ДЕРЖАВНИЙ УНІВЕРСИТЕТ ТЕЛЕКОМУНІКАЦІЙ

НАВЧАЛЬНО–НАУКОВИЙ ІНСТИТУТ
ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

Кафедра інженерії програмного забезпечення

Пояснювальна записка

до магістерської роботи

на ступінь вищої освіти магістр

на тему: «Розробка програмного забезпечення виявлення ознак маніпуляції в повідомленнях з мережі інтернет»

Виконав: студент 7 курсу, групи ППЗМ-71

спеціальність 121 Інженерія програмного забезпечення

Коновал Андрій Сергійович

Керівник: Бондарчук Андрій Петрович

Рецензент: _____

Нормконтроль: _____

м. Київ

2022

ДЕРЖАВНИЙ УНІВЕРСИТЕТ ТЕЛЕКОМУНІКАЦІЙ

Навчально-науковий інститут інформаційних технологій Кафедра Інженерії програмного забезпечення Ступінь вищої освіти -«Магістр» Спеціальність підготовки – 121 «Інженерія програмного забезпечення»

ЗАТВЕРДЖУЮ

Завідувач кафедри

Інженерії програмного забезпечення

Негоденко О.В.

“___” _____ 2022 року

З А В Д А Н Н Я

НА МАГІСТЕРСЬКУ РОБОТУ СТУДЕНТУ

Коновалу Андрію Сергійовичу

1. Тема роботи: Розробка програмного забезпечення виявлення ознак маніпуляції в повідомленнях з мережі інтернет

Керівник роботи: Бондарчук Андрій Петрович, д.т.н., професор кафедри ІПЗ

Затверджені наказом вищого навчального закладу від «__» ____ 2022 року № ___.

2. Строк подання студентом роботи _____

3. Вхідні дані до роботи

Ken Cherven. Network Graph Analysis and Visualization with Gephi. 2013 Packt Publishing, 2013

Главацька Ю.Л. Класифікація «фейкових» новин у сучасному медіапросторі

4. Зміст розрахунково-пояснювальної записки(перелік питань, які потрібно розробити).

5. Перелік демонстраційного матеріалу (назва основних слайдів).

6. Дата видачі завдання _____

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів магістерської роботи	Строк виконання	Примітка
1.	Підбір науково технічної літератури	25.09 - 23.10	
2.	Аналіз існуючих підходів до визначення ознак маніпуляцій в повідомленнях з мережі інтернет	24.10 - 15.11	
3.	Алгоритм автоматизованого виявлення ознак маніпуляції, а також виявлення повідомлень з ознаками маніпуляції в мережі інтернет та фільтрування цих повідомлень	23.11 - 25.12	
4.	Реалізація системи виявлення маніпуляційних повідомлень в мережі інтернет	03.01 - 14.02	
5.	Визначення сили маніпуляційності джерел	18.02 - 20.03	
6.	Вступ, висновки, реферат	21.03 - 28.03	
7.	Розробка презентації	28.03 - 04.04	

Студент: _____

Керівник роботи: _____

РЕФЕРАТ

Питання щодо правдивості інформації стоїть дуже критично. Якщо порівнювати кількість джерел інформації в 20 столітті, та в 21 столітті, то ми побачимо, що їх стало набагато більше. Перевіряти таку кількість інформації просто не можливо і немає сенсу робити це у ручну. Через це і виникла потреба в аналізі інформації за допомогою сучасних методів та технічних рішень.

У цій роботі пропонується оригінальний підхід до виявлення ознак маніпуляцій в повідомленнях, виявлення повідомлень з ознаками маніпуляції та їх фільтрування за певною тематикою що базується на математичній лінгвістиці та машинному навчанні. Враховується, що для максимального залучення уваги користувачів мережі маніпуляційні джерела найчастіше генерують повідомлення, що містять ненормативну лексику, так званні меми, сенсаційні епітети тощо. Такі слова і словосполучення можуть виступати маркерами маніпуляцій.

На цей час ресурси мережі Інтернет стають домінуючим джерелом інформації для людей. В умовах жорсткої конкурентної боротьби, в процес інформування втручаються інформаційні джерела, що створюються з метою маніпулювання свідомістю людей.

Актуальність роботи. На цей час ресурси мережі Інтернет стають домінуючим джерелом інформації для людей. В умовах жорсткої конкурентної боротьби, в процес інформування втручаються інформаційні джерела, що створюються з метою маніпулювання свідомістю людей.

У цій роботі пропонується оригінальний підхід до виявлення ознак маніпуляцій в повідомленнях, виявлення повідомлень з ознаками маніпуляції та їх фільтрування за певною тематикою що базується на математичній лінгвістиці та машинному навчанні. Враховується, що для максимального залучення уваги користувачів мережі маніпуляційні джерела найчастіше генерують повідомлення, що містять ненормативну лексику, так званні меми, сенсаційні епітети тощо. Такі слова і словосполучення можуть виступати маркерами маніпуляцій.

Мета роботи – створення інформаційної технології виявлення ознак маніпуляцій в повідомленнях, а також маніпуляційних повідомлень в мережі Інтернет шляхом автоматизованого аналізу інформації із соціальних мереж.

Ключові слова: Інформаційний маніпуляційний вплив, мережа джерел інформації, ознаки маніпуляції, соціальні мережі, тематична фільтрація.

ЗМІСТ

ПЕРЕЛІК СКОРОЧЕНЬ

ПЗ – програмне забезпечення

ПМ – програмний модуль

МП – маніпуляційне повідомлення

МД – маніпуляційне джерело

МС – маніпуляційне слово

ІП – інформаційний потік

ІВ – інформаційний вплив

ВСТУП

Завдання полягає у розв’язанні часткових поставлених задач:

1. Аналіз існуючих підходів до визначення ознак маніпуляцій в повідомленнях з мережі інтернет;

2. Запропонувати та обґрунтувати алгоритми автоматизованого виявлення ознак МП з мережі інтернет;

3. Запропонувати та обґрунтувати алгоритми автоматизованого виявлення повідомлень з ознаками маніпуляцій в мережі інтернет, фільтрування цих повідомлень за потрібною тематикою;

4. Створити інструмент для виявлення МП в мережі інтернет за потрібною тематикою.

Об’єкт роботи – методи виявлення джерел маніпуляційного інформаційного впливу шляхом автоматизованого аналізу інформації із соціальних мереж (мессенджера Telegram).

Предмет роботи – методи і засоби виявлення МП на основі інтелектуального аналізу тексту, математичної статистики та машинного навчання.

Практичне значення отриманих результатів полягає в створенні програмного забезпечення для виявлення МП за певною тематикою, що надасть можливість подальшого застосування у задачах підтримки прийняття рішень на основі моніторингу соціальних мереж. Крім того, розроблене програмно-алгоритмічне забезпечення можна використовувати на практиці в якості готового засобу виявлення і фільтрації МД інформації в умовах гібридних війн.

РОЗДІЛ 1
АНАЛІЗ ІСНУЮЧИХ ПІДХОДІВ ДО ВИЗНАЧЕННЯ ОЗНАК МАНІПУЛЯЦІЙ В ПОВІДОМЛЕННЯХ З МЕРЕЖІ ІНТЕРНЕТ

1.1 Системи виявлення маніпуляційного інформаційного впливу.

На цей час розвиток засобів автоматизації обробки текстів різко знизили вартість поширення фейкових новин у глобальному інформаційному просторі. Більшість користувачів Інтернету стикаються з фальшивими новинами, принаймні, раз на тиждень. Так, судячи з результатів масштабного європейського дослідження - Євробарометра «Фейк-ньюз і дезінформація онлайн», опублікованим у 2018 році, більше третини респондентів (37%) стикаються з підробленими новинами кожен день. 85% респондентів з різних країн хто вважає фейкові новини проблемою для їхньої країни і 83% розглядають їх як проблему для демократії в цілому. Тому зараз як ніколи важливо навчитися розпізнавати маніпуляційну інформацію, щоб вміти надалі протистояти їй.

Відомо, що «статистична обробка множини підроблених статей дозволяє виділити набори ключових слів, які з певною часткою ймовірності сигналізують про можливість, що стаття є підробленою».

До фейкових новин відносяться новини, які не містять строго фактичної інформації, не підходять стандартам журналістської етики, а також ті, що відповідають багатьом іншим характеристикам, які будуть надані нижче у цій роботі.

Щоб дійсно виявити підроблені новини, у загальному випадку потрібне застосування алгоритмів, що дозволяють інтерпретувати людську мову, у загальному випадку мають здійснюватися:

- Перевірка оригінальності URL-адреси повідомлення на відповідність домену сайту. Сайти-підробки можуть цілком імітувати великі новинні сайти.

- Аналіз фото публікації за принципом збігів такого на сайтах зі списку довірених ресурсів.
- Виконується перевірка дати публікації. Фейкові новини, як правило, не датовані, тому що говорять про події, яких не було.
- Визначення наявності граматичних і пунктуаційних помилок - часто підроблені статті формуються з використанням автоматичних генераторів текстів без якісної вичитки. Якщо статтю становить людина, то зазвичай робиться це недбало і з різних фрагментів текстів за шаблоном, що позначається на грамотності тексту.
- Перевірка на наявність внутрішньої узгодженості аналізованої статті. Підроблені або вводять в оману статті часто мають велику неузгодженість між різними частинами тексту самої статті. Наприклад, алгоритм сканує і оцінює заголовок, основний текст, анотацію новини і т.д. на предмет того, чи узгоджені факти, представлені в статті між собою, чи немає між ними конфліктів і логічних невідповідностей.
- Пошук збігів фрагментів різних статей. Часто основна стаття, що вийшла на авторитетному ресурсі, копіюється повністю або шляхом рерайта з внесенням в неї неправдивих перекручень фактів і подій.
- Аналіз посилань вихідних зі статті, ймовірність того, що стаття буде помилковою вище, якщо вона посилається також на помилкові статті, і навпаки. Схожий принцип діє і при визначенні репутації сайтів.
- Пошук сигнальних слів (сенсаційних). Статті, що містять в своїх заголовках і ключових словах привертають увагу (сенсаційні) затвердження часто мають тенденцію бути підробленими. Статистична обробка безлічі підроблених статей дозволяє виділити набори ключових слів, які з певною часткою ймовірності сигналізують про можливість, що стаття є підробленою.

Слід відмітити, що саме реалізації пункту 8 присвячене програмно-методичне забезпечення, що реалізовано у цій роботі.

До числа компаній, які борються з фальшивими новинами, ботами та тролями, приєдналися Facebook, Youtube, Google. Зокрема, Facebook запустив кампанію, в якій використовуються як повідомлення Facebook, так і рекламні оголошення в газетах, щоб надати споживачам поради про те, як ідентифікувати фейк-ньюз. У липні 2018 року про заходи щодо припинення поширення відео-роликів з фальшивими новинами заявив Youtube.

Нижче наведено перелік ресурсів для виявлення фейкових повідомлень:

Botometer

Веб-сайт Botometer (первинна назва BotOrNot) був створений в Університеті Індіани в відповідь на поширення в Twitter ботів, що публікують фейкові новини. Цей сайт оцінює аккаунти за шкалою від одного до п'яти, де один означає, що обліковий запис належить реальним користувачам, а п'ятіркою позначаються фейкові акаунти. Оцінка проводиться на основі твітів, історії публікацій та згадок іншими користувачами.

Fake News Detection

Ця програма, яку можна знайти на GitHub, використовує технології машинного навчання і байєсівські моделі для пошуку фейкових новин.

FactCheck.org

На цьому сайті користувачі можуть задавати питання про достовірність інформації, що звучить в заявах політиків, а команда сайту проводить розслідування і пропонує докладне пояснення. Пояснення включає інформацію про те, ким була зроблена заява, коли воно прозвучало і як команда його перевірила.

У сайту також є спеціальна функція для перевірки наукової інформації – SciCheck.

Fake Bananas

Розробники з коледжу Суортмор створили Fake Bananas - модель машинного навчання, визначальну фейковий новини з точністю 82 відсотки за допомогою технологій машинного навчання. Програма шукає в авторитетних онлайн-виданнях статті, пов'язані з темою висловлювання, яке потрібно перевірити, і аналізує, чи

згодні автори статей до укладеного в висловлюванні твердженням. Якщо достовірні джерела згодні з ним, програма оцінює затвердження як правдиве.

Hoaxy

Hoaxy – це онлайновий інструмент, що візуалізує поширення статей в Інтернеті. Орієнтований на перевірку фейковий новин сайт створює кольорові інтерактивні графіки, даючи користувачам можливість побачити, як різні заяви поширюються в Twitter. Сайт створено в 2016 році, це спільний проект Центру досліджень комплексних мереж і систем (Center for Complex Networks and Systems Research) та Інституту мережевих наук Університету Індіани (Indiana University Network Science Institute).

Politifact

Лауреат Пулітцерівської премії сайт Politifact перевіряє заяви політиків і блогерів і оцінює ці твердження за шкалою від "правда" до "зовсім забрехався" (pants on fire). Сайт був створений у 2007 році редакцією The Tampa Bay Times, а зараз його роботою керує Інститут Пойнтера. Міжнародна мережа з перевірки фактів включила Politifact в свій список кращих ресурсів.

Snopes

З 1994 року Snopes перевіряє достовірність заяв, статей, постів в соціальних медіа та фотографій. Не обмежуючись простими заявами - "правда" або "брехня", Snopes використовує більш детальні категорії: "правда", "брехня", "суміш того й іншого", "в основному правда", "в основному брехня", "застаріла інформація" , "неправильно зрозуміла інформація" та ін. На сайті також можна знайти список сайтів, що поширюють фейковий новини.

1.2 Ознаки маніпуляційного інформаційного впливу і методи його виявлення.

Хибна аргументація

Головна думка підкріплюється «гнилими» аргументами. Іншими словами, між тезою та аргументами встановлені хибні причинно-наслідкові зв'язки.

Замість аргументів пропагандисти тут звертаються до обивательського здорового глузду, який ґрунтується на незнанні реального стану речей та складних механізмів функціонування сучасного суспільства. Також апелюють до загальновідомих речей, які насправді є поширеними міфами.

Ознаки маніпуляційних новини з хибною аргументацією:

- Висловлена теза не підкріплена достовірними аргументами.
- За «аргументи» видаються поширені у суспільстві міфи.
- Не вказуються конкретні дані та їхні джерела.

Маніпуляційний заголовок

Кожен зустрічав багато заголовків на кшталт: "Ви не повірите...", "Виявилось, що", "СЕНСАЦІЯ", "Читати всім!", "Стало відомо", "Шокуюча правда". Трапляється й чимало заголовків, які емоційно повідомляють про жахливий факт або, навпаки, "суперперемогу", на яку довго очікували.

У заголовках такого типу зазвичай використовують надмірно емоційну лексику або обіцяють сенсацію. Головна мета - аби читач клікнув на них і відкрив новину. У тексті новини найчастіше жодної сенсації немає.

Є й інша мета, для якої використовуються такі заголовки. Новин зараз стільки, що не всі відкривають їх, а проглядають тільки заголовки. У цьому випадку, маніпуляція спрямована на те, щоб читач сформував певну думку лише за заголовком, без перегляду самої новини - наприклад, гортаючи News Feed Фейсбуку або стрічку новин на Укрнеті. У таких заголовках зазвичай повно "зради", в яку читач має повірити, не вдаючись у деталі. При цьому текст новини, у деяких випадках, може бути цілком збалансованим.

Ознаки маніпуляційних заголовків:

- Слова написані великими літерами;
- Використання слів і словосполучень типу “Ви не повірите...”, “Виявилось, що”, “СЕНСАЦІЯ”, “Читати всім!”, “Стало відомо”, “Шокуюча правда”, “Це підірвало мережу”;
- Повідомлення лише на одному сайті про щось дуже жахливе або неймовірно позитивне.

Емоційно упереджена новина

Спосіб з нормальної людини зробити корисного ідіота. Основне завдання – викликати злість і вказати на винного. Використовується емоційна лексика: перебільшення, невідповідні або надумані епітети, негативно забарвлені слова. Повідомляється про щось надзвичайно обурливе.

Ознаки емоційно упередженої новини:

- Ви читаєте - і ваше серце починає битися швидше;
- У тексті багато епітетів, тобто слів, які підкреслюють характерні риси;
- Вам повідомляють про неймовірно жахливі чи дуже радісні речі.

Політичні змови

Описують кулуарні домовленості. Інформація подається буцімто від осіб, які знають ситуацію з середини, написано фамільярно. Серед іншого: «Петя» - замість Петро Порошенко, «Юля» - замість Юлія Тимошенко, «Юра» («Юґа») - замість Юрій Луценко.

Новин цієї категорії небагато - частіше ми побачимо використання подібного прийому у (псевдо) розслідуваннях.

Часто використовують заплутану аргументацію, а інформація подається як інсайдерська. Часто апелюють до неназваних джерел інформації. Перевірити, чи

видання дійсно спілкувалося з якимось «джерелом», і чи можна йому довіряти, неможливо.

Такі тексти будуються наступним чином (тут ми для наочності трохи перебільшуємо): “Політик А дає хабарі антикорупціонеру Б, бо звідки ж іще у Б з’явився новий “Опель”? Політик А, до речі, ходив у дитсадок, де бабуся антикорупціонера Б була подругою вихователькі”.

Ознаки:

- ❖ Написано фамільярно;
- ❖ Використовують заплутану аргументацію.

Сьогодні інформаційна війна Росії проти України ведеться трьома основними «роями»:

1. Респектабельні проросійські ЗМІ, які працюють в Україні. Вони мають гарне фінансування, яке дозволяє найняти гарних редакторів та журналістів, працювати оперативно. Вони якісно «упаковують» свою дезінформацію. З формальної точки зору, їхні повідомлення «чисті».

2. «Зливні бачки». Декілька сотень інтернет-сайтів, які часто під виглядом новин публікують відверті вигадки та напівправду, лише зрідка домішуючи в стрічку правдиві новини.

3. Боти, тролі і групи у соцмережах. Група може мати патріотичну назву - і постити, здавалося б, патріотичний контент. Однак в нього вплітаються меседжі, які потрібні російській пропаганді. Відомий випадок, коли популярну у Фейсбуку групу Патріоти України модерував колишній бойовик ДНР, котрий на той час мешкав у Москві.

Першим кроком, очевидно, для вирішення завдання виявлення джерела фейкових новин можна вважати отримання даних великого обсягу (корпус). Це завдання вирішується за допомогою роботів, реалізованих, зокрема, в системі Cyber Aggregator для десятка соціальних мереж, в тому числі і мессенджера Telegram. Потім можна аналізувати використовувану в тексті лексику. В цьому

випадку тексти корпусу необхідно спочатку піддати автоматичній обробці: розмітити частини мови, виявити емоційно забарвлені слова, імена (в тому числі медіа-персон) і все це порахувати. Робити цю підготовчу роботу лінгвістам допомагають готові програми (наприклад, LIWC, MyStem). Подальше завдання дослідника полягає в тому, щоб, проаналізувавши отримані дані, визначити, які лексичні ознаки є значущими для класифікації. Серед таких маркерів можуть застосовуватися довжина слів, частотність прикметників, спілок, числівників, цитат, знаків оклику та емоційної лексики.

1.3 Актуальність і мета розробки програмного та інформаційного забезпечення.

В умовах інформаційних війн в процес інформування втручаються інформаційні джерела, що створюються з метою маніпулювання свідомістю людей.

Якщо на цей час у світі створюються технології і сервіси, що призначені для виявлення і аналізу окремих фейкових новин, то задачам виявлення джерел, ознак маніпуляції в них та тематичної направленості їх повідомлень приділено суттєво менша увага, тому що ця інформація менш цікава для пересічного користувача. Проте саме виявлення ознак маніпуляції в повідомленнях важливо для підрозділів великих корпорацій, державних установ, засобів масової інформації. У цій роботі пропонується підхід до виявлення ознак маніпуляції в повідомленнях, повідомлень з ознаками маніпуляції та фільтрування цих повідомлень за потрібною тематикою, що базується на математичній лінгвістиці, інтелектуальному аналізу тексту, математичній статистиці та машинному навчанню.

Створення методології і інформаційної технології виявлення ознак маніпуляції в повідомленнях, повідомлень з ознаками маніпуляції та фільтрування цих повідомлень за потрібною тематикою сьогодні актуально для задач змістовного аналізу мережевої інформації, дослідження суспільної думки, виявлення інформаційних атак і операцій, фільтрації впливу на людей. Проблема

сьогодні остаточно не розв’язана, їй займаються дослідники в усьому світі – вона складна і потребує великих витрат.

Висновок до розділу 1.

За результатами аналізу сучасного стану методологій і інформаційних технологій виявлення джерел деструктивного ІВ в мережі Інтернет було встановлено, що побудова відповідних систем і сервісів – складна і витратна проблема, що на цей час спрямована, насамперед, на виявлення ознак фейкових новин, а не на виявлення маніпуляційних інформаційних джерел і на їх тематичну направленість.

Разом з цим, встановлено, що існує декілька підходів до виявлення фейкових новин з текстових корпусів, що приводить, відповідно, до різних видів реалізації технологій. Повідомлення можуть бути визначені як фейкові за допомогою технологій розпізнавання образів, глибинного навчання (Deep Learning), байєсівських алгоритмів машинного навчання.

Разом з цим, інтеграційний підхід, що запропоновано у цій роботі забезпечує визначення саме ознак маніпулювання в повідомленнях, повідомлень з ознаками маніпуляції та фільтрування їх за потрібною нам темою. І цьому факту надано математичне обґрунтування. У цій роботі пропонується підхід до виявлення ознак маніпуляцій в повідомленнях, що базується на математичній лінгвістиці, інтелектуальному аналізу тексту, математичній статистиці та машинному навчанню.

Тому пропонується розробити систему виявлення ознак маніпуляції в повідомленнях на основі аналізу текстових корпусів з мережевих ЗМІ, соціальних мереж, месенджерів.

РОЗДІЛ 2
АЛГОРИТМ АВТОМАТИЗОВАНОГО ВИЯВЛЕННЯ ОЗНАК МАНІПУЛЯЦІЇ, А ТАКОЖ ВИЯВЛЕННЯ ПОВІДОМЛЕНЬ З ОЗНАКАМИ МАНІПУЛЯЦІЇ В МЕРЕЖІ ІНТЕРНЕТ ТА ФІЛЬТРУВАННЯ ЦИХ ПОВІДОМЛЕНЬ

2.1 Методика виявлення ознак маніпуляцій в повідомленнях з мережі інтернет.

Методика виявлення ознак маніпуляцій в повідомленнях з мережі Інтернет передбачає два етапи:

- навчання системи, що передбачає формування словників індикаторів маніпуляційних впливів і стоп-словника (словника незначущої для проблеми лексики);
- застосування навченої системи.

2.1.1 Методика навчання системи.

Навчання системи, що пропонується, базується на автоматизованій обробці ІП, екстрагування слів, їх автоматизованої селекції. Процедура навчання здійснюється на постійній основі, по мірі обробки вхідного ІП. Результати навчання – словники ознаки передаються для виявлення МП – застосування навченої системи. На рис. 2.1 наведено схему навчання системи. Необхідно зазначити, що на цій схемі не передбачено стандартних блоків «Початок» і «Кінець» виходячи із неперервності процесу навчання.

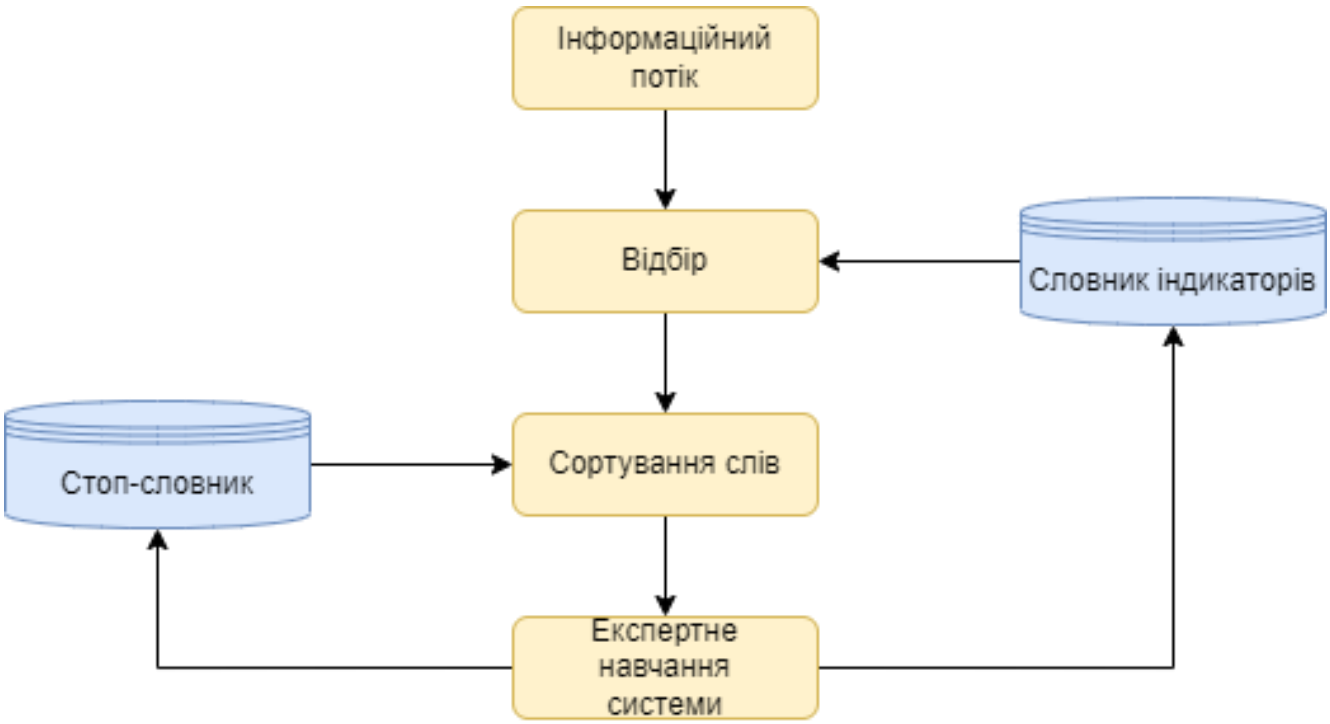


Рисунок 2.1 – Блок-схема етапу навчання системи

- ❖ Формування ІП (масиву вхідних повідомлень), що сканується із соціальних мереж або месенджерів (у цій як полігон для представлення методики використовувалися повідомлення мессенджера Telegram);
- ❖ Формування тимчасового словника слів, що можуть маркувати повідомлення маніпуляційного характеру (термінів);
- ❖ Фільтрація масиву вхідних повідомлень за допомогою цього словника;
- ❖ Сортування слів, та вивід найбільш популярних для подальшого розгляду;
- ❖ Віднесення кожного слова до відповідного словника.

2.1.2 Для виявлення маніпуляцій в повідомленнях на базі виявлених ознак використовується теорема Байєса.

Томас Байєс (англ. Reverend Thomas Bayes, 1702-1761) англійський математик і пресвітеріанський священик, член Лондонського королівського товариства.

Наївні класифікатори Байєса використовуються для фільтрації МП. Використовуються функції набору слів для виявлення маніпуляцій в повідомленнях. Наївні байєсівські класифікатори працюють, співвідносячи використання токенів (зазвичай слів або, іноді, інших речей), з маніпуляцією і не маніпуляцією, а потім використовуючи теорему Байєса для обчислення ймовірності того, що повідомлення є або не є маніпуляційним. Отож, використовуємо наївну байєсівську фільтрацію, як базовий метод боротьби з маніпуляційністю.

Байєсівський фільтр виявлення МП ґрунтується на теоремі Байєса, яка використовується кілька разів в цьому контексті:

- в перший раз, щоб обчислити вірогідність, що повідомлення маніпуляційний, знаючи, що дане слово з'являється в цьому повідомленні;
- вдруге, щоб обчислити вірогідність, що повідомлення маніпуляційний, враховуючи всі його слова (або відповідні їх підмножини);
- іноді в третій раз, коли зустрічаються повідомлення з рідкісними словами.

Обчислення ймовірності того, що повідомлення, що містить дане слово, є маніпуляційним.

Обчислення базуються на формулі, яка отримана з теореми Байєса і формули повної ймовірності:

$$Pr(F|W) = \frac{Pr(W|F)\cdot Pr(F)}{Pr(W)} = \frac{Pr(W|F)\cdot Pr(F)}{Pr(W|F)\cdot Pr(F)+Pr(W|H)\cdot Pr(H)} \quad 2.1$$

- **Pr(F|W)** – умовна ймовірність того, що повідомлення маніпуляційний, за умови, що слово-індикатор знаходиться в ньому;
- **Pr(F)** - повна ймовірність того, що довільне повідомлення є маніпуляційним;
- **Pr(W|F)** – умовна ймовірність того, що слово-індикатор з'являється в повідомленнях, якщо вони є маніпуляційними;
- **Pr(H)** - повна ймовірність того, що довільне сполучення не маніпуляційний;

❖ **Pr(W|H)** - умовна ймовірність того, що слово-індикатор з'являється в повідомленнях, якщо вони є не спамом.

2.1.3 Маніпуляційність слова

Статистика показала, що поточна ймовірність того, що будь-яке повідомлення є маніпуляційним, становить 80%, як мінімум:

$$Pr(F) = 0.8;\ Pr(H) = 0.2 \quad 2.2$$

Тим не менше, більшість байєсівського програмного забезпечення для виявлення маніпуляційності виходить з того, що немає ніяких апріорних причин, щоб будь-яке вхідне повідомлення було маніпуляційним, а не звичайним, і вважає, що обидва випадки мають рівні ймовірності 50%:

$$Pr(F) = 0.5;\ Pr(H) = 0.5 \quad 2.3$$

Фільтри, що використовують цю гіпотезу, називаються «необ'єктивними», що означає, що вони не мають упереджень щодо вхідних повідомлень. Це припущення дозволяє спростити загальну формулу Байеса:

$$Pr(F|W) = \frac{Pr(W|F)}{Pr(W|F) + Pr(W|H)} \quad 2.4$$

Функціонально це еквівалентно запитом "який відсоток випадків появи слова" бандери "з'являється в МП?"

Цю величину назвемо «маніпуляційність» (або «маніпуляційність») слова «бандери» яку можемо обчислити. Pr(W|F) використовувана в цій формулі апроксимується частотою повідомлень, що містять слово «бандери» в повідомленнях, ідентифікованих як маніпуляційні, на етапі навчання. За аналогією, Pr(W|U) апроксимується частотою повідомлень, що містять слово «бандери» в повідомленнях, на етапі навчання. Щоб ці наближення мали сенс, набір вивчених повідомлень повинен бути досить великим і репрезентативним.

Звичайно, визначення того, чи є повідомлення маніпуляційним чи ні, засноване тільки на присутності слова «бандери», схильний до помилок, тому розроблене ПЗ намагається розглянути кілька слів і об'єднати їх маніпуляційність, щоб визначити загальну ймовірність МП.

Точно так же приблизно однаково відносної частоті повідомлень, що містять слово в повідомленнях, ідентифікованих як і маніпуляційний.

Програмні фільтри, побудовані на принципах наївного байєсівського класифікатора, роблять «наївне» припущення про те, що події, відповідні наявності того чи іншого слова в електронному листі або повідомленні, є незалежними по відношенню один до одного. Це спрощення в загальному випадку є невірним для природних мов - таких, як англійська, де ймовірність виявлення прикметника підвищується при наявності, наприклад, іменника. Виходячи з такого «наївного» припущення, для вирішення задачі класифікації повідомлень лише на 2 класи: F (маніпуляції) і $H = \neg F$ (не маніпуляції) з теореми Байєса можна вивести таку формулу оцінки ймовірності «маніпуляційного» всього повідомлення, що містить слова W_1, W_2,…, W_N:

$$p(F|W_1, W_2, \ldots, W_N) = \quad 2.5$$

По теоремі Байеса

$$= \frac{p(W_1, W_2, \ldots, W_N | F)\, p(F)}{p(W_1, W_2, \ldots, W_N)} = \quad 2.6$$

так як W_i, передбачаються незалежними

$$= \frac{\Pi_i p(W_i) p(F)}{p(W_1, W_2, \ldots, W_N)} = \quad 2.7$$

По теоремі Байеса

$$= \frac{\Pi_i \frac{p(F|W_i)p(W_i)}{p(F)} p(F)}{p(W_1, W_2, \ldots, W_N)} = \quad 2.8$$

За формулою повної ймовірності

$$= \frac{\Pi_i \frac{p(F|W_i)p(W_i)}{p(F)} p(F)}{\Pi_i\,(p(W_i|F))p(F) + \Pi_i(p(W_i|\neg F))p(\neg F)} =$$

$$= \frac{\Pi_i\, p(F|W_i)p(W_i)p(F)^{1-N}}{\Pi_i\,(p(F|W_i)p(W_i))p(F)^{1-N} + \Pi_i(p(\neg F|W_i)p(W_i))p(F)^{1-N}} =$$

$$= \frac{\Pi_i\, p(F|W_i)p(W_i)p(F)^{1-N}}{\Pi_i\,(p(F|W_i)) + \left(\frac{p(\neg F)}{p(F)}\right)^{1-N} \Pi_i\,(p(\neg F|W_i)} \quad 2.9$$

Таким чином, припускаючи, що Pr(F)=Pr(H)=0.5, маємо:

$$p = \frac{p_1, p_2, \ldots, p_N}{p_1 p_2 \ldots p_N + (1-p_1)(1-p_2)\ldots(1-p_N)} \quad 2.10$$

- $\boldsymbol{p = \Pr(F|W_1, W_2, \ldots, W_N)}$ - ймовірність, що повідомлення, що містить слова - W_1, W_2,…, W_N маніпуляційне;
- $\boldsymbol{p_1}$ – умовна ймовірність того, що повідомлення маніпуляційне, за умови, що воно містить перше слово;
- $\boldsymbol{p_2}$ – умовна ймовірність того, що повідомлення - маніпуляційне, за умови, що воно містить другий маніпуляційний слово;
- $\boldsymbol{p_N}$ – умовна ймовірність того, що повідомлення маніпуляційне, за умови, що воно містить *N*-е слово.

Результат *p* зазвичай порівнюють з деяким порогом (наприклад, 0.5), щоб вирішити, чи є повідомлення маніпуляційним чи ні. Якщо *p* нижче, ніж поріг, повідомлення розглядають як певно не маніпуляційне, інакше його розглядають як ймовірно маніпуляційне.

2.1.4 Процес виявлення ознак маніпуляції.

Окремі слова мають певну ймовірність появи, як в маніпуляційному так і в звичайному повідомленні. Наприклад, більшість користувачів мессенджера «Telegram» часто зустрічають слово «порохобот» в маніпуляційному, але рідко в звичайному повідомленні. Фільтр не знає цих ймовірностей заздалегідь і повинен спочатку пройти навчання, щоб він міг їх створити. Щоб навчити фільтр, користувач повинен вручну вказати, чи є нове повідомлення маніпуляційним чи ні. Для всіх слів в кожному навчальному джерелі фільтр буде коригувати ймовірність того, що кожне слово з'явиться в маніпуляційному або в легітимному повідомленні в його базі даних. Наприклад, байєсовські фільтри дали високу ймовірність маніпуляційності для слова «порохобот» і «соросята», але дуже низьку ймовірність для слів, що зустрічаються тільки в законних електронних листах, таких як імена друзів і членів сім'ї.

Початкове навчання зазвичай може бути уточнено під час виявлення маніпуляційних ознак від програмного забезпечення (помилкові спрацьовування або помилкові заперечення). Це дозволяє програмному забезпеченню динамічно адаптуватися до постійно мінливої природи маніпуляційності.

2.2 Методика виявлення повідомлень з ознаками маніпуляції в мережі інтернет та їх тематична фільтрація.

2.2.1 Етапи методики виявлення повідомлень з ознаками маніпуляцій та фільтрування цих повідомлень за потрібною тематикою.

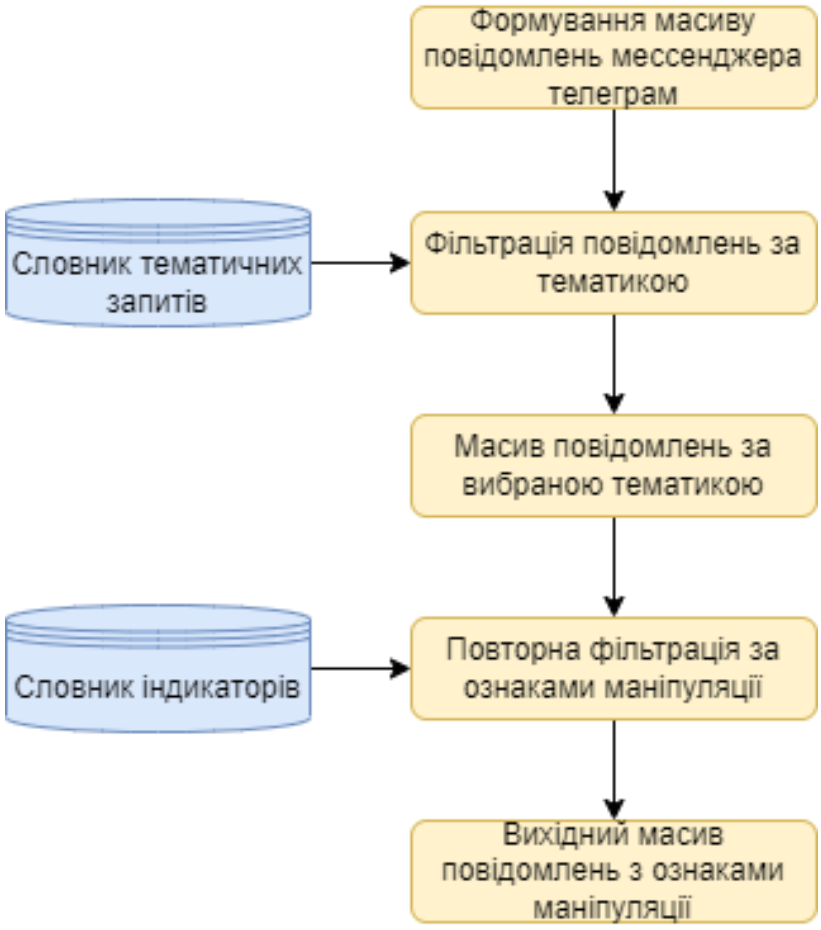


Рисунок 2.2 – Блок-схема етапу сталого функціонування системи

Перший етап (фільтрування), виконує наступне:

1. Формування масиву вхідних повідомлень, що скануються із соціальних мереж або месенджерів (у цій як полігон для представлення методики використовувалися повідомлення мессенджера Telegram);

2. Здійснюється фільтрація за допомогою тематичного словника, слова якого маркують повідомлення за потрібною нам темою.

3. Результати фільтрації, представлені у вигляді записів, що складаються із двох полів – назва джерела і повідомлення, сортування здійснюється за назвами джерел.

Другий етап (етап виявлення), містить такі кроки:

1. Фільтрація масиву вхідних повідомлень за допомогою детального словника слів і словосполучень, що маркують МП;

2. Результати фільтрації записуються так само як і в попередньому етапі. Далі підраховується кількість МП від кожного джерела. Джерела, від яких кількість повідомлень перевищує мінімум, виводяться для подальшого аналізу. Після чого результат виводиться у вигляді кругової діаграми. Яка показує силу спрямованості кожного відібраного джерела за вибраною нами темою.

2.2.2 Фільтрація вхідних повідомлень

Фільтрація масиву вхідних повідомлень за допомогою словника індикаторів, що маркують маніпуляційне повідомлення.

Результати фільтрації, представлені у вигляді записів, що складаються із двох полів – назва джерела і повідомлення, сортування здійснюється за назвами джерел.

Нехай G – множина слів, що маркують маніпуляції: $G = \{g_i\}_{i=1}^{|G|}$

Позначимо множину джерел як S: $S = \{S_k\}_{k=1}^{|S|}$

Ймовірність того, що повідомлення d є маніпуляційним, якщо воно містить слово g_i , позначимо як p_i. Відповідно, ймовірність того, що повідомлення d не є маніпуляційним, якщо воно містить слово g_i , позначимо як q_i $(p_i + q_i = 1)$.

Нехай повідомлення d містить декілька слів з G: $\exists i$: $g_i \in d, g_i \in G$.

Ймовірність того, що повідомлення не є маніпуляційним, у цьому випадку дорівнює:

$$q(d) = \prod_{i:\, g_i \epsilon\, d, g_i \epsilon\, G} (1 - p_i) \qquad 2.11$$

Тоді ймовірність того, що повідомлення є маніпуляційним, дорівнює:

$$p(d) = 1 - \prod_{i:\, g_i \epsilon\, d, g_i \epsilon\, G} (1 - p_i) \qquad 2.12$$

Висновок до розділу 2.

В результаті розробки розділу запропоновано та обґрунтовано методику автоматизованого виявлення ознак маніпуляції в повідомленнях, повідомлень з ознаками маніпуляції з мережі інтернет, базується на математичній лінгвістиці та машинному навчанні.

За допомогою даної методики забезпечується вибір найважливіших термінів, створення словника індикаторів маніпуляційного ІВ. Побудовану мережу можна використовувати в якості основи класифікації, виявлення спрямованості маніпуляційних джерел за певною тематикою. Побудовану методику можна реалізувати в якості інструментального засобу виявлення МП за потрібною тематикою в умовах гібридних війн.

РОЗДІЛ 3
РЕАЛІЗАЦІЯ СИСТЕМИ ВИЯВЛЕННЯ МАНІПУЛЯЦІЙНИХ ПОВІДОМЛЕНЬ В МЕРЕЖІ ІНТЕРНЕТ

3.1 Формування масиву повідомлень мессенджера Telegram

Для формування масиву повідомлень мессенджера Telegram програмно здійснюється сканування відповідних (заздалегідь визначених) Telegram-каналів за адресами вигляду:

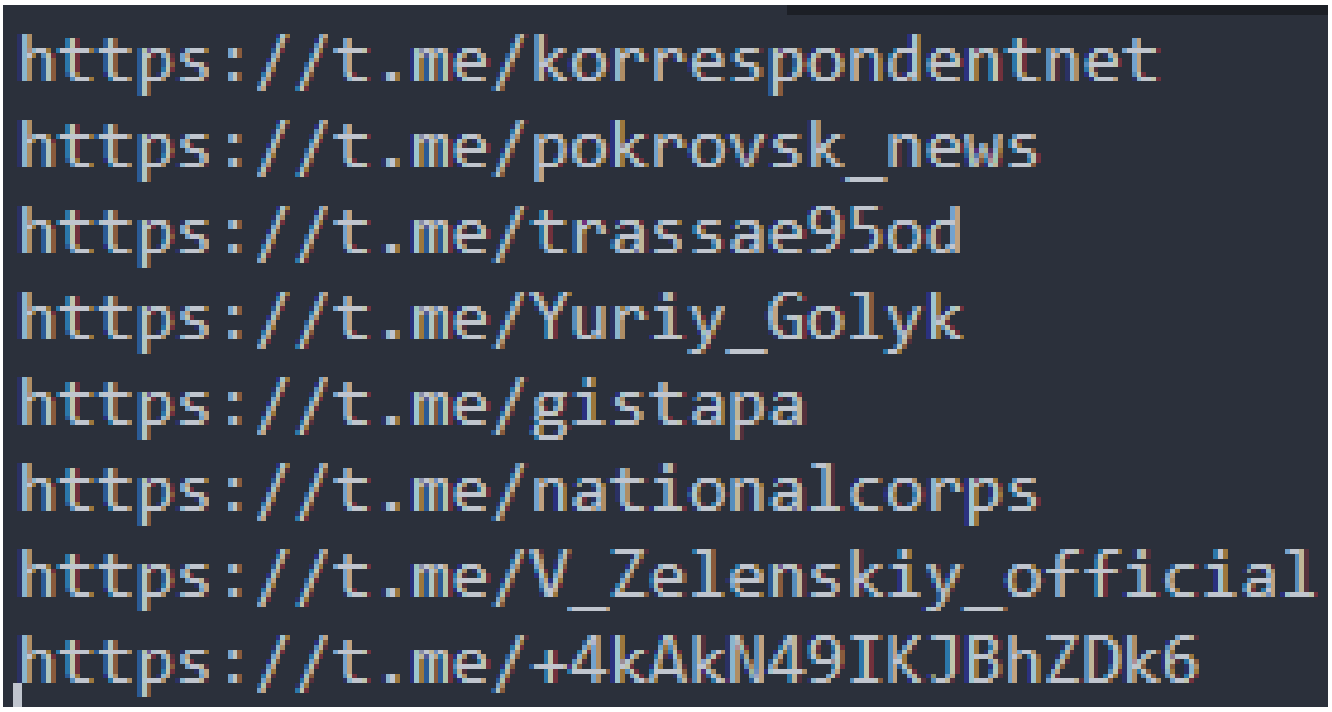


Рисунок 3.1 – приклад адрес Telegram-каналів

Сканування здійснюється за допомогою Selenium WebDriver (див. рис. 3.2).

Selenium WebDriver - це інструмент для автоматизації дій веб-браузера. У більшості випадків використовується для тестування Web-додатків, але цим не обмежується. Зокрема, він може бути використаний для виконання рутинних завдань адміністрування сайту або регулярного отримання даних з різних джерел (сайтів).

Selenium WebDriver - це насамперед набір бібліотек для різних мов програмування.

Проектом Selenium і співтовариством підтримується робота з наступними браузерами: Microsoft Internet Explorer, Google Chrome, Mozilla Suite та Mozilla Firefox під управлінням операційних систем Microsoft Windows, Linux і Apple Macintosh.

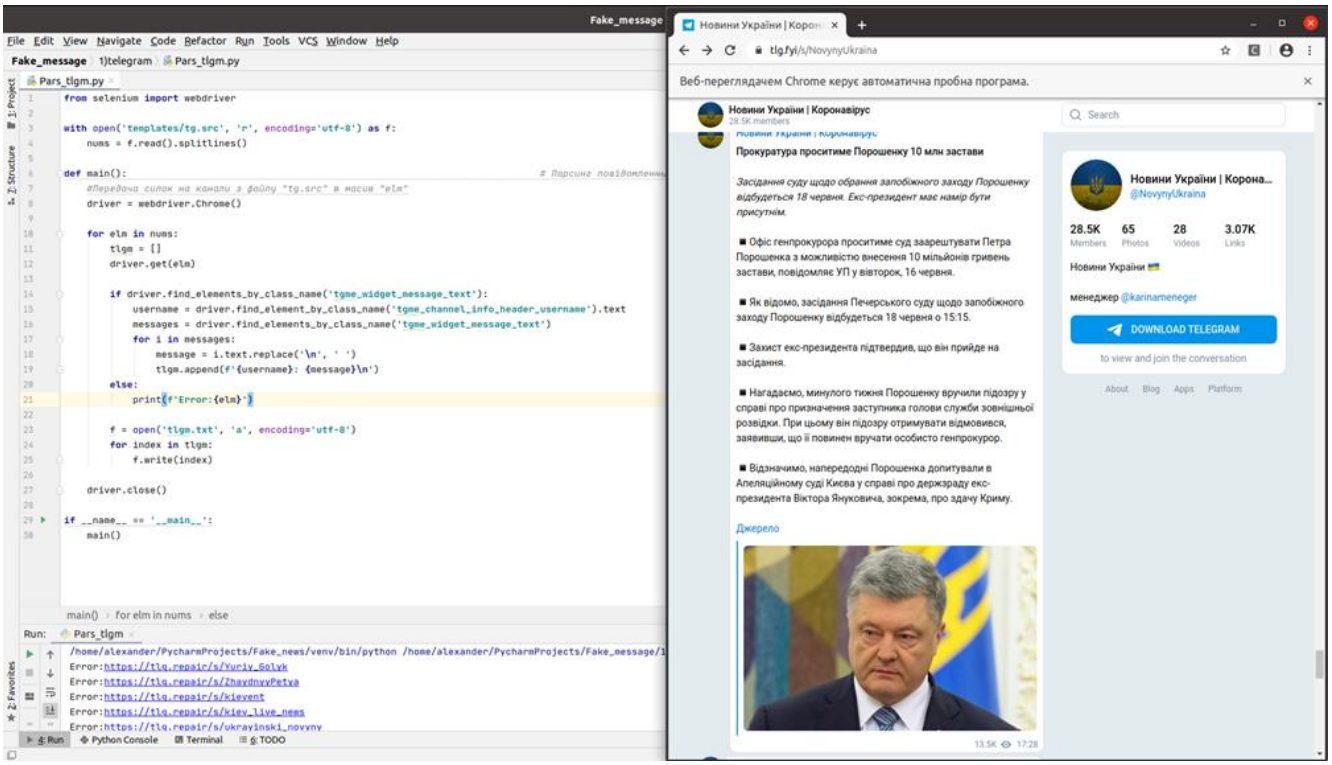


Рисунок 3.2 – приклад роботи парсера

Використовуючи її у Python, парсимо потрібні дані та відразу записуємо їх до потрібного формату. При виявленні помилки, програма виводить інформацію, з яким джерелом сталася помилка. Результати парсингу, представлені у вигляді записів, що склдаються із двох полів – назва джерела і заголовок (див. рис. 3.3)

Текст програми наведено у Додатку 1

Рисунок 3.3 – Формат файлу на виході з програми

3.2 Програмне забезпечення Байєсівського машинного навчання

На початку здійснюється ручний пошук повідомлень маніпуляційного характеру, які подаються на вхід програмного модуля, що наведений у додатку 2. Після чого програма здійснює перевірку поданих повідомлень, та виводить для подальшого розгляду топ-100 слів, що найчастіше зустрічаються.

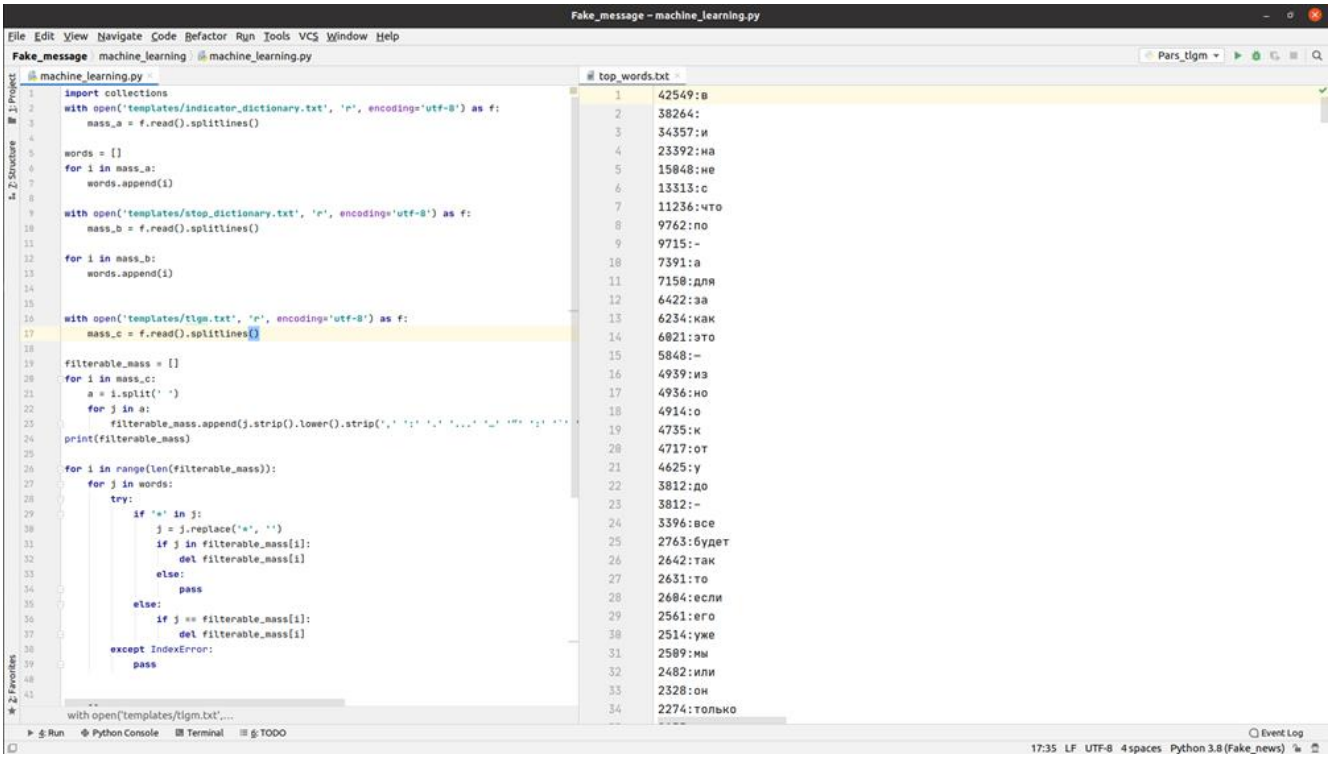


Рисунок 3.4 – Приклад роботи ПЗ байєсівського машинного навчання та файлу на виході з програми

За допомогою Qt Designer створюється інтерфейс у якому буде виведено для розгляду топ-100 слів.

Qt Designer - програма, що служить для розробки графічного користувацького інтерфейсу (GUI), використовуючи бібліотеку Qt. Дозволяє створювати форми й діалоги у візуальному режимі (WYSIWYG). Поставляється з бібліотекою Qt й іншими програмами для Qt: Qt Linguist і Qt Assistant. Являється кросплатформенною.

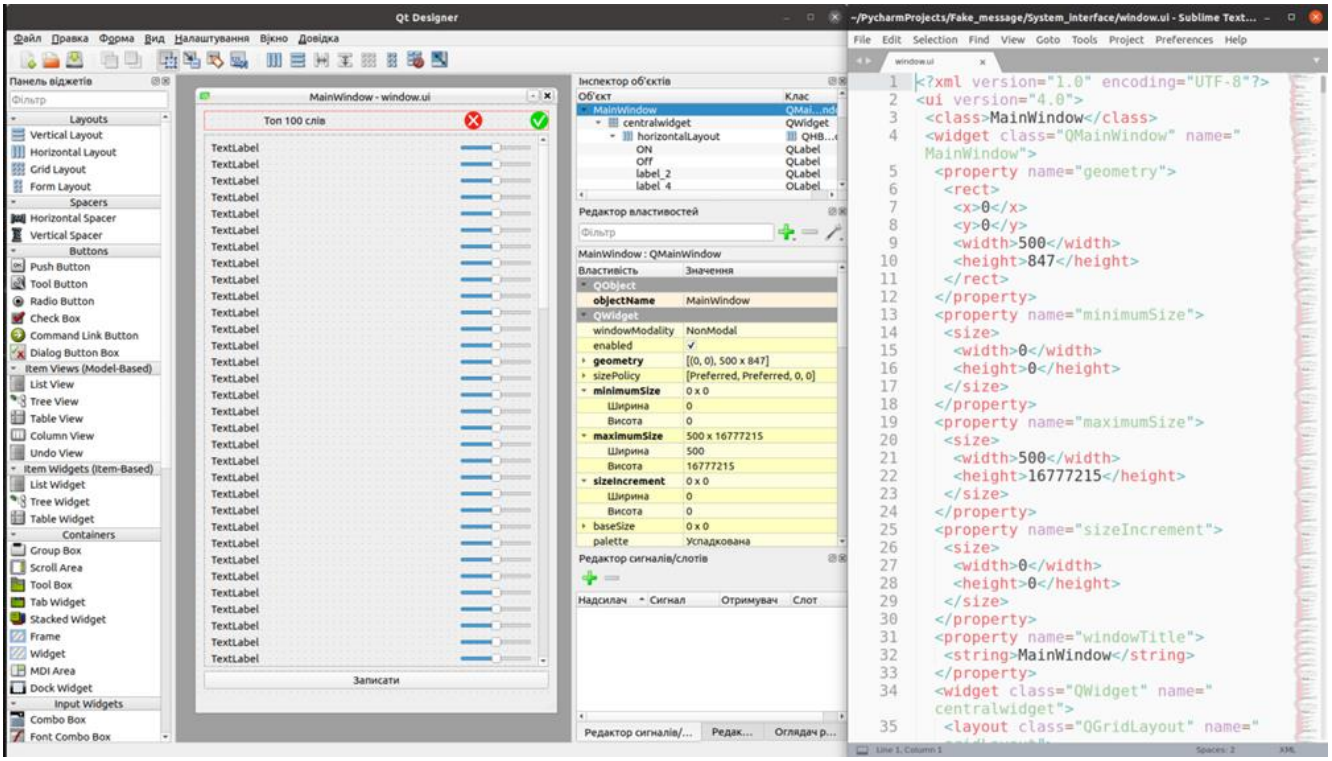


Рисунок 3.5 – Приклад роботи з Qt Designer. Код проекту в форматі .ui

Топ-100 слів, що дає для опрацювання інтерфейс, передаються на обробку експерту, який навчає систему. Експерт передивляючись кожне слово, вирішує у який словник його записати. ПМ інтерфейсу наведений у ДОДАТКУ 3. Інтерфейс побудовано наступним чином: ліва частина – слово, права частина – горизонтальний слайдер, який приймає 3 положення:

1. Ліве – запис у Стоп-словник
2. Центральне – не виконує ніякої дії (у випадку коли важко вирішити до якого словника віднести слово)
3. Праве – запис у словник Індикаторів

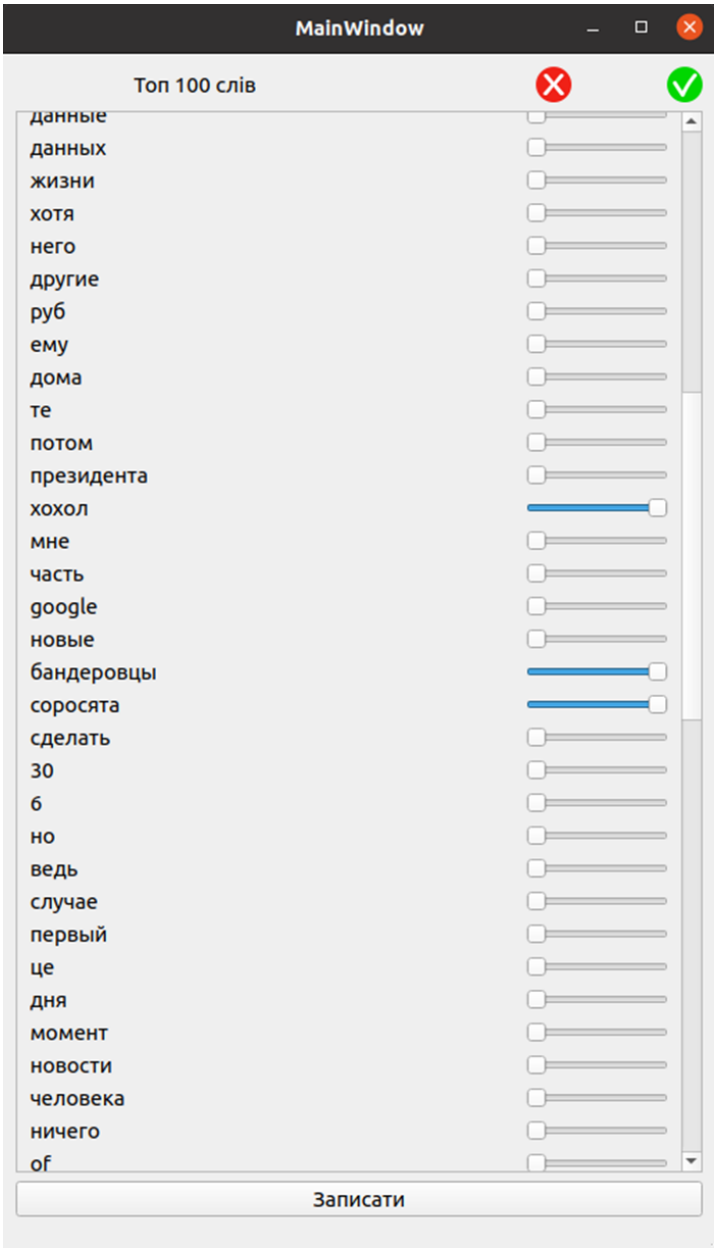


Рисунок 3.6 – Інтерфейс программного модуля машинного навчання

Після тривалого навчання системи отримуємо два словники, а саме словник індикаторів та стоп-словник, які дають можливість виявляти МП. Навчання системи та поповнення словників здійснюється безперервно, так як постійно з’являються нові емоційно-забарвлені слова, що можуть маркувати МП. Приклад словників наведено на рис. 3.6, зліва словник індикаторів, зправа стоп-словник.

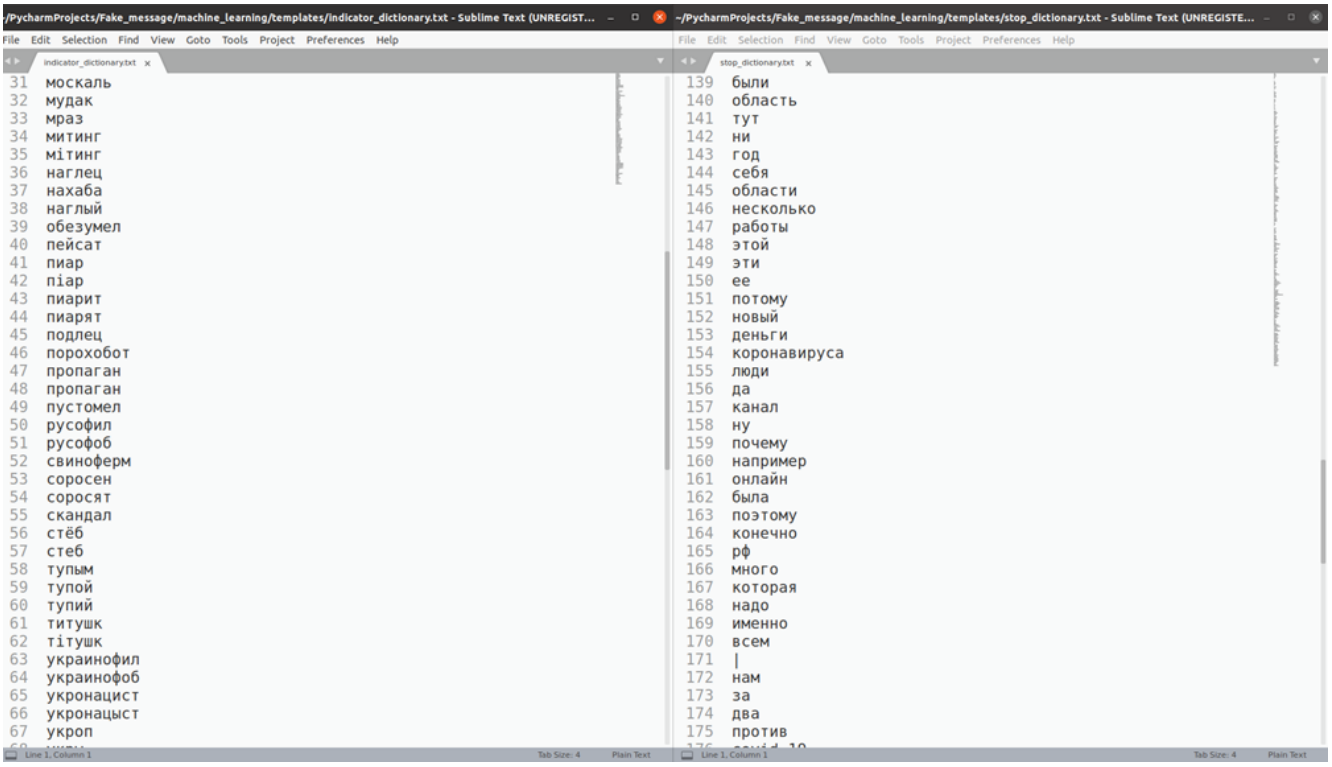


Рисунок 3.7 – Приклад словників отриманих після навчання системи

3.3 Фільтрування повідомлень за обраною тематикою

Для фільтрації вхідних повідомлень за обраною тематикою розроблений окремий програмний модуль, текст якого наведено у Додатку 4.

На вхід програми подається масив повідомлень Telegram сформований при роботі з Selenium WebDriver. Іншим файлом передається словник зі словами та термінами, що маркують потрібну нам тему. Словник створюється вручну.

Для наведення прикладу використовуємо тематику пов’язану з діючим президентом України Зеленським Володимиром Олександровичем.

На виході програми отримуємо масив повідомлень відфільтрованих за потрібною тематикою.

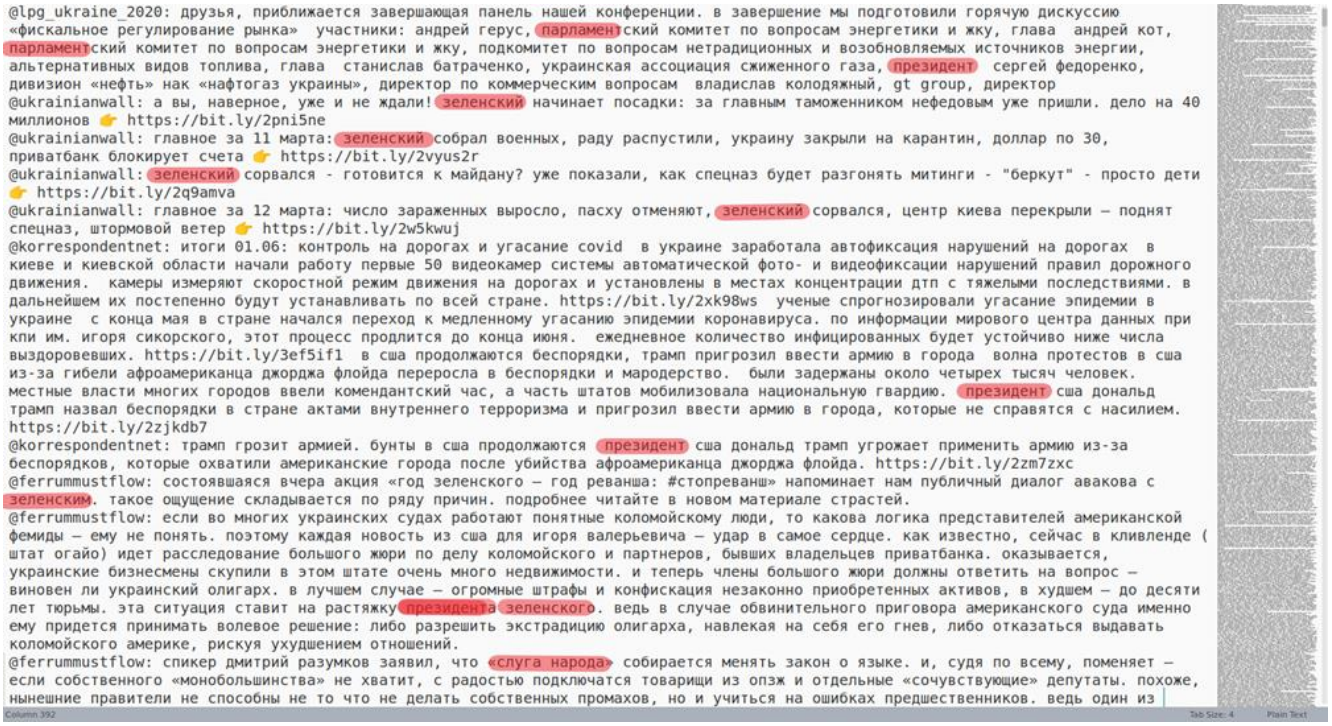


Рисунок 3.8 – Приклад файлу на виході з програми

3.4 Пошук повідомлень з ознаками маніпуляції

Наступним кроком здійснюємо пошук МП у масиві повідомлень попередньо відібраних за тематикою. Використовуємо програмний модуль що наведений у пункті 3.3. На вхід програми передаємо масив повідомлень, що показано на рис. 3.8. Іншим файлом передаємо словник індикаторів сформований під час навчання системи виявлення ознак маніпуляції.

Після опрацювання програмою вхідного масиву повідомлень отримуємо повідомлення з ознаками маніпуляції.

Рисунок 3.9 – Приклад маніпуляційних повідомлень

3.5 Визначення сили маніпуляційності джерел

Визначення сили маніпуляційності джерел здійснюється наступним чином. Програмний модуль (Див. Дод. 5) здійснює підрахунок МП від кожного користувача та записує їх у порядку спадання (див. рис. 3.10). Таким чином визначається сила маніпуляційності кожного джерела.

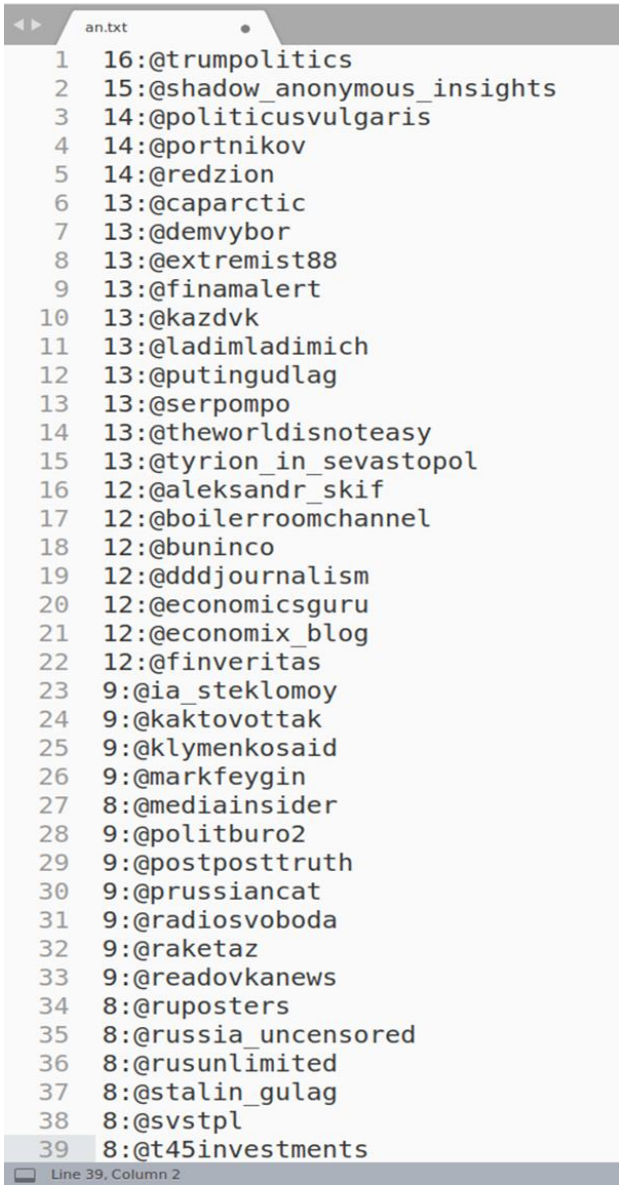


Рисунок 3.10 – Приклад визначення сили маніпуляційності джерел

3.6 Відображення сили маніпуляційності джерел у вигляді кругової діаграми.

Для легшого сприйняття відібрані МД виводяться у вигляді кругової діаграми зображеної на рис. 3.11. Діаграма показує силу кожного МД, що націлене на обрану тематику, а саме джерела які ведуть інформаційну війну проти діючого президента України.

Так як кількість джерел може бути необмежено великою, візуальне сприйняття діаграми буде незручним, тому задаємо умову, якщо кількість повідомлень від джерела менше наприклад 2-х, то такі джерела до уваги не беруться.

Програмний модуль для створення діаграми наведено у ДОДАТКУ 6.

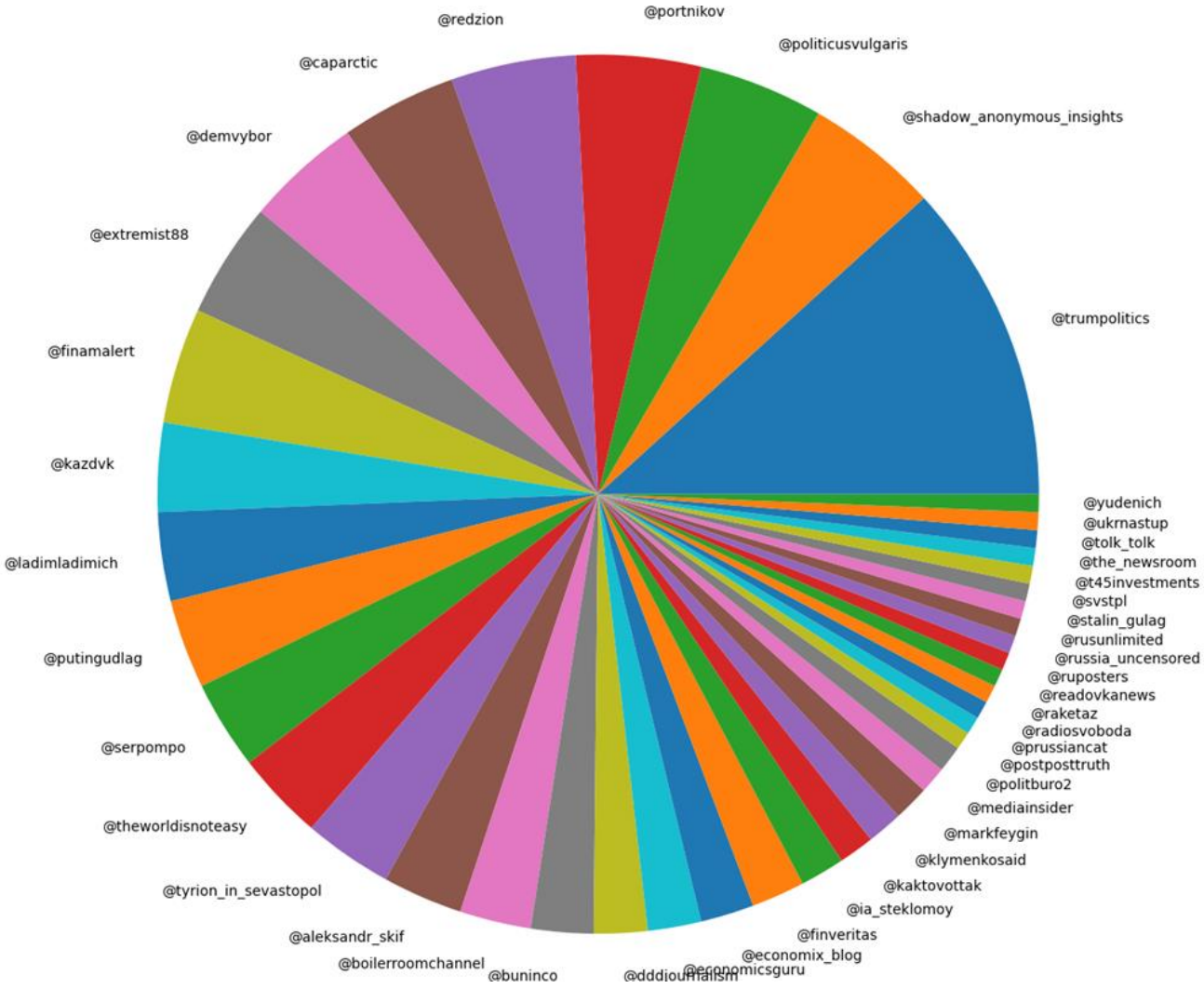


Рисунок 3.11 – Кругова діаграма МД

Висновок до розділу 3.

В результаті розробки розділу було сформовано масив публікацій із мессенджера Telegram. Розроблений комплекс програмних модулів для автоматизованої побудови і візуалізації мережі понять предметної області, за допомогою якого оброблено новини з мережевих ЗМІ та створено мережу понять предметної області.

ВИСНОВКИ

По-перше, на сьогоднішній день приділено мало уваги пошуку ознак маніпуляції в повідомленнях з мережі інтернет, зокрема пошук повідомлень з ознаками маніпуляції - складна і витратна проблема. Тому, визначено, що для пошуку ознак маніпуляцій найбільш результативний підхід – байєсівського машинного навчання.

По-друге, у роботі наведена методика навчання системи виявлення ознак маніпуляції в повідомленнях, автоматизованого виявлення повідомлень з ознаками маніпуляції та фільтрування їх за потрібною тематикою.

По третє, було розроблено комплекс програмних модулів для навчання системи виявлення ознак маніпуляції в повідомленнях, автоматизованого виявлення повідомлень з ознаками маніпуляції та фільтрування їх за потрібною тематикою.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Albert-László Barabási. Network Science. Cambridge University Press, 2016.

2. Al-Heeti A. Facebook Will Fight Fake News with Real Newspaper Ads (and More). CNET. May 23, 2018.

3. Building a better news experience on YouTube, together. URL: https://youtube.googleblog.com/2018/07/building-better-news-experience-on.html

4. Busari S. How fake news does real harm. TED talks conference, 2017.

5. Ken Cherven. Mastering Gephi Network Visualization. Packt Publishing, 2015.

6. Ken Cherven. Network Graph Analysis and Visualization with Gephi. 2013 Packt Publishing, 2013.

7. Niall J. Conroy, Victoria L. Rubin, Yimin Chen. Automatic deception detection: Methods for finding fake news // asis&t, 2015. – Vol. 52, Iss. 1. – pp. 1-4. DOI: 10.1002/pra2.2015.145052010082.

8. Fake news and disinformation online. Report. European Union, 2018. URL: https://ec.europa.eu/digital-single-market/en/news/final-results-eurobarometer-fake-news-and-online-disinformation.

9. John W. Foreman. Data Smart. Using Data Science to Transform Information into Insight. Wiley, 2013.

10. How Content Discovery Platforms Can Fight Fake News via Web Scraping and AI. URL: https://www.promptcloud.com/blog/fight-fake-news-web-scraping-artificial-intelligenc.

11. David M. J. Lazer, Matthew A. Baum, Yochai Benkler etc. The science of fake news. Science, 09 March 2018. – Vol 359, Issue 6380. – pp. 1094-1096. DOI: 10.1126/science.aao2998.

12. Ortuño M., Carpena P., Bernaola P., Muñoz E., Somoza A.M. Keyword detection in natural languages and DNA. Europhys. Lett., 2002. – 57. – P. 759-764.

13. David Sumpter. Outnumbered: From Facebook and Google to Fake News and Filter-bubbles. Bloomsbury Sigma, 2018.

14. Clint Watts. Messing with the Enemy: Surviving in a Social Media World of Hackers, Terrorists, Russians, and Fake News. Harper Paperbacks, 2019.

15. Главацька Ю.Л. Класифікація «фейкових» новин у сучасному медіапросторі: синергетичний аспект. Науковий вісник Херсонського державного університету. 2019. Випуск 1. С. 275-280

16. Горбулін В.П., Додонов О.Г., Ланде Д.В. Інформаційні операции та безпека суспільства: Загрози, протидія, моделювання: монографія. К.: Інтертехнологія, 2009. 164 с.

17. Додонов А.Г., Ландэ Д.В. Моделирование и анализ тематических информационных потоков. Информационное противодействие угрозам терроризма, 2013. № 20. С. 52–59.

18. Кальян Н.А., Матіїшин О.Т. Система контент-моніторингу соціальних мереж з питань кібербезпеки. "Інтелектуальний потенціал - 2019" - збірник наукових праць молодих науковців і студентів з нагоди 30-річчя кафедри кібербезпеки та комп'ютерних систем і мереж ХНУ. Ч.1: Комп'ютерні системи та кібербезпека. Хмельницький: ПВНЗ УЕП, 2019. С. 28-30.

19. Ланде Д.В. Методи оцінки рівня дискримінантної сили слів у текстах з правової тематики. Правова інформатика. 2012. № 3 (35). С. 5-9.

20. Некрасов Г.А., Романова И.И. Разработка поискового робота для обнаружения веб-контента с фейковыми новостями. Инновационные, информационные и коммуникационные технологии. 2017. № 1. С. 128-130.

ДОДАТОК 1

```
from selenium import webdriver

with open('templates/tg.src', 'r', encoding='utf-8') as f:
        nums = f.read().splitlines()
# Парсинг повідомленнь з телеграм каналів та запис у файл "tlgm.txt"
def main():
        driver = webdriver.Chrome()
        #Передача силок на канали з файлу "tg.src" в масив "elm"
        for elm in nums:
        tlgm = []
        driver.get(elm)

        #Звертання до потрібних класів, та парсинг інформації в них
        if driver.find_elements_by_class_name('tgme_widget_message_text'):
        username                                                                =
driver.find_element_by_class_name('tgme_channel_info_header_username').text
        messages = driver.find_elements_by_class_name('tgme_widget_message_text')
        for i in messages:
                message = i.text.replace('\n', ' ')
            tlgm.append(f'{username}: {message}\n')
        else:
        print(f'Error:{elm}')

        f = open('tlgm.txt', 'a', encoding='utf-8')
        for index in tlgm:
        f.write(index)
```

```
        driver.close()

if __name__ == '__main__':
        main()
```

ДОДАТОК 2

```
import collections
with open('templates/indicator_dictionary.txt', 'r', encoding='utf-8') as f:
    mass_a = f.read().splitlines()

# Створення масиву, в який передаються слова з обох словників, для виключення повторень вже перевірених слів
words = []
for i in mass_a:
    words.append(i)

with open('templates/stop_dictionary.txt', 'r', encoding='utf-8') as f:
    mass_b = f.read().splitlines()

for i in mass_b:
    words.append(i)

with open('templates/tlgm.txt', 'r', encoding='utf-8') as f:
    mass_c = f.read().splitlines()

#Запис у масив “filterable_mass” всіх існуючих слів для подальшого розгляду
filterable_mass = []
for i in mass_c:
    a = i.split(' ')
    for j in a:
        filterable_mass.append(j.strip().lower().strip(',' ':' '.' '...' '…' ”” ':' '`' "" """ '/' '//' '(' ')' '?' '!' ';' '№' '<' '>' '*' '«' '»' “‘'))
print(filterable_mass)
```

```
#Перевірка - слова, що записані у будь-який із словників, для подальшого розгляду не виводяться
for i in range(len(filterable_mass)):
    for j in words:
        try:
        if '*' in j:
                j = j.replace('*', '')
                if j in filterable_mass[i]:
                del filterable_mass[i]
          else:
                pass
        else:
                if j == filterable_mass[i]:
                del filterable_mass[i]
        except IndexError:
        pass

# Перевірка та запис у файл ” top_words.txt'”  топ-100 слів, що найчастіше зустрічаються
a = []
for k, cnt in sorted(collections.Counter(filterable_mass).items(), key=lambda x: (-x[1], x[0])):
    a.append(str(cnt) + ':' + k)
print(a)
ax = []
if len(a) > 100:
    for i in range(100):
        ax.append(a[i].split(':')[1])
```

```
else:
    for i in range(100):
        ax.append(a[i].split(':')[1])
print(ax)

f=open('/home/alexander/PycharmProjects/Fake_message/System_interface/templates/top_words.txt', 'w', encoding='utf-8')
for index in ax:
    f.write(index + '\n')
```

ДОДАТОК 3

```
from PyQt5 import QtWidgets
import sys
import window_face
from PyQt5.QtWidgets import QSlider, QLabel

#Запуск вікна, присвоюємо дію кнопці
class ExampleApp(QtWidgets.QMainWindow, window_face.Ui_MainWindow):
    def __init__(self):
        super().__init__()
        self.setupUi(self)
       self.pushButton.clicked.connect(self.check_and_record)

  #Задаємо відповідні параметри дій слайдерам
   def check_and_record(self):
        values = [[], []]
        for i in range(1, 101):
        c = "wh_" + str(i)
        d = "w_" + str(i)
        if i == 1:
                c = 'wh'
                d = "w"
        wh = self.findChild(QSlider, c)
        w = self.findChild(QLabel, d)
        values[0].append(w.text())
        values[1].append(wh.value())

        indct = []
```

```
        stp = []
        for i in range(100):
        if values[1][i] == 1:
                stp.append(values[0][i])
        elif values[1][i] == 3:
                indct.append(values[0][i])
        else:
                pass
        print(indct)
        print(stp)

        f=open('/home/alexander/PycharmProjects/Fake_message/machine_learning/templates/stop_dictionary.txt', 'a', encoding='utf-8')
        for index in stp:
        f.write(index + '\n')

        f=open('/home/alexander/PycharmProjects/Fake_message/machine_learning/templates/indicator_dictionary.txt', 'a', encoding='utf-8')
        for index in indct:
        f.write(index + '\n')
        sys.exit()
def main():
    app = QtWidgets.QApplication(sys.argv)
    window = ExampleApp()
    window.show()
    app.exec_()

if __name__ == '__main__':
    main()
```

ДОДАТОК 4

```
# coding: utf8
with open('templates/tlgm.txt', 'r', encoding='utf-8') as f:
        data = f.read().splitlines()

#Читання файлу " dictionary_2.src'"(словник) та запис у масив "j"
k = open('templates/dictionary_2.src', 'r', encoding='utf-8')
j = []
for l in k:
        if ']' in l:
        j.append(l.replace(']', ' '))
        else:
        j.append(l.strip())

#Перевірка повідомлень на відповідність словнику
thematic_msg = []
for line in data:
        jk = []
        line_2 = line.lower()
        for d in j:
        if d in line_2:
        jk.append(line_2)
   thematic_msg.append(list(set(jk)))
print (thematic_msg)
f=open('/home/alexander/PycharmProjects/Fake_message/filter_2(manipulative_msg)/t
emplates/fn_thematic.txt', 'w', encoding='utf-8')
for index in thematic_msg:
        for i in index:
```

```
    f.write(i + '\n')
```

ДОДАТОК 5

```
import collections
import re

f = open('templates/fn_thematic.txt', 'r', encoding='utf-8').read()
a = []

for k, cnt in sorted(collections.Counter(re.findall(r'^(.+?): ', f, flags=re.M)).items(), key=lambda x: (-x[1], x[0])):
    a.append(str(cnt) + ':' + k)

f = open('an.txt', 'w', encoding='utf-8')
for index in a:
    f.write(index + '\n')
```

ДОДАТОК 6

```
import matplotlib.pyplot as plt
mass = open('templates/an.txt', 'r', encoding='utf-8')

elm = []
for i in mass:
        a = int(i.split(':')[0].strip())
        if a > 1:
        elm.append(i)
        else:
        pass

vals = []
for i in elm:
   vals.append(int(i.split(':')[0].strip()))

labels = []
for i in elm:
   labels.append(i.split(':')[1].strip())

fig, ax = plt.subplots()
ax.pie(vals, labels=labels)
ax.axis("equal")
plt.show()
```

Додаток 7

ДЕРЖАВНИЙ УНІВЕРСИТЕТ ТЕЛЕКОМУНІКАЦІЙ

НАВЧАЛЬНО-НАУКОВИЙ ІНСТИТУТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ
Кафедра Інженерії програмного забезпечення

РОЗРОБКА ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ ВИЯВЛЕННЯ ОЗНАК МАНІПУЛЯЦІЇ В ПОВІДОМЛЕННЯХ З МЕРЕЖІ ІНТЕРНЕТ

Виконавець: Коновал Андрій Сергійович
Керівник роботи: Бондарчук Андрій Петрович
д.т.н., проф.

Київ 2022

Актуальність, мета, об'єкт та предмет роботи

Актуальність	В умовах інформаційної війни дуже важливо отримувати достовірну інформацію
Мета	Створення інформаційної технології для виявлення ознак маніпуляцій в повідомленнях та маніпуляційних повідомлень в мережі Інтернет за допомогою автоматизованого аналізу інформації
Об'єкт	Методи виявлення джерел маніпуляційного інформаційного впливу шляхом автоматизованого аналізу
Предмет	Методи і засоби виявлення МП на основі інтелектуального аналізу тексту, математичної статистики та машинного навчання

Визначення

Фейк - підробка чи імітація новин, які створено з ігноруванням редакційних норм, правил, процесів, прийнятих у ЗМІ для забезпечення відповідності та перевіреності, та яка не витримує жодних, навіть поверхневих, перевірок на відповідність та реальність, але, незважаючи на це, має потужний вплив на свідомість великої кількості людей.

Маніпуляційне повідомлення - повідомлення, що створене з метою маніпулювання свідомістю людини.

3

Аналоги

Botometer	Оцінює аккаунти за шкалою від одного до п'яти, де один означає, що обліковий запис належить реальним користувачам, а п'ятіркою позначаються фейковий Twitter аккаунти. Оцінка проводиться на основі твітів, історії публікацій та згадок іншими користувачами
FactCheck.org	На цьому сайті користувачі можуть задавати питання про достовірність інформації, що звучить в заявах політиків, а команда сайту проводить розслідування і пропонує докладне пояснення. Пояснення включає інформацію про те, ким була зроблена заява, коли воно прозвучало і як команда його перевірила.
Detecting Fake News	Ця програма, яку можна знайти на GitHub, використовує технології машинного навчання і байєсівські моделі для пошуку фейкових новин.

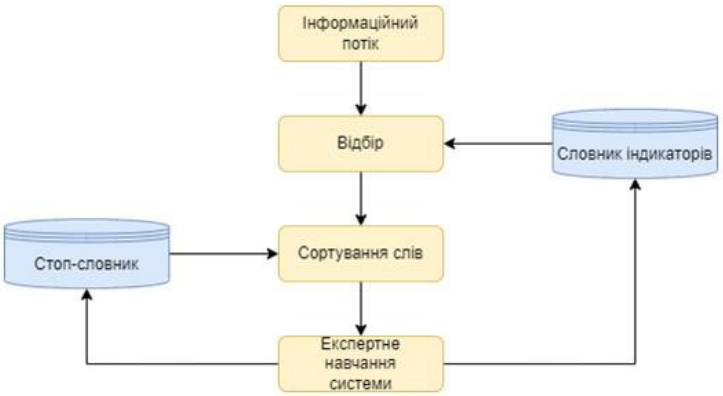
4

Недоліки

Botometer	Має обмеження в кількості мов. Орієнтована на пошук фейкових новин.
Hoaxy	Має обмеження в кількості мов. Орієнтована на пошук фейкових новин.
FactCheck.org	Потребують залучення значної кількості персоналу.
Detecting Fake News	Орієнтована на пошук фейкових новин.
Politifact	Працює з одним джерелом за раз.

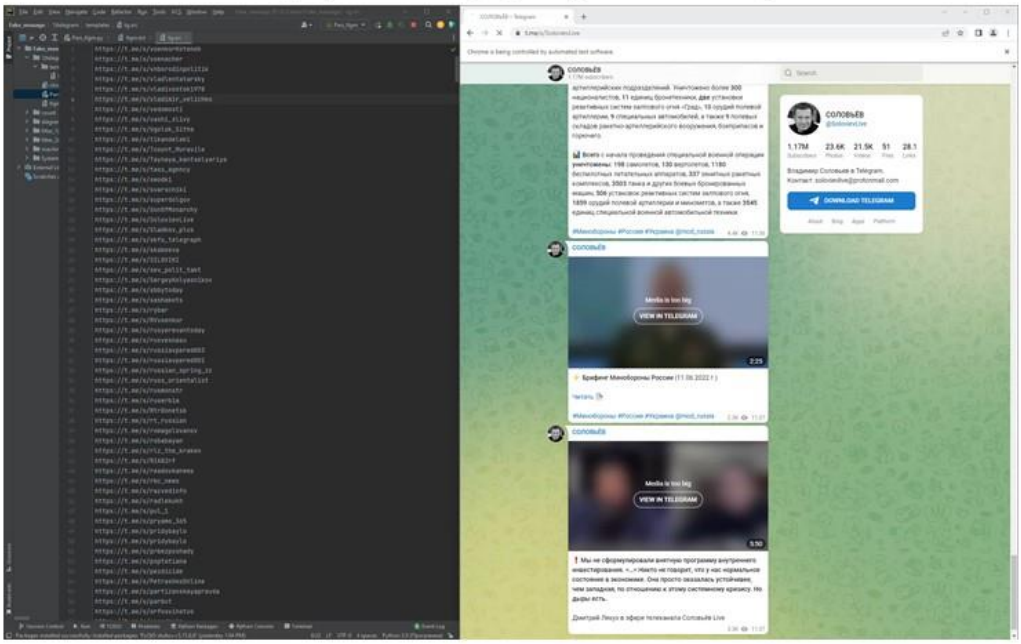
5

Методика навчання системи виявлення ознак маніпуляції в повідомленнях з мережі Інтернет



6

Формування інформаційного потоку

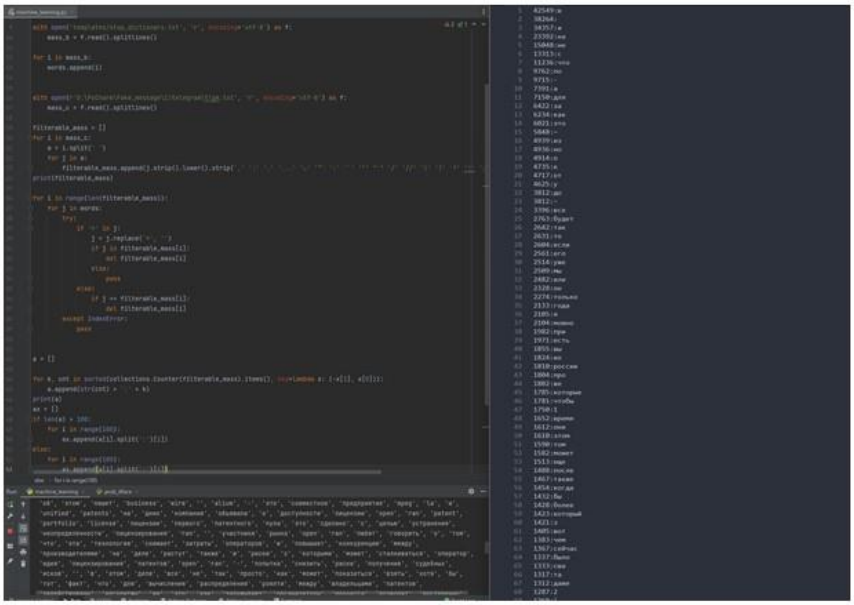


7

Приклад файлу на виході програми

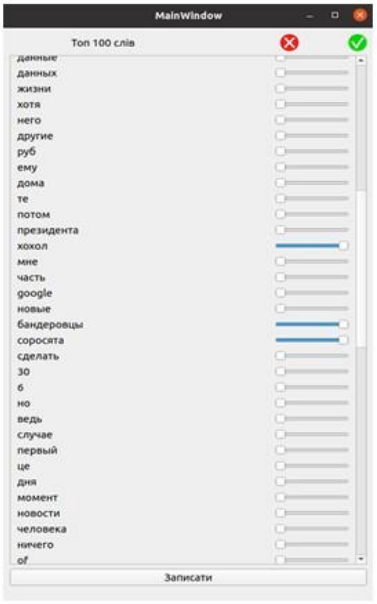
8

Приклад роботи програмного модуля байєсівського машинного навчання та файлу на виході з програми



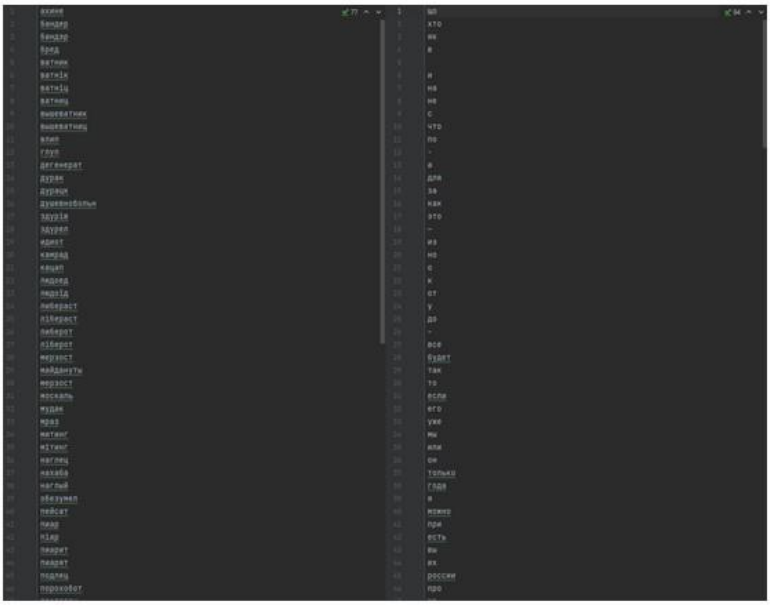
9

Інтерфейс



10

Приклад словників отриманих в результаті навчання системи



11

Функціонування системи

12

Виявлення маніпуляцій в повідомленнях на базі виявлених ознак

$$Pr(F|W) = \frac{Pr(W|F)\cdot Pr(F)}{Pr(W)} = \frac{Pr(W|F)\cdot Pr(F)}{Pr(W|F)\cdot Pr(F)+Pr(W|H)\cdot Pr(H)}$$

➢ **Pr(F|W)** – умовна ймовірність того, що повідомлення маніпуляційний, за умови, що слово-індикатор знаходиться в ньому;

➢ **Pr(F) -** повна ймовірність того, що довільне повідомлення є маніпуляційним;

➢ **Pr(W|F)** – умовна ймовірність того, що слово-індикатор з'являється в повідомленнях, якщо вони є маніпуляційними;

➢ **Pr(H) -** повна ймовірність того, що довільне сполучення не маніпуляційний;

➢ **Pr(W|H) -** умовна ймовірність того, що слово-індикатор з'являється в повідомленнях, якщо вони є не спамом.

13

Маніпуляційність слова

Вірогідність маніпуляційності повідомлення $Pr(F) = 0.5;\ Pr(H) = 0.5$

Спрощена загальна формула Байєса $Pr(F|W) = \frac{Pr(W|F)}{Pr(W|F) + Pr(W|H)}$

14

Визначення маніпуляційності повідомлення

$$p = \frac{p_1, p_2, \dots, p_N}{p_1 p_2 \dots p_N + (1-p_1)(1-p_2)\dots(1-p_N)}$$

- $p = \mathbf{Pr}(F|W_1,W_2,\dots,W_N)$ - ймовірність, що повідомлення, що містить слова - W_1, W_2,…, W_N маніпуляційне;
- p_1 – умовна ймовірність того, що повідомлення маніпуляційне, за умови, що воно містить перше слово;
- p_2 – умовна ймовірність того, що повідомлення - маніпуляційне, за умови, що воно містить другий маніпуляційний слово;
- p_N – умовна ймовірність того, що повідомлення маніпуляційне, за умови, що воно містить N-е слово.

15

Фільтрація повідомлень

Нехай G – множина слів, що маркують маніпуляції: $G = \{g_i\}_{i=1}^{|G|}$

Позначимо множину джерел як S: $S = \{S_k\}_{k=1}^{|S|}$

Ймовірність того, що повідомлення d є маніпуляційним, якщо воно містить слово g_i, позначимо як p_i. Відповідно, ймовірність того, що повідомлення d не є маніпуляційним, якщо воно містить слово g_i, позначимо як q_i ($p_i + q_i = 1$).

Нехай повідомлення d містить декілька слів з G: $\exists i$: $g_i \in d, g_i \in G$.

Ймовірність того, що повідомлення не є маніпуляційним, у цьому випадку дорівнює:

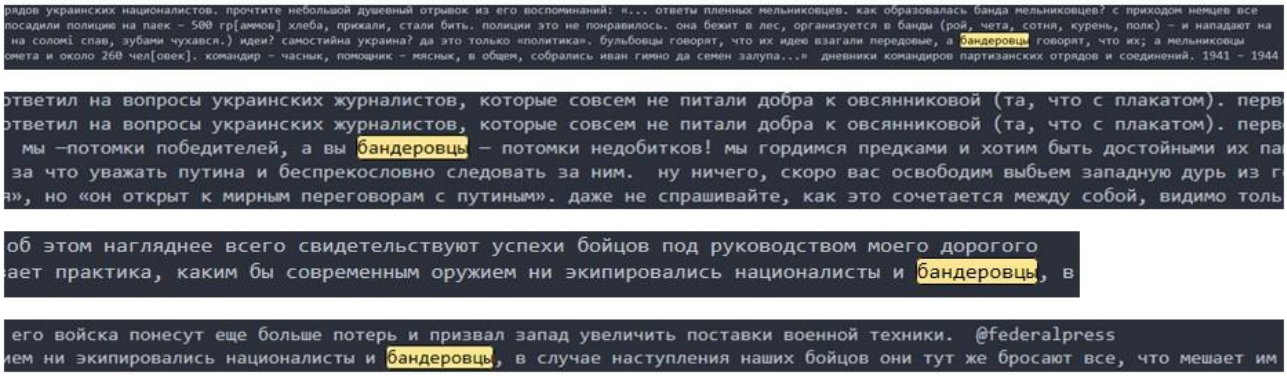
$$q(d) = \prod_{i:\, g_i \in d,\, g_i \in G} (1 - p_i)$$

Тоді ймовірність того, що повідомлення є маніпуляційним, дорівнює:

$$p(d) = 1 - \prod_{i:\, g_i \in d,\, g_i \in G} (1 - p_i)$$

16

Приклад відфільтрованих повідомлень

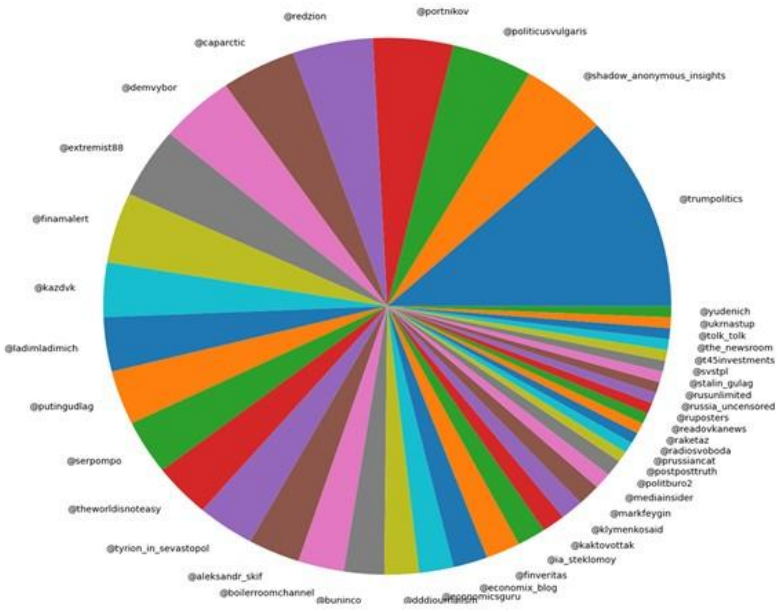


17

Визначення сили маніпуляційності джерела

18

Відображення сили маніпуляційності джерела



19

Висновки

1. Запропоновано новий метод навчання системи виявлення ознак маніпуляції методом Байєса, який має можливість працювати з будь-яким джерелом інформації, не має обмежень у кількості мов, кількості джерел та повідомлень. Також перевагою є відсутність потреби у залучені значної кількості персоналу.
2. Створено програмне забезпечення, яке використовується на практиці в якості засобу виявлення маніпуляційних повідомлень за потрібною тематикою.

20

Апробації

1. Журнал “Телекомунікаційні та Інформаційні Технології” №4

Дякую за увагу!