УДК 004.002

ПОДХОД К АВТОМАТИЧЕСКОЙ КЛАССИФИКАЦИИ КОРОТКИХ ТЕКСТОВЫХ СООБЩЕНИЙ НА ОСНОВЕ МОДИФИЦИРОВАННОГО МЕТОДА БАЙЕСА

А. А. ОВСЯННИКОВ
И.Н. ГРЫЗЛОВ
Е.Ю. ГОЛУБИНСКИЙ
А.А. СМИРНОВ
С.А. ВЛАСОВА

Академия ФСО России, г. Орел

e-mail:
ovsyannikov.aa@mail.ru
igryzlov@gmail.com
darzhek@yandex.ru
al2smi@gmail.com
veton646464@rambler.ru

Востребованность сообщений новостных агентств сети Интернет в деятельности информационных служб, а также их значительный ежедневный поток требует создания средств автоматической обработки сообщений, позволяющих обеспечить их систематизацию и соответственно формирование тематических груш. В статье предлагается подход к автоматической классификации коротких текстовых сообщений, позволяющий обеспечить достаточное качество классификации и высокую скорость при использовании сложных рубрикаторов. В работе также описана разработанная программная система и приведены результаты её тестирования на контрольной выборке коротких текстовых сообщений сети Интернет.

Ключевые слова: короткое текстовое сообщение, автоматическая классификация, рубрикатор, рубрика, методы автоматической классификации текстовых сообщений.

В настоящее время в сети Интернет большое количество событий, явлений, процессов описывается в относительно коротких текстовых сообщениях (до 3000 символов). Их типичными источниками являются сетевые новостные агентства. Накопленные информационными службами массивы таких сообщений активно используются при наблюдении за обстановкой, состояниями объектов, общественным мнением. Современное развитие сетевых СМИ, а также средств автоматического сбора текстовой информации сети Интернет позволяет информационным службам получать тысячи сообщений в день. Так, только одно сетевое новостное агентство федерального уровня, как правило, выпускает несколько десятков сообщений в день, а роботизированные системы сбора сообщений позволяют обрабатывать сотни подобных источников. Информационные потоки большой интенсивности и размерности делают невозможным ознакомление аналитика с каждым сообщением и пониманием его смысла. Одним из решений данной актуальной проблемы является использование средств автоматической классификации текстовых сообщений, получаемых роботизированными системами сбора информации из сети Интернет. Необходимым условием эффективной работы подобных систем является применение алгоритма, обеспечивающего необходимое качество классификации.

Отдельные аспекты проблем автоматизированной и автоматической классификации текстов, поступающих из различных источников информации рассмотрены в работах [4, 5]. Мбайкоджи Э., Драль А. А., Соченков И. В. предложили метод автоматической классификации коротких текстовых сообщений на основе характеристики тематической значимости текста и ее модификации [4]. Данные авторы получили высокие результаты классификации коротких рекламных сообщений (средняя точность классификации превышала 88%), причем обрабатывались только заголовки сообщений. Однако данные результаты были получены при классификации только по пяти рубрикам, что при обработке информационной службой новостных сообщений сети Интернет является явно недостаточным. Так, для систематизированного сбора информации только о произошедших чрезвычайных ситуациях требуется не менее 20 рубрик, соответствующих различным видам аварий, природных и техногенных катастроф. При этом часто необходимо использование иерархического рубрикатора, с несколькими уровнями вложенности.

Вопросам классификации текстовых документов посвящена работа В.И. Шабанова, А.М. Андреева [5]. Однако в этой работе основное внимание авторов уделено проблемам

классификации объемных документов, а также сайтов (по своей сути – наборов текстовых документов).

Анализ перечисленных выше работ показал, что их авторы получали высокие результаты классификации определенного, достаточно узкого класса информационных объектов (рекламных объявлений, текстовых документов значительного объема). Но применение указанных подходов для информационных служб при обработке ресурсов Интернет нецелесообразно из-за несоответствия обрабатываемых информационных объектов и условий проведения экспериментов реалиям информационно-аналитической работы.

Работы [6-8] посвящены сравнению методов автоматической классификации текстов. Анализ данных работ, а также моделирование рассмотренных в них методов показали, что оптимального метода классификации пока не найдено, так как в каждом есть свои достоинства и недостатки. Например, существенными достоинствами метода Байеса (в условиях обработки больших объемов текстовых сообщений) является простота реализации, скорость классификации, быстрота обучения, стабильность на различных данных. Однако точность классификации при использовании данного метода, как правило, уступает точности классификации при использовании более сложных и ресурсоемких методов.

В данной статье рассматривается подход к созданию автоматической системы классификации текстовых сообщений на основе модифицированного метода Байеса, позволяющего в целом сохранить достоинства базового метода Байеса, при условии повышения качества классификации. Оригинальный метод Байеса предполагает содержать в рубрике все слова текстовых сообщений, использованных для обучения, а затем все слова текста сравнивать со всеми словами рубрики. Авторами используется похожий подход, но слова рубрики предварительно подвергаются процедуре дополнительной обработки, направленной на выделение значимых слов в рубрике с использованием процедуры «TF-IDF» («частота термина – обратная документная частота»).

В качестве информационной базы для создания данной системы авторами использовались работы [1-3, 6-8], а также структурированные массивы коротких текстовых сообщений, использованных информационными службами и созданные авторами самостоятельно.

Процедуры первичной обработки данных необходимы для качественного осуществления всего процесса классификации текстовых сообщений в дальнейшем. Первичная обработка текстовых сообщений состоит в следующем. На первоначальном этапе осуществляется импорт текстового сообщения. Далее к текстовому сообщению применяется морфологический анализ, предназначенный для приведения словоформ, встречающихся в тексте, к начальной форме и получения морфологической информации о них. В разработанной системе используется точный подход к морфологическому анализу, основанный на использовании словарей, в которых для каждого слова указано правило изменения его формы. В результате применения морфологического анализа текстовых сообщений формируется перечень слов в тексте в начальной форме. Далее к полученному перечню применяется процедура удаления так называемых «стоп-слов», которые представляют собой малоинформативные слова и служебные части речи, не характеризующие текстовые сообщения по смыслу, например, предлоги, союзы и т. п. Для данной процедуры применяется словарь стоп-слов.

Процедуры морфологического анализа и удаления «стоп-слов» являются широко известными [1, 2] и активно применяются при автоматической обработке текстов различного содержания. Алгоритм первичной обработки сообщений в общем виде представлен на рисунке 1.

Из полученного перечня слов (в начальной форме) формируется вектор в пространстве терминов текстового сообщения. Под терминами текстового сообщения понимаются все одиночные слова текстового сообщения, за исключением стоп-слов. Кроме того, каждой форме слова, обнаруженной в сообщении, будет соответствовать один и тот же термин – данное слово в начальной форме.

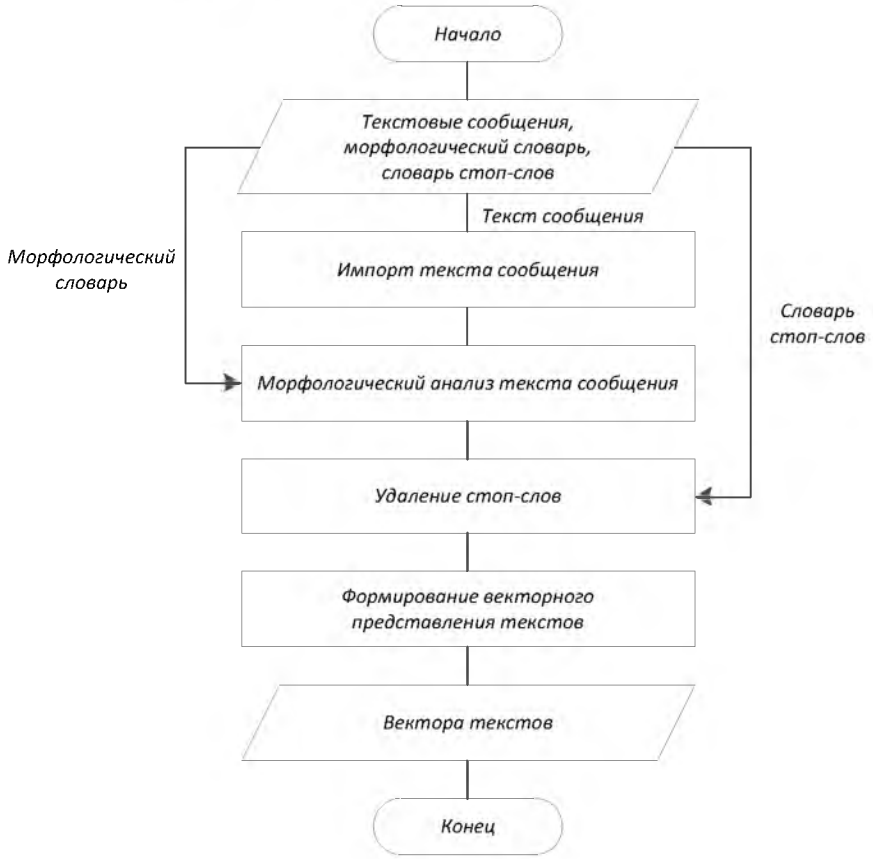

Рис. 1. Алгоритм первичной обработки сообщений

Процедура обучения рубрикатора в качестве входной информации использует вектора текстовых сообщений для обучения. В данном случае осуществляется обработка векторов текстовых сообщений обучающей выборки с использованием процедуры «TF-IDF». Таким способом вычисляются веса терминов, имеющихся во всех векторах текстовых сообщений обучающей выборки.

При этом веса терминов обладают следующими свойствами:

1) имеют высокие значения, если термин часто встречается в небольшом числе текстовых сообщений, тем самым усиливая отличие этих сообщений от других;

2) имеют низкие значения, если термин редко встречается в каком-то текстовом сообщении или встречается во многих сообщениях, тем самым снижая различие между ними.

Результатом применения процедуры «TF-IDF» к векторам текстовых сообщений обучающей выборки (для каждой рубрики рубрикатора) является результирующий вектор рубрики.

Алгоритм классификации позволяет проводить поиск наиболее вероятной рубрики для классифицируемого сообщения. Классификация текстового сообщения проводится в два этапа. На первом этапе осуществляется его первичная обработка. На выходе первого этапа сообщение представляется в виде вектора слов в начальной форме. На втором этапе осуществляется сравнение признаков данного вектора с векторами рубрик обученного рубрикатора. Результаты сравнения сохраняются. Таким образом, формируется массив данных, отражающих соответствие классифицируемого сообщения каждой из рубрик. Затем проводится его ранжирование и перевод значений степени соответствия в шкалу [0,100]. На выходе второго этапа формируется массив, содержащий соответствие классифицируемого сообщения каждой рубрике, выраженное в процентах.

В данной статье не приводится математическая основа используемых авторами методов классификации Байеса, процедуры «TF-IDF», ввиду их достаточно полного описания в работах [7, 8].

Рассмотренные процедуры обработки коротких текстовых сообщений реализованы в системе автоматической классификации коротких текстовых сообщений «TextClassifier-NTR». Данная система позволяет регулировать размер результирующего массива, включая в него информацию о рубриках с наибольшим соответствием.

Система реализована с использованием технологии «клиент-сервер». Серверная часть представлена web-сервисом, реализованным в виде SOAP-приложения. *Simple Object Access Protocol (SOAP)* представляет из себя основанный на *XML* протокол, предназначенный для обмена структурированной информацией между распределенными приложениями поверх существующих *WEB*-протоколов, включая *HTTP, SMTP* и т. д. Клиентская часть посылает запросы сервису через протокол *HTTP*, получает от него ответы и выводит результирующие данные на экран монитора.

Функции, реализованные в программной системе автоматической классификации коротких текстовых сообщений «TextClassifier-NTR»:

– соединение с БД системы – при использовании этой функции серверная часть соединяется с БД, параметры соединения прописаны в ini-файле, и возвращает результат операции;

– получение перечня рубрикаторов системы – при использовании этой функции серверная часть возвращает сообщение в формате XML, содержащее перечень рубрикаторов системы, функция дает возможность использовать индивидуальные рубрикаторы для различных систем;

– получение иерархической структуры рубрик выбранного рубрикатора – при использовании этой функции серверная часть возвращает сообщение в формате XML, содержащее иерархическую структуру рубрик выбранного рубрикатора, реализована возможность использования древовидного рубрикатора с получением полной иерархической цепочки до корневого родительского элемента;

– обучение текущей рубрики выбранного рубрикатора – при использовании этой функции клиентская часть отправляет сервису сообщение в формате XML, содержащий код обучаемого рубрикатора, текст сообщения и код рубрики, которой он соответствует, а серверная часть возвращает результат обучающей операции;

– классификация текстовых сообщений с возможностью отнесения текстового сообщения к нескольким рубрикам с указанием их весов – при использовании этой функции серверная часть возвращает сообщение в формате XML, содержащий код рубрики и весовой коэффициент, при этом возможно получить несколько близких вариантов привязки по рубрикатору.

Интерфейс клиентской части данной системы в режиме классификации сообщений показан на рисунке 2.

Для анализа качества классификации были построены три рубрикатора и сформированы обучающая и контрольная выборки коротких текстовых сообщений для их обучения и оценки точности классификации. Сообщения для формирования обучающих и контрольных выборок отбирались специалистами информационной службы. Основные характеристики рубрикаторов и используемых для их обучения выборок приведены в таблице 1. Характеристики контрольных выборок коротких текстовых сообщений приведены в таблице 2.

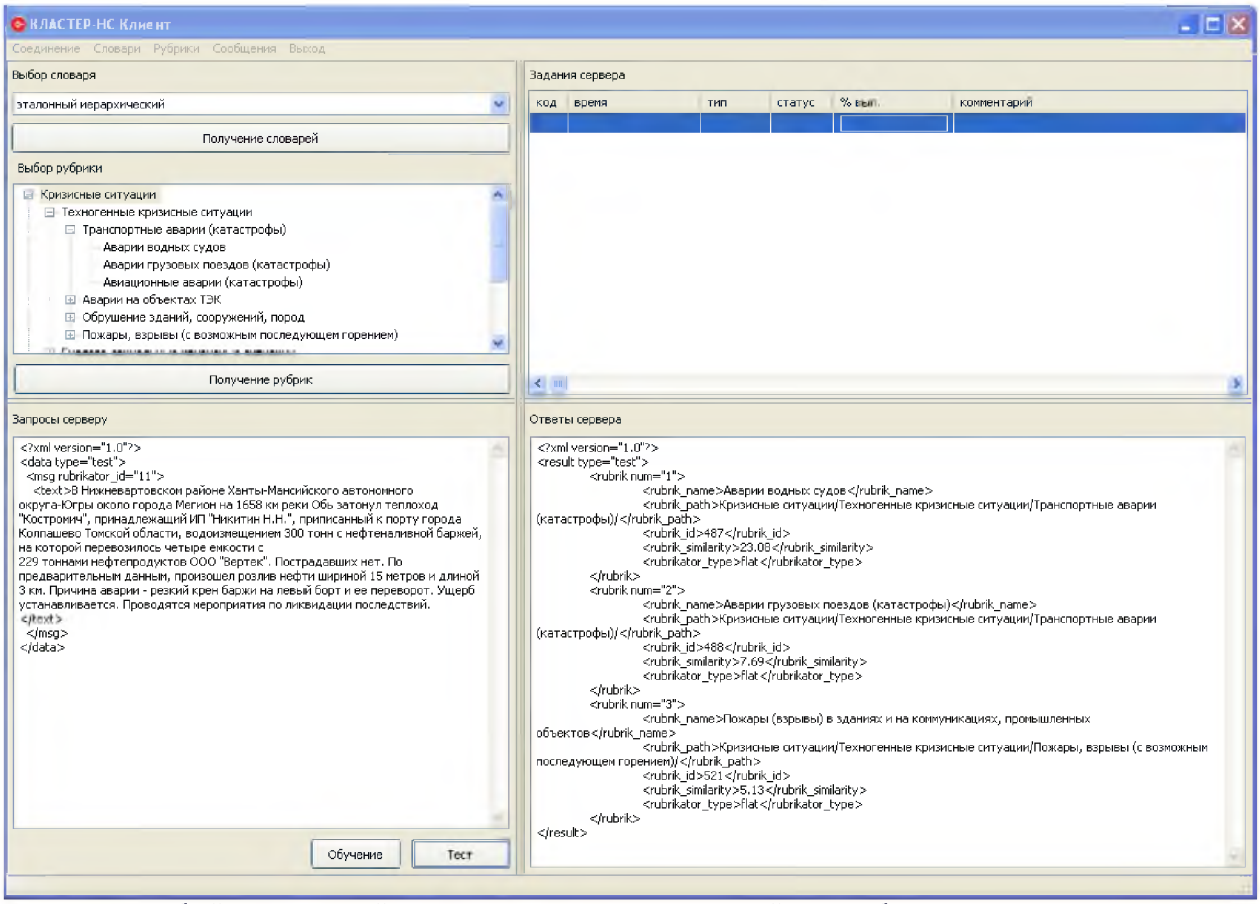

Рис. 2. Интерфейс клиентской части системы автоматической классификации коротких текстовых сообщений «TextClassifier-NTR» в режиме классификации

Таблица 1

Основные характеристики рубрикаторов и обучающих выборок

№ п/п	Тип рубрикатора	Общее количество рубрик (с учетом узлов)	Общее количество рубрик (без учета узлов)	Общее количество сообщений в обучающей выборке	Распределение сообщений для обучения по рубрикам (Количество рубрик / количество сообщений, используемое при обучении каждой из них)
1.	Плоский	31	31	465	31/15
2.	Иерархический	48	31	465	31/15
3.	Иерархический	48	31	2904	16/15; 1/45; 1/47; 1/70; 1/83; 1/86; 1/97; 1/110; 1/131; 1/194; 1/195; 1/255; 1/314; 1/325; 1/333; 1/379

Таблица 2

Основные характеристики контрольной выборки

№ п/п	Тип рубрикатора	Распределение сообщений для определения точности классифицирования по рубрикам (Количество рубрик / количество сообщений, используемое при определении точности классифицирования (для каждой из них))
1.	Плоский	2/3; 1/4; 1/5; 2/8; 1/9; 1/11; 2/12; 1/14; 20/15
2.	Иерархический	2/3; 1/4; 1/5; 2/8; 1/9; 1/11; 2/12; 1/14; 20/15
3.	Иерархический	2/3; 1/4; 1/5; 2/8; 1/9; 1/11; 2/12; 1/14; 20/15

В таблице 3 приведены результаты классификации сообщений контрольной выборки для построенных рубрикаторов.

Таблица 3

Результаты классификации сообщений контрольной выборки (по рубрикаторам)

№ рубрики	Название рубрики	Кол-во сообщений в тестовой выборке	Тип рубрикатора/общее количество рубрик/общее количество рубрик (без учета узлов)/общее количество сообщений в обучающей выборке								
			Плоский/ 31/31/465			Иерархический/ 48/31/465			Иерархический/ 48/31/2904		
			Точность классификации при условии определения 1 наиболее близкой рубрики	Точность классификации при условии определения 2 наиболее близких рубрик	Точность классификации при условии определения 3 наиболее близких рубрик	Точность классификации при условии определения 1 наиболее близкой рубрики	Точность классификации при условии определения 2 наиболее близких рубрик	Точность классификации при условии определения 3 наиболее близких рубрик	Точность классификации при условии определения 1 наиболее близкой рубрики	Точность классификации при условии определения 2 наиболее близких рубрик	Точность классификации при условии определения 3 наиболее близких рубрик
1	2	3	4	5	6	7	8	9	10	11	12
1	Аварии водных судов	14	78,57	92,86	92,86	73,33	86,67	86,67	78,57	100,00	100,00
2	Аварии грузовых поездов (катастрофы)	8	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
3	Аварии на объектах газотранспортной системы	3	66,00	66,00	66,00	66,00	66,00	66,00	66,00	66,00	66,00
4	Авиационные аварии (катастрофы)	15	93,33	100,00	100,00	93,33	100,00	100,00	100,00	100,00	100,00
5	Акции экстремистских организаций	15	86,67	93,33	93,33	86,67	86,67	86,67	66,67	86,67	86,67
...	...	...	...	...	...	...	...	...	...	...	...
27	Разбой	15	86,67	93,33	100,00	73,33	100,00	100,00	93,33	100,00	100,00
28	Смерч, ураган, буря, шторм, тайфун, шквал	4	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
29	Убийство и покушение на убийство	15	80,00	86,67	93,33	86,67	93,33	93,33	86,67	86,67	86,67
30	Хищение оружия, боеприпасов	15	100,00	100,00	100,00	100,00	100,00	100,00	93,33	93,33	93,33
31	Явка с повинной	9	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
	Точность		**83,75**	**93,53**	**94,82**	**83,50**	**91,99**	**94,22**	**85,74**	**91,66**	**93,94**
	F-мера		**91,15**	**96,66**	**97,34**	**91,01**	**95,83**	**97,02**	**92,32**	**95,65**	**96,88**

Приведенные в таблице 3 данные позволяют сделать вывод о достаточно высокой точности классификации при использовании как плоского, так и иерархических рубрикаторов, включающих в себя несколько десятков рубрик. Таким образом, применение процедуры «TF-IDF» при обучении рубрикаторов способствовало улучшению разделяющей способности оригинального метода классификации Байеса.

Апробация разработанной программной системы группой специалистов информационных служб по обработке первичной информации показала возможность его

эффективного практического применения при автоматизированном формировании структурированных массивов коротких текстовых сообщений, характеризующих общественно-политическую и социально-экономическую обстановку в регионах России.

В настоящее время ведутся работы по формированию эталонного структурированного массива коротких текстовых сообщений для исследования возможностей разработанных алгоритмов и системы в целом при условии использования многоуровневого иерархического рубрикатора, включающего в себя более 100 рубрик.

Список литературы

1. Агеев М.С. Методы автоматической рубрикации текстов, основанные на машинном обучении и знаниях экспертов: дис. ... канд. физ.-мат. наук. М., 2004. – 136 с.

2. Васильев В. Г., Кривенко М.П. Методы автоматизированной обработки текстов. М. : ИПИ РАН, 2008. – 304 с.

3. Зайцева Т. В., Васина Н.В., Пусная О.П., Смородина Н.Н. Программная реализация метода деревьев решений для реализации задач классификации и прогнозирования // Научные ведомости БелГУ Серия История Политология. Экономика. Информатика. 2013 № 8 (151). Выпуск 26/1. –С. 121-127.

4. Мбайкоджи Э., Драль А. А., Соченков И. В. Метод автоматической классификации коротких текстовых сообщений // Информационные технологии и вычислительные системы. 2012. № 3. – С. 93-102.

5. Шабанов В. И., Андреев А. М. Метод классификации текстовых документов, основанный на полнотекстовом поиске. Режим доступа: http://romip.ru/romip2003/4_shabanov.pdf

6. Joachim, T. A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization // Proceedings of ICML-97, 14th International Conference on Machine Learning, 1996. – P. 143–151.

7. Dumais S., Platt J., Heckerman D., Sahami M. Inductive learning algorithms and representations for text categorization // In Proc. Int. Conf. on Inform. and Knowledge Manage, 1998. – P. 148–155.

8. Yang Y., Liu X. A re-examination of text categorization methods // Proc. of Int. ACM Conference on Research and Development in Information Retrieval (SIGIR-99), 1999. – P. 42–49.

APPROACH TO AUTOMATIC CLASSIFICATION OF SHORT TEXT MESSAGES BASED ON MODIFIED BAYES'S METHOD

A. A. OVSYANNIKOV
I.N. GRYZLOV
E.Y. GOLUBINSKY
A.A. SMIRNOV
S.A. VLASOVA

The Academy of the Federal Guard Service of the Russian Federation, Orel, Russian Federation,

e-mail:
ovsyannikov.aa@mail.ru
igryzlov@gmail.com
darzhek@yandex.ru
al2smi@gmail.com
veton646464@rambler.ru

The relevance of the Internet news agencies' messages in the activities of information services as well as their significant daily flow require means of automatic processing of messages which can provide message systematization and , therefore, thematic groups formation. The article suggests an approach to automatic classification of short text messages which provides sufficient classification quality and high speed at complex rubricators usage. The article also describes the designed software system and presents the results of its testing on control sample of the Internet short text messages.

Keywords: short text message, automatic classification, rubricator, rubric, methods of automatic classification of text messages.