DOI: https://doi.org/10.31649/2307-5376-2019-4-35-42

УДК 004.855

С. Д. Штовба, д. т. н., проф.; О. В. Штовба, к. е. н., доц.; О. В. Яхимович; М. В. Петричко

ВПЛИВ СИНТАКСИЧНИХ ЗВ'ЯЗКІВ У РЕЧЕННЯХ НА ЯКІСТЬ ІДЕНТИФІКАЦІЇ ТОКСИЧНИХ КОМЕНТАРІВ В СОЦІАЛЬНІЙ МЕРЕЖІ

Соціальні мережі все частіше стають середовищем для погроз, образ та інших складових кібербулінгу. В онлайнових соціальних мережах задіяна величезна кількість людей, тому виникає потреба в автоматизації діяльності із захисту користувачів від антисоціального впливу. Одним із важливих напрямків такої діяльності є виявлення токсичних коментарів, що містять погрози, образи, зневагу до оточуючих тощо. Зазвичай ідентифікацію токсичних коментарів здійснюють за статистикою мішка слів та мішка символів. В статті досліджується вплив синтаксичних зв'язків у реченнях на якість ідентифікації токсичних коментарів в соціальній мережі. Під синтаксичними зв'язками розуміються зв'язки із власними назвами, з особовими займенниками, з присвійними займенниками тощо. Всього перевірено двадцять синтаксичних ознак речень. Встановлено, що додаткове врахування трьох специфічних ознак суттєво покращує якість ідентифікації токсичних коментарів. Цими трьома специфічними ознаками є такі: кількість зв'язків з власними назвами в однині, кількість зв'язків, в яких фігурують погані слова та кількість зв'язків між особовими займенниками та поганими словами. Експерименти проведено на основі даних із kaggle-змагання "Toxic Comment Classification Challenge". Оригінальну kaggle-задачу категоризації токсичних коментарів було модифіковану у задачу класифікації з двома альтернативами: нейтральний коментар та токсичний коментар. Для наших експериментів оригінальну вибірку із 159751 коментарів скорочено до 106590 коментарів через проблеми з автоматичним виділенням синтаксичних ознак тексту. В модифікованій вибірці частка токсичних коментарів становить 12.8%. Для врахування незбалансованості вибірки даних метрикою якості обрано середнє значення частот помилок класифікації кожного типу. Класифікацію здійснено за допомогою дерева рішень. Дерева рішень синтезувалися за двох правил розщеплення: на основі індекса Джині та ентропійного критерію.

Ключові слова: *аналіз тексту, обробка природньої мови, синтаксичні зв'язки, токсичні коментарі, соціальна мережа, ідентифікація, автоматичне навчання, відбір ознак.*

Вступ

Онлайновими соціальними мережами охоплено переважну більшість користувачів інтернету. Для когось соціальна мережа – це відпочинок, дехто не приймає рішень без обговорення в мережі. За даними порталу Statista в жовтні 2018 р. Шість найбільш популярних мереж здолали планку у 1 млрд. активних користувачів. Самою популярною є Facebook, кількість активних користувачів якої перевищила 2.2 млрд. осіб. Соціальні мережі все частіше стають середовищем для погроз, образ та інших складових кібербулінгу. Враховуючи, що в онлайнових соціальних мережах задіяна така величезна кількість людей, виникає потреба в автоматизації діяльності із захисту користувачів від антисоціального впливу. Одним із важливих напрямків такої діяльності є виявлення в соціальних мережах токсичних коментарів, що містять погрози, образи, зневагу до оточуючих тощо.

Для автоматичного виявлення токсичних коментарів застосовують різноманітні підходи, в першу чергу методи статистичного аналізу текстів. Найпростіший варіант – наївний баєсівський класифікатор по всьому лексикону коментарів з навчальної вибірки [1]. Часто використовують статистику мішка слів та мішка символів, тобто підраховують частоти слів та частоти символів без врахування їх порядку у реченні та зв'язків між ними. Зазвичай враховують такі ознаки: довжина коментаря, кількість великих літер, кількість знаків оклику, кількість знаків питання, кількість граматичних помилок, кількість токенів з неабетковими символами, кількість лайливих, агресивних та погрозливих слів у коментарі тощо [2]. Чим більше коментар містить поганих слів, тим вищі шанси класифікувати його як токсичний.

При цьому виникають труднощі зі статистикою поганих слів. Погані слова автори токсичних коментарів навмисно спотворюють, наприклад, замість *shit*, пишуть *shiiit*, *sh1t*, *sh!t*, *shi**, *shyt*, *siht*, тому науковці розроблюють спеціальні технології виявлення замаскованих образливих слів [3, 4] Порядок слів враховують за деякою множиною усталених словосполучень, наприклад, за n-грамами [5], але це суттєво збільшує обчислювальну складність побудови моделей та не завжди розкриває семантику коментаря.

Метою роботи є дослідження впливу синтаксичних зв'язків слів у реченнях на якість ідентифікації токсичних коментарів. Під синтаксичними зв'язками розуміються зв'язки з власними назвами, з особовими займенниками, з присвійними займенниками тощо. На відміну від методу n-грам та наївного байєсівського підходу, модель на основі синтаксичних зв'язків не прив'язана до лексикону навчальної вибірки. В ній усі різноманітні власні назви, особові займенники, присвійні займенники виділяються в окремі групи, тобто використовуються узагальнені ознаки. Якщо в тестовій вибірці буде інший екземпляр із цієї групи, це не вплине на моделювання. Для виділення синтаксичних зв'язків скористаємося інформаційною технологією з роботи [6] одного із співавтора. Для перевірки ефективності порівняємо результати ідентифікації токсичних коментарів на двох наборах ознак: на типовому – на основі статистики мішка слів та мішка символів, та на розширеному, що додатково включає статистику синтаксичних зв'язків. Експерименти проведено за даними "Toxic Comment Classification Challenge".

Вибірки даних

Вибірку даних "Toxic Comment Classification Challenge" надала компанія Jigsaw для kaggle-змагання [7]. Вибірка складається із 159751 текстового коментаря. Коментарі написано переважно англійською [8]. Для кожного коментаря вказана належність до шести категорій токсичності: toxic – токсичний; severe toxic – дуже токсичний; obscene – непристойний; threat – погрозливий; insult – образливий; identity hate – зневага до ідентичності. Коментар може мати кратну токсичність, тобто належати до двох, трьох і навіть до шести категорій токсичності одночасно (рис. 1). Коментар може бути і нейтральним, тобто не належати до жодної категорії токсичності. Наприклад, коментар *"Your vandalism to the Matt Shirvington article has been reverted. Please don't do it again, or you will be be banned."* є нейтральним. Коментар *«Hi! I am back again! Last warning! Stop undoing my edits or die!»* є токсичним та погрозливим, а коментар *«Would you both shut up, you don't run wikipedia, especially a stupid kid.»* є токсичним та образливим. 16225 коментарів є токсичними, а решта – нейтральними. Розподіл коментарів за кратністю токсичності наведено на рис. 2. З нього видно, що рідко зустрічаються лише коментарі з високою кратністю токсичності – 5 та 6.

Додатково до типових ознак на основі статистики мішка слів та мішка символів пропонується низка специфічних ознак, які враховують синтаксичні зв'язки між словами в коментарі. Ми створили відповідний програмний модуль для парсингу англомовних коментарів. Програмний модуль написано на Java в середовищі Eclipse з використанням Maven. Специфічні ознаки вдалося автоматично розрахувати для 106590 коментарів, що становить 66.8% від обсягу початкової вибірки. Частину коментарів не опрацьовано через іншомовність та велику кількість спотворених слів (out-of-vocabulary words). Спотворення слів відбувається через граматичні помилки та помилки друку. Є багато випадків навмисного спотворення слів до фонетично схожих форм. Для цього замінюють англійські буквосполучення *oo* на *u*, *for* – на *4*, *too* – на *2* тощо. Інший варіант – навмисне спотворення до візуально схожих форм таких як *5h1t*, *b!tch*, *b1tch*. Таке спотворення відбувається заміною візуально схожих символів: *i* та *1, i* та *!*, *S* та *5* тощо.

У новій вибірці частота нейтральних коментарів трохи зменшилась – з 89.8% до 87.2%. Розподіл по категоріям токсичності змінився несуттєво (табл. 1).

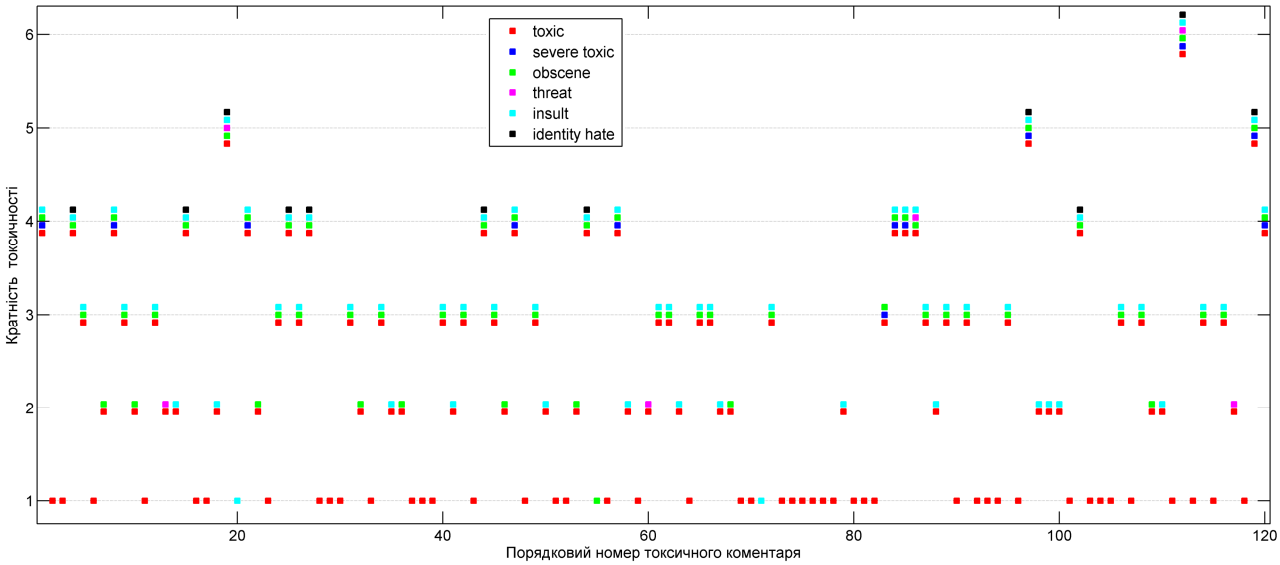

Рис. 1. Категоризація перших 120-ти токсичних коментарів

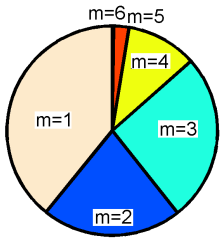

Рис. 2. Розподіл кратності (m) токсичних коментарів

Таблиця 1

Розподіл за категоріями токсичності

Категорія	Коментарів в початковій вибірці	Коментарів у новій вибірці	Частка від початкової вибірки, %
Toxic	15294	12948	84.7
Severe toxic	1595	1492	93.5
Obscene	8449	7303	86.4
Threat	478	442	92.5
Insult	7877	6943	88.1
Identity hate	1405	1251	89

Ознаки коментарів та метрика якості ідентифікації

Формалізований опис кожного коментаря здійснимо за допомогою таких ознак:

x_1 – кількість слів;

x_2 – кількість унікальних слів;

x_3 – частка унікальних слів;

x_4 – кількість токенів без врахування стоп-слів;

x_5 – кількість граматичних помилок;

x_6 – кількість слів, набраних у верхньому регістрі;

x_7 – частка слів, набраних у верхньому регістрі;

x_8 – довжина коментаря;

x_9 – кількість великих літер;

x_{10} – кількість знаків оклику;

x_{11} – кількість знаків питання;

x_{12} – кількість пунктуаційних знаків;

x_{13} – кількість маскувальних символів *, &, $, %.

x_{14} – кількість символів посмішки;

x_{15} – частка знаків оклику;

x_{16} – частка знаків питання;

x_{17} – частка пробілів;

x_{18} – частка заглавних літер;

x_{19} – частка прописних літер;

x_{20} – кількість слів коментаря, які є в списку підозрілих слів на сайті https://www.cs.cmu.edu/~biglou/resources/bad-words.txt;

x_{21} – кількість слів коментаря, які є в списку лайливих слів на сайті http://www.bannedwordlist.com;

x_{22} – кількість слів коментаря, які є в бан-списку фейсбука на сайті https://www.frontgatemedia.com/a-list-of-723-bad-words-to-blacklist-and-how-to-use-facebooks-moderation-tool/;

x_{23} – кількість слів коментаря, які є в бан-списку гугла на сайті https://www.freewebheaders.com/full-list-of-bad-words-banned-by-google/;

x_{24} – кількість слів коментаря, які є в списку підозрілих слів на сайті https://gist.github.com/ryanlewis/a37739d710ccdb4b406d;

x_{25} – кількість слів коментаря, які є у зведеному словнику поганих слів із п'яти вказаних вище списків;

x_{26} - кількість зв'язків з власними назвами в однині;

x_{27} - кількість зв'язків з власними назвами в множині;

x_{28} - кількість зв'язків з особовими займенниками;

x_{29} - кількість зв'язків з присвійними займенниками;

x_{30} - кількість зв'язків із запереченням (з словами never чи not);

x_{31} - кількість зв'язків із запереченням, в яких фігурують власні назви в однині;

x_{32} - кількість зв'язків із запереченням, в яких фігурують власні назви в множині;

x_{33} - кількість зв'язків із запереченням, в яких фігурують особові займенники;

x_{34} - кількість зв'язків із запереченням, в яких фігурують присвійні займенники;

x_{35} - кількість зв'язків між власними назвами в однині та словами, що фігурують у зв'язках із запереченням;

x_{36} - кількість зв'язків між власними назвами в множині та словами, що фігурують у зв'язках із запереченням;

x_{37} - кількість зв'язків між особовими займенниками та словами, що фігурують у зв'язках із запереченням;

x_{38} - кількість зв'язків між присвійними займенниками та словами, що фігурують у зв'язках із запереченням;

x_{39} - кількість зв'язків, в яких фігурують погані слова;

x_{40} - кількість зв'язків із запереченням, в яких фігурують погані слова;

x_{41} - кількість зв'язків між власними назвами в однині та поганими словами;

x_{42} - кількість зв'язків між власними назвами в множині та поганими словами;

x_{43} - кількість зв'язків між особовими займенниками та поганими словами;

x_{44} - кількість зв'язків між присвійними займенниками та поганими словами;

x_{45} - кількість зв'язків між займенниками та поганими словами.

Ознака x_{13} враховує наявність символів *, &, $, %, якими інколи заміщують окремі літери для маскування непристойних слів, наприклад, a$$, $hit тощо. Потреба врахування таких символів обумовлена тим, що користувачі швидко генерують нові варіанти спотворених лайок, які не встигають потрапити у відповідні словники.

Специфічні ознаки x_{26} - x_{45} вперше досліджуються для задачі ідентифікації токсичних коментарів. Для перевірки інформативності нових ознак початкову kaggle-задачу категоризації коментарів зведемо до задачі класифікації з двома класами. Перший клас – нейтральний коментар, а другий клас – токсичний коментар. Вибірка даних є незбалансованою – співвідношення між класами становить приблизно 9 до 1. Тому недоцільно перевіряти якість ідентифікації за частотою помилок класифікації. Як метрику якості ідентифікації застосуємо середнє значення частот помилок класифікації кожного типу:

$$Q_{aver} = \frac{P_{12} + P_{21}}{2},$$

де P_{12} – частота помилок типу 1→2, коли нейтральний коментар визнається токсичним; P_{21} – частота помилок типу 2→1, коли токсичний коментар визнається нейтральним.

Q_{aver} - це проста метрика для перевірки класифікаторів на незбалансованій вибірці даних. Вона підходить для задачі дослідження, тобто для визначення доцільності врахування синтаксичних зв'язків у реченнях для синтезу класифікаторів токсичних коментарів.

Експериментальні дослідження

Класифікатор реалізуємо деревом рішень. Наш вибір обумовлено такими причинами: 1) синтез дерева рішень навіть на великих вибірках є достатньо швидким, що дозволяє провести численні експерименти; 2) під час синтезу дерева рішень здійснюється відбір інформативних ознак, що дозволяє перевірити їх доцільність. Вибірку даних розіб'ємо на навчальну та тестову. У тестову вибірку включимо кожен шостий коментар, а решту – у навчальну. Таким чином, тестова вибірка містить 17765 коментарів, а навчальна – 88825. Дерева рішень синтезуємо на навчальній вибірці та підріжемо таким чином, щоб мінімізувати Q_{aver} на тестовій вибірці. Дослідження проведемо за двох наборів вхідних ознак: типового - x_1 - x_{25} та розширеного - x_1 - x_{44}.

Для вирівнювання балансу проведемо семплування навчальної вибірки, збільшуючи вагу спостережень мінорного класу. Врахуємо, що правильна класифікація коментаря з багатократною токсичністю більш важлива, ніж коментаря, що належить лише до однієї токсичної категорії. Вагу w токсичного коментаря C пропонується встановити у такий евристичний спосіб:

$$w(C) = b + \sqrt{m(C)},$$

де b – базова вага токсичного коментаря; $m(C) \in \{1, 2, ..., 6\}$ – кратність токсичності коментаря C.

На рис. 3 наведено експериментальні залежності якості ідентифікації від базової ваги

токсичного коментаря. Під час експериментів дерева рішень синтезувалися за двох правил розщеплення: на основі індекса Джині та ентропійного критерія. Експерименти засвідчили, що за індексом Джині синтезуються трохи кращі дерева. Малі значення Q_{aver} досягаються коли базова вага токсичного коментаря приймає значення від до 4.5 до 5.8. Мінімум Q_{aver}=0.118 досягнуто, коли $b \in [5.2, 5.5]$. Частота помилок найкращого дерева рішень на усій тестовій вибірці становить 0.0987, при цьому $P_{12} = 0.0919$ та $P_{21} = 0.1442$.

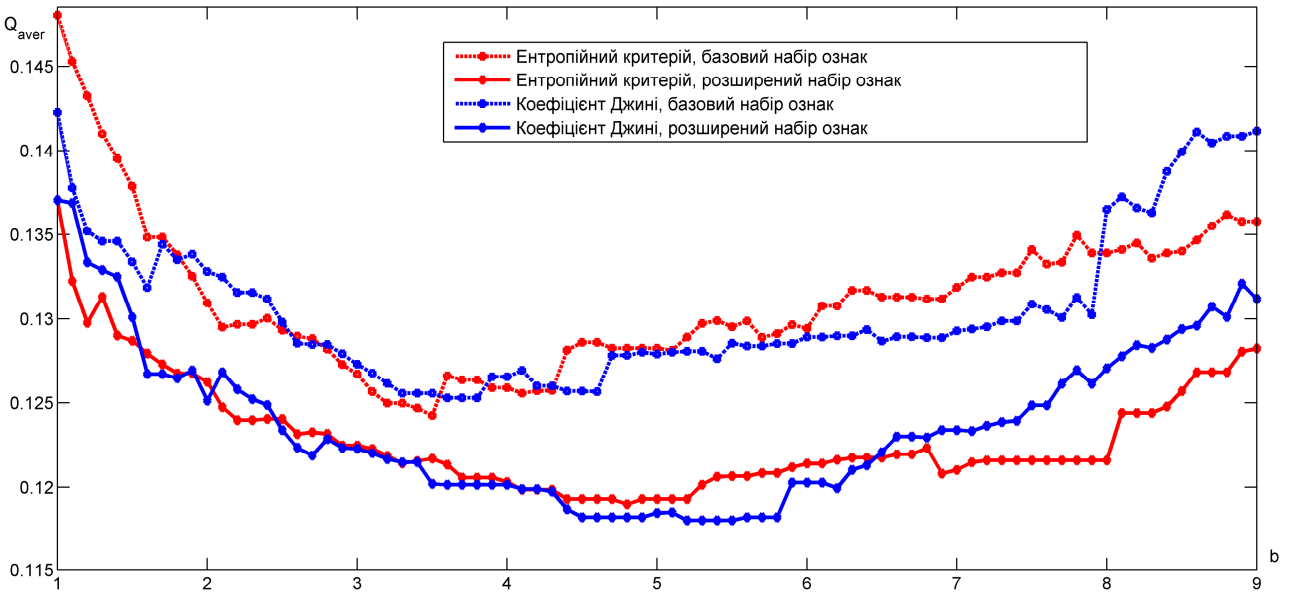

Рис. 3. Залежність якості дерева рішень від базової ваги токсичного коментаря

З рис. 3 видно, що розширений набір ознак значно покращує якість ідентифікації. Перевірка п'яти найкращих дерев рішень показала, що усі вони використовують такі ознаки: x_3-x_9, x_{15}, x_{17}-x_{19}, x_{22}, x_{24}-x_{26}, x_{39} та x_{43}. Чотири із п'яти дерев додатково використовують ознаку x_1. Ці ознаки є найбільш інформативними. Серед них три нові специфічні ознаки: x_{26} - кількість зв'язків з власними назвами в однині, x_{39} - кількість зв'язків, в яких фігурують погані слова та x_{43} - кількість зв'язків між особовими займенниками та поганими словами.

Ще чотири менш важливі ознаки увійшли до кількох найкращих дерев рішень. У двох із п'яти найкращих моделей увійшли типові ознаки x_2, x_{10} та x_{12}. В одну модель потрапила специфічна ознака x_{28} – кількість зв'язків з особовими займенниками. Ці додаткові чотири ознаки можуть бути використані для синтезу більш складних моделей виявлення токсичних коментарів.

Висновки

Розглянуто задачу категоризації коментарів в соціальних мережах для виявлення токсичних текстів. Як експериментальні дані взято вибірку коментарів з kaggle-конкурсу "Toxic Comment Classification Challenge". Для виявлення токсичних коментарів зазвичай використовують типовий набір ознак на основі статистики мішка слів та відповідні словники поганих слів. В статті перевірено ефект від додаткового врахування специфічних ознак. Додатковий набір утворили двадцять специфічних ознак, які описують синтаксичні зв'язки між словами в коментарі.

Встановлено, що врахування специфічних ознак дозволяє суттєво покращити якість ідентифікації токсичних коментарів. Серед запропонованих двадцяти специфічних ознак

найбільш інформативними виявилися наступні три ознаки: кількість зв'язків з власними назвами в однині, кількість зв'язків, в яких фігурують погані слова та кількість зв'язків між особовими займенниками та поганими словами. Відбір трьох специфічних ознак дозволяє суттєво скоротити обчислювальну складність синтаксичного парсингу коментаря, тому що розрахунок усіх двадцяти специфічних ознак вимагає багато ресурсів. Відповідно, якщо до типового набору додати вказані три специфічні ознаки ідентифікація токсичних коментарів може здійснюватися в реальному часі.

Для покращення достовірності ідентифікації токсичних коментарів доцільно перевірити ефект від заміни простого пошуку поганих на спеціальні технології виявлення замаскованих образливих слів [5] на базі нечітких мір схожості та відстані Левенштейна. Перспективним є також поєднання запропонованих моделей на основі статистичного аналізу тексту з моделями інших типів, що враховують ознаки кооперації та загальної активності автора коментаря [9]. Зокрема, доцільно перевірити ефект від врахування таких показників кооперації, як: 1) розподіл кількості відповідей (або інших реакцій) на репліки специфічних учасників дискусії; 2) розподіл кількості коментарів одного і того ж учасника дискусії; 3) схожість імені, часу реєстрації, IP-адрес та адрес електронної пошти автора коментаря з відповідними атрибутами інших користувачів; 4) корельованість активності автора коментаря з іншими учасниками. Також доцільно перевірити ефект від врахування таких показників кооперації, як: 1) розподіл тривалості реакцій користувача – час, за який з'являється його оцінка коментаря, відповідь на звернення тощо; 2) тривалість перебування у соціальній мережі, обсяг створеного контенту та кількість оціночних дій; 3) регулярність перебування в мережі автора коментаря.

Стаття написана за результатами виконання держбюджетної науково-дослідної роботи 46–Д–388 «Ідентифікація прихованих залежностей в онлайнових соціальних мережах на основі методів нечіткої логіки та комп'ютерної лінгвістики».

СПИСОК ЛІТЕРАТУРИ

1. Fine-Grained Classification of Offensive Language / J. Risch, E. Krebs, A. Löser [et all] // Proc. of GermEval 2018, 14th Conference on Natural Language Processing. Vienna, Austria, 2018. – P. 38 – 44.

2. Anatomy of online hate: developing a taxonomy and machine learning models for identifying and classifying hate in online news media/ J. Salminen, H. Almerekhi, M. Milenković [et all] // Twelfth International AAAI Conference on Web and Social Media. – 2018. – P. 330 – 339.

3. Srivastava S. Identifying Aggression and Toxicity in Comments using Capsule Network / S. Srivastava, P. Khurana, V. Tewari // Proc. of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018). – 2018. – P. 98 – 105.

4. Sood S. O. Using Crowdsourcing to Improve Profanity Detection / S. O. Sood, J. Antin, E. F. Churchill // Association for the Advancement of Artificial Intelligence. Spring Symposium: Wisdom of the Crowd. – 2012. – P. 69 – 74.

5. Mohammad F. Is preprocessing of text really worth your time for toxic comment classification? / F. Mohammad // Proc. of Inter. Conference on Artificial Intelligence. CSREA Press. – 2018. – P. 447 – 453.

6. Bisikalo O. Development of the method for filtering verbal noise while search keywords for the English text / O. Bisikalo , A. Yahimovich , Y. Yahimovich // Technology Audit and Production Reserves. – 2018. – № 6. – P. 33 – 41.

7. Toxic Comment Classification Challenge. Available [Електронний ресурс] / Режим доступу : https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge.

8. Stop Illegal Comments: A Multi-Task Deep Learning Approach [Електронний ресурс] / A. Elnaggar, B. Waltl,
I. Glasera [et all] // Software Engineering for Business Information Systems, Technische Universitat¨ Munchen, Germany. – 2018. – Режим доступу : https://arxiv.org/pdf/1810.06665.pdf.

9. Kumar S. Antisocial Behavior on the Web: Characterization and Detection / S. Kumar, J. Cheng, J. Leskovec // Proceedings of the 26th International Conference on World Wide Web Companion. – International World Wide Web Conferences Steering Committee. – 2017. – P. 947 – 950.

Стаття надійшла до редакції 24.12.2018 р.
Стаття пройшла рецензування 18.02.2019 р.

Штовба Сергій Дмитрович – професор, д. т. н., професор кафедри комп'ютерних систем управління, e-mail: shtovba@vntu.edu.ua.

Штовба Олена Валеріївна – доцент, канд. екон. наук, доцент кафедри менеджменту, маркетингу та економіки, e-mail: olena.shtovba@yahoo.com.

Яхимович Олександр Вікторович – аспірант кафедри автоматизації та інтелектуальних інформаційних технологій, yahimovich.olexandr@gmail.com.

Петричко Микола Володимирович – студент факультету комп'ютерних систем та автоматики, e-mail: petrychko.myckola@gmail.com.

Вінницький національний технічний університет.