**УДК 801:001.89**
**DOI https://doi.org/10.24919/2308-4863/39-2-22**

***Марія МАЛИШЕВА,***
*orcid.org/0000-0002-1910-4833*
*аспірантка кафедри прикладної лінгвістики*
*Одеського національного університету імені І. І. Мечникова*
*(Одеса, Україна) mariiamalysheva@onu.edu.ua*

# СТВОРЕННЯ АЛГОРИТМУ ДЛЯ ВИЗНАЧЕННЯ МАРКЕРІВ ВЕРБАЛЬНОЇ АГРЕСІЇ В ДИСКУРСІ СОЦІАЛЬНИХ МЕРЕЖ: ЛІНГВІСТИЧНИЙ ПІДХІД

*Розвідку присвячено вивченню вербальної агресії в мережевому дискурсі. Вербальна агресія як мовленнєве явище є відносно новим напрямом досліджень мовознавців, тому цей термін досі не має усталеної дефініції. Розглянуто низку підходів сучасних науковців до потрактування поняття вербальної агресії для окреслення цього явища. З'ясовано, що дослідження комунікативного простору соціальних мереж як майданчика для розгортання агресивної мовленнєвої поведінки вимагає залучення автоматизованих методів аналізу даних та інструментів для обробки природної мови (natural language processing, NLP), зокрема штучних нейронних мереж і машинного навчання. Проаналізовано низку існуючих методів для автоматизованого визначення деструктивних настроїв у текстах соціальних мереж і виявлено, що залучення нейронних мереж для пошуку агресії в мережевому дискурсі не надає переконливих результатів. Якість роботи нейронної мережі можливо поліпшити за допомогою створення додаткових правил і фільтрів, утім, проблему не можна вважати розв'язаною. Обґрунтовано доцільність розроблення алгоритму для визначення маркерів вербальної агресії в дискурсі соціальних мереж, заснованого на словниках-масивах даних і правилах. Запропоновано створити низку словників (масивів даних) із маркерами вербальної агресії. Кожен маркер у масиві даних має низку характеристик (атрибутів), які можна використовувати для розв'язання інших завдань, наприклад для кількісного аналізу вживання певних видів інвектив тощо. Шляхом пошуку в масивах даних маркерів агресії алгоритм робить висновок про наявність агресії в заданому тексті та наводить типологію маркерів вербальної агресії. У подальшому буде розроблено нові типи масивів даних та вдосконалено атрибути маркерів, зокрема для визначення рівня агресивності заданого тексту.*

***Ключові слова:*** *алгоритм пошуку агресії, вербальна агресія, маркери вербальної агресії, обробка природної мови, штучні нейронні мережі.*

***Mariia MALYSHEVA,***
*orcid.org/0000-0002-1910-4833*
*Postgraduate Student at the Department of Applied Linguistics*
*Odesa I. I. Mechnikov National University*
*(Odesa, Ukraine) mariiamalysheva@onu.edu.ua*

# CREATING THE ALGORITHM FOR DETECTION OF VERBAL AGGRESSION MARKERS IN THE SOCIAL NETWORK DISCOURSE: A LINGUISTIC APPROACH

*The paper is dedicated to the study of verbal aggression in the social network discourse. Verbal aggression as a speech phenomenon is a relatively new field of research for linguists, thus the term is yet to have an established definition. A number of modern scientific approaches to interpreting the concept of verbal aggression were considered to help define this phenomenon. It has been discovered that the study of the communicative space of social networks as a platform for the development of aggressive speech behavior requires the use of automated data analysis methods and tools for natural language processing (NLP), including artificial neural networks and machine learning. A number of existing methods of the automated determination of destructive moods in social network texts were analyzed and it was understood that the use of neural networks to search for aggression in the social network discourse isn't convenient enough. The quality of the neural network can be improved by creating additional precepts and filters, however, the problem cannot be considered solved. The viability of developing an algorithm for identifying markers of verbal aggression in the social network discourse, based on data dictionaries and rules, was vindicated. It was proposed to create a number of dictionaries (data arrays) with markers of verbal aggression. Each marker in the data set has a number of features (attributes) that can be used to solve other problems as well, for example, to quantitatively analyze the use of certain types of invectives and the like. Scanning data arrays for aggression markers, the algorithm concludes that there is aggression in a given text and provides a typology of verbal aggression markers. New types of data arrays will be developed in the future, and marker attributes will be improved, particularly to determine the level of aggression of a given text.*

***Key words:*** *aggression detection algorithm, verbal aggression, verbal aggression markers, natural language processing, artificial neural networks.*

**Постановка проблеми.** Дослідження комунікативного простору мережі Інтернет вимагає залучення відповідних автоматизованих методів аналізу даних, що змушує дослідників-гуманітаріїв набувати навичок програмування. Розвиток інформаційних технологій надає дослідникам-філологам нові можливості для обробки усних та письмових текстів. Цифрові методи обробки інформації дають змогу опрацьовувати великі обсяги даних і вирішувати завдання, що раніше потребували значних ресурсів. Одним із цих завдань є обробка природної мови (*natural language processing, NLP*). Серед актуальних сфер застосування обробки природної мови виокремлюємо створення аналітичних інструментів для розпізнавання деструктивного контенту, зокрема розроблення системи автоматизованого виявлення маркерів вербальної агресії в комунікативному просторі соціальних мереж як майданчику для розгортання деструктивної поведінки користувачів.

**Аналіз досліджень.** Проблеми обробки природної мови висвітлено в наукових розвідках українських (О. Гирин, Д. Дарчук, Н. Чейлитко) та зарубіжних (Д. Гордєєв, Л. Комалова, Р. Потапова) дослідників. Здебільшого у центрі уваги науковців постають особливості обробки англійської мови (О. Гирин, Д. Гордєєв), проте наявні праці, присвячені питанням обробки української мови, зокрема автоматичному синтаксичному аналізу (Д. Дарчук, Н. Чейлитко), розпізнаванню синтаксичних фразеологізмів (Г. Ситар), можливостям сучасних аналітичних інструментів (О. Кислова, І. Кузіна, І. Дирда) тощо. Проблеми аналізу тональності тексту та виявлення стану агресії порушено в наукових розвідках Д. Гордєєва, Л. Комалової та Р. Потапової. Для пошуку вербальної агресії в російськомовних текстах Р. Потапова та Д. Гордєєв використовували класифікатор Random forest і нейронні мережі (Потапова, Гордеев, 2016). Зазначимо, що якість роботи класифікатора оцінюють за допомогою точності або міри f1. Точність – це відношення правильно класифікованих елементів до їхньої загальної кількості; міра f1 є середнім гармонійним значень влучності та повноти. Найвищим значенням міри f1 є 1, найнижчим – 0 (Гущин, Сич, 2018: 265). Отже, Р. Потапова та Д. Гордєєв отримали результат із точністю від 59,13% до 66,68% (Потапова, Гордеев, 2016: 68). Схожі результати було отримано В. Лакустою на матеріалі української мови (міра f1 – 0.6). Утім, за допомогою конструювання ознак (feature engineering), тобто залучення знань з галузі лінгвістики задля розроблення ознак, що поліпшують роботу алгоритму машинного навчання, результат було покращено до 0.77 (Лакуста, 2019). Вищевикладене закономірно приводить нас до думки, що використання нейронних мереж під час обробки української або російської мови дає задовільний результат, проте проблему не можна вважати розв'язаною.

**Мета статті** – створення аналітичного методу для визначення маркерів вербальної агресії в текстах соціальних мереж. Мета розвідки передбачила розв'язання таких завдань: 1) проаналізувати наявні підходи до автоматизованого розпізнавання агресивних настроїв у текстах мережі Інтернет; 2) запропонувати власний алгоритм для пошуку маркерів вербальної агресії в українськомовному комунікативному просторі соціальних мереж.

*Об'єктом дослідження* є конфліктний дискурс, розгорнутий у комунікативному просторі мережі Інтернет, а предметом – маркери вираження вербальної агресії в текстах соціальних мереж.

Джерельною базою слугують дописи і коментарі користувачів українськомовного сегменту Facebook, зібрані у період із початку квітня до кінця травня 2021 р., загальною кількістю понад 150 одиниць.

**Виклад основного матеріалу.** Автоматизоване розпізнавання деструктивного контенту є одним з актуальних завдань обробки природної мови. Серед можливих підходів до автоматизованого виявлення стану агресії найбільш частотним є використання штучних нейронних мереж. Штучні нейронні мережі є обчислювальними системами, що математично моделюють роботу мозку. Особливістю нейронних мереж є здатність до навчання за відсутності попередніх даних про суб'єкт навчання, інакше кажучи, для того щоб навчити нейронну мережу розпізнавати вербальну агресію в текстах, треба лише надати їй велику кількість прикладів текстів, розмічених як «агресія» і «неагресія». Якість роботи класифікатора оцінюють за допомогою точності або міри f1. Точність – це відношення правильно класифікованих елементів до їхньої загальної кількості; міра f1 є середнім гармонійним значень влучності та повноти. Найвищим значенням міри f1 є 1, найнижчим – 0 (Гущин, Сич, 2018: 265).

Яскравим прикладом застосування нейронних мереж для автоматизованого пошуку деструктивних настроїв в українськомовних текстах є спроба В. Лакусти розпізнати образливі дописи в Twitter. Набір даних налічував близько 3 000 твітів, що були власноруч класифіковані на нейтральні і на ті, що містять мову ненависті. Класифікатор було побудовано на основі попередньо натренованої моделі FastLine та отримано резуль-

тат 0.6 для міри f1. Для поліпшення результату було застосовано низку додаткових моделей та інструментів мовного моделювання. Наприклад, після залучення моделі «торба слів» міра f1 досягла значення 0.66. «Торба слів» представляє текст у вигляді «торби», тобто множини слів, не враховуючи граматику та порядок слів, але беручи до уваги їхню кількість, що стає атрибутом для навчання класифікатора. Кінцевий результат було поліпшено до 0.77 за мірою f1 (Лакуста, 2019).

Іншу реалізацію автоматизованого розпізнавання вербальної агресії в текстах мережі Інтернет було запропоновано Д. Гордєєвим. Джерельною базою слугували англомовний іміджборд 4chan.org та російськомовний іміджборд 2ch.hk. Для розпізнавання агресії було запропоновано алгоритм із застосуванням бібліотеки word2vec, бібліотеки Gensim та залученням власноруч натренованого класифікатора Random forest. Для навчання нейронної мережі було використано 654 047 дописів з 4chan.org та 1 148 692 дописів з 2ch.hk. Точність запропонованого методу для англійської мови становить 88%, для російської – 59,13% (Gordeev, 2016: 73). Згодом результат для російської мови було поліпшено до 66,68% після застосування нестатичної згорткової нейронної мережі (Потапова, Гордеев, 2016).

Вищезазначене дає змогу припустити, що застосування нейронних мереж для обробки природної мови має переваги і недоліки. Однією з переваг, на нашу думку, є значна економія людських ресурсів, оскільки значну частину роботи виконує комп'ютер. Утім, маємо відзначити, що отримана точність не є високою, навіть у разі застосування додаткових інструментів.

Ми пропонуємо метод пошуку маркерів вербальної агресії, заснований на словниках та правилах. Вивчення вербальної агресії привернуло увагу лінгвістів нещодавно, і загальноприйнята дефініція цього терміну відсутня, тому на підготовчому етапі нам потрібно чітко окреслити це явище. С. Форманова ототожнює вербальну агресію з інвективою і надає таке визначення: «Вербально виражене ставлення адресанта до адресата, яке має на меті різке звинувачення, осуд із метою образити, принизити й зганьбити опонента та дискредитувати його» (Форманова, 2018: 122). В. Апресян розуміє під вербальною агресією негативне або критичне ставлення мовця до адресата, виражене мовними засобами (Апресян: 1). Ю. Щербініна потрактовує вербальну агресію як особливо неприйнятне в поточній мовленнєвій ситуації вираження негативних почуттів, емоцій або намірів за допомогою слів (Щербинина, 2001: 6). Н. Кондратенко зазначає, що маркером мовної агресії може бути і нейтральна лексика (Кондратенко, 2019: 299).

Для створення масивів даних із маркерами вербальної агресії ми використовуємо власноруч розмічені агресивні коментарі із соцмережі Facebook, напр.:

***Бордель**, **а не*** (1) *верховна рада!* ***І от ці*** (2) ***курви*** (3) *вирішують долю* ***МОЄЇ УКРА\НИ*** (4) ***??????*** (5)

У цьому прикладі ми бачимо такі маркери агресії: (1) іменник з негативної конотацією + *а не;* (2) *і от ці;* (3) *курви;* (4) *МОЄЇ УКРА\НИ;* (5) *??????*. Також наявні різні типи маркерів: правила (1), сталі конструкції (2), лексичні маркери (3), графічні маркери, тобто зловживання великими літерами (4) або знаками оклику та/або питання (5) тощо.

Аналізуючи кожний приклад, додаємо кожний маркер до певного масиву даних. Масиви даних ми зберігаємо у csv-файлах, що надає можливість додавати елементам різноманітних атрибутів для подальшого використання. Першим атрибутом завжди є числове значення, що на поточному етапі вказує на наявність агресії в тексті, втім, у подальшому буде описувати її вагу. Наведемо приклад масиву лексем (рис. 1).

Кількість атрибутів не є регламентованою, додавати атрибути до елементів можна на будь-якому етапі роботи. У подальшому ці атрибути можна використовувати, наприклад, для кількісного аналізу вживання певних видів інвектив тощо. Роботу алгоритму проілюстровано схемою (рис. 2).

| | | | | |
|---|---|---|---|---|
| 80 | хуйня | 1 | оцінка | |
| 81 | кацапія | 1 | політика | країна |
| 82 | москаль | 1 | політика | національність |
| 83 | козойоб | 1 | політика | національність |
| 84 | виблядок | 1 | характер | |
| 85 | пика | 0,5 | зовнішність | частини тіла |
| 86 | морда | 0,5 | зовнішність | частини тіла |
| 87 | дурбецало | 1 | розумові здібності | |

**Рис. 1**

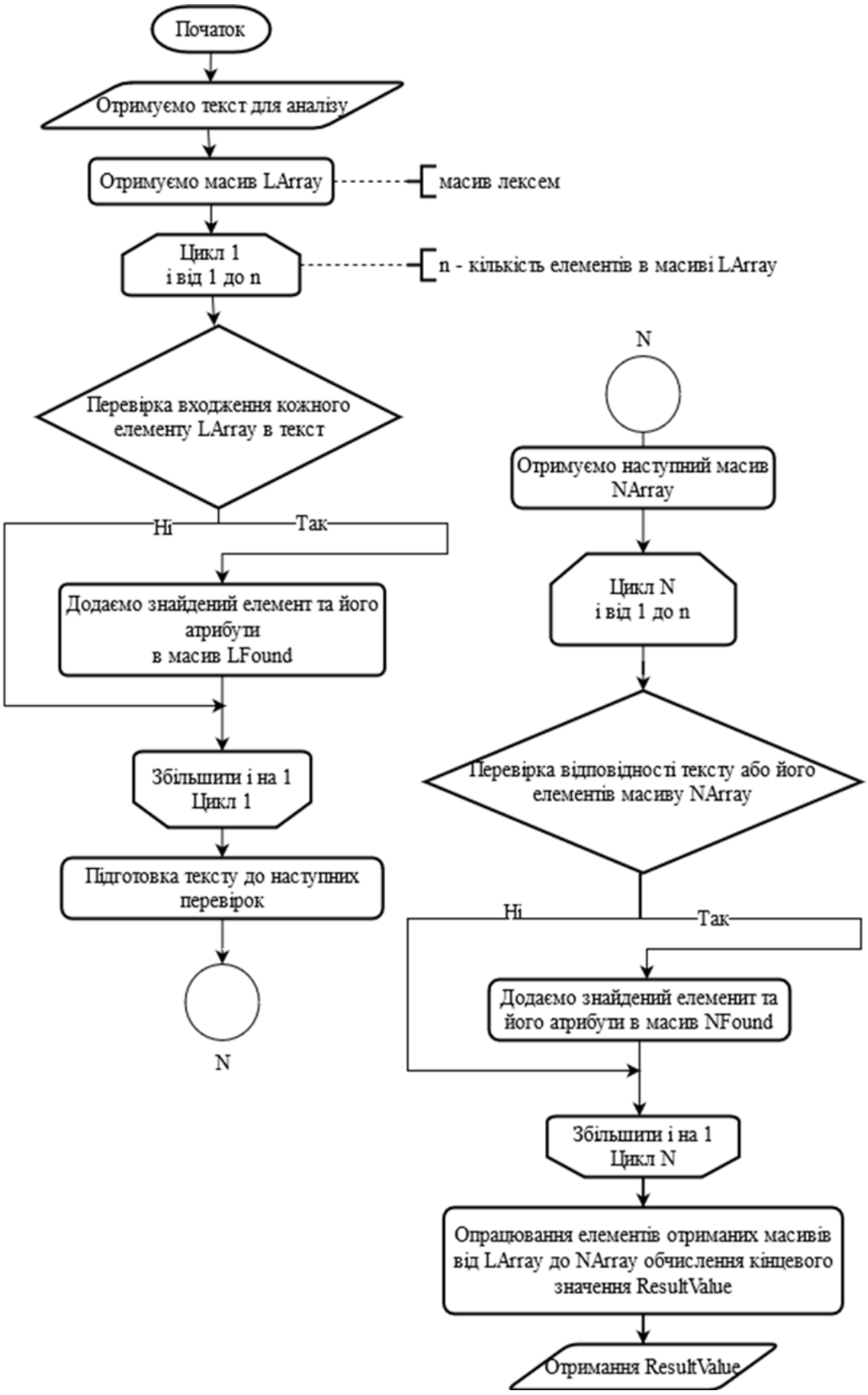
Початок
Отримуємо текст для аналізу
Отримуємо масив LArray
масив лексем
Цикл 1
і від 1 до n
n - кількість елементів в масиві LArray
Перевірка входження кожного елементу LArray в текст
Ні
Так
Додаємо знайдений елемент та його атрибути
в масив LFound
Збільшити і на 1
Цикл 1
Підготовка тексту до наступних перевірок
N
N
Отримуємо наступний масив NArray
Цикл N
і від 1 до n
Перевірка відповідності тексту або його елементів масиву NArray
Ні
Так
Додаємо знайдений елеменит та його атрибути в масив NFound
Збільшити і на 1
Цикл N
Опрацювання елементів отриманих масивів від LArray до NArray обчислення кінцевого значення ResultValue
Отримання ResultValue

**Рис. 2**

На першому етапі ми надаємо програмі текст для аналізу. Програма перевіряє входження кожного з елементів поточного масиву (наприклад, масиву лексем LArray) у текст. Якщо поточний елемент масиву LArray наявний у тексті, програма додає його в новий масив LFound. Після аналізу тексту ми отримуємо масив LFound, що зберігає всі входження елементів масиву LArray у текст та їхні атрибути. Схожі дії проводимо з іншими наявними масивами. Деякі перевірки потребують попередньої обробки тексту, наприклад приведення регістру, лематизації, видалення стоп-слів тощо, тому за необхідності текст проходить обробку. Після перевірки тексту ми отримуємо стільки масивів NFound, скільки було залучено масивів NAarray для перевірки. На наступному етапі ми обчислюємо середнє арифметичне числових елементів кожного масиву окремо і середнє арифметичне отриманих значень. Це число є вагою агресії в заданому тексті. На фінальному етапі роботи алгоритму ми отримуємо вагу агресії в тексті і типологію наявних маркерів вербальної агресії.

**Висновки.** Методи для розпізнавання агресивних настроїв у письмових текстах, засновані на нейронних мережах, не є високоточними, тому було запропоновано алгоритм для пошуку маркерів вербальної агресії на основі словників-масивів даних та правил. Шляхом пошуку в масивах даних маркерів агресії алгоритм робить висновок про наявність агресії в заданому тексті. Перспективу дослідження вбачаємо в реалізації виявлення рівня агресивності тексту та в удосконаленні масивів даних для обробки текстів.

## СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Апресян В. Ю. Имплицитная агрессия в русском языке. *Компьютерная лингвистика и интеллектуальные технологии.* URL: https://goo.su/5sgO.

2. Гущин І. В., Сич Д. О. Аналіз впливу попередньої обробки тексту на результати текстової класифікації. *Молодий вчений.* 2018. № 10(1). С. 264–266. URL: https://goo.su/5yCe.

3. Кондратенко Н. В. Вербальна агресія у спілкуванні в соціальних мережах: актуалізація етнічних гетеростереотипів. *Записки з українського мовознавства*. 2019. Т. 2. № 26. С. 227–233. URL: https://goo.su/65vr.

4. Лакуста В. Конструирование признаков для распознавания оскорбительной речи для украинского языка в Twitter. *Data Science fwdays'19.* 2019. URL: https://goo.su/5ycF.

5. Потапова Р. К., Гордеев Д. И. Определение состояния агрессии с помощью сверточных нейронных сетей. *Речевые технологии.* 2016. № 1(2). С. 65–70. URL: https://goo.su/5yCf.

6. Форманова С. В. Інвектива в сучасних Інтернет-виданнях. *Молодий вчений*. 2018. № 9.1(61.1). С. 121–124. URL: https://goo.su/65Wh.

7. Щербинина Ю. В. Вербальная агрессия в школьной речевой среде : автореф. дис. … канд. пед. наук : спец. 13.00.02. Москва, 2001. 19 с. URL: https://goo.su/65wj.

8. Gordeev D. Automatic Detection of Verbal Aggression for Russian and American Imageboards. *Procedia. Social and Behavioral Sciences.* 2016. № 236. P. 71–75. URL: https://goo.su/5YcG.

## REFERENCES

1. Apresjan V. Ju. Implicitnaja agressija v russkom jazyke [Implicit aggression in language]. Komp'juternaja lingvistika i intellektual'nye tehnologii. URL: https://goo.su/5sgO [in Russian].

2. Gushin I. V., Sych D. O. Analiz vplyvu poperednoi obrobky tekstu na rezultaty tekstovoi klasyfikatsii [Analysis of the impact of text preproccessing on the results of text classification]. Molodyi vchenyi. 2018. № 10(1). pp. 264–266. URL: https://goo.su/5yCe [in Ukrainian].

3. Kondratenko N. V. Verbalna ahresiia u spilkuvanni v sotsialnykh merezhakh: aktualizatsiia etnichnykh heterostereotypiv [Verbal aggression in communication in social networks: actualization of ethnic heterostereotypes]. Zapysky z ukrainskoho movoznavstva. 2019. T 2. № 26. pp. 227–233. URL: https://goo.su/65vr [in Ukrainian].

4. Lakusta V. Konstruirovanie priznakov dlja raspoznavanija oskorbitel'noj rechi dlja ukrainskogo jazyka v Twitter [Feature engineering for abusive language detection for the Ukrainian language on Twitter]. Data Science fwdays'19. 2019. URL: https://goo.su/5ycF [in Russian].

5. Potapova R. K., Gordeev D. I. Opredelenie sostojanija agressii s pomoshh'ju svertochnyh nejronnyh setej [Detection of the state of aggression with convolutional neural networks]. Rechevye tehnologii. 2016. № 1(2). pp. 65–70. URL: https://goo.su/5yCf [in Russian].

6. Formanova S. V. Invektyva v suchasnykh internet-vydanniakh [Invective in modern internet issues]. Molodyi vchenyi. 2018. № 9.1(61.1). pp. 121–124. URL: https://goo.su/65Wh [in Ukrainian].

7. Shherbinina Ju. V. Verbal'naja agressija v shkol'noj rechevoj srede [Verbal aggression in the school speech environment]: avtoref. dis. … kand. ped. nauk: spec. 13.00.02. Moskva, 2001. 19 p. URL: https://goo.su/65wj [in Russian].

8. Gordeev D. Automatic Detection of Verbal Aggression for Russian and American Imageboards. Procedia. Social and Behavioral Sciences. 2016. № 236. pp. 71–75. URL: https://goo.su/5YcG [in English].