НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»
ФІЗИКО-ТЕХНІЧНИЙ ІНСТИТУТ
Кафедра інформаційної безпеки

«До захисту допущено»
В.о. завідувача кафедри

_____ М.В.Грайворонський
(підпис)
“_____”_____2019 р.

Дипломна робота

на здобуття ступеня бакалавра

з напряму підготовки 6.170101 «Безпека інформаційних і комунікаційних систем»
на тему: Методика виявлення потенційно небезпечних повідомлень у соціальних мережах

Виконав: студент 4 курсу, групи ФБ-52
(шифр групи)

Пастущак Даніель Васильович _____
(прізвище, ім’я, по батькові) (підпис)

Керівник к.т.н., доц. каф. ІБ, Коломицев Михайло Володимирович _____
(посада, науковий ступінь, вчене звання, прізвище та ініціали) (підпис)

Рецензент к.т.н., доц. каф. ТК ФІОТ Корнага Ярослав Ігорович _____
(посада, науковий ступінь, вчене звання, науковий ступінь, прізвище та ініціали) (підпис)

Засвідчую, що у цій дипломній роботі немає запозичень з праць інших авторів без відповідних посилань.
Студент _____
(підпис)

Київ - 2019 року

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»
ФІЗИКО-ТЕХНІЧНИЙ ІНСТИТУТ
Кафедра інформаційної безпеки

Рівень вищої освіти – перший (бакалаврський)

Напрям підготовки 6.170101 «Безпека інформаційних і комунікаційних систем»

ЗАТВЕРДЖУЮ
В.о. завідувача кафедри
_____ М.В.Грайворонський
(підпис)
«___»_____2019 р.

ЗАВДАННЯ
на дипломну роботу студенту

Пастущаку Даніелю Васильовичу
(прізвище, ім'я, по батькові)

1. Тема роботи

«Методика виявлення потенційно небезпечних повідомлень у соціальних мережах»,

науковий керівник роботи

доц. каф. ІБ, к.т.н., Коломицев Михайло Володимирович,
(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)

затверджені наказом по університету від 27 травня 2019 р. № 1414-С

2. Термін подання студентом роботи: 10 червня 2019 р.

3. Вихідні дані до роботи: набори даних, що містять повідомлення з соціальної мережі Twitter, методи машинного навчання для класифікації тексту

4. Зміст роботи:

- Аналіз загроз в соціальних мережах спричинених небезпечними повідомленнями

- Особливості розробки системи пошуку потенційно небезпечних повідомлень в соціальних мережах

- Реалізація методики виявлення потенційно небезпечних повідомлень в соціальних мережах

5. Перелік ілюстративного матеріалу (із зазначенням плакатів, презентацій тощо):

- Презентація

6. Дата видачі завдання: 8 вересня 2018 року.

Календарний план

№ з/п	Назва етапів виконання дипломної роботи	Термін виконання етапів дипломної роботи	Примітка
1	Узгодження тематики з науковим керівником	08.09.18 – 30.09.18	
2	Вивчення проблематики	30.09.18 – 01.11.18	
3	Пошук та вивчення літератури	01.11.18 – 15.12.18	
4	Написання першого розділу дипломної роботи	15.12.18 – 15.01.19	
5	Написання другого розділу дипломної роботи	15.01.19 – 05.02.19	
6	Визначення інструментів для розробки власної методики	05.02.19 – 15.02.19	
7	Розробка методики та написання програмної частини	15.02.19 – 15.04.19	
8	Проходження переддипломної практики	15.04.19 – 19.05.19	
9	Аналіз результатів	10.05.19 – 20.05.19	
10	Написання третього розділу проекту	20.05.19 – 29.05.19	
11	Передзахист проекту	30.05.19	
12	Підготовка графічної частини дипломної роботи	30.05.19 – 18.06.19	
13	Захист дипломної роботи	19.06.19	

Студент _____ _____
(підпис) (ініціали, прізвище)

Науковий керівник роботи _____ _____
(підпис) (ініціали, прізвище)

РЕФЕРАТ

Робота обсягом 62 сторінок містить 9 ілюстрацій, 5 таблиць, 1 додаток та 32 літературних посилань.

Метою даної кваліфікаційної роботи є розробка методики виявлення потенційно небезпечних повідомлень у соціальних мережах на прикладі розробки класифікатора повідомлень у соціальній мережі Twitter за терористичним змістом.

Об’єктом дослідження є повідомлення у соціальній мережі Twitter.

Предметом дослідження є зміст повідомлень у соціальній мережі Twitter.

Результати роботи викладені у вигляді таблиці та ілюстрацій, що демонструють правильність вибору методів виділення суттєвих ознак, а також прикладу класифікації повідомлень.

Результати роботи можуть бути використані для подальшої розробки системи пошуку терористичних повідомлень у соціальних мережах, або бути перекваліфікованими для пошуку повідомлень іншого характеру, наприклад, кібербулінгу.

тероризм, соціальні мережі, повідомлення, ІДІЛ, python, машинне навчання, обробка природної мови, метод опорних векторів

ABSTRACT

The work includes 62 pages, 9 illustrations, 5 tables, 1 appendix and 32 literary references.

The purpose of this qualification work is to develop a methodology for detecting potentially undesirable messages in social networks by example of developing a classifier of messages in the social network Twitter for terroristic content.

The object of the research is a dataset of messages in the social network Twitter.

The subject of the study is the content of messages in the social network Twitter.

The results of the work are presented in the form of a table and illustrations, which demonstrate the correctness of the choice of methods for feature extraction, as well as an example classification of messages.

The results of the work can be used for further development of a system for detection of terroristic messages in social networks, or be re-qualified to detect messages of a different nature, such as cyber-bulling.

terrorism, social networks, messages, ISIS, python, machine learning, natural language processing, support vector machine

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

Соціальна мережа — соціальна структура, утворена індивідами або організаціями. Вона відображає розмаїті зв'язки між ними через різноманітні соціальні взаємовідносини, починаючи з випадкових знайомств і закінчуючи тісними родинними зв'язками.

Твіт (англ. Tweet) — повідомлення у соціальній мережі Твіттер довжиною до 140 символів.

НОТД – Національна організація технічних досліджень (Індія).

ЦШІР – Центр штучного інтелекту та робототехніки (Індія).

ЦРУ – Центральне розвідувальне управління (США).

АНБ – Агенція національної безпеки (США).

УРО – Управління радіотехнічної оборони (Австралія).

ЦУЗ – Центр урядового зв’язку (Великобританія).

СБУК – Служба безпеки урядових комунікацій (Нова Зеландія).

ЦБК – Центр безпеки комунікацій (Канада).

СБУ – Служба безпеки України.

ІДІЛ – Ісламська Держава Іраку та Леванта.

Кібербулінг (кіберзнущання) – це знущання, яке відбувається в технологічних комунікаційних платформах, таких як електронні листи, чати, телефонні розмови та соціальні мережі, зловмисником, який використовує платформу для переслідування жертви, неодноразово посилаючи шкідливі повідомлення, сексуальні зауваження або погрози; публікуючи незручні фотографії або відео жертви; або через іншу неприйнятну поведінку.

Python — інтерпретована об'єктно-орієнтована мова програмування високого рівня зі строгою динамічною типізацією. Розроблена в 1990 році Гвідо ван Россумом. Структури даних високого рівня разом із динамічною семантикою та динамічним зв'язуванням роблять її привабливою для швидкої розробки програм, а також як засіб поєднування наявних компонентів.

Хеш-тег – слово або фраза, яким передує символ «#». Користувачі можуть об'єднувати групу повідомлень за темою або типом з використанням хештегів — слів або фраз, які починаються з #.

Машинне навчання (англ. machine learning) — це підгалузь штучного інтелекту в галузі інформатики, яка часто застосовує статистичні прийоми для надання комп'ютерам здатності «навчатися» (тобто, поступово покращувати продуктивність у певній задачі) з даних, без того, щоби бути програмованими явно.

Обробка природної мови (англ. Natural-language processing, NLP) — загальний напрям інформатики, штучного інтелекту та математичної лінгвістики. Він вивчає проблеми комп'ютерного аналізу та синтезу природної мови. Стосовно штучного інтелекту аналіз означає розуміння мови, а синтез — генерацію розумного тексту. Розв'язок цих проблем буде означати створення зручнішої форми взаємодії комп'ютера та людини.

Повідомлення терористичного характеру – в даній роботі це повідомлення, написані сторонниками ІДІЛ, несуть джихадистський зміст. Для тренування класифікатора був використаний набір таких повідомлень з відкритого доступу, що називається isisfanboy.

Метод опорних векторів — це метод аналізу даних для класифікації та регресійного аналізу за допомогою моделей з керованим навчанням з пов'язаними алгоритмами навчання, які називаються опорно-векторними машинами (ОВМ, англ. support vector machines, SVM, також опорно-ве´кторними мережами, англ. support vector networks).

NER – Named Entity Recognition (процес пошуку і зберігання маркерів, які представляють людей, організації, місця розташування, дати, часу, та інших).

POS – Part of Speech (граматичні категорії, що використовуються в різних наборах даних для аналізу)

ВСТУП

Сьогодні мільйони користувачів Інтернету регулярно відвідують тисячі соціальних веб-сайтів, щоб продовжувати зв'язуватися зі своїми друзями, ділитися своїми думками, фотографіями, відео та обговорювати навіть їхнє повсякденне життя.

Оскільки використання соціальних мереж стає все більш впровадженим в повсякденне життя користувачів, особиста інформація стає легко розкритою та використана у злочинних намірах. На сьогоднішній день соціальні мережі є одним з найнебезпечніших місць у кіберпросторі, оскільки ними користуються не тільки доброзичливі для користувача люди.

В соціальних мережах існують загрози, що стосуються дуже багатьох складових благополуччя життя користувача. Недоброзичливці можуть нанести шкоду матеріальному статусу, репутації, технічному/програмному забезпеченню, фізичному та ментальному здоров’ю користувача соціальних мереж.

Протягом останніх років представники терористичних угруповань задля зручності та широкого поширення своїх ідей заполонили соціальні мережі. Вони ведуть блоги, переписки в публічному доступі. Їх повідомлення часто несуть терористичний характер. Очевидною є важливість виявлення таких повідомлень.

Актуальність роботи. Зумовлюється тим, що на даний момент у світі поширилися терористичні акти і в багатьох з цих випадків виконавці ділились своїми намірами у соціальних мережах. Системи моніторингу соціальних мереж переважно залежать від людей, тобто ці повідомлення перевіряються вручну людьми. Представлена робота пропонує автоматичний класифікатор, що базується на вже виявлених повідомленнях терористичного характеру, для виявлення таких повідомлень у режимі реального часу.

Мета роботи. Класифікація повідомлень за терористичним змістом у соціальній мережі Twitter.

Завдання роботи. Розробка класифікатора повідомлень за терористичним змістом у соціальній мережі Twitter.

Об’єкт дослідження. Повідомлення у соціальній мережі Twitter.

Предмет дослідження. Зміст повідомлень соціальній мережі Twitter.

Наукова новизна. Підтверджується тим, що в результаті роботи набуло подальшого розвитку використання машинного навчання, та обробки природної мови у сфері протидії тероризму.

Практичне значення. Результати роботи можуть бути використані для подальшої розробки системи пошуку терористичних повідомлень у соціальних мережах, або бути перекваліфікованими для пошуку повідомлень іншого характеру, наприклад, кібербулінгу.

1 АНАЛІЗ ЗАГРОЗ В СОЦІАЛЬНИХ МЕРЕЖАХ СПРИЧИНЕНИХ НЕБЕЗПЕЧНИМИ ПОВІДОМЛЕННЯМИ

1.1 Види загроз, що спричиняють повідомлення в соціальних мережах

Зі збільшенням використання соціальних мереж, багато користувачів стали вразливими до загроз своєї приватності та безпеки. Ці загрози можуть бути розділені на 4 основні категорії (рис. 1.1). Перша категорія вміщує класичні загрози, зокрема, загрози конфіденційності та безпеки, які не тільки загрожують користувачам соцмереж, але й користувачам Інтернету, які не використовують соціальні мережі. Друга категорія охоплює сучасні загрози, тобто загрози, які в основному є унікальними для середовища соціальних мереж, і які використовують інфраструктуру соціальних мереж для загрози конфіденційності та безпеки користувача. Третя категорія складається з комбінованих загроз, де ми описуємо, як сьогоднішні нападники можуть, і часто роблять, поєднувати різні типи атак для створення більш складних і летальних нападів. Четверта і остання категорія включає в себе загрози, спеціально орієнтовані на дітей, які використовують соціальні мережі. На рис. 1 представлені схеми всіх конкретних загроз, перелічених у вище перелічених розділах. Межі між усіма цими категоріями загроз, однак, можуть стати нечіткими, оскільки методи і цілі часто перетинаються.

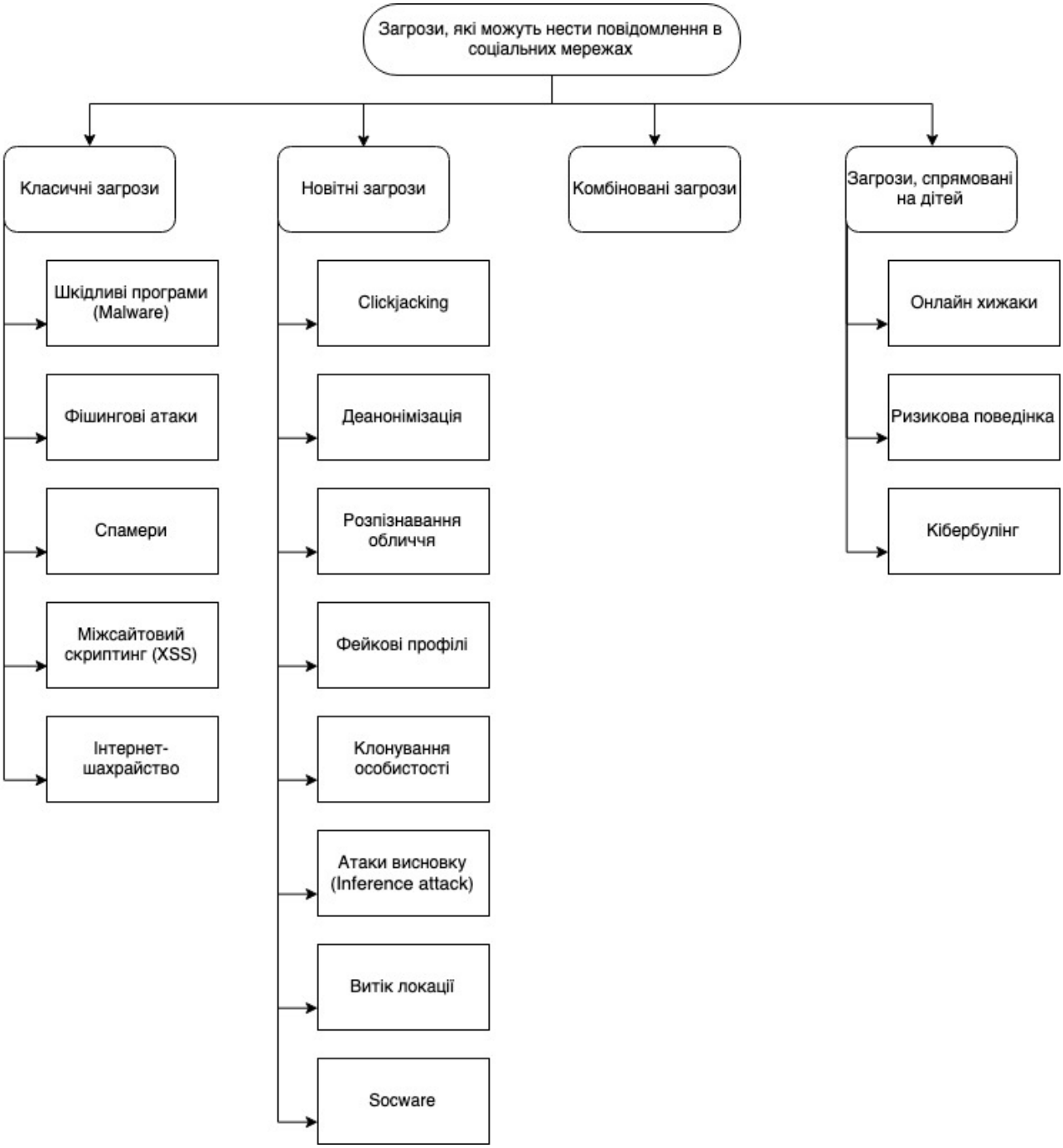

Рисунок 1.1 – Загрози, які можуть нести повідомлення в соцмережах [1]

1.1.1 Класичні загрози

Класичні загрози були проблемою з часів, коли Інтернет отримав широке застосування. Шкідливі програми, спам, міжсайтовий скриптинг (XSS), або фішинг, продовжують бути постійною проблемою. Хоча ці загрози були вирішені у минулому, вони ставали все більш вірусними через структуру і

характер соціальних мереж і можуть швидко поширюватися серед користувачів мережі. Класичні загрози можуть скористатися перевагами особистої інформації користувача, що опублікована в соціальній мережі, оскільки атакується не тільки користувач, а також його друзі, просто змінюючи загрозу вміщувати особисту інформацію користувача.

Наприклад, зловмисник може вставити шкідливий код всередину привабливого спам-повідомлення, в якому використовуються відомості користувача у Facebook [1]. Через особистий характер цього повідомлення, шанси, що користувач відкриє повідомлення і заразиться, є високими. У багатьох випадках ці загрози орієнтовані на основні та повсякденні ресурси користувачів, такі як номери кредитних карт, паролі облікових записів, обчислювальна потужність, і навіть пропускна здатність комп'ютера (для відправки спам-листів).

Тривожно, що ці типи загроз можуть також використовувати інфіковані вкрадені облікові дані користувача для розміщення повідомлень від імені користувача або навіть змінювати особисту інформацію користувача.

Нижче описані різні класичні загрози, а також реальні випадки, коли ці типи загроз поставили під загрозу конфіденційність та безпеку реального користувача.

Шкідливе ПЗ (Malware) в соціальних мережах використовує структуру для свого пропагування серед користувачів та їхніх друзів у мережі. У деяких випадках зловмисне програмне забезпечення може використовувати отримані облікові дані для видавання себе за користувача та надсилати шкідливі повідомлення користувачам, що знаходяться в мережі. Koobface була першою шкідливою програмою, яка успішно розповсюджувалася через соцмережі, такі як Facebook, MySpace і Twitter. Koobface намагається зібрати реєстраційну інформацію та приєднатися до зараженого комп'ютера, щоб бути частиною бот-мережі, так звана "армія зомбі" комп'ютерів, яку часто потім використовують для шкідливої діяльності, наприклад, для надсилання спам-повідомлень та атак на інші комп'ютери та сервери в Інтернеті.

Фішингові атаки - це форма соціальної інженерії для отримання конфіденційної користувацької інформації видаючи себе як третю надійну сторону. Недавні дослідження показали, що користувачі соціальних мереж більш схильні бути ошуканими фішингом через довірливий характер. Більш того, в останні роки фішингові спроби в соцмережах різко зросли. За даними Microsoft Security Intelligence Report [2], 84,5% всіх фішингових атаки націлені на користувачів соціальної мережі. Одна така фішинг-атака сталася на Facebook, заманюючи користувачів на фальшиву сторінку входу. Потім фішинг-атака поширилася серед Facebook користувачів, запрошуючи друзів натискати посилання, розміщене у "життєписі" користувача. На щастя, Facebook зупинив цю атаку.

Спамери - це користувачі, які використовують системи електронних повідомлень для надсилання небажаних повідомлень, наприклад, реклами, іншим користувачам. Спамери використовують соціальні мережеві для відправки рекламних повідомлень іншим користувачам шляхом створення фейкових профілів. Спамери можуть також використовувати соцмережі, щоб додати коментарі до сторінок, які переглядаються багатьма користувачами в мережі. Приклад поширення спаму в мережі можна знайти в Twitter, який постраждав від великої кількості спаму. В серпні 2009 року 11% повідомлень у Twitter були спам-повідомленнями. Однак, до початку 2010 року Twitter успішно зменшив відсоток спам-повідомлень до 1% [3]. Проте, 2013 рік
У статті [3] говориться: "Соціальний спам, як він вже існує в Twitter,
буде продовжувати рости і якщо компанія не зверне увагу на нього
швидше, це може бути єдина річ, яка зруйнує його."

Міжсайтовий скриптинг (XSS) - це напад на веб-застосунки. Зловмисник, який використовує XSS, використовує довіру веб-клієнта у веб-застосунку та змушує веб-клієнт запускати шкідливий код, здатний збирати конфіденційну інформацію. Соцмережі, які відносяться до таких застосунків, можуть страждати від XSS-атак. Крім того, зловмисники можуть використовувати вразливість XSS у поєднанні з інфраструктурою соцмережі та створити черв'як XSS, який може поширюватися вірусно серед користувачів соціальної мережі. У квітні 2009 року

з’явився XSS-хробак Mikeyy, що швидко передавав автоматизовані твіти через Twitter і заразив багатьох користувачів, серед яких і знаменитості, як Опра Вінфрі і Ештон Катчер. Черв'як Mikeyy використовував XSS-вразливість і структуру мережі Twitter, щоб поширюватись через профілі користувачів цієї мережі.

Інтернет-шахрайство, також відоме як кібершахрайство - це використання доступу до Інтернету для шахрайства або отримання переваги над певними людьми. У минулому шахраї використовували очні соціальні мережі, такі як щотижневі збори груп, щоб поступово встановити міцні зв'язки зі своїми потенційними жертвами. Наразі згідно з Північноамериканською Асоціацією Адміністраторів Цінних Паперів, зі зростаючою популярністю онлайн-мереж, шахраї переїхали до онлайнових соцмереж, щоб отримати довіру своїх жертв, для подальшого використання персональних даних, опублікованих в онлайн-профілях жертв. В останні роки, шахраї взламували облікові записи користувачів Facebook, які виїжджають за кордон. Коли шахраям вдалось увійти в обліковий запис користувача, вони звертались до друзів користувача з проханням надати допомогу в переказі коштів на банківський рахунок шахрая. Однією з жертв такого типу шахрайства була Ебігейл Пікетт. Під час подорожі по Колумбії Ебігейл виявила, що її обліковий запис Facebook був захоплений кимось в Нігерії, і його використовували для відправлення запитів на гроші її друзям по мережі, під приводом того, що вона «застрягла» [4].

1.1.2 Новітні загрози

Новітні загрози зазвичай унікальні для середовищ соціальних мереж. Зазвичай ці загрози спеціально орієнтовані на особисту інформацію користувача, а також на особисту інформацію його друзів. Наприклад, зловмисник, який намагається отримати доступ до назви середньої школи користувача Facebook, що є доступним тільки для друзів користувача, може створити підроблений профіль з відповідними деталями та ініціювати запит на дружбу до цільового

користувача. Якщо користувач приймає запит друга, його або її подробиці будуть відкриті зловмиснику. Крім того, зловмисник може збирати дані з друзів користувачів Facebook і використовувати атаку висновку (inference attack) щоб отримати назву школи користувача на основі даних, зібраних у друзів користувача.

Далі буде показано різні новітні загрози та ситуації з реального життя, коли ці типи загроз поставили під загрозу конфіденційність та безпеку користувачів соціальних мереж.

“Викрадення кліку” (Clickjacking): є шахрайською технікою, яка примушує користувачів натискати на щось відмінне від того, що вони хотіли натиснути. Використовуючи clickjacking, зловмисник може маніпулювати користувачем в публікації спам-повідомлень у життєписі Facebook, виконувати дію "мені подобається" без усвідомлення (також називається likejacking) і навіть вмикати мікрофон чи веб-камеру для слідкування за користувачем.

Деанонімізація: у багатьох соціальних мережах, таких як Twitter і MySpace, користувачі можуть захистити свою конфіденційність та анонімність з використанням псевдонімів. Деанонімнізуючі атаки використовують такі методи як відстеження файлів cookie, топології мережі та членство у групах користувачів, щоб розкрити справжню ідентичність користувача.

Розпізнавання обличчя: багато людей використовують соцмережі для завантаження фотографій себе та друзів. Мільйони фотографій щодня завантажуються на Facebook. Більше того, багато фотографій профілю користувача Facebook є загальнодоступними, щоб переглянути та завантажити. Наприклад, завантажена фотографія на профіль у Facebook, дозволена для перегляду більш ніж 1,2 мільярду користувачів. Ці фотографії можна використовувати для створення біометричної бази даних, яку потім можна використовувати для ідентифікації користувачів мережі без їхньої згоди.

Підроблені (фейкові) профілі - це автоматичні або напівавтоматичні профілі, які імітують поведінку людини в соціальній мережі. У багатьох випадках підроблені профілі можуть бути використані для збору особистих даних користувачів у соціальних мережах. Ініціюючи запити друзів іншим

користувачам у соціальній мережі, які часто приймають запити, зловмисники можуть збирати особисті дані користувача, які відкриті лише друзям користувача.

Атаки клонування особистості: використовуючи цю техніку, зловмисники дублюють присутність користувача в тій самій або в інших мережах, для формування довірчих відносин з друзями клонованої особистості. Зловмисник може використовувати цю довіру для збору особистої інформації друзів користувача або виконувати різні типи онлайн-шахрайства. Приклад атаки клонування особистості стався зі старшим командувачем НАТО - адміралом Джеймсом Ставрідісом. Деталі його профілю були клоновані і потім використані для збору даних чиновників міністерства оборони та інших урядових чиновників обманюючи їх, шляхом запиту у друзі від клонованого профілю [5].

Атаки висновку (inference attack): в соціальних мережах використовуються, щоб передбачити конфіденційну інформацію користувача, яку він не збирається розкривати, наприклад, релігійну приналежність або сексуальну орієнтацію. Ці типи атак можуть бути реалізовані з використанням методів інтелектуального аналізу даних у поєднанні з загальнодоступними даними соціальних мереж, такі як топологія мережі та дані від друзів користувача.

Витік локації: зі збільшенням використання смартфонів, які стимулюють обмін інформацією про місцезнаходження, багато людей використовують соціальні мережі, щоб охоче ділитися приватною і іноді чутливою інформацією про їх (або їхніх друзів) поточне або майбутнє місцезнаходження, що може бути використано в різних цілях.

Socware (шкідливі програми націлені на соціальні мережні): тягне за собою фальшиві та, можливо, шкідливі дописи та повідомлення від друзів у соцмережах. Socware може заманити жертв пропонуючи несправжні винагороди користувачам, які встановлюють неблагонадійні застосунки або відвідують сумнівні веб-сайти.

1.1.3 Комбіновані загрози

Сучасні зловмисники можуть також поєднувати класичні та новітні типи загрози, щоб створити більш складну атаку. Наприклад, зловмисник може використовувати фішинг-атаку для збору даних входу у профіль користувача на Facebook, а потім з його профілю розсилати повідомлення друзям, що містять, наприклад, socware. Іншим прикладом є використання клонованих профілів для збору особистої інформації про друзів клонованого користувача. З використанням особистої інформації друзів, зловмисник може надіслати однозначно з її урахуванням повідомлення спаму, що містить вірус. За допомогою особистої інформації, вірус, швидше за все, буде активовано.

1.1.4 Загрози, спрямовані на дітей

Діти, як і зовсім малі так і підлітки, схильні до загроз, описаних вище, як і дорослі користувачі, але є також загрози, які цілеспрямовано і специфічно спрямовані молодших користувачів соціальних мереж. Через критичний характер цієї теми, нижче висвітлено ці загрози, а також описано деякі конкретні висновки з різних досліджень.

Онлайн хижаки: найбільше занепокоєння щодо безпеки особистої інформації дітей викликають педофіли в Інтернеті, які також називаються онлайн хижаками. Лівінгстон і Хаддон [6] з EU Kids Online визначили типологію ризиків та шкоди, пов'язаних з такими видами діяльності в Інтернеті: шкода від вмісту (вплив порнографії або шкідливого сексуального змісту на дитину), шкода від контакту (вплив на дитину, що зв'язалась з дорослим чи іншою дитиною, що намагалась отримати сексуальну вигоду) і шкода від поведінки (дитина, як активний ініціатор ризикованої поведінки). До поведінки, що вважається сексуальною експлуатацією дітей в Інтернеті належить використання дітей дорослими для виробництва дитячої порнографії та її розповсюдження,

споживання дитячого порно і використання Інтернету як засобу ініціювання сексуальної експлуатації в режимі онлайн або оффлайн.

Ризикова поведінка: може включати пряме онлайн спілкування з незнайомими людьми, використання чат-кімнати для взаємодії з незнайомими людьми, розмови інтимного характеру з незнайомими людьми, а також надання приватної інформації та фотографій незнайомцям. Слід зазначити, що в той час як кожна з вищезгаданих поведінкових ситуацій є ризиком, комбінація деяких з цих форм поведінки можуть виправдано викликати величезну тривогу щодо безпеки дитини.

Кібербулінг (кіберзнущання) - це знущання, яке відбувається в технологічних комунікаційних платформах, таких як електронні листи, чати, телефонні розмови та соціальні мережі, зловмисником, який використовує платформу для переслідування жертви, неодноразово посилаючи шкідливі повідомлення, сексуальні зауваження або погрози; публікуючи незручні фотографії або відео жертви; або через іншу неприйнятну поведінку. Сьогодні кіберзнущання стало широко розповсюдженим явищем в соцмережах, в якому зловмисник може використовувати інфраструктуру мережі, щоб поширювати жорстокі чутки про жертву і ділитися незручними фотографіями з мережею друзів жертви [7]. Кібер-залякування зазвичай впливає на дітей, а не дорослих. Недавнє онлайн-опитування, яке включало 18 687 батьків з 24 країн виявили, що 12% батьків стверджувати, що їхня дитина потерпала знущання у кіберпросторі [8]. Крім того, відповідно до результатів опитування, більшість дітей пережили ці знущання на найбільш використовуваних соціальних платформах, таких як Facebook.

Останнім часом також з’явилися загрози радикальних дій, які перед тим як відбутись, сповіщаються виконавцями про своє проведення у соціальних мережах, відкритих каналах у месенджерах. Прикладом можуть слугувати ісламістські угруповання такі як ІДІЛ. Сторонники цієї терористичної організації активно використовують Twitter, канали у Telegram, де активно закликають переходити на свою сторону, а також сповіщають про свої плани на майбутнє. В даній роботі буде наведено приклад боротьби з таким видом злочинної

діяльності, оскільки виявлення таких повідомлень у соціальних мережах, а потім припинення їх поширення, є однією з найактуальніших проблем сьогодення.

1.2 Актуальність задачі виявлення небезпечних повідомлень в соціальних мережах

При аналізі нових загроз користувачам соціальних мереж слід зазначити, що такі мережі широко використовуються як засіб правопорушника або місце для реалізації загроз, про які йде мова в 1.2 протягом тривалого часу. Однією з причин деструктивних процесів у соціальних мережах є те, що формування і розвиток соціальних відносин у цьому новому інформаційному середовищі не регулюються ніякими відповідними правилами або моральними нормами суспільства. Отже, на сьогоднішній день розробка організаційно-правових засад боротьби зі злочинністю в соціальних мережах в нашій країні, яка, на відміну від ряду провідних зарубіжних країн, такого типу правоохоронних органів, практично не має державного значення.

Слід зазначити, що інформація є основним продуктом інтернет-економіки. За даними Boston Consulting, вісім років частка інтернет-послуг у ВВП єврозони досягне 8%, а більшість буде враховувати збір персональних даних і аналіз, який на сьогоднішній день інтернет-компанії перевищує 300 мільярдів. Євро на рік приносить. Європейські компанії все ще заробляють близько 1000 євро на користувача.

Ринок стає не тільки бурхливо зростаючим, але й некерованим. Наприклад, Європейська Комісія розслідує, як Google і Facebook використовують особисту інформацію відвідувачів [9].

Слід зазначити, що практично вся інформація про діяльність терористичної групи була передана у віртуальний світ. Це пов'язано з тим, що він працює безпечніше, ніж традиційні медіа. Основною причиною успіху використання Інтернет-технологій терористичними організаціями є складність виявлення та ліквідація мережевих центрів.

У соціальних мережах люди знаходять одне одного, знайомляться, спілкуються, беруть участь в обговореннях і об'єднуються для формування груп інтересів. Проте інтереси різні. І один з найнебезпечніших «інтересів» - наркотики - не обійшли соціальних мереж. Користувачі соціальних мереж активно створюють групи, які сприяють наркоманії.

У таких групах вони пропонують наркотики, надають консультації щодо можливостей купівлі, дають адреси наркоторговців, проводять кампанію за легалізацію наркотиків, підтримують новачків у всіх відношеннях, друкують статті про наркотики - історію, вплив, характеристики, методи виробництва. Ці статті, до речі, написані науковою мовою з посиланням на конкретні факти та статистику. Тобто, їх навчають фахівці, чия робота полягає в тому, щоб робити явну або замасковану пропаганду наркотиків у соціальних мережах.

Слід також зазначити, що кількість злочинів, скоєних за допомогою інформації, викраденої з соціальних мереж, постійно зростає. Аналітики також вважають, що однією з причин збільшення таких злочинів є збільшення кількості дітей, підлітків та підлітків у соціальних мережах. На жаль, вони навряд чи знають про наслідки відкритості та гіперкомунікації.

Представник «Київстар» зазначив, що, за його даними, 40% українських дітей у соціальних мережах надають особисту інформацію про себе та свої сім'ї, а 60% дітей зустрічаються в реальному світі зі своїми віртуальними друзями без відома батьків [10].

Отже, для використання з метою злочиннної діяльності Інтернет поповнюється протиправним контентом. Хоча не можна сказати, що більшість соціальних мереж використовуються в незаконних цілях, тим не менш очевидно, що збільшення загальної кількості користувачів помножує потенціал злочинця, а також їхній інтерес до незаконного, деструктивного використання соціальних мереж. Потрібно боротись з цим. Коротко обговорюватиметься досвід окремих країн у вирішенні таких проблем.

Державні органи, комерційні структури, громадські організаціямї та громадяни закордоном займаються протидією правопорушенням у соціальних мережах за кордоном.

У Росії працює сайт з реєстром протизаконних сайтів у Рунеті – http://eais.rkn.gov.ru/. До чорного списку вносяться сайти, що містять пропагандистську інформацію про наркотики, заклик до суїциду чи екстремістської діяльності. Ці сайти блокуються без проведення судового процесу, якщо дочасно не буде видаленою суспільно-небезпечна інформація. Також, до реєстру потрапляють такі сайти, за якими було винесено судове рішення про порушення.

Створює і поповнює цей список сайтів Роскомнагляд, приймати рішення про включення до реєстру сайтів можуть також МВС РФ, Федеральна служба з контролю за обігом наркотиків і Росспоживнагляд.

Якщо власник веб-сайту не видаляє інформацію протягом 24 годин після запиту постачальника послуг, дані мережевих ресурсів потраплять до реєстру, а постачальник повинен блокувати доступ до цього сайту. У цьому випадку блокування на IP-адресу буде останньою інстанцією [11].

У Південній Кореї, де сьогодні найбільша кількість випадків самогубств (близько 40 самогубств на день), уряд сформував спеціальну групу експертів, понад 100 осіб, які шукають самогубців-користувачів соціальних мереж і сайтів людей для здійснення таких актів [12].

У 2007 році в Австралії було оголошено план встановлення інтернет-фільтрів, які мали справу з жорстокими сценами, детальними інструкціями щодо вчинення злочинів або терористичних актів, а також використання наркотиків. Але плани австралійських властей викликали критику з боку правозахисників, які вважали, що фільтри в Інтернеті призведуть до цензури в країні.

Водночас слід зазначити, що державні органи окремих країн під приводом боротьби зі злочинами в кіберпросторі намагаються законодавчо закріпити порушення прав людини, а також міжнародного права.

Зокрема, в Нідерландах міністр безпеки і правосуддя І. Опстелтен представив у листі до парламенту план створення законопроекту про збір доказів у процесі розслідування кіберзлочинності. Зокрема, новий закон повинен дозволити слідчому встановити шпигунські програми на підозрілих комп'ютерах, а також "шукати" ці комп'ютери у віддаленому режимі. Слідчі

повинні отримати право видаляти шкідливий вміст, якщо вони його знайдуть під час "пошуку".

Проте цей закон планує надати повноваження встановлювати шпигунські програми на комп'ютери зловмисників не тільки в межах країни, але і за кордоном, за умови, що місцеположення не встановлено. Це небезпечна формулювання, тобто неможливо встановити розташування будь-якої особистої установи через мережу Інтернет зі стовідсотковою точністю.

На думку голландських правозахисників, це створює проблему, а інші країни можуть приймати подібні законодавчі акти, і тоді в Інтернеті почнеться хаос. Хоча експерти відзначають, що такий прецедент вже відбувся з боку Сполучених Штатів у застосуванні цифрової зброї в інших країнах [13].

Прикладом участі громадських організацій у пошуку в Інтернеті змісту, що може становити загрозу для молодих людей і його автора може бути притягнуто до кримінальної відповідальності може служити німецькою організація захисту молоді (jugendschutz.net), яка фінансується всіма землями Німеччини.

Якщо експерти jugendschutz.net знають, що зміст перевірено і він несе в собі негатив, вони намагатимуться видалити його якомога швидше, за умови, що провайдери готові це зробити. У 2011 році, за розрахунками і ініціативами jugendschutz, у різних соціальних мережах було понад 970 разів видалені повідомлення, що несуть в собі неонацистський зміст.

Однак більшість правого контенту часто перезавантажується після видалення - і це є випадком, коли постачальники повинні брати на себе відповідальність. Адже технічно можливо виявити і запобігти перезавантаженню однакових матеріалів [14].

Численні способи боротьби зі злочинністю, що використовують засоби боротьби зі злочинністю та засоби, що використовуються виключно в кіберпросторі, вже давно використовуються працівниками правоохоронних органів ряду держав. Там правоохоронні органи прагнуть використовувати нові технології, у тому числі отримані через наявність соціальних мереж, для виявлення, виявлення, виховання та запобігання злочинності. Вони постійно

контролюють підозрілі блоги, чати, веб-сайти тощо, щоб отримати інформацію, щоб отримати потрібну для правоохоронців інформацію.

Зокрема, сьогодні соціальні мережі широко використовуються правоохоронними органами за кордоном для спілкування з громадськістю, також з метою отримання інформації про злочини. Наприклад, у листопаді 2007 року співробітники Приморського районного управління внутрішніх справ Санкт-Петербурга відкрили групу «Злочинності в Приморському» в соціальній мережі «Вконтакте», яка до цих пір користується великою популярністю. Група містить корисну довідкову інформацію, відеозаписи, які фіксують окремі злочини, фотографії розшукуваних людей та інформацію про зареєстровані злочини. Користувачі оцінюють реалізацію сторінки позитивно і надають допомогу. Через громадську підтримку працівники правоохоронних органів досягли значних результатів у виявленні місць перебування та арешті розшукуваних осіб [15].

У Великобританії правоохоронні органи також використовують соціальні мережі як засіб зв’язку з громадянами. Так, поліція округу Великий Манчестер створила обліковий запис Twitter. У мікроблозі цього ресурсу публікуються важливі новини, відомості та особисту інформацію людей, яких оголсили в розшук. Водночас правоохоронні органи Великої Британії офіційно визнали важливу роль соціальних мереж у запобіганні та виявленні злочинів, а також включили відповідний курс у програму навчання молодих працівників.

Поліція вважає, що соціальні мережі можуть бути особливо корисними для розкриття злочинів, пов'язаних з шантажем і насильством в сім'ї [16].

Дані висновки свідчать про перспективу використання соціальних мереж в протидії деструктивній поведінці.

Моніторинг соціальних мереж новим відділом поліції в Нью-Йорку є досить ефективним. Завданням цього підрозділу є моніторинг винних у соціальних мережах. Співробітники вже арештовували близько 50 членів злочинних угруповань після перегляду їхньої діяльності в соціальній мережі Facebook [17].

Цікавими є програми, розроблені вченими для виявлення незаконних дій і допомоги правоохоронним органам.

Наприклад, швейцарські вчені - дослідницька група з лабораторії аудіовізуальної комунікації Швейцарської федеральної політехнічної школи в Лозанні - розробили алгоритм, що допомагає працівникам правоохоронних органів. За допомогою цього алгоритму можна визначити джерело соціально небезпечної інформації, яка масово дописується в соціальних мережах. Цей метод може також контролювати терористичні атаки, опозиційні політичні заходи, спам, шкідливі програми та комп'ютерні віруси.

На думку вчених, цей метод може бути використаний для пошуку джерела інформаційного сигналу в масиві даних, що циркулюють в соціальній мережі, до якої належить людина. Після того, як проаналізовано фактор часу і деякі інші доступні параметри повідомлення, що були надіслані лише від 15 до 20 учасників, алгоритм відновлює історію поширення цієї інформації та знаходить її джерело [18].

Соціальна мережа Facebook використовує автоматичні алгоритми для сканування чатів та іншої особистої інформації від користувачів у США для пошуку злочинів і виявлення їх на ранній стадії. В основному система налаштована на пошук педофілів, але також може бути налаштована для пошуку ознак інших злочинів, таких як: наприклад, для обговорення купівлі наркотиків, зброї та інших заборонених дій.

Система сканує переписку та дописи користувачів Facebook, а коли виявляє підозрілу активність, відображає профіль і надсилає його спеціалізованому відділу Facebook. Співробітник даного відділу, у свою чергу, оцінює масштаби потенційної небезпеки і, якщо є, повідомляє про правопорушника правоохоронні органи Сполучених Штатів.

Система налаштована з дуже низьким відсотком помилкових спрацьовувань, щоб гарантувати, що особиста переписка між законослухняними користувачами мережі не прослідковується.

За словами представника соціальної мережі, програмне забезпечення, що використовується для моніторингу дій користувачів, зосереджується на діалогах між користувачами з "поганими" зв'язками.

Наприклад, якщо два користувачі не є друзями або нещодавно стали друзями, у той час як у них немає спільних друзів, а інші друзі взаємодіють з користувачем і один з одним надзвичайно рідко, а також якщо два користувачі мають велику різницю у віці, алгоритм програмного забезпечення стає "зацікавленим" у цьому спілкуванні та повідомляє уповноваженої особи на Facebook.

Іноземні правоохоронні органи та спецслужби використовують власні програми для моніторингу щоденних підозрілих блогів, форумів і веб-сайтів, а також отримання великої кількості корисної інформації. Хоч збирання та переробка інформації приносить свою користь, питання ефективності залишається відкритим. Наприклад, велика кількість джихадистських груп напередодні вбивства посла США С. Стівенса у Лівії обговориювали плани нападу на американські дипломатичні місії в Лівії, Єгипті та Алжирі, повідомляє Associated Press. Тим не менше, США не змогли запобігти смерті чотирьох американських дипломатів [19].

Цікавою є розробка веб-сайту Facebook - Connected to the Case, яка допомагає поліції розслідувати злочини з допомогою звичайних громадян - вирішення соціально значущих завдань шляхом волонтерської роботи, Користувачі можуть увійти на веб-сайт, ввівши свій обліковий запис соціальної мережі. Служба отримує дані з профілю на Facebook і визначає, яка особа може проводити розслідування, у яких випадках допомагати поліції. Система запускає кілька ключових тегів для кожного користувача і дізнається, з яких місцях і коли він відвідував, з ким він спілкувався, і так далі.

Потім система порівнює цю інформацію з базою даних нерозкритих правопорушень, щоб визначити, який користувач міг бачити. Крім того, Connected To The Case може надсилати анонімні повідомлення про скоєння злочину [20].

Для моніторингу діяльності в соціальних мережах в Росії була розроблена і використана система моніторингу соціальних мереж «Призма».

Система «Призма» відстежує 60 мільйонів джерел у реальному часі. Вона показує динаміку позитивних і негативних відгуків в блогах для однієї чи іншої

події, а також може створювати графіки для ботів. Водночас відстеження тем моніторингу регулюється індивідуально.

"Призма" контролює діяльність соціальних медіа, що призводить до збільшення соціальної напруженості: проникнення громадських заворушень, протестних настроїв, екстремізму та інших. [21]

Подібна система була розроблена в Україні під керівництвом доктора фізико-математичних наук, професора, члена-кореспондента НАН України А.В. Анісімова. Перша версія системи моніторингу соціальної активності Twitter вже працює. У ній буде збиратися інформація про хід соціальних процесів і явищ через їх представлення в соціальній мережі Twitter. Статистика, зібрана системою, для подальшої обробки експертами. Система може візуалізувати зібрані дані у вигляді діаграм. Збір інформації базується на ключових словах і фразах.

Крім збору даних, система може створити свій первинний семантичний аналіз за шкалою полярностей: позитивний, нейтральний, негативний. Крім того, система містить модуль кластеризації, який використовується для формування подібних текстових груп, які можуть бути використані для ідентифікації найбільш важливих подій в рамках дослідження, що проводиться протягом певного періоду часу.

На жаль, подальший розвиток припинено з фінансових причин.

Сьогодні Україна відстає від цих соціально позитивних процесів. Частково це пов'язано з тим, що правоохоронні органи майже не мають спеціальних систем збору інформації, зокрема моніторингу контенту, аналізу контенту, достатньої кількості фахівців, навчених у цьому напрямку, а також відсутності нормативних прав, повноважень та обов'язки окремих правоохоронних органів нашої держави з метою реалізації відповідних заходів щодо запобігання кіберзлочинності. Слід також закріпити думку, що однією з причин неефективного правового впливу на сучасний кіберпростір є відсталість методів і засобів практичної реалізації існуючої правової бази правоохоронними органами [22, с.15].

На тлі нинішнього стану незаконного використання соціальних мереж та перспективного збільшення кількості та соціальної загрози реальних і

потенційних загроз з боку кібернетичного простору у найближчому майбутньому необхідно створити ефективну систему, яка має бути спрямована на локалізацію та протидію таким деструктивним явищам. Загрози можливі лише тоді, коли правоохоронні органи нашої держави, спільно з відповідальними іноземними органами влади, безпосередньо використовують певні можливості самих соціальних мереж. В результаті, працівники правоохоронних органів повинні постійно використовувати інструменти та інструменти для виявлення, розкриття та запобігання соціальним мережам у своїй діяльності.

В Україні в даний час немає законів, які б регулювали таку діяльність в цілому, і конкретних методів, спрямованих на цільову аудиторію, зокрема.

Однією з найефективніших заходів щодо боротьби з незаконною діяльністю і тим самим захистом прав і свобод наших громадян має бути так званий моніторинг, започаткований правоохоронними органами, функцію якого законодавець має покласти, із відповідним визначенням компетенції, на правоохоронні органи, наділені правом здійснення оперативно-розшукової діяльності, у першу чергу, на поліцію та Службу безпеки України.

Висновки до розділу 1

В даному розділі описується, що соціальні мережі стали частиною повсякденного життя і, в середньому, більшість користувачів Інтернету витрачають більше часу на соціальні мережі, ніж на будь-яку іншу онлайн-діяльність. Людям подобається користуватися соціальними мережами, щоб взаємодіяти з іншими людьми через обмін повідомленнями та відео. Тим не менш, соціальні мережі мають темну сторону, що наповнена шахраями і онлайн хижаками, які використовують соціальні мережі як платформу для приманки жертв. У даному розділі було представлено сценарії, які загрожують користувачам соціальних мереж і можуть поставити під загрозу їх ідентичність, приватність і добробут як у віртуальному світі, так і в реальному житті. Крім того, було надано приклади багатьох представлених загроз для того, щоб

продемонструвати, що ці загрози є реальними і можуть загрожувати кожному користувачеві. Підкреслено певні загрози, які загрожують безпеці дітей та підлітків у кіберпросторі.

Було наведено приклади способів боротьби зі злочинністю в соціальних мережах як на рівні держави, так і приватних корпорацій. Визначено, що для України - це дуже актуальна проблема, оскільки не існує законодавства, яке б здійснювало регулювання такої діяльності взагалі, та спеціальних методів “дослідження цільової аудиторії”, зокрема.

2 ОСОБЛИВОСТІ РОЗРОБКИ СИСТЕМИ ПОШУКУ ПОТЕНЦІЙНО НЕБЕЗПЕЧНИХ ПОВІДОМЛЕНЬ В СОЦІАЛЬНИХ МЕРЕЖАХ

2.1 Аналіз мети інтернет-моніторингу

Розміщення в Мережі негативної інформації може здійснюватися різними способами: багато користувачів вважають за краще створювати спеціалізовані сайти; сайти "Громадської думки", на яких споживачі висловлюють невдоволення на адресу певного виду продукції або послуг (consumer opinion sites); "електронні плітки": саме за допомогою Інтернет поширюються найнеймовірніші чутки, "качки" та інша недостовірна інформація; а також створювати різні групи в соціальних мережах, блогах, форумах. І саме Інтернет дозволяє їм поширюватися по всьому світу з неймовірною швидкістю. Незважаючи на те, що більшість з них з першого погляду сприймаються як повна нісенітниця, вони залишають незгладимий слід в душі користувачів, що в результаті не може не завдавати шкоди компаніям-виробникам.

В результаті компанії змушені захищати свою репутацію і торгові марки шляхом моніторингу (регулярного відстеження) інформації для з'ясування ринкових переваг, коригування недостовірної інформації та поліпшення сервісу обслуговування споживачів. Якісно розроблена система моніторингу може забезпечити компанію інформацією, яка б дозволила заздалегідь виявити небезпечні моменти шляхом відстеження стану ринку, при якому особлива увага приділяється вивченню настроїв споживачів багато в чому за рахунок вивчення різних Інтернет-публікацій, що допомагає мінімізувати кількість скарг, що надходять і запобігає багатьом судові розгляди .

Система Інтернет-моніторингу дозволяє якомога раніше зрозуміти можливі проблеми і спробувати мінімізувати негативні наслідки при економному витрачанні тимчасових ресурсів клієнтів.

Інтернет-моніторинг вельми ефективний при проведенні антикризових заходів, бо останнім часом реакція громадськості в першу чергу проявляється саме в Інтернеті, так що своєчасне виявлення електронних публікацій подібного роду дає можливість оперативно відстежити тенденції, які намічаються і вжити необхідних заходів.

Інтернет-моніторинг також є найважливішою частиною досліджень політики існуючих та потенційних конкурентів і тенденцій розвитку галузі в цілому.

Також слід зазначити, що багато в чому саме моніторинг дозволяє відстежити канали витоку стратегічно важливої інформації про власну компанію.

Моніторинг соціальних мереж - спеціально організоване, систематичне спостереження за станом соціальних мереж, явищ і процесів, що відбуваються в даному середовищі, з метою їх оцінки, контролю і прогнозу.

Постійний моніторинг мережі Інтернет, блогосфери і форумів - перший крок з протидії інформаційному нападу.

Виявлення початку інформаційної війни на ранніх стадіях значно знижує витрати на протидію інформаційній атаці.

Слід чітко розуміти той факт, що активність користувача в соціальних мережах може впливати на його подальше життя, кар'єру і т.д .. Внаслідок цього почали з'являтися компанії, що аналізують активність користувача в соціальних мережах.

Джоел Джевітт, віце-президент компанії Rapleaf, яка спеціалізується на інтелектуальному аналізі даних, стверджує, що може вивчити список контактів будь-якої людини в соціальних мережах і передбачити, які рекламні оголошення привернуть його увагу, наскільки ризиковано видавати йому гроші або кредитну карту. І зламувати акаунти не треба - достатньо відкритої інформації[6].

«Виходячи з того, з ким ви спілкуєтеся, можна зробити практичні висновки про ваші звички, - говорить Джевітт. - Це новий тип інформації. Її дослідження поки на ранній стадії, але результати є вже зараз ».

Rapleaf - один з безлічі стартапів, що ведуть роботу в області моніторингу соціальних мереж (SMM), нового і зростаючого напрямки бізнесу.

Основний масив даних аналітики отримують на таких сайтах, як Twitter, Facebook і MySpace. Тут можна зустріти справжнє розмаїття користувацького контенту, де зміни статусу, фотографій і будь-яка інша інформація про політичні, релігійні або сексуальні переваги, якщо користувач її Спеціально не закрив, доступна всім. Більша її частина - безглузде сміття: повідомлення про настрої, коментарі з купою смайликів і інтернет-акронімів.

Але в моніторингу соціальних мереж немає такого поняття, як «занадто багато інформації».

До недавнього часу такі дані в основному використовувалися для «управління репутацією» і допомагали брендам, рекламним агентствам і PR-компанії з'ясовувати, що ми про них думаємо. Але обсяг користувацького контенту зріс, технологія для збору, фільтрації та аналізу була вдосконалена, і тепер фірми перевернули завдання - вони хочуть знати, що призначені для користувача дані говорять про самих користувачів.

Надходить все більше і більше підтверджень того, що персональні контакти можуть розповісти про людей все - від його ваги до рівня щастя. Дослідники з університету Вікторії недавно запропонували використовувати SMM, щоб «виявити і стежити за тими блогерами, які знаходяться в депресії, можуть накласти на себе руки або нанести шкоду собі або іншим». Навіть ЦРУ і СБУ стали робити спроби дослідити соціальні мережі в інтересах громадської безпеки.

Інтернет з моменту своєї появи дуже швидко став одним з найбільш популярних джерел інформації. Всього кілька років знадобилося глобальної З ети для того, щоб завоювати аудиторію в кілька мільйонів: для досягнення тих же показників популярності радіо і телебаченню знадобився не один десяток років. Безумовно, Інтернет - це дуже зручне і ефективний джерело інформації. Мільйони людей по всій земній кулі за допомогою Інтернету щодня, щогодини і щохвилини спілкуються, працюють, отримують найактуальніші новини та інформацію. Інтернет сьогодні глобальний. Саме найширша поширеність

Інтернету і його загальна доступність стали тими факторами, завдяки яким Інтернетом користуються не тільки в позитивних цілях, але і в цілях корисливих. Інтернет - злочинність (кіберзлочинність) сьогодні стає реальним джерелом суспільної небезпеки, сучасним відгалуженням злочинності.

Основний контингент інтернет-користувачів - це представники молодіжного середовища. Молоді люди - активні користувачі глобальної мережі, які отримують левову частку інформації і саме з інтернет-простору. Важливим фактором в даному випадку є той факт, що світогляд молодих людей перебуває ще на стадії становлення і розвитку. Тому Інтернет з його спектром думок і поглядів, і може представляти реальну небезпеку. Особливо дане судження вірно відносно тих порталів та інтернет-джерел, які створені для поширення ідей тероризму і релігійно політичного екстремізму.

В даний час з урахуванням розвитку глобальної інформатизації екстремізм і тероризм тільки посилює свої позиції. Держави всього світу, порушені небезпечним соціальним і політичним явищем, протистоять екстремізму і тероризму різними методами і способами, як в реальній, так і у віртуальній дійсності.

Філігранно використовуючи інтернет-ресурси, особливо соціальні мережі, різні екстремістські і терористичні групи і організації поширюють в мережі свої програми, ідеї і думки, пропаганду і інформацію екстремістського толку. Ведучи інформаційну атаку, особи, які сповідують екстремізм і тероризм, дезінформують молодих людей з несформованим світоглядом, повних в силу віку протесту і незгоди з деякими реаліями. Інтернет-екстремісти і терористи залучають молодь, використовуючи і направляючи їх енергію в деструктивне русло, піддаючи агітації і пропаганді. І варто визнати, їм вдалося домогтися певного успіху. Тому боротьба з інтернет-екстремізмом і тероризмом сьогодні як ніколи актуальна і важлива. Якщо ми не будемо приділяти належної уваги поширенню в Інтернеті екстремістських і терористичних ідей і концепцій, то в підсумку суспільство може стати плацдармом для трагічних негативних подій, котрі мають місце з 2011 року і до цього дня в країнах Близького Сходу і

Північної Африки. Інтернет-тероризм і екстремізм - реальне явліня, що представляє небезпеку для молодого покоління.

Сьогодні соціальні мережі - це реальний плацдарм, на якому повертається активна політична і соціальна боротьба. Боротьба сфер впливу та інтересів. У політологів існує офіційний термін «твіттер-революція». Так, подібний термін стосується масових заворушень в Кишиневі в 2009 р, акції тощо протесту в Ірані в 2009 р, революція в Тунісі в 2010-2011 рр., Революції в Єгипті в 2011 р та й взагалі до всього «комплексу» подій, які іменуються «арабська весна». Тобто політичними дослідниками офіційно визнається, що соціальна мережа Твіттер (Twitter) - це реальний інструмент, який вміло використовують протестувальники, що створюють заворушення і т.п. в різних державах. В одних державах Твіттер менш ефективний при організації суспільстві н них хвилювань, в інших більше. Так, на відміну від Тунісу, де Твіттер не мав вагомого впливу, в Єгипті він зіграв важливу роль в організації виступів проти режиму Хосні Мубарака.

У цій країні культура використання соціальних медіа є набагато більш поширеною і більш розвиненою, ніж в Тунісі. Влада відповіла на виступи, віддавши мобільним операторам і заборонила пересилку смс. Те ж згодом зробив в Лівії Муаммар Каддафі. В Єгипті інформація про місце, час проведення демонстрацій поширювалася Фейсбук і Твіттер. Тоді режим Мубарака - як раніше Бен Алі в Тунісі - відключив Інтернет і 3G-мережу. Після чого соціальні мережі замінив аналоговий еквівалент Твіттера - плакати з інформацією про те, де і коли збиратися на наступний день, які демонстранти тримали над головою.

Facebook неодноразово стикався з критикою через дії терористів. У 2014 році влада Британії звинуватили Facebook у тому, що він не допоміг спецслужбам заздалегідь знайти екстремістів, по-звірячому вбили військовослужбовця Лі Рігбі в травні 2013 року: соцмережа блокувала акаунти терористів, але не повідомляла про це правоохоронцям. У Німеччині Facebook звинуватили в невидаленого ультраправої пропаганди. У Франції вбив поліцейських терорист вів трансляцію своїх дій в соцмережі. За день до цього влаштував бійню в клубі Pulse (Орландо, Флорида) Омар Мартін продовжував

користуватися Facebook під час теракту і шукав новини про себе, а до того він написав кілька постів про ісламську загрозу і з вимогами до Росії і США перестати бомбити позиції бойовиків «ісламської держави». Facebook нерідко допомагає спецслужбам. У липні 2016 року соцмережа передала бразильської поліції переписку бойовиків, які готували теракти на Олімпіаді-2016. У тому ж році Facebook і Twitter видалили аккаунти радикальної палестинської організації «Хамас».

Google теж неодноразово звинувачували в потуранні діяльності терористів. У червні The Times повідомила , що Салман Абеді, який влаштував вибух на концерті в Манчестері, дивився на YouTube відео про виготовлення бомб. Наприклад, в травнем 2017 роки родичі жертв теракту в Сан-Бернардіно подали в суд на Twitter, Facebook і Google. Нібито недбалість адміністрації сервісів допомагала радикалізації подружжя Сайеда Ризваном Фарука і Ташфін Малик, які влаштували теракт. У грудні 2016-го родичі жертв стрільця з Орландо подали в суд на ті ж компанії, тому що вони «надавали терористичному угрупованню ІД облікові записи, які воно використовувало для розповсюдження екстремістської пропаганди, збору фінансових коштів і вербування нових прихильників». Компанії-відповідачі назвали звинувачення «безглуздими і маніпулятивними».

Останні події червня 2013 р. в Туреччині також проходять мало не під егідою твіттер-революції. Координація дій проти протестуючих також відбувається через соціальні мережі. Лабораторія соціальних мереж за участі при університеті Нью-Йорка провела дослідження задіяння Твіттера в турецьких заворушеннях і прийшла до висновку, що його вплив досяг феноменального рівня.

Які висновки можна зробити зі сказаного? По-перше, якщо у випадку з усіма попередніми громадськими заворушеннями можна було сміливо відстежувати «руку Вашингтона», спецслужби якого використовували соціальні мережі (в першу чергу Твіттер) для створення сприятливого міжнародного фону під події в Алжирі, Лівані, Тунісі, Ємені, Єгипті та Лівії, в Туреччині відбувається радикальне зміщення акцентів. Твіттер використовується в

Туреччині саме для рекогносцировки: куди перемістилися загони поліції, які їхні сили, в яких районах, кварталах, вулицях скупчилися протестують, скільки їх, які шанси на успішне протистояння і т.п.

Турецький Твіттер не просто протистоїть офіційним ЗМІ, які сьогодні інтенсивно фільтрують інформацію про суспільство н них заворушеннях, а безпосередньо нав'язує себе суспільству в якості повної альтернативи. 1 червня набрав оборот хештег # BugünTelevizyonlarıKapat («Вимкни сьогодні ТВ!»),

Даний приклад використання Twitter, ефективності його використання, масовість і чисельність користувачів даної соціальної мережі показує, як зручно, швидко, а головне - продуктивно можна використовувати даний ресурс в досягненні поставлених соціальних, політичних і будь-яких інших цілей. В цьому і цінність, і величезна небезпека Твіттера. Тепер уявіть, що якесь екстремістське угруповання вийшло на стежку агресивної війни. І щоб залучити до своїх лав прибічників і прихильників якомога більше людей, почало використовувати Твіттер.

2.2 Огляд наявних методів боротьби з потенційно небезпечними повідомлень в соціальних мережах

Ще в травні 2016 року Facebook, Google, Twitter і Microsoft підписали так званий поведінковий кодекс, придуманий Єврокомісією. Його головна мета - боротьба з мовою ворожнечі в ЄС. Компанії зобов'язалися видаляти всі повідомлення про «розпалюванні ненависті» і з «мовою ворожнечі» протягом доби. Також вони пообіцяли ділитися з правоохоронними органами Євросоюзу інформацією про те, як вони борються з вербальним екстремізмом, щоб поліція і спецслужби самі навчилися боротися з агресією в мережі.

Alphabet (материнська компанія Google) в 2015 році разом з американським Інститутом стратегічного діалогу (ISD) почали боротися з тероризмом через рекламу в пошуковій видачі. Її бачили ті, хто шукав інформацію про ісламський радикалізм.

YouTube, що належить Alphabet, разом з Twitter і Facebook в 2016 році брав участь в експериментах того ж інституту. Метою було дізнатися, який тип повідомлень і таргетингу дозволить вийти на потенційних екстремістів до того, як їх погляди радикалізуються. Сервісів вдалося серйозно поліпшити показники перегляду антитерористичних відео, однак представники Інституту стратегічного діалогу не можуть сказати, скількох людей вони реально змогли врятувати від перетворення в ісламських терористів.

Нещодавно Google разом з ISD заявили, що роздадуть $ 1,3 млн проектам, націленим на вирішення проблеми тероризму в Великобританії. Одночасно представники Google повідомили, що компанія буде інвестувати в технології, що перешкоджають поширенню екстремістської пропаганди в інтернеті.

У червні 2017 року Facebook заявив, що буде розвивати штучний інтелект для виявлення і видалення терористичного контенту. Однак до цієї ініціативи є чимало питань: наприклад, що саме буде вважатися «тероризмом», звідки візьмуть базу зображень, чи буде до результатів роботи штучного інтелекту доступ у некомерційних організацій, журналістів і т.п. У ті ж дні про посилення боротьби з терористичним контентом повідомили в Google і YouTube. Спеціальний алгоритм буде знаходити відповідні відео і не давати їх завантажити вдруге. Twitter теж поліпшила свої алгоритми і відзвітувала , що за перші шість місяців 2017 року видалила 300 000 облікових записів терористів.

2.2.1 NETRA

У рамках зусиль уряду Індії щодо розробки системи моніторингу в Інтернеті, дві різні установи, ЦШІР і НОТД, отримали завдання на розробку технічних систем, які будуть зосереджені на скануванні даних через Інтернет та виявленні підозрілих слів[23]. Приватна організація - Paladion, взяла на себе відповідальність за розробку системи, допомагаючи НОТД, тоді як ЦШІР створила внутрішню команду з 40 членів для розробки системи NETRA. Система

моніторингу та нагляду Vishwarupal розроблювана НОТД, зіткнулася з декількома проблемами, такими як участь деяких зовнішніх приватних компаній у питаннях безпеки, а також спроможність і компетентність НОТД в управлінні такою системою незалежно без допомоги Paladion. Поряд з цим, інші платформи було започатковано урядом, як NATGRID, створений урядом Індії для моніторингу терористичних операцій шляхом розробки структури для посилення боротьби з тероризмом в Індії. Крім того, тестування Vishwarupal шляхом дослідження та аналізу не закінчилося задовільно, посилаючись на причину, що система давала збої на регулярній основі. З іншого боку, система NETRA була розроблена і експлуатується групою вчених, призначених урядом, і жодна частина її операції не включала жодних зовнішніх або приватних агентств, що ведуть до страхування безпеки. Тестування NETRA здійснювалося Службою розвідки, яка була задоволена роботою NETRA та постійними інвестиціями ЦШІР у його крило досліджень і розробок, щоб задовольнити зміни технологій. Таким чином, NETRA отримала перевагу над Vishwarupal як офіційна система моніторингу Інтернету уряду Індії.

В даному контексті аналіз трафіку може аналізувати потік даних через Інтернет, навіть коли передані повідомлення шифруються таким чином, що важко розшифрувати. NETRA моделюється для обробки та відфільтровування контенту, який генерується. Вона спрямована на виявлення і виявлення слів уловлювання (як показано на рис. 2.1), таких як "атака", "бомба", "вибух", "вбити", "джихад", "вбивство", "терорист" серед інших подібних.

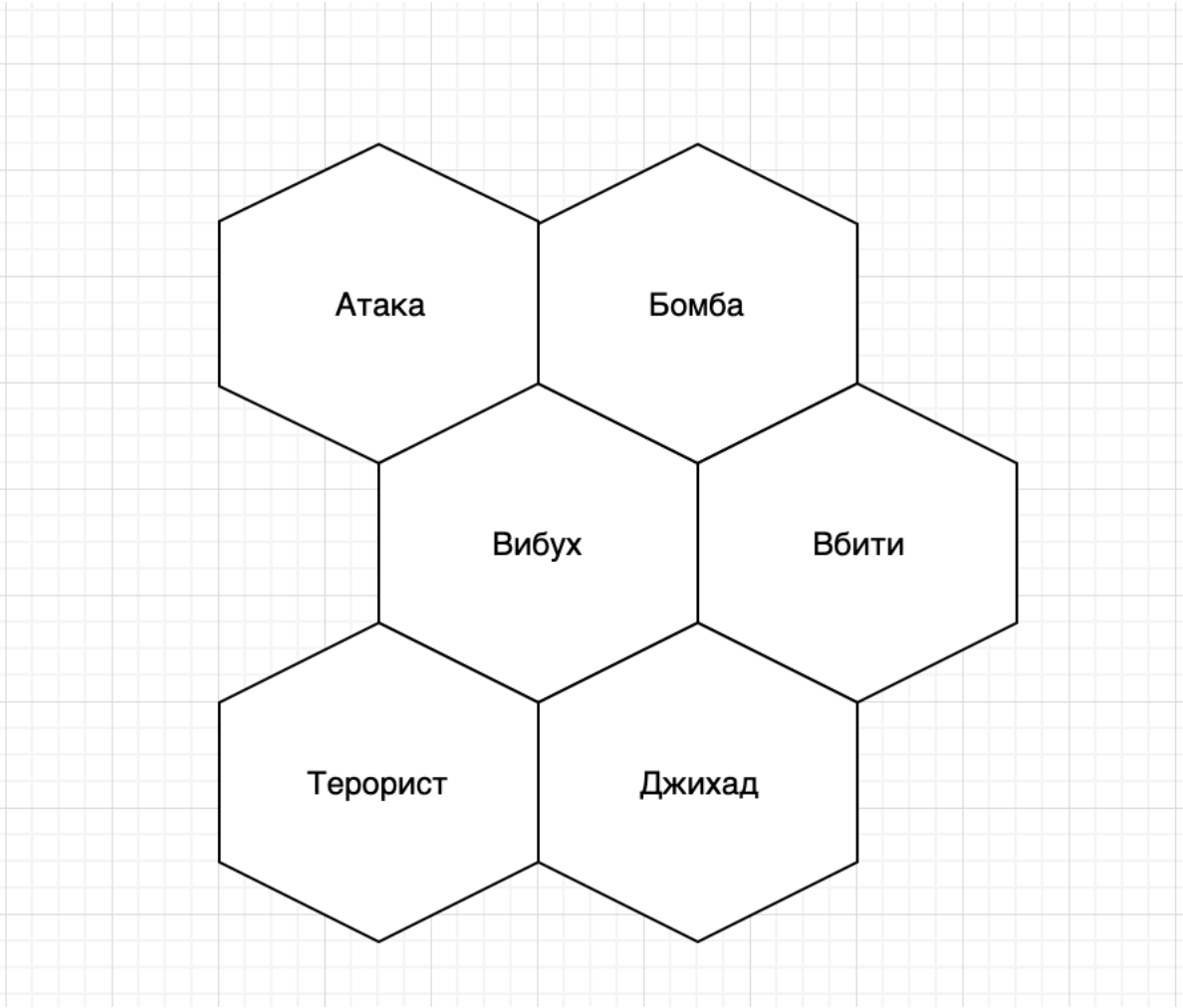

Рисунок 2.1 – Ключові слова, що відслідковуються системою NETRA

NETRA запрограмована повідомляти про будь-яке використання такого слова та IP-адреси, з яких було використано дане слово, відповідним органам та установам, які можуть негайно вжити відповідних заходів. Система спрямована на моніторинг переважно терористичної діяльності і призначена для реалізації в трьох основних робочих зонах, які включають основні органи безпеки країни, Секретаріат Кабінету Міністрів, ВДА та Службу розвідки. Останні дві - зовнішні та внутрішні розвідувальні органи країни. Виділення 300 ГБ пам'яті було зроблено для згаданих трьох агентств безпеки для зберігання інтернет-трафіку, який вони перехоплюють для аналізу. Ще 100 ГБ було призначено для того ж самого закону правоохоронних органів.

Система NETRA більшою мірою спрямована на спостереження за діяльністю в Інтернеті та тенденціями підозрілих і сумнівних людей,

підприємств та організацій, які мають історію або схильність до здійснення огидних і недоброзичливих дій. Канали (як показано на рис. 2.2), яких торкається впровадженням системи NETRA, включають твіти в підозрілих облікових записах Twitter, оновлення статусу від Facebook, електронної пошти, служби миттєвих повідомлень, Інтернет-дзвінки, послуги Blackberry, блоги, форуми, а також перехоплення підозрілого голосового трафіку від Google Talk Services і Skype Messenger з відповідними IP-адресами.

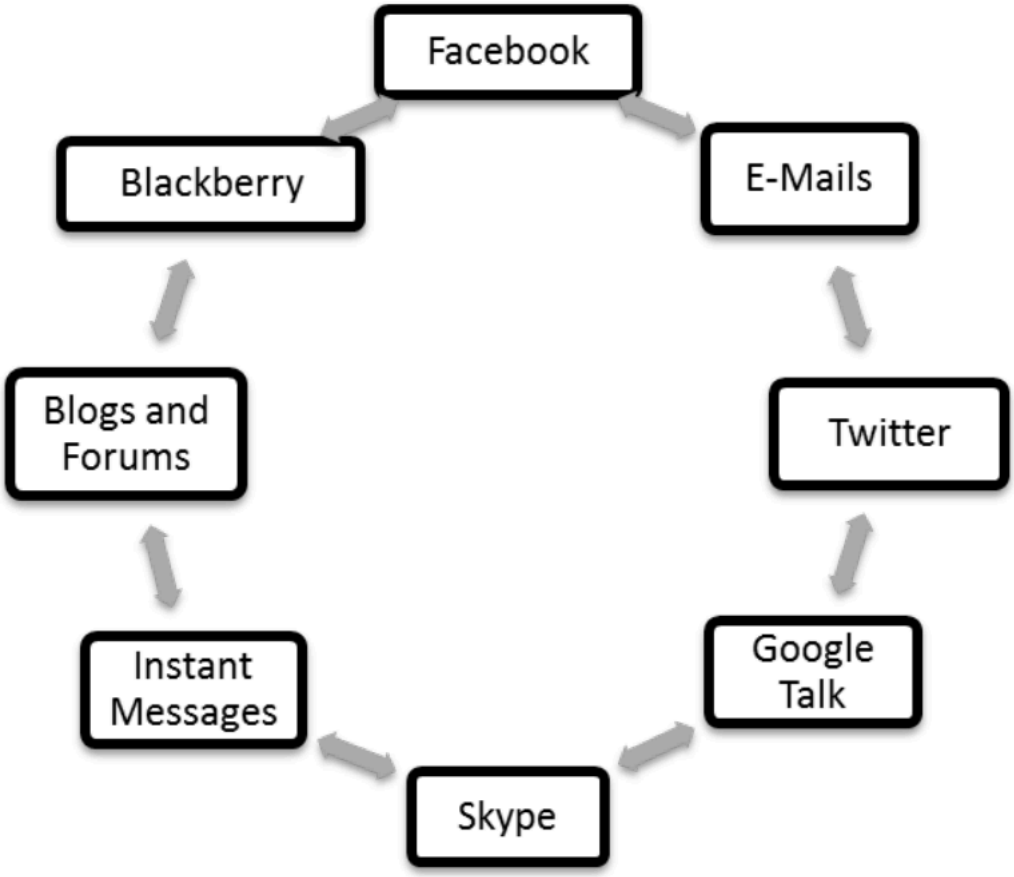

Рисунок 2.2 – Канали, яких торкається впровадження NETRA

NETRA, хоча і є передовим заходом безпеки на національному рівні, приречена не бути на 100% успішним з точки зору аналізу через велику кількість викликів, з якими вона повинна зіткнутися. NETRA, на даний момент схильна

тільки до часткового успіху завдяки численним проблемам, які перешкоджають очікуваному успіху, як показано на рис. 2.3. Вони розглядаються нижче.

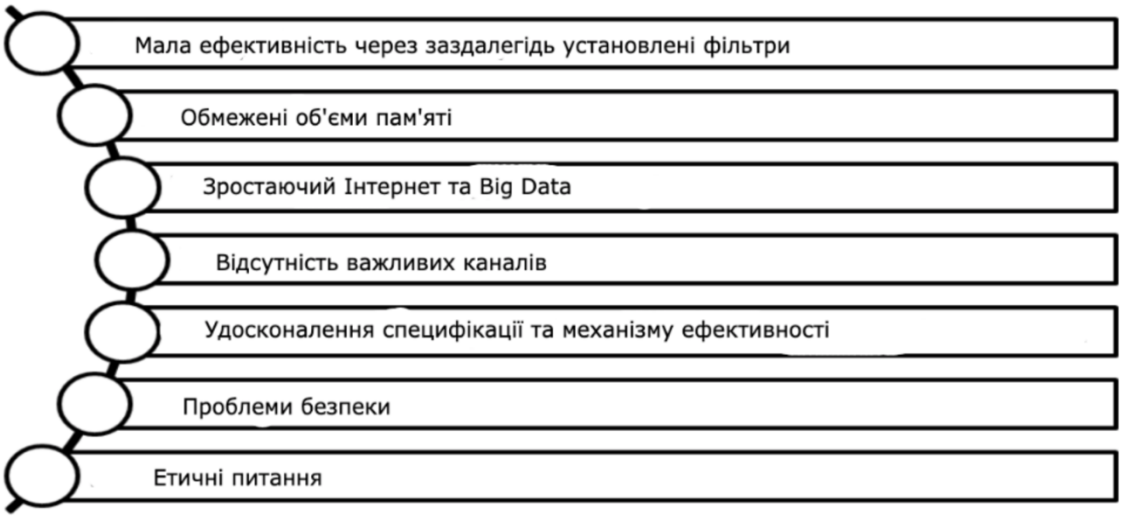

Рисунок 2.3 – Проблеми існуючої системи NETRA

1) Мала ефективність через заздалегідь установлені фільтри

NETRA як система працює за допомогою попередньо визначених фільтрів, які відповідають ключовим словам "атака", "бомба", "вибух", "терорист" і "джихад". Але це може призвести до неефективної системи, оскільки розмова або дані, які перехоплюються, можуть мати мале значення для національної безпеки. Буде важко зробити висновок про наміри, що стоять за ними, використовуючи такі слова, оскільки тільки частота не може бути придатним методом для цієї мети. Система повинна бути більш адаптивною до аналізу слів.

2) Обмежені об’єми пам’яті

Технічні характеристики системи NETRA показують, що трьом органам національної безпеки, Секретаріату Кабінету Міністрів, Службі розвідки та Відділу досліджень та аналізу буде надано обмежене об'ємом 300 ГБ, а правоохоронні органи отримають додаткові 100 ГБ. Статистика показує, що в Індії є 216,5 мільйона користувачів соціальних медіа [24], що є суттєвим збільшенням з 134-ох мільйонів користувачів соціальних медіа в 2015 році. 6000

твітів кожну секунду дня, що приблизно близько 350 000 твітів за хвилину і понад 500 мільйонів твітів на день, які зростають з кожним днем через введення нових облікових записів Twitter на день в країні (інтернет-статистика). Масштабність даних, які необхідно розшифрувати та вивчити системою NETRA, є величезною і навіть якщо ми ігноруємо постійно зростаючий характер даних, NETRA не зможе керувати та перехоплювати велику кількість даних, передаються і передаються щодня. Дуже ймовірно, що система вийде з ладу, тому що простір для зберігання 300 ГБ для трьох агентств безпеки вичерпається дуже швидко в часі [23].

3) Зростаючий Інтернет та Big Data

Інтернет - це велика мережа, яка зростає експоненційною швидкістю. Її моніторинг і нагляд можуть бути набагато складнішим і зручнішим завданням, ніж було передбачено спочатку. Є мільйони користувачів соціальних мереж, які щодня передають і передають мільярди повідомлень і голосових повідомлень. Дослідження та перевірка такої великої кількості користувачів вимагатимуть ресурсів і технологій, які є більш просунутими, ніж те, що в даний час володіє Індія. Завдяки тому, що стільки комунікаційних програм зростає з експоненційною швидкістю, і велика кількість користувачів, які виходять в Інтернет через мобільні телефони та ноутбуки, погіршать ситуацію. Велика технологія передачі даних повинна бути реалізована таким чином, щоб масштабування даних не впливало на систему спостереження. Таким чином, зростаюче використання Інтернету та постійно зростаючі великі дані становлять серйозну загрозу для роботи NETRA. Зі збільшенням використання Інтернету веб-сайти, а також постачальники послуг почали надавати просунуті рівні шифрування даних, щоб забезпечити конфіденційність користувачів, але оскільки система NETRA залежить від випробуваних даних, не працюючих інструментів дешифрування, які використовуються на шлюзах, і, отже, система не може розшифрувати зашифровані дані, що робить всю вправу безглуздою. Крім того, із зростанням популярності інтернет-дзвінків виникає проблема

розшифровки інтернет-викликів, в якій влада не має успіхів, що робить всю мету відстеження викликів марною[25].

4) Відсутність важливих каналів

Популярні веб-сайти для спільного використання файлів і канали, такі як Dropbox, Rapidshare і Fileshare [26], не були включені в канали, що мають вплив ЦШІР. Мессенджери, що мають високий попит, такі як Whatsapp та інші мобільні додатки також не були включені. Такі прогалини залишають величезний пробіл у правильній реалізації системи і можуть перешкоджати повному перехопленню онлайн-діяльності підозрілих осіб і груп. Більш того, постійно розвиваються такі платформи та канали, які також змушують систему моніторингу бути більш динамічною у своєму підході. А просування технологій шифрування цих каналів соціальних медіа призвело до того, що NETRA зіткнулася з низкою проблем, які можна пояснити обмеженим досвідом криптології керівниками програм в Індії. Ситуація така, що оборонні відомства мають обмежені можливості, окрім надання допомоги від провідних провайдерів соціальних медіа, таких як Facebook, Google і Twitter тощо, для надання їм оновленої бази даних комунікацій [25].

5) Удосконалення специфікації та механізму ефективності

Під час невеликої демонстрації в січні 2012 року, NETRA змогла передати лише 3 Гб трафіку з загальної кількості 28 Гб через свої зонди. Це єдина найвища точка, яка працювала як перевага в той час. Це була єдина система з тих, що тестувалися, яка змогла захопити трафік даних в Інтернеті без будь-якого зриву. Крім того, для такої системи, як NETRA, яка спрямована на забезпечення безпеки, швидка реакція є обов'язковою і, мабуть, найважливішою особливістю. Система може врятувати багато життів за останній час і, отже, необхідна швидка реакція. Базуючись тільки на ключових словах, фільтрація не вирішить проблему

ефективності, і для майбутніх заходів потрібно буде підтримувати вдосконалені технології збору даних.

6) Проблеми безпеки

ЦШІР має захищати систему NETRA від зовнішнього злому. Зовнішній взламувач системи NETRA може призвести до хаосу в інтернет-трафіку і можливих передач загроз без будь-якого перехоплення, що робить основну функцію використання системи NETRA марною. Крім того, атака ключових слів може призвести до паніки для системи. Таким чином, повинен існувати розпізнавач шаблонів, який повинен мати можливість контролювати фільтри системи та відстежувати аномальну поведінку.

7) Етичні питання

Нагляд за приватними переговорами і повідомленнями на платформах, таких як Google Talk і Skype, вважається дуже неетичним. Навіть в значній мірі розвинені економіки, такі як США та Китай, не були повністю успішними у впровадженні своїх систем моніторингу Інтернету з тієї ж причини. Інтернет-спостереження може працювати і бути успішним лише в тій мірі, за якою рівень успіху та ефект стає стагнацією. Навіть один з батьків-засновників World Wide Web, доктор Вінт Серф вважає, що інтернет-спостереження ніколи не може бути успішним, оскільки йде врозріз з основами Інтернету, який, в першу чергу, ґрунтується на свободі слова і дії [23].

2.2.2 ECHELON

ECHELON – це не просто система, а кодова назва, яка використовувалася для опису колекції інтелекту будь-якого сигналу та аналізу його мережі. Кодова назва була розроблена та розроблена спільною операцією п'яти різних країн, що

підписали її, а саме: Австралії, Нової Зеландії, Великобританії, США та Канади, що також посилається на ряд скорочень, таких як AUSCANNZUKUS. Через секретний договір близько кінця 1940-х років ці п'ять країн утворили ECHELON під англо-саксонським клубом без будь-яких комерційних наслідків. Було вирішено розділити комунікаційну діяльність у різних регіонах такими країнами, що було б корисним для проведення підслуховування у всьому світі. Система була головним чином орієнтована на військову розвідку та різні дипломатичні зв'язки під час холодної війни 1960-х і 1970-х років. ECHELON простими словами, можна сказати, що це глобальна мережа різних станцій моніторингу, які можуть таємно слухати спілкування на телефонах, факсах і електронних листах. ECHELON безпосередньо пов'язаний зі штаб-квартирою Агентства національної безпеки США (АНБ) у Форт-Меді в штаті Меріленд [27].

Система ECHELON була розроблена для вирішення багатьох каналів зв'язку для передачі повідомлень. Він може перехоплювати та перевіряти передачу даних через факсиміле, телефонні розмови, повідомлення через телетекст, використання Інтернету, електронну пошту та інші форми цифрового зв'язку, що відбуваються в різних регіонах світу. SILKWORTH і SIRE - це функціональні програми, що складають основу ДНК системи ECHELON, і перехоплення здійснювалося за допомогою супутника VORTEX [27].

Робота системи ECHELON, як показано на рис. 2.4, не є складним завданням [28] [29]. Вона має три основні кроки, пов'язані з системою та її роботою. Перший - це збір даних, другий - обробка даних, а третє - узгодження даних. Захоплення даних здійснюється через різні точки заходу по всьому світу і захоплює через різні канали зв'язку. Після цього АНБ здійснює обробку всіх типів даних за допомогою суперкомп'ютерів. Після того, як обробка завершена, словник поєднується з підозрілими словами та фразами, які дозволяють виявити будь-який зловмисний шаблон і працювати над ним. Система ECHELON добре розбирається з різними каналами та платформами зв'язку, які можуть захоплювати різноманітні передачі даних.

Рисунок 2.4 – Порядок роботи системи ECHELON

Кожне агентство, пов'язане з компанією ECHELON, охоплює частину світу, розділяючи світ між партіями Великобританії та США. АНБ США охоплює сигнали обох континентів Америки. ЦУЗ Британії охоплює регіон Африки, Європи і західних частин Росії. УРО Австралії охоплює регіони, такі як східна сторона Індійського океану, південно-західні береги Тихоокеанського регіону і південно-східну частину азіатського регіону. СБУК Нової Зеландії охоплює частини острівних країн з південної частини Тихого океану. ЦБК Канади охоплює північні райони Росії та Європи разом з американськими комунікаціями.

Система ECHELON з'явилася, коли Європейський парламент опублікував доповідь про ECHELON у 2001 році, а звинувачення в тому, що він прослідковує конфіденційну інформацію з європейського регіону, було зроблено проти конкурентів з Америки та Великобританії. Йдеться про те, що система потенційно перешкоджає конфіденційності та таємниці окремих осіб та організацій, вторгаючись у їхнє середовище спілкування.

2.2.3 DishFire

Dish Fire є системою спостереження, що працює на глобальному рівні, контролюючи трафік разом з підтримкою бази даних для зв'язку. Її здійснюють агенції безпеки Сполучених Штатів Америки та Сполученого Королівства - АНБ і ЦУЗ, відповідно. Вона працює щодня і збирає сотні і мільйони текстових повідомлень по всьому світу. Для аналізу зібраних даних використовується аналогічний аналітичний інструмент, відомий як PREFER. Цей інструмент використовується для обробки SMS, що надсилається через стільникову мережу і витягує життєво важливу інформацію з телефону, як сповіщення про виклик, розташування телефону, фінансові деталі виконаних платежів і електронну обробку карт. Dish Fire працює шляхом збору та аналізу автоматизованих текстових повідомлень, таких як оповіщення про пропущені виклики або тексти, що надсилаються для інформування користувачів про міжнародні тарифи на роумінг [30]. Ця система також порушує конфіденційність користувачів і громадян. Крім того, ця система також вилучає текстові дані для аналізу будь-якої потенційної загрози.

2.2.4 PRISM

PRISM - це система, яку АНБ використовує для отримання доступу до приватних комунікацій користувачів дев'яти популярних Інтернет-послуг. Ця система, запущена в 2007 році, також використовується для спостереження за допомогою моніторингу та інтелектуального аналізу даних АНБ у США в асоціації з ЦЗУ Великобританії. Офіційна номенклатура PRISM - SIGAD US-984XN, призначена для збору та аналізу даних. Система повністю підпадає під юридичні зобов'язання і дотримується положень Акту про негласне спостереження в цілях зовнішньої розвідки. Вона більш схильна до оцінки зашифрованих наборів даних. Програми АНБ збирають два типи даних: метадані та вміст. Метадані - чутливий побічний продукт зв'язку, таких як час, кількість

викликів, телефонні записи та вміст - електронні листи, чати, VoIP-дзвінки, файли, що зберігаються у хмарі, та багато іншого [31]. Система також має подібний порівняльний стиль шуканого контенту і затвердженого змісту в межах юридичних меж.

Висновки до розділу 2

В даному розділі був розглянутий стан суспільства під впливом соцмереж, наслідки публічних повідомлень з потенційно небезпечним змістом і методи боротьби з ними.

Саме соціальні мережі в наш час є основним способом обміну інформацією, а інформація в них може бути як і корисною, так і шкідливою. Інтернет моніторинг - одна з найперспективніших методик у забезпеченні безпеки як і на приватному, так і на державному рівні. Підняті етичні питання систем інтернет-моніторингу. Завдяки тому, що були розглянуті особливості розробки сервісів інтернет-моніторингу різними корпораціями чи державними установами, а також після вивчення їх проблем, з’явилась можливість розробити власну методику виявлення потенційно небезпечних повідомлень.

3 РЕАЛІЗАЦІЯ МЕТОДИКИ ВИЯВЛЕННЯ ПОТЕНЦІЙНО НЕБЕЗПЕЧНИХ ПОВІДОМЛЕНЬ В СОЦІАЛЬНИХ МЕРЕЖАХ

3.1 Постановка задачі та вибір інструментів для її вирішення

Завдання даної роботи буде виконане на прикладі виявлення повідомлень терористичного характеру в соціальній мережі Twitter.

Класифікація повідомлень терористичного характеру на Twitter

Хештеги забезпечують групування подібних повідомлень, оскільки можна шукати хештег і отримувати набір повідомлень, які його містять. Сьогодні вони стали немодерованим форумом для обговорення. Наприклад, слідування хештегу показує користувачів або суб'єкти, які зацікавлені в темі, позначеною хештегом. У нашому дослідженні ми використовуємо силу цих хештегів, оскільки вони з'являються як потужний грубозернистий фільтр для тем для обговорення, захоплюючи дані, що відповідають конкретним хештегам, а потім виконують тонкозернисту класифікацію для виявлення прихованих категорій.

Метою цього розділу є створення класифікатора повідомлень терористичного характеру у Twitter. Основними етапами роботи є:

1. Попередня обробка даних (dataset preprocessing).
2. Побудова визначення характеристик.
3. Оцінка правильності роботи вибраного алгоритму машинного навчання (SVM) в залежності з використаними характеристиками.
4. Тестування роботи класифікатора

Інструментами для виконання роботи вибрано:

1. NLTK (набір модулів Python, що дозволяє працювати з даними природної мови)

2. Pandas (бібліотека Python, розроблена для аналізу даних, яка забезпечує легке використання даних, робить роботу з даними гнучкою)
3. Scikit-learn (вільна бібліотека програмного забезпечення Python, створена на Numpy, SciPy і matplotlib, яка має інструменти для інтелектуального аналізу даних і добування даних, що будуть корисні в нашій моделі машинного навчання з учителем)
4. Meaning Cloud (API для для семантичного аналізу)

3.2 Розробка рішення для поставленої задачі

Набір даних для виконання проекту взято з відкритого доступу (kaggle - платформа, що організовує змагання, навчальні програми у сфері data science). Перша фаза побудови конвеєра - це попередня обробка даних. Друга фаза - це визначення характеристик. Нарешті, остання фаза побудови конвеєра є кульмінацією того, що було зроблено раніше, використовувались алгоритми класифікації, які надаються scikit-learn.

3.2.1 Набір даних

Змагання kaggle, з якого були взяті дані, називається “Tweets targeting Isis”. Міститься 2 набори даних пов’язаних з діяльністю терористичної організації ІДІЛ (Ісламська держава Іраку та Леванта). Перший набір (isisfanboy) містить дані, зібрані з облікових записів сторонників ІДІЛ, та містить приблизно 17000 твітів. Другий набір даних (aboutisis) містить нейтральні твіти (приблизно 120000) пов’язані з темою ІДІЛ.

У обох з них містяться наступні стовпці:

- username - ідентифікатор облікового запису
- tweets - текст, розміщений у Twitter.

Передача обох наборів даних була зроблена з використанням функціоналу Pandas read_csv, тобто повертається об'єкт DataFrame, що містить стовпці наборів даних.

Для обох наборів даних додається новий стовпець з назвою “radical”, значення якого “yes” у наборі даних isisfanboy та “no” у даних aboutisis, для полегшення навчання моделі з використанням навчання з учителем. Після цього ці набори даних об'єднані за допомогою Pandas і з допомого методу values об'єкта DataFrame, дані перетворюються на об'єкт numpy для того, щоб зробити визначення характеристик та застосування класифікатора. Ці дані будуть передані у конвеєр, що описаний нижче.

3.2.2 Конвеєр (Pipeline)

Використовується модуль pipeline, що надається Scikit-Learn, для автоматизації робочого процесу класифікатора.

У цій фазі отримуються сирі дані в якості вхідних даних, ці дані повинні бути підготовлені для алгоритму машинного навчання і тому необхідно зробити деякі перетворення даних у фазі попередньої обробки, яка має повернути дані, підготовлені для використання алгоритмом.

Після попередньої обробки сирих даних, можна зробити потрібні операції для покращення класифікатора. Буде зроблено визначення ознак, яке отримує в якості вхідних даних попередньо оброблені дані, виконує в них деякі операції і повертає нову колонку з новою ознакою, яка може бути використана алгоритмом машинного навчання.

Фінальною стадією конвеєра є класифікатор і останнім кроком є отримання точності алгоритму класифікації.

На рисунку 3.1 показано всі вищеописані стадії.

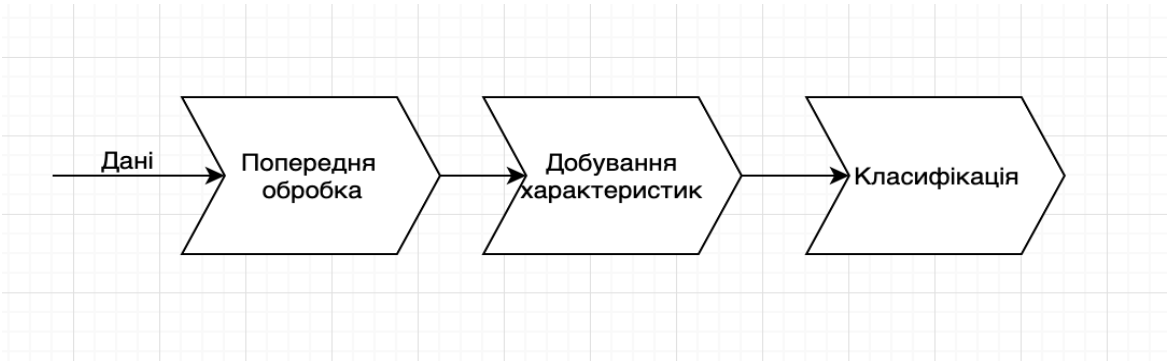

Рисунок 3.1 – Схема моделі

3.2.3 Попередня обробка

На даному етапі відбувається очищення даних від символів, що не впливають на зміст повідомлення (твіта). Для того, щоб видалити не важливі частини твітів, важливою є концепція маркування. Це процес розбиття речення на частини, що називається маркерами(tokens), чим можуть бути знаки пунктуації, слова, числа та багато інших. Кожен маркер - символ, чи група символів, розділена пробілами. У цьому проекті використовується TweetTokenizer з NLTK, який повертає список маркерів, що містяться в твіті. За допомогою цього функціоналу здійснено процес маркування, і як тільки отримано список маркерів, видалено наступні:

- Знаки пунктуації.
- Стоп-слова: це слова, які не змінюють змісту речення, вони не мають значення для класифікації твітів, наприклад стоп-слова - це "a", "the" і "other". Насправді, пошукові системи ігнорують ці слова. Ми використовували стоп-слова, надані NLTK.
- Цифри: номери були видалені, оскільки вони не несуть користі для класифікації тексту. Для того, щоб видалити цифри використовується метод isdigit, що надається модулем Python String.
- Смайли
- URL-адреси: маркери, що містять URL-адреси, також були видалені, оскільки вони не корисні для нашого класифікатора.

Говорячи про Twitter, обов’язково потрібно згадати про хеш-теги. Хештег - це рядок, що складається з одного або кількох зв'язаних слів, яким передує хеш-символ (#). Твіти можуть бути згруповані з хештегами, якщо ми шукаємо хештег у панелі Twitter, всі твіти, що містять ці хеш-теги, будуть показані і будуть перемішані, що є однією з цілей ІДІЛ.

TweetTokenizer є специфічним маркувальником для твітів, з ним, хештеги можуть бути видалені; якщо його параметр strip handles має логічне значення true, ми видалимо їх, значення false залишить їх. Залежно від бажаної особливості, хеш-теги будуть видалені чи ні. У цьому проекті, хештеги були видалені у всіх методах виділення суттєвих ознак, за винятком методу хеш-тегів.

Після отримання маркерів існують слова, які є спільнокореневими. У завданнях класифікації ці слова повинні бути уніфіковані для того, щоб дозволити алгоритму машинного навчання аналізувати лише ознаки, що містять всі варіації слів, це досягається за допомогою стемера. Метою використання стемера є нормалізація даних, перетворення в загальне слово всі спільнокореневі, що мають одинаковий сенс. Наприклад, якщо у нас є слово “убивця”, а також “убиває”, то ці слова можна згрупувати за їх коренем “убив”. У цьому проекті використовується SnowballStemmer, який перетворює слова в загальний корінь, який не обов'язково повинен бути існуючим словом, а WordNetLemmatizer використовується для перетворення кореня в загальне слово.

3.2.4 Виділення суттєвих ознак

Цей етап полягає у створенні нових наборів ознак шляхом виконання операцій для отриманих даних з початкового набору даних з метою надання допомоги в розмежуванні категорій алгоритму класифікації.

Всі методи виділення суттєвих ознак отримують дані після попередньої обробки, здійснюють певні операції, відповідно до ознак(характеристик), які вони повинні отримати.

Для кожної отриманої ознаки до нашого набору даних додано новий стовпець, щоб потім визначити чи збільшує ця ознака точність алгоритму чи ні.

Для об'єднання методів виділення суттєвих ознак ознак необхідно використовувати клас FeatureUnion, наданий scikit-learn, в конструктор якого передається список методів, які він має об’єднати в один.

3.2.4.1 Виділення слів

Цей метод виділення суттєвих ознак був зроблений за допомогою інструменту TfidfVectorizer, наданого scikit-learn.

TfidfVectorizer створює матрицю частоти частотно-інверсної термінології (tf-idf) це означає, що вага маркерів, які часто з'являються, буде меншою ніж вага маркерів, які з'являються менше разів, тому що вони більш інформативні і можна визначити мітку, коли йдеться про класифікацію цієї функції.

Застосування CountVectorizer з наступним TfidfTransformer має той же результат.

Перш за все, TfidfVectorizer створює матрицю відліку маркерів від наших даних до матриці лічильника маркерів (наприклад, CountVectorizer) і після, він перетворює матрицю лічильників на нормоване представлення tf-idf (наприклад, TfidfTransformer).

TfidfVectorizer має деякі параметри, які можна редагувати для адаптації до нашого проекту. Ми використовували в якості формату кодування utf-8 і як аналізатор, який є параметром для вилучення послідовності лексем, ми вибрали метод попередньої обробки, який пояснений в 3.2.3.

Нарешті, кожне речення розглядається класифікатором як вектор з вагою tf-idf кожного маркера, що має текст.

Найбільш вживані 10 слів в обох наборах даних можна побачити в таблиці 3.1.

Таблиця 3.1 – найчастіше вживані 10 слів в обох наборах даних

isisfanboy	aboutisis
Isi	Islamic
Kill	Retake
Islamic	Mosul
US	Troop
Attack	Iraq
Muslim	Muslim
Say	Daesh
Today	Defence
Fight	Help

3.2.4.2 Виділення n-грам

N-грама являє собою підпослідовність, складену n елементами заданої послідовності. При створенні класифікатора твітів, так само важливо аналізувати n-грами як і слова, оскільки з n-грамних виразів або груп слів, які використовуються в isisfanboy або в aboutisis можна зробити більш точну класифікацію.

3.2.4.3 Виділення POS

Метою цього добування є аналіз граматичних категорій, що використовуються в різних наборах даних. Було створено метод виділення ознак, який отримує кількість маркерів з категоріями POS (Part of speech/частини мови).

3.2.4.4 Виділення NER

Концепція NER (named entity recognition) відноситься до процесу пошуку і зберігання маркерів, які представляють людей, організації, місця розташування, дати, часу, та інших. Використаний Stander NER Tagger знаходить лише імена, організації та місця розташування.

Найбільш вживані 10 об’єктів NER в обох наборах даних можна побачити в таблиці 3.2

Таблиця 3.2 – найчастіше вживані 10 об’єктів NER в наборі isisfanboy та aboutisis

isisfanboy	aboutisis
Isi	Isi
Syria	Mosul
Assad	Carter
Aleppo	Iraq
City	US
Muslim	France
Abu	Daesh
Village	Muslim
Palayra	Cnet

3.2.4.5 Виділення хеш-тегів

Метою виділення хештегів є аналіз хештегів, що входять до кожного твіту. Найбільш вживані 10 хеш-тегів в обох наборах даних можна побачити в таблиці 3.3

Таблиця 3.3 – найчастіше вживані 10 хеш-тегів в наборі isisfanboy та aboutisis

isisfanboy	aboutisis
Isi	Isi
Syria	Daesh
Assad	Syria
Aleppo	Iraq
City	Mosul
Breaking	Indiaisisandbangladesh
Iraq	Islamicstate
Amaqagency	Poke
Lybia	Dontbanpeacetv

3.2.4.6 Визначення настрою повідомлення

Цей метод виділення ознак був розроблений з використанням Meaning Cloud API для аналізу настроїв. Цей API повертає JSON з результатом аналізу настроїв твітів. У першу чергу, запит на кожен твіт до Meaning Cloud API був зроблений з використанням ключа API, за яким слідує текст, що міститься в твітті. Результатом запиту є JSON, включаючи поля, що аналізують полярність, іронію та суб'єктивність.

Для кожного твіту в JSON було додано значення полів тегів для кожного твіту, що вказує на знайдену в аналізованому тексті полярність, може бути шість можливих значень:

- P +: сильна позитивна полярність.
- P: полярність позитивних настроїв.
- N +: сильна негативна полярність.
- N: полярність негативних настроїв.
- NEU: ні позитивна, ні негативна полярність.
- NONE: не знайдено настрою.

Таблиця 3.4 – значення полярності для десяти твітів

Твіт (можливе скорочення)	Полярність
T @jangojadoon: After 10 years we will be ashamed of coz of his connection with ISIS #SackDoval he is only friend to @narendramodi https:/…	NEU
Come on @UNESCO . Nobody believed that statement to be true. No even the ARDENT ISLAMISTS themselves. Not even ISIS. @DavidBCohen1	N
RT @alfonslopeztena: Muslim man hugs ISIS suicide bomber moments before explosion, saves hundreds of lives: http://www.indiatimes.com/news/world/muslim-man-hugs-isis-militant-armed-wearing-suicide-vest-before-explosion-saves-hundreds-of-lives-258126.html # via @…	P
RT @peddoc63: When Japan attacked America, Truman bombed Japan✔ When ISIS strikes America, Obama threatens to take our guns❌ https://t.co/…	N
RT @moscow_ghost: #Syria 'expert ' tweets 'reliable ' article from western press And some wonder why people turn to RT? #ISIS #Russia https:/…	P+
RT @MahirZeynalov: A Turkish daily publishes stories on ISIS, and then the Turkish govt blocks their web-site. The newspaper outraged. http…	N+
@SlametMuslim nyadar kalo isis ternyata seperti itu	NONE
RT @nytimes: The U.S. will deploy 560 troops to Iraq to help retake Mosul from ISIS http://www.nytimes.com/2016/07/12/world/middleeast/us-iraq-mosul.html	P
rt: RT waheedgul: International Criminal Owned isis #SackDoval pic.twitter.com/6GimZAkw5A	N+
@PageSix @lindsaylohan maybe try ISIS?	NONE

3.2.3 Класифікатор

Останньою частиною нашого конвеєра є алгоритм машинного навчання, який вивчає закономірності для того, щоб правильно присвоїти мітку вхідним даним. У цьому випадку алгоритм отримує характеристики, що були добуті, в якості вхідних даних.

Алгоритми потребують навчальних даних, які є попередньо опрацьованими, та тестових даних, які використовуються для перевірки, чи алгоритм призначив правильну мітку для даного входу.

Модель була натренована з використанням перехресної перевірки, яка полягає в оцінці результатів статистичного аналізу, що гарантує незалежність процесу навчання та перевірки одне від одного. Використовувана методика перехресної перевірки називається k-fold, вона складається з поділу даних на k частин одного розміру, кожна частина використовується один раз як тестові дані і (k-1) разів як навчальні дані, що можна бачити на рисунку 3.2.

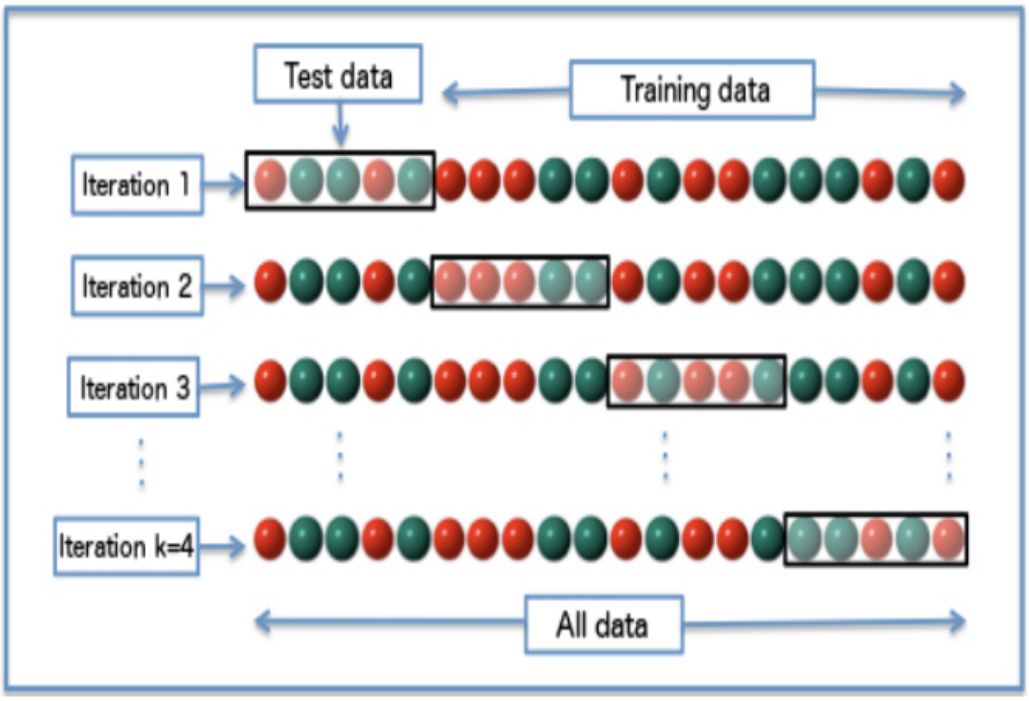

Рисунок 3.2 – Перехресна перевірка K-Fold [32]

Support Vector Machine (метод опорних векторів). (Рисунок 3.3)

Цей алгоритм використовується для задач класифікації, а також завдань регресії. Цей алгоритм також використовується для багатокласових класифікаційних завдань.

Постановка задачі класифікації SVM описується далі. Часто в алгоритмах машинного навчання виникає необхідність класифікувати дані. Кожен об'єкт даних представляється як вектор (точка) в p-вимірному просторі (упорядкований набір p чисел). Кожна з цих точок належить тільки одному з двох класів. Питання полягає в тому, чи можна розділити точки гіперплощиною розмірності p-1. Це - типовий випадок лінійної роздільності. Шуканих гіперплощин може бути багато, тому вважають, що максимізація відступу між класами сприяє більш впевненою класифікації. Тобто, чи можна знайти таку гіперплощину , щоб відстань від неї до найближчої точки було максимальним. Це еквівалентно тому, що сума відстаней до гіперплощини від двох найближчих до неї точок, що лежать по різні боки від неї, має бути максимальною. Якщо така гіперплощина існує, вона називається оптимально розділяючою гіперплощиною, а відповідний їй лінійний класифікатор називається оптимально розділяючим класифікатором.

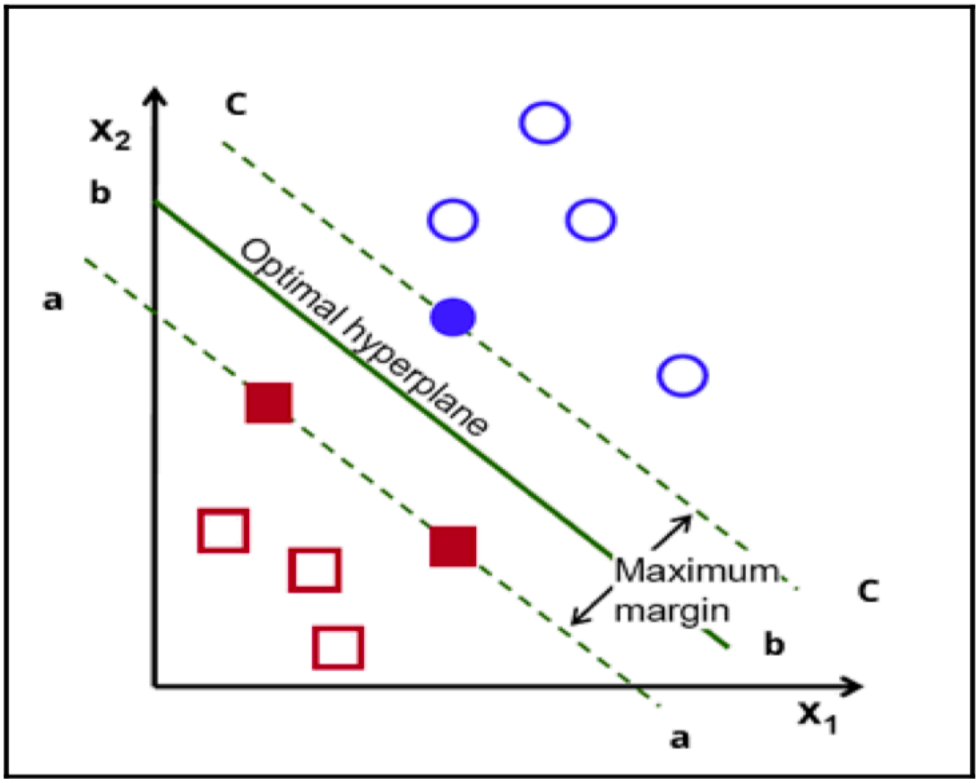

Рисунок 3.3 – Графік базової класифікації SVM [33]

SVM надає два типи алгоритмів. SVR використовується для регресійних проблем, і SVC, який використовується в даній роботі, зазвичай використовується для класифікації. Параметри SVC, значення яких визначено за допомогою GridSearchCV, такі:

C: це параметр покарання терміну помилки. Цьому параметру надано значення 10.

Kernel: існує п'ять типів ядер, які є лінійними, поліноміальними (poly), радіальними базисними функціями (rbf), сигмоїдними та попередньо обчисленими ядрами.

Gamma: вона являє собою поширення ядра і, отже, область прийняття рішення. Цьому параметру надане значення 1.

Probability: параметр, значення якого є істинним або помилковим; вона вказує, чи потрібно включати ймовірнісні оцінки. Задано значення true.

Середнє значення і відхилення правильної роботи моделі SVM з кожним методом виділення особливостей індивідуально можна побачити в Таблиці 3.2.

Найкращих результатів можна досягти, якщо включити до конвеєра тільки виділення слів та n-грам. Тоді правильність становитиме 0.96 з відхиленням +/- 0.01.

Таблиця 3.5 – Правильність та відхилення SVM з кожним з методів виділення особливостей

Метод виділення	Правильність та відхилення
Слів	0.96 (+/- 0.01)
N-грам	0.96 (+/- 0.01)
POS	0.59 (+/- 0.01)
NER	0.86 (+/- 0.01)
Хеш-тегів	0.67 (+/- 0.01)
Настроїв	0.55 (+/- 0.02)

3.3 Приклади застосування класифікатора

Для того, щоб упевнитись, що класифікатор натреновано правильно, а також правильно працює, потрібно його протестувати. Потрібно вигадати твіти, що несуть різний зміст. Після класифікації введеного твіта буде виведено “yes”, якщо він несе про-терористський зміст, та “no”, якщо терористичного змісту немає. Для першого варіанту було вибране повідомлення таке: “kill non muslims in the wake of allah”, і було правильно класифікованим. Для другого: “we are against ISIS, coallition should get all forces to defeat them” і отриманий результат запевнив нас у відсутності терористичних нахилів у даному повідомленні. Випробування приведені на рисунку 3.4.

```
(diploma) Daniels-MacBook-Pro:diploma kranzer$ python main.py
Enter tweet to be classified
kill non muslims in the wake of allah
['yes']
(diploma) Daniels-MacBook-Pro:diploma kranzer$ python main.py
wEnter tweet to be classified
we are against ISIS, coallition should get all forces to defeat them
['no']
```

Рисунок 3.4 – Випробування класифікатора на випадках протилежного змісту

Висновки до розділу 3

В даному розділі був запропонований спосіб класифікації повідомлень у соціальній мережі Twitter за допомогою методу опорних векторів.

На початку розділу зроблена постановка задачі та вибір необхідних інструментів. Проведено детальний аналіз кожного з етапів вирішення даної задачі. Розглянуто роботу даного алгоритму з шістьма методами виділення суттєвих ознак, а саме виділення слів, n-грам, настрою, POS, NER, настрою та хеш-тегів. Дослідження за допомогою методу перехресної перевірки показало, що найбільший показник правильно класифікованих об’єктів показало використання методу виділення слів та n-грам одночасно. Відповідно до цього можна зробити висновок, що для класифікації тексту за допомогою алгоритму SVM найкраще використовувати ці методи виділення суттєвих ознак. Також було показано приклад використання даного класифікатора.

ВИСНОВКИ

Результатом даної роботи є методика виявлення потенційно небезпечного змісту в повідомленнях у соціальній мережі Twitter. За допомогою розробленого класифікатора було перевірено повідомлення різного вмісту. Критерієм правильності роботи вибрано - accuracy (частка правильно класифікованих об'єктів). Для визначення правильності класифікації було випробувано 6 методів виділення суттєвих ознак і в результаті вони показали такі показники:

1. Метод виділення слів - 0.96 (+/- 0.01)
2. N-грам - 0.96 (+/- 0.01)
3. POS - 0.59 (+/- 0.01)
4. NER - 0.86 (+/- 0.01)
5. Хеш-тегів - 0.67 (+/- 0.01)
6. Настроїв - 0.55 (+/- 0.02)

Підрахунок оцінки правильності класифікації відбувався за допомогою методу перехресної перевірки (k-fold).

В результаті найбільший показник класифікатор отримує при використанні методів виділення суттєвих ознак таких як виділення слів та n-грам.

Результати свідчать про те, що використання SVM (методу опорних векторів) є одним з ефективних методів класифікації, а також його можна використовувати і в проблемах регресії.

Можна зробити висновок, що використання машинного навчання, а також обробки природної мови є перспективним напрямом у галузі інтернет-моніторингу та кібербезпеки у цілому, оскільки методику представлену у даній роботі можна використовувати у багатьох інших цілях. Для цього просто потрібно вибрати набори даних, які стосуються цікавої нам тематики, натренувати модель, що представлена у цій роботі, працювати з повідомленнями, що стосуються вже, наприклад, кібербулінгу.

СПИСОК ДЖЕРЕЛ ПОСИЛАНЬ

1) Michael Fire, Roy Goldschmidt and Yuval Elovici: “Online Social Networks: Threats and Solutions”, IEEE Communication Surveys & Tutorials [Електронний ресурс]. – 2014. – Режим доступу до ресурсу: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6809839

2) D. Cavit et al., Microsoft Security Intelligence Report Volume 10, 2010 [Електронний ресурс]. – 2010. – Режим доступу до ресурсу: http://www.microsoft.com/ en-us/download/details.aspx?id=17030

3) L. Tristan, Twitter’s Growing Spam Problem, Forbes. [Електронний ресурс]. – 2013. – Режим доступу до ресурсу: http://www.forbes.com/sites/ tristanlouis/2013/04/07/twitters-growing-spam-problem/

4) J. Halliday, “Facebook fraud a ‘Major Issue’,” The Guardian, London, U.K., [Електронний ресурс]. – 2010. – Режим доступу до ресурсу: http://www.theguardian.com/ technology/2010/sep/20/facebook-fraud-security

5) J. Lewis, “How spies used facebook to steal NATO chiefs’ details,” The Telegraph, London, U.K. [Електронний ресурс]. – 2012. – Режим доступу до ресурсу:

http://www.telegraph.co.uk/technology/9136029/How-spies-used-Facebook-tosteal-Nato-chiefs-details.html

6) S. Livingstone and L. Haddon, Child Safety Online: Global Challenges and Strategies [Електронний ресурс]. – 2012. – Режим доступу до ресурсу: http://www.newyorker.com/online/blogs/culture/2011/10/amanda-todd-michael-brutsch-and-free-speechonline.html

7) M. Deans, The Story of Amanda Todd, The New Yorker [Електронний ресурс]. – 2012. – Режим доступу до ресурсу:

http://www.newyorker.com/online/ blogs/culture/2012/10/amanda-todd-michael-brutsch-and-free-speechonline.html

8) Lpsos, One in Ten (12%) Parents Online, Around the World Say Their Child has Been Cyberbullied, 24% Say They Know of a Child Who has Experienced Same

in Their Community [Електронний ресурс]. – 2012. – Режим доступу до ресурсу: http://www.ipsos-na.com/news-polls/pressrelease. aspx?id=5462

9) Персональные данные – золотая жила интернет-компаний [Електронний ресурс]. – 2012. – Режим доступу до ресурсу : http://intemetua.com/personalnie-dannie—zolotaya-jila-internet-kompanii.

10) Україна скопіює у Росії ідею списку шкідливих сайтів [Електронний ресурс]. – 2012. – Режим доступу до ресурсу:

http://ua.ht.omments.ua/2012/11/06/187682/ukraina-skopiyuie-u-rosii-ideyu. html

11) Российские власти составили черный список противозаконных сайтов [Електронний ресурс]. – 2012. – Режим доступу до ресурсу:

http://for-ua.com/world/2012/11/01/100120. html

12) Корея боротиметься із суїцидами через соціальні мережі [Електронний ресурс]. – 2012.– Режим доступу до ресурсу :

http://www.lenta.ru/news/2012/06/05/congressmen/.

13) Голландская полиция хочет взламывать компьютеры за границей [Електронний ресурс]. – 2012.– Режим доступу до ресурсу :

http://internetua.com/gollandskaya-policiya-hocset-vzlamivat-kompuateri-za-granicei.

14) Неонацисти все частіше використовують соцмережі для пропаганди [Електронний ресурс]. – 2012 – Режим доступу до ресурсу :

http://www.dw.de/dw/article/0,,16091158,00.html

15) Милиция осваивает сеть “В Контакте” [Електронний ресурс]. – 2009. – Режим доступу до ресурсу :

http://net.compulenta.ru/milizia-osvaievaet-set-vkontakte.html

16) Британскую полицию научат находить преступников в Twitter и Face-book [Електронний ресурс]. – 2012. – Режим доступу до ресурсу :

http://internet-search.ru/britanskuu-policiu-nauchat-nakhodit-prestupnikov-v-twitter.html

17) Поліція США затримала півсотні злочинців завдяки Facebook / [Електронний ресурс]. – 2012. – Режим доступу до ресурсу :

http://ua.korrespondent.net/world/1394544-policiya-ssha-zatrimala-pivsotni-zlochinciv-zavdyaki-facebook

18) Найден способ борьбы со спамом в соцсетях [Електронний ресурс]. – 2012. – Режим доступу до ресурсу: http://www.fromua.com/news/97d8bbf410e3c.html.

19) Белоус Н. Киберджихад [Електронний ресурс]. – 2012. – Режим доступу до ресурсу: http://2000.net.ua/2000/derzhava/ekspertiza/84034.

20) Стартап сразится с преступностью при помощи Facebook [Електронний ресурс]. – 2012. – Режим доступу до ресурсу :

http://internetua.com/startap-srazitsya-s-prestupnostua-pri-pomosxi-Facebook.

21) Кремль уличили в использовании новейших систем мониторинга соцсетей [Електронний ресурс]. – 2012. – Режим доступу до ресурсу :

http://lb.ua/news/2012/08/16/166125_kreml_ulichili_ispolzovanii.html.

22) Горовий В. М. ІТ-субкультура як фактор розвитку сучасного правотворення / В. М. Горовий // Актуальні проблеми управління інформаційною безпекою держави: зб. мат. наук-практ. конф. (30 берез. 2012 р.). - К. : Наук-вид. відділ НА СБ України, 2012. - 308 с.

23) Ghosh, M., “Beware, Government Plans To Spy On Your Internet Activity Using Netra” [Електронний ресурс]. – 2013 – Режим доступу до ресурсу: http://trak.in/tags/business/2013/12/17/govt-spy-internet-netra/

24) “Statistics and facts about Facebook”. Statista. Home. Industries. Internet. Social Media & User-Generated Content. Facebook - Statistics & Facts. [Електронний ресурс]. – 2016 – Режим доступу до ресурсу: https://www.statista.com/topics/751/facebook/

25) Singh, V.P., “Myopic Netra: why the new system has failed to deliver”. [Електронний ресурс]. – 2015 – Режим доступу до ресурсу: http://www.governancenow.com/gov-next/egov/myopic-netra-new-cyber-tracking-systemfailed-deliver

26) UNODC, “UNODC Report on the Use of Internet for Terrorist purposes” [Електронний ресурс]. – 2016 – Режим доступу до ресурсу:

http://www.unodc.org/documents/frontpage/Use_of_Internet_for_Terrorist_Purposes.pdf

27) Chandler, P., “ECHELON -- The Spy System That Knows Everything, philipcfromnyc [Електронний ресурс]. – 2013 – Режим доступу до ресурсу: http://philipcfromnyc.hubpages.com/hub/ECHELON----The-Spy-System-That-Knows-Everything/

28) Bomford, A., “The Echelon spy network”. Echelon spy network revealed. [Електронний ресурс]. – 1999 – Режим доступу до ресурсу: http://news.bbc.co.uk/2/hi/503224.stm

29) ECHELON. “Exposing the NSA’s Global Spy Network”Alexandra Valiente [Електронний ресурс]. – 2013 – Режим доступу до ресурсу:

https://libya360.wordpress.com/2013/06/22/echelon-exposing-the-nsas-global-spy-network/

30) Hahn, D, J., “DISHFIRE: The Program That Lets the NSA Capture Almost 200 Million Texts a Day” [Електронний ресурс]. – 2014 – Режим доступу до ресурсу: http://www.complex.com/pop-culture/2014/01/dishfire-nsa-collects-texts

31) Greenwald, G. & MacAskill, E., “NSA Prism program taps in to user data of Apple, Google and others [Електронний ресурс]. – 2013 – Режим доступу до ресурсу: http://www.theguardian.com/world/2013/jun/06/us-tech-giants-nsa-dataon

32) Towards Data Science Blog [Електронний ресурс]. – 2018 – Режим доступу до ресурсу:

https://towardsdatascience.com/cross-validation-explained-evaluating-estimator-performance-e51e5430ff85

33) Jalaj Thanaki. Python Natural Language Processing – Packt Publishing Ltd. – 2017. – с.320

ДОДАТОК А

Лістинг коду

```
from collections import Counter

import pandas as pd
import string
from nltk.tokenize import TweetTokenizer
from nltk.corpus import stopwords
from nltk import pos_tag
import re

from sklearn.base import BaseEstimator, TransformerMixin
from sklearn.feature_extraction import DictVectorizer
from sklearn.feature_extraction.text import TfidfVectorizer, TfidfTransformer
from sklearn.feature_extraction.text import CountVectorizer



from nltk.stem.snowball import SnowballStemmer
from sklearn.pipeline import Pipeline, FeatureUnion

from preprocessor import preProcessSerie
import pickle

from sklearn.svm import SVC
from sklearn.model_selection import cross_val_score, KFold



def preProcessor(tweet):
    emoji_pattern = re.compile("["
```

```
            u"\U0001F600-\U0001F64F"
            u"\U0001F300-\U0001F5FF"
            u"\U0001F680-\U0001F6FF"
            u"\U0001F1E0-\U0001F1FF"
                        "]+", flags=re.UNICODE)
        tweet = str(emoji_pattern.sub(r'', str(tweet)))
        pal = tokenizer_1.tokenize(str(tweet))
        urls = re.compile(r'.http[s]?://(?:[a-zA-Z]|[0-9]|[$-_@.&+]|[!*\(\),]|(?:%[0-
9a-fA-F][0-9a-fA-F]))+')
        line = urls.sub('', str(tweet))
        ht = re.compile(r'http.')
        bar = re.compile(r'//*')
        punctuation = set(string.punctuation)
        stoplist = stopwords.words('english')
        pr = ["rt","@","http","https","'s",'...', 'english', 'translation','):', '. .', '..']
        pal = [stemmer.stem(str(i)) for i in pal if i not in pr
            if i not in stoplist if i not in punctuation
            if not bar.search(i) if not ht.search(i)
            if not i.isdigit() if not i.startswith('#')]
        tweet = pal
        return tweet

    tokenizer_1 = TweetTokenizer(preserve_case=False, reduce_len=True,
strip_handles = True)
    tokenizer_2 = TweetTokenizer(preserve_case=False, reduce_len=True,
strip_handles = False)
    stemmer=SnowballStemmer("english")

    listwords = []

    ngrams_featurizer = Pipeline([
```

```
        ('count_vectorizer', CountVectorizer(ngram_range=(1, 3), encoding='ISO-
8859-1',
                                             analyzer=preProcessor)),
        ('tfidf_transformer', TfidfTransformer())
    ])
    isis = pd.read_csv('isisfanboy.csv')
    about = pd.read_csv('aboutisis.csv')

    isis = isis[:17392]
    about = about[:17392]



    dataframe = pd.concat([isis, about])
    X = dataframe['tweets'].values.astype('U')
    y = dataframe['radical'].values

    class POS(BaseEstimator, TransformerMixin):

        def stats(self, tweet):
            tokens = tokenizer_1.tokenize(str(tweet))
            tagged = pos_tag(tokens, tagset='universal')
            counts = Counter(tag for word, tag in tagged)
            total = sum(counts.values())
            pos_fts = {
                'PRON': 0, 'NUM': 0,
                'NOUN': 0, 'ADJ': 0,
                'CONJ': 0, 'ADP': 0,
                'VERB': 0, 'ADV': 0
            }
```

```
        pos = dict((tag, float(count) / total) for tag, count in counts.items())
        for key in pos:
            pos_fts[key] = pos[key] if key in pos_fts else None
        return pos_fts

    def transform(self, data, y=None):
        dataproc = preProcessSerie(data)
        result = [self.stats(tweet) for tweet in dataproc]
        return result

    def fit(self, data, y=None):
        return self


class Hashtags(BaseEstimator, TransformerMixin):
    listwords = []

    def fit(self, X, y=None):
        return self

    def notinlist(self, item, list_hashtags):
        return False if item not in list_hashtags else True

    def get_hashtags(self, tweet, list_hashtags):
        list_hashtags = pickle.load(open("hashtags.pkl", "rb"))
        return list_hashtags

    def hashtags(self, tweet, all_hashtags, result_list):
        all_hashtags_dict = dict((ht, 0) for ht in all_hashtags)
        sent = tokenizer_2.tokenize(str(tweet))
        for term in sent:
```

```
            all_hashtags_dict[term]=1 if term in all_hashtags else None
        result_list.append(all_hashtags_dict)

        return (result_list)

    def transform(self, data):
        dataproc = preProcessSerie(data)
        list_ = []
        result_list = []
        list_ht = [self.get_hashtags(tweet, list_) for tweet in data]
        result_list = [self.hashtags(tweet, list_ht, result_list) for tweet in dataproc]
        return result_list


class NER(BaseEstimator, TransformerMixin):

    def fit(self, X, y=None):
        return self

    def notinlist(self, item, list_ner):
        return False if item not in list_ner else True

    def get_ner(self, dataproc, list_ner):
        list_ner = pickle.load(open("ner.p", "rb"))
        return list_ner

    def ner(self, tweet, all_ner_list, ner_s, result_list):
        all_ner_dict = dict((entity, 0) for entity in all_ner_list)
        words = tokenizer_2.tokenize(str(tweet))

        for i in words:
```

```
            for k in all_ner_dict:
                all_ner_dict[k] = 1 if i.lower() == k[0] and i.lower() in ner_s else None
        result_list.append(all_ner_dict)
        return (result_list)

    def transform(self, data):
        dataproc = preProcessSerie(data)
        list_ner = []
        result_list = []
        ner_s = []
        list_ner = self.get_ner(dataproc, list_ner)
        [ner_s.append(k[0]) for k in list_ner]
        for tweet in dataproc:
            result_list = self.ner(tweet, result_list)
        print(result_list[0])

        return result_list



class Sentiment(BaseEstimator, TransformerMixin):
    def fit(self, data, y=None):
        return self

    def transform(self, data, y=None):
        sentiments = pickle.load(open("sentiments.pkl", "rb"))
        list_sntms = []

        for tweet in data:
            if tweet != "nan":
                try:
                    print(tweet, sentiments[tweet])
```

```
                tweetsent = sentiments[tweet]
                list_sntms.append(tweetsent)
            except:
                list_sntms.append("NONE")
        else:
            list_sntms.append("NONE")
    return list_sntms


pipelinesvm = Pipeline([
    ('features',
     FeatureUnion([
        ('words', TfidfVectorizer(encoding='utf-8', analyzer=preProcessor)),
        ('ngrams', ngrams_featurizer),
        ('pos_stats', Pipeline([
                        ('pos_stats', POS()),
                        ('vectors', DictVectorizer())
                    ])),
        ('ner', Pipeline([
            ('ner_recogniser', NER()),
            ('vectors', DictVectorizer())
        ])),
        ('hashtags', Pipeline([
            ('gethashtags', Hashtags()),
            ('vect', DictVectorizer())
        ])),
        ('sentiments', Pipeline([
            ('getsentiments', Sentiment()),
            ('vector', TfidfVectorizer(encoding='utf-8'))
        ]))
    ])),
```

```
    ('clf', SVC(C=10, gamma= 1, kernel='rbf', probability=True)
)

])
cv = KFold(2, shuffle=True, random_state=33)
print(cv)
#
print(type(X.shape[0]))
scores = cross_val_score(pipelinesvm, X, y, cv=cv)
print("Scores in every iteration", scores)
print("Accuracy: %0.2f (+/- %0.2f)" % (scores.mean(), scores.std() * 2))

pipelinesvm.fit(X,y)
pickle.dump(pipelinesvm, open('model1.pkl', 'wb'))
pipelinesvm = pickle.load(open('model1.pkl', 'rb'))
print("Enter tweet to be classified")
tweet = input()

print(pipelinesvm.predict([tweet]))
```