varepsilon_t}$, где ε_t - взвешенная ошибка классификатора h_t
   6. Обновить веса $W_{t+1}(i) = \frac{W_t(i) e^{-\alpha_t y_t h_t(d_t)}}{Z_t}$, где Z_t – нормализующий параметр

Выход: $H(d) = sign(\sum_{t=1}^{T} \alpha_t h_t(d))$ ансамбль классификаторов

Рис. 5. Алгоритм AdaBoost

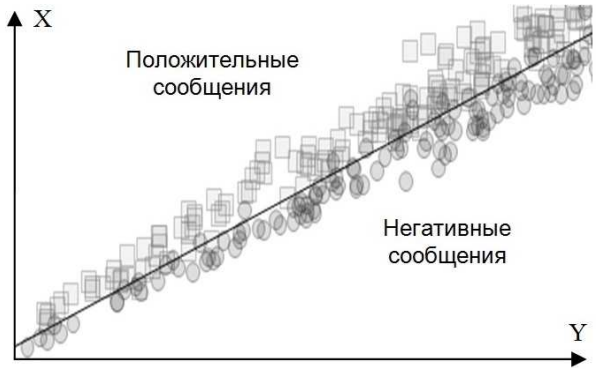

Рис. 6. Распределение классов положительных и негативных сообщений

Алгоритм бэггинга

Другой алгоритм повышения качества классификации называется бэггингом. Впервые был предложен Л. Брейманом Негативные сообщения работе [16]. Алгоритм бэггинга пр рис. 7.

Из исходного обучающего множества сообщений *D* длины |*D*| формируются различные обучающие подмножества сообщений D_t той же длины |*D*| с помощью бутстрепа – случайного выбора с возвращениями. При этом некоторые сообщения попадают в подмножества по нескольку раз, некоторые – ни разу. Далее строятся контрольные множества сообщений вычитанием из множества *D* подмножества D_t.

Вход: $(d_1, y_1), (d_2, y_2), \dots, (d_m, y_m)$, где $d_i \in D$ - множество размеченных сообщений;
$y_i \in Y = \{+1, -1\}$, $y_i = +1$, если позитивная и $y_i = -1$, если негативная тональность текста;
T –число классификаторов в ансамбле;
$|D|$ - длина обучающей выборки;
e – допустимая ошибка классификации

1. Для $t = 1$ *до* T делать
   2. Случайно выбрать $|D|$ сообщений из D для построения обучающего множества D_t
   3. Построить контрольное множество сообщений из множества D/D_t сообщений
   4. Построить классификатор $h_t: D \to \{+1, -1\}$
   5. Оценить ошибку e_t классификатора h_t на контрольном множестве
   6. Если $e \geq e_t$ тогда добавить классификатор h_t в ансамбль

Выход: $H(d) = \text{sign}(\sum_{t=1}^{T} h_t(d))$ ансамбль классификаторов

Рис. 7. Алгоритм бэггинга

По обучающим подмножествам строятся классификаторы h_t. Их ошибка классификации e_t оценивается по контрольной подвыборке и затем сравнивается с допустимой ошибкой классификации e. Если ошибка построенного классификатора меньше допустимой ошибки, то он добавляется в ансамбль. Классификация сообщений производится ансамблем построенных и допущенных классификаторов с помощью простого голосования.

4. РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТА

На базе рассмотренных алгоритмов авторами была разработана программа «Text Analyzer» на языке программирования C#. Для обучения и оценки точности классификации использовался тестовый набор, состоящий из отзывов клиентов российских банков взятых с интернет-сайта [13]. Он включает 304 положительных отзыва и 850 негативных отзывов на русском языке. Примером отзыва о банке с положительной тональностью является отрывок: «Заявку на кредит оформил быстро, без лишних вопросов, в течение 20 минут». Пример отзыва с негативной тональностью: «Рассмотрение заявки заняло по времени два месяца». Пример работы программы представлен на рис. 8.

В предварительную обработку текста входила лемматизация всех встречающихся слов. Лемматизацией называется приведение различных слов к их начальной форме, например, для существительного это именительный падеж, единственное число. Мотивация к лемматизации текста обусловлена тем фактом, что различные строки могут зачастую выражать один и тот же смысл. В связи с этим оправдано приведение слов к единой форме. Были использованы библиотеки LemmaGen, написанные на языке C# и предназначенные для лемматизации слов. Данные библиотеки доступны на интернет-странице разработчика [14].

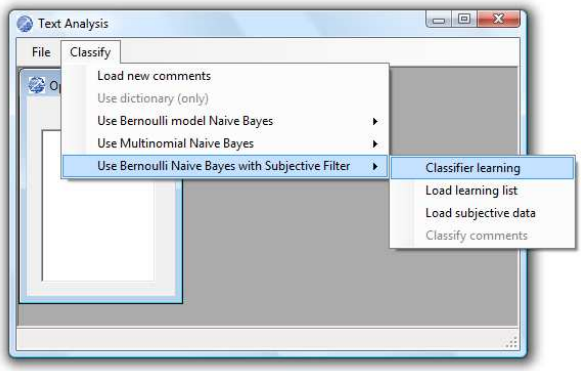

Рис. 8. Интерфейс программы

Для оценки обобщающей способности алгоритма использовался скользящий контроль или кросс-проверка. Фиксировалось множество, состоящее из 10 разбиений исходной выборки, каждое из которых в свою очередь состояло из двух подвыборок: обучающей и контрольной. Для каждого разбиения выполнялась настройка алгоритма по обучающей подвыборке, затем оценивалась его средняя ошибка на объектах контрольной подвыборки. Оценкой скользящего контроля являлась средняя по всем разбиениям величина ошибки на контрольных подвыборках. Для алгоритма бэггинга принята допустимая ошибка классификатора равная e = 25 %.

Для оценки точности классификации каждого контрольного блока используется показатель «точность классификации», который вычисляется по формуле:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \cdot 100\%, \quad (18)$$

где TP – число верно классифицированных положительных отзывов; TN – число верно классифицированных негативных отзывов; FP – число неверно классифицированных положительных отзывов; FN – число неверно классифицированных негативных отзывов.

Результаты вычислительных экспериментов приведены в табл. 2. Наивные байесовские классификаторы, реализованные по мультиномиальной модели и модели Бернулли, показали

приблизительно одинаковую точность классификации, 86,49 и 86,82 % соответственно. Применение алгоритма бэггинга дало прирост точности на 0,33 % для модели Бернулли и 0,86 % для мультиномиальной модели. На рис. 9 изображен график зависимости точности классификации от числа классификаторов в ансамбле. На графике видно, что максимальная точность классификации достигается при числе классификаторов от 8 до 18. Дальнейшее увеличение этого числа ведет к снижению точности классификации.

ЗАКЛЮЧЕНИЕ

Авторами было установлено, что применение алгоритма бэггинга для построения ансамбля классификаторов на основе наивного байесовского обучения повышает точность классификации тональности отзывов. Оптимальное число классификаторов в ансамбле, при котором достигается максимальная точность классификации, равно тринадцати. Сравнение двух способов реализации наивного байесовского обучения для рассматриваемой задачи показало, что мультиномиальная модель имеет большую точность классификации на используемых наборах данных.

Задача автоматической классификации тональности текстовых сообщений имеет сложную природу и требует нестандартных подходов к ее решению. Сложность ее природы заключается в том, что исходными данными являются тексты на естественном языке. Каждое слово такого текста несет свой смысл, а комбинация слов представляет собой сложное взаимодействие смысла каждого слова. В настоящее время не существует универсального метода моделирования такого взаимодействия на языке машины или на языке чисел.

Таблица 2

Алгоритм	Точность, %
Модель Бернулли	86,49
Мультиномиальная модель	86,83
Бэггинг (модель Бернулли)	86,82
Бэггинг (мультиномиальная модель)	87,69

Несмотря на сложность задачи, она привлекает большое количество исследователей по всему миру. Поиски в данной области активно ведутся, и есть некоторые достижения. Многие разработанные алгоритмы достигают точности классификации более 85 %. Но нужно учитывать, что эти результаты получены на тестовых данных в условиях эксперимента. Официальной информации о реальном успешном практическом применении систем, решающих подобную задачу, к сожалению, пока нет.

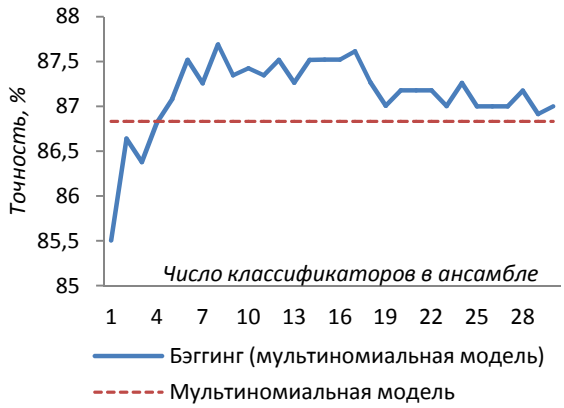

Рис. 9. График зависимости

Дальнейшее исследование будет продолжено в рамках научных работ, проводимых по теме «Разработка инструментальных средств поддержки принятия решений для различных видов управленческой деятельности в промышленности в условиях слабоструктурированной информации на основе технологий распределенного искусственного интеллекта» по заданию Министерства образования РФ на проведение научных исследований, поддержано грантами РФФИ 12-07-00377-а, 11-07-00687-а, 12-07-00377-а «Алгоритмическое и программное обеспечение поддержки принятия решений в задачах управления сложными социально-экономическими системами при наличии слабо структурированных данных», 11-07-00687-а «Интеграция интеллектуальных информационных технологий на примере мониторинга банкротств» и будет посвящено выявлению лингвистических особенностей русского языка с целью их использования для увеличения точности классификации. Отдельный интерес представляет комбинирование подходов, основанных на лингвистике и на машинном обучении.

СПИСОК ЛИТЕРАТУРЫ

1. Marketing research of cosumer opinions with using information technologies / M. Boyko [et al.] // Proc. of the 13th Intern. Workshop on Computer Science and Information Technologies, Germany. 2011. P. 103–105.

2. **Пазельская А. Г., Соловьев А. Н.** Метод определения эмоций в текстах на русском языке //

Компьютерная лингвистика и интеллектуальные технологии: сб. научных статей. Вып. 10 (17). М.: Изд-во РГГУ, 2011. С. 510–522.

3. Официальный сайт компании «Эр Си О». Компонент определения тональности текста [Электронный ресурс] (http://x-file.su/tm/Default.aspx)

4. **Ермаков А. Е., Киселев С. Л.** Лингвистическая модель для компьютерного анализа тональности публикаций СМИ. Компьютерная лингвистика и интеллектуальные технологии: междунар. конф. Диалог'2005. М.: Наука, 2005.

5. **Yi J. Nasukawa T., Niblack W., Bunescu R.** Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques // Proc. of the 3rd IEEE international conference on data mining (ICDM 2003) Florida. USA. P. 427–434.

6. **Pang B., Lee L.** Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval. 2. No. 1–2 (2008). P. 1–135.

7. **Pang B., Lee L.** Thumbs up? Sentiment Classification using Machine Learning Techniques. Proceedings of the Conference on Empirical Methods in Natural. Language Processing (EMNLP). Philadelphia. 2002. P. 79–86.

8. **Pang B., Lee L.** A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts // Proc. of the ACL. 2004.

9. **O'Keefe T., Koprinska I.** Feature selection and weighting methods in sentiment analysis // Australasian Document Computing Symposium. 2009.

10. **Prabowo R., Thelwall M.** Sentiment analysis: A combined approach // Journal of Informatics. 2009.

11. **König A., Brill E**. Reducing the Human Overhead in Text Categorization // Proc. of KDD. 2006.

12. **Manning C., Raghavan P., Schuetze H.** An Introduction to Information Retrieval. Cambridge University Press. Cammbridge, England (2009).

13. Интернет-портал, посвященный российским банкам [Электронный ресурс] (http://banki.ru).

14. Портал, посвященный лемматизации [Электронный ресурс] (http://lemmatise.ijs.si/Software/Version3).

15. **Freund Y., Schapire R.** Experiments with New Boosting Algorithm // Machine Learning: Proc. of the 13th Intern. Conference. 1996. P. 148–156.

16. **Breinman L.** Bagging Predictors // Machine Learning. 24. 1996. P. 123–140.

ОБ АВТОРАХ

Юсупова Нафиса Исламовна, проф., зав. каф. вычислительн. математики и кибернетики, декан факультета информатики и робототехники. Дипл. радиофизик (Воронежск. гос. ун-т, 1975). Д-р техн. наук по управлению в техн. системах (УГАТУ, 1998). Иссл. в обл. критическ. ситуационного управления, информатики.

Богданова Диана Радиковна, доц. той же каф. Канд. тех. наук (УГАТУ, 2008). Дипл. экономист-математик (УГАТУ, 2005). Иссл. в обл. управления в социальных и экономических системах.

Бойко Максим Викторович, асп. той же каф. Дипл. экономист-математик (УГАТУ, 2011). Иссл. в обл. управления в социальных и экономических системах.