УДК 004.913
DOI: 10.46548/21vek-2020-0950-0026

**ОБНАРУЖЕНИЕ СПАМА В СМС-СООБЩЕНИЯХ С ИСПОЛЬЗОВАНИЕМ ТЕХНОЛОГИЙ WORD EMBEDDING И TERM FREQUENCY- INVERSE DOCUMENT FREQUENCY (TF-IDF)**



**Аббаси Мохсин Маншад**, аспирант кафедры теоретических основ информатики (ТОИ),
*Удмуртский государственный университет (УдГУ)*
*(426034, Россия , Ижевск, ул. Университетская, 1, e-mail: mohsinmanshad@gmail.com )*
*Университет AJ & K, Музаффарабад*
**Бельтюков Анатолий Петрович**, профессор, доктор физико-математических наук,
заведующий кафедрой теоретических основ информатики (ТОИ)
*Удмуртский государственный университет (УдГУ)*
*(426034, Россия , Ижевск, ул. Университетская, 1, e-mail: belt.udsu@mail.ru)*
**Лал Хусейн**, доцент кафедры компьютерных наук и информационных технологий
*Университет AJ & K, Музаффарабад*
*(13100, Пакистан, Азад Кашмир, e-mail: lall_hussain2008@live.com)*
**Аббаси Аниес Qамар**, аспирант
*Технологический институт Карлсруэ, Германия*
преподаватель кафедры компьютерных наук
*Женский Университет AJ & K;*
*(12500, Пакистан, Баг, Азад Кашмир,e-mail: itsabbassi@gmail.com.)*

**Аннотация**. Обнаружение спама - это идентификация нежелательной части информации из текстового корпуса. Она включает в себя классификацию нежелательного фрагмента текста, называемого спамом. Это важное направление исследований в области анализа текста. Наиболее распространенными видами спама являются спам в электронной почте и короткие текстовые сообщения (СМС). Они рассматриваются организациями как серьезное неудобство для клиентов, а также вредны для компьютерных систем. Традиционно целью спама была реклама продуктов и услуг потенциальному клиенту. Однако со временем люди стали использовать спам в качестве механизма взлома или атаки на системы с помощью вирусов. Ученые и исследователи предложили различные методологии для обнаружения спама и его фильтрации в электронных письмах. Но выявление спама в коротких текстовых сообщениях не привлекло большого внимания. В данной работе основное внимание уделяется разработке программы обнаружения спама в коротких текстовых сообщениях (СМС). Для этой цели используются две известные модели: Word embedding и Термин Частота - Обратная частота документа (TF-IDF). Эти Модели анализируют текст путем преобразования дискретного текстового сообщения в непрерывную числовую векторную форму. Вектор представляет каждое слово в тексте, а числовое значение размеров слова основано на контексте слова. Результаты исследования подробно изложены в разделах эксперимент и обсуждение.

**Ключевые слова**:спам, tf-idf, Word embedding, СМС, обнаружение, машинное обучение, особенности, размеры.

**SPAM DETECTION IN SHORT TEXT MESSAGES (SMS) USING WORD EMBEDDING AND TERM FREQUENCY- INVERSE DOCUMENT FREQUENCY (TF-IDF)**



**Abbashi Mohsin Manshad**, postgraduate student of the
Department of Theoretical Foundations of Informatics (TOI),
*Udmurt State University (Udmurt State University)*
*(426034, Russia, Izhevsk, Universitetskaya St., 1, e-mail: mohsinmanshad@gmail.com)*
*AJ & K University, Muzaffarabad*
*(13100, Pakistan, Azad Kashmir)*
**Beltyukov Anatoly Petrovich**, professor, doctor of Physical and Mathematical Sciences,
Head of the Department of Theoretical Foundations of Informatics (TOI)
*Udmurt State University (Udmurt State University)*
*(426034, Russia, Izhevsk, Universitetskaya St., 1, e-mail: belt.udsu@mail.ru)*
**Lal Hussein**, associate Professor, Department of Computer Science and Information Technology
*AJ & K University, Muzaffarabad*
*(13100, Pakistan, Azad Kashmir, e-mail: lall_hussain2008@live.com)*
**Abbasi Anies Qamar**, postgraduate student
*Institute of Technology Karlsruhe, Germany*
lecturer, Department of Computer Science
*Women's University AJ & K*

*(12500, Pakistan, Bagh, Azad Kashmir, e-mail: itsabbassi@gmail.com.)*

**Abstract**. Spam detection is the identification of unwanted piece of information from a text corpus. It includes clas-sification of non-desirable piece of text as spam. It is an important field of research in text analysis. The most common types of spam are email spam and short text messages (SMS) spam. They are considered as a serious inconvenience for the clients and are harmful for the computer systems. Traditionally the purpose of spam was doing advertisement of products and services to a potential customer. However, with time people are using spam as a mechanism of doing hacking or attacking the systems with viruses. The scientists and researchers have proposed different methodologies for spam detection and its filtration in emails. But the spam identification in short text messages did not got much attention. In this work, the focus is to develop a model for spam detection in short text messages (SMS). For this purpose, two famous models Word embedding and Term Frequency – Inverse Document Frequency (TF-IDF) models are used. These Models analyze text by converting the discrete text message to a continuous numerical vector form. A vector represent each word in the text, and the numerical value of the dimensions of a word is based on the context of the word. The results of the research are detailed in methodology and experiment sections of the paper.

**Keywords**: spam, tf-idf, Word embedding,SMS, detection, machine learning, features, dimensions

**Введение**. Технология и ее развитие сделали процесс общения эффективным и действенным. Наиболее распространенной формой обмена информацией являются электронные письма и короткие текстовые сообщения (СМС). Короткие текстовые сообщения могут быть разных форм. Они могут передаваться через локальную мобильную сеть, могут быть приложениями *Viber* или сообщениями *WhatsApp*. Компании используют эти короткие текстовые сообщения (СМС) как средство рекламы своих продуктов и услуг для клиентов. Люди часто находят раздражающим получение спама в огромном количестве и считают это некомпетентностью поставщиков услуг по обнаружению и фильтрации этих сообщений. Помимо рекламы, спам также содержит подозрительные ссылки и файлы. Загрузка и запуск подозрительного файла может привести к нарушению безопасности системы или сервера. Аналогичным образом, нажатие на ссылку может перевести систему на сайт с вирусами или вредоносными программами. Такие спам-сообщения используют большую часть пропускной способности сети. Существуют различные методы предотвращения спама и его маршрутизации через Интернет, но, поскольку Интернет общедоступен, его нельзя полностью контролировать.

В случае электронной почты существуют различные модели и правила, которые используются для класси-фикации электронной почты под меткой спам или не спам перед отправкой клиенту. Он включает в себя такие фильтры, как массовый почтовый фильтр, явную блокировку, размещение нулевого отправителя и проверку заголовка пустого отправителя. Обнаружение спама в коротких сообщениях (СМС) не привлекло большого внимания исследователей. Имеется ряд существенных различий между обнаружением спама в электронной почте и в текстовых сообщениях. Для коротких текстовых сообщений доступно несколько баз данных с ограниченной длиной текстовых сообщений. Небольшая длина текста препятствует ухудшению классификации и способности прогнозирования алгоритма. В основном это короткие сообщения без заголовков. В некоторых случаях первые несколько предложений используются для определения темы текстового сообщения. В коротких сообщениях используется много символов и неформальных слов. Эти факторы увеличивают сложность процесса обнаружения спама в таких сообщениях.

**Материалы и результаты исследования**. В этой исследовательской работе для обнаружения спама в коротких текстовых сообщениях используются два известных механизма обработки естественного языка *Word embeddings* и *TF-IDF*.

***Word embedding***. *Word embedding* - это механизм представления текста в виде векторов числовых значений. Он использует словарь *FastText*, который состоит из миллиона уникальных токенизированных слов в английском языке вместе с 300 измерениями (*dimensions*). Пространственный вектор используется для идентификации отношений между двумя словами. Например, слово мужчина и женщина имеют аналогичное измерение, чем слово мужчина и яблоко.

*N*-мерный вектор из текста генерируется с помощью функции *Word2Vec*. *Word2Vec* преобразовывает дискретные символы в непрерывные векторы числовых значений. Данная система работает на гипотезе распределения. В функции *Word2Vec* большой текстовый корпус использует алгоритм *Unsupervised*. Здесь для каждого слова в тексте его смысл прогнозируется с использованием контекста *BAG-OF-WORDS (CBOW)*. *Word2Vec* - это нейронная сеть с одним скрытым слоем (с размером *d* (*dimensions*)) и функцией оптимизации отрицательная выборка (*Negative-Sampling*).

***Term Frequency*** – I*nverse Document Frequency (TF-IDF)*. Похожий на *Word embedding, TF-IDF* представляет собой механизм для представления текстовых документов в матричной форме. Каждый текстовый доку-мент преобразуется в строку матрицы *TF-IDF*. *TF-IDF* это разреженная матрица, в которой количество ненулевых элементов в векторах равно количеству уникальных слов в документах. Для создания матрицы *TF-IDF* из текстового документа предварительно обрабатывается документ, который включает в себя удаление из него *STOPWORDS*, чисел, знаков пунктуации и использование заглавных букв всех слов документа.

Например, для предложений. *King is great. He is not*

*stupid. He is intelligent.* Матрица *TF-IDF* в ее простом виде будет:

*Таблица 1 – TF-IDF матрица предложений*

| Sentence/ Document | King | is | Great | He | Not | stupid | Intelligent |
|---|---|---|---|---|---|---|---|
| S1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| S2 | 0 | 0 | 0 | 1 | 1 | 1 | 0 |
| S3 | 0 | 1 | 0 | 1 | 0 | 0 | 1 |

*S1* представляет текст или предложение. Каждое слово в тексте имеет свой собственный вектор-столбец в зависимости от его вхождения в разных предложениях текста. *TF-IDF* генерирует вектор для каждого слова в тексте.

***История анализа эмоций***. Развитие общей системы запросов (1996) (Филиппа Стоун, [1]) стала первой для определения эмоций в тексте. Обычно система подсчитывает примеры положительных и отрицательных эмоций. После этого было проделано много работы для определения эмоций, выраженных в текстах на различных языках. Важным вкладом стали труды Яниса Виби, Петера Терни и Василеуса Хацивасилоглу в ранние 90-е. Янис Виби (1990, [2]) определяет термин «Субъективность» для исследования поиска информации. Позднее, в 1997 году, Хацивасилоглу и др. 1997, [3] определил семантическую ориентацию прилагательных. Спустя несколько лет Петер Терни [4] нашел революционный подход *Thumbs Up* (знак поднятый большой палец) и *Thumbs Down* (знак опущенный большой палец) для классификации положительных и отрицательных отзывов. Авторы Пэнг и др. в 2002 году [5] предложили вручную создать лексикон настроений для обзоров фильмов. Они пришли к выводу, что методы машинного обучения лучше выполняются на созданном вручную лексиконе для анализа настроений. Автор Денеке в 2009 году [6] объявил интересное исследование нескольких доменов, чтобы продемонстрировать преимущество предварительной оценки полярности в *SentiWordNet. SentiWordNet* - это онлайн-ресурс, содержащий список слов, используемых для выражения эмоций. Каждое слово имеет определенное значение, называемое знаком полярности, которое может быть положительным, отрицательным или нейтральным в зависимости от типа эмоций, которые оно выражает. В России исследования, посвященные анализу чувств, до 2011 года не так многочисленны. Ермаков представил в 2009 году [7] систему анализа чувств, выявляющую мнения об автомобилях в российских блогах. Анализ эмоций в тексте на русском языке в основном появляется в многоязыковых экспериментах. В международных исследованиях анализ русского настроения проявляется главным образом в многоязычных экспериментах. Загибалов и др, в 2010 году [8] представил сопоставимые экземпляры книжных обзоров, состоящие из двух частей: русской и английской, представляющих два очень разных языка. Корпусы сопоставимы по размеру и стилю. Они также содержат краткое описание специфики языка и области, которые наблюдаются в этих корпусах. В книге Штайнбергера и др в 2011 году [9] описана конструкция вокабулярий общих эмоций для различных языков.

Первая попытка анализа шаблонов спама в тексте была сделана в 2001 году Фабрицио Себастьяни. Он использовал анализ текста для изучения и обнаружения спам-паттернов в большом количестве текста. Он сделал классификацию текс-та для извлечения полезной информации, связанной с содержанием сообщений [10]. В 2004 году Делани и др. использовали наивный байесовский классификатор и алгоритм опорных векторов для представления признаков коротких текстовых сообщений. Согласно им, идентификация спама и его фильтрация из текста является важным механизмом в мире беспроводной связи [11]. В 2006 г. Идальго и соавторы для проведения эксперимента использовали набор данных СМС на испанском и английском языках. Они предположили, что байесовские методы фильтрации выполнялись лучше, чем другие методы машинного обучения [12]. В 2007 году Кормак и соавторы использовала методы представления особенностей для выявления спама. Они пришли к выводу, что для выявления спама необходим хороший уровень адаптации алгоритма к набору функций [13]. В 2011 году Джунаид и соавторы представили механизм рассылки спама в коротких текстовых сообщениях (СМС). Они изучили поведение клиента при получении спама СМС и наблюдали за нарушениями политики конфиденциальности [14]. В 2012 году Делани и др. изучили увеличение использования коротких текстовых сообщений в развивающихся странах. Они определили причины резкого увеличения его использования и проблемы, возникающие из-за этого роста [15]. В течение того же года Идальго и соавторы изучали прикладные алгоритмы машинного обучения на СМС-спаме. Они отметили, что машина опорных векторов работала лучше, чем другие алгоритмы идентификации спама СМС [16]. В 2013 году Алмейда и др. использовали набор данных для сбора неповторяющихся сообщений. Они применяли разные алгоритмы для классификации спама по разным категориям [17]. В 2015 году Дипак и соавторы сделали классификацию спама СМС с использованием алгоритма, названного «Слабым алгоритмом». Они создали программу, используя «Слабый алгоритм», чтобы рассчитать ее точность для классификации спама. По их словам, данная программа обладает высокой точностью обнаружения спама в СМС [18]. В 2016 году Сулиман и соавторы использовали общие символы в качестве механизма идентификации текстовых сообщений со

спамом [19]. В 2018 и 2019 годах Аббаси и соавторы изучили внутреннюю структуру эмоций в тексте. Они проанализировали логические, семантические и синтаксические характеристики текстового документа [20-23]. Было отмечено, что в основном исследователи использовали классический метод обнаружения спама в коротких текстовых сообщениях. Размер набора данных был очень ограничен. Они ис-пользовали словарь тэгов общих символов или слов, которые считаются спамом. Когда размер набора данных увеличивается, эффективность и точность алгоритмов уменьшается. В этой работе предложенная методология использовала современный словарь *Word embedding*, который преобразует каждое слово текста в непрерывный вектор числовых значений с *N* номером измерения, представляющим глобальное содержимое. Принимая во внимание, что матрица *TF-IDF* со-держит векторные значения слова на основе его локального контекста в конкретномтекстовом документе. Производительность обеих моделей для обнаружения спама анализируется и сравнивается. Для эксперимента на обеих моделях используется набор данных спама из (*https://www.kaggle.com/uciml/CMC-spam-collection-dataset*). Набор данных представляет собой набор из 5 574 коротких сообщений СМС на английском языке, которые помечены как *Ham* или *Spam*. Набор данных в формате *CSV* был загружен и реализован с использованием набора инструментов языка программирования *Matlab*. Модели *Word embedding* и *TF-IDF* были обучены и протестированы на наборе данных спама. *Word embedding* состоит из токенизированных слов с 300 измерениями каждого слова. В случае обучения модели для обнаружения спама требуется фраза составных слов из набора данных спама. Для этой цели токенизированные слова из вложения *Word* были объединены для создания словосочетания. Каждая фраза состоит из 15 слов, которые показаны в таблице 2.

*Таблица 2 – Фразы и их категория для обучения модели*

| Category | Phrase |
|---|---|
| Ham | Go until jurong point crazy Available only in bugis great world buffet Cine there got |
| Spam | Free entry 2 wkly comp to win FA Cup final tkts 21st May 2005 Text |
| Ham | Even my brother not like to speak with me They treat me like aids patent |
| Ham | I m gonna home soon I don't want to talk about this stuff anymore tonight |
| Ham | Even my brother not like to speak with me They treat me like aids patent |
| Spam | England v Macedonia dont miss the goals team news Txt ur national team to 87077 |
| Spam | Had your mobile 11 months more entitled Update latest colour mobiles with camera for Free |
| Ham | Is that seriously how you spell his name and where do you meet him everyday |
| Ham | Thats cool I am gentleman and will treat you with dignity and respect as I. |
| Ham | Yes I started send requests make it but pain came back so I m back |

Размеры каждой фразы в этом случае становятся 15 × 300 = 4500 размеров. В случае, если в последнем предложении или фразе не останется 15 слов, место пропущенных слов будет заменено векторами нулей для уравновешивания фразы. Текст был преобразован в вектор непрерывных числовых значений и был сохранен в таблицу, которая состоит из номера документа × вектор признаков. Модель создана для проведения обучения и тестирования набора данных спама, который можно увидеть на рисунке 1 ниже.

```
emb = fastTextWordEmbedding

filename = "spam.csv";
data = readtable(filename,'TextType','string');
textData = data.text;
documents = tokenizedDocument(textData);
sequences = doc2sequence(emb,documents,'Length',15);

% Initialize the table and add the data
data = table;
data.word = [label;sequences];
pred = [label;sequence];
data = [data array2table(pred)];
data.resp = zeros(height(data),1);
data.resp(1:length(sequence)) = 1;

% Preview the table
head(data(:,[1,end,2:8 ]))


rng('default') % for reproducibility
c = cvpartition(data.resp,'Holdout',0.6);
train = data(training(c),2:end);
Xtest = data(test(c),2:end-1);
Ytest = data.resp(test(c));
Ltest = data(test(c),1);
Ltest.label = Ytest;
```

*Рисунок 1 – Код для создания модели Word embedding для обучения и тестирования набора данных*

Для преобразования текста в документ «Матрица Термин Частота - Обратная частота документа» (*TF-IDF*) вычисление было выполнено с использованием формул:

$$TFIDF = TF \times IDF$$

$$TF = \frac{No.of\ repetitions\ of\ aword\ in\ a\ document}{Total\ No.of\ words\ in\ a\ document}$$

$$IDF = Log\left(\frac{No.of\ documents}{No.of\ documents\ containing\ word}\right)$$

Например, для предложений: *King is great. He is not stupid. He is intelligent. IDF* представлен в таблице 3.

*Таблица 3 Матрица IDF для слов из трех предложений*

| Word | IDF value | Doc 1 | Doc 2 | Doc 3 |
|---|---|---|---|---|
| King | Log (3/1) = 1.09 | 1/3 | 0/4 | 0/3 |
| Is | Log (3/3) = 0 | 1/3 | 1/4 | 1/3 |
| Great | Log (3/1) = 1.09 | 1/3 | 0/4 | 0/3 |
| He | Log (3/2) =0.41 | 0/3 | 1/4 | 1/3 |
| Not | Log (3/1) =1.09 | 0/3 | 1/4 | 0/3 |
| Stupid | Log (3/1) =1.09 | 0/3 | 1/4 | 0/3 |
| Intelligent | Log (3/1) =1.09 | 0/3 | 0/4 | 1/3 |

The *TF-IDF* значения матрицы для трех приведенных выше предложений представлены в таблице 4.

*Таблица 4 – TF-IDF Матрица для слов из трех предложений*

| Doc. | King | is | Great | He | Not | stupid | Intelligent |
|---|---|---|---|---|---|---|---|
| S1 | 0.363 | 0 | 0.363 | 0 | 0 | 0 | 0 |
| S2 | 0 | 0 | 0 | 0.102 | 0.272 | 0.272 | 0 |
| S3 | 0 | 0 | 0 | 0.446 | 0 | 0 | 0.363 |

Для обнаружения спама из коротких текстовых сообщений было выбрано 4500 наиболее часто встречающихся уникальных слов для обучения модели *TF-IDF.* Была создана таблица, состоящая из номера документа × вектор объекта. Каждый документ был представлен строкой, а каждое слово имело свой собственный вектор-столбец. *TF-IDF* был намного эффективнее *Word embedding* в использовании. Код для создания TF-IDF можно наблюдать на рисунке 2.

```
data = table;
data.word = [label;sequences];
pred = [label;sequence];
data = [data array2table(pred)];
data.resp = zeros(height(data),1);
data.resp(1:length(sequence)) = 1;

head(data(:,[1,end,2:8 ]))


rng('default')
c = cvpartition(data.resp,'Holdout',0.6);
train = data(training(c),2:end);
Xtest = data(test(c),2:end-1);
Ytest = data.resp(test(c));
Ltest = data(test(c),1);
Ltest.label = Ytest;

% Train model
mdl = fitcdiscr(train,'resp');
% Predict on test data
Ypred = predict(mdl,Xtest);
cf = confusionmat(Ytest,Ypred);


% Display results
```

*Рисунок 2 – Код для обучения и тестирования набора данных спама*

Для целей обучения использовалось 60% набора данных, а остальные – для целей тестирования. *SVM* применяется для обучения обеих моделей, и результаты приведены ниже в таблице 5.

*Таблица 5 Сравнение модели TF-IDF и Word embedding для обнаружения спама*

| Model | Precision | Recall | F1 Score |
|---|---|---|---|
| TF-IDF | 0.9726 | 0.8567 | 0.91052 |
| Word embedding | 0.9612 | 0.8236 | 0.89012 |

Из приведенной таблицы 5 можно наблюдать, что *TF-IDF* работает лучше и имеет высокие значения точности, отзыва и *F1 Score*. *Word embedding* является более сложным по своей природе и требует больше времени для выполнения.

**Заключение**. Исследование состоит из анализа проделанной работы по обнаружению спама в коротких текстовых сообщениях (СМС). Было отмечено, что эта область до сих пор не исследована, и предлагаемых методов недостаточно для больших наборов данных с неформальными символами и словами. В этом исследовании две современные модели обработки естественного языка использовались для обнаружения спама из набора коротких текстовых сообщений. Было отмечено, что модель *TF-IDF* работает лучше, чем *Word embedding*. В *TF-IDF* меньше шансов на подгонку (*Over fitting*) по сравнению с *Word embedding*, так как словарь небольшой, и правила могут быть сделаны только для часто встречающихся примеров в наборе обучающих данных. Вложение слов содержит много неиспользуемых и ненужных символов для анализа текста.

**СПИСОК ЛИТЕРАТУРЫ:**

1. Stone P.J., Dunphy D.C., Smith M.S. The General Inquirer: A Computer Approach to Content Analysis / MIT Press - Cambridge, 1966. 519 p.

2. Wiebe, Janyce M. Identifying Subjectivity characters in Narrative // Proc. 13th International Conference on Compu-tational Linguistics. Helsinki, 1990, pp. 401-408.

3. Vasileios H., Kathleen R. M. Predicting the Semantic Orientation of Adjectives // Proc. 8th Conference on European chapter of the Association for Computational Linguistics. Spain, 1997, pp 174-181.

4. Peter D. T. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews // Proc. of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia, 2002, pp. 417-424.

5. Pang B., Lee L. Thumbs up? Sentiment Classification using Machine Learning Techniques // Proc. Conference on Empirical Methods in Natural Language Processing. Philadelphia, 2002, pp. 79-86.

6. Denecke K. Are SentiWordNet scores suited for multi-domain sentiment classification? // Proc. 4th International Conference on Digital Information Management. USA, 2009, pp. 33-38.

7. Ermakov A. Knowledge extraction from text and its processing: Current state and prospects // Proc. of the Computational Linguistics and Intellectual Technologies. 2009, pp. 50–55.

8. Zagibalov., Taras., Belyatskaya et al. Comparable English-Russian Book Review Corpora for Sentiment Analysis. Russia, 2010. 67 p.

9. Steinberger J., Lenkova P., Kabadjov M. Multilingual Entity-Centered Sentiment Analysis Evaluated by Parallel Corpora // Proc. of Recent Advances in Natural Language Processing. Bulgaria, 2011, pp. 770–775.

10. Fabrizio Sebastiani. Machine Learning in Automated Text Categorization // arXiv:cs/0110053 [cs.IR] Oct 2001

11. Delany. Sarah Jane., Pádraig Cunningham. An analysis of case-base edit-ing in a spam filtering system //Advances in Case-Based Reasoning, Spring-er Berlin Heidelberg, 2004. pp. 128-141.

12. Gómez H.J., Cajigas B.G., Puertas S.E., Carrero G.F. Content Based SMS Spam Filtering // Proceedings of the 2006 ACM Symposium on Document Engineering (ACM DOCENG'06), Amsterdam, The Netherlands, 2006, pp. 10-13.

13. Cormack G.V., Gómez J. M., Puertas S.E. Feature engineering for mobile (SMS) spam filtering // Proceedings of the 30th Annual international ACM Conference on Research and Development in information Retrieval (ACM SIGIR'07), New York, NY,

2007, pp. 871-872.

14. Bilal J., Muddassar F. Using Evolutionary Learning Classifiers to Do Mo-bile Spam (SMS) Filtering // GECCO'11, July 2011, Dublin, Ireland.

15. Delany S.J., Buckley M., Greene D. SMS Spam Filtering: Methods and Da-ta // Expert Systems with Applications, 2012, vol. 39, issue 10, pp. 9899-9908.

16. Gómez H., J.M Almeida., Yamakami A. On the Validity of a New SMS Spam Collection // Proceedings of the 11th IEEE International Conference on Machine Learning and Applications (ICMLA'12), Boca Raton, FL, USA, 2012.

17. Almeida T.A., Gómez J.M., Silva, T.P. Towards SMS Spam Filtering: Re-sults under a New Dataset // International Journal of Information Security Science (IJISS), 2013, vol. 2, issue 1, pp. 1-18.

18. Dipak R., Kawade., Kavita S. SMS Spam Classification using WEKA // In-ternational Journal of Electronics Communication and Computer Technolo-gy (IJECCT), 2015, vol. 5, issue 1, pp. 43-47.

19. Sulaiman N.F., Jali M.Z. (2016) A New SMS Spam Detection Method Using Both Content-Based and Non Content-Based Features // Advanced Computer and Communication Engineering Technology. Lecture Notes in Electrical Engineering, Springer, 2016, vol. 362. pp. 505-514.

20. Beltiukov A. P., Abbasi M.M. Logical analysis of Emotions in Text from Natural language // Vestnik Udmurtskogo Universiteta. Matematika. Mek-hanika. Komp'yuternye Nauki, Ижевск, 2019, vol. 1, issue. 29, pp. 106-116.

21. Abbasi M.M., Beltiukov A.P. Identifying the strength of emotions in re-lation with the topic of text using Word space // Proceedings of the 21th in-ternational workshop on computer science and information technologies, Austria, Vienna, 2019 // Journal of Atlantis Highlights in Computer Sci-ences, 2019, vol. 3, pp. 1-5.

22. Abbasi M.M., Beltiukov A.P., Hussain Lal., Abbasi A.Q. Analysis of emotions from texts for managing society // Infocommunication technolo-gies Journal, Academy of Telecommunications and In-formatics, Samara, 2019, vol. 2, issue. 17, pp. 246-254.

23. Abbasi M.M., Beltiukov A.P. Summarizing emotions from text using Plutchik wheel of emotion// Proceedings of the 7th All Russian Conference on Information technology for intelligent decision making support (ITIDS), Ufa, Russian Federation, 2019, vol. 166, pp. 291-294.

*Статья поступила в редакцию 10.05.2020*
*Статья принята к публикации 10.06.2020*