УДК 004.048 DOI: 10.26102/2310-6018/2020.30.3.016

Модели и методы анализа тональности в текстах на башкирском языке

А. К. Сулейманов, М. А. Шарипова, О. Н. Сметанина, Е. Ю. Сазонова, К. В. Миронов

Федеральное государственное бюджетное образовательное учреждение высшего образования «Уфимский государственный авиационный технический университет», Уфа, Российская Федерация

Резюме: Исследования в области автоматического извлечения мнений по-прежнему остаются актуальными. В статье представлено формальное описание термина мнение, постановки задач в зависимости от определяемых свойств мнения. Описаны проблемы решения задачи анализа тональности текста, подходы к ее решению и готовые программные реализации. Приведены имеющиеся корпусы текстов на башкирском языке, а также постановка задачи анализа тональности в текстах на башкирском языке. Рассмотрена методика ее решения, включающая алгоритм разметки текста, методы предобработки, выбора признаков классификации, методы классификации, приведены результаты эксперимента с целью выбора наиболее эффективного метода классификации для программной реализации с учетом метрик качеств. Полученные в работе результаты и разработанное программное решение на основе SVM со стохастическим градиентным спуском, продемонстрировавшим наиболее высокие показатели в критериях точности, полноты и *F*-меры, могут быть использованы для оценки тональности текстов новостных сайтов на башкирском языке.

Ключевые слова: анализ тональности текста, компьютерная лингвистика, машинное обучение, признаки классификации, гибридный подход, метод опорных векторов, случайный лес.

Для цитирования: Сулейманов А. К., Шарипова М. А., Сметанина О.Н., Сазонова Е. Ю., Миронов К. В. Модели и методы анализа тональности в текстах на башкирском языке. *Моделирование, оптимизация и информационные технологии.* 2020;8(3). Доступно по: https://moit.vivt.ru/wp-content/uploads/2020/08/SuleimanovSoavtors_3_20_1.pdf DOI: 10.26102/2310-6018/2020.30.3.016

Models and methods for sentiment analysis of texts in Bashkir language

A.K. Suleymanov, M.A. Sharipova, O.N. Smetanina, Y.Y. Sazonova, K.V. Mironov

Federal State Budgetary Educational Institution of Higher Education "Ufa State Aviation Technical University", Ufa, Russian Federation

Abstract: The research works on automatic opinion extraction are still relevant. The article presents a formal description of the term opinion, setting tasks depending on the determined properties of opinion. The problems of solving the tasks of sentiment analysis, approaches to its solution and ready-made software implementations are described. Available corpora of texts in the Bashkir language are presented, and also task statement for sentiment analysis in the Bashkir language. Presented solution, which include an algorithm for tagging the texts, a preprocessing algorithm, a choice of classification features, and classification algorithms. Also, the results of computational experiment, which aimed to define the most effective classifier based on quality metric, are present. The results in this work and the developed software solution based on SVM with stochastic gradient descent, which demonstrated the highest indicators in the criteria of accuracy, completeness, and *F*-measure, can be used to

sentiment analysis of news sites in the Bashkir language. The results of the research presented in this article were supported by Grants RFBR 19-07-00709, 20-08-00668 and Ministry of Science and Higher Education of the Russian Federation in the framework of the work under the State Assignment of Ufa State Aviation Technical University # FEUE-2020-0007.

Keywords: sentiment analysis, computational linguistics, machine learning, classification features, hybrid intelligent system, support vector machine, random forest.

For citation: Suleymanov A.K., Sharipova M.A., Smetanina O.N., Sazonova E.Y., Mironov K.V. Models and Methods for Sentiment Analysis of Texts in Bashkir Language. *Modeling, Optimization and Information Technology*. 2020;8(3). Available from: https://moit.vivt.ru/wp-content/uploads/2020/08/SuleimanovSoavtors_3_20_1.pdf DOI: 10.26102/2310-6018/2020.30.3.016 (In Russ).

Введение

Исследования в области автоматического извлечения мнений по-прежнему остаются актуальными. Это связано увеличением количества размеченных текстов на разных языках, развитием моделей и методов решения задач, а также готовых программных решений.

Возникающие проблемы при решении задач рассматриваются в рамках проводимых исследований в области анализа тональности текста, как зарубежными, так и российскими специалистами Beltiukov A. [27], Kadam S. [29], Joglekar S. [29], Liu B. [31], Гаршина В.В. [6], Ермаков А.Е. [7], Киселев С.Л. [7], Лукашевич Н.В. [12, 13], Пазельская А.Г. [17], Посевкин Р.В. [18], Соловьев А.Н. [17], Толкунов А.А. [22]и др.

Несмотря на широкий диапазон уже исследуемых вопросов в этой области, возникают новые задачи и потребности, как в части совершенствования инструментария для решения задач, так и в части решения задачи оценки тональности в текстах на языках, для которых еще не проводился анализ, с учетом их особенностей.

В статье описаны проблемы решения задачи анализа тональности текста и подходы к ее решению, а также готовые программные решения. Рассмотрены постановка задачи анализа тональности в новостных текстах на башкирском языке и методика решения. Описаны имеющиеся корпусы текстов на башкирском языке, приведены алгоритмы и методы разметки текста, предобработки, выбора признаков классификации, результаты эксперимента с целью выбора наиболее эффективного метода классификации для программной реализации на основе метрик качества. Полученные в работе результаты могут быть использованы для оценки тональности текстов новостных сайтов на башкирском языке.

Современное состояние проблемы анализа тональности текста

Согласно [31] мнение в задаче оценки тональности текста формально может быть представлено объектом, его свойствами, тональной оценкой, субъектом, моментом времени (Рисунок 1). Однако есть задачи, требующие более полного описания характеристик мнения [22]. Наличие нескольких характеристик формального представления мнения в задаче анализа тональности позволяет оперировать ими и формулировать разные постановки задач. Необходимо учитывать, что задачи анализа тональности текста требуют, в зависимости от ее постановки (Таблица 1), самих данных, и связанных с ними проблем (Рисунок 2), применения различных подходов к решению задач (Рисунок 3).

При разработке подходов к решению задачи анализа тональности текста различают ряд проблем (Рисунок 2) [5]. В основу тех или иных подходов положены

принципы (как правило, лингвистические), методы и модели (теоретико-графовые модели [24], векторная модель слова [8, 34], методы предобработки текста (удаление стоп-слов, стемминг), алгоритмы обучения нейронной сети и многие другие), ограничения (например, длинный или короткий текст [3, 4], временные ограничения [1, 12]).

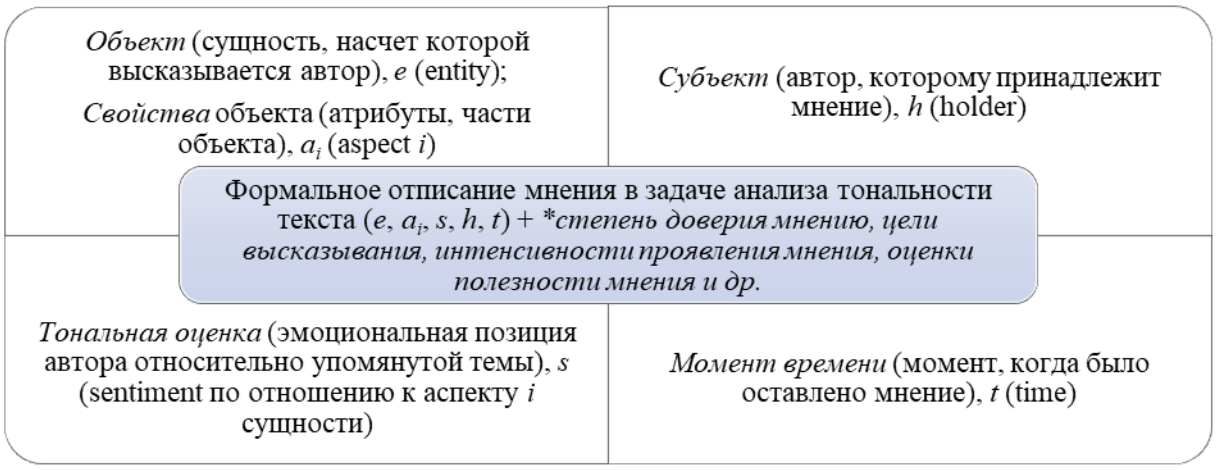

Рисунок 1 – Формальное представление мнения в задаче анализа тональности текста
(* – дополнительные характеристики формализации термина мнение)
Figure 1 – Formal representation of opinion in the problem of text sentiment analysis
(* - additional characteristics of the formalization of term opinion)

Таблица 1 – Постановка задач анализа тональности текста
Table 1 – Statement of text sentiment analysis tasks

№	Задача	Необходимо найти характеристики мнения, [2] - характеристики *x* можно не выявлять	Авторы
1	Определение исключительно тональности: Дано: D^1 - <d_1, d_2, …, d_n>, [1]множество данных	(x^2, x^2, *s*-?, x^2, x^2)	Краснов Ф. В. [11], Васильева М. И., Куртукова А.В., Лопатин Д. В., Мещеряков Р. В., Романов А. С. [19], Чиркин Е. С. [26]
2	Определение субъекта оценки: Дано: D^1 - <d_1, d_2, …, d_n>	(x^2, x^2, *s*-?, *h*-?, x^2)	Wiebe J., Wilson T., Cardie C. [35]
3	Определение тональности, субъекта и объекта: Дано: D^1 - <d_1, d_2, …, d_n>	(*e*-?, x^2, *s*-?, *h*-?, x^2)	Минина М. А. [15], Толкунов А.А. [22]
3	Определение мнения в целом: Дано: D^1 - <d_1, d_2, …, d_n>	(*e*-?, a_i-?, *s*-?, *h*-?, *t*-?)	Ананьева М. И., Кобозева М. В., Поляков И. В., Соловьев Ф.Н., Чеповский А.М. [1], Лукашевич Н.В. [12].

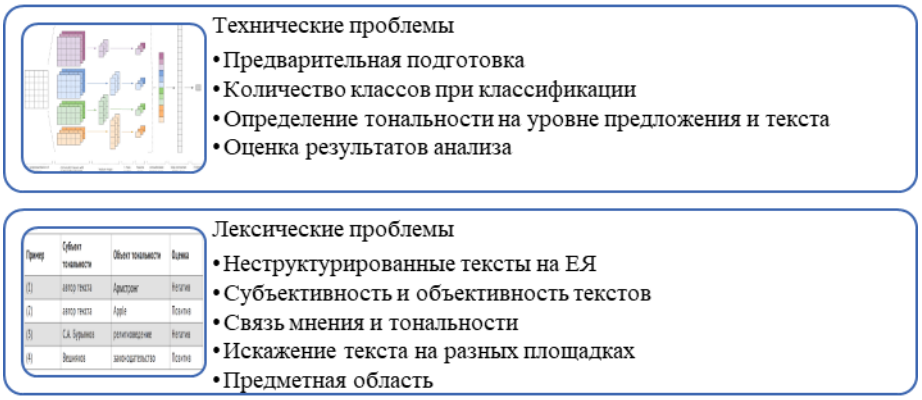

Рисунок 2 – Проблемы анализа тональности текста
Figure 2 – Problems of text sentiment analysis

При классификации, как отмечает ряд авторов [9, 20], следует обозначить методы, основанные на правилах; методы, основанные на машинном обучении; гибридные методы и методы, основанные на теоретико-графовых моделях (Рисунок 3).

1. Методы, основанные на правилах и словарях [8, 20, 23, 28, 34]	•генерация *правил* для определения тональность текста (текст разбивается на слова или последовательности слов (N-grams); полученные данные используются для выделения часто встречающихся шаблонов с присвоением положительной или отрицательной оценки). Выделенные шаблоны применяются при создании правил вида «ЕСЛИ условие, ТО заключение»; •*тональные словари* составляются заранее.
2. Методы, основанные на машинном обучении [19, 20,30]	•*обучение с учителем* (требуется наличие обучающей коллекции размеченных в рамках эмотивного пространства текстов для построения статистического или вероятностного классификатора (например, байесовского); •*обучение без учителя* (наибольший вес в тексте имеют термины, которые чаще встречаются в тексте и в то же время присутствуют в небольшом количестве текстов всей коллекции; выделив термины и определив их тональность, можно сделать вывод о тональности всего текста.
3. Гибридные методы [10, 32]	•сочетает все или несколько методов 1 и 2, в основе применение классификаторов с их использованием.
4. Методы, основанные на теоретико-графовых меделях [24]	•*предположение*: не все слова в текстовом корпусе документа равнозначны, т.е. имеют разный вес и по разному влияют на тональность текста; •*этапы*: построение графа на основе исследуемого текста; ранжирование его вершин; классификация найденных слов; вычисление результата.

Рисунок 3 – Классы методов для оценки тональности текстов
Figure 3 – Classes of methods for sentiment analysis

Количество классов тональности системы связано с поставленными задачами и шкалами (например: «положительная»/ «отрицательная»; оценка может быть дополнена усилительными словами «сильно» / «слабо»; может быть определено численное

значение, например, «-5»-«+5» и «0» для нейтральной оценки). Меньшиков И. Л., Кудрявцев А. Г. [14] описывают ряд наиболее известных систем в области анализа тональности в текстах, имеющих возможность работать с текстами на русском языке и для сравнения – на английском языке (Рисунок 4).

Программная разработка исследователей Стенфордского университета (язык английский)	•*В основе:* обучение на корпусе оценочных высказываний (в виде деревьев синтаксического разбора); глубокий синтактико-семантический анализ входящего текста. •*Выход*: результаты анализа тональности в отношении каждого объекта внутри предложения. •Скорость обработки у компонента анализа тональности - 1.5 кб/сек; точность классификации тональности на текстах Твиттера - 62-68% и на негативных комментариях из социальных сетей – 37%,; положительных комментариях - 95-96%; покрытие - 50-70%.
Аналитический курьер и X-files/ Компания «Ай-Теко» http://www.iteco.ru/solutions/business_intelligence_products/analytical_courier [16] (язык русский)	•*В основе*: метод на словарях и правилах (шкала: позитивный /негативный /нейтральный). •*Выход*: массив размеченных предложений (объекты тональности (при наличии) и цепочка слов, несущая в себе тональность по отношению к ним); подсчитана тональность для каждого предложения. •*Недостатки*: отсутствие количественной оценки текста. •Скорость обработки у компонента анализа тональности - в 30-75 кб/сек на больших предложениях и в 7.5-9 кб/сек - на коротких предложениях (твиттах); точность классификации тональности на текстах Твиттера - 94-94% и на негативных комментариях из соц. сетей - 72%, на положительных комментариях - 95-96%; покрытие - 30-40%.
Senti Strength/ http://sentistrength.wlv.ac.uk (языки английский, греческий, немецкий, русский и др.)	•Возможна конфигурация для работы с текстами на ряде языков. •*В основе*: поиск максимального значения тональности в тексте для каждой шкалы. При работе учитывается простейшее взаимодействие слов и идиоматические выражения. •*Выход*: 2 оценки – позитивная и негативная, шкалы могут быть разными. •*Недостатки*: реализованные в системе алгоритмы не учитывают специфику русского языка, считает лишь общую тональность текста.
Ваал/ В. Шалак http://www.vaal.ru (язык русский)	•*В основе*: превращение текста в частотный словарь и отнесение слов к определенным психолингвистическим категориям. •*Выход*: набор оценок по ряду критериев, относящихся к данному тексту/слову и т.д. • *Недостатки*: нет анализа семантики текста; рекомендуется к использованию только специалистами в области психолингвистики.
Компонент в системе RCO Fact Extracto/ RCO http://www.rco.ru/?page_id=3554 (язык русский)	•*В основе*: подход, основанный на правилах; учитывает синтаксическую структуру текста и взаимодействие различных типов слов. •*Выход*: оценка общей тональности текста на основе тональностей всех входящих в него пропозиций

Рисунок 4 – Системы в области анализа тональности в текстах
Figure 4 – Systems in the field of sentiment analysis in texts

Анализ других систем показал отсутствие систем анализа тональности в текстах на башкирском языке.

Постановка задачи и ее решение для анализа тональности в текстах на башкирском языке

В последнее время в области создания корпусов текстов на башкирском языке работают несколько научных групп [2, 21, 25]. Наиболее близкими могли бы быть корпусы [21], однако авторами было решено самостоятельно собрать и разметить данные с новостных сайтов на башкирском языке (Башинформ и Bashnews). Сбор осуществлялся специально разработанной программой-роботом. Собираемые тексты новостей сохранялись в файл формата csv.

Постановка задачи.

Дано: новостные сайты на башкирском языке с множеством новостей D–<d_1, d_2, …, d_n>, множество категорий классов K–<отрицательный, положительный>.

Необходимо построить: классификатор Φ:K×D→{0,1}.

Для решения задачи анализа тональности в текстах на башкирском языке использован гибридный подход, включающий методы машинного обучения с учителем на основе обучающей выборки в виде набора объектов (новости на башкирским языке) для классификации и соответствующих им меток тональности по шкале «положительная»/ «отрицательная».

При составлении обучающей выборки объекты представляются в виде n-мерных векторов, где n – количество признаков, используемых для классификации. Далее происходит обучение классификатора и его валидация на новых данных (в векторном виде). По результатам валидации может потребоваться настройка параметров классификатора, изменение размеров обучающей выборки.

В процессе разработки возможно использование множества классификаторов, и после их сопоставления выбирается один с наиболее высокими показателями качества классификации.

Решение задачи по обучению алгоритма классификации можно представить последовательностью этапов (Рисунок 5).

Рисунок 5 – Этапы обучения алгоритма классификации и проведение экспериментов
Figure 5 – Stages of training classification algorithm and conducting experiments

Этап 1. Для осуществления разметки составлен словарь тональности позитивных (всего 650, например: көслө, ҡыҙыҡ, ихтирам) и негативных (всего 980, например, ауырыу, насар, боҙолған) слов на башкирском языке (частично переведены с

английского [33]). Далее осуществляется разметка тональности новостей (позитивная, негативная). В основу положен метод, с использованием правил и словарей. Для разметки текста используется алгоритм (Рисунок 6).

В процессе разметки получено позитивных меток 22209, негативных – 4404. Во избежание несбалансированности классов для обучения использовано одинаковое количество позитивных и негативных новостей.

Этап 2. Среди методов предобработки текста рассмотрены приведение к нижнему регистру; удаление символов, не являющихся буквами, удаление стоп-слов, стемминг (Рисунок 7).

Метод стемминг для башкирского языка не очень эффективен, поскольку часть «отрицательных» слов образуются путем добавления окончаний. Слова «аҡыллы» («мудрый»), «аҡылһыҙ» («глупый») могут быть преобразованы в основу «аҡыл» («ум»), что в контексте недопустимо.

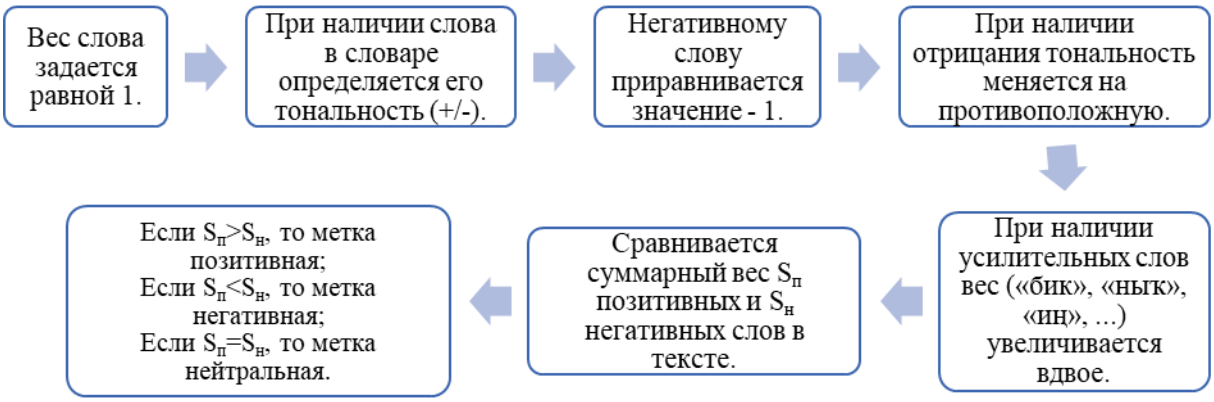

Рисунок 6 – Алгоритм разметки текста
Figure 6 – Algorithm for tagging the texts

Рисунок 7 – Методы предобработки текста
Figure 7 – Text preprocessing methods

Этап 3. Этап выбора признаков для классификации опирается на метод представления текста в векторном виде Bag-of-words и статистический показатель TF-

IDF, отражающий значимость слова в документе из некоторой коллекции документов или корпуса (Таблица 2).

Таблица 2 – Характеристики методов
Table 2 – Characteristics of the methods

Методы	Характеристика
Bag-of-words	- Текст задается как неупорядоченный набор терминов без указания связей между ними. - Термины представлены *n*-граммами или другими последовательностями символов. - Строится словарь выборки, включающий уникальные слова всех текстов. - Документы представляются в виде матрицы термин-документ, что позволяет получить векторные документы, например: d_i=[1,1,1,1,1,0,0]. - Векторные документы получаются разреженными, поскольку в тексте редко встречаются все слова из словаря выборки, что требует ограничения размерности словаря выборки.
TF-IDF	Мера TermFrequency (TF) – частотность термина вычисляется как $tf(t,d) = \frac{n_i}{\sum_k n_k}$ –отношениечисла вхождений термина в документ d к количеству равнозначных слов документа. Мера InverseDocumentFrequency (IDF) – снижения важности часто употребляемых, не влияющих на смысл документа $idf(t,D) = log\frac{\lvert D\rvert}{\lvert\{d_i \in D \lvert t \in d_i\}\rvert}$ слов, определяемая через логарифм (с произвольным основанием) отношения количества документов в коллекции к числу документов из коллекции, в которых содержится термин t. Мера TF-IDF – есть произведение: $tf\text{–}idf(t,d,D) = tf(t,d) * idf(t,D)$.

С помощью модели векторного представления документов можно получить наборы векторов, которые впоследствии могут быть сравнены между собой с помощью вычисления расстояний (например, Евклидово) между ними.

Этап 4. Выбор и реализация алгоритмов классификации. К таким алгоритмам можно отнести: наивный байесовский классификатор, метод опорных векторов (SVM), случайный лес (Таблица 3).

Этап 5. Выбор метрик для оценки качества классификации. Результаты классификации тональности текста оцениваются в терминах точности (характеризует, сколько из классифицированных системой положительных ответов действительно являются истинными, $Precision = \frac{TP}{TP+FP}$), полноты (характеризует, все ли истинные ответы вернула система, $Recall = \frac{TP}{TP+FN}$) и F-меры (гармоническое среднее между полнотой и точностью, $F = 2\frac{Precision*Recall}{Precision+Recal}$), где *TP*, *FP*, *FN*, *TN* – типы объектов в результатах классификации (Таблица 4).

Чем выше показатели полноты и точности, тем лучше качество классификации. *F*-мера может использоваться с учетом приоритета той или иной меры (полнота может быть важнее точности, например, в задаче выдачи релевантных новостей). Тогда можно рассчитать *F*-меру, установив важность каждой из метрик $F = (\beta^2 + 1)\frac{Precision*Recall}{Precision+Recal}$, где β принимает значения в диапазоне $0 < \beta < 1$, если приоритет отдается точности, а

при β>1 приоритет отдается полноте. При β=1 формула сводится к ранее приведенной, с получением сбалансированной *F*-меры (илиF_1).

Таблица 3 – Характеристики методов классификации

Table 3 – Characteristics of classification methods.

Методы	Характеристика
Наивный байесовский классификатор	Метод эффективен для классификации с большим количеством признаков. *Дано*: документ $d=\{w_1, w_2, \ldots, w_n\}$, где w_i - вес *i*-ого термина, *n*– размер словаря выборки, класс *c*. Выбор наиболее вероятного класса (*max* условной вероятности принадлежности документа *d* классу *c*) $c^* = argmax_c P(w_1, w_2, \ldots, w_n|c) * P(c)$.При условной независимости признаков$P(w_1|c)*(w_2|c)*\ldots*(w_n|c)=\Pi_i P(w_i|c_j)$. Условные вероятности принадлежности документа *d* определяются для каждого из классов и выбирают класс, имеющий *max* вероятность $C_{NB} = argmax_c\ [P(c_j) * \prod_i P(w_i|c_j)\ ]$.При большом количестве слов в документах используется свойство логарифма произведения: $C_{NB} = argmax_c\ [\ln\left(P(c_j)\right) + \sum_{i=1}^{n} \ln\left(P(w_i|c_j)\right)]$.Оценка вероятностей событий осуществляется на обучающей выборке. Вероятность класса $P(c)$ оценивается по формуле:$P(c)=D_c/D$, где D_c- количество документов класса *c*, *D* – общее количество документов в обучающей выборке.Условные вероятности для признаков определяются отношением количества терминов w_iв классе c_j кобщему количеству терминов в этом классе: $\hat{P}(w_i|c_j) = \frac{count\ (w_i,c_j)}{\sum_{w\in V} count\ (w,c_j)}$, где *V* – словарь обучающей выборки. При обнаружении отсутствующего слова в обучающей выборке, но присутствующего в документе при его классификации, они будут относиться к любому из классов с вероятностью =0. Проблему решают путем добавления 1 к частотам появления терминов словаря), что позволяет классифицировать тексты. $\hat{P}(w_i|c) = \frac{count(w_i,c)+1}{\sum_{w\in V}(count(w,c)+1)} = \frac{count(w_i,c)+1}{(\sum_{w\in V}(count(w,c))+|V|}$,где $|V|$- количество слов в словаре обучающей выборки.
Метод опорных векторов (Support Vector Machine, SVM)	*В основе* лежит построение гиперплоскости для оптимального разделения объектов обучающей выборки $(x_1, y_1), \ldots, (x_k, y_k), x_i \in \mathbb{R}^n$ на два класса $y_i \in \{-1,1\}$. Классифицирующая функция: $F(x) = sign(< w, x > +b)$, где$< w, x >$ – скалярное произведение, w – нормальный вектор к разделяющей плоскости, b – вспомогательный параметр. Один класс – объекты со значениемфункции $F(x) = 1$, другой класс - объекты с $F(x) = -1$. Любая гиперплоскость задается в виде $< w, x > +b = 0$, для некоторых w и b, выбираемых для максимизации расстояния от гиперплоскости до объектов каждого класса $\frac{1}{||w||}$. Учитывая, что проблемы нахождения *max*$\frac{1}{||w||}$ и нахождения *min*$||w||^2$ аналогичны, можно записать задачу оптимизации: $\begin{cases} arg\ \min\limits_{w,b} ||w||^2, \\ y_i(< w, x > +b) \geq 1, i = 1, \ldots, m \end{cases}$,и ее решение с помощью множителей Лагранжа. Стохастический градиентный спуск используют для данных большого объема.
Случайный лес	*В основе*: дерево решений. *Цель построения дерева решений*: создание модели для классификации и решения о значениях целевой функции нескольких переменных. Узлы дерева решений (не листья), содержат атрибуты, по которым различаются случаи. В

Методы	Характеристика
	листьях находятся значения целевой функции. Спускаясь по ребрам, можно классифицировать различные случаи. *Особенности*: Верхние уровни дерева решений влияют на итоговый результат, поэтому используется модель на основе *случайного леса*. Обучающая выборка делится на подмножества. Для каждого из подмножеств строится дерево решений. Итог - ансамбль деревьев для классификации. Объект для классификации прогоняется через все деревья. Каждое дерево голосует за принадлежность объекта к определенному классу. Классификация происходит путем выбора большинства голосов.

Таблица 4 – Матрица ошибок
Table 4 – Error matrix

Документ		**Оценка эксперта**	
		Positive	Negative
Оценка системы	Positive	*TP* (TruePositive) - истинно-положительные (классифицированные системой и экспертом как положительные)	*FP* (FalsePositive) – ложноположительные (классифицированные экспертом как отрицательные, а системой - положительны)е
	Negative	*FN* (FalseNegative) – ложноотрицательные (классифицированные экспертом как положительные, системой – отрицательные	*TN* (TrueNegative) – истинно-отрицательные (классифицированные системой и экспертом как отрицательные)

Для тестирования качества классификации эмоциональной окраски новостей был использован метод 5-ти кратной кросс-валидации, также известный как метод перекрестной проверки на сформированной ранее выборке (4404 положительных и 4404 отрицательных новостей):

- для каждого алгоритма выборка случайным образом разбивается на обучающую и тестовую подвыборки 5 раз;
- для каждого разбиения классификатор обучается на обучающей подвыборке, затем происходит валидация на тестовой подвыборке;
- результатом метода кросс-валидации являются усредненные показатели качества классификации на тестовых подвыборках.

Для проведения эксперимента были выбраны наиболее популярные алгоритмы классификации, используемые для анализа тональности: мультиномиальный наивный байесовский классификатор (MNB); SVM со стохастическим градиентным спуском; случайный лес (RF).

Согласно результатам, в среднем наиболее высокие показатели в критериях точности, полноты и F-меры продемонстрировал SVM со стохастическим градиентным спуском (Таблица 5).

Таким образом, метод опорных векторов со стохастическим градиентным спуском с набором параметров эксперимента – «униграммы, биграммы, предобработка» показал наилучшие показатели точности, полноты и F-меры, который и был реализован программно (Рисунок 8).

Таблица 5 – Результаты экспериментов
Table 5 – Experimental results

Алгоритм	Параметры эксперимента			5-fold crossvalidation		
	Униграм-мы	Биграм-мы	Предобра-ботка	Точность, %	Полнота, %	F-мера, %
MNB	+	+		0,89	0,86	0,87
SVM	+	+	+	**0,93**	**0,96**	**0,94**
RF	+	+	+	0,86	0,93	0,9

Заключение

Свойства характерные для формального описания термина «мнение» позволяют сформулировать различные постановки задач, наиболее сложной из которых является выявление всех свойств объекта и характеристик мнения. Хотя для многих задач этого не требуется. Описанные проблемы решения задачи анализа тональности текста, как других задач компьютерной лингвистики, связаны как с характеристиками самого текста, так и качеством инструментария для его обработки.

Результаты анализа готовых программных реализаций подходов к решению задачи анализа тональности текста показывают отсутствие таких решений для текстов на башкирском языке.

Предложенная постановка задачи анализа тональности в текстах на башкирском языке и ее решение позволяют разделить новостные тексты на положительные и отрицательные. Имеющиеся в открытом доступе корпусы текстов на башкирском языке оказались недостаточными для решения данной задачи. Некоторые методы предобработки текста, например, стемминг, оказался непригодным ввиду особенностей башкирского языка.

Разработанное программное решение на основе SVM со стохастическим градиентным спуском, продемонстрировавшим наиболее высокие показатели в критериях точности, полноты и *F*-меры, может быть использовано для оценки тональности текстов новостных сайтов на башкирском языке.

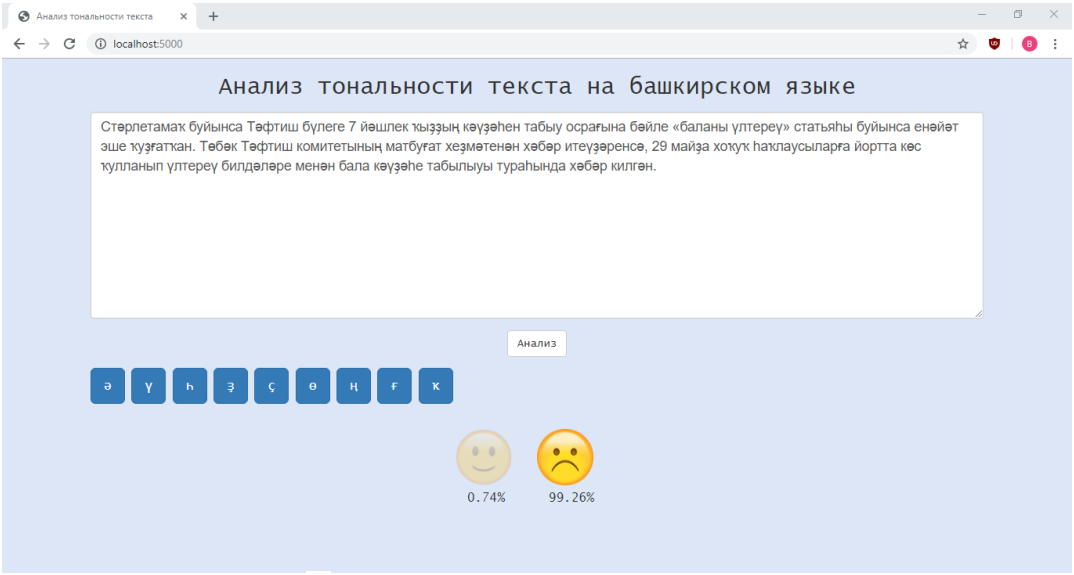

Рисунок 8 – Программная реализация классификатора
Figure 8 – Software implementation of the classifier

БЛАГОДАРНОСТИ

Результаты исследования, представленные в статье, получены при поддержке грантами РФФИ 19-07-00709, 20-08-00668 и Министерства науки и высшего образования РФ в рамках выполнения работ по Государственному заданию ФГБОУ ВО УГАТУ # FEUE-2020-0007.

ЛИТЕРАТУРА

1. Ананьева М. И., Кобозева М. В., Соловьев Ф. Н., Поляков И. В., Чеповский А. М. О проблеме выявления экстремистской направленности в текстах. *Вестник Новосибирского государственного университета. Серия: Информационные технологии.* 2016;14(4):5–13.
2. Башкирский поэтический корпус. Доступно по адресу: http://web-corpora.net/bashcorpus/search/?interface_language=ru (дата обращения 30.04.2020).
3. Бодрунова С.С. Кросс-культурный тональный анализ пользовательских текстов в Твиттере. *Вестник Московского университета Серия 10. Журналистика.* 2018;6:191-212.
4. Воронина И. Е., Гончаров В. А. Анализ эмоциональной окраски сообщений в социальных сетях (на примере сети «вконтакте»). *Вестник ВГУ. Серия :Системный анализ и информационные технологии.* 2015;4:151-158.
5. Горбушин Д. А., Гринченков Д. В., Мохов В. А., Нгуен Фук Хау Системный анализ подходов к решению задачи идентификации тональности текста. *Известия вузов. Северо-кавказский регион. Технические науки*. 2016;2:36-41.
6. Гаршина В. В., Калабухов К. С., Степанцов В. А., Смотров С. В. Разработка системы анализа тональности текстовой информации. *Вестник ВГУ, Серия: Системный анализ и информационные технологии.* 2017;3:185-194.
7. Ермаков А. Е., Киселев С. Л. Лингвистическая модель для компьютерного анализа тональности публикаций СМИ. *Компьютерная лингвистика и интеллектуальные технологии*:труды Международной конференции Диалог'2005. Москва:Наука, 2005. Доступно по адресу: http://www.dialog-21.ru/media/2365/ermakov-kiselev.pdf (дата обращения 30.04.2020).
8. Клековкина М. В., Котельников Е. В. Метод автоматической классификации текстов по тональности, основанный на словаре эмоциональной лексики. *Электронные библиотеки:перспективные методы и технологии, электронные коллекции (RCDL-2012*):труды 14-й Всероссийской научной конференции (Переславль-Залесский, Россия, 15-18 октября 2012 г.). 2012:81-86. Доступно по адресу: http://ceur-ws.org/Vol-934/paper15.pdf (дата обращения 30.04.2020).
9. Колмогорова А. В., Калинин А. А., Маликова А. В. Лингвистические принципы и методы компьютерной лингвистики для решения задач сентимент-анализа русскоязычных текстов. *Актуальные проблемы филологии и педагогической лингвистики.* 2018;1(29):139-148.
10. Котельников, Е.В. Комбинированный метод автоматического определения тональности текста. *Программные продукты и системы.* 2012;3:189-195.
11. Краснов Ф. В. Анализ тональности текста научно-практических статей по нефтегазовой тематике с помощью искусственных нейронных сетей. *Вестник Евразийской науки*. 2018;3(10). Доступно по адресу: https://esj.today/PDF/43ITVN318.pdf (дата обращения 30.04.2020).

12. Лукашевич Н. В. Автоматический анализ тональности текстов по отношению к заданному объекту и его характеристикам. *Russian Digital Libraries Journal*. 2015;18b(3-4):88-119.
13. Лукашевич Н. В., Четверкин И. И. Комбинирование тезаурусных и корпусных знаний для извлечения оценочных слов. *Системы и средства информатики.* 2015;25(1):20–33.
14. Меньшиков И. Л., Кудрявцев А. Г. Обзор систем анализа тональности текста на русском языке. *Молодой ученый*. 2012;12(47):140-143. Доступно по адресу: https://moluch.ru/archive/47/5951/ (дата обращения 30.04.2020).
15. Минина М. А. Психолингвистический анализ семантики оценки (на материале глаголов движения):*автореферат дис. ... кандидата филологических наук:*10.02.19. Москва, 1995:22.
16. Официальный сайт компании Ай-Теко. Доступно по адресу: https://www.i-teco.ru/solutions/business_intelligence_products/analiz_tonalnosti_teksta/(дата обращения 30.04.2020).
17. Пазельская А. Г., Соловьев А. Н. Метод анализа эмоций в текстах на русском языке. *Компьютерная лингвистика и интеллектуальные технологии:*материалы ежегодной Международной конференции «Диалог» (Бекасово, 25–29 мая 2011 г.). М.: Изд-во РГГУ. 2011;10 (17):510-552.
18. Посевкин Р.В. Автоматизация сентимент-анализа текста. *Междисциплинарный диалог: современные тенденции в гуманитарных, естественных и технических науках*:сборник трудов IV Всероссийской научно-практической конференции преподавателей, ученых, специалистов и аспирантов. Издательство: Общество с ограниченной ответственностью "Полиграф-мастер" (Челябинск). 2015:242-244.
19. Романов А. С., Васильева М. И., Куртукова А.В., Мещеряков Р. В. Анализ тональности текста с использованием методов машинного обучения. Доступно по адресу:http://ceur-ws.org/Vol-2233/Paper_8.pdf (дата обращения 30.04.2020).
20. Сарбасова А.Н. Исследование методов сентимент-анализа русскоязычных текстов// *Молодой ученый.* 2015;8(88):143-146. Доступно по адресу: https://moluch.ru/archive/88/17413/. (дата обращения 30.04.2020).
21. Сиразитдинов З. А., Полянин А.И., Ибрагимова А. Д., Ишмухаметова А.Ш. Корпусы башкирского языка: принципы разработки. *Проблемы востоковедения.* 2013;4 (62):65-72.
22. Толкунов А. А. Модель оперативной аналитической обработки текстовых комментариев к законопроектам: *автореферат дис. ... кандидата технических наук*: 05.13.17. Орел:Академия ФСО, 2014:24.
23. Тутубалина Е.В., Иванов В. В., Загулова М., Мингазов Н., Алимова И., Малых В. Тестирование методов анализа тональности текста, основанных на словарях. *Электронные библиотеки*. 2015;18(3-4):138-162.
24. Усталов Д. В. Извлечение терминов из русскоязычных текстов при помощи графовых моделей. Доступно по адресу:http://koost.eveel.ru/ science/ CSEDays2012.pdf. (дата обращения 30.04.2020).
25. Устный корпус башкирского языка. Доступно по адресу: https://linghub.ru/oral_bashkir_corpus/ (дата обращения 30.04.2020).
26. Чиркин Е. С., Лопатин Д. В. Подходы к нечеткому поиску нежелательного контента на веб-странице. *Вестник Тамбовского университета. Серия Естественные и технические науки.* Тамбов. 2016;21(6):2358-2365.
27. Abbasi M. M., Beltiukov A. P. Анализ эмоций из текста на русском языке с использованием синтаксических методов. *Information Technology and Systems*:7th

International Science Conference. At Khanty-Mansiysk. Russian Federation. 2019. Доступно по адресу: https://www.researchgate.net/publication/333489703Analiz_emocijiz_teksta_na_russkom_azyke_s_ispolzovaniem_sintaksiceskih_metodov (дата обращения 30.04.2020).
28. Yan G. et al. A bilingual approach for conducting Chinese and English social media sentiment analysis. *Computer Networks*. 2014;75(PB):491-503.
29. Kadam S.A., Joglekar S.T. Sentiment Analysis:An Overview. International *Journal of Research in Engineering & Advanced Technology*. 2013;1(4).
30. Kennedy A., Inkpen D. Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence*. 2006;22:110-125.
31. Liu B. Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies #16*. 2012;XIV:165.
32. Moilanen K., Pulman S., Zhang Y. Packed Feelings and Ordered Sentiments:Sentiment Parsing with Quasi-compositional Polarity Sequencing and Compression. *Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2010)*:proceedings of the 1st Workshop at the 19th European Conference on Artificial Intelligence (ECAI 2010).2010:36-43.
33. Opinion lexicon English Доступно по адресу: https://github.com/jeffreybreen/twitter-sentiment-analysis-tutorial-201107/tree/master/data/opinion-lexicon-English (дата обращения 30.04.2020).
34. Potapova R., Komalova L. Multimodal perception of aggressive behavior. *Lecture Notes in Computer Science*. 2016;9811:499-506.
35. Wiebe J.M., Wilson, T., Cardie, C. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*. 2005;39 (2-3):165-210.

REFERENCES

1. Ananeva M. I., Kobozeva M. V., Solovev F. N., Poliakov I. V., Chepovskii A. M.On the problem of revealing extremist ideology in texts. *Bulletin of Novosibirsk State University. Series:Information Technologies*. 2016;14 (4):5–13. (in Russian)
2. Bashkir poetry corpus. Available from: http://web-corpora.net/bashcorpus/search/?interface_language=ru (Accessed:30th April 2020). (in Russian)
3. Bodrunova S. S. Cross-Cultural Sentiment Analysis of Users' Texts in. *Bulletin of Moscow State university. Series 10. Journalism.* 2018;6:191-212. (in Russian)
4. Voronina I. E., Goncharov V. A. Analysis of Emotional Sentiments in Social Network Messages (using Vkontakte Network as an example). *Bulletin of Voronezh State University. Series:System Analysis and information Technologies*. 2015;4:151-158. (in Russian)
5. Gorbushin, D. A., Grinchenkov D.V., Mokhov V.A., Nguen Fuk Khau System Analysis of Approaches on Solving the Task of Text Sentiment Identification. *The tidinngs of HEI. North Caucasus Region. Technical Sciences*. 2016;2:36-41. (in Russian)
6. Garshina V.V., Kalabukhov K. S., Stepantsov V. A., Smotrov S.V. Development of a System for Text Information Sentiment Analysis. *Bulletin of Voronezh State university. Series:System Analysis and information Technologies*. 2017;3:185-194. (in Russian)
7. Ermakov A. E., Kiselev S.L. Linguistic model for Computational Sentiment Analysis of Mass media Publications. *Computational Linguistics and Intelligent technologies*:proceedings of the International conference Dialog'2005. Moscow:Nauka.

2005:616. Available from: http://www.dialog-21.ru/media/2365/ermakov-kiselev.pdf (Accessed:30th April 2020). (in Russian)
8. Klekovkina M.V., Kotelnikov E.V. A Method for Automated Sentiment Classification of Texts Based on the Dictionary of Emotional Lexicon. *Digital Libraries:Advanced Methods and Technologies, Digital Collections:*proceedings of the 14th All-Russian Scientific Conference (RCDL-2012). (Pereslavl-Zalesskii, Russia, 15-18 October 2012). 2012:81-86. Available from: http://ceur-ws.org/Vol-934/paper15.pdf (Accessed:30th April 2020). (in Russian)
9. Kolmogorova A.V., Kalinin A. A., Malikova A.V. Linguistic Principles and Methods of Computational Linguistics for Solving the Tasks of Sentiment Analysis of Russian Texts. *Open Issues of Philology and Pedagogic Linguistics*. 2018;1(29):139-148.(in Russian)
10. Kotelnikov E.V. Combined Method for Automatic Sentiment Identification of a Text. *Software Products and System*. 2012;3:189-195. (in Russian)
11. Krasnov F.V. Sentiment Analysis of Applied Scientific Articles on Oil and Gas Industry with use of Artificial Neural. *Bullentin of Eurasian Science*. 2018;3(10). Available from: https://esj.today/PDF/43ITVN318.pdf (Accessed:30th April 2020). (in Russian)
12. Lukashevich N.V. Automatic Sentiment Analysis of the Text with Respect to the Predefined Object and its Characteristics. *Russian Digital Libraries Journal*. 2015;18(3-4):88-119. (in Russian)
13. Lukashevich N.V., Chetverkin I. I. Combining of Thesaurus and Corporal Knowledge for Extracting the Words of Characteristics. *Systems and Means of Informatics* .2015;25(1): 20–33. (in Russian)
14. Menshikov I. L., Kudriavtsev A. G. A Survey on Sentiment Analysis for Texts in Russian. *Young Scientist*. 2012;12(47):140-143. Available from: https://moluch.ru/archive/47/5951/ (Accessed:30th April 2020). (in Russian)
15. Minina M. A. Psycholinguistic Analysis of Evaluative Sematics (on the Material of the verbs of Movement):10.02.19. *Thesis for the degree of candidate of philological sciences*, Moscow. 2005. (in Russian)
16. i-Teco official Website. Available from: https://www.i-teco.ru/solutions/business_intelligence_products/analiz_tonalnosti_teksta/ (Accessed:30th April 2020). (in Russian)
17. Pazelskaia A. G., Solovev A. N. Method for emotion Analisys in Russian. *Computational Linguistics and Intelligent Technologies*:proceedings of the International conference Dialog (Bekasovo, 25–29 May 2011). Publishing House of the Russian State University for the Humanities. 2011;10(17):510-552.(in Russian)
18. Posevkin R.V. Automation of the Sentiment analysis of the Text. *Inter disciplinary Dialog:Novel Trends in Humanities, Natural, and Technical Sciences. Proceedings of the IV All-Russian Conference on Applied Sciences for Lecturers, Scientists, Experts, and doctoral Students*. 2015:242-244.(in Russian)
19. Romanov A. S., Vasileva M. I., Kurtukova A.V., Meshcheriakov R.V. Sentiment Analysis of the text with use of Machine learning Methods. Available from: http://ceur-ws.org/Vol-2233/Paper_8.pdf (Accessed:30th April 2020). (in Russian)
20. Sarbasova A. N. Exploration of the Sentiment Analysis Methods for the texts in Russian. *Young Scientist*. 2015;8 (88):143-146. Available from: https://moluch.ru/archive/88/17413/ (Accessed:30th April 2020). (in Russian)
21. Sirazitdinov Z. A., Polianin A. I., Ibragimova A. D., Ishmukhametova A. Sh. Corpora of Bashkir Language:Development Principles. *Problems of Orientalism.* 2013;4 (62):65-72. (in Russian)

22. Tolkunov A. A. Model for Realtime Analytical Processing of Textual Comments on Legislative Proposals. *Thesis for the degree of candidate of technical sciences*: 05.13.17. Academy of Federal Protective Service, Orel. 2014: 24. (in Russian)
23. Tutubalina E.V., Ivanov V.V., Zagulova M., Mingazov N., Alimova I., Malykh V. Testing Dictionary-Based Methods of Sentiment Analysis. *Digital Libraries.* 2015;18(3-4):138-162. (in Russian)
24. Ustalov D. V. Extracting terms from Russian texts with use of Graph-Based Methods. Available from: http://koost.eveel.ru/_science/_CSEDays2012.pdf (Accessed:30th April 2020). (in Russian)
25. Oral Corpus of Bashkir Language. Available from: https://linghub.ru/oral_bashkir_corpus/ (Accessed:30th April 2020). (in Russian)
26. Chirkin E.S., Lopatin D.V. Approaches on Fuzzy Search for Unintended Content on a Web-Page. *Bulletin of Tambov University. Series:Natural and Technical Sciences.* Tambov. 2016;21(6): 2358-2365. (in Russian)
27. Abbasi M. M., Beltiukov A. P. Emotion Analysis in Russian Text with use of Syntactic Methods. Information Technology and Systems:7th International Science Conference. At Khanty-Mansiysk. Russian Federation. 2019. Available from: https://www.researchgate.net/publication/333489703Analiz_emocijiz_teksta_na_russkom_azyke_s_ispolzovaniem_sintaksiceskih_metodov (Accessed:30th April 2020).
28. Yan G. et al. A bilingual approach for conducting Chinese and English social media sentiment analysis. *Computer Networks*. 2014;75(PB):491-503.
29. Kadam S.A., Joglekar S.T. Sentiment Analysis:An Overview. International Journal of Research in Engineering & Advanced Technology. 2013;1(4).
30. Kennedy A., Inkpen D. Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence*. 2006. 22:110-125.
31. Liu B. Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies #16*. 2012;XIV:165.
32. Moilanen K., Pulman S., Zhang Y. Packed Feelings and Ordered Sentiments:Sentiment Parsing with Quasi-compositional Polarity Sequencing and Compression. *Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2010)*:proceedings of the 1st Workshop at the 19th European Conference on Artificial Intelligence (ECAI 2010).2010:36-43.
33. Opinion lexicon English Доступно по адресу: https://github.com/jeffreybreen/twitter-sentiment-analysis-tutorial-201107/tree/master/data/opinion-lexicon-English (дата обращения 30.04.2020).
34. Potapova R., Komalova L. Multimodal perception of aggressive behavior. *Lecture Notes in Computer Science.* 2016;9811:499-506.
35. Wiebe J.M., Wilson, T., Cardie, C. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*. 2005;39 (2-3):165-210.

ИНФОРМАЦИЯ ОБ АВТОРАХ / INFORMATION ABOUT THE AUTHORS

Сулейманов Азамат Каримович, магистрант кафедры вычислительной математики и кибернетики, Уфимский государственный авиационный технический университет, Факультет информатики и робототехники, Уфа, Российская Федерация
e-mail: azamat-sul2010@yandex.ru

Azamat K. Suleimanov ., Master Student At The Department Of Computational Mathematics And Cybernetics, Ufa State Aviation University, Faculty Of Computer Science And Robotics, Ufa, Russian Federation.

Шарипова Миляуша Амировна, аспирант кафедры вычислительной математики и кибернетики, Уфимский государственный авиационный технический университет, Факультет информатики и робототехники, Уфа, Российская Федерация
e-mail: mamirovna@yandex.ru

Milyausha. A. Sharipova, Phd Student At The Department Of Computational Mathematics And Cybernetics, Ufa State Aviation University, Faculty Of Computer Science And Robotics, Ufa, Russian Federation

Сметанина Ольга Николаевна, доктор технических наук, доцент,профессор кафедры вычислительной математики и кибернетики, Уфимский государственный авиационный технический университет, Факультет информатики и робототехники, Уфа, Российская Федерация.
e-mail: smoljushka@mail.ru

Olga N. Smetanina, Professor at the Department of Computational Mathematics and Cybernetics, Ufa State Aviation Technical University, Faculty of Computer Science and Robotics, Ufa, Russian Federation.

Сазонова Екатерина Юрьевна, кандидат технических наук, доцент кафедры вычислительной математики и кибернетики, Уфимский государственный авиационный технический университет, Факультет информатики и робототехники, Уфа, Российская Федерация.
e-mail: rassadnikova_ekaterina@mail.ru

Ekaterina. Y. Sazonova, Assistant professor of the Department of Computational Mathematics and Cybernetics, Ufa State Aviation Technical University, Faculty of Computer Science and Robotics, Ufa, Russian Federation

Миронов Константин Валерьевич, старший преподаватель кафедры вычислительной техники и защиты информации, Уфимский государственный авиационный технический университет, Факультет информатики и робототехники, Уфа, Российская Федерация.
e-mail: mironovconst@gmail.com

Konstantin V. Mironov, PhD, Senior Lecturer at the Department of Computer Technology and Information Security, Ufa State Aviation Technical University, Faculty of Computer Science and Robotics, Ufa, Russian Federation