УДК 004.9

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ РОЗПІЗНАВАННЯ ПРОПАГАНДИ, ФЕЙКІВ ТА ДЕЗІНФОРМАЦІЇ У ТЕКСТОВОМУ КОНТЕНТІ НА ОСНОВІ МЕТОДІВ NLP ТА МАШИННОГО НАВЧАННЯ

Висоцька В. А. – д-р техн. наук, доцент, доцент кафедри «Інформаційні системи та мережі», Національний університет «Львівська політехніка», Львів, Україна.

АНОТАЦІЯ

Актуальність. Дослідження спрямоване на застосування штучного інтелекту для розроблення та вдосконалення засобів кіберборотьби, зокрема для боротьби з дезінформацією, фейками та пропагандою в Інтернет-просторі, виявлення джерел дезінформації та неавтентичної поведінки (боти) скоординованих груп. Реалізація проекту сприятиме вирішенню важливого та актуального у наш час питання інформаційної маніпуляції у медіа, адже для ефективної боротьби із викривленням та дезінформацією необхідно отримати ефективний інструмент розпізнання цих явищ у текстових даних для вироблення подальшої стратегії запобігання розповсюдження таких даних.

Метою дослідження є розробка інформаційної технології для автоматичного розпізнавання політичної пропаганди у текстових даних, яка побудована на основі машинного навчання з учителем та реалізована за допомогою методів опрацювання природної мови.

Метод. Розпізнання наявності пропаганди відбуватиметься на двох рівнях: на загальному рівні, тобто рівні документу, та на рівні окремих речень. Для реалізації проекту використано такі методи конструювання ознак, як статистичний показник TF-IDF, модель векторизації «Торба слів», розмічування частин мови, моделі word2vec для отримання векторних представлень слів, а також розпізнавання тригерних слів (підсилюючі слова, абсолютні займенники та «блискучі» слова). У якості основного алгоритму моделювання використана логістична регресія.

Результати. Розроблено моделі машинного навчання для розпізнавання пропаганди, фейків та дезінформації на рівні документу (статті) та на рівні речень. Обидві оцінки моделі є задовільними, проте модель для розпізнання пропаганди на рівні документу впоралася в майже 1,2 разів краще (на 20%).

Висновки. Створені моделі показує відмінні результати розпізнавання пропаганди, фейків та дезінформації у текстовому контенті на основі методів NLP та машинного навчання. Аналіз вихідних даних показав, що моделі розпізнання пропаганди на рівні документу (статті) вдалося коректно класифікувати 6097 не пропагандистських статей та 694 статті пропагандистського характеру. 123 пропагандистські статті та 285 не пропагандистських статей були класифіковані невірно. Отримана оцінка моделі: 0,9433254618697041. Модель розпізнання пропаганди на рівні речень успішно класифікувала 1917 не пропагандистських статей та 205 пропагандистських статей, проте 585 пропагандистських статей та 146 не пропагандистських статей були класифіковані невірно. Оцінка моделі становить: 0,7437784787942516.

КЛЮЧОВІ СЛОВА: дезінформація, фейк, пропаганда, лінгвістичний аналіз, опрацювання природньої мови, машинне навчання, кіберборотьба, штучний інтелект, семантичний аналіз, інформаційна безпека.

АБРЕВІАТУРА

ЗМІ – засоби масової інформації;
ІПСО – інформаційно-психологічна операція;
ІС – інтелектуальна система;
ІТ – інформаційна технологія;
ПЗ – програмне забезпечення;
ПО – предметна область;
СД – сховище даних;
IDF – Inverse Document Frequency;
ML – machine learning;
NLP – Natural Language Processing;
nPMI – Normalized pointwise mutual information;
SVM – Support Vector Machine;
TF – Term Frequency.

НОМЕНКЛАТУРА

S – система розпізнавання пропаганди;
I – множина вхідних даних;
O – множина вихідних даних;
R – основні правила опрацювання вхідних даних;
U – параметри опрацювання вхідних даних;
L_R – метод машинного навчання;
α – оператор скачування вхідних даних;
β – оператор опрацювання вхідних даних;
γ – оператор аналізу статей на основі ML;
μ – оператор ідентифікації тематичних статей;
χ – оператор формування датасету статей;
ω – оператор маркування статті;
λ – оператор прийняття рішення;
i_1 – множина даних із Інтернет-джерел;
i_2 – сховище даних публікацій;
i_3 – словники слів-маркерів пропаганди;
i_4 – множина тематичних ключових слів фейків;
o_1 – періодичні запити на збір публікацій;
o_2 – результат застосування NLP;
o_3 – результат застосування ML;
r_1 – правила збору даних з Інтернет-джерел;
r_2 – правила NLP текстового контенту;
r_3 – правила ML для розпізнавання пропаганди;
r_4 – правила маркування статті як пропаганди;
u_1 – множина умов збору статей в Інтернет-джерелах;
u_2 – множина вимог фільтрування датасету від шуму;
u_3 – множина умов опрацювання датасету статей;
u_4 – множина умов ML для розпізнавання фейку;
u_5 – множина вимог формування висновків.

DOI 10.15588/1607-3274-2024-2-13

ВСТУП

Дезінформація визначається як «фактично невірна інформація, яка не підтверджена доказами». Дезінформація в Інтернет є актуальною та життєво важливою проблемою, особливо в сферах, пов'язаних з війною в Україні. Така інформація, отримана з соціальних медіа, включаючи тематичні онлайн-спільноти, впливає на результати формування громадської думки, керування настроями суспільства та, відповідно, на хід війни в цілому. Занепокоєння з приводу дезінформації зросло із збільшенням кількості запитів на відповідну інформацію в Інтернет, зокрема, в ЗМІ та соціальних мережах. Відсутність захисних механізмів під час обговорень в онлайн-спільнотах сприяє поширенню та зміцненню дезінформації, фейків та пропаганди. Існуюча література здебільшого зосереджена на виявленні фальшивих оглядів і фейкових новин. Однак у літературі бракує комплексної теоретичної основи, розробленої для виявлення дезінформації, особливо в контексті онлайн-спільноти. Враховуючи величезний обсяг дезінформації про війну в Україні, що поширюється в відповідних онлайн-спільнотах, існує необхідність розробити ефективну модель автоматичного виявлення потоку дезінформації для подальшої ідентифікації неавтентичної поведінки скоординованих груп людей/ботів-розповсюджувачів.

Метою дослідження є розроблення інформаційної технології виявлення дезінформації для підвищення рівня інформаційної безпеки держави шляхом розроблення математичних моделей, методів та засобів кіберборотьби з дезінформацією. Зокрема, це сприятиме для автоматичного виявлення джерел дезінформації та неавтентичної поведінки (боти) скоординованих груп в Інтернет на основі стилістичного аналізу та лінгвістичного опрацювання тексту фейків та пропаганди, особливостей їх розповсюдження та репостів на основі ML-методів.

Розробка методів та засобів моніторингу та виявлення дезінформації в Інтернет вимагає розв'язку відповідних задач, зокрема:

– лінгвістичне опрацювання дезінформації для виявлення спільних характерних ознак пропаганди;

– розпізнання пропаганди на рівні статті;

– розпізнання пропаганди на рівні речення;

– тренування моделей для формування прогнозів на основі тестової вибірки;

– розроблення модулів ІС для аналізу текстових потоків контенту для виявлення пропаганди;

– експериментальна апробація розробленої ІТ розпізнавання пропаганди, фейків та дезінформації у текстовому контенті на основі методів NLP та ML.

Наукова новизна полягає у розробленні методів:

– стилістичного аналізу та лінгвістичного опрацювання дезінформації для виявлення спільних характерних ознак фейків одного авторського колективу основі методів опрацювання природньої мови та штучного інтелекту, лінгвістичного аналізу повідомлень, класифікації/кластеризації тексту тощо для виявлення лінгвістичних ознак деструктивного та маніпулятивного спроб впливу на читача;

– виявлення потенційно подібних за стилістикою дезінформації для формування множини потенційних авторів та учасників розповсюдження пропаганди на основі збору/моніторингу/виявлення/класифікації інформаційних загроз в Інтернет-просторі.

Практична новизна полягає у розробленні ІС виявлення пропаганди, а також експериментальна апробація, збір/опрацювання/аналіз отриманих результатів для розрахунку точності/ефективності функціонування на основі реалізації модулів ПЗ як:

– модуль інтелектуального пошуку, збору, маркування, лінгвістичного аналізу та класифікації інформаційних повідомлень для подальшого формування множини потенційних фейків, а також моніторингу, керування, виявлення та відстеження даних інформаційних загроз на основі ML;

– модуль стилістичного аналізу множини фейків для ідентифікації подібних за стилем для одного авторського колективу з подальшим їх класифікацією (людина/бот) на основі методів ML та NLP.

Проект спрямований на застосування штучного інтелекту для розроблення та вдосконалення засобів кіберборотьби, зокрема для боротьби з дезінформацією в Інтернет, а саме для автоматичного виявлення джерел дезінформації та неавтентичної поведінки (боти) скоординованих груп. Необхідно дослідити явище політичної пропаганди у новинних медіа, розпізнати наявність пропаганди у текстових даних. Необхідно також розробити алгоритм підготовки та виокремлення ознак текстових даних, а також побудувати модель машинного навчання, котра розпізнаватиме наявність політичної пропаганди у текстах за допомогою цих ознак. Об'єкт дослідження процеси пошуку, виявлення та класифікації політичної пропаганди, фейків та дезінформації у медіа, зокрема у ЗМІ в Інтернет-середовищі. Предмет дослідження – це методи та засоби розпізнання пропаганди, фейків та дезінформації у текстових даних. Дослідження сконцентроване на розробці системи розпізнання пропаганди, фейків та дезінформації на основі машинного навчання через опрацювання природної мови як на рівні речення, так і на рівні документу.

1 ПОСТАНОВКА ПРОБЛЕМИ

Зростання темпів розповсюдження дезінформації в ЗМІ, зокрема в Інтернет, під час інформаційнішої війни вже давно викликає занепокоєння суспільства, оскільки поширення такої дезінформації має негативний вплив на населення як споживача цього контенту та відповідно хід самої війни. Зазвичай виявлення тематичної онлайн-дезінформації ПО ґрунтується на лінгвістичних особливостях змісту текстового контенту публікацій (статей). Але вони множаться та розповсюджуються швидше, ніж їх ідентифікують та блокують. Тому виявлення джерел подібного контенту, потенційних авторів, механізмів

DOI 10.15588/1607-3274-2024-2-13

розповсюдження, зокрема аналіз та ідентифікація поведінки потенційних генераторів фейків є задачею першочерговою для вдосконалення засобів кіберборотьби з дезінформацією на просторах Інтернету. А це базується на результатах точного та оперативного виявлення стилістично подібного тексту в публікаціях пропаганди та фейків ПО.

Систему розпізнавання пропаганди, фейків та дезінформації у текстовому контенті на основі методів NLP та машинного навчання подамо як:

$$S = < I, O, R, U, L_R, \alpha, \beta, \gamma >, \quad (1)$$

де $I = \{i_1, i_2, i_3, i_4\}$, $O = \{o_1, o_2, o_3\}$, $R = \{r_1, r_2, r_3, r_4\}$, $U = \{u_1, u_2, u_3, u_2, u_3\}$.

Основними процесами моделі аналізу текстового контенту статей із Інтернет-джерел для розпізнавання пропаганди, фейків та дезінформації є «Збір статей для формування датасету», «NLP текстового контенту статей для виділення лінгвістичних ознак», «Машине навчання для розпізнавання пропаганди» та «Формування висновків наявності пропаганди».

Процес «Збір статей для формування датасету» опишемо суперпозицією:

$$C_{AU} = \mu \circ \beta \circ \alpha, \quad (2)$$

$$C_{AU} = \mu(\beta(\alpha(i_1, i_2, i_4), r_1, u_1), u_2). \quad (3)$$

Особливості онлайн-дезінформації можна класифікувати на два рівні: центральний(включаючи особливості теми) і периферійний (включаючи лінгвістичні особливості, особливості настроїв і особливості поведінки користувачів). Необхідно знайти особливості поведінки, щоб відобразити характеристики взаємодії користувачів: початок обговорення, залучення до взаємодії, сфера впливу, посередництво у відносинах та інформаційна незалежність. Тому процес «NLP текстового контенту статей для виділення лінгвістичних ознак» опишемо суперпозицією:

$$C_{CU} = \chi \circ \beta \circ \alpha, \quad (4)$$

$$C_{CU} = \chi(\beta(\alpha(C_{AU}, i_2, i_3, i_4), r_1, u_3), r_2). \quad (5)$$

Щоб побудувати моделі та методи ідентифікації дезінформації в Інтернеті, багато дослідників присвятили себе виявленню особливостей дезінформації. Дезінформацію в соціальних мережах можна розглядати як повідомлення, які публікуються, щоб переконати інших користувачів. Щоб виявити ефективні функції виявлення дезінформації в онлайн-спільнотах, необхідно використали модель, яка зможе допомогти зрозуміти, як дезінформація в Інтернет, зокрема в соціальних мережах та онлайн-спільнотах переконує користувачів. Користувачі зазвичай будують ставлення до повідомлення як центральним, так і периферійним маршрутом. У центральному маршруті користувачі ретельно перевіряють якість і силу інформації; тоді як у периферійному маршруті користувачі більше дбають про поверхневі фактори, такі як репутація джерела, візуальна привабливість і презентація. Окрім змісту повідомлення, деяка вторинна інформація (наприклад, кількість лайків і зірочок) суттєво підвищує валідність і надійність повідомлень. Тому функції центрального рівня повідомлень переконують користувачів на основі вмісту повідомлень, тоді як функції периферійного рівня переконують користувачів через вплив авторів повідомлень. Найкращими функціями для виявлення дезінформації в соціальних мережах можуть бути ті, які розглядають особливості користувача, повідомлення, теми та поведінки користувача.

Процес «Машине навчання для розпізнавання пропаганди» опишемо як:

$$C_{UL} = \omega \circ \gamma \circ \beta \circ \alpha, \quad (6)$$

$$C_{UL} = \omega(\gamma(\beta(\alpha(C_{CU}, L_R, i_2), i_3), u_4), r_3). \quad (7)$$

Створення моделі виявлення дезінформації, яка об'єднує функції центрального рівня (зокрема особливості теми) та функції периферійного рівня (зокрема лінгвістичні особливості, особливості настрою та особливості поведінки користувачів), потребує подальших досліджень. На основі цих функцій необхідно оцінювати їхню здатність автоматично відрізняти дезінформацію від правдивої в межах тематичної онлайн-спільноти за допомогою різних методів машинного навчання.

Процес «Формування висновків наявності пропаганди» опишемо як:

$$C_{US} = \lambda \circ \gamma \circ \beta \circ \alpha, \quad (8)$$

$$C_{US} = \lambda(\gamma(\beta(\alpha(C_{US}, i_2), i_4), u_5), r_4). \quad (9)$$

Розроблена система швидкої ідентифікації джерел дезінформації має базуватися на аналізі неавтентичної поведінки учасників розповсюдження фейків. Результати не лише мають продемонстрували ефективність поведінкових особливостей у виявленні дезінформації, але й запропонували як методологічний, так і теоретичний внесок у виявлення дезінформації з точки зору інтеграції особливостей повідомлень, а також особливостей авторів повідомлень.

На фоні інформаційної війни витрачається багато ресурсів та часу на оперативний збір контенту, його опрацювання та аналіз, а також генерування рішень/висновків щодо його наповнення. На це також впливає мова публікацій, при перекладі якої суттєво/частково спотворюється зміст. ІС не зможе повністю замінити діяльність фахівців кібербезпеки та кіберборотьби. Але вона буде допоміжним інструментом для оперативного формування відповідних датасетів/корпусів фейкового контенту та їх джерел, стилістичного та лінгвістичного аналізі тексту дезінформації для формування інформаційного портрету авторів, пошук авторів та розповсюджувачів через аналіз неавтентичної поведінки та результатів аналізу стилю написання контенту, а також реагування на динаміку змін або локальні зміни в

DOI 10.15588/1607-3274-2024-2-13

контентому потоці, маркуючи відповідний контент як ймовірносно фейковий.

2 АНАЛІЗ ЛІТЕРАТУРНИХ ДЖЕРЕЛ

Онлайн ЗМІ та соціальні мережі дозволяють швидко обмінюватися інформацією, в тому числі і дезінформацією як цілеспрямовано, так і випадково/хаотично. Поряд з основною перевагою як організація швидкого доступу для всіх бажаючих до оперативної та актуальної інформації, онлайн медіа часто використовують для поширення навмисно оманливого контенту як фейків та пропаганди про конкретні події, людей або організацій, в тому числі уряди [1]. Останнім часом яскравими прикладами розповсюдження дезінформації є спроби російського уряду контролювати інформацію під час війни в Україні з 2014 року, наприклад, авіакатастрофа MH17 [2]. Паралельно на багато онлайн-інформації накладається регіональна цензура на певних територіальних регіонах із-за політичних, економічних, соціальних, релігійних та інших чинників, наприклад, для контролю/управління думкою людей цього регіону, наприклад на окупованих територіях росії для контролю майбутніх виборців бункерного президента [3]. Загубитися та зорієнтуватися в цій масі потоку контенту з протилежними фактами та причинами подій/явищ пересічній людині легко [4]. Контролювати, що показувати/скривати (накладати цензуру) серед Інтернет-контенту пересічному користувачу в демократичних державах є неетично, незаконно та недоцільно без прямих доказів щодо наявності дезінформації/ фейку/ пропаганди для цілеспрямованого порушення інформаційної безпеки організації/країни [5–6]. Це один із перших кроків переходу до тоталітаризму. А надавати інформацію, наприклад, журналістам про можливий тематичний фейк для проведення журналістичного розслідування або попередження пересічного читача про можливість наявності в цьому контенті/ресурсі дезінформації є з одного боку підтримкою свободи слова, з іншого надання можливості людині обирати чому вірити. Це дає змогу отримувати розуміння подій та орієнтування в потоці інформації для вирішення буденних задач і корегування бізнес-стратегій тощо.

Політична пропаганда є спрямованим та навмисним поширенням інформації, метою якого є вплив на громадську думку суспільства на користь певної громадської позиції чи спільної справи. Пропаганда може відбуватися як у формі дезінформації, тобто фабрикування недостовірних та фальшивих новин, так і використовувати більш складні та комплексні методики. Пропаганда, фейки та дезінформація зазвичай генерується, формується та розповсюджується за допомогою ЗМІ, є дотичною до тих чи інших політичних подій – передвиборна кампанія, фінансова криза тощо. Отже, у деяких випадках, для розпізнання пропаганди необхідно знати контекст політичного клімату у світі.

Значне та масове розповсюдження фейків, пропаганди та дезінформації на фоні війни в Україні без систематичного та ґрунтовного аналізу ймовірно впливає на формування думки суспільства та керує нею, а також призводить до панічних настроїв серед відповідного регіону/верства населення, значно впливає на корегування стратегій/планів державних органів, соціальних служб, бізнесу, тощо. Блокуванням дезінформації та джерел її розповсюдження, а також ідентифікацій потенційних авторів на основі аналізу неавтентичної поведінки зазвичай є функціональним обов’язками уповноважених органів, особливо під час інформаційної війни. Але вона настільки зараз швидко та оперативно генерується/розповсюджується на основі застосування сучасних інформаційних технологій та штучного інтелекту, що справитися з цією задачею на 100% ніхто не спроможний без використання нових методів та засобів на базі машинного навчання.

Для повного аналізу всього нового/старого контенту не вистачить ресурсів. Та і поки буде проведений системний аналіз даних, сама дезінформація стане застарілою. А ось швидке формування/модифікування/поповнення баз/сховищ даних маркованого контенту як блокований/ неблокований в певному регіоні, відсортованого за відповідними метриками (час, тема, регіон блокування, мова тощо) від актуального/релевантного до менш актуального для подальшого аналізу методами/технологіями NLP/ML значно пришвидшить процес орієнтування серед хаосу нової інформації в Інтернет. Визначення теми/причини блокування контенту (накладання цензури) на певному регіоні дозволить покращити якість ідентифікації фейків/пропаганди/дезінформації відповідної тематики. Тому актуальним та необхідним є розроблення ІС автоматичного виявлення джерел дезінформації та неавтентичної поведінки користувачів чатів в кібернетичному просторі. ІС має бути реалізована на нових принципах інформаційної безпеки (моніторинг даних, виявлення загроз, прогнозування), що дасть змогу ідентифікувати, моніторити, повідомляти про рівень загрози та прогнозувати кіберзагрози, а також ступінь ймовірнісного інформаційно-психологічного впливу на громадську думку. З огляду на це, такий проект є релевантним, актуальним, перспективним і своєчасним для збільшення ступеня інформаційної безпеки держави на основі ідентифікації, моніторингу, прогнозування та аналізу загроз у кіберпросторі України.

Проблема політичної пропаганди стає усе більш актуальною за рахунок того, що інформація стає усе більш доступною, поширюється практика діджиталізації суспільних медіа, продукувати та редагувати новини стає усе простіше, зростає вплив соціальних мереж. Наприклад, дослідження поширення пропаганди за допомогою соціальних

DOI 10.15588/1607-3274-2024-2-13

мереж, підтверджує, що пропаганда, розповсюджена таким чином, поширюється швидше та на більш широкий демографічний спектр, а також є більш стійкою до розпізнання за рахунок використання підтверджувального упередження [7]. Окрім цього, результати нещодавнього дослідження Science Magazine показали, що плітки та фейкові новини розповсюджуються приблизно у шість разів швидше, ніж достовірна та правдива інформація [8]. Це вказує на те, що явище політичної пропаганди не лише згубно впливає на сучасний політичний клімат, але й частково формує його. На даний момент більшість проектів, пов'язаних із розпізнанням пропаганди, виконуються за допомогою статистичних досліджень із залученням спеціалістів, проте, за рахунок швидкого поширення пропаганди за допомогою онлайн-медіа, є доволі неефективним та дорогим з точки зору використання ресурсів. Саме тому побудова ефективної модель ML для оптимізації цього процесу є як ніколи актуальною, особливо враховуючи той факт, що на сьогоднішній день ML-системи, базовані на NLP, набувають усе більшої популярності як у академічному, так і у прикладному середовищі науки про дані.

Для успішного аналізу та опрацювання природного тексту, необхідно зазначити певні ознаки пропаганди, котрі використовують при класифікації тексту. Для цього необхідно проаналізувати, яку саме структуру текст повинен мати для того, аби бути маркованим як ймовірно пропагандистський. Основні методи поширення пропаганди є наступними [9]:

– Відволікаючий маневр. Презентація недоречного матеріалу, котрий не має відношення до обговорюваного питання у тексті.

– Викривлення позиції. Заміна подібною, але не аналогічною, позицією.

– Whataboutism. Позиція опонента дискредитується шляхом звинувачення його у лицемірстві без безпосередньої аргументації.

– Причинне спрощення. Виділення та презентація лише одної гіпотетичної причини певного явища, коли таких причин є декілька.

– Навмисна розпливчатість, збиття з пантелику. Використання навмисно абстрактних термінів та слів таким чином, щоб інтерпретація сказаного не була єдиною та очевидною.

– Апелювання до авторитету. Ствердження, що заявка є вірною, бо чинний авторитет/експерт її підтримує, зазвичай без подання жодних доказів.

– Чорно-біла оманливість. Презентація двох альтернативних варіантів або точок зору як єдиних можливих, навіть якщо існують інші варіанти.

– Навішування ярликів. Спосіб, коли джерело пропаганди надає явищу, проти котрого виступає, негативних зміст, зазвичай апелюючи до того, чого цільова аудиторія боїться, що ненавидить.

– Навантажена мова. Використання певних фраз та слів з сильним емоційним підтекстом (позитивним або негативним) для впливу на цільову аудиторію.

– Перебільшення або мінімізація. Певне явище репрезентовано або у надмірному вигляді, або ж як щось менш важливе, ніж є насправді.

– Розмахування прапором. Маніпуляція патріотичними поглядами/почуттями цільової аудиторії для виправдання/поширення явища/ідеї.

– Сумнів. Ставлення під сумнів авторитетність та надійність певної людини або явища.

– Апелювання до страху або упередження. Спроба сприяти підтримці певної ідеї шляхом вселення страху та тривоги, або ж апелювання до певних соціальних упереджень цільової аудиторії.

– Слогани. Використання коротких ударних фраз, для навішування ярликів або стереотипізації через апелювання до емоцій цільової аудиторії.

– Кліше без суті. Слова/фрази для перешкоджання аргументованому обговоренню ситуації та критичному мисленню.

– Загальна платформа. Спроба переконати цільову аудиторію приєднатися до справи та прийняті певне рішення як всі інші або більшість.

– Повторення. Повторення одного посилу декілька разів для зомбування цільової аудиторії.

Пропаганда є комплексним комунікативним явищем, котре використовує різноманітні методи та підходи для досягнення своєї мети. Задача розпізнання пропаганди у текстах не є новою та у цій сфері проведено багато досліджень. Для аналізу обрано декілька аналогів, різних за методами дослідження та моделювання.

– Framing and Agenda-setting in Russian News: a Computational Analysis of Intricate Political Strategies. Дослідження проведено колективом вчених із університету Карнегі Мелон, університету Хайфа та Стенфордського університету у 2018 році [10]. Основна мета роботи – дослідити «фреймінг» у новинних джерелах росії протягом економічно несприятливого становища у країні та те, як це пов'язано із висвітленням новин, що стосуються США, у російських ЗМІ (бо більшість із них керуються державними органами та знаходяться під їх безпосереднім впливом). Був встановлений сильний негативний кореляційний зв'язок між економічної ситуацією у росії та кількістю новин, що фокусуються на висвітленні подій у США. Наступним кроком досліджено те, у якому контексті та з яким фреймом ці новини подаються. Фрейм у даному контексті визначає те, яким чином (із якими конотаціями) висвітлюється та чи інша новина у медіа. Потім висунута гіпотеза про те, що кореляційний зв'язок є причинно спрямованим, і що несприятлива економічна ситуація у росії є безпосереднім чинником, котрий викликає посилену увагу до США у державних ЗМІ. За допомогою імовірнісних методів (причинність за Грейнджером та лінійної регресійної моделі) встановлено, що причинний зв'язок між цими двома змінними є наявним у наборі обраних даних. Далі за допомогою імовірнісного методу Баєса сформовано так звані лексикони фрейму, з

DOI 10.15588/1607-3274-2024-2-13

урахуванням конкретного слова та фрейму, до якого воно належить. Для кожного фрейму обрано 250 слів, котрі мають найбільшу імовірність зустрітися у відповідному контексті. У якості початкового, базового рішення, обрано логістичну регресію та модель «торба слів». Для того, аби визначити, які фрейми стосуються США, обчислено nPMI. Виділено п'ять фреймів, котрі надалі використовуються для дослідження. У дослідженні також використовується Structure Topic Modelling для того, аби концептуалізувати статті на тому рівні, на якому б їх сприймав читач. Таким чином, досліджено метод пропаганди, націлений на відволікання публіки від скрутного стану внутрішніх справ у країні за допомогою перекладання уваги на зовнішню політику.

– The Use of Supervised Learning Algorithms in Political Communication and Media Studies: Locating Frames in the Press. Дослідження фокусується на дослідженні фреймінгу у сучасних медіа [11] та побудовано на наборі даних із іспанських новинних джерел за 2015 рік, зокрема увагу сфокусовано на висвітленні кризи біженців. Обрано два фрейми, згідно з якими проводилася класифікація – human rights та security. Основний алгоритм ML, котрий використовується у дослідженні для побудови моделі – метод SVM, який застосовують для розв'язку задач класифікації та побудований на припущенні, що дані є лінійно розподілені таким чином, що можна знайти таку гіперплощину, котра могла б ефективно розділити їх один від одного. Однак, лінійне ядро моделі не показало задовільних результатів, тому в остаточній версії алгоритму використано метод опорних векторів із радіальним ядром. Для видалення семантично незначних слів із корпусу даних використано алгоритм IDF.

– Фейкогриз – це інструмент для розпізнання пропаганди, створений українською платформою дослідження даних Texty.org [12–13]. Система побудована у вигляді додатку для веб-браузера, а також у якості чат-боту для платформи Telegram. У якості вхідних даних Фейкогриз приймає посилання на новину і, використовуючи веб-скрейпінг, визначає, чи у тексті, поданому за джерелом, є наявною маніпуляція або пропаганда. Модель побудована за принципом трансферного навчання, що означає, що використовується модель загального призначення, котра, можливо, не була натренована для виконання специфічного завдання, проте згодом відбувається відповідне до задачі налаштування гіперпараметрів моделі [14–15]. Замість того, щоб фокусуватися на проблемі класифікації та витрачати ресурси на позначення даних відповідними класами, Фейкогриз побудований на алгоритмі кластеризації, котрий автоматично визначає клас даних, котрі отримує.

3 МАТЕРІАЛИ ТА МЕТОДИ

Оскільки розпізнавання пропаганди у текстових даних [16–18] відбуватиметься на двох рівнях – на рівні документу та на рівні речення, обрано два окремі набори даних для кожної задачі.

– Розпізнання пропаганди на рівні документу. Набір даних для розпізнавання пропаганди для цієї задачі складається із 35993 статей (включно із заголовками) англійською мовою, кожна із яких промаркована як «пропаганда» або «не-пропаганда». Також в наборі присутній унікальний ідентифікатор для кожної із статей. Дані подані як текстовий файл, у якому текст статті є відділеним від категорії та ідентифікатору знаками табуляції. Після завантаження файлу, видалення атрибуту ідентифікатору та перетворення даних до формату pandas.DataFrame (рис. 1).

	article	label
0	Et tu, Rhody? A recent editorial in the Provi...	non-propaganda
1	A recent post in The Farmington Mirror — our t...	non-propaganda
2	President Donald Trump, as he often does while...	non-propaganda
3	February is Black History Month, and nothing l...	non-propaganda
4	The snow was so heavy, whipped up by gusting w...	non-propaganda
5	Four months after the Sandy Hook School shooti...	non-propaganda
6	The first major newspaper article about Donald...	non-propaganda
7	For three years, starting in 2008, New York ar...	non-propaganda
8	President Donald Trump's tumultuous administra...	non-propaganda
9	With Hartford on edge about the future of Aetn...	non-propaganda
10	An employee at a Hibachi Express in Florida ha...	non-propaganda
11	With the toll of the carnage from the country'...	non-propaganda
12	The State Department's point-man on North Kore...	non-propaganda
13	The Trump Organization announced Monday that i...	non-propaganda
14	Aer Lingus' service from Bradley International...	non-propaganda
15	For its show "Constellations," which ends its ...	non-propaganda
16	The Corporation for Public Broadcasting (CPB) ...	non-propaganda
17	All five members of New Britain's state legisl...	non-propaganda
18	The leader and second in command of a credit-c...	non-propaganda

Рисунок 1 – Датасет для розпізнання на рівні статті

– Розпізнання пропаганди на рівні речень. Набір даних для цього типу задачі містить у собі близько 450 англомовних статей (включно із заголовками), розбитих на окремі речення. Кожне речення марковане як «пропагандистське» або «не пропагандистське». Також набір даних містить у собі унікальний ідентифікатор для кожної статті, а також унікальний ідентифікатор для речення у межах статті, до якої воно належить. Дані подано як набір текстових файлів, окремий для кожної статті. Атрибути даних також зберігаються окремо. Загалом набір даних налічує 15168 речення (рис. 2). Після завантаження кожної колекції файлів, конкатенуємо їх та формуємо єдиний pandas.DataFrame, усуваючи із набору даних унікальні ідентифікатори статей/речень. Оскільки обидва набори даних є промарковані за двома класами, під час реалізації проекту вирішуватиметься задача бінарної класифікації. Існує багато відомих методів вирішення цієї задачі,

DOI 10.15588/1607-3274-2024-2-13

включно із алгоритмами нейронних мереж, проте заради економії обчислювальних ресурсів зупинимося на класичних методах ML із вчителем. Одними із найбільш відомих та ефективних моделей є SVM, наївний класифікатор Баєса та логістична регресія. Розглянемо кожен із них.

	sentence	label
0	US bloggers banned from entering UK	non-propaganda
2	Two prominent US bloggers have been banned fro...	non-propaganda
4	Pamela Geller and Robert Spencer co-founded an...	propaganda
6	They were due to speak at an English Defence L...	non-propaganda
8	A government spokesman said individuals whose ...	non-propaganda
...	...	...
15164	This is a Moon of Alabama fundraiser week.	non-propaganda
15165	No one pays me to write these blog posts.	non-propaganda
15166	If you appreciated this one, or any of the 7,0...	non-propaganda
15167	Posted by b on November 29, 2018 at 10:23 AM |...	non-propaganda
15168	Comments	non-propaganda

14263 rows × 2 columns

Рисунок 2 – Датасет для розпізнання на рівні речення

– Метод SVM – це модель, призначена для бінарної класифікації, зокрема, для класифікації текстових даних (рис. 3) із використанням неімовірнісного лінійного бінарного класифікатору. Модель SVM є поданням зразків як точок у просторі, де зразки з окремих категорій розділено найбільш оптимальною гіперплощиною. Модель може мати декілька видів ядр, проте найчастіше використовується із лінійним ядром. Для того, аби класична модель SVM із лінійним ядром показувала хороші результати, дані повинні бути лінійно розділеними. Працює ефективно у тому випадку, коли між даними різних класів є чіткий розподіл, дані є багатовимірними та тоді, коли кількість вимірів є більшою за загальну кількість екземплярів даних. Проте не призначена для роботи із наборами даних великого обсягу та не є ефективною у випадку, коли у наборі даних є багато «шуму», а також тоді, коли кількість екземплярів даних одного класу перевищує кількість екземплярів іншого класу, тобто дані повинні бути збалансованими. Як бачимо на рис. 3, оскільки наші набори даних не є збалансованими, лінійна модель SVM погано впоралась із своєю задачею – багато зразків пропагандистських текстів були класифіковані як не пропагандистські.

– Наївний класифікатор Баєса для визначення ймовірності приналежності екземпляру до одного з класів, приймаючи гіпотезу незалежності змінних (рис. 4). Наївний класифікатор Баєса є не надто чутливим до відсутності певних значень атрибутів у наборі, швидше працює у тому випадку, коли розмір тренувальної вибірка є відносно великим. Проте приймає гіпотезу незалежності даних, тому може бути неефективним у тому випадку, якщо вони пов’язані між собою, а також є дуже чутливим до форми вхідних даних. Як бачимо на рис. 4, оскільки текстові дані мають багато шуму, наївний класифікатор Баєса також погано впорався із своєю задачею.

– Логістична регресія побудована на основі лінійної регресії, проте, на відміну від неї, логістична регресія призначена для задачі класифікації (рис. 5).

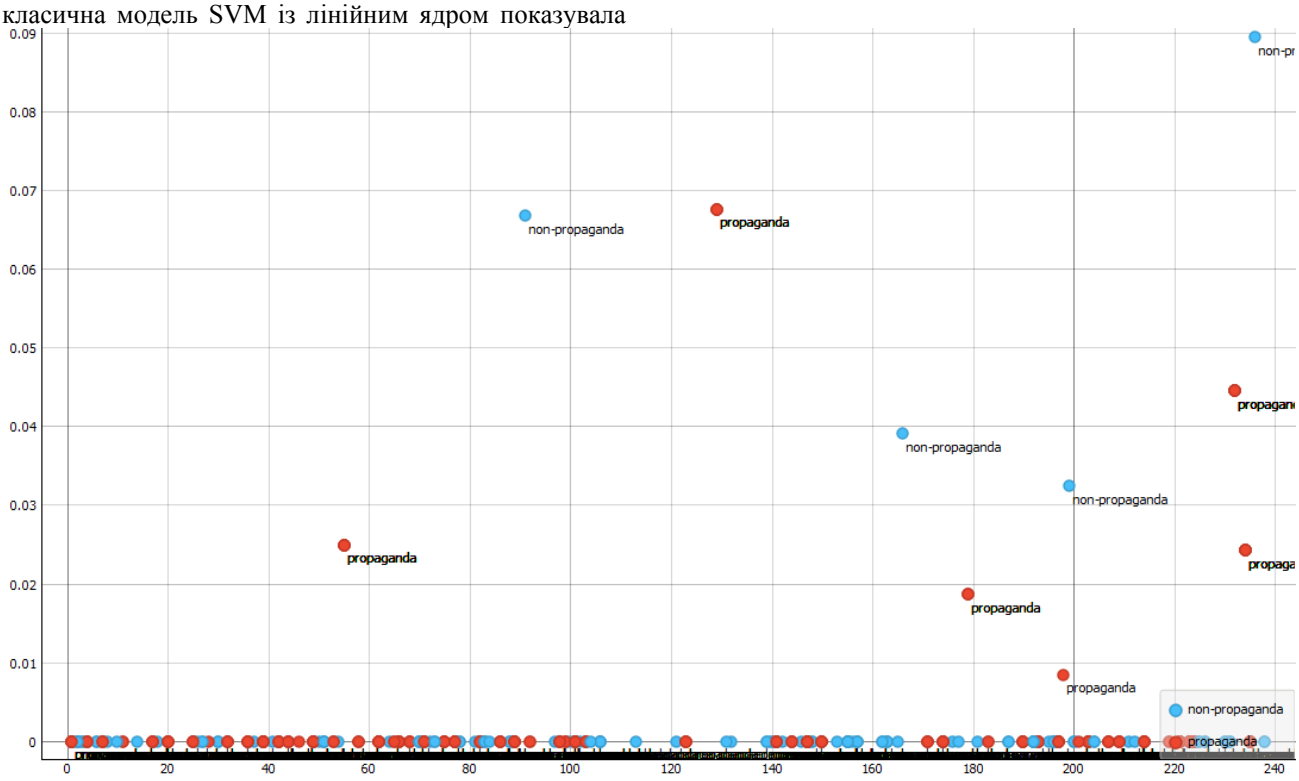

Рисунок 3 – Візуалізація роботи алгоритму SVM

DOI 10.15588/1607-3274-2024-2-13

Прогнозування імовірності належності змінної до того чи іншого класу визначається шляхом порівняння значення із логістичною кривою. Модель логістичної регресії не є чутливою до перенавчання у тому випадку, коли набір даних не є багатовимірним, проте навіть у такому випадку можна використати алгоритм регуляризації.

Зупиняємо вибір на моделі логістичної регресії, оскільки не можемо гарантувати незалежність змінних одна від одної для текстових даних (вимога наївного класифікатору Баєса), а також відсутність шуму (може викликати неефективність у роботі SVM). До того ж, наші дані не збалансовані з точки зору розподілу за класами. Оскільки сирий текст сам по собі не має жодних ознак та атрибутів та не є придатним для використання у моделі ML, необхідно також визначити методи вилучення ознак.

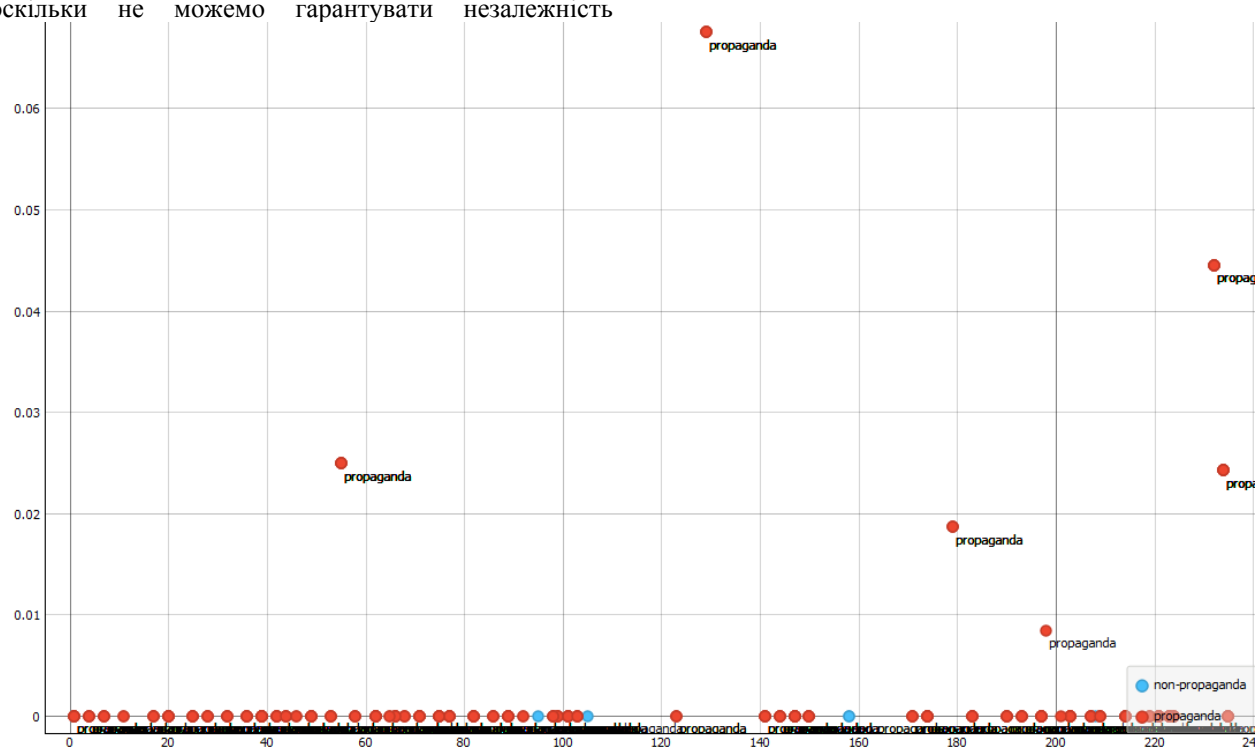

Рисунок 4 – Візуалізація роботи алгоритму наївного Баєса

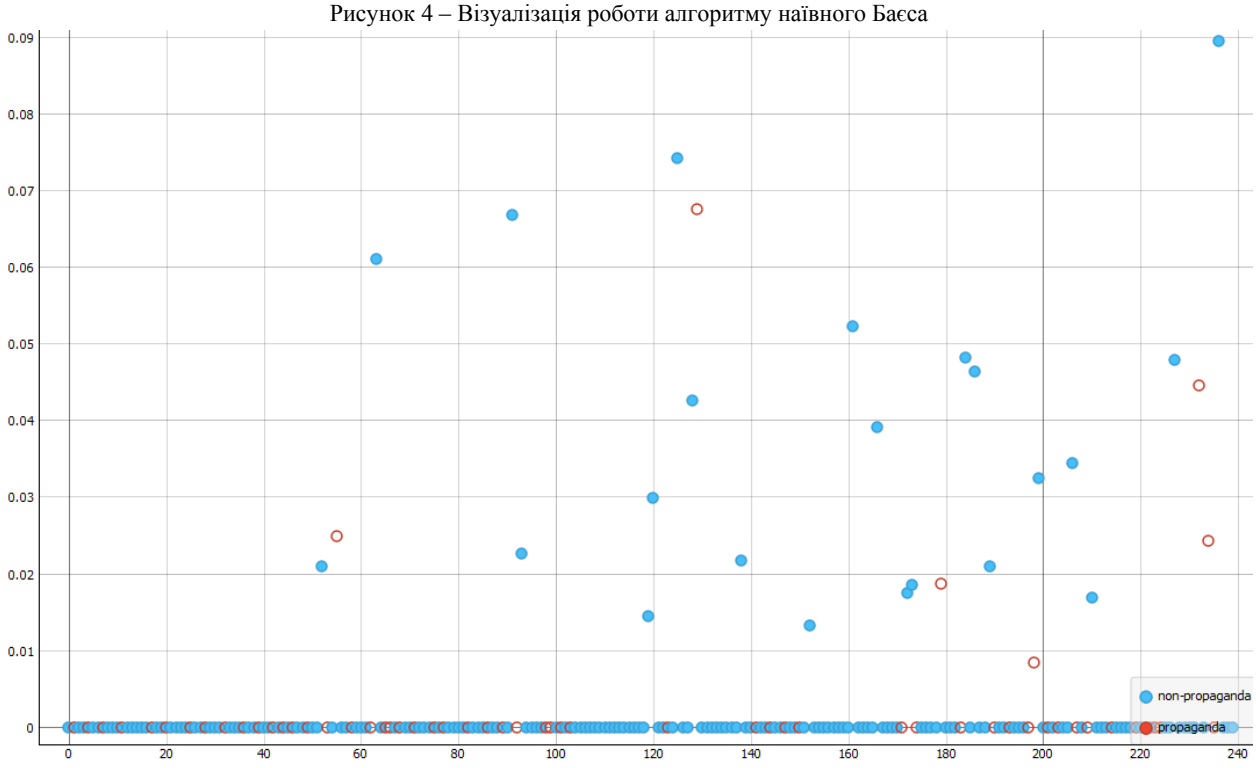

Рисунок 5 – Візуалізація роботи моделі логістичної регресії

DOI 10.15588/1607-3274-2024-2-13

Перелік методів є наступним:

– Векторизація текстових даних за моделлю «Торба слів». Кожному слову (терму) у корпусі присвоюється певне число, а текст перетворюється на вектор розмірністю N, де N – це загальна кількість слів у корпусі, у якому значення кожного елемента дорівнює частоті терму.

– TF-IDF трансформація на основі оцінки важливості слів у контексті статті/речення, що є частиною колекції статей/речень.

– Розмічування частин мови. Форматування текстових даних у вигляді «%слово%_%частина мови, до якої належить слово%_%лема слова%».

– Використання Word2Vec моделі для вбудовування слів. Використання неглибокої двошарової нейронної мережі для векторизації слів із одночасним зменшенням кількості вимірів.

4 ЕКСПЕРИМЕНТИ

Основою методології дослідження пропонуємо синтезовану технологію на основі методів штучного інтелекту, комп'ютерної лінгвістики, машинного навчання, інтелектуального аналізу даних, статистичної обробки даних, теорії систем та системного аналізу, комп'ютерного та імітаційного моделювання тощо. Проблема складається з двох основних складових – визначення множини інформації як фейкової та основі неї знайти джерела та проаналізувати неавтентичну поведінку учасників.

Принцип функціонування ІС автоматичного виявлення джерел дезінформації та неавтентичної поведінки користувачів чатів:

Етап 1. Визначення множини інформації як фейкової:

Крок 1.1. Збір та інтеграція контенту відповідної мови з відповідних ресурсів у СД.

Крок 1.2. Перевірка чи блокований контент з конкретного ресурсу в певному регіоні.

Крок 1.3. Маркування кожного контенту як блокований/неблокований на певному регіоні з відповідним додаткових метриками (час, ресурс, частота появи блокованих/неблокованих дублів, наявність в назві/дайджесті/анотації відповідних маркованих слів, наприклад власні назви, тощо)

Крок 1.4. Формування проміжної бази маркових відсортованих даних.

Крок 1.5. Застосування до топового контенту NLP методів для розрахунку потенційності фейку і/або теми як причини блокування контенту на певному регіоні на основі словників та множини метрик. Схема NLP процесу визначення теми контенту:

1.5.1. Визначення множини ключових слів відповідного контенту та множини наявних слів-маркерів (власних назв, абревіатур, топ-слів відповідної теми тощо). Визначення якщо можливо теми контенту (метод класифікації тексту).

1.5.2. Якщо за ключовими словами складно визначити тему – ідентифікація стійких словосполучень. Визначення якщо можливо тему контенту

1.5.3. Якщо за стійкими ключовими словами складно визначити тему – проведення семантичного аналізу та побудова онтології. Визначення якщо можливо тему контенту.

1.5.4. Якщо за результатами семантичного аналізу це зробити неможливо – відповідно маркувати та передати в список для роботи модератора контенту

1.5.5. При визначеній темі якщо контент маркований як блокований перевірити з списком тем блокованих на цьому регіоні раніше тем. Якщо немає – поновити список. Якщо є поновити кількість блоків цієї теми як цензури в конкретному регіоні.

Крок 1.6. Застосування технологій ML для покращення аналізу/маркування/ NLP даних. Попередньо тренування моделей ML на перевіреному тренувальному датасеті.

Крок 1.7. Формування моделей/шаблонів потенційних фейків для поновлення списку метрик сортування маркованого контенту на кроці 1.3 та метрик/словників для NLP.

Крок 1.8. Постійне поновлення проміжної бази маркових відсортованих даних та переведення в архів застарілого контенту.

Крок 1.9. Поновлення тренувального датасету для вдосконалення моделей ML. Загальна схема процесу навчання та тренування модуля аналізу дезінформації:

Конвеєр 1.9.1. Попередньо марковані дані → NLP → ML → Моделі/шаблони/метрики

Конвеєр 1.9.2. Вхідні нові дані → Маркування даних (блоковані/неблоковані) → NLP → ML → Маркування контенту (фейк/не фейк) або знаходження потенційної причини блокування (не фейк, але саме ця подія/тема є забороною на певному регіоні для пересічної аудиторії). Загальна схема ІС розпізнавання пропаганди подана на рис. 6. Процес розпізнання пропаганди на рівні статті подано на рис. 7, а на рівні речення – на рис. 8.

Етап 2. Ідентифікація джерел та аналіз неавтентичної поведінки учасників

Крок 2.1. Створення моделі виявлення дезінформації, яка об'єднує функції центрального рівня (зокрема особливості теми) та функції периферійного рівня (зокрема лінгвістичні особливості, особливості настрою та особливості поведінки користувачів),

Крок 2.2. Оцінювання здатності функцій центрального та периферійного рівня автоматично відрізняти дезінформацію від правдивої в межах тематичної онлайн-спільноти за допомогою різних методів машинного навчання.

DOI 10.15588/1607-3274-2024-2-13

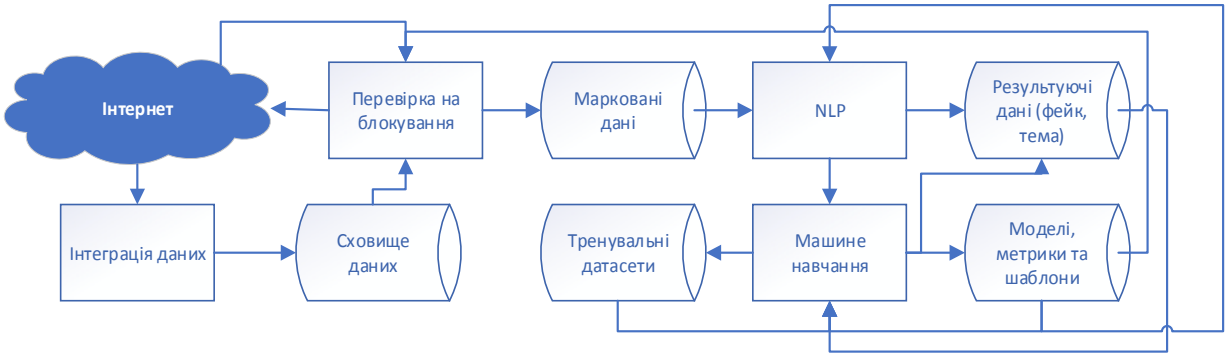

Рисунок 6 – Загальна схема системи розпізнавання пропаганди

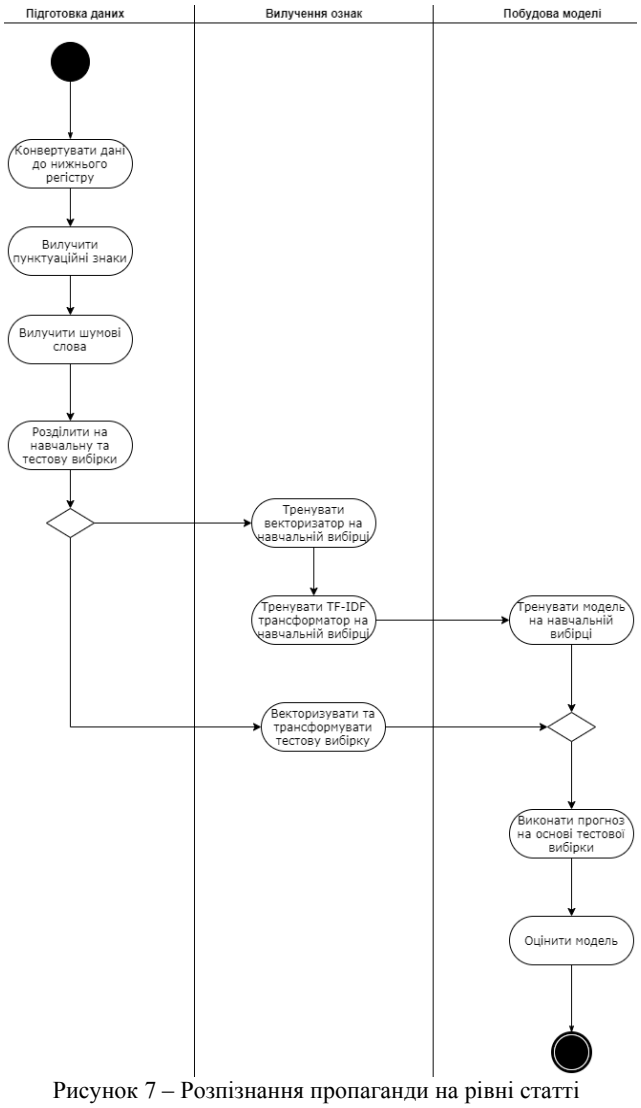

Рисунок 7 – Розпізнання пропаганди на рівні статті

Крок 2.3. Інтелектуальний пошук фейків на основі машинного навчання.

Крок 2.4. Знаходження множини стилістично подібних фейків для одного автора.

Крок 2.5. Знаходження першоджерел фейку на основні аналізу графу розповсюдження.

Крок 2.6. Аналіз поведінки автора/колективу/бота за тривалий проміжок часу для формування множини основних характерних поведінкових рис.

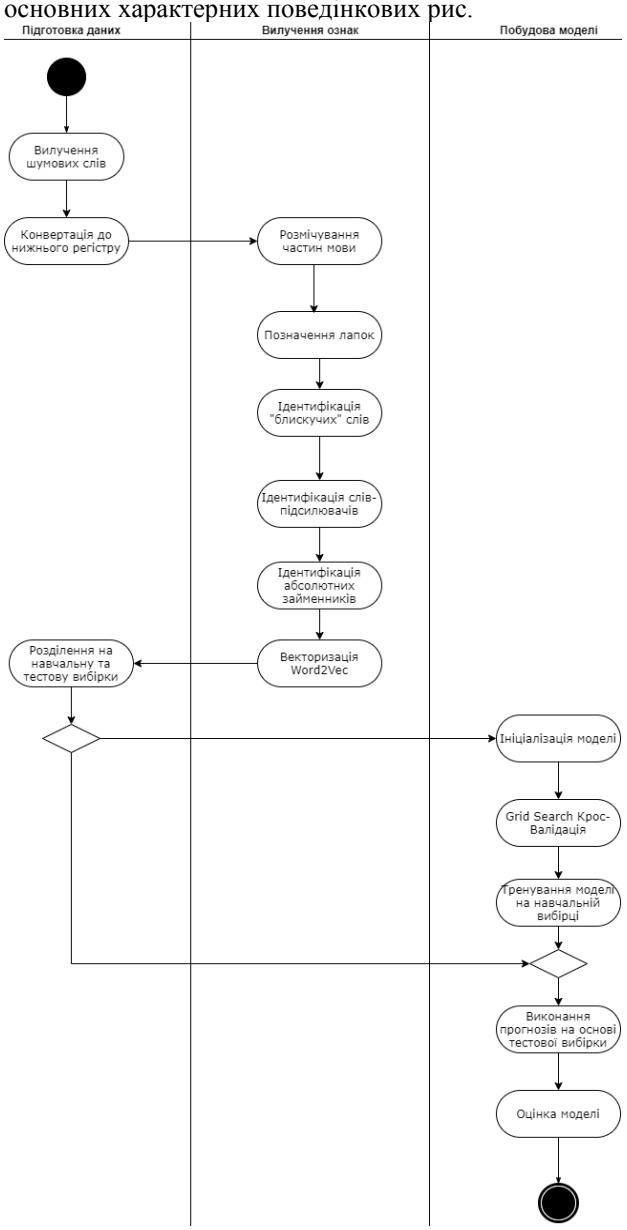

Рисунок 8 – Розпізнання пропаганди на рівні речень

DOI 10.15588/1607-3274-2024-2-13

Крок 2.7. Знаходження інших фейків автора за його стилем написання та поведінки.

Крок 2.8. Формування портрету поведінки автора та моделі передбачення поведінки.

Крок 2.9. На основі аналізу інформаційних портретів різних авторів формувати прогнози розвитку та розповсюдження фейків (частота, густина, тематика), наприклад для ІПСО.

Розпізнання пропаганди на рівні статті (рис. 7):

– Підготовка даних – очищення даних, усунення непотрібних слів та атрибутів (шумові слова та символи пунктуації), конвертування даних до нижнього регістру, розділення даних на навчальну та тестову вибірки;

– Вилучення ознак – виокремлення атрибутів із текстових даних – ініціалізація векторизатора та TF-IDF трансформатора, їх тренування та відповідні перетворення для начальної та тестової вибірок;

– Побудова моделі – ініціалізація моделі, тренування, виконання прогнозів на основі тестової вибірки та оцінка ефективності роботи моделі.

Розпізнання пропаганди на рівні речення (рис. 8):

– Підготовка даних – вилучення шумових слів та конвертування до нижнього регістру (знаки пунктуації знадобляться нам для ідентифікації певних ознак пропаганди), а також розділення даних на навчальну та тестову вибірки;

– Вилучення ознак – розмічення частин мови, позначення лапок, ідентифікація так званих «блискучих» слів, слів-підсилювачів та абсолютних займенників, а також векторизація за допомогою нейронної мережі Word2Vec;

– Побудова моделі – ініціалізація моделі, виконання Grid Search крос-валідації, тренування моделі та виконання прогнозів через тестову вибірку.

5 РЕЗУЛЬТАТИ

Зупиняємо вибір на мові Python за рахунок більш простої майбутньої інтеграції, а також більшої кількості обчислювальних можливостей. Зокрема, використовуватимемо наступні бібліотеки Python: scikit-learn (для побудови моделі, виокремлення ознак та застосування метрик); pandas, numpy (для збереження та маніпуляції даними); spacy (для розмічування частин мови); nltk (для видалення шумових слів); genism (для використання Word2Vec моделі); seaborn, matplotlib (для візуалізації).

– Розпізнання пропаганди на рівні статті (рис. 9). Переглянемо розподіл даними за їх класами для того, аби зрозуміти рівень їх збалансованості.

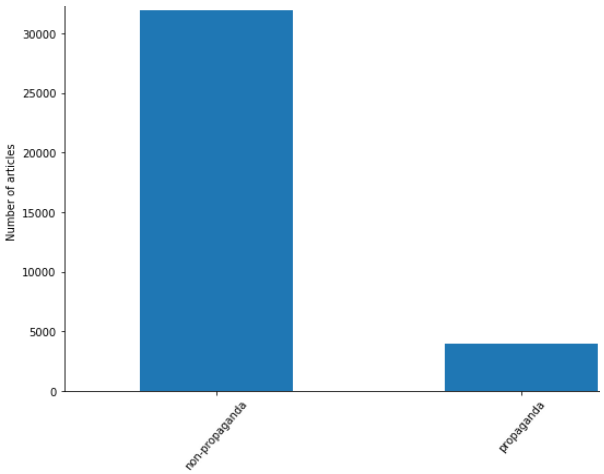

Рисунок 9 – Графік розподілу даних для розпізнання пропаганди на рівні документу за класами

Як бачимо, дані не є збалансованими та у наборі суттєво переважають статті не пропагандистського характеру. А саме, набір даних містить 31972 статті, марковані як «не пропаганда» та 4021 статті, марковані як «пропаганда». Перед тим, як розпочати процес виділення ознак, необхідно виконати деякі операції для очистки та підготовки даних. В першу чергу, необхідно виконати операцію конвертування кожного слова у наборі даних до нижнього регістру для того, аби під час процесу векторизації два ідентичних слова, котрі починаються із різного регістру літер, не враховувалися у якості окремих токенів. Для цього виконуємо наступне перетворення:

```
data['article'] = data['article'].apply(lambda
x: " ".join(x.lower() for x in x.split()))
```

Далі вилучаємо із набору даних пунктуаційні знаки, оскільки на даному рівні вирішення задачі пунктуація не є інформативною ознакою, проте під час процесу векторизації буде рахуватися у якості окремого токену, що може призвести до формування зайвого шуму у даних. Виконуємо перетворення:

```
data['article'] =
data['article'].str.replace('[^\w\s]','')
```

Вилучаємо шумові слова (не несуть змістовного навантаження). Для цього використаємо вбудований корпус шумових слів бібліотеки nltk та у циклі вилучимо їх із статей.

```
stop = nltk.stopwords.words('english')
data['article'] = data['article'].apply(lambda
x: " ".join(x for x in x.split() if x not in
stop))
```

Після цього можемо розділити дані на навчальну та тестову вибірки.

```
X = data['article']
y = data['label']
X_train, X_test, y_train, y_test =
train_test_split(X, y, test_size=0.2,
random_state=42)
```


DOI 10.15588/1607-3274-2024-2-13

Вилучення ознак. Виконуємо векторизацію тексту. Використаємо клас CountVectorizer із scikit-learn.

```
vectorizer = CountVectorizer(analyzer='word',
token_pattern=r'\w{1,}', ngram_range=(1,2),
strip_accents='unicode', min_df=3, max_df=0.5)
X_train = vectorizer.fit_transform(X_train)
X_test = vectorizer.transform(X_test)
```

У якості вихідного результату методи fit та transform екземпляру класу CountVectorizer продукують розріджену матрицю, розмірність якої дорівнює кількості унікальних слів у тексті. Переконаємося у тому, що після трансформації кількість атрибутів навчальної та тестової вибірки співпадають (рис. 10).

```
: X_train.shape
: (28794, 605048)

: X_test.shape
: (7199, 605048)
```

Рисунок 10 – Перевірка розмірностей навчальної та тестової вибірок даних на рівні документу

Далі виконаємо TF-IDF трансформацію. Використовуємо клас TfidfTransformer із scikit-learn.

```
transformer = TfidfTransformer(use_idf=True,
smooth_idf = True)
X_train = transformer.fit_transform(X_train)
X_test = transformer.transform(X_test)
```

Перетворення відбується для попередньо сформованої розрідженої матриці векторизованих текстових даних. Побудова моделі.

```
LogisticRegression(C=1.0,
    class_weight='balanced',
    dual=False, fit_intercept=True,
    intercept_scaling=1,
    l1_ratio=None, max_iter=100,
    multi_class='auto', n_jobs=None,
    penalty='12', random_state=None,
    solver='lbfgs', tol=0.0001,
    verbose=0, warm_stert=False)
```

Для моделювання використаємо клас LogisticRegression із бібліотеки scikit-learn.

```
model = LogisticRegression(penalty='12',
class_weight='balanced', solver='lbfgs')
model.fit(X_train, y_train)
```

Серед параметрів вказуємо penalty='l2', тобто для регуляризації модель використовуватиме метод гребеневої регресії, а solver='lbfgs' означає, що для оптимізації модель використовуватиме алгоритм Бройдена-Флетчера-Гольдфарба-Шанно із обмеженим використанням пам'яті.

Підготовка даних. Побудуємо графік розподілу даних за категоріями для оцінки рівня незбалансованості даних (рис. 11). Набір даних не є збалансованим та у ньому знову переважують речення не пропагандистського характеру.

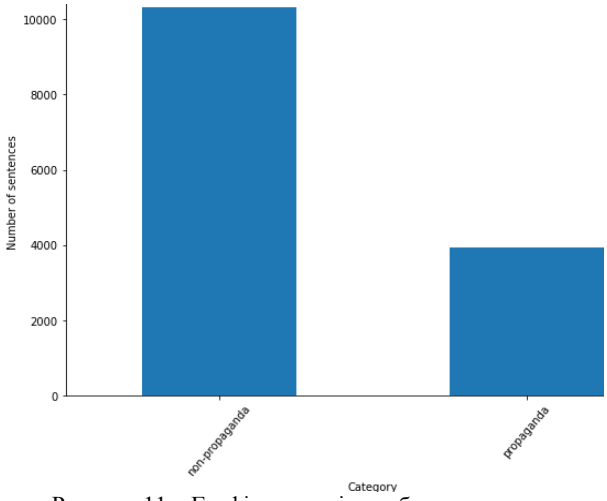

Рисунок 11 – Графік розподілу набору даних для розпізнання пропаганди а рівні речення за класами

Далі повторюємо процес видалення шумових слів із набору даних. Після цього знову конвертуємо текстові дані до нижнього регістру.

```
stop = stopwords.words('english')
data['sentence'] = data['sentence'].apply(lambda
x: " ".join(x for x in x.split() if x not in
stop))
data['sentence'] = data['sentence'].apply(lambda
x: " ".join(x.lower() for x in x.split()))
```

На даному рівні задачі не видаляємо пунктуаційні знаки із речень, оскільки вони знадобляться на етапі виділення ознак. Виконуємо розмічення частин мови. Для цієї операції використовуємо бібліотеку spacy.

```
nlp = spacy.load('en')
def tag(sentence):
                global nlp
                doc = nlp(sentence)
                return "
".join([f'{x.text}_{x.tag_}_{x.lemma_}' for x in
doc])
sentences_pos = copy.deepcopy(data['sentence'])
tagged = pos_tagging(sentences_pos)
data['tagged'] = tagged
```

Маркуємо кожне речення у наборі даних за наявністю у ньому фігурних лапок. Це дозволяє відслідкувати техніку пропаганди як «Апеляція до авторитету». Ініціалізуємо відповідну функцію.

```
def get_quotations(sentences):
    result = []
    for sentence in sentences:
        match = 1 if '"' in sentence else 0
        result.append(match)
    return np.array(result).reshape(-1, 1)
```

Далі перевіряємо кожне речення на наявність так званих «блискучих слів». Прикладами таких слів є «патріотизм», «свобода», «сила», «ідея» тощо. Виконуємо цю операцію для ідентифікації таких методики пропаганди, як «Розмахування прапором» та «Слоган». Для цього обчислюємо кількість співпадінь між словами у речення та словами у лексиконі «блискучих слів» та розділяємо це значення на загальну кількість слів у речення для того, щоб нормалізувати коефіцієнт. Формуємо відповідний

DOI 10.15588/1607-3274-2024-2-13

лексикон у вигляді текстового файлу із «блискучими словами» та ініціалізуємо відповідну функцію.

```
def get_glitter(tagged):
    filename = 'glitter_words.txt'
    glitters = []
    append = glitters.append
    with open(os.path.join(LEXICONS_PATH,
filename), encoding='utf-8') as f:
        for line in f.readlines():
            append(line.replace('\n', ''))
    result = []
    for sentence in tagged:
        words = 0
        matches = 0
        for wline in sentence.split():
            try:
                w, t, l = wline.split("_")
            except:
                continue
            w = w.lower()
            l = l.lower()
            words+=1
            if l in glitters or w in glitters:
                matches+=1
        if words == 0:
            result.append(0)
        else:
            result.append(matches/words)
    return np.array(result).reshape(-1, 1)
```

Використовуємо аналогічний підхід для пошуку у реченнях слова-підсилювачів («неймовірно», дуже», «абсолютно», «тощо) та абсолютних займенників («усі», «ніхто» тощо). Намагаємося відповідно ідентифікувати таки методики пропаганди як «Загальна платформа» та «Навантажена мова». Формуємо аналогічні лексикони та іцініалізуємо відповідні функції. У якості останнього кроку, векторизуємо текстові дані за допомогою моделі Word2Vec. Для цієї задачі використовуємо заздалегідь натреновану на наборі даних із соціальної мережі Twitter модель, котра має 200 вимірів. Використовуємо бібліотеку gensim.

```
w2v_file = os.path.join(WORD2VEC_PATH,
'twitter.27B.200d.txt')
w2v_model =
KeyedVectors.load_word2vec_format(w2v_file,
binary=False)
def w2v_vectorize(tagged):
    global w2v_model
    X = []
    ndims = 200
    for sentence in tagged:
        words = []
        for wline in sentence.split():
            try:
                w, t, l = wline.split("_")
            except:
                continue
            words.append(w)
        row_data = np.mean([w2v_model[w] for w
in words if w in w2v_model] or
                           [np.zeros(ndims)],
axis=0).tolist()
        X.append(row_data)
    X = np.array(X)
    X_std = (X - X.min(axis=0)) / (X.max(axis=0)
- X.min(axis=0))
    X_scaled = X_std * (1 - 0) + 0
    return X_scaled
```

Формуємо новий DataFrame на основі уже існуючого з використанням усіх описаних операцій.

```
word2vec_features =
w2v_vectorize(data['tagged'])
word2vec_columns = [f'dim{x}' for x in
range(200)]
glitter_words = get_glitter(data['tagged'])
quotations = get_quotations(data['sentence'])
intensifiers = get_intensifiers(data['tagged'])
absolutes = get_absolutes(data['tagged'])
X = pd.DataFrame(word2vec_features,
columns=[word2vec_columns])
X['quotations'] = quotations
X['glitter_words'] = glitter_words
X['intensifiers'] = intensifiers
X['absolutes'] = absolutes
y = data['label']
```

Розділяємо датасет на навчальну/тестову вибірки.

```
X_train, X_test, y_train, y_test =
train_test_split(X, y, test_size=0.2,
random_state=42)
```

Використовуємо алгоритм крос-валідації Grid Search для пошуку найкращих параметрів, ініціалізуємо та тренуємо модель.

```
lr_model = LogisticRegression()
penalty = ['l1', 'l2']
C = np.logspace(0, 4, 10)
hyperparameters = dict(C=C, penalty=penalty)
clf = GridSearchCV(lr_model, hyperparameters,
refit='f1', cv=5)
best_model = clf.fit(X_train, y_train)
```

Отримані найкращі параметри моделі є наступними: penalty=’l2’, C=7.74.

6 ОБГОВОРЕННЯ

Проаналізуємо та оцінимо роботу моделі розпізнання пропаганди на рівні статті. Для цього виконуємо прогнози на основі тестової вибірки. Побудуємо матрицю помилок (рис. 12). Матрицю помилок можемо інтерпретувати наступним чином: моделі вдалося коректно класифікувати 6097 не пропагандистських статей та 694 статті пропагандистського характеру. 123 пропагандистські статті та 285 не пропагандистських статей були класифіковані невірно. Отримана оцінка моделі: 0.9433254618697041.

```
n = len(conf_matrix)
plt.figure(figsize = (n, n))
sns.heatmap(conf_matrix, annot=True)
plt.title("Confusion Matrix")
plt.ylabel("Actual Value")
plt.xlabel("Predicted Value")
plt.show()
```

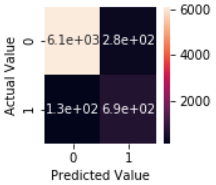

Рисунок 12 – Побудована матриця помилок для моделі розпізнання пропаганди на рівні статті

Після цього виконаємо аналогічні дії для моделі розпізнання пропаганди на рівні речень. Отже, знову виконуємо прогнозування на основі тестової вибірки. Будуємо аналогічну матрицю помилок (рис. 13).

DOI 10.15588/1607-3274-2024-2-13

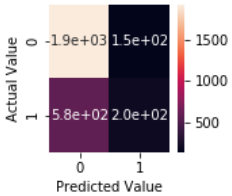

Рисунок 13 – Матриця помилок на рівні речень

Матриця помилок може бути інтерпретована наступним чином: модель успішно класифікувала 1917 не пропагандистських статей та 205 пропагандистських статей, проте 585 пропагандистських статей та 146 не пропагандистських статей були класифіковані невірно. Оцінка моделі становить: 0.7437784787942516. Обидві оцінки є задовільними, проте модель для розпізнання пропаганди на рівні документу впоралася краще. До того, враховуючи незбалансованість обидвох наборів даних бачимо, що друга модель (для розпізнання пропаганди на рівні речень) невірно класифікувала більше екземплярів даних, котрі були марковані як пропагандистські. Це може свідчити про те, що модель є недостатньо специфічною, тобто у майбутніх розробках необхідно використати складніший, більш комплексний алгоритм та вилучити більше специфічних ознак для того, аби мати можливість точніше ідентифікувати інші методики пропаганди.

ВИСНОВКИ

На етапах опрацювання дезінформації пропонується новий метод аналізу пропаганди для ідентифікації ознак та зміни динаміки поведінки скоординованих груп на основі машинного навчання.. Впровадження отриманих результатів дозволить суттєво скоротити час на прийняття найбільш адекватного рішення щодо впровадження заходів боротьби з дезінформацією стосовно ідентифікованих скоординованих груп генерування, дезінформації фейків і пропаганди.

Під час виконання роботи реалізовано дві моделі для розпізнання пропаганди у текстових даних – на рівні статті та на рівні речення. Для цього розв'язано задачу бінарної класифікації тексту. Обидві моделі побудовані на основі логістичної регресії, у процесі підготовки даних та виокремлення ознак застосовано такі методи, як векторизація за моделлю «Торба слів», TF-IDF векторизація, розмічування частин мови, вбудовування слів за допомогою двошарової нейронної мережі Word2Vec, а також ручні методи виокремлення ознак, котрі націлені на ідентифікацію конкретних методик політичної пропаганди у текстах. Проаналізовано аналоги розроблюваного проекту, досліджено ПО (застосування пропаганди у ЗМІ та основі методики її продукування). Програмну реалізацію виконано за допомогою Python, із використанням бібліотек scikit-learn, pandas, numpy, spacy, nltk, genism, matplotlib, seaborn. Отримана оцінка моделі для розпізнання пропаганди на рівні статті: 0.9433254618697041, а на рівні речень: 0.7437784787942516.

Очікувані результати виконання проєкту:

– вперше розроблені основи та основні принципи синтезованої інформаційної технології автоматичного виявлення джерел дезінформації та неавтентичної поведінки користувачів чатів, що дозволить своєчасно виявляти деструктивні і підозрілі спільноти в різних соціальних мережах, визначати їх лідерів і кураторів, виявляти інформаційні загрози в повідомленнях користувачів, попереджати поширення фейковий і шкідливої інформації.

– вперше розроблено метод стилістичного аналіз та лінгвістичного опрацювання дезінформації для формування інформаційного портрету автора/бота генерування текстового контенту як частини параметрів пошуку як подібного авторського контенту, так і шляхів розповсюдження.

– вироблені критерії та параметри неавтентичної поведінки користувачів чатів для формування інформаційних портретів потенційних розповсюджувачів дезінформації та виявлення маршрутів та механізмів розповсюдження, частоти генерування фейків, тематики та ключові слова, характерні для відповідної групи.

NLP процесу визначення контенту як фейку/не феку є складним процесом, так як дуже залежить не лише від швидкості/якості попередньо зібраного/ інтегрованого та опрацьованого контенту (блокований/неблокований на певному регіоні, теми контенту) але від ефективно підібраної моделі машинного навчання на тренувальних датасетах. Зазвичай фейк не блокується. Мета його створення – по швидше розповсюдити як у всьому світі, так і на тих регіонах, де зазвичай правдива інформація (не фейк) потенційна може блокуватися (не гарантовано). Якщо не фейк інформація заблокована на певній території, а протилежна інформація (фейк) з цієї території розповсюджується – то шанс ідентифікувати фейк збільшується. Якщо і нефейк не заблокований та фейк паралельно вільно розповсюджується, тут методи NLP не допоможуть. Вони лише можуть промаркувати дві множини протилежними поясненнями щодо події/явища. І лише при додаткових статистичних дослідженнях можна ідентифікувати як множина з фейками, а яка ні. Складність ще полягає в самій мові контенту, зокрема в українській. Для порівняння з англомовним контентом українська/російська мови є досить складними для автоматичного опрацювання, особливо при аналізі семантики та розбудові онтології. Стандартні та традиційні методи, які застосовують для опрацювання англомовних тестів, не підходять для опрацювання слов'янських мов, в тому числі для виявлення дезінформації та особливостей стилістики авторів-генераторів фейків та пропаганди. Аналогічно крім того що неавтентична поведінка користувачів чатів як людей і ботів відрізняється, так і відрізняється людей з різною вмотивованістю (віра в

DOI 10.15588/1607-3274-2024-2-13

пропаганду, робота за гроші, просто одна з видів вандалізму та так би мовити дозвілля), національністю, освітою, статтю, ментальністю, рівнем знання мови тексту, ступенем віри, інтелектом, тощо. Це все значно впливає на процес визначення критеріїв поведінки різних спільнот та в межах однієї спільноти, що в свою чергу значно впливає на формування інформаційного портрету неавтентичної поведінки користувачів різних чатів (те що властиве для пропагандиста-мусульманина, суттєво відрізняється для представника рашки або днр/лнр).

Обґрунтування практичної цінності запланованих результатів проєкту для економіки та суспільства.

– Зменшення обсягів дезінформації, фейків та пропаганди та частоти/регулярності публікації за рахунок відстежування стилістично подібного контенту та маршрутів розповсюдження.

– Зменшення негативного впливу дезінформації на настрої суспільства та зменшення ступеня керування громадською думкою через розповсюдження пропаганди в інформаційній війні. Наприклад пригнічення психіки молоді (у т.ч. порушення психіки, доведення до летальних наслідків), спонукання їх до асоціальної поведінки, формування груп громадської непокори чи агресивної поведінки за сфабрикованими приводами, аналіз соціальних наслідків кібератак тощо.

– Зменшення ціни пошуку, ідентифікації та блокування дезінформації, авторів/цільових розповсюджувачів та джерел.

Розробка вищеописаних методів спрямована на виявлення загроз, стороннього втручання (атак) на ранніх стадіях, класифікацію загроз за видами та подальше протистояння кожному виду загроз.

ЛІТЕРАТУРА

1. Zhao Y. Detecting health misinformation in online health communities: Incorporating behavioral features into machine learning based approaches / Y. Zhao, J. Da, J. Yan // Information Processing & Management. – 2021. – Vol. 58(1). – P. 102390. DOI: 10.1016/j.ipm.2020.102390
2. Hartmann M. Mapping (dis-)information flow about the MH17 plane crash / M. Hartmann, Y. Golovchenko, I. Augenstein, // arXiv. – Access mode: https://arxiv.org/abs/1910.01363.
3. Prokipchuk O. Ukrainian Language Tweets Analysis Technology for Public Opinion Dynamics Change Prediction Based on Machine Learning / O. Prokipchuk, V. Vysotska // Radio Electronics, Computer Science, Control. – 2023. – Vol. 2(2023). – P. 103–116. DOI: 10.15588/1607-3274-2023-2-11
4. Ahmed S. Classification of Censored Tweets in Chinese Language using XLNet / S. Ahmed, A. Kumar // Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda :, Association for Computational Linguistics, Online, 2021 : proceedings. – Online: ACL, 2021. – P. 136–139. DOI: 10.18653/v1/2021.nlp4if-1.21
5. NLP Tool for Extracting Relevant Information from Criminal Reports or Fakes/Propaganda Content / [V. Vysotska, S. Mazepa, L. Chyrun et al.] // Computer Sciences and Information Technologies : 17th International Conference, Lviv, 2022, November. – Lviv: IEEE, 2021. – P. 93–98. DOI: 10.1109/CSIT56902.2022.10000563
6. Oliinyk V. A. Propaganda Detection in Text Data Based on NLP and Machine Learning / [V. A. Oliinyk, V. Vysotska, Y. Burov et al.] // CEUR Workshop Proceedings. – 2020. – Vol. 2631. – P. 132–144.
7. Bjola C. Propaganda in the digital age / C. Bjola // Global Affairs. – 2017. – Vol. 3(3). – P. 189–191. DOI: 10.1080/23340460.2017.1427694
8. Vosoughi S. The spread of true and false news online / S. Vosoughi, D. Roy, S. Aral // Science. – 2018. – Vol. 359(6380). – P. 1146–1151. DOI: 10.1126/science.aap9559
9. Propaganda Definitions. – Access mode: https://propaganda.qcri.org/annotations/definitions.html
10. Framing and Agenda-setting in Russian News: a Computational Analysis of Intricate Political Strategies /. [A. Field, D. Kliger, S. Wintner et al.] // arXiv. – Access mode: – https://arxiv.org/abs/1808.09386
11. Garcia-Marín J. The Use of Supervised Learning Algorithms in Political Communication and Media Studies: Locating Frames in the Press / J. Garcia-Marín, A. Calatrava // Pamplona. – 2018. – Vol. 31(3). – P. 175–188. DOI: 10.15581/003.31.3.175-188
12. nginx. – Access mode: https://fgz.texty.org/
13. texty.org.ua. How Texty detects and makes sense of manipulative news. – Access mode: https://medium.com/@texty.org.ua/how-texty-detects-and-makes-sense-of-manipulative-news-1f43d33936eb
14. Hein V. Propaganda detection in Russian and American news coverage about the war in Ukraine through text classification / V. Hein // Diploma Thesis, Technische Universität Wien. – 2023. DOI: 10.34726/hss.2023.104640
15. Ceuşan I. F. European Union policies and strategies to counter Russian propaganda and disinformation / I. F. Ceuşan // L'Europe Unie. – 2023. – Vol. 19(19). – P. 113–122.
16. Perdoor S. Fake News Detection with LSTM and NLP – ProRew1 / S. Perdoor. – Access mode: https://www.kaggle.com/code/superrajdoor/fake-news-detection-with-lstm-and-nlp-prorew1/input //
17. Duratnir İ. Fake News Detection with NLP and LSTM / İ. Duratnir. – Access mode: https://www.kaggle.com/code/ilaydadu/fake-news-detection-with-nlp-and-lstm
18. propaganda-detection-our-data. – Access mode: https://www.kaggle.com/datasets/vladimirsydor/propaganda-detection-our-data

Accepted 09.02.2024.
Received 27.04.2024.

UDC 004.9

INFORMATION TECHNOLOGY FOR RECOGNIZING PROPAGANDA, FAKES AND DISINFORMATION IN TEXTUAL CONTENT BASED ON NLP AND MACHINE LEARNING METHODS

Vysotska V. – Dr. Sc., Associate Professor of Information Systems and Networks Department, Lviv Polytechnic National University, Lviv, Ukraine.

DOI 10.15588/1607-3274-2024-2-13

ABSTRACT

Context. The research is aimed at the application of artificial intelligence for the development and improvement of means of cyber warfare, in particular for combating disinformation, fakes and propaganda in the Internet space, identifying sources of disinformation and inauthentic behavior (bots) of coordinated groups. The implementation of the project will contribute to solving the important and currently relevant issue of information manipulation in the media, because in order to effectively fight against distortion and disinformation, it is necessary to obtain an effective tool for recognizing these phenomena in textual data in order to develop a further strategy to prevent the spread of such data.

Objective of the study is to develop or automatic recognition of political propaganda in textual data, which is built on the basis of machine learning with a teacher and implemented using natural language processing methods.

Method. Recognition of the presence of propaganda will occur at two levels: at the general level, that is, at the level of the document, and at the level of individual sentences. To implement the project, such feature construction methods as the TF-IDF statistical indicator, the "Bag of Words" vectorization model, the marking of parts of speech, the word2vec model for obtaining vector representations of words, as well as the recognition of trigger words (reinforcing words, absolute pronouns and "shiny" words). Logistic regression was used as the main modeling algorithm.

Results. Machine learning models have been developed to recognize propaganda, fakes and disinformation at the document (article) and sentence level. Both model scores are satisfactory, but the model for document-level propaganda recognition performed almost 1.2 times better (by 20%).

Conclusions. The created model shows excellent results in recognizing propaganda, fakes and disinformation in textual content based on NLP and machine learning methods. The analysis of the raw data showed that the propaganda recognition model at the document (article) level was able to correctly classify 6097 non-propaganda articles and 694 propaganda articles. 123 propaganda articles and 285 non-propaganda articles were misclassified. The obtained estimate of the model: 0.9433254618697041. The sentence-level propaganda recognition model successfully classified 205 propaganda articles and 1917 non-propaganda articles. The model score is: 0.7437784787942516 (but 731 articles were incorrectly classified).

KEYWORDS: disinformation, fake, propaganda, linguistic analysis, natural language processing, machine learning, cyber warfare, artificial intelligence, semantic analysis, information security.

REFERENCES

1. Zhao Y., Da J., Yan J. Detecting health misinformation in online health communities: Incorporating behavioral features into machine learning based approaches, *Information Processing & Management*, 2021, Vol. 58(1), P. 102390. DOI: 10.1016/j.ipm.2020.102390
2. Hartmann M., Golovchenko Y., Augenstein I. Mapping (dis-)information flow about the MH17 plane crash, *arXiv*. Access mode: https://arxiv.org/abs/1910.01363.
3. Prokipchuk O., Vysotska V. Ukrainian Language Tweets Analysis Technology for Public Opinion Dynamics Change Prediction Based on Machine Learning, *Radio Electronics, Computer Science, Control,* 2023, Vol. 2(2023), pp. 103–116. DOI: 10.15588/1607-3274-2023-2-11
4. Ahmed S., Kumar A. Classification of Censored Tweets in Chinese Language using XLNet, *Fourth Workshop on NLP for Internet Freedom. Censorship, Disinformation, and Propaganda, Association for Computational Linguistics, Online, 2021, proceedings. Online: ACL,* 2021, pp. 136–139. DOI: 10.18653/v1/2021.nlp4if-1.21
5. Vysotska V., Mazepa S., Chyrun L., Brodyak O., Shakleina I., Schuchmann V. NLP Tool for Extracting Relevant Information from Criminal Reports or Fakes/Propaganda Content, *Computer Sciences and Information Technologies : 17th International Conference,* Lviv, 2022, November. Lviv, IEEE, 2021, pp. 93–98. DOI: 10.1109/CSIT56902.2022.10000563
6. Oliinyk V. A., Vysotska V., Burov Y., Mykich K., Fernandes V. B. Propaganda Detection in Text Data Based on NLP and Machine Learning, *CEUR Workshop Proceedings*, 2020, Vol. 2631, pp. 132–144.
7. Bjola C. Propaganda in the digital age, *Global Affairs,* 2017, Vol. 3(3), pp. 189–191. DOI: 10.1080/23340460.2017.1427694
8. Vosoughi S., Roy D., Aral S. The spread of true and false news online, *Science*, 2018, Vol. 359(6380), pp. 1146–1151. DOI: 10.1126/science.aap9559
9. Propaganda Definitions. Access mode: https://propaganda.qcri.org/annotations/definitions.html
10. Field A. Kliger D., Wintner S., Pan J., Jurafsky D., Tsvetkov Y. Framing and Agenda-setting in Russian News: a Computational Analysis of Intricate Political Strategies, *arXiv*. Access mode: https://arxiv.org/abs/1808.09386
11. Garcia-Marín J., Calatrava A. The Use of Supervised Learning Algorithms in Political Communication and Media Studies: Locating Frames in the Press, *Pamplona,* 2018, Vol. 31(3), pp. 175–188. DOI: 10.15581/003.31.3.175-188
12. nginx. – Access mode: https://fgz.texty.org/
13. texty.org.ua. How Texty detects and makes sense of manipulative news. Access mode: https://medium.com/@texty.org.ua/how-texty-detects-and-makes-sense-of-manipulative-news-1f43d33936eb
14. Hein V. Propaganda detection in Russian and American news coverage about the war in Ukraine through text classification, *Diploma Thesis*, Technische Universität Wien, 2023. DOI: 10.34726/hss.2023.104640
15. Ceuşan I. F. European Union policies and strategies to counter Russian propaganda and disinformation, *L'Europe Unie,* 2023, Vol. 19(19), pp. 113–122.
16. Perdoor S. Fake News Detection with LSTM and NLP – ProRew1. Access mode: https://www.kaggle.com/code/superrajdoor/fake-news-detection-with-lstm-and-nlp-prorew1/input //
17. Duratnir İ. Fake News Detection with NLP and LSTM / İ. Duratnir. Access mode: https://www.kaggle.com/code/ilaydadu/fake-news-detection-with-nlp-and-lstm
18. propaganda-detection-our-data. Access mode: https://www.kaggle.com/datasets/vladimirsydor/propaganda-detection-our-data

DOI 10.15588/1607-3274-2024-2-13