УДК 004.85

G. M. Mutanov, Zh. D. Mamykova, V.I. Karyukin, A.Zh. Zhaksykeldi
(al-Farabi Kazakh National University, Almaty, Republic of Kazakhstan
Email: zhmamykova@kaznu.kz, vladislav.karyukin@gmail.com)

DEVELOPMENT OF A MACHINE-LEARNING ALGORITHM FOR DETERMINING THE SENTIMENT OF THE USER'S PERCEPTION OF SOCIAL MEDIA CONTENT

Abstract. The analysis of textual data is essential in order to get information about events taking place in the world. Due to the rapid development of the Internet, the increase in the number of websites, blogs and social networks, the problem of automatic data processing arises. The use of machine learning algorithms has an important role for the analysis of the emotional aspect of the news topics posted on the network and user opinions on the events described in them. This article has reviewed systems monitoring social media content and the development of a module for automatic classification of the sentiment of emotional aspects of the news topics and user comments using machine learning algorithms for OMSystem.

Keywords: Internet resources, social media, OMSystem, news topics, user opinions, machine learning, supervised learning, data classification, metrics, ROC curve.

Г.М. Мутанов, Ж.Д. Мамыкова, В.И. Карюкин, А.Ж. Жақсыкелді
(Казахский Национальный Университет имени аль-Фараби, г. Алматы, Республика Казахстан
Email: zhmamykova@kaznu.kz, vladislav.karyukin@gmail.com)

РАЗРАБОТКА МАШИННО-ОБУЧАЕМОГО АЛГОРИТМА ОПРЕДЕЛЕНИЯ ТОНАЛЬНОСТИ ПОЛЬЗОВАТЕЛЬСКОГО ВОСПРИЯТИЯ КОНТЕНТА СОЦИАЛЬНЫХ МЕДИА

Аннотация. Анализ текстовой информации имеет важное значение для получения сведений о событиях, происходящих в мире и обществе. В связи с быстрым развитием Интернета, увеличением числа сайтов, блогов и социальных сетей возникает задача автоматической обработки данных. Применение алгоритмов машинного обучения имеет важную роль для анализа эмоционального окраса размещенных в сети новостных топиков и мнений пользователей на события, описанные в них. В данной статье были рассмотрены системы мониторинга контента социальных медиа и разработка модуля автоматической классификации тональности эмоционального окраса новостных топиков и пользовательских комментариев с применением алгоритмов машинного обучения для системы мониторинга OMSystem.

Ключевые слова: Интернет-ресурсы, социальные медиа, OMSystem, новостные топики, пользовательские мнения, машинное обучение, обучение с учителем, классификация данных, метрики, ROC-кривая.

Введение

Активное развитие социальных медиа привело к значительному росту числа новостных порталов, статей и пользователей, активно читающих и комментирующих содержание текстов [1]. На сегодняшний день в казахстанском сегменте сети Интернет определение тональности статей и комментариев в основном производится вручную или на основе лингвистических правил, учитывающих структурные особенности и семантику слов [2]. При этом лингвистические правила подразумевают наличие большого корпуса слов, покрывающих большинство используемых в языке конструкций. Хотя данный подход позволяет получить достаточно точные оценки, он занимает много времени и ресурсов. Решением данной проблемы является применение широко распространенных алгоритмов машинного обучения для автоматической классификации текстовых данных. В данной статье рассматривается создание классификатора, обученного на наборе статей и комментариев пользователей из казахстанских социальных медиа. Результаты работы алгоритма оцениваются на тестовом наборе данных путем сравнения значений.

Методы

В автоматической обработке данных выделяют следующие важные подходы, находящие широкое применение: правила языка, тональные словари, алгоритмы машинного обучения (с учителем и без учителя) [3].

1. Первый тип включает набор правил, с помощью система определяет тональность текста;
2. Второй тип основывается на базовых тональных словарях. Создается словарь, в котором слова содержат определенный оттенок эмоционального окраса (положительный, нейтральный, отрицательный) и относятся к определенной категории тем;
3. Обучение с учителем. Методы обучения с учителем применяются тогда, когда для имеющихся объектов обучающей выборки уже выставлены значения меток (labels);
4. Обучение без учителя. В данном случае не выставлены значения меток. Для определения тональности текста необходимо найти зависимости и связи между объектами.

Существуют следующие казахстанские и зарубежные системы мониторинга контента социальных медиа: iMAS, Alem media monitoring, Brand Analytics, Microsoft Engagement и другие.

iMAS позволяет анализировать общественное мнение в режиме реального времени, получать информацию о критических замечаниях общества в СМИ, соцсетях, блогах, форумах. Alem media monitoring выполняет мониторинг всех актуальных для Казахстана социальных сетей и позволяет быть в курсе различных событий. Brand Analytics – российская система мониторинга мнений пользователей о брендах компаний в социальных медиа. Microsoft Social Engagement – платформа для наблюдения и взаимодействия с социальным медиа пространством. Предоставляется возможность взаимодействия с общественностью путем обсуждения продуктов и услуг компании.

В настоящее время исследователями Казахского национального университета им. аль-Фараби разработана платформа мониторинга казахстанского медиа пространства OMSystem (Opinion monitoring system) [4], выполняющая анализ мнений и упоминаний в информационном пространстве. Мониторинг системы охватывает новостные порталы Казахстана, социальные сети: ВКонтакте, Twitter, Facebook, Instagram, YouTube, а также блоги и сайты отзывов. OMS позволяет выполнять широкий круг функций: автоматизация рутинных операций мониторинга информационного пространства, поиск и обработка большого массива данных; оперативный мониторинг социальных сетей по горячей теме; анализ мнений пользовательского восприятия событий в обществе; отслеживания упоминания бренда; выявление существующих и потенциальных источников негатива и острых дискуссий; отслеживание динамики вовлеченности пользователей в ту или иную тему; оценка уровня социального самочувствия в обществе.

Система хорошо адаптирована под казахский и русский языки, в основе анализа данных лежит применение динамически наполняемого словаря «положительных» и «отрицательных» слов. Она способна за короткий промежуток времени оценить фон проявления социального самочувствия с целью оценить уровень настроения общества. Однако система проводит анализ данных только с использованием словарного подхода, требующего значительного промежутка времени на наполнение. Вычислительные мощности компьютера и широко используемые при анализе данных алгоритмы машинного обучения позволяют существенно ускорить процесс обработки данных, работая с полными текстами без необходимости применения огромных словарей. Разработка модуля машинного обучения OMSystem [4], основанного на обучении с учителем размеченного экспертами корпуса текстов казахстанских новостных порталов, позволит улучшить точность оценки тональности данных.

Реализация

Для выполнения анализа данных была сформирована база данных, включающая полные тексты статей из таких известных казахстанских новостных порталов, как tengrinews.kz, zakon.kz, nur.kz, informburo.kz, times.kz, today.kz, bnews.kz. Отдельно были сформированы таблицы, содержащие комментарии пользователей к соответствующим статьям. Было оценено 2430 статей и 11486 комментариев.

Разработанный машинно-обучаемый алгоритм определения тональности пользовательского восприятия контента социальных медиа включает 4 этапа, показанных на рис. 1.

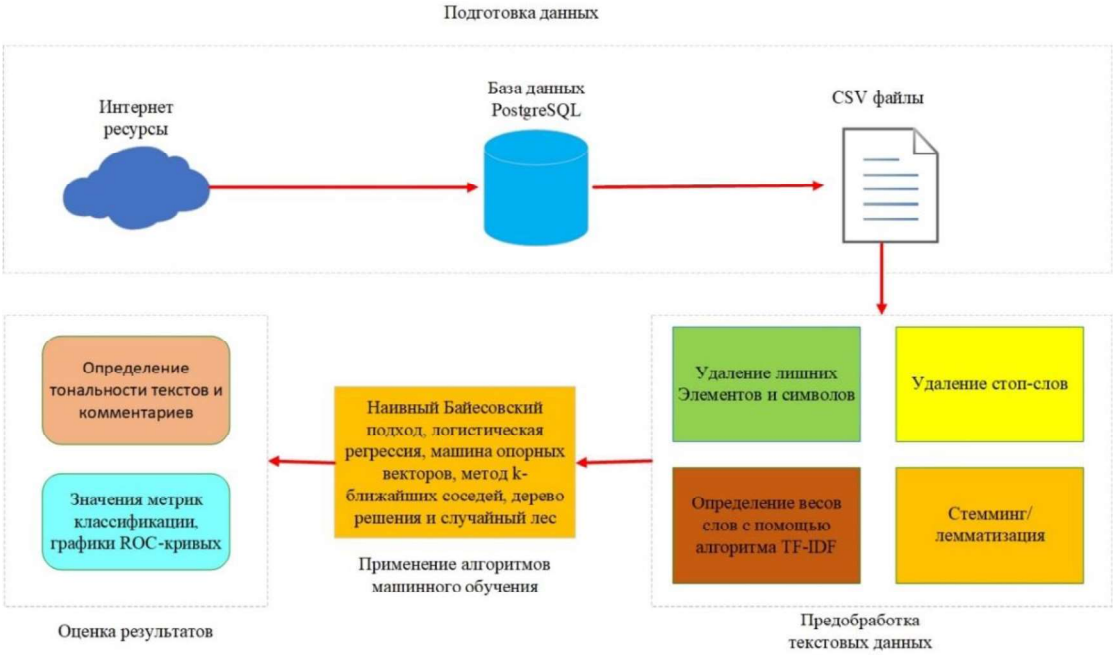

Рис. 1. Этапы машинно-обучаемого алгоритма

Этап «Подготовки данных» включает сбор информации поисковым механизмом системы OMSystem из казахстанских новостных порталов, формирование базы данных размеченного экспертами корпуса текстов и создание документов, содержащих параметры, необходимые для алгоритмов классификации данных: текст статьи/комментария, тональность и язык.

Этап «Предобработки текстовых данных» начинается с удаления символов, знаков, слов, не несущих какой-либо смысловой нагрузки [5]. После этого для уменьшения числа различных слов выполняется стемминг путем удаления из слова суффиксов и окончаний, чтобы оставалась лишь основа, которая одинакова для всех его форм. Для русских слов был использован SnowballStemmer из библиотеки NLTK [6-7] на языке программирования Python [8-9]. Для слов на казахском языке был разработан собственный стеммер Kazakh_Stemmer. Данный стеммер выполняет удаление лишних символов, стоп-слов, с использованием регулярного выражения [^a-zA-Za-яA-ЯӘІҢҒҮҰҚӨәіңғүұқөh]. Для выполнения стемминга имеется база данных окончаний и суффиксов казахского языка. Производится их сопоставление с частями слова и удаление найденных частей в случае совпадения. Если слово имеет длину меньше или равную двум, то оно остается без изменения. Данное ограничение вызвано тем, что некоторые слова, состоящие из двух букв, схожи с окончаниями и суффиксами.

После очистки текстов требуется перевести их в векторное представление (vectorization). Эффективным способом векторного представление является применение метрики *TF-IDF* [10] для слов в каждом документе (одна запись в таблице является одним текстовым документом). Расчет метрики *TF-IDF* представлен следующими формулами:

$$TF_{d,w} = \frac{count(w,d)}{count(N,d)} \tag{1}$$

где $count(w,d)$ – количество раз слово w встречается в документе d, а $count(N,d)$ – общее количество слов N в документе d.

$$IDF_{d,w} = \log \frac{count(D)}{count(w,d)} \tag{2}$$

где $count(D)$ – общее количество документов, $count(w,d)$ – количество раз слово w встречается в документе d [11].

Этап «Применения алгоритмов машинного обучения» реализуется после векторизации метрикой *TF-IDF* и заключается в использовании следующего ряда алгоритмов машинного обучения для классификации данных: Наивный Байесовский классификатор (Naïve Bayes) [12], логистическая регрессия, машина опорных векторов, метод k-ближайших соседей, дерево решения и случайный лес [13-14] его.

Наивный Байесовский классификатор – очень распространенный метод векторного анализа, который показывает результаты не хуже, чем другие более сложные классификаторы. Этот классификатор берет за основу условную вероятность принадлежности документа *d* к классу *c* и использует формулу Байеса. Для модели анализа текстов формула имеет следующий вид

$$P(c \mid d) = \frac{P(c) * P(d \mid c)}{P(d)} \tag{3}$$

где вектор: $d = \{x_1, x_2, \ldots, x_n\}$, x_i – вес *i*–ого слова документа, а *c* – класс документа.

Логистическая регрессия сводится к применению линейного классификатора, который оценивает вероятность отношения объектов к определенному классу. Для бинарной классификации значения меток будут $y = \{0,1\}$. Применяется функция логистической регрессии

$$f(z) = \frac{1}{1 + e^{-z}} \tag{4}$$

Здесь *z* определяется по следующей формуле

$$z = \theta_1 * x_1 + \theta_2 * x_2 + \ldots + \theta_n * x_n \tag{5}$$

где $x_1, x_2, \ldots, x_n$ –веса, $\theta_1, \theta_2, \ldots, \theta_n$ – значения коэффициентов регрессионной функции.

Классификатор на основе метода машины опорных векторов SVM определяет каждый объект входных данных с помощью вектора $d = \{x_1, x_2, \ldots, x_n\}$. Для разделения объектов на определенное число классов с метками $y = \{0,1\}$ находится гиперплоскость с максимальным расстоянием между опорными векторами.

В методе k-ближайших соседей определяется расстояние от векторов тестовой выборки до векторов обучающей выборки. В качестве функции расстояния выбрано Евклидово расстояние.

$$\sqrt{\sum_{i}^{n} (x_i^{test} - x_i^{train})} \tag{6}$$

В разрешающем дереве при его построении применяются следующие правила: выбирается слово, и документы, содержащие его, откладываются в одну сторону, а не содержащие его – в другую. Таким образом, документы относятся к двум непересекающимся множествам. Для каждого множества выбирается новое слово, и вновь выполняется описанный выше этап. Процедура повторяется, пока не получится однородное множество, в котором все документы относятся к одному классу.

Алгоритм случайного леса использует множество разрешающих деревьев. Деревья строятся независимо друг от друга. При классификации документу присваивается тот класс, который определило наибольшее количество деревьев.

Этап «Оценки результатов» заключается в определении точности алгоритма классификации данных с помощью метрик и графиков ROC-кривых, дающих характеристику полученных результатов работы определенного алгоритма. Их вычисления основаны на истинно-положительных (TP), истинно-отрицательных (TN), ложно-положительных (FP) и ложно-отрицательных (FN) значениях данных.

Для бинарной классификации метриками являются правильность (accuracy), точность (precision), полнота (recall) и *F*-мера (*F*-measure).

• Правильность – это отношение количества верных предсказаний T к общему числу предсказаний $N = T + F$. Вычисление производится по следующей формуле

$$accuracy = \frac{T}{N} \tag{7}$$

• Точность – это отношение объектов, действительно принадлежащих определенному классу ко всем объектам, отнесенным к нему алгоритмом классификации данных. Вычисление производится по следующей формуле

$$precision = \frac{TP}{TP + FP} \tag{8}$$

• Полнота – это отношение объектов, верно определенных классификатором, ко всем объектам этого класса в тестовой выборке. Вычисление выполняется по следующей формуле

$$recall = \frac{TP}{TP + FN} \tag{9}$$

• F-мера – метрика, определяющая баланс точности и полноты. Вычисления выполняются по следующей формуле

$$F = 2\frac{precision * recall}{precision + recall} \tag{10}$$

Разделяем корпус текстовых данных на тренировочные в размере 70% и тестовые 30%. После этого выполняем обучение на тренировочных данных с применением каждого из указанных алгоритмов машинного обучения. Для оценки результатов производим тестирование на оставшихся 30% данных и выполняем оценку алгоритма классификации данных с помощью метрик и ROC-кривых. ROC-кривая, или кривая ошибок, – удобная графическая характеристика качества бинарного классификатора, показывающая зависимость доли верных положительных классификаций *TRP* от доли ложных положительных классификаций *FPR*, где

$$TPR = \frac{TP}{TP + FN} \tag{11}$$

$$FPR = \frac{FP}{FP + TN} \tag{12}$$

Полученные значения метрик классификации для каждого алгоритма представлены в таблице 1 и таблице 2:

Таблица 1. **Метрики алгоритмов классификации топиков**

Алгоритмы			
Логистическая регрессия	**Значения**	**Наивный Байесовский подход**	**Значения**
Accuracy	0.88	Accuracy	0.88
Precision	0.86	Precision	0.87
Recall	0.96	Recall	0.94
F-measure	0.90	F-measure	0.90
Метод К-ближайших соседей	**Значения**	**Машина опорных векторов**	**Значения**
Accuracy	0.83	Accuracy	0.90
Precision	0.83	Precision	0.89
Recall	0.91	Recall	0.95
F-measure	0.87	F-measure	0.92
Дерево решений	**Значения**	**Случайный лес**	**Значения**
Accuracy	0.75	Accuracy	0.81
Precision	0.80	Precision	0.78
Recall	0.77	Recall	0.96
F-measure	0.79	F-measure	0.86

Таблица 2. **Метрики алгоритмов классификации комментариев**

Алгоритмы			
Логистическая регрессия	**Значения**	**Наивный Байесовский подход**	**Значения**
Accuracy	0.73	Accuracy	0.71
Precision	0.79	Precision	0.85
Recall	0.62	Recall	0.50
F-measure	0.69	F-measure	0.64
Метод К-ближайших соседей	**Значения**	**Машина опорных векторов**	**Значения**
Accuracy	0.69	Accuracy	0.74
Precision	0.65	Precision	0.75
Recall	0.80	Recall	0.71
F-measure	0.72	F-measure	0.73
Дерево решений	**Значения**	**Случайный лес**	**Значения**
Accuracy	0.62	Accuracy	0.71
Precision	0.67	Precision	0.79
Recall	0.44	Recall	0.55
F-measure	0.53	F-measure	0.65

Оценим качество алгоритмов классификации данных с помощью ROC-кривых. Чем ближе значение кривой к единице, тем лучше результат классификации. Графики ROC-кривых для алгоритмов классификации топиков представлены на рис.2, а комментариев – на рис. 3.

Исходя из показаний метрик и графиков наилучших значений удалось достичь с применением алгоритмов машинного обучения: Наивный Байесовский подход, логистическая регрессия и машина опорных векторов. В то же время точность алгоритмов классификации топиков выше точности алгоритмов классификации комментариев.

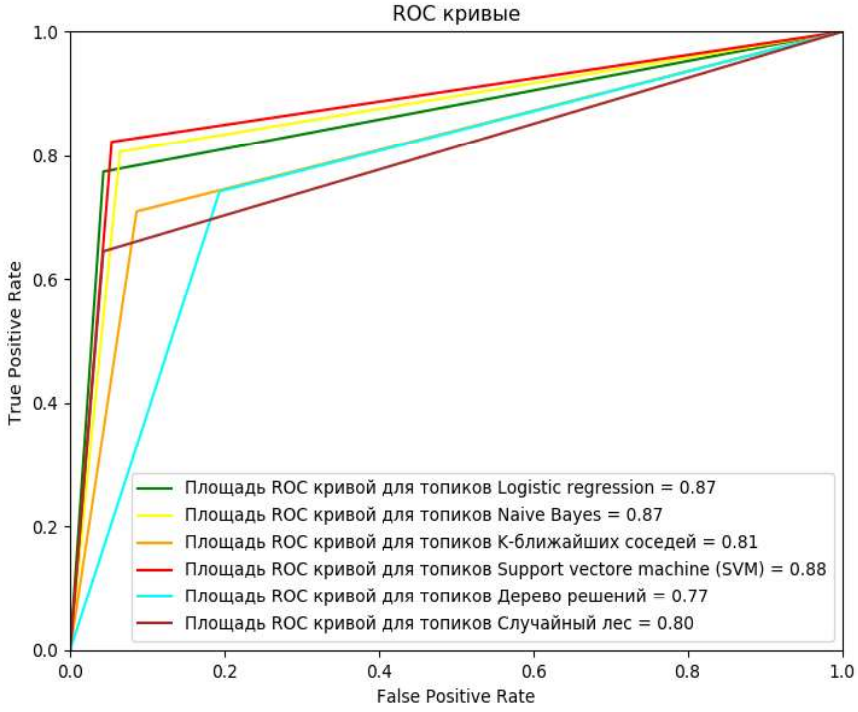

Рис. 2. ROC кривые алгоритмов классификации топиков

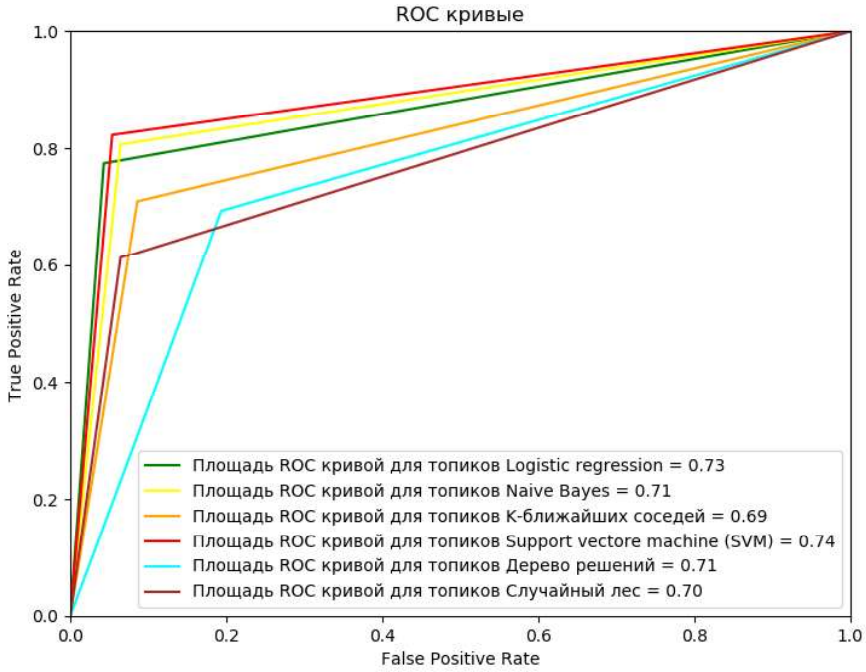

Рис. 3. ROC кривые алгоритмов классификации комментариев

Обсуждение

Выполнение классификации тональности текстов и пользовательских комментариев показало, что значения метрик оказались лучше для полных текстов статей, чем для комментариев. Это вызвано тем, что тексты статей имеют значительно больший размер, они хорошо структурированы и содержат малое количество ошибок. В то же время комментарии часто бывают очень короткими, включающими всего несколько слов, различных символов, многие из которых написаны с ошибками. Среди алгоритмов машинного обучения наилучшие значения метрик дали Наивный Байесовский подход, логистическая регрессия и машина опорных векторов. В статьях [14-15] и учебниках [6, 11, 13] использованные алгоритмы хорошо зарекомендовали себя при классификации текстовых данных, что показывает релевантность оценок полученных результатов данной работы. Хотя классические алгоритмы машинного обучения позволяют добиться хороших результатов, современные технологические достижения в области искусственного интеллекта широко используют искусственные нейронные сети. В связи с этим в дальнейшем планируется их применение в дополнение к классическим алгоритмам машинного обучения.

Результаты

По результатам разработки машинно-обучаемого алгоритма определения тональности пользовательского восприятия контента социальных медиа были выполнены все этапы: формирование данных, предобработка текстов, применение алгоритмов машинного обучения и анализ выходных данных в виде значений метрик классификации данных и графиков ROC-кривых. Размеченный корпус текстовых данных был сформирован в таблицах базы данных СУБД PostgreSQL, откуда он был выгружен в CSV файлы. Созданный модуль был добавлен в платформу OMSystem. Программный модуль создан с применением библиотек обработки естественных языков Natural language toolkit (NLTK), обработки и анализа данных Pandas, машинного обучения Scikit-learn и визуализации двумерной графики Matplotlib на языке программирования Python 3.7 дистрибутива Anaconda.

ЛИТЕРАТУРА

[1] Vaishali Kalra, Dr. Rashmi Aggarwal. Importance of text data preprocessing and implementation in RapidMiner // Proceedings of the First International Conference on Information Technology and Knowledge Management – New Dehli, India, – 2017 – Volume 14. – Pages 71-75.

[2] Mita K. Dalal, Mukesh A. Zaveri. Automatic Text Classification: A Technical Review // International Journal of Computer Applications. – 2011. – Volume 28. – Pages 37-40.

[3] Ubale Swati, Chilekar Pranali, Sonkamble Pragati. Sentiment analysis of news articles using machine learning approach // International Journal of Advances in Electronics and Computer Science. – 2015. – Volume 2 – Issue-4. – Pages 114-116.

[4] Мамыкова Ж.Д., Мутанов Г.М., Сундетова Ж.Т., Торекул С. М. Подходы к разработке информационной системы мониторинга мнений и оценки социального самочувствия // Вестник КазНУ. Серия математика, механика, информатика. – 2018. – N.4(100). – С. 63-77.

[5] Said A. Salloum, Mostafa Al-Emran, Azza Abdel Monem and Khaled Shaalan. Using text mining techniques for extracting information from research articles // Intelligent Natural Language Processing: Trends and Applications, Studies in Computational Intelligence. 2018 – Pages 373-397.

[6] Steven Bird, Ewan Klein, Edward Loper. Natural language processing with python. – 1st edition. – Sebastopol: O'Reilly Media, 2009. – Pages 504.

[7] Dipanjan Sarkar. Text analytics with Python. – New York: Apress, 2016. – Pages 385.

[8] Madhura Anil Zende, Megha Bhaskar Tuplondhe, Shalan Baban Walunj, Sujata Vasudev Parulekar. Text mining using Python // ISSN. – 2016. – Volume 3, Issue 3. – Pages 54-56.

[9] Benjamin Bengfort, Rebecca Bilbro, Tony Ojeda. Applied Text Analysis with Python: Enabling Language-Aware Data Products with Machine Learning. – 1st edition. – Sebastopol: O'Reilly Media, 2018. – Pages 332.

[10] Muthe Sandhya, Shitole Sarika, Sinha Anukriti, Aghav Sushila. Automatic text categorization on news articles // International Journal of Engineering and Techniques. – 2016. – Volume 2 – Issue 3. – Pages 33-38.

[11] Daniel Jurafsky, James H. Martin. Speech and Language Processing, an Introduction to Natural Language Processing, computational Linguistics and speech recognition. – 3rd edition. – Prentice Hall, 2018. – Pages 558.

[12] Aruna Gunda, Varsha Teratipally. Sentiment Analysis of Political News articles and the effect of negation scope // International Research Journal of Engineering and Technology (IRJET). – 2016. – Volume 3 – Issue 10. – Pages 1105-1109.

[13] Gavin Hackeling. Mastering Machine Learning with scikit-learn. – 1st edition. – Birmingham: Packt Publishing Ltd., 2014. – Pages 238.

[14] Megha Joshi, Purvi Prajapati, Ayesha Shaikh, Vishwa Vala. A Survey on Sentiment Analysis // International Journal of Computer Applications. – 2017. – Volume 163 – No 6. – Pages 34-38.

[15] Jochen Hartmann, Juliana Huppertz, Christina Schamp, Mark Heitmann. Comparing automated text classification methods // International Journal of Research in Marketing. – 2019. – Volume 36 – Issue 1. – Pages 20-38.

Мутанов Г.М., Мамыкова Ж.Д., Карюкин В.И., Жақсыкелді А.Ж.

Әлеуметтік медиа мазмұнын пайдаланушылардың тоналдығын анықтау үшін машина-оқыту алгоритмін әзірлеу

Түйіндеме. Мәтіндік ақпаратты талдау әлемдегі және қоғамдағы оқиғалар туралы ақпарат алу үшін маңызды. Интернеттің жылдам дамуына байланысты веб-сайттардың, блогтардың және әлеуметтік желілердің санын көбейтудің арқасында автоматты түрде деректерді өңдеу мәселесі туындайды. Машиналық оқыту алгоритмдерін пайдалану желіде орналастырылған жаңалықтар тақырыбының эмоционалдық түсінін талдау және оларда сипатталған оқиғалар туралы пайдаланушылардың пікірін анықтауда маңызды рөл атқарады. Бұл мақалада әлеуметтік медиа-контент мониторингі жүйелері және OMSystem мониторингі жүйесі үшін компьютерлік оқыту алгоритмдерін пайдаланатын жаңалықтар тақырыбының эмоционалдық түсі мен пайдаланушылық түсініктемелердің автоматты түрде жіктелуін қарастыратын модуль әзірленді.

Түйінді сөздер: Интернет ресурстары, әлеуметтік медиа, OMSystem, жаңалықтар тақырыптары, пайдаланушы пікірлері, машина жасау, мұғалімдерді оқыту, деректерді жіктеу, метрикалар, ROC қисықтары.

УДК 530.1

I.S. Tleubayeva, K.K. Dikhanbayev, Y. Shabdan

TECHNOLOGY OF OBTAINING NANOSTRUCTURED SILICON FILMS AND COMPARISON OF THEIR REFLECTIVE PROPERTIES

Abstract. The work is devoted to technology of producing silicon films containing nanostructures. The dependence of reflective properties of experimentally obtained nanostructures of silicon films on incident radiation wavelength was studied and comparative analysis was performed. The results were compared with similar dependence of monocrystalline silicon films without nanostructures.

Key words: porous silicon, nanostructure, silicon nanowires, chemical deposition method, electro-chemical anodizing, reflection, morphology.