ISSN 2079-5459 (print)
ISSN 2413-4295 (online)

УДК 004.773.2 doi:10.20998/2413-4295.2018.09.15

СИСТЕМА КРИТЕРІЇВ ДЛЯ ВИЯВЛЕННЯ ФРАГМЕНТІВ ОНЛАЙН-ДИСКУСІЙ З ПІДОЗРОЮ НА НАЯВНІСТЬ ІНФОРМАЦІЙНО-ПСИХОЛОГІЧНОЇ МАНІПУЛЯЦІЇ

З. Д. ГОЛУБ

Кафедра соціальних комунікацій та інформаційної діяльності, НУ «Львівська політехніка», Львів, УКРАЇНА
e-mail: zorianaholub@gmail.com

АНОТАЦІЯ *У статті розглянуто критерії для виявлення і виділення підозрілих фрагментів дискусій онлайн-спільнот. Критерії поділено за темпоральною характеристикою на динамічні та статичні. Згідно з формальною моделлю онлайн-спільноти виділено організаційно-структурні рівні, на яких критерії відрізняються формою реалізації та механізмом виявлення. На основі критеріїв побудовано фільтри для виявлення підозрілих фрагментів дискусії, описано будову системи фільтрів, застосовано метод вагових показників для визначення підозрілого фрагменту дискусії на основі результатів проходження системи фільтрів.*
Ключові слова: *інформаційно-психологічна маніпуляція; динамічний критерій; статичний критерій; система фільтрів; онлайн-спільнота.*

SYSTEM OF CRITERIA FOR DETECTING ONLINE DISCUSSION FRAGMENTS SUSPECTED OF CONTAINING INFORMATION AND PSYCHOLOGICAL MANIPULATION

Z. HOLUB

Social Communication and Informational Activities Department, Lviv National Polytechnic University, Lviv, UKRAINE

ABSTRACT *This paper is devoted to the problem of detecting and outlining discussion fragments of online communities suspended of containing information and psychological manipulation (IPM). Criteria are divided into dynamic and static on the basis of temporal features. Static criteria are participant-centred, they constitute an overall characteristic of the participants activities in light of IPM. Static criteria are divided into three classes in accordance to the entitlement of a participant to influence and create data for the criteria. The latter are profile data, behavioural characteristics, activities history. Dynamic criteria are exploit to analyse a particular act of information activity and are used for analysing communication at the message level. Dynamic criteria are divided into to classes on the basis of level of processing and analysis required to obtained data necessary for the criteria. The latter are shallow and deep dynamic criteria. In accordance to the formal model of an online community organizational and structural levels are defined. Suspected discussion fragments are detected by means of analysing data extracted from different levels of organizational structure of online communities. They are community level, discussion level, message level. Criteria differs by the form of realization and detection mechanism at each level. The system of filters for detecting suspicious discussion fragments is devised on the basis of the criteria, taking into consideration that criteria have different significance and stay in different cause-result relations with the IPM precedent. The structure of the system of filters is described, the method of weighted indices for detecting a suspicious discussion fragment on the basis of filtering procedure results is applied. Every criterion has different significance for the detecting IPM. Considering that fact, for every filer the weight was assigned. A discussion fragment that is trapped by a predefined quantity of filters is identified as a suspicious discussion fragment.*
Keywords*: informational and psychological manipulation; dynamic criteria; static criteria; the system of filters; online communities*

Вступ

Забезпечення ефективності та безпеки інформаційної діяльності в онлайн-спільнотах належить до ключових проблем сьогодення. Внаслідок популярності онлайн-спільнот, як платформ для обміну інформацією, негативні явища традиційного спілкування перемістилися в онлайн-дискусії. Постала проблема захисту учасників онлайн-спільнот від деструктивного впливу інформаційно-психологічної маніпуляції (ІПМ).

Процес виявлення ІПМ в онлайн-спільнотах передбачає ґрунтовий та глибокий аналіз інформаційного наповнення онлайн-спільноти, тобто обробку вербальної, невербальної та метаграфічної складових текстових повідомлень [1, 2]. Цей процес буде часо- та ресурозатратними, якщо не обмежити область аналізу. Виникає потреба у розробленні критеріїв для поверхневого аналізу онлайн-спільнот з метою виділення та окреслення областей для глибинного аналізу на предмет наявності ІПМ. Ці критерії мають враховувати багаторівневу будову онлайн-спільноти, особливості текстової онлайн-комунікації та види інформаційної активноті, які доступні у онлайн-спільнотах.

Сучасні дослідження спрямовані на вирішення проблеми виявлення деструктивної інформаційної діяльності в онлайн-спільнотах пропонують застосовувати інструменти виділення ставлення [3-4], інструменти тональнісного аналізу [5-7] для вирішення дотичних проблем, але ці інструменти дозволяють вирішити лише певні аспекти проблеми виявлення ІПМ в онлайн-спільнотах.

Ціль роботи

У роботі поставлено мету розробити критерії для виявлення підозрілих фрагментів дискусії. На основі цих критеріїв необхідно побудувати систему фільтрів для виявлення фрагментів дискусії, які містять найбільше ознак ІПМ.

Виклад основного матеріалу

Критерії, які свідчать про потенційну наявність ІПМ поділяються на два темпоральні види динамічні і статичні. В основу даної класифікації покладено часовий період необхідний для збору інформації для розрахунку критерію. Введемо наступні групи критеріїв.

• Статичні критерії — це критерії, які розраховуються на основі діяльності учасника протягом встановленого періоду часу.

• Динамічні критерії — це критерії, на основі яких можна робити висновки зразу ж після їх ідентифікації, які не потребують спостережень протягом певного періоду часу.

Статичні критерії наявності ІПМ

Статичні критерії зосереджені навколо діяльності учасника онлайн-спільноти. Статичні критерії поділяються на три класи залежно від спроможності учасника впливати на та створювати дані для критеріїв.

Для розрахунку статичних критеріїв небхідно дані наступних типів:

• дані;

• поведінкові характеристики користувача;

• історія дій.

Дані профілю — це дані, які учасник самостійно додає до профілю і змінює. Відсутність або мала кількість даних цього типу (аватар, місце навчання, робота і т.д.) є підставою для виникнення підозр щодо цього учасника [8].

Дані профілю, зокрема, використовуються на підготовчому етапі алгоритму виявлення ІПМ в соціальних середовищах інтернету, з метою створення списку релевантних спільнот, адже на основі даних профілю визначають соціально-демографічі характеристики [9].

Аналіз таких даних профілю, як імена учасників використовується на підотовчому етапі під час сортування дискусій за сприятливістю до здійснення ІПМ та на етапі виявлення, з метою ідентифікації підозрілих фрагментів дискусі. Наприклад, виявити потенційних маніпуляторів на основі імені можна наступним чином. Якщо ім'я містить провокативний меседж, набори символів, які не мають жодного змісту, чи структура або елементи імені подібні до імен інших учасників, то від цих профілів можна очікувати маніпулятивної діяльності [10].

Поведінкові характеристики учасника – це змінні в часі характеристики. З метою ефективної ідентифікації маніпуляторів, їх доцільно визначати через встановлені часові періоди. До цієї групи ознак, які виявляються на основі аналізу активності учасника в спільноті, це: частота відвідуваності спільноти та дописів у дискусії, період дня, протягом якого користувач бере участі в інформаційних процесах.

До поведінкових характеристик учасника належать дані, які є наслідками діяльності учасника. Рейтинг, кількість друзів, кількість послідовників характеризують діяльність учасника, але учасник не може їх безпосередньо контролювати та змінювати.

Ефективність системи виявлення ІПМ залежить від актуальності даних для розрахунку значень статичних критеріїв. З усіх типів статичних критеріїв дані про поведінкові характеристики учасника вимагають регулярного оновлення, оскільки це найбільш змінні в часі статичні критерії.

Історія дій — це своєрідна біографія профілю учасника. Критеріями цього типу є характеристики етапів активності учасника у спільноти. Етапи активності — це є часові проміжки між важливими віхами діяльності учасника у спільноті. Критеріїв цього типу базуються на таких даних: дата реєстрації, дата першого повідомлення в обговоренні, дата ініціації дискусії.

Статичні та динамічні критерії відносно організаційно-структурних рівнів онлайн-спільноти

За допомогою статичних критеріїв розглядають інформаційну діяльність учасника в проекції на структуру онлайн-спільноти, тобто відносно трьох рівнів організації інформаційного наповнення спільноти (рис. 1). Відповідно до формальної моделі онлайн-спільноти цими трьома рівнями є рівень спільноти, дискусії та повідомлення []. Критерії цих трьох організаційно-структурних рівнів відрізняються механізмом розрахунку та значимістю.

Значення критеріїв рівня обговорення та спільноти прив'язані до учасника спільноти. За допомогою цих критеріїв виявляють підозрілого учасника спільноти, діяльність якого необхідно перевірити на наявність ІПМ.

В той час як статичні критерії рівня повідомлення безпосередньо вказують на елементи інформаційного наповнення дискусії, які потрібно проаналізувати на наявність ІПМ.

Динамічні критерії використовуються для аналізу конкретного акту інформаційної-активності. Вони не містять узагальненої інформації про роль та поведінку учасника в онлайн-спільноті.

На відміну від статичних, динамічні критерії не поділяються за структурно-організаційними рівнями. Динамічні критерії використовуються для аналізу комунікації на рівні повідомлення. Вони вказують на наявність у повідомленні слідів та наслідків застосування прийому ІПМ. На основі слідів та наслідків застосування ІПМ виявляють використанні прийоми ІПМ та стани учасників-жерт ІПМ,

ISSN 2079-5459 (print)
ISSN 2413-4295 (online)

відповідно. Ця інформація необхідна для визначення застосуваної тактики ІПМ та ідентифікації прецеденту ІПМ.

Рис.1 – Класифікація криитеріїв ІПМ

Значення критеріїв рівня обговорення та спільноти прив'язані до учасника спільноти. За допомогою цих критеріїв виявляють підозрілого учасника спільноти, діяльність якого необхідно перевірити на наявність ІПМ.

В той час як статичні критерії рівня повідомлення безпосередньо вказують на елементи інформаційного наповнення дискусії, які потрібно проаналізувати на наявність ІПМ.

Динамічні критерії використовуються для аналізу конкретного акту інформаційної-активності. Вони не містять узагальненої інформації про роль та поведінку учасника в онлайн-спільноті.

На відміну від статичних, динамічні критерії не поділяються за структурно-організаційними рівнями. Динамічні критерії використовуються для аналізу комунікації на рівні повідомлення. Вони вказують на наявність у повідомленні слідів та наслідків застосування прийому ІПМ. На основі слідів та наслідків застосування ІПМ виявляють використанні прийоми ІПМ та стани учасників-жерт ІПМ, відповідно. Ця інформація необхідна для визначення застосуваної тактики ІПМ та ідентифікації прецеденту ІПМ.

На підготовчому етапі використовуються лише ті динамічні критерії, фільтри на основі яких не потребують глибинного аналізу інформаційного наповнення дискусії. Наприклад, тональнісні маркери емоцій, а не складені маркери психічних станів, які задаються за допомогою векторів елементаим, яких є емоції. За допомогою фільтрів на основі динамічних критеріїв поверхневого рівня визначають лише значні і часті перепади емоцій учасників дискусії.

Формальна модель фільтрів для виявлення підозрілих фрагментів дискусій

Поставивши за ціль підвищити ефективність алгоритму моніторингу дискусії з метою виявлення ІПМ, уникнуто перебору всього інформаційного наповнення дискусії. Цього досягнуто за рахунок окреслення областей дискусії, які мають найбільше ознак наявності ІПМ, тобто виділення підозрілих фрагментів дискусії.

Підозрілі фрагменти дискусії — це набори логічно пов'язаних повідомлень, кількісні і якісні характеристики яких та кількісні і якісні характеристики профілів авторів цих повідомлень властиві повідомленням з ІПМ.

Систему характеристик підозрілих фрагментів дискусії розроблено на основі закономірностей виявлених внаслідок аналізу комунікації у онлайн-спільнотах, їх узагальненої структури та способів презентації учасників. На основі цих характеристик та формальної моделі спільноти, дискусії, повідомлення та учасника [11] запропоновано систему фільтрів, яка рекомендуватиме для подальшого аналізу на наявність ІПМ підозрілі фрагменти дискусії, тобто уривки, які потенційно містять ІПМ.

Систему фільтрів розроблено на основі статичних критеріїв наявності ІПМ та частково на основі динамічних. Ознаки рівнів спільноти та дискусії створюють узагальнену картину діяльності учасника в дискусіях Інтернету, в той час як критерії рівня повідомлення вказують на ознаки ІПМ в конкретних актах інформаційної діяльності.

Виявляючи маніпуляцію потрібно враховувати той факт, що кожна інтернет-дискусія має свої правила, стиль спілкування і аудиторію [12]. Тому, для конкретного фільтра, критерії кожного рівня необхідно адаптувати до особливостей комунікації в конкретній дискусії та встановити відповідне порогове значення критеріїв. На основі цього порогового значення фільтри поділяють інформаційне наповнення дискусій на безпечне та з підозрою на наявність ІПМ.

Порогові значення визначаються для кожної спільноти окремо на основі експертного аналізу комунікації в конкретній онлайн-спільноті. Порогове значення статичних ознак розраховується на основі аналізу профілю, історії дія учасника та поведінкових особливостей його діяльності.

Для прикладу розглянемо статичну ознаку ІПМ рівня обговорення $RepliesNumberRatio$. $RepliesNumberRatio$ — це відношення кількості відповідей учасника на повідомлення до усіх повідомлень опублікованих учасником (1).

$$RepliesNumberRatio = \frac{RepliesNumber}{MessagesTotal} * 100\%, \qquad (1)$$

де $RepliesNumber$ — це кількість відповідей учасника на повідомлення інших; $MessagesTotal$ — це кількість усіх повідомлень написаних учасником.

У дослідженні [10] визначено, для 84.3% учасників, які розміщували повідомлення з визначеним замовниками змістом $RepliesNumberRatio < 40\%$, в той час як для користувачів, які розміщували не замовлену інформацію $RepliesNumberRatio \approx 90\%$. Відповідно на основі цих даних оптимальне порогове значення для спільноти вибирають у діапазоні [30%,

50%] залежно від особливостей спільноти. Якщо $\mathrm{Re}\,pliesNumber$ є нижчим за порогове значення, то даний фільтр відбирає всі повідомлення цього користувача для подальшої перевірки.

Згідно з дослідженням [12] 60% учасників спільноти, які розміщують проплачені тексти реагують на коментарі з інтервалами часу 200 с, в той час лише 40% нормальних учасників реагуть на повідомленння зі швидкістю близькою до 200. Зазвичай, вони виступають так бурхливо і їхня швидкість публікації повідомлень є меншою. З точки зору ІПМ в онлайн-спільнотах, цей критерій потрібно трактувати дещо інакше: аномально швидкі відповіді на повідомлення є або продуктом маніпуляторів, або реципієнтів, які вже «клюнули на гачок», але так чи інакше зменшення інтервалу часу між публікаціями повідомлень є свідченням того, що в уривку потенційно наявна ІПМ.

Додатковим критерій, тривалість часницької діяльності, розраховується а основі наступної формули (2):

$$MembershipPeriod = CurrentDate - RegistrationDate \quad (2)$$

де $CurrentDate$ – це поточна дата; $RegistrationDate$ – це дата реєстрації учасника.

Якщо $MembershipPeriod < 30$, то діяльність учасника потребує додаткової перевірки.

Всі ці критерії мають різну важливість і знаходяться в різній причинно-наслідковій залежності з фактом здійснення маніпулятивної діяльності. Саме тому для кожного з критеріїв було встановлено вагові показники. Наприклад, $\mathrm{Re}\,pliesNumberRatio$ є важливішим з точки зору виявлення підозрілих уривків, ніж $MembershipPeriod$.

Кількість обговорень спільноти, в яких учасник розміщує тематично релевантні повідомлення — це складений критерій ІПМ. Розглядаючи цей критерій виділимо три варіанти поведінки: користувач веде активну інформаторську активність лише в кількох тематично пов'язаних дискусія, учасник є активним в багатьох дискусіях, учасник є активним у багатьох дискусіх, але всі його пости є одного тематиного спрямування і, тому є неревантнми у багатьох дискусіях, де вони розміщені. Останній варіант поведінки свідчить про маніпулятивну діяльність. Звичайно, теоретично можливий ще такий варіант: більшість постів, які розміщує учасник не релевантні до тематики обговорень, але такий тип діяльності не можна розглядати як маніпуляція, адже така поведінка більше схожа нескладного бота і інформаційне наповнення створене ним не буду серйозно розглядатися реципієнтами.

Для виявлення ІПМ на основі активності учасника та релевантності його постів розроблено наступний фільтр (3):

$$\begin{cases} DiscussionActivenessRatio > 50\% \\ IrrelevantMessagesRatio > 50\% \end{cases} \quad (3)$$

де $DiscussionActivenessRatio$ — це відношення кількості дискусій, в яких учасник веде активну діяльність до кількості усіх дискусій, яке відображає відносну активність учасника спільноти (4); $IrrelevantMessagesRatio$ — це відношення кількості нерелевантних повідомлень розміщених учасником до усіх створених ним повідомлень (5).

$$DiscussionActiveness = \frac{ActiveDiscussionNumber}{CommunityDiscussionTotal} \quad (4)$$

де $ActiveDiscussionNumber$ – це кількість дискусій, в яких учасник веде активну інформаторську діяльність; $CommunityDiscussionTotal$ – це загальна кількість усіх дискусій в спільноті.

$$IrrelevantMessagesRatio = \frac{IrrelevantMessages}{MessagesTotal} \quad (5)$$

де $IrrelevantMessages$ – це кількість розміщених учасником повідомлень, які є нерелевантні до тематики дискусії; $MessagesTotal$ – це загальна кількість усіх повідомлень розміщених учасником.

Якщо подвійна умова (3) виконується для учасника, то він потенційно веде маніпулятину діяльність.

Кожен критерій, на основі якого побудовано фільтр, має різну значущість з точки зору виявлення ІПМ. Тому для кожного фільтра визначений ваговий коефіцієнт, значення якого вказують експерти. Значення вагових коефіцієнтів становлять множину вагових коефіцієнтів критеріїв підозрілих уривків (6).

$$Weight_i = \left\{Weight_{ij}\right\}_{j=1}^{N^{Criterion}} \quad (6)$$

де $Weight_{ij}$ – це вага критерію потенційної ІПМ. При чому для вагових коефіцієнтів критеріїв вірні наступні рівності (7), (8):

$$\sum_i Weight_i = 1 \quad (7)$$

$$Weight_i \geq 0 \quad (8)$$

Кожен критерій має визначений експертами з галузі менеджменту онлайн-спільнот ваговий показник, тому кількість елементів множини вагових показників дорівнює кількості критеріїв потенційної ІПМ.

Формальна модель критерію наявності ІПМ задається наступною формулою:

ISSN 2079-5459 (print)
ISSN 2413-4295 (online)

$$Criterion_i = \langle CriterionValue_i, Weight_i \rangle$$

де $CriterionValue_i$ – значення критерію наявності ІПМ у фрагменті; $Weight_i$ – це встановлений ваговий показник критерію наявності ІПМ в уривку.

Індикатор наявності потенційної ІПМ, $PotentialIPMIndicator_i$, може приймати значення 0 або 1, відповідно до того чи фільтром було затриманоінформаційний контент чи ні (9).

$$PotentialIPMIndicator_i = \begin{cases} 0, CriterionValue_i \in ThresholdValue_i \\ 1, CriterionValue_i \notin ThresholdValue_i \end{cases} \quad (9)$$

де $ThresholdValue_i$ – порогове значення критерію, для конкретної спільноти.

Порогове значення критерію задається інтервалом (10):

$$ThresholdValue_i \in \left[\min^i, \max^i\right] \quad (10)$$

Підозрілість фрагменту дискусії оцінюється на основі наступної формули (11):

$$IPMLikelyhood = \sum_{j=1}^{N^{Criterion}} Criterion_{ij} Weight_{ij} \quad (11)$$

Якщо повідомлення користувача будуть відібраними наперед встановленою кількістю фільтрів, то певне інформаційне наповнення створене цим користувачем та пов'язані з ним повідомлення інших буде ідентифіковане як підозрілий уривок дискусії (12).

$$PotentialManipulator = \begin{cases} 0, IPMLikelyhood = [0;0,5) \\ 1, IPMLikelyhood = [0,5;1] \end{cases} \quad (12)$$

Фрагменти дискусії, які містять акти інформаційної діяльності, які внаслідок застосування системи фільтрів ідентифіковано як з наявністю елемементів ІПМ, підлягають глибинному аналізу інформаційного наповнення на наступному етапі алгоритму моніторингу онлайн-спільноти з метою виявлення ІПМ.

Висновки

Завдяки критеріям розробленими на основі даних профілів учасників, поведінкових особливостей, історії діяльності, та поверхневим динамічні характеристикам повідомлень, розроблено систему фільтрів, яка затримує повідомлення онлайн-дискусій з підозрою на наявність ІПМ. Система фільтрів уможливлює реалізацію ефективного моніторинг з онлайн-спільнот з метою виявлення ІПМ, виключивши, такі недоліки як значні затрати часу та ресурсів на перевірку великих об'ємів інформації. Результати системи фільтрів передаються на наступний етап алгоритму моніторингу онлайн-спільноти, який перебачає глибинний аналіз інформаційного наповнення з метою виявлення прецендентів ІПМ та подальшої ідентифікації тактик ІПМ.

Список літератури

1. **Peleschyshyn, A.** Development of the System for Detecting Manipulation in Online Discussions / **A. Peleschyshyn, Z. Holub** // *International Conference on Systems, Control and Information Technologies 2016*. – Springer International Publishing. – 2016. – vol. 543 – pp. 111-117. – doi: 10.1007/978-3-319-48923-0_15.
2. **Peleschyshyn, A.** Methods of real-time detecting manipulation in online communities / **A. Peleschyshyn, Z. Holub, and I. Holub** // *Scientific and Technical Conference "Computer Sciences and Information Technologies (CSIT)*, 2016 XIth International. – IEEE. – 2016. – pp. 15-17. – doi: 10.1109/STC-CSIT.2016.7589857.
3. **Dey, L.** Opinion Mining from Noisy Text Data / **L. Dey, S. K. Haque, A. Mirajul** // *Proceedings of the second workshop on Analytics for noisy unstructured text data.* – 2008. – p. 83-90. – doi: 10.1007/s10032-009-0090-z.
4. **Iqbal, S.** The survey of sentiment and opinion mining for behavior analysis of social media / **S. Iqbal, A. Zulqurnain, Y. Wani et al.** // *International Journal of Computer Science & Engineering Survey (IJCSES)*. – 2015. – vol. 6. – p.21-27. – doi: 10.5121/ijcses.2015.6502.
5. Case Study: Advanced Sentiment Analysis [Електронний ресурс]. URL: chrome-extension://padcapdkhelngdelppbbjmkmkfceoikg/content/pdf/viewer/viewer.html?file=http%3A%2F%2Fwww.aclweb.org%2Fanthology%2FW04-2328.
6. **Pozzi, F. A.** Sentiment Analysis in Social Networks / **F. A. Pozzi, E. Fersini, E. Messina, et al.** Morgan Kaufmann, 2016, 284 p. – doi: 10.1016/B978-0-12-804412-4.00001-2.
7. **Oraby, S.** And That's a Fact: Distinguishing Factual and Emotional Argumentation in Online Dialogue / **S. Oraby, L. Reed, R. Compton, et al.** In The 2nd Workshop on Argumentation Mining, at The North American Chapter of the Association for Computational Linguistics (NAACL), Denver, Colorado, 2015. – doi: 10.3115/v1/W15-0515.
8. **Anwar, T.** Modeling a Web Forum Ecosystem into an Enriched Social Graph / **T. Anwar, M. Abulaish** // *Ubiqutos Social Media Analysis*. – LNCS, Springer, 2013. – Vol. 8329. – pp. 152-173. – doi: 10.1007/978-3-642-45392-2_8.
9. **Fedushko, S.** Development of a software for computer-linguistic verification of socio-demographic profile of web-community member / **S. Fedushko.** - Webology, 2014. – doi: 10.6084/m9.figshare.2056647.
10. **Benevenuto, F.** Characterizing user behavior in online social networks / **F. Benevenuto, T. Rodrigues, M. Cha, et al.** // *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement*. – ACM New York, USA 2009. – pp. 49-62. – doi: 10.1145/1644893.1644900.
11. **Peleschyshyn, A.** Formal Model and Key Features of an Online Community Fundamental for Detecting Informational and Psychological Manipulation / **A. Peleschyshyn, Z. Holub, I. Holub** // *Scientific and Technical Conference "Computer Sciences and Information*

Technologies (CSIT), 2017 XIth International. IEEE. – 101-104 pp. – doi: 10.1109/STC-CSIT.2017.8098746.
12. **Liu, D.** User Interest and Interaction Structure in Online Forums / **D. Liu, D. Percival, S. Flenberg** // *The AAAI Press,* 2010.

Bibliography (transliterated)

1. **Peleschyshyn, A., Holub, Z.** Development of the System for Detecting Manipulation in Online Discussions. *International Conference on Systems, Control and Information Technologies 2016*, Springer International Publishing, 2016, **543**, 111-117, doi: 10.1007/978-3-319-48923-0_15.
2. **Peleschyshyn, A., Holub, Z., Holub, I.** Methods of real-time detecting manipulation in online communities. *Scientific and Technical Conference "Computer Sciences and Information Technologies (CSIT)*, 2016 XIth International, IEEE, 2016, 15-17, doi: 10.1109/STC-CSIT.2016.7589857.
3. **Dey, L., Haque, S. K., Mirajul, A.** Opinion Mining from Noisy Text Data. *Proceedings of the second workshop on Analytics for noisy unstructured text data,* 2008, 83-90, doi: 10.1007/s10032-009-0090-z.
4. **Iqbal, S., Zulqurnain, A., Wani, Y. et al.** The survey of sentiment and opinion mining for behavior analysis of social media. *International Journal of Computer Science & Engineering Survey (IJCSES)*, 2015, **6**, 21-27, doi: 10.5121/ijcses.2015.6502.
5. Case Study: Advanced Sentiment Analysis Available at: chrome-extension://padcapdkhelngdelppbbjmkmkfceoikg/content/pdf/viewer/viewer.html?file=http%3A%2F%2Fwww.aclweb.org%2Fanthology%2FW04-2328.
6. **Pozzi, F. A., Fersini, E., Messina, E., et al.** Sentiment Analysis in Social Networks. Morgan Kaufmann, 2016, 284, doi: 10.1016/B978-0-12-804412-4.00001-2.
7. **Oraby, S., Reed, L., Compton, R., et al.** And That's a Fact: Distinguishing Factual and Emotional Argumentation in Online Dialogue. In The 2nd Workshop on Argumentation Mining, at The North American Chapter of the Association for Computational Linguistics (NAACL), Denver, Colorado, 2015, doi: 10.3115/v1/W15-0515.
8. **Anwar, T., Abulaish, M.** Modeling a Web Forum Ecosystem into an Enriched Social Graph. *Ubiqutos Social Media Analysis*, LNCS, Springer, 2013, **8329**, 152-173, doi: 10.1007/978-3-642-45392-2_8.
9. **Fedushko, S.** Development of a software for computer-linguistic verification of socio-demographic profile of web-community member. - Webology, 2014, doi: 10.6084/m9.figshare.2056647.
10. **Benevenuto, F., Rodrigues, T., Cha, M., et al.** Characterizing user behavior in online social networks. *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement*, ACM New York, USA, 2009, 49-62, doi: 10.1145/1644893.1644900
11. **Peleschyshyn, A., Holub, Z., Holub, I.** Formal Model and Key Features of an Online Community Fundamental for Detecting Informational and Psychological Manipulation. *Scientific and Technical Conference "Computer Sciences and Information Technologies (CSIT)*, 2017 XIth International. IEEE, 101-104, doi: 10.1109/STC-CSIT.2017.8098746.
12. **Liu, D., Percival, D., Flenberg, S.** User Interest and Interaction Structure in Online Forums. *The AAAI Press,* 2010.

Відомості про авторів (About authors)

Голуб Зоряна Дмитрівна - аспірант, Національний університет «Львівська політехніка», м. Львів., Україна; e-mail: zorianaholub@gmail.com.

Zoriana Holub – postgraduate student, Lviv National Polytechnic University, Lviv, Ukraine; e-mail: zorianaholub@gmail.com.

Будь ласка, посилайтесь на цю статтю наступним чином:

Голуб, З. Д. Формалізація система критеріїв для виявлення фрагментів онлайн-дискусій з підозрою на наявність інформаційно-психологічної маніпуляції / **З. Д. Голуб** // *Вісник НТУ «ХПІ», Серія: Нові рішення в сучасних технологіях*. – Харків: НТУ «ХПІ». – 2018. – № 9 (1285). – С. 106-111. – doi:10.20998/2413-4295.2018.09.15.

Please cite this article as:

Holub, Z. System of criteria for detecting online discussion fragments suspected of containing information and psychological manipulation. *Bulletin of NTU "KhPI". Series: New solutions in modern technologies*. – Kharkiv: NTU "KhPI", 2018, **9** (1285), 106-111, doi:10.20998/2413-4295.2018.09.15.

Пожалуйста, ссылайтесь на эту статью следующим образом:

Голуб, З. Д. Формализация система критериев для выявления фрагментов онлайн-дискуссий с подозрением на наличие информационно-психологической манипуляции /**З. Д. Голуб** // *Вестник НТУ «ХПИ», Серия: Новые решения в современных технологиях*. – Харьков: НТУ «ХПИ». – 2018. – № 9 (1285). – С. 106-111. – doi:10.20998/2413-4295.2018.09.15.

АННОТАЦИЯ *В статье рассмотрены критерии для выявления и выделения подозрительных фрагментов дискуссий онлайн-сообществ. Критерии разделяются за их темпоральными характеристиками на динамические и статические. В соответствии с формальной моделью онлайн-сообществ выделены организационно-структурные уровни, на которых критерии различаются формой их реализации и механизмом выявления. На основании критериев созданы фильтры для выявления подозрительных фрагментов дискуссии, описано строение системы фильтров, применено метод весовых показателей для определения подозрительного фрагмента дискуссии на основании результатов прохождения системы фильтров.*

Ключевые слова: *информационно-психологическая манипуляция; динамический критерий; статический критерий; система фильтров; онлайн-сообщество.*

Надійшла (received) 12.03.2018