DOI 10.34185/1562-9945-6-155-2024-05
УДК 004.94

П.В. Герасимчук, Т.А. Ліхоузова, Ю.О. Олійник

ЗАСТОСУВАННЯ МЕТОДІВ МАШИННОГО НАВЧАННЯ ДЛЯ АНАЛІЗУ ПОВІДОМЛЕНЬ В TWITTER

Анотація. В сучасному світі соціальні мережі стали невід'ємною частиною життя багатьох людей. Вони є зручним та доступним засобом повсякденного спілкування. Однак з розвитком соціальних мереж з'явилися й проблеми, однією з яких є кібербулінг. В роботі виконано попередню обробку текстових повідомлень, розглянуто різні методи бінарної класифікації текстів та застосовано їх для визначення кібербулінгу в повідомленнях Twitter. Проведено дослідження ефективності кожного з методів та виконано їх порівняльний аналіз. Програмне забезпечення реалізоване мовою Python із застосуванням бібліотек spaCy, gensim, scikit-learn, pandas, numpy, matplotlib.
Ключові слова: класифікація текстів, обробка природної мови, соціальні мережі, кібербулінг, бінарна класифікація, оцінка моделі, word2vec, Bag-of-Words, TF-IDF.

Булінг в інтернеті, зокрема в соціальних мережах, є серйозною загрозою для ментального здоров'я користувачів. Агресивні, образливі або принизливі повідомлення можуть викликати стрес, тривогу, депресію або інші психічні розлади. Через це виявлення та запобігання кібербулінгу є пріоритетним завданням для організацій, що розробляють комунікаційні платформи.

Метою дослідження є пошук моделей для класифікації повідомлень як таких, що містять ознаки булінгу або ні (бінарна класифікація).

Матеріали та методи. Автоматичне визначення булінгу у твітах за допомогою методів машинного навчання є важливим кроком у боротьбі з цією проблемою. Ефективні методи класифікації допоможуть швидко ідентифікувати образливі повідомлення й оперативно вжити відповідні заходи для їх усунення. Це дозволить захистити користувачів від негативного впливу й створити безпечне та комфортне середовище для спілкування [1, 2].

Для того, щоб ефективно протидіяти булінгу в мережі, необхідно розробити моделі для бінарної класифікації текстових повідомлень. Вхідними даними для цієї задачі є самі повідомлення, вихідними - позитивний чи негативний висновок щодо наявності ознак кібербулінгу в конкретному повідомленні.

Для роботи було обрано датасет [3]. Кількість унікальних записів у датасеті – 46017, 39747 твітів з ознаками булінгу і 7945 нейтральних повідомлень. Програмне забезпечення реалізоване мовою Python [4] із застосуванням бібліотек spaCy [5], gensim [6], scikit-learn [7], pandas, numpy, matplotlib.

Першим етапом попередньої обробки даних є очищення текстів від html-тегів, гіперпосилань, імен користувачів, хештегів, небуквених символів та приведення літер до нижнього регістру. Для цього було використано регулярні вирази. Після цього було проведено токенізацію, лематизацію та очищення від стоп слів за допомогою засобів бібліотеки spaCy.

Для застосування класифікаційних моделей було утворено 3 набори вхідних даних: матрицю Bag-of-Words, TF-IDF матрицю та матрицю word2vec. Кожен із наборів було поділено на навчальний та тестовий у співвідношенні 80% та 20% відповідно. Матриця Bag-of-Words містить кількість входжень кожного слова у кожен із документів, TF-IDF матриця – TF-IDF метрики слів у кожному документі, а матриця word2vec – векторне представлення кожного із документів.

Для класифікації текстових повідомлень було використано наступні методи:

- логістична регресія (LogisticRegression бібліотеки scikit-learn);
- метод k найближчих сусідів (KNearestNeighbors бібліотеки scikit-learn);
- випадковий ліс (RandomForestClassifier бібліотеки scikit-learn);
- метод опорних векторів (SVC бібліотеки scikit-learn);
- наївний баєсівський класифікатор (MultinomialNB бібліотеки scikit-learn).

Результати та основний матеріал дослідження. Для оцінки ефективності класифікації текстів моделями було використано показники влучності(precision), повноти (recall), f1-метрики (f1-score) та точності (accuracy), а також матриці невідповідності.

Логістична регресія (рисунки 1-3).

	precision	recall	f1-score	support
0	0.62	0.56	0.59	1591
1	0.91	0.93	0.92	7440
accuracy			0.86	9031
macro avg	0.77	0.74	0.75	9031
weighted avg	0.86	0.86	0.86	9031

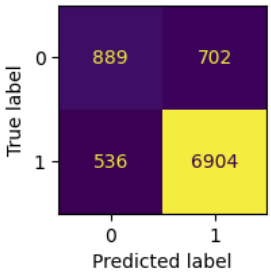

Рисунок 1 – LogisticRegression, вхідні дані – матриця Bag-of-words

	precision	recall	f1-score	support
0	0.68	0.43	0.52	1591
1	0.89	0.96	0.92	7440
accuracy			0.86	9031
macro avg	0.78	0.69	0.72	9031
weighted avg	0.85	0.86	0.85	9031

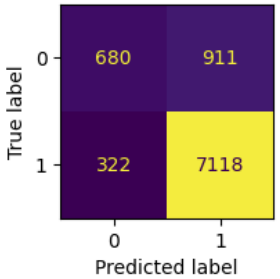

Рисунок 2 – LogisticRegression, вхідні дані – матриця TF-IDF

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

	precision	recall	f1-score	support
0	0.63	0.28	0.39	1591
1	0.86	0.97	0.91	7440
accuracy			0.84	9031
macro avg	0.75	0.62	0.65	9031
weighted avg	0.82	0.84	0.82	9031

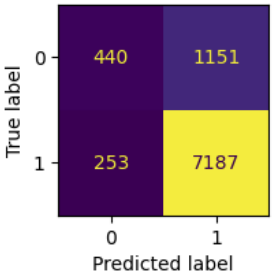

Рисунок 3 – LogisticRegression, вхідні дані – матриця word2vec

Метод k найближчих сусідів (рисунки 4-6). Найвища точність досягнута при k = 55.

	precision	recall	f1-score	support
0	0.51	0.32	0.39	1591
1	0.86	0.93	0.90	7440
accuracy			0.83	9031
macro avg	0.69	0.63	0.64	9031
weighted avg	0.80	0.83	0.81	9031

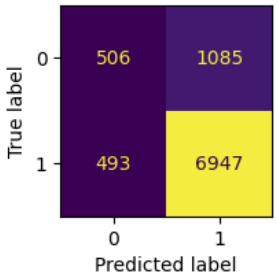

Рисунок 4 – KNearestNeighbors, вхідні дані – матриця Bag-of-words

	precision	recall	f1-score	support
0	0.71	0.29	0.41	1591
1	0.87	0.98	0.92	7440
accuracy			0.85	9031
macro avg	0.79	0.63	0.66	9031
weighted avg	0.84	0.85	0.83	9031

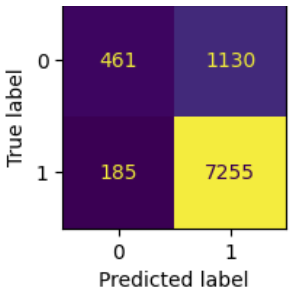

Рисунок 5 – KNearestNeighbors, вхідні дані – матриця TF-IDF

	precision	recall	f1-score	support
0	0.67	0.22	0.33	1591
1	0.85	0.98	0.91	7440
accuracy			0.84	9031
macro avg	0.76	0.60	0.62	9031
weighted avg	0.82	0.84	0.81	9031

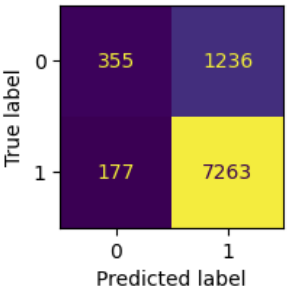

Рисунок 6 – KNearestNeighbors, вхідні дані – матриця word2vec

Метод випадкового лісу (рисунки 7-9).

	precision	recall	f1-score	support
0	0.74	0.30	0.43	1591
1	0.87	0.98	0.92	7440
accuracy			0.86	9031
macro avg	0.81	0.64	0.68	9031
weighted avg	0.85	0.86	0.83	9031

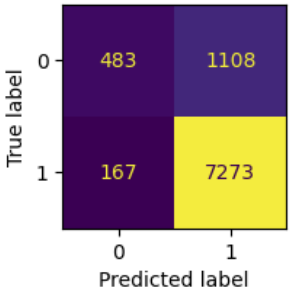

Рисунок 7 – RandomForestClassifier, вхідні дані – матриця Bag-of-words

	precision	recall	f1-score	support
0	0.76	0.29	0.41	1591
1	0.87	0.98	0.92	7440
accuracy			0.86	9031
macro avg	0.81	0.63	0.67	9031
weighted avg	0.85	0.86	0.83	9031

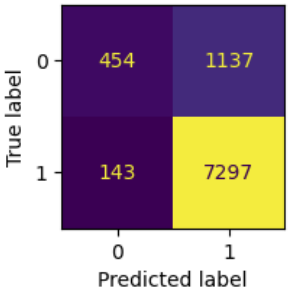

Рисунок 8 – RandomForestClassifier, вхідні дані – матриця TF-IDF

	precision	recall	f1-score	support
0	0.65	0.30	0.41	1591
1	0.87	0.97	0.91	7440
accuracy			0.85	9031
macro avg	0.76	0.63	0.66	9031
weighted avg	0.83	0.85	0.83	9031

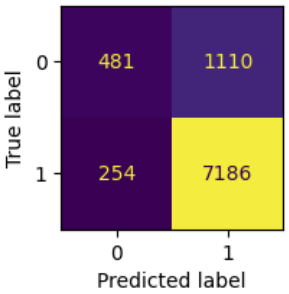

Рисунок 9 – RandomForestClassifier, вхідні дані – матриця word2vec

Метод опорних векторів (рисунки 10-12). Для класифікації було обрано модель LinearSVC через велику розмірність вхідних наборів даних.

	precision	recall	f1-score	support
0	0.57	0.57	0.57	1591
1	0.91	0.91	0.91	7440
accuracy			0.85	9031
macro avg	0.74	0.74	0.74	9031
weighted avg	0.85	0.85	0.85	9031

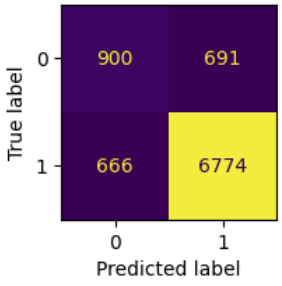

Рисунок 10 – LinearSVC, вхідні дані – матриця Bag-of-words

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

	precision	recall	f1-score	support
0	0.62	0.52	0.57	1591
1	0.90	0.93	0.92	7440
accuracy			0.86	9031
macro avg	0.76	0.73	0.74	9031
weighted avg	0.85	0.86	0.86	9031

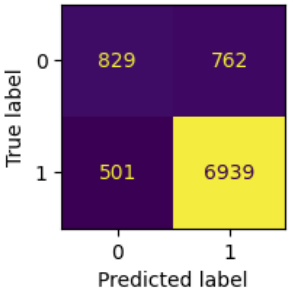

Рисунок 11 – LinearSVC, вхідні дані – матриця TF-IDF

	precision	recall	f1-score	support
0	0.64	0.22	0.33	1591
1	0.85	0.97	0.91	7440
accuracy			0.84	9031
macro avg	0.75	0.60	0.62	9031
weighted avg	0.82	0.84	0.81	9031

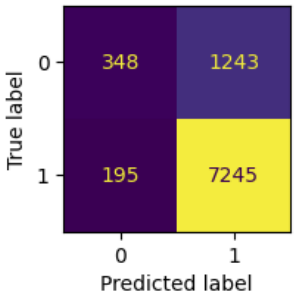

Рисунок 12 – LinearSVC, вхідні дані – матриця word2vec

Наївний баєсівський класифікатор (рисунки 13-15). Для класифікації було обрано модель MultinomialNB через поліноміальну природу розподілу частот слів у текстах.

	precision	recall	f1-score	support
0	0.66	0.33	0.44	1591
1	0.87	0.96	0.91	7440
accuracy			0.85	9031
macro avg	0.76	0.65	0.68	9031
weighted avg	0.83	0.85	0.83	9031

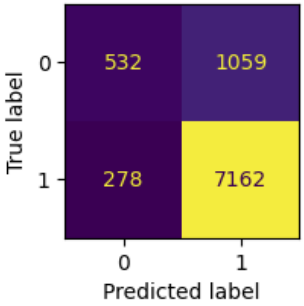

Рисунок 13 – MultinomialNB, вхідні дані – матриця Bag-of-words

	precision	recall	f1-score	support
0	0.90	0.12	0.20	1591
1	0.84	1.00	0.91	7440
accuracy			0.84	9031
macro avg	0.87	0.56	0.56	9031
weighted avg	0.85	0.84	0.79	9031

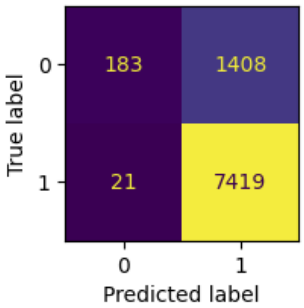

Рисунок 14 – MultinomialNB, вхідні дані – матриця TF-IDF

Порівняння результатів моделей.

За результатами тестування моделей можна зробити такі висновки:

- ефективність класифікації для усіх моделей мало залежить від вибору типу вхідних даних (Bag-of-Words, TF-IDF, word2vec);

- логістична регресія та метод опорних векторів показали трохи кращу ефективність, ніж інші моделі;
- наївний баєсівський класифікатор показав відносно низьку ефективність при тренуванні на TF-IDF матриці;
- метод k найближчих сусідів показав відносно низьку ефективність при тренуванні на матриці Bag-of-Words.

Враховуючи те, що найкращу ефективність показали моделі на основі логістичної регресії та методу опорних векторів та швидкодію логістичної регресії, найдоцільнішим для розв'язання поставленого завдання є використання саме класифікатора LogisticRegression на матриці Bag-of-Words.

Висновки. За результатами порівняльного аналізу ефективності моделей виокремлено логістичну регресію на вхідних даних Bag-of-Words як найефективнішу модель для задачі бінарної класифікації текстових повідомлень із обраного набору.

Отримані в ході дослідження результати можуть бути використані для розробки систем автоматичного виявлення ознак кібербулінгу в повідомленнях користувачів соціальних мереж та оперативного вживання відповідних заходів.

ЛІТЕРАТУРА / REFERENCES

1. Писаренко О. А. Інтелектуальна система фільтрації коментарів з використанням машинного навчання. – 2019.
2. Іванов О. А. Розробка сервісу для боротьби з кібербулінгом //Автоматизація та комп'ютерно-інтегровані технології у виробництві та освіті: стан, досягнення, перспективи розвитку. – С. 298.
3. https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification/
4. Документація мови програмування Python. [Електронний ресурс] – Режим доступу до ресурсу: https://docs.python.org/3/
5. Документація бібліотеки spaCy. [Електронний ресурс] – Режим доступу до ресурсу: https://spacy.io/usage
6. Документація бібліотеки Gensim. [Електронний ресурс] – Режим доступу до ресурсу: https://radimrehurek.com/gensim/

Документація бібліотеки scikit-learn. [Електронний ресурс] – Режим доступу до ресурсу: https://scikit-learn.org/stable/

Received 04.09.2024.
Accepted 09.09.2024.

Application of machine learning methods for analysis of Twitter messages

The paper investigates the problem of binary classification of text messages for the presence of bullying. Bullying on the Internet, in particular in social networks, is a serious threat to the mental health of users. Aggressive, offensive or humiliating messages can cause stress, anxiety, depression or other mental disorders. Because of this, identifying and preventing cyberbullying is a priority for organizations developing communication platforms.

A dataset with Twitter messages was prepared and pre-processed, including cleaning, tokenization, and lemmatization. 3 sets of input data for classification models were created: Bag-of-Words, TF-IDF matrix, word2vec matrix.

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

Models based on various machine learning methods were built and tested: logistic regression, k nearest neighbors, random forest, support vector, naive Bayesian classifier methods on each of the input data sets.

Based on the results of testing the models, a comparative analysis of their effectiveness was carried out, logistic regression on Bag-of-Words input data was singled out as the most effective model for the task of binary classification of text messages from the selected set.

The results obtained in the course of the study can be used for the development of systems for automatic detection of signs of cyberbullying in the messages of users of social networks and the prompt use of appropriate measures.

Keywords: text classification, natural language processing, social networks, cyberbullying, binary classification, model evaluation, word2vec, Bag-of-Words, TF-IDF.

Герасимчук Павло Вікторович – студент кафедри інформатики та програмної інженерії, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

Ліхоузова Тетяна Анатоліївна – к.т.н., доцент кафедри інформатики та програмної інженерії, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

Олійник Юрій Олександрович – к.т.н., доцент кафедри інформатики та програмної інженерії, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

Herasymchuk Pavlo – student, Department of Informatics and Software Engineering National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute».

Likhouzova Tetiana – PhD, associate professor, Department of Informatics and Software Engineering National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute».

Oliinyk Yurii – PhD, associate professor, Department of Informatics and Software Engineering National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute».