ISSN 2224-087X. Електроніка та інформаційні технології. 2023. Випуск 22. С. 34–45
Electronics and information technologies. 2023. Issue 22. P. 34–45

УДК 004.932 DOI: https://doi.org/10.30970/eli.22.4

ВИЯВЛЕННЯ АГРЕСИВНОЇ РИТОРИКИ В ТЕКСТІ З ВИКОРИСТАННЯМ АЛГОРИТМІВ МАШИННОГО НАВЧАННЯ

М. Притула, І. Оленич

Львівський національний університет імені Івана Франка,
вул. Драгоманова, 50, 79005 Львів, Україна
marianna.prytula@lnu.edu.ua, igor.olenych@lnu.edu.ua

У роботі реалізовано моделі класифікації текстової інформації, яка містить агресивну лексику та емоційні вирази. За допомогою алгоритмів машинного / глибокого навчання проведено аналіз новинних повідомлень в електронних ЗМІ та соціальних мережах українською та російською мовами, пов'язаних з повномасштабним російським вторгненням на територію України. Досліджено ефективність розпізнавання агресивної риторики у текстових повідомленнях за допомогою наївного класифікатора Байєса, методів опорних векторів, *k*-найближчих сусідів, випадкового лісу, логістичної регресії, рекурентних нейронних мереж з архітектурою LSTM і Bidirectional LSTM. Встановлено, що збалансованість навчальної вибірки текстових даних суттєво впливає на точність класифікації. Виявлено лінійну кореляцію між фейковими новинами і агресивною риторикою в інформаційних повідомленнях.

Ключові слова: комп'ютерний аналіз тексту, сентимент-аналіз, глибоке навчання, машинне навчання з учителем.

1. Вступ

Автоматична обробка новин та визначення їх емоційного забарвлення відіграють важливу роль у сучасному медіапросторі. Через постійне зростання обсягу інформації, яка розповсюджується у засобах масової інформації (ЗМІ) та соціальних мережах, необхідно мати ефективні інструменти для швидкого та об'єктивного аналізу новинних матеріалів у реальному часі без підкреслення певних думок чи позицій. Автоматична обробка новин дає змогу зібрати, класифікувати та аналізувати великі обсяги даних для виявлення фейків, зміни настроїв, агресивної риторики та провокаційних закликів до несанкціонованих дій, які можуть зашкодити безпеці країни [1, 2]. Важливість протидії дезінформації та забезпечення інформаційної безпеки значно зросли з початком повномасштабного російського вторгнення на територію України.

З іншого боку, серйозну небезпеку для користувачів соціальних мереж становить кібербулінг, оскільки його жертви стають сприйнятливими до багатьох негативних факторів, таких як низька самооцінка, страх, тривога, гнів тощо [3]. Тому аналіз емоційного забарвлення текстового контенту та виявлення віртуальної агресії є серед найважливіших завдань обробки природної мови (Natural Language Processing, NLP). Застосування комп'ютерного сентимент-аналізу дає змогу виявити інтенсивність та тип емоцій текстових повідомлень [4, 5].

Існує низка різноманітних методів та алгоритмів для виявлення агресії в повідомленнях та аналізу емоційного забарвлення тексту загалом, які виражають основні підходи NLP:

1. Використання моделей які навчаються на позначених даних, що містять текст із вказаними емоційними категоріями, включаючи агресію. Моделі використовують такі методи, як наївний класифікатор Баєса, метод опорних векторів (support vector machines, SVM), метод *k*-найближчих сусідів (*k*-nearest neighbor, KNN) та логістичної регресії, щоб навчитися розпізнавати емоційні ознаки тексту, пов'язані з агресією [6–8].

2. Глибоке навчання: Глибокі нейронні мережі можуть бути використані для аналізу тексту та виявлення агресії. Наприклад, рекурентні нейронні мережі (recurrent neural networks, RNN) або моделі-трансформери, які ґрунтуються на концепції перенесеного навчання (Transfer Learning) [9], можуть бути навчені на великому обсязі тексту з позначеними емоційними категоріями, включаючи агресію. Ці моделі можуть автоматично виявляти контекстуальні зв'язки та емоційні ознаки, пов'язані з агресією [4].

3. Підхід на основі словників і лексичних ресурсів базується на використанні попередньо побудованих сентимент-словників, що містять емоційно забарвлені слова або фрази. Аналіз текстової інформації здійснюється шляхом виявлення цих слів або фраз. Залежно від їх кількості або контексту можна визначити емоційну тональність тексту, зокрема і рівень агресії [10–12]. Для врахування суб’єктивних факторів, які притаманні вираженню людських емоцій, часто використовують методи та підходи нечіткої логіки для формалізації здатності людини до наближених міркувань [13, 14]. Проте, вагомою проблемою методів, заснованих на словниках і правилах, є трудомісткість процесу створення словника. Значною мірою це стосується сентимент-аналізу україномовних текстів, оскільки словників тональності українською мовою є обмежена кількість.

Важливою частиною NLP є попередня обробка даних для видалення розділових знаків і стоп-слів та маркування даних для подальшого аналізу [15]. Точність моделей машинного навчання загалом та у галузі NLP зокрема суттєво залежить від кількісних та якісних характеристик наборів даних, які використовуються для навчання моделей. Основна частина доступних наборів даних придатні для тренування лише англомовних моделей NLP. Тому в роботі значна увага була приділена підготовці та аналізу текстових наборів даних українською та російською мовами, що може сприяти нейтралізації цифрових загроз в умовах російської агресії. Мета дослідження полягала у розробці програмних модулів аналітичної системи для виявлення агресивного забарвлення тексту на основі виявлення грубих, образливих слів, загроз, розпалювання конфліктів тощо з використанням різних методів статистичного аналізу текстів, машинного та глибокого навчання.

2. Методи та засоби реалізації

Реалізація запропонованих моделей класифікації інформаційних потоків здійснювалася відповідно до алгоритму визначення агресивного забарвлення новинних повідомлень, який схематично зображений на рис. 1. Розроблена аналітична система дає змогу завантажити новинні матеріали українською та російською мовою, одержані з різних джерел: публікації в електронних ЗМІ, повідомлення в соціальних мережах тощо. На початковому етапі текст розділяється на прості одиниці (tokens), тобто осмислені групи символів, що відповідають заданим шаблонам (наприклад, слова) [12].

У результаті було отримано масив слів, який після процедури стемінгу та виключення розділових знаків і стоп-слів використовувався для подальшого аналізу. Векторизація тексту здійснювалась з використанням статистичного показника Term Frequency – Inverse Document Frequency (TF–IDF), який відображає важливість слова у тексті, а не тільки частоту його появи [16].

Рис. 1. Алгоритм виявлення агресивної риторики у текстових повідомленнях.

Наступним кроком алгоритму виявлення агресивної риторики у новинних матеріалах є власне реалізація методів машинного / глибокого навчання, використовуючи тестові набори даних українською та російською мовами. У роботі було застосовано наступні методи: SVM, KNN, наївний класифікатор Байєса, ансамблевий метод Random forest, логістичну регресію, а також RNN з архітектурою Long short-term memory (LSTM) і Bidirectional LSTM. Фінальним етапом запропонованого алгоритму є оцінка ефективності реалізованих моделей.

Для створення програмних модулів було використано мову програмування Python 3.9 та середовище Jupyter Notebook. Реалізація моделей машинного навчання мовою Python здійснювалась за допомогою бібліотеки Scikit-learn. Використана бібліотека надає широкий спектр інструментів для підготовки даних, зменшення їх

розмірності, класифікації, регресії, кластеризації та інших завдань машинного навчання, в тому числі й алгоритми для оцінки моделей. Scikit-learn містить засоби реалізації багатьох популярних алгоритмів машинного навчання, зокрема й тих, які використовуються для виявлення емоційного забарвлення текстової інформації.

Набір текстових даних із новинами, пов'язаними з повномасштабною російською агресією, був сформований у співпраці з науково-педагогічними працівниками факультету журналістики Львівського національного університету імені Івана Франка у період з червня по грудень 2022 року і складається з понад 4 тис. повідомлень українською та російською мовами. Дані були розділені на два класи: з агресивною та нейтральною риторикою. Варто зазначити, що обсяг класів суттєво відрізнявся: було класифіковано 2962 емоційно нейтральних новин і 1330 повідомлень з агресивною риторикою. Крім того, новинні матеріали були класифіковані за критерієм дезінформації, що можна використати як додатковий параметр для сентимент-аналізу текстів. Співвідношення між обсягом класів за цим критерієм було приблизно однаковим.

3. Результати та їх аналіз

Класифіковані текстові дані були використані для тренування запропонованих у роботі моделей машинного / глибокого навчання та оцінки ефективності розпізнавання агресивної риторики розробленою аналітичною системою. Для цього набір класифікованих новинних матеріалів був розділений на навчальну та тестову вибірки, обсяг яких складав 80 і 20 % від загальної кількості текстових повідомлень, відповідно. Навчені моделі використовувались для класифікації тестової вибірки новинних повідомлень, у результаті чого були одержані звіти про ефективність моделей, які зображені на рис. 2 – рис. 8. Тут клас 0 відповідає повідомленням з агресивною риторикою, а клас 1 – нейтральним повідомленням.

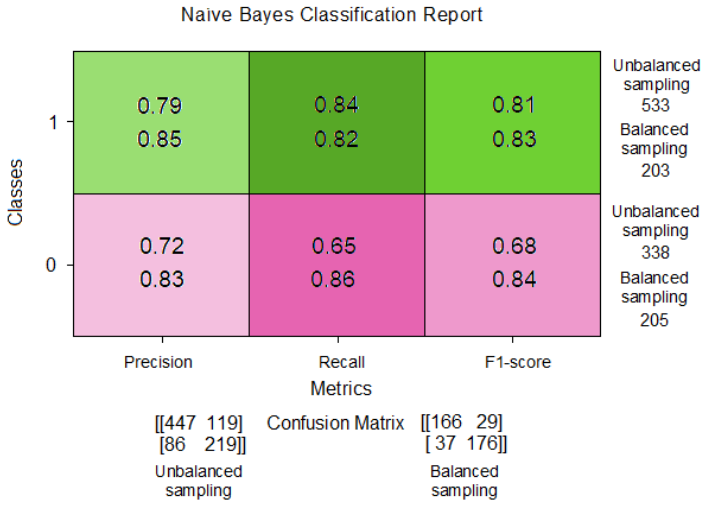

Рис. 2. Звіт для оцінки ефективності виявлення агресивної риторики з використанням наївного класифікатора Баєса.

Зазвичай результати розв'язання задачі бінарної класифікації позначають як позитивні або негативні, тобто правильно або неправильно класифіковані екземпляри. Ці розв'язки візуально представляють у вигляді матриці невідповідностей (Confusion Matrix) [17]. На рис. 2 – рис. 8 також зображені матриці невідповідностей для кожної з розглянутих моделей. На основі матриці невідповідностей та її значень обчислюються різноманітні метрики ефективності класифікації.

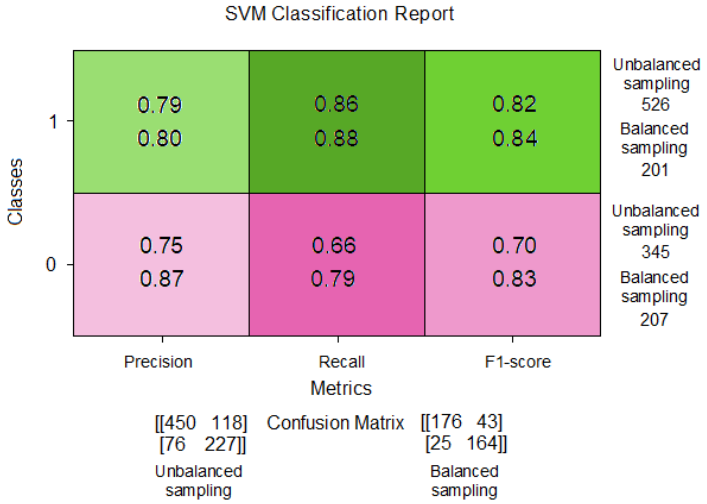

Рис. 3. Звіт для оцінки ефективності виявлення агресивної риторики методом SVM.

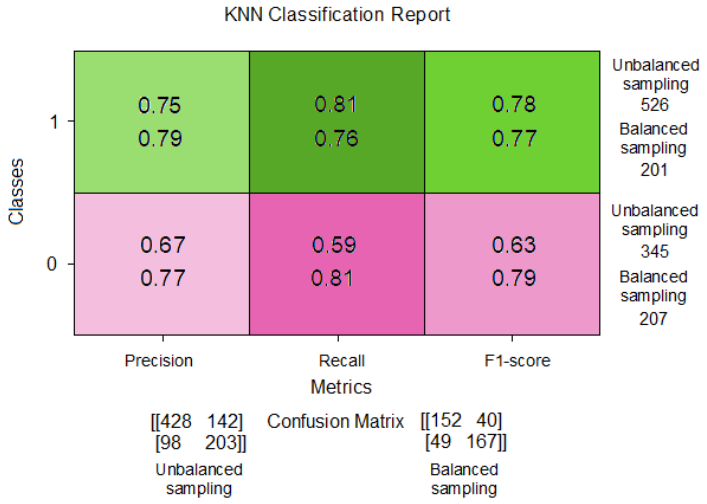

Рис. 4. Звіт для оцінки ефективності виявлення агресивної риторики методом KNN.

Оцінювання ефективності запропонованих моделей здійснювалось засобами бібліотеки Scikit-learn за допомогою наступних метрик:

- accuracy, що вимірює відношення правильно класифікованих екземплярів до загальної кількості екземплярів:

$$accuracy = (TP + TN) / (TP + TN + FP + FN),$$

де TP і FP – кількість правильно і неправильно класифікованих екземплярів позитивного класу, а TN і FN – кількість правильно і неправильно класифікованих екземплярів негативного класу, відповідно;

- precision, яка визначає відношення правильно класифікованих екземплярів позитивного класу до загальної кількості екземплярів позитивного класу:

$$precision = TP / (TP + FP);$$

- recall, що визначає частку правильно класифікованих екземплярів позитивного класу до загальної кількості дійсно позитивних екземплярів:

$$recall = TP / (TP + FN);$$

- F1-score, як міра узгодженості між влучністю (precision) і повнотою (recall), що демонструє наскільки багато екземплярів прогнозується моделлю правильно і скільки істинних екземплярів модель не пропустить, тобто дає змогу одержати узагальнену оцінку ефективності моделі:

$$F1\text{-}score = 2 * (precision * recall) / (precision + recall).$$

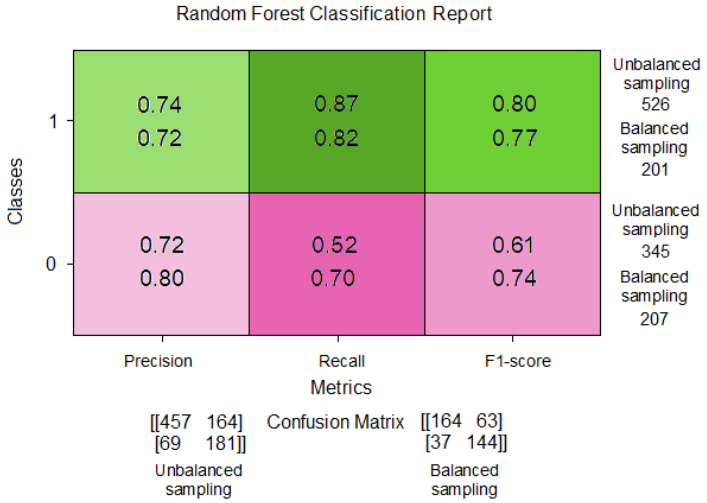

Рис. 5. Звіт для оцінки ефективності виявлення агресивної риторики методом Random Forest.

Аналіз одержаних результатів дає змогу зробити висновок, що класичні методи машинного навчання розпізнають агресивну риторику у новинних повідомленнях українською та російською мовами із задовільною точністю. Проте надзвичайно важливе значення має якість класифікованих даних і збалансованість вибірки, яка використовувалась для тренування моделей. У разі, коли кількість нейтральних новин значно перевищує кількість агресивних повідомлень, значення precision, recall та F1-

score для такої вибірки суттєво відрізняються: точність розпізнавання агресивної риторики є меншою. Підвищення точності сентимент-аналізу було досягнуто збалансуванням навчальної вибірки шляхом зменшення кількості нейтральних повідомлень внаслідок випадкового їх виключення з набору даних.

Іншим способом підвищити точність розпізнавання агресивної риторики було застосування RNN з архітектурою LSTM і Bidirectional LSTM (див. рис.7 і рис.8, відповідно). Зазначені моделі глибокого навчання доволі успішно використовують в галузі NLP для розпізнавання природньої мови, покращення машинного перегляду тощо.

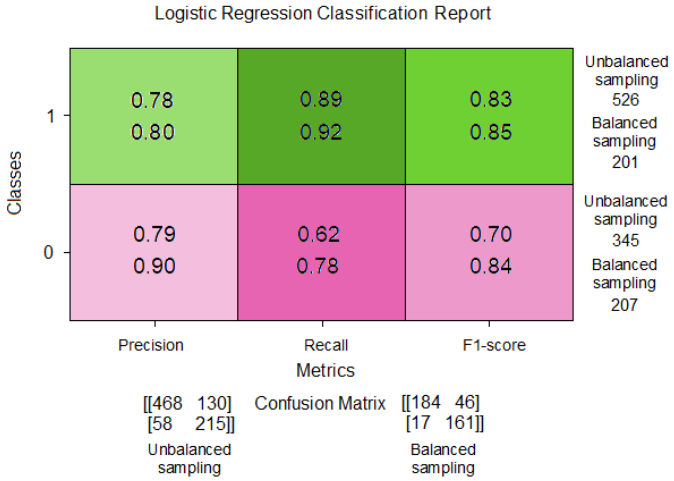

Рис. 6. Звіт для оцінки ефективності виявлення агресивної риторики з використанням алгоритму Logistic Regression.

Для побудови LSTM-моделі класифікації тексту використовувалась бібліотека Keras. Модель LSTM складається з Embedding-шару для перетворення індексів слів у вектори фіксованої довжини; шару LSTM, який приймає послідовності, вбудовує їх і передає до повнозв'язаного Dense-шару з sigmoid-активацією для класифікації. Модель компілюється з оптимізатором "adam" та функцією втрати "mean_squared_error".

Модель для класифікації тексту з використанням Bidirectional LSTM-шару дає змогу шару LSTM працювати в двох напрямках: вперед і назад по вхідній послідовності. Аналізуючи текст, модель може одночасно враховувати слова, які з'явилися раніше і пізніше в тексті. Це покращує здатність моделі розуміти контекст і залежності в тексті.

Варто зазначити, що моделі LSTM і Bidirectional LSTM вимагають значний обсяг навчальної вибірки текстових даних. Імовірно, саме через недостатньо великі набори даних, які використовувались у роботі для аналізу текстових повідомлень, зазначені моделі глибокого навчання не демонстрували помітного підвищення ефективності розпізнавання агресивної риторики за метриками accuracy, precision, F1-score порівняно з іншими методами машинного навчання. Спостерігалося тільки незначне збільшення повноти (recall) класифікації.

Результати оцінювання точності виявлення агресії у текстових повідомленнях розглянутими моделями за звичайною метрикою accuracy наведені у табл. 1. В усіх випадках спостерігалося збільшення точності класифікації внаслідок збалансування навчальної вибірки. За винятком методів KNN і Random Forest, які демонструють найменші значення параметру accuracy, точність запропонованих моделей машинного / глибокого навчання, натренованих на вибірці новинних повідомлень українською та російською мовами, становить 81–85 %.

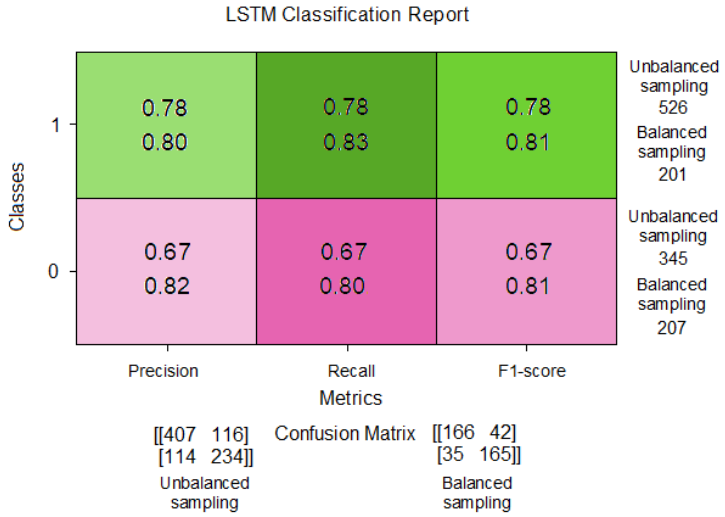

Рис. 7. Звіт для оцінки ефективності виявлення агресивної риторики з використанням LSTM-нейромережі.

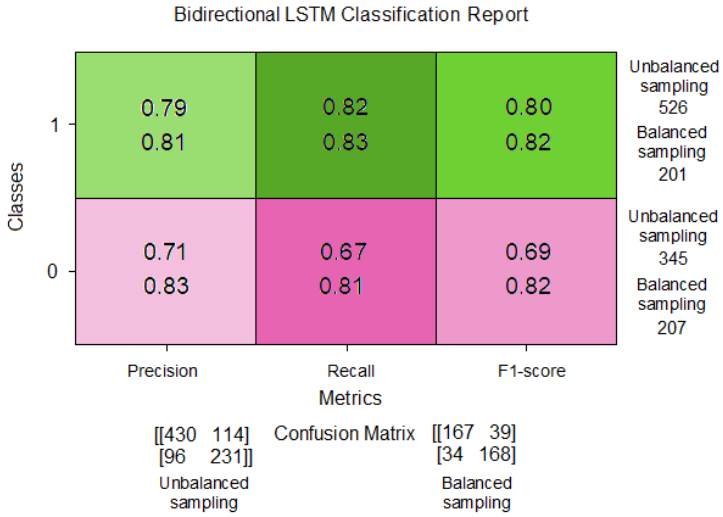

Рис. 8. Звіт для оцінки ефективності виявлення агресивної риторики з використанням Bidirectional LSTM-нейромережі.

Подальше збільшення точності розпізнавання агресивної риторики у текстових повідомленнях потребує розширення збалансованого набору тренувальних даних та/або реалізацію нових підходів і методів сентимент-аналізу. Одним з потенційних способів підвищення ефективності моделей є врахування додаткових ознак для класифікації. Зокрема, кореляційний аналіз набору текстових даних виявив суттєвий лінійний статистичний зв'язок між фейковими новинами і агресивною риторикою: коефіцієнт Пірсона дорівнював 0,6. Ба більше, аналіз російськомовного сегменту текстових повідомлень виявив більш тісний зв'язок між дезінформацією та агресією. У цьому випадку коефіцієнт Пірсона становив 0,81.

Таблиця 1. Точність (accuracy) розпізнавання агресивної риторики у текстових повідомленнях використаними алгоритмами машинного / глибокого навчання.

Method	Accuracy	
	Unbalanced sampling, %	Balanced sampling, %
Naive Bayes	76.5	83.8
SVM	77.7	83.3
KNN	72.4	78.2
Random Forest	73.2	75.5
Logistic regression	78.4	84.6
LSTM	73.6	81.1
Bidirectional LSTM	75.9	82.1

4. Висновки

Запропонована у роботі аналітична система дає змогу опрацьовувати україномовні та російськомовні тексти та автоматично виявляти агресивну риторику. За допомогою алгоритмів SVM, KNN, наївного класифікатора Байєса, логістичної регресії, методу Random forest та рекурентних нейромереж з архітектурою LSTM і Bidirectional LSTM проведено аналіз новинних повідомлень в електронних ЗМІ та соціальних мережах українською та російською мовами, які пов'язані з повномасштабним російським вторгненням на територію України. Здійснено порівняння ефективності різних моделей машинного / глибокого навчання для класифікації інформаційних матеріалів як нейтральні та агресивно забарвлені повідомлення.

Встановлено, що збалансованість навчальної вибірки суттєво впливає на точність класифікації. За винятком методів KNN і Random Forest, які демонструють найменші значення параметру accuracy, точність запропонованих моделей машинного / глибокого навчання досягає 81–85 %. Виявлено лінійну кореляцію між фейковими новинами і агресивною риторикою інформаційних повідомлень, що можна використати як додаткову ознаку для класифікації новинних матеріалів. Розроблена аналітична система може бути корисною для модерації контенту у соціальних мережах, аналізу публічної думки щодо певних тем, розпізнавання загроз та попередження конфліктів у веб-спільнотах.

Список використаних джерел

1. *Mansour S.* Social Media Analysis of User's Responses to terrorism using sentiment analysis and text mining // Procedia Computer Science. – 2018. – Vol. 140. – P. 95–103.
2. *Hao J., Dai H.* Social Media Content and Sentiment Analysis on Consumer Security Breaches // Journal of Financial Crime. – 2016. – Vol. 23, No 4. – P. 855–869.
3. *Hosseinmardi H., Rafiq R.I., Han R., Lv Q., Mishra S.* Prediction of cyberbullying incidents in a media-based social network // Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), San Francisco, CA, USA. – 2016. DOI: 10.1109/ASONAM.2016.7752233.
4. *Khan A., Baharudin B., Lee L.H., Khan K.* A Review of Machine Learning Algorithms for Text-Documents Classification // Journal of Advances in Information Technology. – 2010. - Vol. 1. – P. 4–20.
5. *Luo X.* Efficient English text classification using selected Machine Learning Techniques // Alexandria Engineering Journal. – 2021. – Vol. 60. – P. 3401–3409.
6. *Gasparetto A., Marcuzzo M., Zangari A., Albarelli A.* A Survey on Text Classification Algorithms: From Text to Predictions // Information. – 2022. – Vol. 13. – P. 83.
7. *Hassan S.U., Ahamed J., Ahmad K.* Analytics of machine learning-based algorithms for text classification // Sustainable Operations and Computers. – 2022. – Vol. 3. – P. 238–248.
8. *Kowsari K., Meimandi K.J., Heidarysafa M., Mendu S., Barnes L., Brown D.* Text Classification Algorithms: A Survey // Information. – 2019. – Vol. 10. – P. 150.
9. *Raza S., Ding C.* Fake news detection based on news content and social contexts: a transformer-based approach // Int. J. Data Sci. Anal. – 2022. – Vol. 13. – P. 335–362.

10. *Chopra F.K., Bhatia R.* Sentiment Analyzing by Dictionary based Approach // International Journal of Computer Applications. – 2016. – Vol. 152, No.5. – P. 32–34.
11. *Оленич І., Притула М., Сінькевич О., Хамар О.* Система автоматичного визначення тональності тексту // Електроніка та інформаційні технології. – 2021. – Випуск 15. – С. 16–23.
12. *Thelwall M., Buckley K., Paltoglou G., Kappas A., Cai D.* Sentiment strength detection in short informal text // Journal of the American Society for Information Science and Technology. – 2010. – No. 61. – P. 2544–2558.
13. *Olenych I., Prytula M., Sinkevych O., Khamar O.* System of Automatic Determination of Ukrainian Text Tone // IEEE 12th International Conference on Electronics and Information Technologies (ELIT). – 2021. – P. 80–83.
14. *Olenych I., Sinkevych O., Salamakha M., Prytula M.* Text Tone Determination using Fuzzy Logic // Applied Computer Systems. – 2021. – Vol. 26 – P. 158–163.
15. *Khan U., Khan S., Rizwan A., Ghada A., Jamjoom M.M., Samee N.A.* 2022. Aggression Detection in Social Media from Textual Data Using Deep Learning Models // Applied Sciences. – 2022. – Vol. 12, No. 10. – P. 5083.
16. *Robertson S.* Understanding Inverse Document Frequency: On Theoretical Arguments for IDF // Journal of Documentation. – 2004. – Vol. 60, No. 5. – P. 503–520.
17. *Vijayarani S., Nithya M.N.* Efficient machine learning classifiers for automatic information classification // Int. J. Mod. Trends Eng. Res. – 2015. – Vol. 2. – P. 685–694.

DETECTION OF AGGRESSIVE RHETORIC IN TEXT USING MACHINE LEARNING ALGORITHMS

M. Prytula, I. Olenych

Ivan Franko National University of Lviv,
50 Drahomanov St., UA–79005 Lviv, Ukraine

marianna.prytula@lnu.edu.ua, igor.olenych@lnu.edu.ua

Automated news processing enables the classification and analysis of large volumes of data to detect fakes, aggressive rhetoric, and provocative calls for unauthorized actions. Cyber aggression has a negative effect, such as harming, threatening, or harassing of person. The importance of countering disinformation and ensuring information security has grown significantly since the beginning of the full-scale Russian invasion of Ukraine. Increasing volumes of news, messages, and tweets require choosing the optimal model of aggression detection in textual data to alleviate its impact on people's lives.

The paper implements models of textual information classification containing aggressive vocabulary and emotional expressions. The dataset with more than 4,000 news in electronic media and social network items in Ukrainian and Russian languages was used in the development process. The development of aggression detection models took place in several stages: data pre-processing (tokenization, stemming, lemmatization, and stop words elimination); text vectorization using TF–IDF; implementation of machine learning and deep learning algorithms; comparisons of models using classification reports and confusion matrices.

The effectiveness of recognizing aggressive rhetoric in text messages using a naive Bayes classifier, support vector methods, k-nearest neighbors, random forest, logistic regression, and recurrent neural networks with LSTM and bidirectional LSTM architecture was studied. The classification models have been evaluated according to the accuracy, precision, recall, and F1-score metrics. It was established that the balance of the training sampling of textual data significantly affects the classification accuracy. More errors were made in the prediction of aggressively labeled text due to the predominance of non-aggressive specimens in the case of the unbalanced dataset. The accuracy of machine / deep learning models trained on balanced data reaches 81-85%. A linear correlation between fake news and aggressive rhetoric in information messages was found, which can be used as an additional feature for the classification of news materials.

Key words: computer text analysis, sentiment analysis, deep learning, supervised machine learning.

Стаття надійшла до редакції 11.05.2023
Прийнята до друку 15.05.2023