МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Київський національний університет імені Тараса Шевченка

Навчально-науковий інститут філології
Кафедра української мови та прикладної лінгвістики

Автоматичне визначення мови ворожнечі на матеріалі українськомовних текстів соціальних мереж

Кваліфікаційна робота
освітнього ступеня «бакалавр»
за спеціальністю 035 «Філологія»,
спеціалізацією 035.10 «Прикладна лінгвістика»,
галузі знань 03 «гуманітарні науки»
ОПП «Прикладна (комп'ютерна) лінгвістика та англійська мова»
студентки IV курсу
Катерини ДЬОГТЯР

Наукові керівники:
д.філол.н., проф. Наталія ДАРЧУК,
к.техн.н., доц. Микола КОСТІКОВ

«Допущено до захисту»
Протокол № 11 засідання кафедри
української мови та прикладної лінгвістики
ННІФ від 01.06.2023
Завідувач кафедри_____**Сергій Різник**

КИЇВ – 2023

Зміст

ВСТУП

Актуальність теми дослідження. Після появи Інтернету та соціальних мереж, які є просторами та засобами розповсюдження, отримання та трансформації інформації, відбулися зміни у повсякденному житті не тільки користувачів, а й всього суспільства. У сучасному інформаційному світі конфлікти, як і будь-яка інформація, розповсюджуються швидко. Разом з ними набувають актуальності такі поняття як «інформаційна війна», «мережева війна», «ворожнеча», «ненависть», «мова ворожнечі» і т.д.

Українське суспільство сьогодні стикається зі значною кількістю випадків використання мови ненависті, що є наслідком подій російсько-української війни. Агресивна реакція громадян на росіян та воєнні злочини рашистів (напади, погрози, вбивства, зґвалтування, руйнування, грабежі) є захисною реакцією суспільства, яка допомагає витримати величезний тиск на фізичному та моральному, психічному та психологічному, фінансовому рівнях, а також утрату членів родин та майна. У зв'язку з цим актуалізується проблема — суперечність між активною увагою до мови ворожнечі та відсутністю валідного адаптованого інструментарію для її виміру у текстових повідомленнях.

Метою даної роботи є виявлення змістовних складових унаявлення мови ворожнечі та створення програмного засобу, який автоматично визначатиме мову ворожнечі в українськомовних текстах.

Об'єктом дослідження є українськомовні медіадописи. **Предметом** – особливості виявлення змістовних складових виділення мови ворожнечі.

Для досягнення мети дослідження слід виконати такі **завдання**:

1. Охарактеризувати поняття «мова ворожнечі».
2. Описати контекст появи мови ворожнечі
3. Висвітлити зв'язок вживання мови ворожнечі з іншими суміжними поняттями.
4. Проаналізувати корпус створений на базі лабораторії комп'ютерної лінгвістики

5. Укласти корпус текстів воєнного періоду (березень-квітень 2022 року)

6. Автоматично укласти базу даних текстів воєнного періоду на основі корпусу

7. Проаналізувати корпус текстів воєнного періоду

8. Автоматично витягнути тексти з Телеграм-каналу «BUCHA LIVE» (період вересень 2022 року – травень 2023 року)

9. Вручну вибрати тексти з мовою ворожнечі, нейтральні тексти, які не відносяться до російсько-української війни та додати їх до наявного корпусу

10. Створити програмний засіб, який визначає мову ворожнечі

11. Створити графічний інтерфейс програми автоматичного визначення ворожнечі у текстах українською мовою

РОЗДІЛ 1. СТАНОВЛЕННЯ ПОНЯТТЯ «МОВА ВОРОЖНЕЧІ» В СОЦІОЛОГІЧНОМУ ТА ЛІНГВІСТИЧНОМУ ДИСКУРСАХ

1.1.Експлікація поняття «мова ворожнечі»

1.1.1Визначення мови ворожнечі в лінгвістиці

«Ворожнеча – відносини і дії між ким-небудь, пройняті ненавистю, недоброзичливістю, ворогування.»[15] «Ненависть – почуття великої неприхильності, ворожості до кого-, чого-небудь, нелюбов, ворожість, ворожнеча, ненависть» [14]. Грунтуючись на філологічному тлумаченні наведених понять, можна стверджувати, що їх сполучення «мова ненависті» за своїм змістом виражає дію, яка полягає у висловлюванні думок, пройнятих ненавистю, недоброзичливістю, ворожістю до кого-небудь чи чого-небудь, що сприяє формуванню негативних світоглядних настанов та соціальних настроїв тощо і має негативні наслідки для суспільства (окремої особи чи групи осіб) [О.Л.Львова]

З початком війни українські медіа стали більш пристрасними, іноді з порушенням журналістських стандартів. Цю тенденцію посилили соціально-економічна та політична кризи. Медійники почали використовувати "мову ворожнечі" для підсилення позиції свого видання або для привернення цільової аудиторії. В результаті виникає поділ на "наших" і "чужих", на добрих і поганих, на розумних і менш розумних. Гіперболізація негативних ознак опонентів і вигадання виняткових чеснот представників свого "табору" спричинює розбрат у сприйнятті світу, що створюють споживачі масмедіа. Лінгвістичні засади «мови ворожнечі» в медіа ґрунтуються на чотирьох напрямах мовознавства: прагмалінгвістичному, лінгвокогнітивному, структурно-семантичному і стилістичному на стику з юриспруденцією та психологією.

«Мова ненависті стала однією з технологій і лінгвістичним маркером так званої «гібридної війни», коли образ опонента позбавляється людських рис та

наділяється абсолютно не властивою людині поведінкою, руйнуються способи ідентифікації особистості по відношенню до соціальної групи, які призводять до зміни самоідентифікації. Створюється уявлення про моральну неповноцінність, кримінальність і негативний вплив на суспільство, формуються певні лінгвокультурні, когнітивно-прагматичні установки, спрямовані на очорнення культури й ідеалів учасника протилежного боку конфлікту» [І. Богданова, О.Лептуга, 2020]

1.1.2 Визначення поняття «мова ворожнечі» у соціології

З розвитком суспільства відбувається поділ громадянза різними видами відмінностей – статтю, кольором шкіри, релігійною приналежністю тощо. Так у результаті становлення правил поділу на кращих і гірших, вищих і нижчих і з'явилась дискримінація, ксенофобія та расизм. Одним з інструментів утиску є мова ненависті(мова ворожнечі)

Мова ворожнечі, за визначенням Комітету міністрів Ради Європи, – це «усі форми самовираження, які включають поширення, підбурення, сприяння або виправдання расової ненависті, ксенофобії, антисемітизму чи інших видів ненависті на ґрунті нетерпимості, у тому числі: нетерпимість висловів у формі радикального націоналізму та етноцентризму, дискримінації та ворожості щодо меншин, мігрантів і людей з числа іммігрантів» [26]. Відповідно до цього визначення можна виділити такі групи за тим, на кого спрямована мова ненависті: 1) національна належність; 2) релігійна ідентифікація; 3) сексуальні меншини.

Наразі, групи, на які може бути напрямлена мова ворожнечі, вже не обмежується названими основними сферами, оскільки вона може поширюватись на будь-які висловлення, які мали негативне значення, тобто були спрямовані проти певної особи, соціальної групи чи суспільства. На основі проаналізованої літератури (С.Лихової, Г.Рибальченко, О.Коробкової, О.Гладіліна) можна виділити певні **форми** прояву мови ворожнечі: 1) висловлювання, які можуть бути усними чи письмовими; 2) фотографії, які налаштовують читачів чи глядачів проти певної

групи людей (наприклад, перекреслені фото етнічних чи релігійних груп); 3) відеоматеріали як синтез слів, музики та зображення змінили об’єм інформації, яка сприймається. Тобто, за допомогою відеоматеріалів легше маніпулювати, застосовуючи мову ворожнечі [Лихова, 2013; Коробкова, 2009; Гладилин, 2012]. Характеризуючи мову ворожнечі, варто виділити наступну особливість: вона може ставати засобом спілкування у повсякденні [Бурдье, 1995]. П’єр Бурдьє вводить поняття «габітус» і визначає його як систему «стійких набутих схильностей». Мова ворожнечі трансформується та перетворюється «з навмисної на випадкову і переходить з глобального рівня на індивідуальний» [Бурдье, 1995]. Тобто індивіди як споживачі інформації ненавмисно стають ретрансляторами мови ненависті. Прикладом такого може бути ситуація, коли жителів Західної України називають «западенцями», тим самим ображаючи почуття групи людей за територіальною ознакою, не розуміючи цього. Приклад про «западенців» варто розглянути з точки зору соціального стереотипу, адже стереотип з’являється раніше, ніж з’являється уявлення.

Характеристика мови ворожнечі також включає часову залежність в поєднанні з етноцентризмом, що означає, що різні суспільства на різних етапах розвитку мають відмінні норми та правила. Наприклад, у моніторингу Інституту соціології НАНУ можуть бути використані такі терміни, як "негри" та "цигани", що можна сприйняти як мову ворожнечі. Проте ці слова не можуть бути просто замінені через їх непорівнянність. Вони вживаються в повсякденному житті, і заміна їх може призвести до неправильного розуміння (наприклад, розуміння афроамериканців замість нігерійців). В інших суспільствах, наприклад у США, переважає політика коректності, і терміни для позначення "темношкірих" людей змінилися з "негрів" на "афроамериканців" або "особи з темною шкірою".

Поняття мова ворожнечі вживається у різних сферах та науках. У сфері захисту прав людини мова ворожнечі визначається як «некоректні висловлювання на адресу етнічних, конфесійних чи певних соціальних груп як спільнот і на адресу конкретних людей як представників цих спільнот» [Лихова, 2013]. У

журналістиці мову ненависті висвітлюють з іншого боку та визначають як «ознаку дегуманізації, яка допомагає диференціювати суспільство та ідентифікувати його представників за принципом «ми - вони», «свій - чужий»» [Бойко, 2014]. У психології мову ворожнечі визначають як «інструмент впливу на свідомість аудиторії, формування потрібного владним структурам світогляду» [Попова, 2016].

Мова ворожнечі можна розглядати як соціологічне поняття з кількох причин. По-перше, вона пов'язана з соціальною реальністю та взаємодією людей у спільнотах та суспільствах. По-друге, це соціальна дія, оскільки вона є формою поведінки людини, що регулюється суспільними цінностями і нормами, спрямована на інших та має певну мету. По-третє, вона відображає соціальні відносини між різними соціальними групами у суспільстві. По-четверте, вона є елементом взаємодії соціальних акторів з різними культурними приналежностями. Нарешті, вона є результатом взаємопроникнення культур.

Мова ненависті поширюється на різні соціальні групи всередині суспільства або між різними суспільствами. Вона виступає як засіб передачі соціальних проблем, таких як бездомність, злочинність, примусове переміщення, інвалідність та інші, і може призвести до стигматизації конкретної соціальної групи або суспільства. Соціальними наслідками мови ворожнечі є заглиблення соціальних проблем, присутніх у суспільстві.

В контексті теми, мети та завдань даної роботи термін «мова ворожнечі» вживається у такому значенні: ***мова ворожнечі– це промови, висловлювання, коментарі, які спрямовані на виявлення агресії проти певних осіб, соціальних груп чи суспільств.***

1.2 ЗВ'ЯЗОК МОВИ ВОРОЖНЕЧІ З ІНТОЛЕРАНТНІСТЮ, КСЕНОФОБІЄЮ, СОЦІАЛЬНОЮ НАПРУЖЕНІСТЮ, СОЦІАЛЬНОЮ АГРЕСІЄЮ ТА ПРОПАГАНДОЮ

Мова як засіб комунікації є також транслятором та засобом реалізації мови ворожнечі. Політолог та історик Бенедикт Андерсон поділяє мови на вищі та нижчі, сакральні та «плебейські» (розмовні) відповідно [Андерсон, 2001]. Загалом можна виділити два види об'єднання за та проти чогось або когось. Мову ворожнечі за цією класифікацією можна віднести до типу об'єднання проти чогось або когось, тому що вона спрямована на утиски певної соціальної групи за певною ознакою.

Мова ворожнечі розглядається як застосування некоректних висловів щодо представників різних груп, наприклад, мова ненависті щодо релігійних чи конфесійних груп. Найчастіше ці некоректні виловлювання використовуються журналістами. Приблизно 40-61% всієї мови ворожнечі є продуктом навмисного чи випадкового висловлювання журналістів [Мельников, 2006].

Мова ворожнечі – це прояв глобальних проблем певного суспільства чи суспільств. Вона є ретранслятором таких проблем, як ксенофобія, інтолерантність, соціальна напруженість, соціальна агресія та пропаганда.

Ксенофобія, за визначенням Лариси Апанасюк, є страх інакшості, нетерпимість до всього чужого і може розглядатися як підвид соціальних фобій [Апанасюк, 2013], які є індикаторами соціокультурних конфліктів.

Мова ворожнечі є вагомим чинником, який сприяє розвитку ксенофобії. Зараз у соцмережах, медіа можна побачити критичні зображення і прочитати багато критичних матеріалі стосовно російської армії, громадян і взагалі всього, що відноситься до росії. Зараз ще залишається нетерпимість і до інших груп населення, не тільки до росіян і білорусів. Багато українців зараз погано ставляться до людей, які переїхали закордон. Київський міжнародний інститут соціології опублікував статистичні дані, за якими 36% українців не хочуть жити поряд з українцями, які є росіянами за походженням [КМІС]

При визначенні поняття «інтолерантність» можна відштовхнутись від визначення толерантності за Г.К.Селевко [Селевко, 2006]. Відповідно до цього *інтолерантність – це неповага прав людини та плюралізму, в тому числі культурного плюралізму.* Інтолернатність заважає людям вступати в діалог та

ускладнює процес комунікації, тобто, за цією ознакою, мова ворожнечі є ретранслятором та засобом прояву інтолерантності.

Є.Ю.Кольцова та Є.Є.Таратута провели дослідження, згідно з яким визначили два типи інтолерантності: допустиму та недопустиму [Кольцова, 2003]. Самозахист є прикладом допустимої інтолерантності, коли він використовується для захисту від агресії. Однак, якщо проаналізувати джерела походження та розповсюдження мови ворожнечі, то можна зрозуміти, що політика та ЗМІ, які допускають інтолерантність у соціальних взаємодіях, є основними джерелами поширення цієї мови. Екстремізм, радикалізм, фундаменталізм, фанатизм та інші є найбільш розповсюдженими проявами інтолерантності, де мова ворожнечі є засобом здійснення агресії. Наприклад, вживання таких слів, як «ватніки», «колоради», «лугандони», «донбасівці» і т.д. є прикладами такої мови. Соціальна напруженість визначається як «психічний стан соціуму, що виникає у відповідь на екстремальні ситуації» [Бєлай, 2012]. Саме мова ворожнечі може стати причиною такого стану суспільства, де переважає незадоволеність та агресивність громадян, соціальна та економічна нерівність. Соціальна агресія є однією з найгостріших проблем сучасності, яка з'являється у відповідь на певні соціальні обставини та деструктивність суспільства. Мову ворожнечі можна віднести до вербальної агресії, яка проявляється у словесній формі. Дійсно, мова ненависті може мати негативні наслідки і приводити до різних форм агресії. Це може бути як фізична, так і психологічна агресія, яка може призвести до шкоди для індивіда або групи. Важливим є і те, що мова ненависті посилює стереотипи та дискримінацію, що теж негативно впливає на суспільство в цілому.

Є ще одне явище – пропаганда, яке використовується з метою впливу на думки, переконання та поведінку людей і для опису різних методів та технік, в різних контекстах, включаючи політику, рекламу, релігію тощо. На думку В.П. Таркіна, пропаганда – це «метод інформаційно-психологічної війни, що характеризується комплексом дій – переконання, розповсюдження ідей, думок, доктрин, ідеології, чуток та іншої інформації (чи дезінформації), що впливають

на цільову авдиторію із метою її примушування до певного способу поведінки й мислення» [Таркін].

Історично пропаганда була пов'язана з політичними режимами та військовими конфліктами і в сучасному світі використовується з різною метою: від продажу товарів до впливу на поведінку людей в соціальних мережах. Важливо бути свідомими того, що пропаганда використовується з метою маніпуляції та формування негативних стереотипів, тому важливо бути критичним до інформації, яку ми отримуємо, та перевіряти її джерела.

1.3Мова ворожнечі: українські реалії

Сучасний інформаційний простір України є середовищем формування нових значень слів та поведінки українців. Розвиток інформаційно-комунікативних технологій та засобів зв'язку, поєднаний зі зростанням внутрішньополітичних процесів та агресивною політикою Росії, стали потужними генераторами трансформацій комунікаційних форм самовираження громадян. Закони вживання слів змінюються разом із їхніми значеннями. Нейтральні слова можуть набути зневажливого сенсу та викликати цілеспрямовану емоційну реакцію. Механізми створення нових значень можна назвати семіотичними. Фізичні об'єкти можуть зберігати своє значення, але отримувати нові позначення, які не просто негативні, а доволі точно програмують негативну реакцію аудиторії, надаючи їй чіткі мотивації для дій.

1.3.1 Російсько-українська інформаційна війна

Російсько-українська інформаційна війна є прикладом когнітивної війни, де боротьба ведеться і за програмування мислення як своєї сторони, так і протилежної в потрібному напрямі, і за перепрограмування, в межах якого повністю міняються місцями поняття «друг» і «ворог».[Почепцов] Більшість

людей здатні запам'ятовувати погані події і негативні асоціації набагато краще і яскравіше. Це пояснюється їхнім інстинктом самозбереження, коли вони перестраховуються та перебільшують події в разі потенційної майбутньої небезпеки. У таких випадках мова може перетворитися на інструмент агресивного захисту, а норми мови стають етичними, формуючи масову несвідому та пасивну свідомість, яка не здатна до об'єктивного критичного судження. Мова ворожнечі може навіть привести до фізичної форми вияву насильства у вчинках. Саме через це сучасне інформаційне середовище нагадує поле боротьби між "своїми" і "чужими", як у традиційних медіа, так і в нових інтерактивних медіа.

Дослідниця Ісакова дає перелік слів, що мають ознаки «мови ворожнечі», присвячених війні *аквафреш, бандерлоги, беркулята, боняри, ванговать, вангую, вата, ватники, вишеватники, гейропа, гиркнувся, дачинг, диванні війска, диванна сотня, домбанатя, домбасс, єрмолки, каламойша, колоради, кровополитика, кримнаш, кримнашки, ленінопад, луганда, лугандон, няшмаш, руіна, майдануті, майданята, мотороли, нанор¤оссия, намкриш, новопендосия, отдонбасить, псакнуть, псакинг, правосеки, підораша, реконструктори, решеткін, свидомит, укробойці, укроп, укропія, укри, хунта, фашизм* тощо. Різноманітні метафори на позначення ворога – необхідний компонент будь- якого збройного конфлікту.[Ісакова]

1.3.2. Меми як спосіб ведення інформаційної війни

Український інформаційний простір містить багато ворожих висловлювань, серед яких особливе місце займає мем – колективне несвідоме, що набуває словесної чи знакової форми. Успіх мему полягає в тому, що він привертає увагу до тем, які раніше були пригнічені або приховані, і пробиває бар'єри в свідомості багатьох людей. Мем не може просто викласти інформацію,

його справжній зміст залишається прихованим і дозволяє відвернути критичне мислення та проникнути у підсвідомість людей за допомогою опосередкованих асоціацій. [Ісакова]

З погляду психології, меми є своєрідними культурними позначками та відображають поляризацію між групами: «Принижуючи суперника, ми вивищуємося над ним, він зменшується в розмірі, стає слабким, а ми відчуваємо владу і силу» [Карп'юк]. Дегуманізація противника є одним з методів психологічної війни, коли намагаються зменшити значущість іншої людини, зробити її менш людяною і, отже, менш важливою. Це впливає на наше мислення, сприйняття світу, нашу поведінку і ставлення до інших. Тому важливо бути обережним зі словами і мемами, які можуть викликати дегуманізувальний ефект і нашкодити нашим стосункам з іншими людьми. Іронічний мем «кримнаш» використовується на адресу росіян, які експресивно виражають захоплення анексією Криму. «Майданутий» – зневажлива назва патріотично налаштованих українців, які наголошують на своїй відмінності від росіян і готові емоційно відстоювати свою проукраїнську позицію. [Карп'юк]

Станом на 2023 рік українці дуже багато приділяють уваги чинному президенту України – Володимиру Зеленському. В його бік є дуже багато негативних висловлювань та хейту. Не залишається поза увагою і колишній президент – Петро Порошенко. Досліджуючи медіа дописи, газети, журнали, коментарі у соціальних мережах, видно, що українці поділилися на два табори: одних називають *порохоботи,* що також відноситься до мови ворожнечі, а інших – *зелебобіки* та ін.

1.4 Класифікація мовних засобів

У мові ворожнечі можна виділити кілька принципів класифікації мовних засобів, які використовуються для вираження ворожнечі між мовами. Найбільш поширеними принципами класифікації є наступні:

Етнічний принцип: ворожнеча може виражатися через національні прогалини в мовленні, а також через використання слів і виразів, які є ознаками національного самовизначення.

Соціальний принцип: ворожнеча може виражатися через різницю в соціальному статусі мовців, яка може виявлятися у виборі слів, тоні мовлення та іншій лінгвістичній поведінці.

Культурний принцип: ворожнеча може виражатися через різницю в культурних нормах та цінностях, які можуть мати вплив на мовлення та вибір мовних засобів.

Історичний принцип: ворожнеча може виражатися через історичні події та сприйняття історичних подій в різних країнах та культурах.

Політичний принцип: ворожнеча може виражатися через політичні конфлікти та ідеологічні розбіжності між країнами та культурами.

Ці принципи можуть використовуватися окремо або в поєднанні один з одним для класифікації мовних засобів у мові ворожнечі.

Класифікацій мовних засобів є багато. Оглянемо, ті які найбільш вживані у суспільстві, класифікація визначена Богдановою та Лептугою [23]
Ця класифікація є загальною і зрозумілою. У класифікації таксонів, написаної мною, я розширила список, через те, що не всі слова підходять до загальної класифікації. Але для базового аналізу вона є досить простою і зрозумілою.

1. Заклики до насильства. Це дійсно можливе під час гострих фаз протистояння між різними групами людей. Прямі заклики до насильства можуть викликати агресивні реакції в середині спільноти та сприяти посиленню напруження. Це небезпечно, оскільки може спричинити фізичну та психологічну шкоду.

У той же час, загальні гасла або приховану дискримінацію складніше помітити, але вони так само можуть викликати негативні наслідки, напр., сприяти формуванню стереотипів у відношенні до різних груп людей, що спричинює дискримінацію та несправедливість у суспільстві.

*«Невідомі вандали чорною фарбою замалювали очі скульптури, також на монументі з'явилися нацистські символи і написи «**Хайль Гітлер**» і «**Смерть жидам**».»* Використання прецедентних феноменів, що спільні для автора і читача, створює враження взаєморозуміння у думках. Цей психолінгвістичний підхід досить ефективний у пропагандистських матеріалах, які намагаються видати себе за журналістські. Проте, в таких матеріалах можуть міститися непрямі, але нав'язливі висловлення, які можуть спонукати до припущень про бажання насильства проти певної групи.

2. Створення негативного образу етнічної, релігійної чи певної соціальної групи. Це поширений прийом заграваання перед аудиторією видання. У цих випадках позиція журналіста теж простежується через психолігвістичні ознаки тексту. *Якщо минулого року Літаючий Макаронний Монстр приєднався до ходи попів **москальського патріархату** лише за допомогою фотошопу, — цього року вже ікона **пастафаріанського божества** приєдналася до ходи. **Захід пройшов у теплій, дружній атмосфері**, а пастафаріанців супроводжували **приємні хлопці з міліції**», — написала запорожчанка Арина Мосягіна, яка при цьому була присутня* (061.ua, 05.03.2017). Так, цитати з такою лексикою мають явно негативну конотацію та створюють враження, що автор статті спрямовує свої негативні емоції на групу людей згідно з їхнім етнічним або релігійним походженням. Використання такої «мови ворожнечі» стимулює появу негативного стереотипу про певної релігійної групи та спричинює її дискримінацію. Крім того, іронія в реченні може бути сприйнята як сарказм або насмішка з певної групи і викликати негативні емоції та подалішу дискримінацію

3. Виправдання історичних випадків насильства. *«–А він таких не любить, бо наш президент хоче заборонити російську мову». Коли Вікторія написала, що політикою не цікавиться взагалі, то це не подіяло і хлопець продовжував писати, що **«бандерівці» винищували народи разом із фашистами**. Ректор Закарпатського угорського університету в Берегово Ілдико Орос на мітингу в Будапешті порівняла Україну з фашистами. На її думку, сьогоднішні події в Україні «**нагадують події минулого-позаминулого століть з переслідуванням***

євреїв» [Depo Харків», 26.10.2018]. "Мова ворожнечі" використовується для створення негативного образу українського народу, зокрема його влади, як людей, що відсталі розумово, без моральних принципів і підтримують нацистські або фашистські режими. У прикладі використана жорстка лексика з негативним забарвленням та вислови, що мають негативний відтінок.

4. Твердження про неповноцінність певної соціальної чи національної групи. *Побитий незрячий біженець скаржиться на бездіяльність правоохоронців. Незрячий переселенець-інвалід з окупованого Донбасу Сергій Пономаренко запевняє, що його протиправно виселили із модульного містечка у Нікополі, при цьому жорстоко побили. Поліція ж відмовилась відкрити кримінальну справу* [«Історична правда», 24.10.2018]. Використання нетолерантного слова «інвалід» замість «люди з особливими потребами», чи «люди з обмеженою мобільністю», чи «люди, які мають інвалідність», а також юридично й етично неграмотне використання слова «біженець» замість «вимушений переселенець» чи «внутрішньо переміщена особа» створюють відчуття навішеного ярлика, мовляв, таку людину інакше й не можна назвати. Таке принизливо формалізоване ставлення викликає у читачів, які не належать до вказаних груп, внутрішній шовінізм. Це відчуття може загостритися аж до неадекватного сприйняття іншої нації чи групи в цілому.

5. Твердження про історичні злочини тієї чи тієї етнічної або релігійної групи як такої. Російсько-українська війна спровокувала глибше вивчення спільної історії держав. Відновлення національної пам'яті, аналіз неоднозначних рішень тогочасних і сучасних політиків, звичайно, позначилися на картині світу журналістів і реципієнтів. Чого тільки вартий сформований у медіадискурсі концепт «рускій мір». Наведемо кілька прикладів. «*Не Росія — тільки Московія! Чого б не узяти й не обізвати росіянина* ***москалем****? Скажете: некоректно… Та ж ні, переконують експерти, історично виправдано і безальтернативно! Етнонім «росіяни»* ***ницо вкрали*** *у нас, українців, ще 300 років тому. З нього донині* ***злочинно*** *користає* ***одвічний ворог*** *Української держави —* ***Московія****. І саме на ньому заснована* ***путінська міфологія «руського міра»****.* У прикладах

проголошене агресивне ставлення до іншої країни, що виявляється у використанні згрубілої лексики, діалектизмів, негативних оцінних висловів. З точки зору психолінгвістики — чітко простежується позиція як журналіста, так і цільової аудиторії видання.

6. Твердження про кримінальність тієї чи тієї етнічної або релігійної групи. Стереотипи, що частково є підґрунтям для виникнення прецедентних феноменів і концептів, також провокують «мову ворожнечі» в українських медіа. *У Володимирі* ***цигани обчистили будинок*** *пенсіонера. У Володимирі-Волинському у 80-річного пенсіонера* ***цигани*** *вкрали 14 тисяч гривень. Пенсіонер розповів: ще 10-го лютого цього року, близько 15 години, до нього прийшли три* ***циганки****. Вони хотіли подивитися будинок, який продає чоловік. Під час огляду будинку, як припускає заявник, зникли його кошти. Пенсіонер вважає, що заощадження в сумі 14 тисяч гривень були викрадені* ***циганами*** [Волинь 24, 16.02.2018]. Передусім є порушення в цитуванні — наявні нетолерантні вислови, а також у тенденційному доборі фактів — наводяться лише негативні риси ромів, що створює їхній негативний образ.

7. Звинувачення в негативному впливі тієї чи тієї етнічної, релігійної або певної соціальної групи на суспільство, державу. *У Луцьк приїде* ***циганська відьма-екстрасенс****. До Луцька приїде учасниця шоу «Битва екстрасенсів»,* ***циганська відьма*** *Рубіна Цибульська. Про це повідомляє Таблоїд Волині з посиланням на Інстаграм Рубіна Цибульська родом з Росії. Виховує з чоловіком 6-річного сина, який,* ***як і всі цигани****, володіє даром* [«Волинь 24», 09.02.2018]. Ознаки «мови ворожнечі»: використання ксенофобних назв національностей, образливих слів, проголошення агресивного ставлення до іншої країни. Психолінгвістична подача тексту — агресивно негативна з елементами знецінення.

8. Згадка певної групи або її представників у принизливому або образливому контексті: *У закинутому дитсадочку —* ***бомжі, наркомани*** *і гори сміття. Закинуте приміщення дитсадку, що на вул. Кривоноса, 7 А, облюбували* ***бомжі і наркомани****. Всюди — гори сміття та сумнівні компанії* [«20 хвилин»,

09.11.2015]. Порушення: крім некоректного використання слова «бомж», бачимо нетолерантне вживання слова «наркомани» (правильно — «наркотично залежні» або «люди з наркотичною залежністю»). На відміну від попереднього прикладу «мова ворожнечі» тут чітко простежується. Натомість ефект багатослів'я, коли негативний сенс фрази приховується великим за обсягом висловлюванням, робить складним ідентифікацію одиниць «мови ворожнечі».

1.5 Корпуси мови ворожнечі

Існує декілька корпусів мови ворожнечі, що можуть використовуватись для дослідження різних аспектів ворожнечі в мові. Ось кілька з них:

Корпус мови ворожнечі автоматизованого діалогу (Conversation-based Hostility Corpus) - це корпус, який містить транскрипти діалогів, в яких є елементи ворожнечі. Корпус складається з більше 500 діалогів, відібраних зі збірників даних автоматизованого діалогу.

Корпус мови ворожнечі в соціальних медіа (Social Media Hostility Corpus) - це корпус, який містить повідомлення зі соціальних мереж, що містять елементи ворожнечі. Корпус складається з більше 10 000 повідомлень, відібраних з різних платформ соціальних медіа.

Корпус мови ворожнечі в політичному дискурсі (Political Hostility Corpus) - це корпус, який містить тексти політичних мовців та коментаторів, що містять елементи ворожнечі. Корпус складається з більше 1 500 текстів, відібраних зі збірників даних політичного дискурсу.

Корпус мови ворожнечі в міжнародних конфліктах (International Conflict Hostility Corpus) - це корпус, який містить тексти пов'язані з міжнародними конфліктами та воєнними діями, що містять елементи ворожнечі. Корпус складається з більше 5 000 текстів, відібраних з різних джерел, таких як засоби масової інформації та офіційні заяви урядів.

1.6 Висновок до Розділ 1

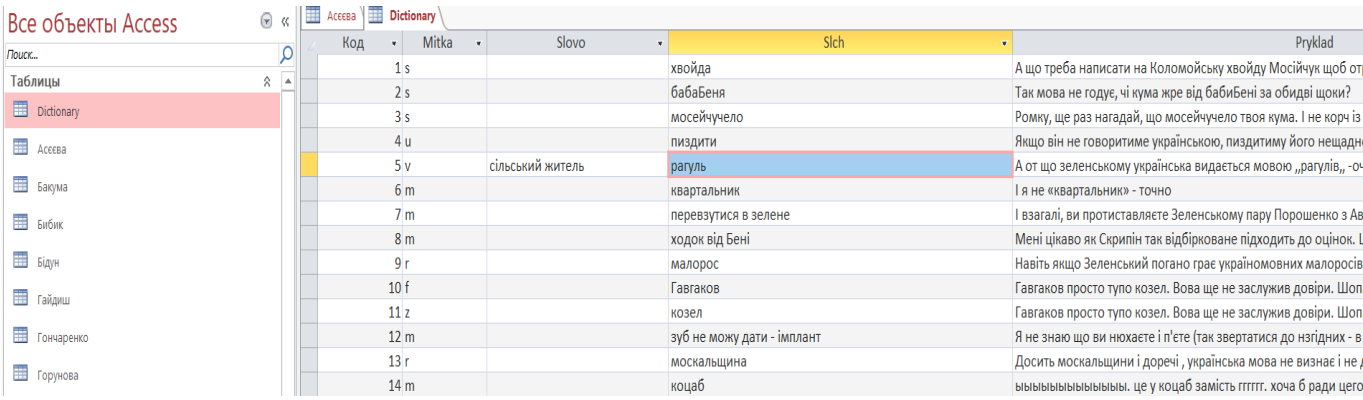
Мова ворожнечі відрізняється від загальновживаної мови за допомогою специфічної лексики, фразеології, експресивності виразів та особливого

використання словотвірних засобів. Сутність мови ворожнечі сильно залежить від соціальних стереотипів та дискримінації. Негативні оцінки, що виявляються в мові ворожнечі, походять від світогляду, культурно-історичного досвіду та способу інтерпретації поточних подій проросійськими медіа. Її використання є одним із напрямів планомірної роботи з поширення негативного ставлення до України, українців, законної української влади, способу життя українців, культури, мови, історії України.

РОЗДІЛ 2. РОБОТА З КОРПУСОМ

Корпус: movavorozhnechi.mdb **див. Додаток 6**

Цей корпус я отримала на другому курсі навчання. Моя робота полягала у тому, що я аналізувала наданий мені корпус, написала та додала найменування таксонів мови ворожнечі



Фрагмент бази даних(корпусу)

2.1 Класифікація таксонів мови ворожнечі(мітки до корпуса)

Таксон — група у класифікації, яка складається з дискретних об'єктів, об'єднаних на підставі спільних властивостей і ознак

1. **s** - Сексизми – до них відносять мізогічні сексизми (спрямовані проти жінок), мізандризми (спрямовані проти чоловіків), а також вияви гомофобії.

(«*Порох здохни* ***підаре***», «*Халявні гроші* ***шкурам*** *у вигляді аліментів подавай.*»)

2. **r -** Расизми – «вияви національного шовінізму. Лексико-фразеологічна система української мови має низку усталених ксенофобських найменувань: *хохол, москаль, кацап, жид, негритос* та інше («***Угорський йобік*** *сказав а завтра українські йобіки будуть кричати ми угорці, точно так як кричали Донбас расєя.*»)

3. **e** -Ейджизми – упередження до людей, пов'язані з їх належністю до вікової категорії.(« *Він же ж не лох якийсь і не* ***бабулєта*** *з кравчучкою, щоб так мучитись*»)

4. **l** - Лукізми є «виявом упередженого ставлення до людей, чиї зовнішні дані відрізняються від сучасних уявлень про красу» («***Дрищ****, не переводь тему у ситуацію невирішеності після вашого проседу.*»)

5. **b** - Ейблізми – мовне вираження упередженого ставлення до людей з психічними хворобами та обмеженими фізичними можливостями. («*афганський карлик*», «*дальтонік*», «*від стресу зійшов з розуму, а значить- всі його призначення підлягають відміні, як прийняті людиною* ***психічно хворою***.»)

6. **m** - Менталізми – негативні висловлювання про людей, які мають невисокі, на погляд мовця, розумові здібності та рівень грамотності. («*Навіть кіт в шоці за кого ви тут понаголусовували,* ***ідіоти***»)

7. **k** – мовне вираження негативного ставлення до людей через їх професійну приналежність або майновий стан. («*комік*», «*мєнт*», «***Баризі*** *не було коли відпочивати, ночами не спав, все думав, як би ще хоч гривню з кожного українця здерти!*», «*Тому що їх* ***комік*** *сам ніколи не прийме правильно рішення*.»)

8. **v** -Локалізми є виявом упередженого ставлення до мешканців певної території. Наприклад, в Україні, залишається актуальним протиставлення села та міста.(«*западенці*», «*Сироїд тупорила* ***селючка***»)

9. **n** -Матримоніалізми відображають упередженість до людей через їх сімейний статус – до родин з одним з батьків, самотніх людей тощо.(« *Це ваше право на владу…. Наче сходка* ***бомжів****…. Чесне слово……*»)

10. **z** – зоологізми – порівняння з тваринами («*В нашій країні є їоч хтось окрім цих* ***свиней****?», «****Криси*** *бояться без корма лишитись*»)

11. **t** – порівняння з міфічними істотами («*чувак звинувачує колишню дружину в тому що вона* ***відьма***»)

12. **u** – за назвами частин тіла і фізіологічні процеси(«*світ заповнила реклама. і ти в ній* ***гамно***»)

13. **p** – порівняння з рослинами(«*На колінах хай ця* ***зелена пліснява*** *вибачається !!!*»)

14. **d**- спонукання до дій(«*йди нафіг*», «*Петя знов танки ,знов Раша нападає?!* ***Вали вже за грати****,чмо*»)

15. **c**- сарказм, завуальоване висловлювання(«*соблюдатєлі традіций*», «*Унікальна ви наша*.»)

16. **x**- порівняння з казковими та мультигероями(«*тролі*», «*Треба братися за розум, всі знають, що цей* ***гном*** *нічого не вирішить!*», «*У вас пластинку заїло з вашими тупими новинами,* ***зомбі*** *тупорилі!!*»)

17. **g**-задовбувати, роздратовувати(можна віднести і до емоцій(w))(« ***Задовбав*** *своїми роликами!!!», «Ви* ***задрали*** *називати всіх моральних уйобків даунами!»*)

18. **j**- упередженне ставлення до релігії(«*Мабуть ці працівники суду слуги евреев* ***(сатаністів)***», «*Це упороті* ***сектанти*** *і переконувати їх немає жодного сенсу.*»)

19. **w**- емоції(«*сьогоднішня антитерористична операція - це якись* ***йобаний стид****, вибачте на слові.*», «*Чмо, зрадник держави,* ***ганьба*** *цілої країни.*»)

20. **y**- порівняння з предметами («***Пустушка*** *одним словом*», «*Як йому беня скаже то він і робить. Які ви зебіли, все ж таки зебіли. Цей клоун* ***лялька***»)

21.**i** – мовне вираження негативного ставлення про людей з залежностями(«*Може варто Макрона попросити, щоби притримав того* ***торчка****, поки Порох під'їде?*», «***Алкаші*** *Донбасу здали пляшки й придбали собі буки та гради.»*)

22. **q**- фізичні дії(заклики до насильства) («*Того ти не знаєш і ніколи знати не будеш???вас заселили замість мільйонів українців яких комунисти*

знищилиии*!!!!!*», «*А кого турбує якоби «утиснение русского язика»(хоча це не правда),прошу* ***валити*** *в рашку!*»)

23. **f-** політика(«*Зборище* ***зебілів****, скоро будете вити дивлячись на своє дзеркальне відображення.*», «*Зробіть МРТ... можливо у вас пухлина..., бо іншого пояснення вашого захисту* ***Пєті****, я не знаходжу...*»)

2.2 Аналіз корпусу, створеного в лабораторії комп'ютерної лінгвістики

Цей корпус я отримала на другому курсі навчання. Моя робота полягала у тому, що я аналізувала наданий мені корпус, написала та додала найменування таксонів до класифікації мови ворожнечі (пункт 2.1).

Корпус, проаналізований мною в рамках дипломного проєкту, нараховує близько 3700 слів/словосполучень. База даних розміщена в програмному середовищі ACCESS, таблиця складається з 5 колонок в "Dictionary"- "Код(№)", "Мітка", "Слово", "Словосполучення", "Приклад(контекст)".

Корпус містить не лишу мову ворожнечі написаною українською мовою, а й англійською.

У корпусі багато висловлювань емоційних, політичних(це в основному про чинного президента, та його команди і його прихильників) і про Порошенка, його прихильників. Два табори, про які писалося вище, негативно висловлюються один проти одного, що викликає великий хаос в коментарях.

Також не мало расизму та сексизму.

І далі показано покриття таксонами мови ворожнечі тексту

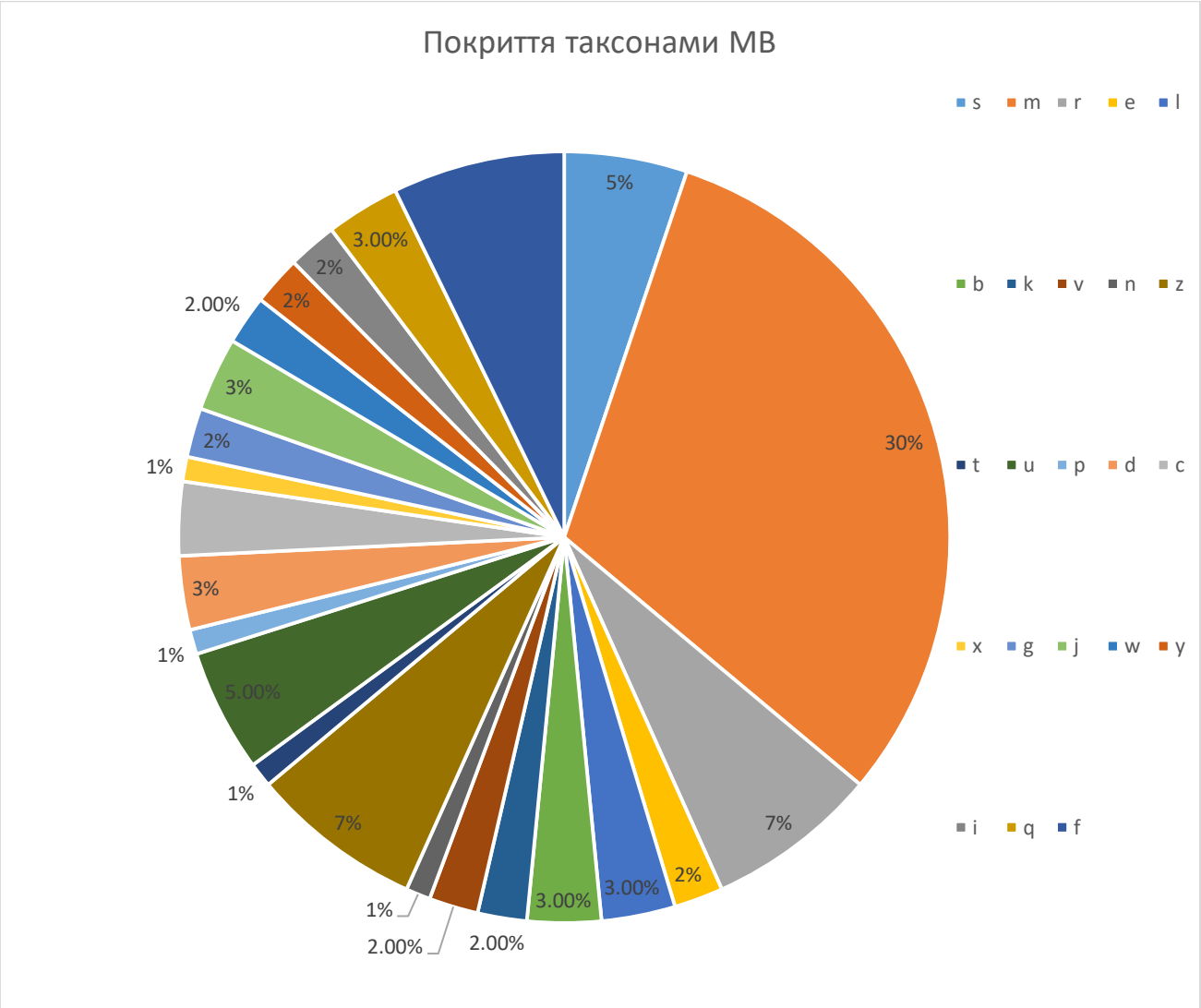


Рис.1 Покриття таксонами мови ворожнечі для першого корпусу текстів

Отже, за результатами бачимо, що найбільше слововживань відноситься до менталізмів.

«Вони хочуть зруйнувати нашу Одесу, але побачать тільки дно чорного моря.»

*«Бо вони – **дно»***

*«Пам'ятаєте я писав, що ці **дибіли** в Білорусі відправляли те, що накрали в наших містах і цим самим задокументували свої злочини»*

***«покидьок** суспільства заявив, що трупи в Бучі розкидала розвідка МІ6.»*

Негативні висловлювання про людей з невисоким рівнем розумових здібностей та грамотності можуть мати серйозний вплив на їх майбутнє. Ось деякі можливі наслідки таких висловлювань:

Зниження самооцінки: Негативні коментарі можуть підірвати впевненість у собі та самооцінку людини. Вони можуть відчувати себе менш цінними або нездатними досягати успіху, що може обмежити їх потенціал.

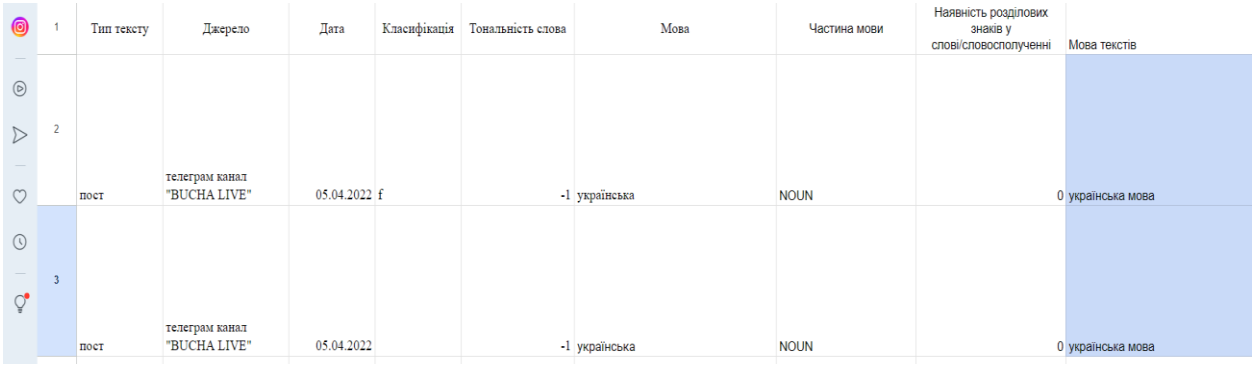
Відсутність мотивації: Коли людина постійно почувається критикованою або зневаженою, це може позбавити її мотивації розвиватися та покращувати свої здібності. Вони можуть втратити інтерес до навчання та розвитку, що обмежує їх можливості.

Соціальна відокремленість: Негативні висловлювання можуть призвести до соціальної відокремленості і викликати відчуття відчуження від оточуючих людей. Це може підірвати їх соціальну інтеграцію та можливості для побудови взаємовідносин.

2.3 Побудова корпусу текстів воєнного періоду

Для створення корпусу було обрано один телеграм канал – “BUCHA LIVE” і вручну зібрано тексти з наявністю мови ворожнечі. Вручну я обробила 250 текстів і зафіксувала 280 прикладів мови ворожнечі, так як в одному тексті могло бути по декілька слововживань мови ворожнечі. Тексти від 1 до 10 речень. Є пости де текст містить одне слово, яке і є мовою ворожнечі. На основі зафіксованих слів ворожнечі, я створила список новоутворених слів, які виникли під час повномасштабного вторгнення у 2022 році.

	A	B	C	D	E	F	G	H	I
	2	кат	Російський окупант із Владивостока 20-річний Михайло Ткач, який вбивав українців у Бучі, загрожує повернутися до нашої країни та продовжити різати голови мирних громадян. Українці знайшли сторінку Ткача у соцмережах та почали писати йому, що він кат та окупант. У відповідях росіянин не тільки не заперечує своїх злочинів, а й відкрито заявляє, що повторить свої звірства.	пост	телеграм канал "BUCHA LIVE"	05.04.2022		-1	українська
	3	тварюка	Смачного тварюко	коментар	телеграм канал "BUCHA LIVE"	27.02.2022	z	-1	українська
	4	свинособака	Свинособаки застосовують проти мирного населення гранати і вогнепальну зброю. Ці російські іроди не достойні повернення на росію навіть грузом-200	пост	телеграм канал "BUCHA LIVE"	02.03.2022	z	-1	українська
	5	ірод	Свинособаки застосовують проти мирного населення гранати і вогнепальну зброю. Ці російські іроди не достойні повернення на росію навіть грузом-200	пост	телеграм канал "BUCHA LIVE"	02.03.2022	l	-1	українська
	6	москаль	Здається українці запустили челлендж «віджми БТР у москаля»	коментар	телеграм канал "BUCHA LIVE"	28. 02.2022	r	1	українська
	7	руський мір	Охтирка після спроби встановлення "руського міра". РОСІЯ – ВБИВЦЯ	коментар	телеграм канал "BUCHA LIVE"	28.02.2022	c	-1/0	українська
	8	русня	Русня не пропускає накормить і напоїть коней, погрожують зброєю	коментар	телеграм канал "BUCHA LIVE"	28.02.2022	r	0	українська
	9	Могуча російська армія	"Могуча російська армія" веде бій із пам'ятником афганцям в Бучі	коментар під відео	телеграм канал "BUCHA LIVE"	28.02.2022	c	1	українська



	1	Тип тексту	Джерело	Дата	Класифікація	Тональність слова	Мова	Частина мови	Наявність розділових знаків у слові/словосполученні	Мова текстів
	2	пост	телеграм канал "BUCHA LIVE"	05.04.2022	f	-1	українська	NOUN	0	українська мова
	3	пост	телеграм канал "BUCHA LIVE"	05.04.2022		-1	українська	NOUN	0	українська мова

Фрагменти корпусу текстів воєнного періоду

Повну версію корпусу див. Додаток 4
https://docs.google.com/spreadsheets/d/1kHGT_CgxJ3ZwNNPahntfDhS0rA2AHfMJSGmExNr2riM/edit#gid=0

2.4 Автоматичне укладання бази даних на основі корпусу текстів воєнного періоду

Дивитися **Додатки(Додаток 1, Додаток 3, Додаток 4)**

Було створено базу даних на основі зібраного і проаналізованого корпусу текстів воєнного періоду. Було використано SQlite- система управління базами даних; морфологічний аналізатор -pymorphy2.

База даних має таку структуру: 6 таблиць – «Bad words», «Classifications», «Contex_Types», «General Union», «Languages», «Sources»

Перша таблиця називається "Bad_Words" і містить дві колонки: "Word_ID" - це цілочисельний первинний ключ, а "Word" - це текстова колонка. Вона призначена для зберігання поганих слів або словників для фільтрації.

Друга таблиця називається "Context_Types" і містить дві колонки: "Type_ID" - це цілочисельний первинний ключ, а "C_Type" - це текстова колонка. Вона призначена для зберігання типів контекстів, які можуть бути пов'язані з текстом.

Третя таблиця називається "Sources" і містить дві колонки: "Source_ID" - це цілочисельний первинний ключ, а "C_Source" - це текстова колонка. Вона призначена для зберігання джерел, з яких було отримано текст.

Четверта таблиця називається "Classifications" і містить дві колонки: "Classification_ID" - це цілочисельний первинний ключ, а "Classification" - це текстова колонка. Вона призначена для зберігання класифікацій або тегів, які можуть бути пов'язані з текстом.

П'ята таблиця називається "Languages" і містить дві колонки: "Lang_ID" - це цілочисельний первинний ключ, а "Lang" - це текстова колонка. Вона призначена для зберігання мов, на яких написано текст.

Шоста табиця – “General _Union”, яка видає дані всіх таблиць

База даних є важливим елементом для машинного навчання, оскільки вона містить набір даних, які можуть бути використані для тренування та валідації моделей машинного навчання.

Основна користь баз даних полягає в тому, що вони дозволяють зберігати великі обсяги даних у структурованому форматі. Це дозволяє легко виконувати операції пошуку, фільтрації, сортування та обробки даних.

Також база даних може містити дані, які можуть бути використані для виконання різних завдань машинного навчання, таких як класифікація, кластеризація, регресія та інші. Крім того, бази даних дозволяють зберігати метадані про дані, такі як їхній тип, довжина, розмір та інші властивості, що можуть бути важливі для ефективного використання цих даних в моделях машинного навчання.

Іншою користю баз даних є їхня можливість зберігати дані в реальному часі, що може бути корисним для багатьох додатків машинного навчання, які потребують постійного оновлення даних.

2.5 Аналіз корпусу текстів воєнного періоду

Корпус містить 280 слововживань. Має 12 колонок: “ID”, “Слово/Словосполучення”, “Контекст”, “Тип тексту”(коментар або пост)

“Джерело”- це є один Telegram-канал «BUCHA LIVE» , “Дата”, “Класифікація”- таксони, “Тональність”(Текст, який має чітко виражену ворожнечу має позначку «-1», текст, який не несе ворожнечі у бік когось або чогось, позначається «1», нейтральний текст, який не містить інформації російсько-української тематики, позначається «0»), “Мова слова”, “Частина мови”, “Наявність розділових знаків у слові/ словосполученні” (при створенні програмного засобу треба робити токенізацію і маркування наявності розділових знаків полегшить роботу), “Мова тексту”. Класифікацію використано з поданого вище списку. Корпус містить не лише мову ворожнечі, вживаною українською мовою і суржиком. Наявні слова в яких голосні літери зашифровані знаками: “*”, “#”, “…” “”. Ці слова зашифровують для того, щоб не заблокували користувача і щоб машина не побачила мови ворожнечі у цих висловлюваннях. У корпусі багато емоційних, політичних, расистських висловлювань. Майже вся лексика відноситься до російсько-української війни та російської нації.

Корпус проаналізований за такими критеріями:

-частотність таксонів

Див. Додаток 8

- частиномовна характеристика лексем МВ

NOUN	195	Ці російські **іроди** не достойні повернення на росію навіть грузом-200
ADJ+NOUN	48	Охтирка після спроби встановлення "**руського міра**"

ADJ	11	Вони що **тупі**?
PART	7	Знову **бл@ь** самі себе бомбим
VERB	7	**Ідіть** разом зі своїм кораблем **н@й**!
NOUN+NOUN	4	Моторошне відео з наслідками злочинів фашистів рф.
NOUN+PREP+NOUN	2	гівно в голові
NOUN+ADJ	2	
ADJ+ADJ+NOUN	1	"**Могуча російська армія**" веде бій із пам'ятником афганцям в Бучі
VERB+PREP+ NOUN	1	
VERB+NOUN	1	Кемеровський СОБР **отримав піздюлєй** і згорів живцем
NOUN+VERB	1	

-грамотність

В деяких текстах були лексичні помилки- вони підсвічені червоним кольором в таблиці в колонці “ Контекст”

-наявність розділових знаків у словах/словосполученнях(*Ху# вам за щоку, пид..расты еб..чие, Ху#ло, ху# смоктати, Пизду#те наху#, х*рня, "руська орда")* .Так зашифровується в більшості випадків лише обценна лексика, для запобігання блокування користувача

-наявність абревіатур у текстах

Підсвічені зеленим кольором у таблиці у колонці “Контекст”

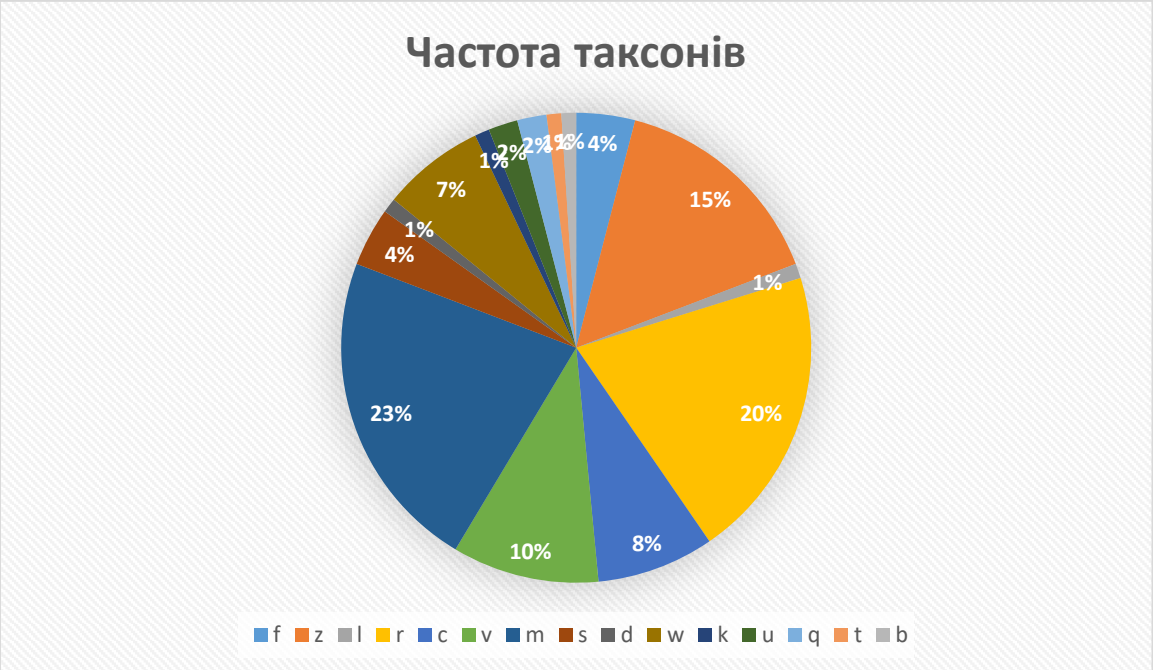
(*СРСР, НКВС, ОВА, БМП, ЗСУ, НАТО, МЗС РФ, ЦВО, РФ, ОДА, ОПЖЗ*)

-вживання мови(українська/російська/суржик), їх частотність

Мова текстів	
українська мова	218
українська+суржик	28
українська+російська	29
українська+суржик+російська	4

-список новоутворених слів під час війни (див. Додаток 7)

-Частота таксонів:



2.6 Доповнення корпусу текстів воєнного періоду та аналіз

Актуальною роботою в рамках дипломного проєкту є автоматично зібрані тексти з Telegram-каналу і доповнення корпусу певними текстами. Було додано тексти з негативною та нейтральною тональністю. Корпус містить 400 текстів.
Для створення програми автоматичного визначення ворожнечі у текстах у корпус були додані тексти з нейтральною тональністю для того, щоб програма могла визначити чи є текст токсичним чи він навпаки токсичності не має, в інших випадках текст май нейтральну тональність. Тексти з нейтральною тональністю не аналізувалися. Нейтральність позначається «0» у колонці «Тональність».

2.7 Порівняння покриття таксонів у двох корпусах

Мною було проаналізовано два корпуси, зокрема, на частоту вживання таксонів. Всі результати, вище, подані у діаграмах. Корпус, наданий мені на другому курсі, укладався 5 років тому, більшість вживань відносились до **менталізмів** – 30%. І цікавим є те, що **менталізми** з кількістю -23% також переважають у 2022-2023 році за результатами покриття таксонів корпусу, укладеного мною. Також порівнюючи дві діаграми, таксон **сексизму**(«s») значно зріс у 2022-2023 роках. 5 років тому **сексизм** займав 5%, у 2022 році – 20%. Таксон «z»- **зоологізми** зріс у 2022 році до 15% , а у 2016-2017 роках був 7%. Таксон «v» - **локалізми** зріс у 2022 році до 10%, 2% було у 2017 році.

2.8 Висновки

У цьому розділі було проаналізована два корпуси текстів. Один був проаналізований за покриттям таксонів мови ворожнечі. Другий корпус був укладений мною та проаналізований за такими критеріями: частотність таксонів, частиномовна характеристика лексем мови ворожнечі, грамотність, наявність розділових знаків у словах/словосполученнях, наявність абревіатур у текстах, вживання мови та її частотність, виведений список новоутворених слів під час

війни, частота таксонів. Для корпусу текстів воєнного періоду (2022-2023) було створено базу даних.

РОЗДІЛ 3.МЕТОДИ АВТОМАТИЗОВАНОГО ВИЗНАЧЕННЯ ОБРАЗЛИВОГО ВМІСТУ ТЕКСТОВИХ ПОВІДОМЛЕНЬ

3.1 Мова програмування Python

Python - це високорівнева інтерпретована мова програмування, яка була створена Гвідо ван Россумом і вперше випущена в 1991 році. Python був розроблений з урахуванням простоти і зрозумілості, щоб зробити програмування приємним і доступним для початківців, але він також є потужним інструментом для професіоналів.

Однією з найбільших переваг Python є його простота. Він має простий і зрозумілий синтаксис, що дозволяє розробникам швидко створювати програми. Python заснований на концепції "читабельного коду", тому їх легше розуміти і змінювати. Це допомагає зменшити кількість помилок і спрощує процес розробки програм.

Інтерпретована природа Python означає, що програми написані на цій мові не потребують компіляції перед виконанням. Це полегшує швидкість розробки і тестування програм, оскільки розробники можуть відразу ж бачити результати своєї роботи.

Python має велику кількість вбудованих бібліотек і модулів, які дозволяють розробникам виконувати різноманітні завдання без необхідності написання великої кількості коду. Наприклад, бібліотека NumPy дозволяє працювати з

масивами даних і виконувати наукові обчислення, а бібліотека Pandas дозволяє зручно обробляти та аналізувати дані.

Python також популярний у галузі штучного інтелекту і машинного навчання. Бібліотека TensorFlow, яка використовується для створення нейронних мереж, і бібліотека scikit-learn, яка надає інструменти для машинного навчання, є лише декількома прикладами потужних інструментів, які доступні у Python.

3.2 Робота з Telegram

Робота з телеграм API мені потрібна, для того, щоб автоматично зібрати тексти з телеграм-каналу “BUCHA LIVE”

Telegram API - це набір інструментів, які дозволяють розробникам створювати програми, що взаємодіють з Telegram-платформою. API дозволяє розробникам створювати програми для Telegram, що можуть виконувати різноманітні завдання, включаючи надсилання повідомлень, створення груп, каналів, ботів, розсилку повідомлень та багато іншого.

Telegram API підтримує різні мови програмування, включаючи Python, JavaScript, Ruby, PHP, Java, C# і багато інших. Інтерфейс API надає розробникам доступ до різних функцій Telegram, таких як:

надсилання повідомлень до користувачів чи груп;

отримання інформації про користувачів, групи та канали;

створення ботів;

створення груп та каналів;

розсилка повідомлень в групи та канали;

отримання повідомлень та взаємодія з ними;

використання різних типів повідомлень, таких як фото, відео, аудіо, документи та інші.

Telegram API дозволяє розробникам створювати різноманітні програми, такі як боти, клієнти для Telegram, інструменти аналізу даних та багато іншого. API є

дуже потужним та може бути використаний для створення різних проектів на базі Telegram-платформи.

В цій роботі, було використано бібліотеки telegram-test-api та Telegram-Client. Знання про Telegram-API допомогли мені автоматично зібрати тексти з Telegram-каналу.

3.3 Визначення тональності в текстах образливого вмісту

Зазвичай визначення образливого вмісту є однією з завдань комп'ютерної лінгвістики. Це означає, що ми можемо використовувати інструменти обробки природної мови, такі як тегери, парсери і т. д., щоб знайти та класифікувати текст як образливий чи ні. Існують кілька підходів до автоматичної класифікації, зокрема підходи, засновані на правилах, словниках і машинному навчанні (з учителем та без учителя).

Перший підхід, оснований на правилах, полягає в написанні правил, які дозволяють віднести текст до певної категорії. Наприклад, для класифікації текстів за предметом шкільної програми можна використовувати правила, які враховують наявність певних термінів. Цей підхід вимагає спеціаліста з розумінням предметної області та навичок написання регулярних виразів. Регулярні вирази - це формалізована мова для пошуку та маніпуляцій з текстом, що базується на використанні метасимволів або символів підстановки. Перевагами цього підходу є висока точність та його широке використання в комерційних системах. Однак, недоліками є складність створення правил, потреба у великій кількості продуманих правил та високі витрати на підтримку системи.

Цей підхід найкраще підходить для сфер людської діяльності, де текстові документи підпорядковані певним законам та алгоритмам побудови речень, наприклад, в юриспруденції та законотворчості. Однак він не є ефективним для

визначення образливого вмісту, особливо в контексті людської комунікації та соціальних мереж. Це пов'язано з постійним появою нових термінів та інфоприводів, що потребує постійного оновлення правил. Затрати на написання та підтримку такої системи правил перевищують її переваги в точності.

Отже, автоматична класифікація, заснована на правилах, не є найкращим підходом для визначення образливого вмісту у соціальних мережах та сферах людської комунікації.

Другий підхід, заснований на словниках, використовує так звані тональні словники або словники тональності для аналізу тексту. У простій формі, тональний словник представляє собою список слів разом з їх тональністю, яка в даному випадку відображає ступінь образливості. Для аналізу тексту застосовується наступний алгоритм: спочатку кожному слову в тексті присвоюється значення тональності зі словника, якщо таке слово присутнє, а потім обчислюється загальна тональність всього тексту. Для розрахунку загальної тональності можна використати різні підходи. Найпростіший з них - обчислити середнє арифметичне значень образливості слів, які зустрічаються в тексті. Більш складним підходом є навчання класифікатора, наприклад, нейронної мережі, для визначення ступеня образливості вхідного тексту.

Перевагою цього методу є його простота в застосуванні. Однак у нього є деякі недоліки, такі як його обмежена універсальність і необхідність великого словника з різноманітними елементами специфічної лексики, яка активно використовується в Інтернеті. Наразі існують словники ненормативної лексики, але використання нецензурних слів у поєднанні з різними допоміжними словами може надавати словам різних відтінків і змісту. Тому для ефективного визначення образливого вмісту необхідний великий та розносторонній словник, що враховує контекст та варіативність використання слів.

Висновок: другий підхід, хоча має свої переваги, все ж потребує значних зусиль для створення та підтримки великого словника, зокрема для розпізнавання образливого вмісту в соціальних мережах та Інтернеті.

Третій підхід - машинне навчання - це галузь штучного інтелекту, яка досліджує методи створення програм та алгоритмів, які можуть навчатися на основі даних, та підбирати самостійно правильні відповіді на задані запитання та завдання.

Одним з основних підходів до машинного навчання є навчання з вчителем, при якому система отримує вхідні дані та зв'язані з ними правильні відповіді, і на основі цього будує математичну модель, яка може передбачати відповіді для нових даних. Також існують підходи без вчителя, які дозволяють системі виявляти закономірності у вхідних даних самостійно, без надання правильних відповідей.

Одним з ключових аспектів машинного навчання є обробка даних, оскільки правильність та якість вхідних даних суттєво впливає на результати моделей. Також важливим є вибір алгоритму навчання, який залежить від характеру завдання та типу даних.

Незважаючи на значні досягнення в галузі машинного навчання, вона також має свої обмеження та виклики, такі, як проблема переносу навчання, коли модель не може ефективно застосовуватися на нових даних, а також етичні питання, пов'язані з використанням машинного навчання: проблема біасу в моделях, що може призвести до дискримінації та нерівності у суспільстві. Тому важливим є розроблення етичних стандартів та правил використання машинного навчання, а також постійна робота над удосконаленням алгоритмів та методів, що використовуються в цій галузі.

Ще одним викликом для машинного навчання є великі обсяги даних, з якими необхідно працювати. Це може призвести до труднощів у збиранні та обробці даних, а також до потреби в потужних обчислювальних ресурсах для тренування та застосування моделей.

Незважаючи на ці виклики, машинне навчання має значний потенціал у вирішенні складних завдань та проблем у багатьох галузях. Дослідження в цій галузі продовжуються, і очікується, що з часом машинне навчання буде застосовуватися ще більш широко та ефективно, що сприятиме подальшому розвитку науки, технологій та інших галузей людської діяльності.

У своїй роботі я використовувала 3 методи машиноого навчання – Naïve Bayes, random forest classifier та Linear SVM.

Naive Bayes (наївний Баєс) - це статистичний алгоритм машинного навчання, який використовуєся для класифікації тексту, фільтрації спаму, аналізу настроїв і багатьох інших завдань. Він базується на теоремі Баєса та припущенні наївності (наївне припущення) про незалежність функціональних характеристик.

Наївний Баєс є досить простим і швидким алгоритмом, який часто використовується для базової класифікації тексту із великої кількості можливих категорій. Він має свої обмеження, особливо коли припущення про незалежність функціональних характеристик не виконується, але в багатьох випадках він продемонструє прийнятну ефективність.

Переваги Naive Bayes: легко і швидко передбачає клас тестового набору даних; добре справляється із багатокласовим прогнозуванням; добре працює з категоріальними ознаками (порівняно з числовими), продуктивність наївного байєсовського класифікатора краща, ніж в інших простих методів, більш того, йому потрібно менше навчальних даних. Недоліки Naive Bayes: обмеженням даного алгоритму є припущення незалежності ознак, однак у реальних завданнях цілком незалежні ознаки трапляються вкрай рідко; якщо змінна має категорію (в тестовому наборі даних), яка не спостерігалася в навчальному наборі даних, то модель надасть 0 (нульову) ймовірність і не зможе зробити прогноз.[1: Вознюк, Левківський]

Random Forest Classifier (випадковий ліс класифікаторів) - це алгоритм машинного навчання, який використовується для задач класифікації і регресії.

Він поєднує декілька рішень дерева рішень, що утворюють "ліс", із застосуванням поняття випадковості.

Алгоритм Random Forest є ансамблевим методом, який використовує кілька класифікаторів дерева рішень під час навчання моделі. Він повертає результат у вигляді модальності класів при класифікації або середнього значення класів для регресійного аналізу. Випадковий ліс поєднує декілька алгоритмів одного типу, тобто декілька дерев рішень, що утворює "ліс". Він може застосовуватись як для задач регресії, так і для класифікації. Random Forest є одним з найпопулярніших алгоритмів, оскільки він є простим, гнучким і має різноманітні застосування. Хоча використання Random Forest має свої переваги і недоліки, він не є упередженим, оскільки використовує декілька дерев, кожне з яких навчається на підмножині даних. В основному, алгоритм Random Forest розраховує на силу "натовпу", що знижує загальну упередженість моделі.

3.4 Визначення токсичності у текстах образливого вмісту

Токсичність у текстах образливого вмісту відноситься до рівня негативних емоцій, які вони викликають у читача. Це може включати образи, які можуть викликати страх, огиду, гнів, обурення або інші негативні емоції у читача. Такі текстові повідомлення можуть містити образливі або зневажливі коментарі, агресивні або нарругливі висловлювання, або сприяти шовіністичним та расистським переконанням.

Існує кілька підходів для визначення токсичності у текстах образливого вмісту, одним із найбільш поширених є машинне навчання та аналіз сентиментів. Цей підхід базується на створенні моделей, які можуть автоматично визначати токсичність текстів на основі певних ознак, таких як використання конкретних слів або відтінків емоцій в тексті.

Інші методи включають в себе ручну оцінку тексту від експертів, аналіз контексту та інших факторів, які можуть впливати на сприйняття тексту. Визначення токсичності у текстах образливого вмісту допомагає розуміти, які

повідомлення можуть бути шкідливими для людей, і виявляти можливість їхнього покращення або видалення.

3.5 Створення програмного забезпечення

3.5.1 Методологічна база

Для створення системи автоматичної класифікації тексту(сентимент-аналіз) було обрано мову програмування Python версії 3.11, головним чином через її модульність та можливість суттєвого розширення функціоналу шляхом залучення готових вбудованих та зовнішніх бібліотек (наприклад, **sklearn, scipy, pandas**). Для створення графічного інтерфейсу (діалогового застосунку) використовувалась бібліотека **tkinter**(бібліотека є стандартним набором інструментів для розробки графічного інтерфейсу користувача (GUI) в мові програмування Python.). Основна ідея застосування Tkinter полягає у створенні вікна, яке містить різні графічні елементи. Це досягається шляхом створення об'єктів різних класів, таких як Tk (головне вікно програми), Frame (контейнер для інших елементів), Button (кнопка), Label (напис) RadioButton. Розробка здійснювалась у редакторі PyCharm.

3.5.2 Джерельна база

Телеграм канал «BUCHA LIVE»

3.5.3 Опис програми:

Програмне забезпечення складається з 2 частин: частина тренування моделі та частина прикладного використання натренованої моделі.

Частина, що відповідає за тренування моделі тренує на базі одного і того самого тренувального датасету 3 різні класифікаційні моделі

- модель на базі Naive Bayes підходу
- модель на базі Random Forest підходу
- Модель на базі SVM підходу. Якщо більш конкретно, то це Linear SVM

Натреновані моделі потім зберігаються у файли для подальшого використання в прикладній частині.

Також робиться пошук гіперпараметрів що покажуть найкращу точність для Random Forest моделі.

Прикладна частина являє собою простий діалоговий застосунок, котрий дає можливість користувачу ввести текст і перевірити його “токсичність” згідно до зібраного тренувального корпусу.

Використання бібліотеки tkinter для створення графічного інтерфейсу:

GROOVE: Це одна з констант, яка використовується для встановлення стилю рамки для віджетів в Tkinter. Встановлення значення GROOVE для параметра relief рамки створює глибоку рамку з піднятими краями.

StringVar: Цей клас є спеціальним класом змінної, який зберігає рядкове значення. Він забезпечує зв'язок між рядком та віджетом Tkinter, таким як Label або Entry, що дозволяє вам оновлювати значення змінної та автоматично оновлювати відображення віджета.

Text: Цей клас використовується для створення багаторядкового текстового віджета в Tkinter. Він надає можливість відображати та редагувати текст у вікні.

Tk: Цей клас є головним вікном програми Tkinter. Він ініціалізує основне вікно програми і обробляє всі події та взаємодії з користувачем.

Toplevel: Цей клас використовується для створення нового вікна у програмі Tkinter. Він дозволяє створювати додаткові вікна, які можуть бути незалежними або зв'язаними з головним вікном.

WORD: Ця константа використовується для налаштування параметра wrap текстового віджета Tkinter. Встановлення значення WORD забезпечує перенесення слова на новий рядок, якщо воно не поміщається на поточному рядку, замість горизонтальної прокрутки.

showinfo: Ця функція є частиною модуля messagebox в Tkinter і використовується для відображення інформаційного діалогового вікна з повідомленням.

Button: Цей клас використовується для створення кнопок у Tkinter. Кнопка може мати текст або зображення і виконувати певну дію, коли на неї натискають.

Label: Цей клас використовується для створення міток (наприклад, текстових надписів) у Tkinter. Він дозволяє відображати статичний текст або змінну.

Labelframe: Цей клас використовується для створення рамки з заголовком у Tkinter. Він дозволяє групувати та організовувати віджети в логічні групи з заголовком.

Radiobutton: Цей клас використовується для створення радіокнопок у Tkinter. Радіокнопки дозволяють користувачеві обрати один варіант з декількох доступних.

Ці бібліотеки налаштовуються шляхом імпортування їх з модуля tkinter (import tkinter або from tkinter import *) та використання їх класів, методів і функцій для створення та налаштування віджетів і вікон у графічному інтерфейсі користувача.

3.5.4 Програмне забезпечення технології: опис програми та коди програми

Повні коди програми наведені у **Додатку 9**

Мова програмування: Python

Бібліотеки: sklearn, numpy, scipy, pandas, click, tkinter

Код програми:

Тренувальна частина

```
import pandas as pd
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer
from sklearn.model_selection import train_test_split, RandomizedSearchCV
from sklearn.naive_bayes import MultinomialNB
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, confusion_matrix, precision_score,
recall_score
from sklearn.svm import LinearSVC
from scipy.stats import randint
import pickle
data = pd.read_csv('./model/corpus.csv')

def preprocess_text(text):
    text = str(text)
    text.replace(r"[^А-Яа-яіїєґ']+", " ")
```

```
    return text.lower()

data['Контекст'] = data['Контекст'].apply(preprocess_text)

# Векторизація тексту
vectorizer = TfidfVectorizer()
X = vectorizer.fit_transform(data['Контекст'])
X_train, X_test, Y_train, Y_test = train_test_split(X, data['Токсичність'],
test_size=0.25, random_state=37)

filename_vectorizer = "./model/vectorizer.sav"
pickle.dump(vectorizer, open(filename_vectorizer, 'wb'))

MNB = MultinomialNB()
MNB.fit(X_train, Y_train)
predicted_MNB = MNB.predict(X_test)
accuracy_MNB = accuracy_score(Y_test, predicted_MNB)
print(f"MultinomialNB: {accuracy_MNB}")
filename_MNB = "./model/MNB.sav"
pickle.dump(MNB, open(filename_MNB, 'wb'))

RF = RandomForestClassifier()
RF.fit(X_train, Y_train)
predicted_RF = RF.predict(X_test)
accuracy_RF = accuracy_score(Y_test, predicted_RF)
print(f"RandomForestClassifier: {accuracy_RF}")
filename_RF = "./model/RF.sav"
pickle.dump(RF, open(filename_RF, 'wb'))

LSVC = LinearSVC()
LSVC.fit(X_train, Y_train)
predicted_LSVC = LSVC.predict(X_test)
accuracy_LSVC = accuracy_score(Y_test, predicted_LSVC)
print(f"LinearSVC: {accuracy_LSVC}")
filename_LSVC = "./model/LSVC.sav"
pickle.dump(LSVC, open(filename_LSVC, 'wb'))


param_dist = {'n_estimators': randint(50,500),
              'max_depth': randint(1,20)}
# Створення Random Forest класифікатору
rf = RandomForestClassifier()
# Застосування рандомізованого пошуку для знаходження найкращих гіперпараметрів
rand_search = RandomizedSearchCV(rf,
                                 param_distributions = param_dist,
                                 n_iter=10,
                                 cv=5)
# Тренування пошуку на даних
rand_search.fit(X_train, Y_train)
# отримання найкращої моделі
best_rf = rand_search.best_estimator_
# та її гіперпараметрів
print('Best hyperparameters:',  rand_search.best_params_)
predicted_BRF = best_rf.predict(X_test)
accuracy_BRF = accuracy_score(predicted_BRF, Y_test)
print(accuracy_BRF)
)
```

Прикладна частина

```
from collections.abc import Iterable
from enum import StrEnum
from pathlib import Path
```

```
from pickle import load
from tkinter import GROOVE
from tkinter import StringVar
from tkinter import Text
from tkinter import Tk
from tkinter import Toplevel
from tkinter import WORD
from tkinter.messagebox import showinfo
from tkinter.ttk import Button
from tkinter.ttk import Label
from tkinter.ttk import Labelframe
from tkinter.ttk import Radiobutton
from typing import cast, Self, Any, Protocol
class Vectorizer(Protocol):
    def transform(self, content: Iterable[str]) -> Any:
        ...
class Classifier(Protocol):
    def predict(self, data: Any) -> Any:
        ...
def deserialize(filename: Path) -> Any:
    with open(filename, 'rb') as file:
        return load(file)
class ClassierType(StrEnum):
    LinearSupportVector = 'Linear SVM'
    RandomForest = 'Random Forest'
    MultinomialNaiveBayes = 'Naive Bayes'
    Default = LinearSupportVector
class Model:
    def __init__(self, classifier: Classifier) -> None:
        self._classifier = classifier
    @classmethod
    def load(cls, filename: Path) -> Self:
        classifier = cast(Classifier, deserialize(filename))
        return cls(classifier)
    def predict(self, input_data: Any) -> Any:
        return self._classifier.predict(input_data)
def load_vectorizer(filename: Path) -> Vectorizer:
    return cast(Vectorizer, deserialize(filename))
def load_models(models_dir: Path) -> dict[ClassierType, Model]:
    models_dir = models_dir.resolve(strict=True)
    return {
        ClassierType.LinearSupportVector: Model.load(models_dir / 'LSVC.sav'),
        ClassierType.RandomForest: Model.load(models_dir / 'RF.sav'),
        ClassierType.MultinomialNaiveBayes: Model.load(models_dir / 'MNB.sav'),
    }
class MainWindow(Toplevel):
    def __init__(self, parent: Tk | Toplevel, *, title: str = "") -> None:
        super().__init__(parent, title)
        self._text: Text
        self._model = StringVar(self, value="Linear SVM", name='model')
        self._create_controls()
    def _create_controls(self) -> None:
        frame = Labelframe(self, text="Model", relief=GROOVE, borderwidth=5)
        Radiobutton(frame, text=ClassierType.LinearSupportVector,
value=ClassierType.LinearSupportVector, variable=self._model).grid(row=0,
column=0, sticky="ewns", padx=5, pady=5)
        Radiobutton(frame, text=ClassierType.RandomForest,
value=ClassierType.RandomForest, variable=self._model).grid(row=1, column=0,
sticky="ewns", padx=5, pady=5)
        Radiobutton(frame, text=ClassierType.MultinomialNaiveBayes,
value=ClassierType.MultinomialNaiveBayes, variable=self._model).grid(row=2,
column=0, sticky="ewns", padx=5, pady=5)
        frame.grid(row=0, column=0, columnspan=3, sticky="ewns", padx=5, pady=5)
```

```
        frame = Labelframe(self, text="Text", relief=GROOVE, borderwidth=5)
        Label(frame, text="Text").grid(row=0, column=0, sticky="ewns", padx=5,
pady=5)
        self._text = Text(frame, wrap=WORD)
        self._text.grid(row=1, column=0, sticky="ewns", padx=5, pady=5)
        frame.grid(row=1, column=0, columnspan=3, sticky="ewns", padx=5, pady=5)
        Button(self, text="Check", command=self._check).grid(row=2, column=1,
sticky="ewns", padx=5, pady=5)
    def _preprocess_text(self, text):
        text.replace(r"[^А-Яа-яіїєґ']+", " ")
        return text.lower()
    def _check(self) -> None:
        text_output = ["Negative", "Neutral", "Positive"]
        vectorizer =
load_vectorizer(Path('../../model/vectorizer.sav').resolve(strict=True))
        models = load_models(Path('../../model').resolve(strict=True))
        # showinfo(message=self._model.get())
        text = self._text.get("1.0","end-1c")
        text = self.preprocess_text(text)
        input_data = vectorizer.transform([text])
        if selected_model := models.get(ClassierType(self._model.get())):
            result = selected_model.predict(input_data)
            showinfo(message=text_output[result[0] + 1])
```

Опис роботи програми

Тренувальна частина

Програма тренування виконує тренування вибраних моделей (Naive Bayes, Random Forest, Linear SVM)

1. Завантажуємо корпус даних з corpus.csv файлу за допомогою функції read() бібліотеки pandas.

2. Застосовуємо до даних з колонки “Контекст” препроцесінг, щоб прибрати звідти всі нетекстові і нечисельні дані, що не належать до української мови. А також всі слова записуємо у нижньому регістрі для спрощення подальшої роботи.

3. Створюємо векторизатор тексту TfidfVectorizer().

4. За допомогою функції fit_transform() перетворюємо весь корпус тексту на його векторизоване TF-IDF представлення.

5. Розділяємо векторизований датасет на трейн та тест підмножини за допомогою функції train_test_split() беручи на трейнову частину 75% усього датасету, а решту на тест частину.

6. Записуємо векторизатор до файлу vectorizer.sav за допомогою функції dump бібліотеки pickle.
7. Створюємо екземпляр класу MultinomialNB що представляє собою Naive Bayes підхід.
8. Тренуємо його за допомогою функції fit().
9. Отримуємо результати тренованої моделі на тестових даних за допомогою функції predict().
10. Рахуємо якість роботи за допомогою функції accuracy_score(), використовуючи Y_test як основу для порівняння.
11. Виводимо показник якості в консоль функцією print().
12. Зберігаємо модель у файл MNB.sav .
13. Створюємо екземпляр класу RandomForestClassifier, що являє собою Random Forest підхід.
14. Тренуємо його за допомогою функції fit().
15. Отримуємо результати тренованої моделі на тестових даних за допомогою функції predict().
16. Рахуємо якість роботи за допомогою функції accuracy_score(), використовуючи Y_test як основу для порівняння.
17. Виводимо показник якості в консоль функцією print().
18. Зберігаємо модель у файл RF.sav.
19. Створюємо екземпляр класу LinearSVC що представляє собою Linear SVM Classification підхід.
20. Тренуємо його за допомогою функції fit().
21. Отримуємо результати тренованої моделі на тестових даних за допомогою функції predict().
22. Рахуємо якість роботи за допомогою функції accuracy_score(), використовуючи Y_test як основу для порівняння.
23. Виводимо показник якості в консоль функцією print().
24. Зберігаємо модель у файл LSVC.sav.
25. Готуємо набір можливих значень для підбору оптимальних значень

гіперпараметрів для Random Forest Classification алгоритму.

26. Створюємо екземпляр класу RandomForestClassifier.

27. Створюємо екземпляр класу для рандомізованого пошуку RandomizedSearchCV().

28. Тренуємо його функцією fit() на тренувальній вибірці.

29. Потрім отримуємо найкращу модель з параметру best_estimator_ та його параметри зі значення змінної best_params_.

30. Опрацьовуємо тестову вибірку найкращим естіматором за допомогою функціїї fit().

31. Обчислюємо якість роботи знайденого класифікатора за допомогою функціїї accuracy_score() та виводимо її в консоль.

Прикладна частина

Складається з простої програми на базі модального діалогового вікна. Функціонування цього діалогового вікна забезпечується за допомогою імплементації класу MainWindow

Цей клас містить функцію конструктор __init__() що проводить базову ініціалізацію параметрів.

Також у класу є функції _create_controls(), вона створює елементи керування на графічному інтерфейсі: радіобаттони, назви груп, поле вводу та кнопку “Check”, натискання на яку призводить до, власне, виконання прикладного функціоналу по перевірці, до якого класу належить введений текст.

Вконання перевірки відбувається за допомогою функції _check(). При виклику цієї функції, ми отримуємо статус групи радіобаттонів, що відповідають за визначення того, яку саме модель користувач хоче використовувати для проведення класифікаціїю Потім за допомогою функції load_vectorizer() завантажуємо векторайзер, що був отриманий на тренувальному кроці. За домопогою функції load_model() завантажуємо з файлу модель, відповідну до

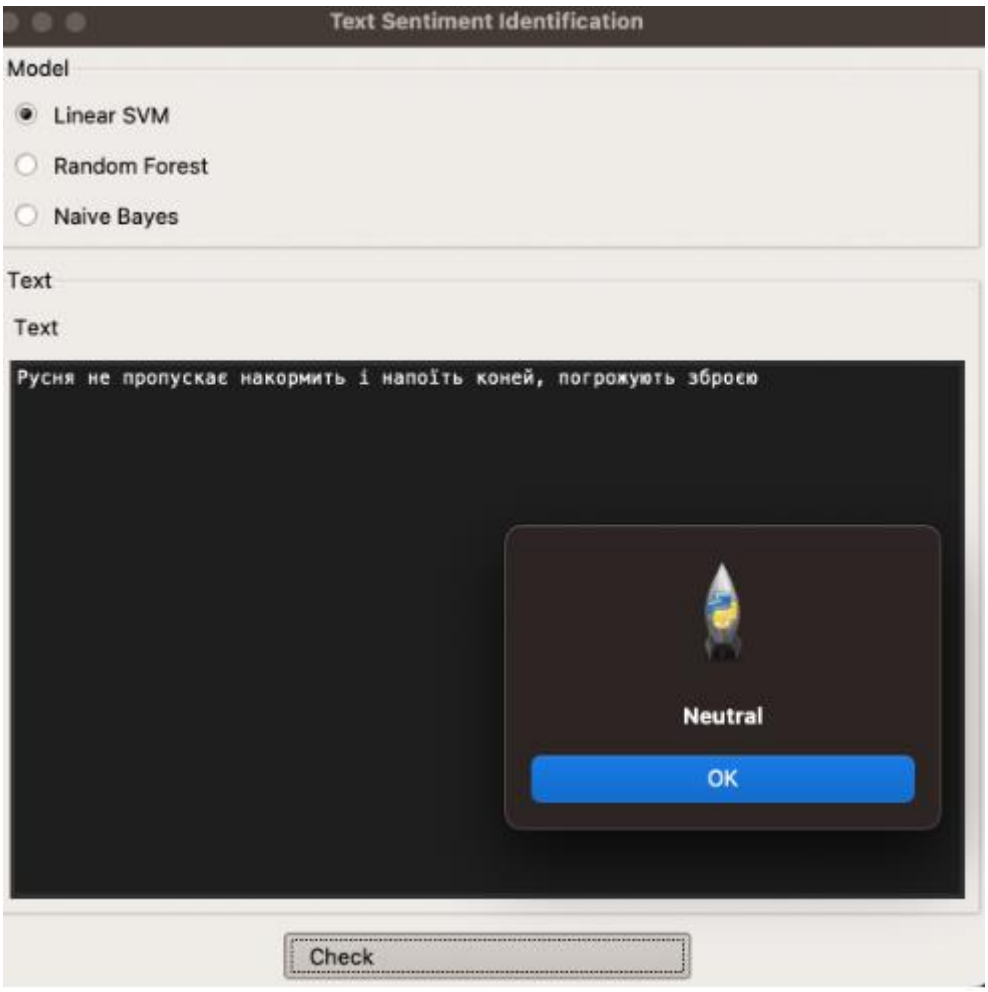
обраної користувачем. Отримуємо з текстового поля введений користувачем текст, що потрібно класифікувати використовуючи метод get() змінної _text.

Після отримання тексту робимо його препроцесінг функцією _preprocess_text() котра прибирає всі символи, що не відносяться до слів і чисел української мови та переводить текст у нижній регістр. Це робиться для того, щоб класифікація тексту відбувалася на даних максимально наближених до даних, що використовувалися для тренування.

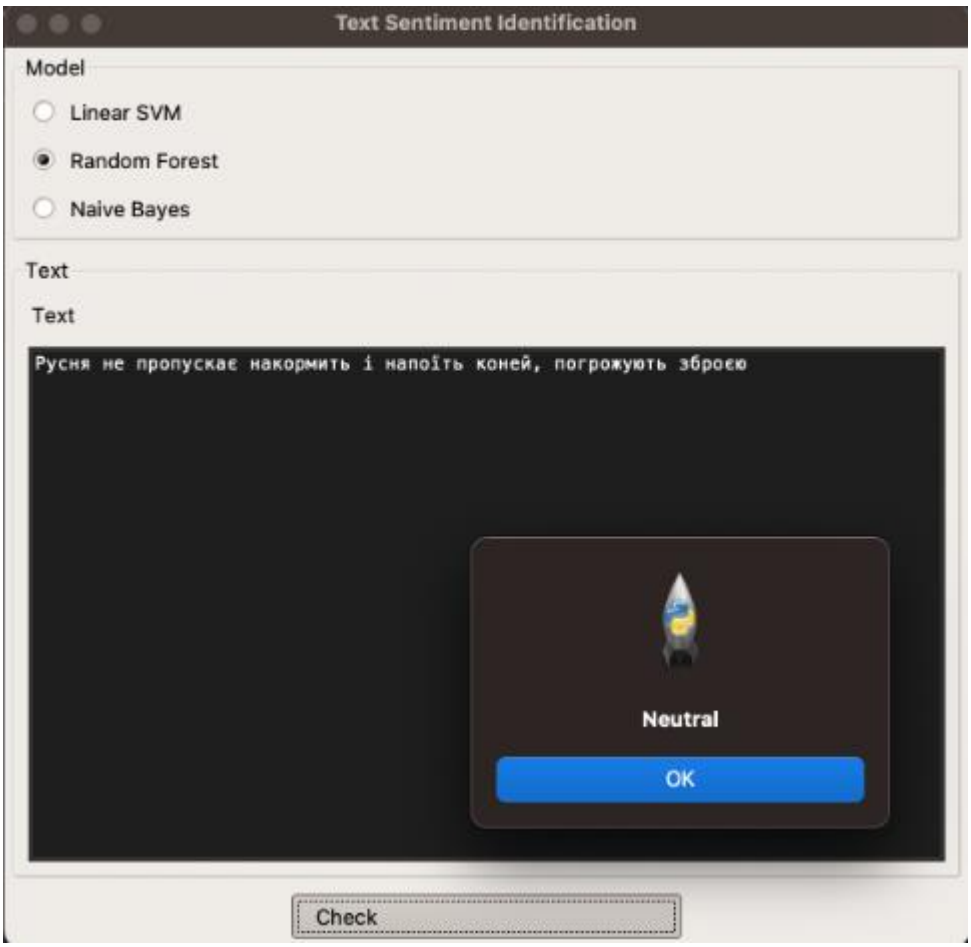
Коли текст підготовано, ми можемо його опрацювати векторайзером і перевести у цільовий фазовий простір за допомогою функції transform(), а потім класифікувати використовуючи функцію predict() обраної і завантаженої моделі.

Результат роботи функції предікт потім переноситься з чисельного вигляду [-1, 0, 1] до відповідного строкового [“Negative”, “Neutral”, “Positive”] і виводиться на екран у окремому діалоговому вікні за допомогою функції бібліотеки tkinter showinfo().

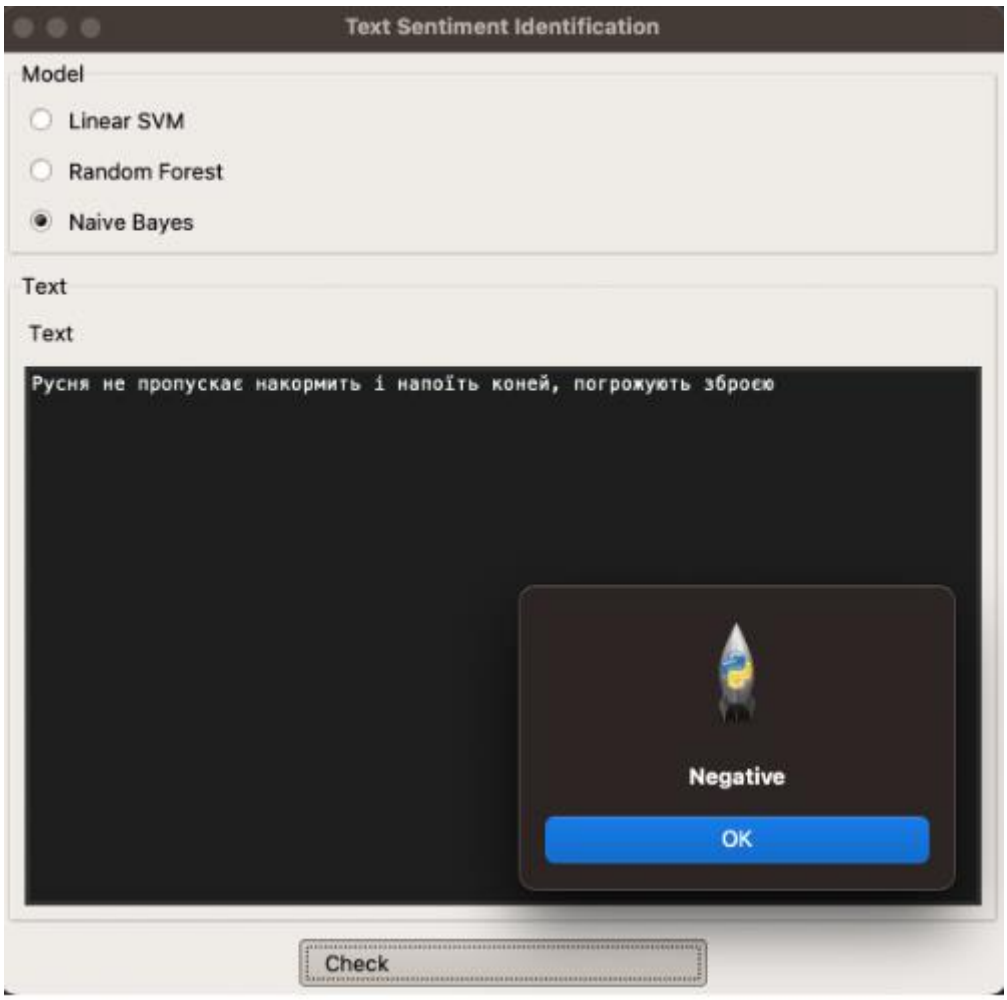
3.5.5 Результат роботи програми:



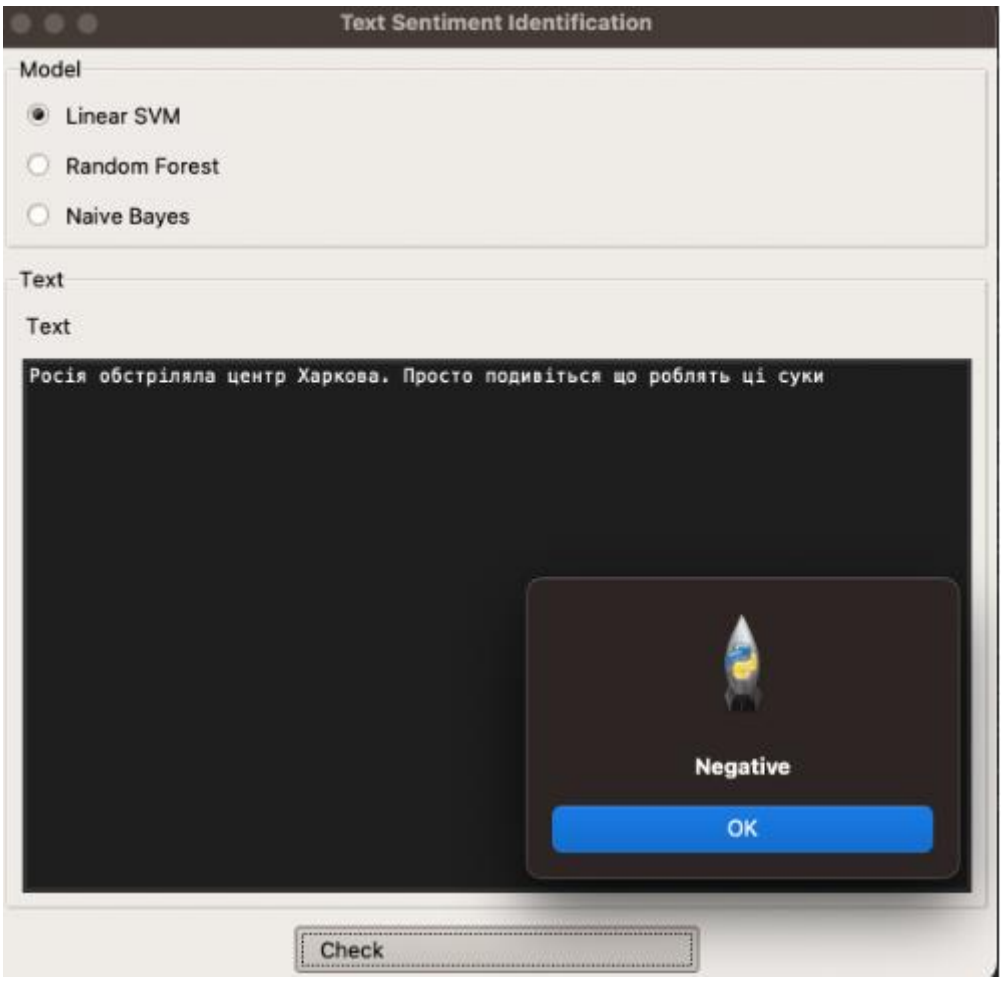
Мал.1 Коректна ідентифікація нейтрального тексту за допомогою Linear SVM



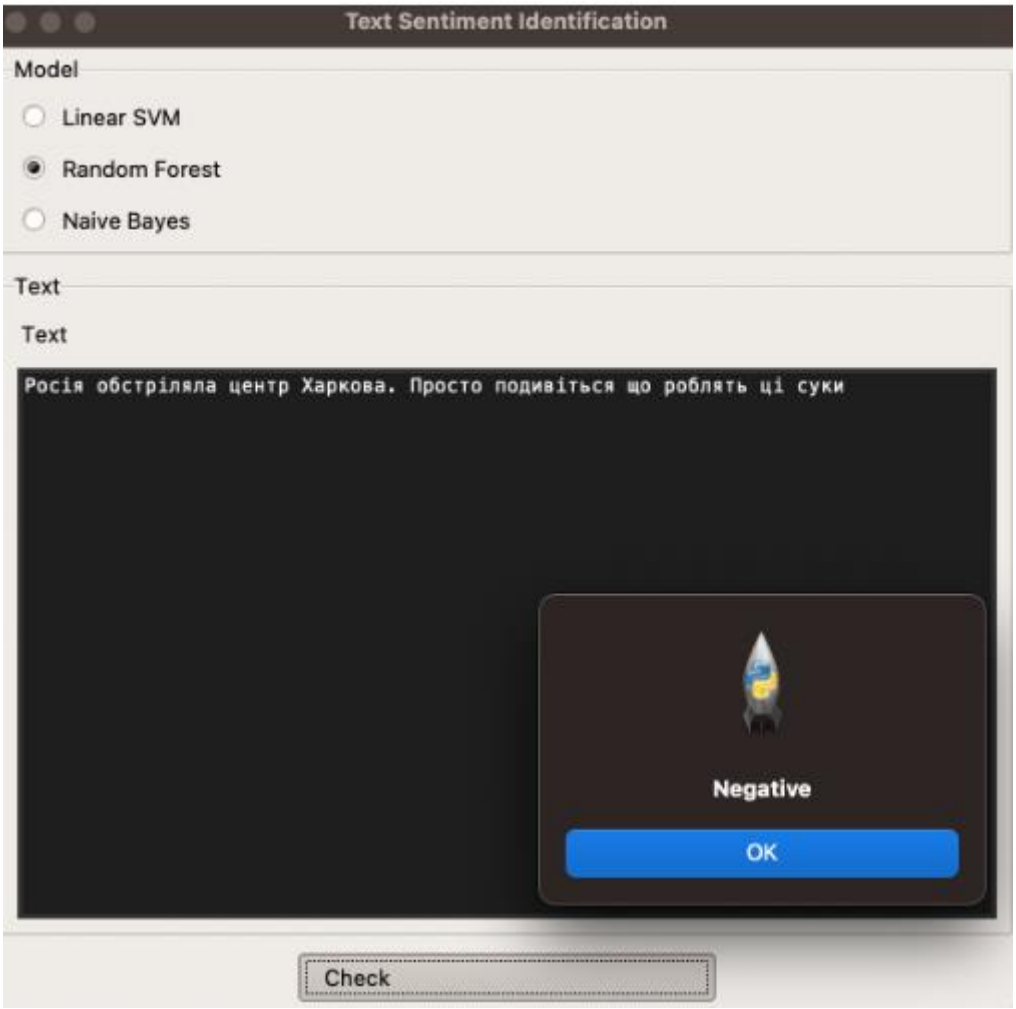
Мал.2 Коректна ідентифікація нейтрального тексту за допомогою Random Forest



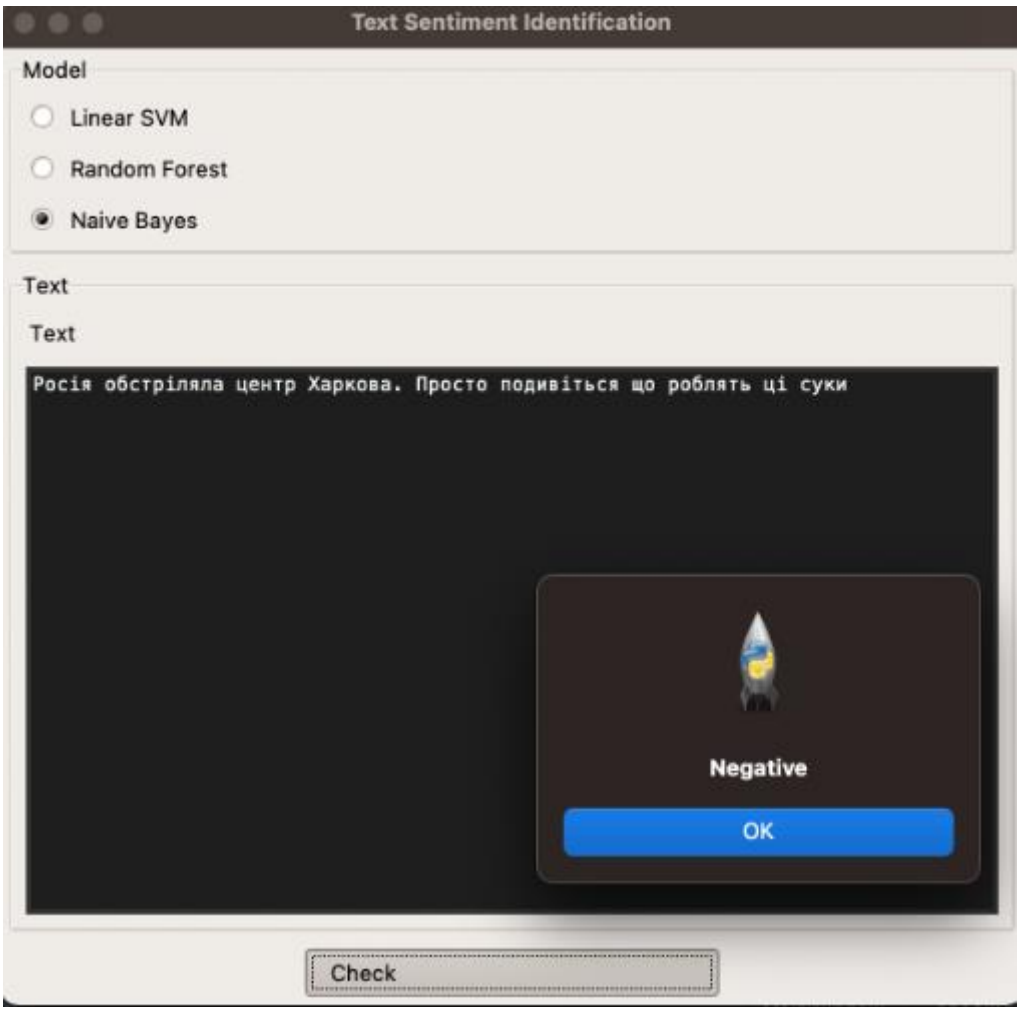
Мал.3 Некоректна ідентифікація нейтрального тексту за допомогою Naive Bayes



Мал.4 Коректна ідентифікація негативного тексту за допомогою Linear SVM



Мал.5 Коректна ідентифікація негативного тексту за допомогою Random Forest



Мал.6 Коректна ідентифікація негативного тексту за допомогою Naive Bayes

3.5.6 Висновки

В процесі роботи над задачею було обрано 3 підходи до класифікації тональності українських текстів - Naive Bayes, Random Forrest та Linear SVM. Для кожного з підходів було натреновано модель-класифікатор, використовуючи корпус даних, що був зібраний з використанням телеграм-каналу “BUCHA LIVE”. Всі три моделі показали високий рівень точності ідентифікації (метрика accuracy). Він у них був більше 75%. Найкраще за все показала модель Linear SVM підходу, оскільки вона була менш всього залежна від вибору підходів до векторизації фазового простору в той час як інші дві при зміні векторизатора на більш простий CountVectorizer(). У цьому випадку SVM модель втрачала до 3% акуратності, в той час як для інших двох метрика падала нижче 50%.

На додачу варто зауважити, що в прикладній частині не була використана модель Random Forest, що була отримання після перебору гіперпараметрів, оскільки ця

модель давала метрику акуратності 76% в той час як модель без тюнінгу параметрів показувала акуратність на рівні 80%.

Також варто відмітити, що специфіка зібраних даних (обсценна лексика стосовно представників росії), час збору даних (активна фаза війни після 24.02.2022) та обмеженість ресурсу збору призвели до того, що тренувальний корпус є дуже незбалансованим з перекосом у тексти, що мають негативний окрас. Це робить результати тренування та валідації частково неконсистентними і дає напрямки для подальших досліджень або для розширення і балансування корпусу даних або пошуку інших моделей класифікації, що зможуть показати кращу поведінку в умовах незбалансованих датасетів.

Список використаних джерел

1.АНАЛІЗ МОДЕЛЕЙ КЛАСИФІКАЦІЇ В PYTHON ДЛЯ СТВОРЕННЯ СИСТЕМИ АВТОМАТИЧНОЇ КАТЕГОРИЗАЦІЇ ПУБЛІКАЦІЙ БЛОГУ- Вознюк М. Ю., Левківський В. Л

2. Андерсон Б. Уявлені спільноти. Міркування щодо походження й поширення націоналізму. — К., 2001. — 272 с.

3. Бєлай С. В. Дослідження соціальної напруженості як передумови виникнення кризових явищ / С. В. Бєлай // Розвиток системи державного управління в Україні / С. В. Бєлай., 2012. – С. 27–34.

4. Бойко Н.Л. «Hatespeech» в соціальних мережах як показник соціальної напруженості /Нові нерівності – нові конфлікти: шляхи подолання. тези доповідей та виступів учасників III конгресу Соціологічної Асоціації України, (Харків, 12-13 жовтня 2017 р.) – Харків: ХНУ імені В.Н.Каразіна, 2017. – С.58-59.

5. Дуцик Д. «Мова ворожнечі» в дискурсі українських медіа [Електронний ресурс] / Д. Дуцик. – 2012. – Режим доступу до ресурсу: http://ekmair.ukma.edu.ua/handle/123456789/2332.

6. *Заболотня Т.М., Бартков'як А.Ю*. Програмна бібліотека методів аналізу тональності відгуків Інтернет-користувачів // "Informatics and Computer Technics Problems". Тези доповідей. – Чернівці. 21-24 травня 2016. – С. 85-87.

7. *Заболотня Т.М., Соколовська А.В.* Підхід до визначення образливого вмісту текстових повідомлень в соціальних мережах // ”Прикладна математика та комп’ютинг. ПМК 2018”.

8. Кольцова Є. Ю. Вимір толерантності / Є. Ю. Кольцова, Є. Є. Таратута // Журнал соціології та соціальної антропології. – 2003. – Т. 6., № 4. – С. 113-129.

9. Конвенція про захист прав людини і основоположних свобод, Рада Європи; Конвенція, Міжнародний документ від 04.11.1950. – Режим доступу: http://zakon2.rada.gov.ua/laws/show/995 004.

10. Карп’як О. Вата з укропом: мова політичних мемів. URL: http://discourse.in.ua/library/vata- z- ukropommova- politychnyh- memiv

11. Лихова С. Я. Мова ворожнечі як ознака порушення рівноправності громадян залежно від їх расової, національної належності або ставлення до релігії / С. Я. Лихова, Г. Г. Рибальченко // Актуальні проблеми вдосконалення чинного законодавства України. - 2013. - Вип. 31. - С. 274-282.

12. Мова ворожнечі та ЗМІ: міжнародні стандарти та підходи. URL: http://noborders.org.ua/wp- content/ uploads/2015/11/guidelines_hatespeech_media.pdf

13. Мова ворожнечі у ЗМІ: якою вона буває та до чого призводить (інфографіка) [Електронний ресурс] // Інститут масової інформації. – 2016. – Режим доступу до ресурсу: http://imi.org.ua/advices/mova-vorojnechi-u-zmi-yakoyu-vona-buvae-ta-do-chogo-prizvodit-infografika/.

14. Новий словник української мови: у 4 т. С.853.

15. Новий словник української мови: у 4 т. Київ: Аконіт, 2000. Т. 1 / укладачі: В. В. Яременко, О. М. Сліпушко. С. 520.

16. Паніотто В. Динаміка ксенофобії й антисемітизму в Україні (1994-2007) [Електронний ресурс] / В. Паніотто // КМІС. – 2007. – Режим доступу до ресурсу: http://www.kiis.com.ua/materials/articles/xenophobia_antisemitism.pdf.

17. Попова О. М. Прояви вербальної агресії у ЗМІ / О. М. Попова, Л. М. Солодун // Молодий вчений / О. М. Попова, Л. М. Солодун. – Маріуполь, 2016. – С. 654–658.

18. Почепцов Г. Г. Особенности пропагандистских механизмов с двух сторон российско- украинского конфликта. URL: http://osvita.mediasapiens.ua/material/33530.

19. Правий екстремізм і толерантність: з досвіду України та Німеччини / Фонд ім. Фрідріха Еберта, Регіональне представництво в Україні та Білорусі. К.: Заповіт, 2008. — 76 с.

20. Селевко Г. К. Енциклопедія освітніх технологій. – М.: НИИ шкільних технологій, 2006. – Т. 1. – 816 с.

21. Таркін В.П. ПРОПАГАНДА ЯК МЕТОД ІНФОРМАЦІЙНО-ПСИХОЛОГІЧНОЇ ВІЙНИ: ПОЛІТИЧНИЙ КОНТЕКСТ, 2019- 2с

22. Толокольнікова К. Уникання мови ворожнечі наблизить мир [Електронний ресурс]. – 2015. – Режим доступу до ресурсу: http://osvita.mediasapiens.ua/ethics/standards/unikannya_movi_vorozhnechi_nablizit_mir/.

23. Український світ у наукових парадигмах: Збірник наукових праць Харківського національного педагогічного університету імені Г. С. Сковороди. — Харків : ХІФТ, 2020. — Вип. 7. — 211, [6] с.

24. Черненко Г.А. ВИДИ МОВИ ВОРОЖНЕЧІ ЗА ТИПОМ ... www.irbis-nbuv.gov.ua › irbis_nbuv › cgiirbis_64

25. *D. Jurafsky,* and *J.H. Martin* Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. / *Jurafsky, D.* and *Martin, J.H.* // — Pearson Prentice Hall, 2009. — 2009 — P. 988.

26. Recommendation NO. R (97) 20 of the committee of ministers to member stateson. URL: https://wcd.coe.int/ com.instranet.InstraServlet?command=com.instranet.CmdBlobGet&InstranetImage= 1658005&SecMode=1&D ocId=582600&Usage=2 «HATE SPEECH»

27. SQlite-СУБД

28. Електронний ресурс https://internews.ua/opportunity/antybot-notes

29. Електронний ресурс https://cedem.org.ua/consultations/mova-vorozhnechi-yaki-vyslovlyuvannya-ne-mozhna-poshyryuvaty-v-zmi/

30.Електронний
Ресурс http://mdu.in.ua/Nauch/Konf/2016/suchasni_mediakomunikaciji.pdf ст 112

31. Електронний ресурс: https://naub.oa.edu.ua/2017/мова-ворожнечі-в-засобах-масової-ін/

32. Електронний ресурс: https://m.day.kyiv.ua/uk/article/media/keruvannya-nenavystyu

33. Електронний ресурс: https://detector.media/infospace/article/121803/2016-12-27-mova-vorozhnechi-chy-ie-protyrichchya-mizh-profesiynym-ta-gromadyanskym-obovyazkom/

34. Електронний ресурс: https://www.radiosvoboda.org/a/24740474.html

35. Інтерет-ресурс: https://www.volyn24.com

36. Корпус:  movavorozhnechi.mdb

ДОДАТКИ

Додаток 1

https://drive.google.com/drive/u/0/folders/1opZZQHT_ie_LLM_nI-iJHbeN19dnjtqA

У цьому додатку ви можете перейти за посиланням і побачити папки з Додатками, вам потрібно перейти в папку Додаток 1, щоб подивитись код автоматичного укладання Бази даних на основі корпусу текстів воєнного періоду

```
import sqlite3
import pymorphy2
from csv import reader
from telethon import TelegramClient, events, sync
from telethon.errors import SessionPasswordNeededError


morph_analyzer = pymorphy2.MorphAnalyzer(lang="uk")
```

```
def check_db():
    db = sqlite3.connect("./data/db/corpus.db")
    cursor = db.cursor()

    cursor.execute("""CREATE TABLE IF NOT EXISTS Bad_Words
                      (Word_ID INTEGER PRIMARY KEY,
                       Word TEXT);""")

    cursor.execute("""CREATE TABLE IF NOT EXISTS Context_Types
                         (Type_ID INTEGER PRIMARY KEY,
                          C_Type TEXT);""")

    cursor.execute("""CREATE TABLE IF NOT EXISTS Sources
                            (Source_ID INTEGER PRIMARY KEY,
                             C_Source TEXT);""")

    cursor.execute("""CREATE TABLE IF NOT EXISTS Classifications
                               (Classification_ID INTEGER PRIMARY KEY,
                                Classification TEXT);""")

    cursor.execute("""CREATE TABLE IF NOT EXISTS Languages
                                  (Lang_ID INTEGER PRIMARY KEY,
                                   Lang TEXT);""")

    db.commit()
    db.close()


def to_db_record(input):
    db = sqlite3.connect("./data/db/corpus.db")
    cursor = db.cursor()
    tmp = input[:]

    # Change word to its ID in Bad_Words table. Add if not exists
    bad_word = tmp[0]
    res = cursor.execute("""SELECT Word_ID from Bad_Words WHERE Word=?""",
(bad_word,)).fetchall()
    if len(res) == 0:
        cursor.execute("""INSERT INTO Bad_Words (Word) VALUES (?)""", (bad_word, ))
        res = cursor.execute("""SELECT Word_ID from Bad_Words WHERE Word=?""",
(bad_word,)).fetchall()
```

```
    tmp[0] = res[0][0]

    type = tmp[2]
    res = cursor.execute("""SELECT Type_ID from Context_Types WHERE C_Type=?""",
(type,)).fetchall()
    if len(res) == 0:
        cursor.execute("""INSERT INTO Context_Types (C_Type) VALUES (?)""", (type,))
        res = cursor.execute("""SELECT Type_ID from Context_Types WHERE C_Type=?""",
(type,)).fetchall()
    tmp[2] = res[0][0]

    source = tmp[3]
    res = cursor.execute("""SELECT Source_ID from Sources WHERE C_Source=?""",
(source,)).fetchall()
    if len(res) == 0:
        cursor.execute("""INSERT INTO Sources (C_Source) VALUES (?)""", (source,))
        res = cursor.execute("""SELECT Source_ID from Sources WHERE C_Source=?""",
(source,)).fetchall()
    tmp[3] = res[0][0]

    classification = tmp[5]
    res = cursor.execute("""SELECT Classification_ID from Classifications WHERE
Classification=?""", (classification,)).fetchall()
    if len(res) == 0:
        cursor.execute("""INSERT INTO Classifications (Classification) VALUES (?)""",
(classification,))
        res = cursor.execute("""SELECT Classification_ID from Classifications WHERE
Classification=?""", (classification,)).fetchall()
    tmp[5] = res[0][0]

    lang = tmp[7]
    res = cursor.execute("""SELECT Lang_ID from Languages WHERE Lang=?""",
(lang,)).fetchall()
    if len(res) == 0:
        cursor.execute("""INSERT INTO Languages (Lang) VALUES (?)""", (lang,))
        res = cursor.execute("""SELECT Lang_ID from Languages WHERE Lang=?""",
(lang,)).fetchall()
    tmp[7] = res[0][0]

    lang = tmp[10]
    res = cursor.execute("""SELECT Lang_ID from Languages WHERE Lang=?""",
```

```
(lang,)).fetchall()
    if len(res) == 0:
        cursor.execute("""INSERT INTO Languages (Lang) VALUES (?)""", (lang,))
        res = cursor.execute("""SELECT Lang_ID from Languages WHERE Lang=?""",
(lang,)).fetchall()
    tmp[10] = res[0][0]

    tmp[9] = bool(int(tmp[9]))
    words = bad_word.split( )
    tmp[8] = ''
    for w in words:
        w.strip()
        tmp[8] = tmp[8] + ' + ' + str(morph_analyzer.parse(w)[0].tag.POS)
    tmp[8].strip(' + ')

    db.commit()
    return tuple(tmp)


def main():
    with open("./data/data.csv", "r") as input:
        csv_reader = reader(input)
        # Iterate over each row in the csv using reader object
        records = [row[1:] for row in csv_reader]

    records.pop(0)
    print(records)

    records = [to_db_record(r) for r in records]
    print(records)
    #records = [tuple(r) for r in records]
    db = sqlite3.connect("./data/db/corpus.db")
    cursor = db.cursor()
    cursor.execute("""DROP TABLE IF EXISTS General_Union;""")
    cursor.execute("""CREATE TABLE General_Union
                      (ID INTEGER PRIMARY KEY,
                       Word INTEGER,
                       Context TEXT,
                       Text_Type INTEGER,
                       Text_Source INTEGER,
                       Date_Gathered TEXT,
```

```
                        Classification INTEGER,
                        Tonality TEXT,
                        Word_Language INTEGER,
                        POS TEXT,
                        Delimiters BOOL,
                        Context_Language INTEGER);""")

    cursor.executemany("""INSERT INTO General_Union
                        (Word, Context, Text_Type, Text_Source, Date_Gathered,
Classification, Tonality,
                        Word_Language, POS, Delimiters, Context_Language)
                        VALUES (?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?)""", records)
    db.commit()
    db.close
    pass


if __name__ == "__main__":
    check_db()
    main()
```

Даний код містить програму, яка реалізує операції зберігання та обробки даних з текстового файлу формату CSV.

Спочатку виконується імпорт необхідних модулів для роботи з базою даних SQLite, для морфологічного аналізу текстів за допомогою бібліотеки pymorphy2 та для отримання даних з месенджера Telegram за допомогою бібліотеки telethon.

Далі визначається функція check_db(), яка перевіряє, чи існує база даних, якщо ні - створює її та таблиці для зберігання інформації про погані слова, контексти, джерела текстів, класифікації, мови тощо.

Функція to_db_record() перетворює запис з CSV-файлу у відповідний запис для бази даних та здійснює деякі операції над цим записом, зокрема заміну слова на його ідентифікатор у таблиці поганих слів, отримання ідентифікаторів для

інших полів тощо. Також в цій функції проводиться морфологічний аналіз слів та визначення їх частин мови.

Головна функція main() відкриває CSV-файл, з якого зчитує дані та виконує ітерацію по рядках файлу, вилучаючи перший рядок (з заголовками стовпців). Потім записи з CSV-файлу перетворюються в записи бази даних за допомогою функції to_db_record(). Останнім кроком є збереження отриманих записів в базу даних SQLite за допомогою команди executemany().

Ці процедури забезпечують зручний та ефективний доступ до даних, які можна використовувати для дослідження та аналізу різних аспектів мови та текстів.

Ця база даних містить 5 таблиць:

"Bad_Words" - містить список поганих слів. Кожне погане слово має унікальний ідентифікатор (Word_ID) та текстове значення (Word).

"Context_Types" - містить типи контекстів, в яких можуть зустрічатися погані слова. Кожен тип контексту має унікальний ідентифікатор (Type_ID) та текстове значення (C_Type).

"Sources" - містить джерела, з яких були отримані дані про погані слова та їх контексти. Кожне джерело має унікальний ідентифікатор (Source_ID) та текстове значення (C_Source).

"Classifications" - містить класифікації поганих слів за різними критеріями. Кожна класифікація має унікальний ідентифікатор (Classification_ID) та текстове значення (Classification).

"Languages" - містить список мов, які використовуються в базі даних. Кожна мова має унікальний ідентифікатор (Lang_ID) та текстове значення (Lang).

Таблиці пов'язані між собою за допомогою зовнішніх ключів. Наприклад, в таблиці "Bad_Words" є стовпець "Lang_ID", який посилається на стовпець "Lang_ID" таблиці "Languages". Це дозволяє пов'язувати погані слова з мовою, на якій вони вживаються.

Крім того, в таблиці "Bad_Words" є стовпець "Context_Type_ID", який посилається на стовпець "Type_ID" таблиці "Context_Types". Це дозволяє пов'язувати погані слова з типами контекстів, в яких вони можуть зустрічатися.

Так само, в таблиці "Bad_Words" є стовпець "Classification_ID", який посилається на стовпець "Classification_ID" таблиці "Classifications". Це дозволяє класифікувати погані слова за різними критеріями.

Додаток 2

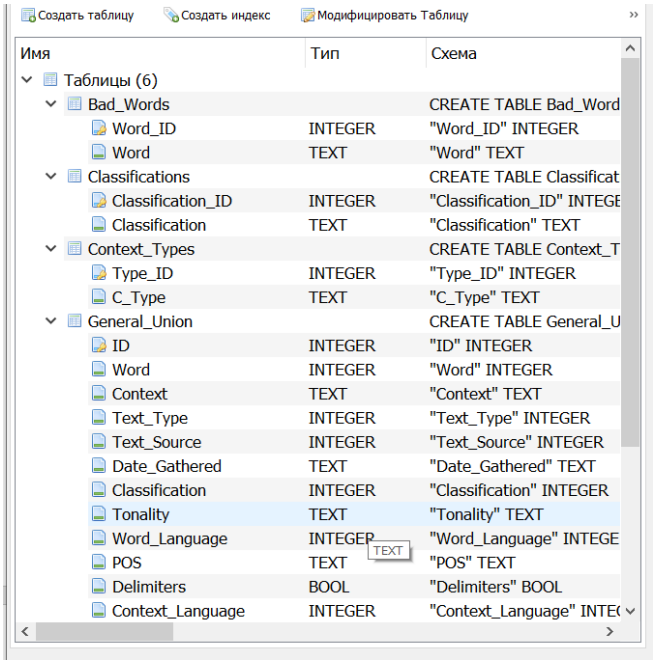
Це скрипт, який дозволяє зчитувати нові повідомлення, які надходять до телеграм каналу

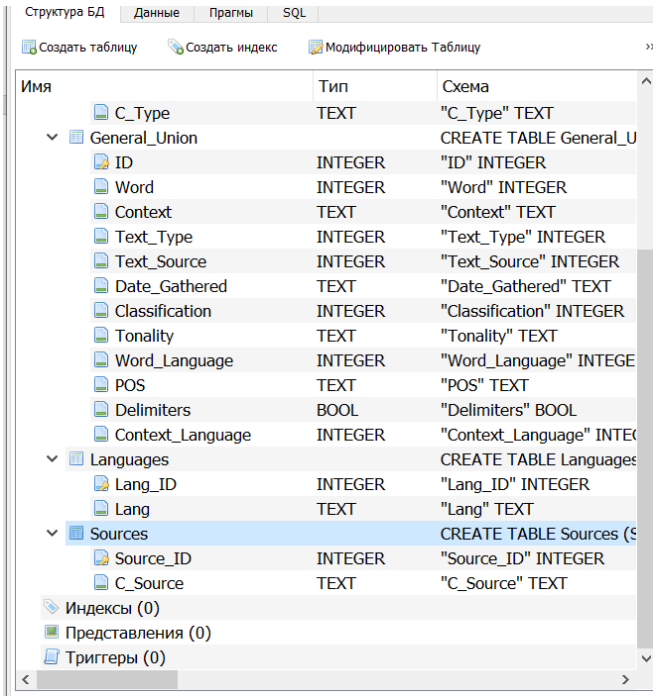
```
def telegram_client():
    api_id = 17896098
    api_hash = "eadeb246b9478bd22fd50c5b7f705846"

    phone = "+380501333745"
    username = "katyadegtyar"

    # Create the client and connect
    client = TelegramClient(username, api_id, api_hash)

    @client.on(events.NewMessage(chats='BUCHA LIVE'))
    async def my_event_handler(event):
        print(event.raw_text)

    client.start()
    print("Client Created")
    # Ensure you're authorized
    if not client.is_user_authorized():
        client.send_code_request(phone)
        try:
            client.sign_in(phone, input('Enter the code: '))
        except SessionPasswordNeededError:
            client.sign_in(password=input('Password: '))

    client.run_until_disconnected()
```

Додаток 3

corpus.db

У цьому Додатку ви можете ознайомитися з скріншотами роботи програми по укладанню бази даних для корпусу текстів воєнного часу





Bad_Words (119 rows)

```
SELECT * FROM 'Bad_Words' LIMIT 0,30
```

Execute

Word_ID	Word
1	окупант
2	кат
3	тварюка
4	свинособака
5	ірод
6	москаль
7	руський мір
8	русня
9	Могуча російська армія
10	сука
11	щур
12	орк
13	"руська орда"
14	дно
15	"визволителі"
16	руська орда

Classifications (32 rows)

```
SELECT * FROM 'Classifications' LIMIT 0,30
```

Execute

Classification_ID	Classification
1	f
2	
3	z
4	l
5	r
6	c
7	v
8	v/r
9	m
10	s
11	d
12	w
13	k
14	p/r
15	u
16	c
17	m/c

Context_Types (5 rows)

```
SELECT * FROM 'Context_Types' LIMIT 0,30
```

Execute

Type_ID	C_Type
1	пост
2	коментар
3	коментар під відео
4	коментар,пост
5	

General_Union (280 rows)

```
SELECT * FROM 'General_Union' LIMIT 0,30
```

Execute

ID	Word	Context	Text_Type	Text_Source	Date_Gathered	Classification	Tonality	Word_Language	POS	Delimiters	Co
1	1	Російський окупант із Вла…	1	1	05.04.2022	1	-1	1	+ NOUN	0	2
2	2	Російський окупант із Вла…	1	1	05.04.2022	2	-1	1	+ NOUN	0	2
3	3	Смачного тварюко	2	1	27.02.2022	3	-1	1	+ NOUN	0	2
4	4	Свинособаки застосовуют…	1	1	02.03.2022	3	-1	1	+ NOUN	0	2
5	5	Свинособаки застосовуют…	1	1	02.03.2022	4	-1	1	+ NOUN	0	2
6	6	Здається українці запусти…	2	1	28. 02.2022	5	1	1	+ NOUN	0	2
7	7	Охтирка після спроби вста…	2	1	28.02.2022	6	-1/0	1	+ ADJF + NOUN	0	2
8	8	Русня не пропускає накор…	2	1	28.02.2022	5	0	1	+ NOUN	0	2
9	9	"Могуча російська армія" в…	3	1	28.02.2022	6	1	1	+ ADJF + ADJF + NOUN	0	2
10	7	Ще одна підбірка фотогра…	2	1	28.02.2022	6	0	1	+ ADJF + NOUN	0	2
11	10	Росія обстріляла центр Ха…	2	1	01.03.2022	3	-1	1	+ NOUN	0	2
12	11	Росія імовірно використов…	1	1	02.03.2022	3	0	1	+ NOUN	0	2
13	7	Ще кадри "руського міра" …	2	1	02.03.2022	6	0	1	+ ADJF + NOUN	0	2
14	12	Буча, розбита колона росі…	2	1	02.03.2022	7	-1	1	+ NOUN	0	2
15	13	Тисячі людей не хочуть пр…	1	1	02.03.2022	8	0	1	+ ADJF + None	1	2
16	14	Вони хочуть зруйнувати на…	2	1	03.03.2022	9	0	1	+ NOUN	0	2
17	7	Страшні кадри… Так вигля…	2	1	04.03.2022	6	0	3	+ ADJF + NOUN	0	3

General_Union (280 rows)

```
SELECT * FROM 'General_Union' LIMIT 0,30
```

Execute

ext	Text_Type	Text_Source	Date_Gathered	Classification	Tonality	Word_Language	POS	Delimiters	Context_Language
ський окупант із Вла…	1	1	05.04.2022	1	-1	1	+ NOUN	0	2
ський окупант із Вла…	1	1	05.04.2022	2	-1	1	+ NOUN	0	2
ного тварюко	2	1	27.02.2022	3	-1	1	+ NOUN	0	2
особаки застосовуют…	1	1	02.03.2022	3	-1	1	+ NOUN	0	2
особаки застосовуют…	1	1	02.03.2022	4	-1	1	+ NOUN	0	2
ться українці запусти…	2	1	28. 02.2022	5	1	1	+ NOUN	0	2
жка після спроби вста…	2	1	28.02.2022	6	-1/0	1	+ ADJF + NOUN	0	2
а не пропускає накор…	2	1	28.02.2022	5	0	1	+ NOUN	0	2
ча російська армія" в…	3	1	28.02.2022	6	1	1	+ ADJF + ADJF + NOUN	0	2
дна підбірка фотогра…	2	1	28.02.2022	6	0	1	+ ADJF + NOUN	0	2
обстріляла центр Ха…	2	1	01.03.2022	3	-1	1	+ NOUN	0	2
імовірно використов…	1	1	02.03.2022	3	0	1	+ NOUN	0	2

Languages (12 rows)

```
SELECT * FROM 'Languages' LIMIT 0,30
```

Execute

Lang_ID	Lang
1	українська
2	українська мова
3	українська+суржик
4	українська мова+суржик
5	суржик
6	українська+російська
7	російська мова
8	українська +російська
9	українська+ російська
10	
11	українська+суржик+російська
12	українська мова+російська

Sources (2 rows)

```
SELECT * FROM 'Sources' LIMIT 0,30
```

Execute

Source_ID	C_Source
1	телеграм канал "BUCHA LIVE"
2	

Додаток 4

1. Корпус 2:

У цьому додатку, щоб подивитись на повний вміст корпусу текстів воєнного періоду, ви можете перейти за посиланням.

Структура корпусу: Корпус містить 280 слововживань. Має 12 колонок: “ID”, “Слово/Словосполучення”, “Контекст”, “Тип тексту”, “Джерело”, “Дата”, “Класифікація”, “Тональність слова”, “Мова слова”, “Частина мови”, “Наявність розділових знаків у слові/ словосполученні”, “Мова тексту”. Класифікацію використано з поданого вище списку. Корпус містить не лише мову ворожнечі, вживаною українською мовою, а російською й суржиком. Наявні слова в яких голосні літери зашифровані знаками: “*”, “#”, “…” “”. У корпусі багато емоційних, політичних, расистських висловлювань. Майже вся лексика відноситься до російсько-української війни та російської нації.

https://docs.google.com/spreadsheets/d/1kHGT_CgxJ3ZwNNPahntfDhS0rA2AHfMJSGmExNr2riM/edit#gid=0

2. https://inloop.github.io/sqlite-viewer/

3. База даних:  corpus.db

Додаток 5

У цьому додатку ви можете перейти за посиланням, зайти в папку Додаток 5 і подивитись скрипт, за яким я автоматично збирала тексти воєнного періоду, а саме з вересня 2022 року по квітень 2023 року. Перейшовши в папку Додаток 5, вам потрібно зайти у файлк “скрипт-парсер для телеграм каналу”. **https://drive.google.com/drive/u/0/folders/1opZZQHT_ie_LLM_nI-iJHbeN19dnjtqA**

Це зразок Python-скрипту для обробки даних, які зберігаються в форматі JSON. Скрипт призначений для обробки даних з месенджера Telegram.

Код містить функцію parse, яка приймає один аргумент - ім'я файлу для обробки. Файл повинен бути в форматі JSON, що містить інформацію про повідомлення в месенджері Telegram.

У функції спочатку відкривається файл з даними, потім використовується функція json.load для зчитування даних у форматі JSON. Далі зчитування даних з файлу відбувається через доступ до ключа 'messages' у зчитаних даних.

Далі скрипт обробляє кожне повідомлення в месенджері, витягуючи текст повідомлення та дату, коли воно було відправлено. Оброблені дані записуються в CSV-файл з іменем 'res.csv'. Для цього використовується бібліотека csv.

Нарешті, скрипт виводить список всіх повідомлень, які були зчитані з файлу.

Щоб запустити скрипт, необхідно викликати функцію parse та передати їй ім'я файлу для обробки. Файл можна передати як аргумент командного рядка при запуску скрипту.

Цей скрипт було створено в середовищі PyCharm, але може бути використаний в будь-якому іншому редакторі Python-коду.

Додаток 6

У цьому додатку ви можете перейти за посиланням, зайти в папку Додаток 6 і продивитись базу даних, яка була створена в лабораторії комп'ютерної лінгвістики.

https://drive.google.com/drive/u/0/folders/1opZZQHT_ie_LLM_nI-iJHbeN19dnjtqA

Додаток 7

Список новоутворених слів/словосполучень у текстах воєнного часу:

1. Свинособака
2. Руський мір
3. Свинособача армія
4. Путінська мерзота
5. Рашист
6. Бомжо-армія
7. Русак
8. Жирік
9. Орк
10. Русняві питушки
11. Пукін
12. Дегенератська федерація
13. Руснява сволота
14. Мухосранск
15. Руснявий
16. Руснява сволота
17. Рашистський виродок

18. Фашист рф
19. Ручний пьос
20. Армія свинособак
21. Путінська орда

Додаток 8

Частотність слів у корпусі текстів воєнного періоду

орк	23
рашист	19
руський мір	15
сука	15
русня	12
свинособака	10
підарас	9
пукін	8
свинособача армія	7
піздец	7
мразота	7
блядь	6
рашка	6
тварюка	5
кончений	5
русак	4
тварь	4
бомжо-армія	3
пропагандон	3
кацап	3
русский мір	3
"руська орда"	2

свиня	2
виблядок	2
хуйло	2
зомбі	2
сволота	2
дибіл	2
пиздюк	2
мухосранск	2
мухосранськ	2
ручний пьос	2
армія свинособак	2
виродки	2
окупант	1
кат	1
ірод	1
москаль	1
Могуча російська армія	1
щур	1
дно	1
"визволителі"	1
підараст	1
іди на хуй	1
гандон	1
кадирівець	1
піздєц	1
путінська мерзота	1
блядьский терорист	1

картопляний фюрер	1
хуй	1
х*рня	1
бандерівець	1
асвабадітєль	1
отримати піздюля	1
тупий	1
косоглазий дебіл	1
жирік	1
відсосати	1
низькопробне походження	1
мразь	1
бомжо армія рф	1
нахуй	1
укронацысты	1
покидьок	1
кончена мерзота	1
кремлебот	1
іти нахуй	1
в'їбати	1
клоун	1
русняві питушки	1
пиздить	1
мерзота	1
овочебаза	1
пиздец	1

дегенератська федерація	1
кончене бидло	1
Пизду#те наху#	1
свинособача техніка	1
руснява сволота	1
йобаний	1
кацапський урод	1
піздєцовий	1
руснявий	1
тупоголовий	1
бомжо армія	1
убогі бляді	1
рашистський виродок	1
кончений зомбак	1
рашистський	1
ху# смоктати	1
шизік	1
заїбати	1
кончені тварі	1
пошёл нахуй	1
Ху#ло	1
путінська орда	1
тварина	1
гівно в голові	1
бидло	1
гнида	1

питушари	1
бандеровци	1
блядські морди	1
бля	1
блядський(блядська війна)	1
Ху# вам за щоку	1
говнорашка	1
подонки вонючие	1
пид..расты еб..чие	1
залупа	1

Додаток 9

Повні коди програмної реалізації для автоматичного визначення мови ворожнечі

За цим посиланням можна ознайомитись з тренувальними моделями https://drive.google.com/file/d/1R-clUoxx-Dt-Qk3UV5J4CODnYRtFZnpg/view?usp=drive_link

А за цими посиланнями можна побачити реалізацію графічного інтерфейсу https://drive.google.com/file/d/1SefVsEAtKNwxZpflIefZTF3NxRS1vu9x/view?usp=drive_link

https://drive.google.com/file/d/1gwzEMMwIb_ToK3hPxjG9ZN-4CF30EX6l/view?usp=drive_link

Всі коди наведені у папці “Автоматичне визначення мови ворожнечі” https://drive.google.com/drive/folders/150FAEmGoFelRbtlIfLOH24Gt7kNwhC-Y?usp=drive_link