# Метод Виявлення Тролінгу як Інформаційно-Психологічної Операції в Кіберпросторі

Вероніка Островська
кафедра захисту інформації
Вінницький національний технічний університет
Вінниця, Україна
nika.ostrovska21@gmail.com

Олеся Войтович, Леонід Куперштейн
кафедра захисту інформації
Вінницький національний технічний університет
Вінниця, Україна
voytovych.op@gmail.com, kupershtein.lm@gmail.com

# Method of Detecting Trolling as Informational and Psychological Operationin Cyberspace

Veronika Ostrovska
dept. of Information Protection
Vinnytsia National Technical University
Vinnytsia, Ukraine
nika.ostrovska21@gmail.com

Olesia Voitovych, Leonid Kupershtein
dept. of Information Protection
Vinnytsia National Technical University
Vinnytsia, Ukraine
voytovych.op@gmail.com,kupershtein.lm@gmail.com

***Анотація*—У статті наведено основні види тролів. Запропоновано метод виявлення інформаційно-психологічної операції – тролінгу – у текстовому контенті на основі сучасних методів інтелектуального аналізу: контент-аналізу та методів машинного навчання. Особливістю розробленого методу є те, що вона націлена на контент з негативною тональністю. Це дозволило підвищити ефективність розпізнавання тролінгу.**

***Abstract*—The main types of trolls are presented at the article. It was determined that professional trolls are the greatest danger. The method of detecting information-psychological operation – trolling – in text content on the basis of modern intellectual analysis methods: content analysis methods and methods of machine learning is proposed. The method of trolling detection is to determine the sentiment analysis of the text content of social networks; obtaining indicators that characterize the presence of trolling signs in the text, and computing information entropy of text content for these indicators. The formalization of the task for determining the sentiment analysis of text content is given. In order to determine the text emotion analysis, the signs of the questionable statements in the published facts and sensationality in the detected negative text content, the formulas for calculating relative indices of partial features were proposed. The decision on the presence of trolling signs in the text content of social networks is based on the calculated value of information entropy, which provides automation of decision-making procedures, increasing the efficiency and speed of cyberspace monitoring processes. The peculiarity of the developed method is the focus on content with a negative sentiment analysis. It has allowed to increase the efficiency of trolling recognition.**

***Ключові слова—троль; тролінг; інформаційно-психологічні впливи; соціальні мережі; машинне навчання; контент-аналіз***

***Keywords—troll; trolling; informational and psychological influences; social networks; machine learning; content analysis***

## I. Вступ

У ході проведення інформаційних воєн сучасні спеціалісти стали активно використовувати мережу Інтернет, зокрема, технології соціальних мереж та інших соціальних Інтернет-сервісів [1-2].

Кожен із користувачів соціальних мереж стає не тільки об'єктом для здійснення інформаційно-психологічного впливу, але також сприяє його подальшому поширенню. Дослідження свідчать, що соціальні мережі можуть виступати як інструмент маніпулювання суспільною свідомістю та є сприятливим середовищем для формування громадської думки [3].

Одним із найпоширеніших у соціальних мережах різновидів інформаційно-психологічних операцій є тролінг (англ. trolling – «виспівувати»), що застосовується для формування суспільної думки з актуальних питань та активного обговорення другорядних подій. Результати аналізу текстового контенту, поширюваного тролями, свідчать, що він містить дезінформацію з елементами маніпулювання [4].

Метою роботи є розробка методики для виявлення тролінгу в соціальних мережах для підвищення ефективності його розпізнавання.

## II. РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

Кожна людина стикається з когнітивними обмеженнями від упереджень та ефектів обмеження рамками. Когнітивне упередження – це структура відхилення у судженні, коли висновки про інших людей і ситуації можуть бути зроблені нелогічно [5].

Упередження, що досліджені шляхом проведення психологічних експериментів, свідчать про те, що люди часто не виконують усі вимоги на шляху до раціональності їхніх дій. І роблять це систематично та в певному напрямку, який можливо спрогнозувати. Когнітивні упередження також змушують громадськість бути відкритою до сприйняття інформаційно-психологічних впливів. Знаючи особливості когнітивних упереджень людини, тролі створюють спеціально підготовлену інформацію – меми – з елементами маніпулятивного впливу.

В соціальних Інтернет-сервісах тролі як засоби агресивного впливу поділяються на: природних, професійних та ботів. Під природними тролями будемо розуміти користувача, який, як правило, спеціально тролінгом не займається. Професійні тролі є найманими особами, які залишають коментарі в соціальних мережах за завданням урядових структур. Такі тролі за гроші залишають коментарі, призначені для здійснення пропаганди або розміщення політичної реклами під новинами на форумах, блогах або на інших соціальних ресурсах в мережі. Бот – це шкідлива програма, яка імітує діяльність людини через користувацькі інтерфейси [6].

Оскільки професійні тролі становлять найбільшу небезпеку, їх розпізнавання є найбільш актуальною задачею на сьогодні з точки зору кібербезпеки. Тому для виявлення тролінгу в соціальних мережах необхідно виокремлювати в публікаціях такі важливі ознаки його застосування: тональність повідомлень, емоційність повідомлень, сумнівність наведених фактів, сенсаційність повідомлення, повідомлення у великій кількості та дублікати повідомлень [7].

Методика виявлення тролінгу ґрунтується на методах машинного навчання і контент-аналізу та полягає в наступному.

**Перший етап** – це визначення тональності публікацій $F$ на основі сучасних методів машинного навчання з учителем, без учителя, використанням правил або словників. Задача визначення тональності публікацій розв'язується за допомогою методів машинного навчання. Аналіз сучасних підходів показав, що для вирішення задач визначення тональності контенту соціальних мереж з метою виявлення інформаційно-психологічних операцій одним із найкращих інструментів є нейронні мережі [8-10].

Для аналізу тональності текстових даних доцільно застосовувати глибоке навчання рекурентних нейронних мереж, яке не викликає складнощів із перенавчанням, на відмінну від згорткових та повнозв'язних нейронних мереж. Етап полягає у віднесенні тональності повідомлення $d_j$, $j = \overline{1,n}$ до заздалегідь визначеного класу $c_i$, $i = \overline{1,m}$ – негативний або позитивний [8]

$$(d_j, c_i) \in D \times C, \tag{1}$$

де $D$ – колекція документів; $C$ – множина класів тональності повідомлень.

У формалізованому вигляді завданням визначення тональності текстового контенту публікації є класифікація $i$-го документу $D_i$ при аналізі вектора ознак тональності $X_i$. При цьому виконується синтез ієрархічної структури документів (бінарне дерево рішень $S_D$), в кожній вершині якого застосовується вирішальне правило, яке реалізує метод послідовного аналізу при визначенні тональності текстових документів між класами $D_{pos}$ та $D_{neg}$. На кожному $i$-му етапі вирішального правила аналізується чергова ознака тональності $X_i$ і обчислюється відношення правдоподібності

$$\Theta = \prod_i P(x_{ik}/D_{neg}) \Big/ P(x_{ik}/D_{pos}), \tag{2}$$

яке порівнюється з порогами $\Theta > A$, $\Theta < B$.

При виконанні однієї з умов приймається рішення про $D_{pos}$ та $D_{neg}$ відповідно і виконується перехід на більш низький рівень ієрархії $S_D$ з метою уточнення тональності. При виконанні обох нерівностей додається $i + 1$ ознака і процедура повторюється.

У результаті виконання першого етапу отримаємо визначений клас тональності публікації і нормоване числове значення відповідно до шкали, що поділяється на чотири інтервали: виражено позитивна (1.00 – 0.76), помірно позитивна (1.75 – 0.51), помірно негативна (0.50 – 0.26), виражено негативна (0.25 – 0.00) [11].

Оскільки повідомлення тролів містять критику, самовпевнені висловлювання, нецензурну лексику та інші негативні засоби впливу, доцільно звернути увагу на контент саме з негативною тональністю. Тому на наступних етапах буде досліджуватися контент з негативною тональністю, що був відібраний на цьому етапі.

**Другий етап** полягає в визначенні емоційного забарвлення негативного контенту $F_1$ соціальних мереж. На цьому етапі відбувається визначення наявності у повідомленні проявів емоцій чи почуттів автора стосовно досліджуваних об'єктів або подій і полягає у встановленні кількості окличних речень, вигуків, прислівників, вживання лексем емоційного характеру.

Окличні речення $F_{11}$ – відношення числа окличних речень $S_{dec}$ до всієї кількості речень $S$ в текстовому контенті

$$F_{11} = \frac{S_{dec}}{S}. \quad (3)$$

Вигуки $F_{12}$ – показник вживання у публікації вигуків (наприклад, ага, ну-ну, овва, от тобі і на тощо)

$$F_{12} = \frac{W_{int}}{W}, \quad (4)$$

де $W_{int}$ – кількість знайдених вигуків у публікації; $W$ – загальна кількість слів.

Прислівники $F_{13}$ – кількість прислівників $W_{adv}$ у текстовому контенті, що застосовуються для порівняння та зосередження читача публікації на його емоціях (наприклад, немов, більше, сором, на жаль, на щастя, назавжди тощо)

$$F_{13} = \frac{W_{adv}}{W}. \quad (5)$$

Лексеми емоційного характеру $F_{14}$ – показник вживання у коментарях лексем емоційного характеру $W_{emot}$ (наприклад, посміховисько, жертва, жахливий тощо)

$$F_{14} = \frac{W_{emot}}{W}. \quad (6)$$

**Третій етап** полягає у виявленні ознак сумнівності викладених у негативному контенті соціальних мереж фактів $F_3$, який зводиться до розрахунку частки, що показує ступінь відсутності аргументації, частки запитальних речень та частки сумнівних висловлювань.

Відсутність аргументації $F_{21}$ – показник використання слів, які відкидають необхідність обґрунтування та підтвердження правдивості контенту (наприклад, явно, незаперечний факт, поза сумнівом, вочевидь, певна річ, само собою зрозуміло тощо

$$F_{21} = \frac{W_{un\,arg}}{W}. \quad (7)$$

де $W_{un\,arg}$ – кількість слів із запереченням необхідності підтвердження контенту.

Наявність запитальних речень $F_{22}$ – показник наявності запитальних речень $S_q$ у текстовому контенті:

$$F_{22} = \frac{S_q}{S}. \quad (8)$$

Сумнівні висловлювання $F_{23}$ – показник вживання слів, які можуть трактуватися по-різному (наприклад, можливо, ймовірно, постійно):

$$F_{23} = \frac{W_{amb}}{W}, \quad (9)$$

де $W_{amb}$ – кількість неоднозначних висловлювань.

**Четвертий етап** – встановлення сенсаційності негативного контенту $F_3$ внаслідок підвищення уваги користувачів соціальних мереж, оперативності контенту в результаті використання слів для створення атмосфери скороминущості й першочерговості явищ. Етап зводиться до виявлення наступних ознак.

Підвищення уваги $F_{31}$ – показник використання слів, що здатні привернути увагу читача, зумовлюють зростання тривоги (наприклад, напад, жах, небезпека)

$$F_{31} = \frac{W_{atten}}{W}, \quad (10)$$

де $W_{atten}$– кількість виявлених слів, що підвищують увагу.

Оперативність $F_{32}$ – показник вживання слів, які створюють атмосферу скороминущості й першочерговості явищ (наприклад, відразу, терміново, раптово)

$$F_{32} = \frac{W_{effic}}{W}, \quad (11)$$

де $W_{effic}$ – кількість знайдених слів для демонстрації оперативності.

**П'ятий етап** – визначеннякількості повідомлень від одного користувача та дублікатів повідомлень $F_4$. У соціальних Інтернет-сервісах користувачі звертають увагу на контент з великою кількістю репостів, коментарів та «лайків» [12]. Публікуючи багато коментарів, тролі спричиняють соціалізацію цього контенту та створюють видимість активного обговорення, їх важливості та критичності [13].

Сутність алгоритму визначення дублікатів повідомлень полягає у знаходженні повторень конструкцій слів у контенті, що аналізується, та наведена нижче.

*Крок 1* полягає в приведенні тексту повідомлень до канонічного вигляду. Для цього необхідно видалити смайли, хештеги, HTML-теги, гіперпосилання, розділові знаки, прийменники, сполучники й інші компоненти, які не несуть змістовного навантаження контенту. В деякихвипадках необхідно здійснювати нормалізацію іменників до однини називного відмінка.

На *Кроці 2* здійснюється розбиття нормалізованого тексту на фрагменти. Вибір значення довжини текстового фрагменту залежить від довжини самого тексту і лежить в інтервалі 5-10. Зростання довжини вихідного тексту вимагає збільшення цього показника.

На *Кроці 3* обчислюється хеш-сума фрагменту тексту, яка порівнюється, з використанням функцій (SHA1, SHA2, SHA3, MD5 тощо) і записується в двовимірний масив даних. Після цього випадково обирають значення хешів зі збережених для порівняння між собою.

*Крок 4* зводиться до розрахунку показника відповідності порівнюваного текстового контенту як співвідношення кількості хешів фрагментів з однаковими значеннями до їх загальної кількості.

В загальному вигляді зв'язок між ознаками застосування тролінгу в соціальних мережах, що розглянуті на попередніх етапах, зображено у вигляді ієрархії на Рис. 1.

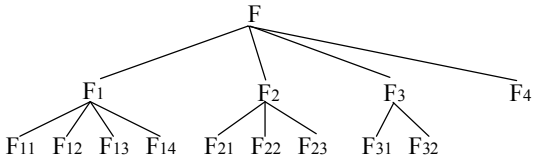


Рис. 1. Ієрархія ознак застосування тролінгу в соціальних мережах.

**Шостий етап** – розрахунок інформаційної ентропії застосування засобів тролінгу в соціальних мережах, використовуючи показники, що отримані на попередніх етапах. Суть полягає у встановленні рівня невизначеності щодо наявності у негативному контенті прихованого впливу на користувачів соціальних мереж. Числове значення порівнюється із шкалою оцінки для прийняття рішення про рівень загрози. Шкала оцінки застосування засобів тролінгу в соціальних мережах поділяється на п'ять інтервальних значень ентропії: дуже висока – 0.00-0.20, висока – 0.21-0.49, звичайна – 0.50-0.74, низька – 0.75-0.90, дуже низька – 0.91-1.00 [11].

Таким чином, зміст методу виявлення тролінгу зводиться до визначення тональності текстового контенту соціальних мереж; отримання показників, які характеризують наявність ознак тролінгу в тексті; обчислення для цих показників інформаційної ентропії текстового контенту та порівняння її числового значення із допустимим граничним. Інформаційна ентропія зменшується при зростанні частот появи ознак тролінгу у текстовому контенті соціальних мереж. У випадку малих частот прояву ознак інформаційна невизначеність зростає [14].

## Висновки

Запропонований метод виявлення ознак застосування тролінгу ґрунтується на методах інтелектуального аналізу текстового контенту. Рішення про наявність ознак тролінгу у текстовому контенті соціальних мереж приймається на основі обчисленого значення інформаційної ентропії, що забезпечує автоматизацію процедур прийняття рішень, підвищення ефективності та швидкодії процесів моніторингу кіберпростору.

## Література References


[1] Китов П. Совершенствование способов и средств ведения психологических операцийвооружённых сил США / П. Китов // Зарубежное военное обозрение. – 2013. – № 3 (792). – С. 19–22.

[2] Александр Ольшанский: как Украине победить в информационной войне с РФ и чем опасны Google, Twitter и Facebook [Электронный ресурс]. – Режим доступа : http://ain.ua/2015/04/21/576406.

[3] Пенченко В. М. Інформаційна безпека особи в умовах соціалізації Інтернет-сервісів / В. М. Пенченко // Актуальні проблеми управління інформаційною безпекою держави : зб. матер. наук.-практ. конф., 20 березня 2012 року. – Київ : Наук.-вид. відділ НА СБ України, 2012. – С. 81–83.

[4] Чернишук С. В. Методика виявлення кібернетичних загроз у природномовних текстах / С. В. Чернишук // Проблеми створення, випробування, застосування та експлуатації складних інформаційних систем. – 2013. – Вип. 8. – С. 112–121.

[5] Haselton M. G. The evolution of cognitive bias / M. G.Haselton, D. Nettle, P. W. Andrews ; D. M. Buss (Ed.) // The Handbook of Evolutionary Psychology. – Hoboken, NJ, US : John Wiley & Sons Inc., 2005. –P. 724–746.

[6] Войтович О.П., Дудатьєв А.В., Головенько В.О. Модель та засіб для виявлення фейкових облікових записів у соціальних мережах // Вчені записки таврійського національного університету ім. В.І. Вернадського. Серія: Технічні науки. Частина 1 – 2018. – № 1 Том 29 (68). – С. 112 – 119.

[7] Островська В. М. Тролінг як засіб інформаційної війни / В. М. Островська, О. П. Войтович [Електронний ресурс]. – Режим доступу : https://conferences.vntu.edu.ua/index.php/all-fitki/all-fitki-2018/paper/view/4071/4535. – Назва з екрану.

[8] Faraz A. A comparison o f text Categorization methods / A. Faraz // International Journal on Natural Language Computing. – 2016. – № 5(1). – P. 31–44.

[9] Волосюк Ю. В. Методи класифікації текстових документів в задачах Text Mining / Ю. В. Волосюк // Наукові записки Українського науково-дослідного інституту зв'язку. – 2014. – № 6(34). – С. 76–81.

[10] Fernandez-Martinez F. Text categorization methods for automatic estimation o f verbal intelligence / F. FemandezMartinez, K. Zablotskaya, W. Minker// Expert Systems with Applications. – 2012. – № 9 (10). – P. 9807–9820.

[11] Грищук Р. В. Метод оптимізації розмірності потоку вхідних даних для систем захисту інформації / Р. В. Грищук, В. М. Мамарев // Інформаційна безпека. – 2012. – № 2 (8). – С. 27–34.

[12] Voitovych O. Badania sieci społecznych jako źródła informacji w czasie wojny [Електронний ресурс] / Voitovych O., Holovenko V. // Inżynier XXI wieku projectujemy przyszlosc : monografia / [pod red. : Jacek Rysiński] [Електронний ресурс]. – Режим доступу : https://ir.lib.vntu.edu.ua/bitstream/handle/123456789/17254/2688.pdf?sequence=3. – Назва з екрану.

[13] Молодецька К. В. Підхід до виявлення організаційних ознак інформаційних операцій у соціальних інтернет-сервісах / К. В. Молодецька // Пріоритетні напрямки розвитку телекомунікаційних систем та мереж спеціального призначення. Застосування підрозділів, комплексів, засобів зв'язку та автоматизації в АТО : зб. матер. IX наук.-практ. конф., 25 листоп. 2016 р. – Київ : ВІТІ, 2016. – С. 130–131.

[14] Молодецька-Гринчук К. В. Методика виявлення маніпуляцій суспільною думкою у соціальних інтернет-сервісах / К. В. Молодецька-Гринчук // Інформаційна безпека. – 2016. – № 4(24). – С. 80–92