А. Б. НУГУМАНОВА*, Д. Ж. АХМЕД-ЗАКИ², Е. М. БАЙБУРИН¹, К. С. АПАЕВ³

¹*Восточно-Казахстанский университет имени Сарсена Аманжолова,
Усть-Каменогорск, Казахстан*

²*Astana IT University, Нур-Султан, Казахстан*

³*Восточно-Казахстанский университет имени Даулета Серикбаева,
Усть-Каменогорск, Казахстан
yalisha@yandex.kz*

СЕНТИМЕНТ-АНАЛИЗ ОТЗЫВОВ ПОЛЬЗОВАТЕЛЕЙ В ФЕЙСБУКЕ: СРАВНЕНИЕ БИБЛИОТЕК TEXTBLOB И DOSTOEVSKY

В данной статье рассматривается работа двух библиотек Python TextBlob и Dostoevsky для определения тональности сообщений на русском языке. Библиотека TextBlob напрямую не поддерживает sentiment-анализ текстов на русском языке, но обладает возможностью передавать такие тексты на вход переводчику Google Translate, чтобы затем оценивать полярность этих текстов на основе сформированных английских переводов. Вторая библиотека Dostoevsky создана специально для sentiment-анализа текстов на русском языке и обучена на крупнейшем русскоязычном sentiment-корпусе RuSentiment, который содержит более 30 тысяч аннотированных вручную сообщений. Для сравнения указанных библиотек в данной статье используются реальные отзывы жителей города Нур-Султан о качестве жизни в столице, в том числе о деятельности городских служб и жилищно-коммунального сектора, работе объектов сферы обслуживания, образования, здравоохранения, культуры и т. д., опубликованные в общедоступных группах Фейсбука. Результаты проведенного sentiment-анализа оцениваются с помощью общепринятых метрик точности, полноты и F-меры и демонстрируют незначительное превосходство библиотеки Dostoevsky.

Ключевые слова: *сентимент-анализ, обработка естественного языка, анализ социальных сетей, умный город.*

Введение. Сентимент-анализ (анализ настроений) и извлечение мнений – это две активно развивающиеся отрасли обработки естественного языка, которые изучают проблемы автоматического анализа общественного мнения в масштабе больших данных. Как правило, эти отрасли ориентируются на социальные сети как основной источник данных для анализа, но также могут извлекать тренды общественного мнения из новостей, выступлений политиков, сайтов Интернет-торговли и сайтов рекомендаций. Существует определенное различие между сентимент-анализом и извлечением мнений. Методы извлечения мнений (Opinion Mining) выявляют мнение людей о тех или иных сущностях (товарах, услугах, событиях, персоналиях и организациях), в то время как методы сентимент-анализа (Sentiment analysis) выявляют настроения, выраженные в тексте [1-2]. Соответственно, сентимент-анализ можно рассматривать как задачу классификации сообщений как минимум по трем классам: сообщения с позитивной полярностью, сообщения с негативной полярностью и нейтральные сообщения. При этом полярность может определяться как на уровне всего текста или предложений, так и на уровне отдельных аспектов – характеристик сущности, о которой дается отзыв.

* E-mail корреспондирующего автора: yalisha@yandex.kz

Сентимент-анализ, проводимый на уровне аспектов, или, как его еще называют, аспектно-ориентированный сентимент-анализ, является наиболее востребованным в сфере Интернет-торговли, маркетинга и таргетинга, поскольку именно в этих сферах характерны отзывы с разными полярностями в одном и том же высказывании. Например, в высказывании «камера у данного телефона отличная, а вот звук оставляет желать лучшего» встречаются сразу обе полярности – позитивная по отношению к аспекту «камера» и негативная по отношению к аспекту «звук». В этом случае схема сентимент-анализа усложняется, к задаче определения полярности добавляется задача извлечения сущности и ее аспектов.

Сентимент-анализ на уровне всего текста или его отдельных предложений в основном применяется для измерения общественного мнения, для мониторинга социальных сетей, для оценки удовлетворенности граждан или потребителей оказываемым уровнем услуг. В данной работе мы исследуем технологии сентимент-анализа именно этого второго типа. Мы измеряем удовлетворенность жителей и гостей Нур-Султана качеством жизни в столице посредством сентимент-анализа сообщений в соответствующих открытых группах Фейсбука. Для реализации этой задачи мы используем две современные библиотеки Python TextBlob и Dostoevsky и сравниваем их между собой.

Методы и материалы. Получение отзывов пользователей мы выполняем путем скрапинга сообщений в заданных группах Фейсбука с использованием библиотеки Selenium [4]. Чтобы воспользоваться этой библиотекой, необходимо сначала установить ChromeDriver или другой драйвер, который отвечает за программное управление браузером, а затем настроить доступ, чтобы осуществлять навигацию в браузере, действуя как человек. Всего с помощью библиотеки Selenium мы извлекли 100 постов в группах, посвященных городу Нур-Султану, и более 1000 комментариев пользователей к этим постам. Полученные данные мы загрузили в среду Google Colab, на рисунке 1 показан фрагмент датасета с полученными данными.

ive already mounted at /content/gdrive; to attempt to forcibly remour

id	Text	Polarity
1	100 лошадей и 1000 баранов были куплены за сч...	1
2	Утром гулял с ребенком. Заметил очень много мо...	-1
3	Подполковник Абикен лишился поста начальника л...	-1
4	БАНКОМАТ-БАНДИТ Неделю назад, чтобы оплатить к...	-1
5	Добрый день! Есть ли тут охотники и рыбаки?Под...	0
6	Кто придумал такие дебильные правила?? Прием п...	-1
7	В городе стало много таких автомашин, причём р...	-1
8	Участок транспортной развязки на пересечении п...	0
9	В Нур-Султане благоустраивается новое общество...	0
10	В Нур-Султане более 200 волонтеров помогают пр...	0
11	Обладатель уникального голоса, выдающийся каза...	1
12	До и после: Как преобразилась столица за год С...	0
13	Международный турнир памяти Дениса Тен под эги...	0
14	В столице прошла патриотическая акция «Мені? т...	0
15	Яндекс такси, вы там водителей как-то контроли...	-1
16	Про вредность стаканчиков одноразовых.	-1
17	Казахстан - нефтедобывающая страна с огромными...	-1
18	Дорогие астанчане, будьте внимательны, смотрит...	0
19	Ничего нового, опять пишу про старую проблему....	-1
20	Здесь будет второй городской пляж.	0

Рисунок 1 – Фрагмент датасета с отзывами и комментариями пользователей

Как отмечалось выше, в качестве инструментов сентимент-анализа высказываний пользователей мы использовали две конкурирующие библиотеки языка Python:

1) библиотеку TextBlob [5], основанную на пакете NLTK и словаре слов с размеченной полярностью;

2) библиотеку Dostoevsky, обученную на русском корпусе образцов высказываний с размеченной полярностью [6].

Библиотека TextBlob не выполняет сентимент-анализ высказываний на русском языке, но зато поддерживает функцию их перевода на английский язык. Таким образом, для измерения полярности используются не оригинальные тексты, а их переводы. На рисунке 2 представлен фрагмент исходного датасета, к которому был присоединен столбец, содержащий переводы на английский язык.

	id	...	TextEn
0	1	...	The Seilkhanov family bought 100 horses and 10...
1	2	...	Walked with the child in the morning. I notice...
2	3	...	Lieutenant Colonel Abiken lost the post of hea...
3	4	...	ATM-BANDIT A week ago, in order to pay for the...
4	5	...	Good day! Are there hunters and fishermen here...
5	6	...	Who came up with such moronic rules ?? Receivi...
6	7	...	There are many such cars in the city, and they...
7	8	...	A section of a traffic intersection at the int...
8	9	...	A new public space is being improved in Nur-Su...
9	10	...	In Nur-Sultan, more than 200 volunteers help t...
10	11	...	The owner of a unique voice, the outstanding K...
11	12	...	Before and after: How the capital has changed ...
12	13	...	International tournament in memory of Denis Te...
13	14	...	A patriotic action "Meni? tuym! "In honor of t...
14	15	...	Yandex taxi, do you somehow control the driver...
15	16	...	About the harmfulness of disposable cups.
16	17	...	Kazakhstan is an oil-producing country with hu...
17	18	...	Dear Astana residents, be careful, look not on...
18	19	...	Nothing new, again I am writing about an old p...
19	20	...	There will be a second city beach here.

Рисунок 2 – Фрагмент датасета с отзывами и комментариями пользователей, а также с присоединенным столбцом их переводов

Анализатор TextBlob оценивает полярность и субъективность высказывания. Полярность оценивается по шкале от -1 до 1, отрицательные значения указывают на негативное настроение в тексте высказывания, а положительные указывают на позитивное настроение. Субъективность оценивается в диапазоне от 0 до 1, уровень субъективности равный 1 обычно означает выражение личного мнения или эмоций. Анализатор TextBlob действует на основе правил и заранее заданного лексикона тональностей. Если в тексте встречаются слова, тональность которых неизвестна, то анализатор пропускает такие слова. Сравните 2 примера на рисунках 3 и 4.

```

text_blob_ru = TextBlob('Связь в этом районе очень плохая')
text_blob_en = text_blob_ru.translate(from_lang='ru', to='en')

print(text_blob_en, '->', text_blob_en.sentiment)

Communication in the area is very poor -> Sentiment(polarity=-0.52, subjectivity=0.78)

```

Рисунок 3 – Полярность высказывания определена верно, т. к. определение «плохая» является негативно окрашенным и присутствует в лексиконе TextBlob

```

text_blob_ru = TextBlob('Завтраки в этом отеле выше всяких похвал')
text_blob_en = text_blob_ru.translate(from_lang='ru', to='en')

print(text_blob_en, '->', text_blob_en.sentiment)

Breakfasts at this hotel are beyond praise -> Sentiment(polarity=0.0, subjectivity=0.0)

```

Рисунок 4 – Полярность высказывания определена неверно, т. к. анализатор TextBlob не сумел распознать сложное определение «выше всяких похвал»

Мы выполнили sentiment-анализ высказываний с помощью TextBlob и вывели на экран 5 самых позитивных и 5 самых негативных высказываний (см. рисунок 5). Ни одно из высказываний не получило максимальной оценки ни в положительном, ни в отрицательном направлении.

id	Text	...	BlobPolarity	Pred
76	«Мир Керамики» на Бараева - не рекомендую нико...	...	-0.800000	-1.0
72	Очень опасный поворот тут, и дети ходят через ...	...	-0.780000	-1.0
21	Жирентаева 13, во дворе горка стоит, акимат убе...	...	-0.600000	-1.0
31	Когда уже решится проблема 61 маршрута? Это же...	...	-0.441667	-1.0
86	Опять молодой парень не справился с управление...	...	-0.300000	-1.0
..	...	...	...	...
11	Обладатель уникального голоса, выдающийся каза...	...	0.437500	1.0
93	Картошечка по 95 тг за кг\nВыглядит хорошо и м...	...	0.450000	1.0
83	Всем здравствуйте! Подскажите пожалуйста хорош...	...	0.450000	1.0
45	Доброе утро страна...	...	0.700000	1.0
66	Добрыи? день! Взяли вчера на вынос 2 пиццы с @...	...	0.750000	1.0

Рисунок 5 – Пять самых позитивных (сверху) и самых негативных (снизу) высказываний согласно TextBlob

В отличие от библиотеки TextBlob, библиотека Dostoevsky основана на машинном обучении. Для обучения анализатора Dostoevsky был использован крупнейший русскоязычный sentiment-корпус RuSentiment [], содержащий более 30 тысяч аннотированных вручную сообщений. Мы также выполнили sentiment-анализ высказываний с помощью Dostoevsky и вывели на экран 5 самых позитивных и 5 самых негативных высказываний (см. рисунок 6).

	text	dostoevsky	pred
79	Ул. Толеби, выезд на ул.Сыганак, до конца года...	-0.822199	-1.0
94	Акимат воовще не работает ...зачем нам акимат ес...	-0.760661	-1.0
68	Вот акимат города ровно так же как и граждане ...	-0.629785	-1.0
41	Вандалы на городском кладбище совсем страх пот...	-0.622469	-1.0
30	Когда уже решится проблема 61 маршрута?!это же...	-0.592677	-1.0
..	...	...	...
27	Месяц назад с акимом района "Алматы" Бекмурзае...	0.000000	0.0
26	Месячник осведомленности о раке молочной желез...	0.000000	0.0
99	Безобразие в Евразии. Вообще не проверяют QR н...	0.000000	0.0
76	Мне тоже не очень понравилось обслуживание, ос...	0.327678	1.0
74	А мне нравится. Все же лучше, чем матерные сло...	0.665421	1.0

Рисунок 6 – Пять самых позитивных (сверху) и самых негативных (снизу) высказываний согласно Dostoevsky

Результаты и их обсуждение. Для проверки качества работы сентимент-анализаторов мы вручную разметили полярность 100 высказываний (-1 – негативные, 0 – нейтральные, 1 – позитивные), а затем сравнили с оценками, предсказанными обеими моделями: TextBlob и Dostoevsky. Матрицы ошибок этих моделей представлены на рисунках 7–8 соответственно. Затем на основе этих матриц были рассчитаны точность, полнота и F1-мера для каждого из методов (см. таблицу 1).

Рисунок 7 – Матрица ошибок сентимент-анализа посредством TextBlob

Рисунок 8 – Матрица ошибок sentiment-анализа посредством Dostoevsky

Таблица 1 – Сравнение точности, полноты и F1-меры, определенных с помощью библиотек TextBlob и Dostoevsky, для всех трех классов высказываний

	TextBlob			Dostoevsky		
	Pr	Re	F1	Pr	Re	F1
Negative (-1)	0.79	0.34	0.47	0.95	0.31	0.47
Neutral (0)	0.24	0.16	0.19	0.32	1.00	0.49
Positive (1)	0.15	0.80	0.25	0.50	0.10	0.17

Анализ результатов, приведенных в таблице 1, показывает, что для негативных высказываний библиотека Dostoevsky дает более точные результаты, но при этом жертвует полнотой, в то время как библиотека TextBlob дает большую полноту, но при этом захватывает лишние высказывания, т. е. жертвует точностью. На позитивных высказываниях результаты гораздо хуже, но также наблюдается перекося в сторону точности у библиотеки Dostoevsky, и в сторону полноты у библиотеки TextBlob. Что касается нейтральных высказываний, к ним также отнесены и высказывания, для которых тональность оказалось невозможным определить, поэтому оценить качество измерения нейтральных высказываний не представляется возможным.

Закключение. Социальные сети (Twitter, Facebook и Instagram) стали такими же полноценными источниками информации, как и цифровые датчики и системы, генерирующие данные в умных городах в режиме реального времени. В [8] пользователей социальных сетей сравнивают с социальными датчиками, которые передают боль-

шие и сложные наборы данных с целью поддержки процесса принятия решений. И для того, чтобы эти данные превратились в ценную информацию, необходимую для управления в самых разных сферах умных городов, включая транспорт, потребление энергии, мониторинг окружающей среды, жилищное хозяйство и т. д., они должны быть обработаны и правильно интерпретированы.

В этой работе мы исследовали, как данные социальной сети Фейсбук могут использоваться для измерения общественных настроений. Мы оценили точность и полноту определения полярности в высказываниях горожан в социальных сетях с помощью базовых библиотек Python, и пришли к выводу, что эти технологии недостаточно точны для того, чтобы информация, полученная на их основе, могла использоваться для поддержки принятия решений. В наших будущих работах мы планируем исследовать новые технологии сентимент-анализа, включая аспектно-ориентированный сентимент-анализ и тематическое моделирование.

Благодарности. Работа поддержана грантом ПЦФ BR10965311 «Разработка интеллектуальных информационно-телекоммуникационных систем для городской инфраструктуры: транспорт, экология, энергетика и аналитика данных в концепте Smart City».

ЛИТЕРАТУРА

- 1 Ravi K., Ravi V. A survey on opinion mining and sentiment analysis: tasks, approaches and applications // Knowledge-based systems. – 2015. – Vol. 89. – P. 14-46.
- 2 Malathi M. et al. A Systematic Review on Opinion Mining and Sentiment Analysis // International Journal of Data Structures. – 2020. – Vol. 6. – No. 2. – P. 21-29.
- 3 Medhat W., Hassan A., Korashy H. Sentiment analysis algorithms and applications: A survey // Ain Shams engineering journal. – 2014. – Vol. 5. – No. 4. – P. 1093-1113.
- 4 García B. et al. A survey of the selenium ecosystem // Electronics. – 2020. – Vol. 9. – No. 7. – P. 1067.
- 5 Gujjar J. P., HR P. K. Sentiment Analysis: Textblob For Decision Making // International Journal of Scientific Research & Engineering Trends. – 2021. – Vol. 7. – P. 1097-1099.
- 6 Moskin N. D. et al. Comparative Analysis of Texts by FM Dostoevsky and Other Authors Using Decision Trees // 2021 IEEE International Conference on Smart Information Systems and Technologies (SIST). – IEEE, 2021. – P. 1-5.
- 7 Rogers A. et al. RuSentiment: An enriched sentiment analysis dataset for social media in Russian // Proceedings of the 27th international conference on computational linguistics. – 2018. – P. 755-763.
- 8 Al Nuaimi A. et al. Social media analytics for sentiment analysis and event detection in smart cities // Proc. 4th Int. Conf. Natural Lang. Comput.(NATL). – 2018. – P. 57-64.

А. Б. НУГУМАНОВА, Д. Ж. АХМЕД-ЗАКИ², Е. М. БАЙБУРИН¹, К. С. АПАЕВ³

¹ Сәрсен Аманжолов атындағы Шығыс Қазақстан университеті,
Өскемен, Қазақстан

² Астана ІТ университеті, Нұр-Сұлтан, Қазақстан

³ Дәулет Серікбаев атындағы Шығыс Қазақстан университеті, Өскемен, Қазақстан

FACEBOOK-та ПАЙДАЛАНУШЫЛАРДЫҢ ПІКІРЛЕРІНІҢ СЕНТИМЕНТ-ТАЛДАУЫ: TEXTBLOB ЖӘНЕ DOSTOEVSKY КІТАПХАНАЛАРЫН САЛЫСТЫРУ

Бұл мақала екі Python кітапханасының TextBlob және Достоевскийдің орыс тіліндегі пікірлердің сезімін анықтауға арналған жұмысын қарастырады. TextBlob кітапханасы орыс тіліндегі мәтіндердің сезімдік талдауына тікелей қолдау көрсетпейді, бірақ жасалған ағылшын тіліндегі аудармалар негізінде осы мәтіндердің полярлығын бағалау үшін мұндай мәтіндерді Google Translate аудармашысының енгізуіне жіберу мүмкіндігі бар. Достоевскийдің екінші кітапханасы орыс тіліндегі мәтіндердің сезімдік талдауы үшін арнайы құрылған және 30 мыңнан астам қолмен аннотацияланған хабарламаларды қамтитын ең үлкен орыс тілді RuSentiment сезім корпусында оқытылады. Осы кітапханаларды салыстыру үшін бұл мақалада Нұр-Сұлтан қаласының тұрғындарының елорданың өмір сүру сапасы, оның ішінде қалалық қызметтер мен тұрғын үй-коммуналдық сектордың қызметі, қызмет көрсету нысандарының жұмысы, білім беру, қоғамдық Facebook топтарында жарияланған денсаулық сақтау, мәдениет және т.б. Жүргізілген көңіл-күйді талдау нәтижелері жалпы қабылданған дәлдік, толықтық және F-өлшеу көрсеткіштері арқылы бағаланады және Достоевский кітапханасының шамалы артықшылығын көрсетеді.

Түйін сөздер: сентимент-талдау, табиғи тілді өңдеу, әлеуметтік медиа талдауы, ақылды қала.

A. B. NUGUMANOVA, D. ZH. AKHMED-ZAKP, E. M. BAYBURIN¹, K. S. APAEV³

¹Sarsen Amanzholov East Kazakhstan University, Ust-Kamenogorsk, Kazakhstan

² Astana IT University, Nur-Sultan, Kazakhstan

³Daulet Serikbayev East Kazakhstan University, Ust-Kamenogorsk, Kazakhstan

SENTIMENT ANALYSIS OF USERS REVIEWS IN FACEBOOK: COMPARISON OF TEXTBLOB AND DOSTOEVSKY LIBRARIES

This article examines the work of two Python libraries TextBlob and Dostoevsky for determining the sentiment of reviews in Russian. The TextBlob library does not directly provide sentiment analysis of texts in Russian, but it has the ability to transfer such texts to the input of the Google Translate translator; in order to then evaluate the polarity of these texts based on the generated English translations. The second Dostoevsky library was created specifically for sentiment analysis of texts in Russian and is trained on the largest Russian-language sentiment corpus RuSentiment, which contains more than 30 thousand manually annotated messages. To compare these libraries, this article uses real feedback from residents of the city of Nur-Sultan about the quality of life in the capital, including about the activities of city services and the housing and communal sector, the work of service facilities, education, health care, culture, etc., posted on public Facebook groups. The results of the sentiment analysis carried out are evaluated using generally accepted metrics of accuracy, completeness and F-measure and demonstrate a slight superiority of the Dostoevsky library.

Keywords: sentiment analysis, natural language processing, social media analysis, smart city.