Державна служба України з надзвичайних ситуацій
Львівський державний університет безпеки життєдіяльності
Навчально-науковий інститут цивільного захисту
Кафедра інформаційних технологій та систем електронних комунікацій

«Допущено до захисту»
Начальник кафедри ІТтаСЕК
підполковник служби цивільного
захисту
_____ Олександр ПРИДАТКО
“____” лютого 2024 року

КВАЛІФІКАЦІЙНА РОБОТА МАГІСТРА

на тему **«Розроблення системи виявлення небезпечних повідомлень у месенджерах на основі методів розпізнання та класифікації текстової інформації»**

Виконав:
здобувач VI курсу, групи КН-61м
спеціальності (освітньої програми)
122 «Комп’ютерні науки» (Комп’ютерні науки)
(шифр і назва спеціальності (освітньої програми)
Іван-Роман Іванішин
(ім’я та прізвище)
Керівник Олександр Хлевной
(ім’я та прізвище)

Рецензент _____
(прізвище та ініціали)

Львів – 2024 року

Державна служба України з надзвичайних ситуацій
Львівський державний університет безпеки життєдіяльності
Навчально-науковий інститут цивільного захисту
Кафедра інформаційних технологій та систем електронних комунікацій
Освітньо-кваліфікаційний рівень магістр
Спеціальність 122 «Комп'ютерні науки»

ЗАТВЕРДЖУЮ
Начальник кафедри ІТтаСЕК
кандидат технічних наук
доцент
_____ Олександр ПРИДАТКО
"21" листопада 2023 року

ЗАВДАННЯ
на кваліфікаційну роботу

Здобувачу Івану-Роману ІВАНІШИНУ
(прізвище, ім'я, по батькові)

1. Тема Розроблення системи виявлення небезпечних повідомлень у месенджерах на основі методів розпізнання та класифікації текстової інформації.

керівник роботи Олександр ХЛЕВНОЙ, к.т.н.
(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)

затверджені наказом ЛДУ БЖД від "20" листопада 2023 року № НС-144/90.
2. Термін подання здобувачем роботи 5 лютого 2024 року.
3. Початкові дані до роботи

1. Відкриті набори даних для різних областей, таких як Kaggle, UCI Machine Learning Repository, або Open Data portals.
2. "Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit" - Стівен Бердж, О'Рейлі Медіа, 2009.
3. "Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS" - Гарі Вінсі, Джі Гі Хо, Сас Інститут, 2013.

4. Зміст кваліфікаційної роботи/проекту (перелік питань, які потрібно розробити)

Вступ
Розділ 1. Аналіз літературних джерел
Розділ 2: Аналіз методів та матеріалів
Розділ 3. Експерементальні осліджження якості роботи моделей
Розділ 4: Обговорення результатів дослідження
Висновки
Список використаних джерел
Додатки

5. Консультанти розділів роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв

6. Дата видачі завдання _____

КАЛЕНДАРНИЙ ПЛАН

	Назва етапів виконання дипломної роботи/проекту	Термін виконання етапів роботи	Примітка
	Аналіз літературних джерел		
	Аналіз методів та матеріалів		
	Експериментальні дослідження якості роботи моделей		
	Обговорення результатів дослідження		

Здобувач _____ _Іван-Роман Іванішин___
(підпис) (прізвище та ініціали)

Керівник роботи _____ _Олександр ХЛЕВНОЙ_____
(підпис) (прізвище та ініціали)

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ І СКОРОЧЕНЬ

BoW – Bag-of-Words
TF – Term frequency
IDF – Inverse Document Frequency
TF-IDF – Term Frequency-Inverse Document Frequency
SVM – Support Vector Machine

АНОТАЦІЯ

Іванішин Іван-Роман «Розроблення системи виявлення небезпечних повідомлень у месенджерах на основі методів розпізнання та класифікації текстової інформації» Дипломна робота за спеціальністю 122 “Комп’ютерні науки” складається з текстової частини, що містить 4 розділи, 82 с., 22 рис., 20 джерела.

Об’єкт дослідження: повідомлення месенджерів.

Предмет дослідження: виявлення воєнно заборонених повідомлень месенджерів.

Мета кваліфікаційної роботи: визначення ефективності алгоритмів класифікації, для ідентифікації заборонених повідомлень.

Отож ми зосередимося на одній із таких технік як **NLP Text classification**. А саме завдання, які ми реалізовуватимемо на кваліфікаційній роботі, зі стратегічної точки зору, будуть:

- Проаналізувати датасет та визначити правильні кроки етапу попереднього обробки даних (Data Preprocessing stage), саме для нашого датасету та які важливі для задачі класифікації.
- Порівняти різні технології виділення ознак (feature extraction techniques) та вибрати ту, яка найкраще підійде для нашої поставленої задачі.
- Використати різні методи класифікації, та визначити для них такі гіперпараметри, які дадуть їм найкращу точність.

МІШОК СЛІВ, ТЕРМІН ЧАСТОТА, ІНВЕРСІЙНА ЧАСТОТА ДОКУМЕНТА, ТЕРМІН ЧАСТОТА-ІНВЕРТА ДОКУМЕНТА, МАШИНА ОПОРНИХ ВЕКТОРІВ

ABSTRACT

Ivanyshyn Ivan-Roman "Development of a system for detecting dangerous messages in messengers based on methods of recognition and classification of textual information" Diploma work in the specialty 122 "Computer science" consists of a text part containing 4 chapters, 82 pp., 22 figures, 20 sources.

The object of the study: messages of messengers.

The subject of the study: detection of war-prohibited messenger messages.

The purpose of the qualification work: to determine the effectiveness of classification algorithms to identify prohibited messages.

So, we will focus on one such technique as **NLP Text classification**. And the tasks that we will implement in the qualification work, from a strategic point of view, will be:

- Analyze the dataset and determine the correct steps of the Data Preprocessing stage, specifically for our dataset and which are important for the classification task.
- Compare different feature extraction techniques and choose the one that best suits our task.
- Use different classification methods and define hyperparameters for them that will give them the best accuracy.

BAG-OF-WORDS, TERM FREQUENCY, INVERSE DOCUMENT FREQUENCY, TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY, SUPPORT VECTOR MACHINE

ВСТУП

У нинішню епоху інформації велика кількість джерел, таких як соціальні медіа, урядові та промислові операції, блоги та відеоблоги, генерують величезні обсяги даних. Більшість комунікацій відбувається через відео та текстові дані, причому останні, зазвичай, походять із таких джерел, як інтернет-ЗМІ, блоги, публікації в соціальних мережах, повідомлення в загальних чи приватних чатах. Таким чином, підприємства, військові чи уряди прагнуть структурувати ці невпорядковані дані та тримати над ними контроль. Проте, задача аналізу тексту, кількість якої росте експоненційно, з різних джерел інформації є дуже складною для людини, яка потребує багато часу, сил та ресурсів.

На щастя, за останні роки штучний інтелект революціонізував наш підхід до автоматизації. Використовуючи машинне навчання, обробку природної мови та інші методи штучного інтелекту, ми можемо навчити машини виконувати найрізноманітніші завдання, які інакше потребували б втручання людини. Одним із таких завдань є визначення тематики деякого тексту, а саме виявлення воєнно заборонених повідомлень месенджерів.

Актуальність обраного дослідження:

Класифікація тексту має безліч корисних застосувань у різних сферах. У Обробці природної мови (NLP) класифікація тексту використовується для аналізу настрою, визначення теми тексту, категоризації документів чи чи здійснення автоматичного підсумовування. Також методи класифікації тексту застосовуються для пошуку та індексування інформації, вони допомагають створювати ефективні пошукові системи, рекомендувати відповідні статті або контент на основі аналізу текстової інформації. Ще одна галузь , де використовується класифікація- це соціальні науки та гуманітарні дослідження для аналізу соціальних медіа, літературних текстів, класифікація допомагає виявляти тематичні та емоційні тенденції, проводити аналіз контенту та досліджувати культурні аспекти.

Зараз, коли соціальні мережі є дуже поширеним джерелом пошуку інформації, однією з найбільш актуальних сфер застосування класифікації є виявлення спаму та

небажаного контенту, оскільки на даний момент в Україні вже більше року йде повномасштабна війна , виявлення воєнно заборонених повідомлень месенджерів є актуальним завданням.

Новизна обраного дослідження:

Проведення дослідження та аналізу методів класифікації , які я обрав, а саме: k-найближчих сусідів (k-NN) , Decision Tree Classifier, та Random Forest Classifier допоможе мені зробити аргументований підсумок щодо вибору актуальних та ефективних методів для задачі визначення тематики тексту, зокрема для виявлення воєнно заборонених повідомлень у месенджерах. Результати моєї дослідницької роботи будуть корисні у багатьох галузях науки, які я описав вище, де використовується класифікація тексту.

Об'єкт дослідження: повідомлення месенджерів.

Предмет дослідження: виявлення воєнно заборонених повідомлень месенджерів.

Мета кваліфікаційної роботи: визначення ефективності алгоритмів класифікації, для ідентифікації заборонених повідомлень.

Отож ми зосередимося на одній із таких технік як **NLP Text classification**. А саме завдання, які ми реалізовуватимемо на кваліфікаційній роботі, зі стратегічної точки зору, будуть:

- Проаналізувати датасет та визначити правильні кроки етапу попереднього обробки даних (Data Preprocessing stage), саме для нашого датасету та які важливі для задачі класифікації.
- Порівняти різні технології виділення ознак (feature extraction techniques) та вибрати ту, яка найкраще підійде для нашої поставленої задачі.
- Використати різні методи класифікації, та визначити для них такі гіперпараметри, які дадуть їм найкращу точність.

ВИСНОВКИ

Дана кваліфікаційна робота була присвячена визначенню алгоритмів класифікації, для задачі ідентифікації воєнно-заборонених повідомлень. Для цього було використано 3 методи класифікації: : k-найближчих сусідів (k-NN) , Decision Tree Classifier, та Random Forest Classifier. Також для виконання виділення ознак повідомлень було обрано дві техніки: BoW та LSA. Після дослідження отриманих результатів точностей моделей, можна зробити наступні висновки:

1. Decision Tree Classifier продемонстрував найнижчі результаті точностей на тренувальній та тестувальній вибірках, незалежно від попереднього способу векторизації повідомлень, а саме:

- Використовуючи векторизовані повідомлення за допомогою методу Мішка слів з максимальним розміром словника, ми отримали незбалансоване дерево, що сигналізує про погану його здатність враховувати усі передані йому ознаки повідомлень. Також при спробі знайти оптимізований розмір словника BoW, дана модель не змогла показати кращих результатів.
- Однак використовуючи повідомлення зменшеної розмірності за допомогою техніку Латентно-семантичного аналізу, DTC змогла збудувати не тільки збалансоване дерево, але й досягти майже однакової високої точності як модель k-NN з представленням LSA.

Отож для датасету обраного в цій кваліфікаційній роботі, для побудови моделі Дерева рішень, яка показуватиме достатньо хороші результати, необхідно налаштувати його гіперпараметри та надати повідомлення зменшеної розмірності за допомогою техніки LSA, аніж векторизовані методом Мішка слів.

2. Метод k-найближчих сусідів показав низькі результати точностей класифікації на 2-ох вибірках, з переданими йому векторизованими повідомленнями максимальної довжини за допомогою методу Мішка слів. Однак налаштувавши гіперпараметри BoW, ми змогли зменшити розмір словника, що дозволило нашій моделі k-NN збільшити точність на тренувальних та тестувальних даних. Зрештою

передавши моделі повідомлення зменшеної розмірності за допомогою техніки LSA, k-NN змогла перевершила свої результати із векторизатором BoW навіть з оптимізованим розміром словника та показати навіть трішки кращу точність аніж DTC.

Отож для датасету обраного в цій кваліфікаційній роботі, для побудови моделі k-NN, найкраще використовувати векторизовані повідомлення за допомогою BoW, проте попередньо потрібно подбати про гіперпараметри векторизатора та самої моделі k-NN. Також модел найкращу точність модель показує

Отож для датасету обраного в цій кваліфікаційній роботі, для побудови якісної моделі k-NN необхідно налаштувати її параметр k, і найкращим рішенням буде використати повідомлення зменшеної розмірності за допомогою техніки LSA. Однак також можна використовувати векторизовані повідомлення за допомогою BoW, проте попередньо потрібно подбати про гіперпараметри векторизатора.

3. Ансамблевий метод класифікації повідомлень показав найкращі результати серед усіх моделей, незалежно від способу представлення повідомлень. Отож для датасету обраного в цій кваліфікаційній роботі, для побудови якісної моделі Випадкового лісу необхідно налаштувати його гіперпараметри, і використати будь-який з 2-ох представлених методів векторизації.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Wisam Abdulazeez Qader, Musa M. Ameen Bilal I. Ahmed. An Overview of Bag of Words; Importance, Implementation, Applications, and Challenges. URL:https://www.researchgate.net/publication/338511771_An_Overview_of_Bag_of_WordsImportance_Implementation_Applications_and_Challenges
2. Krishna Juluru, Hao-Hsin Shih, Krishna Nand Keshava Murthy, Pierre Elnajjar. Bag-of-Words Technique in Natural Language Processing: A Primer for Radiologists. URL: https://doi.org/10.1148/rg.2021210025
3. Peter W. Foltz. Latent semantic analysis for text-based research. URL:https://link.springer.com/article/10.3758/bf03204765
4. Thomas K Landauer, Peter W. Foltz, Darrell Laham. An Introduction to Latent Semantic Analysis.URL: http://lsa.colorado.edu/papers/dp1.LSAintro.pdf
5. Subhasis Dasgupta. Latent Semantic Analysis and its Uses in Natural Language Processing. URL:https://www.analyticsvidhya.com/blog/2021/09/latent-semantic-analysis-and-its-uses-in-natural-language-processing/
6. Xin Kang, Fuji Ren, Yunong Wu. Exploring latent semantic information for textual emotion recognition in blog articles. URL:https://ieeexplore.ieee.org/document/7833266
7. Deepanshi. Text Preprocessing in NLP with Python Codes. URL:https://www.analyticsvidhya.com/blog/2021/06/text-preprocessing-in-nlp-with-python-codes/
8. Purva Huilgol. Quick Introduction to Bag-of-Words (BoW) and TF-IDF for Creating Features from Text. URL:https://www.analyticsvidhya.com/blog/2020/02/quick-introduction-bag-of-words-bow-tf-idf/
9. Sumit Dua. Text Classification using K Nearest Neighbors. URL:https://towardsdatascience.com/text-classification-using-k-nearest-neighbors-46fa8a77acc5

10. Prshntkmr112. Embedding Techniques on Text Data using KNN. URL:https://www.analyticsvidhya.com/blog/2022/01/embedding-techniques-on-text-data-using-knn/
11. JYOTISHMAN DAS. Spam Text Classification using decision tree.URL: https://www.kaggle.com/code/jyotishmandas/spam-text-classification-using-decision-tree
12. Prabha Karan. Email Spam Classification. URL:https://medium.com/@yesprabhakaran98/email-spam-classification-92b661d3b700
13. Saifuldeen H Abdulrahman, Mohammad Salim. Using Decision Tree Algorithms in Detecting Spam Emails Written in Malay: A Comparison Study.URL: https://www.itm-conferences.org/articles/itmconf/pdf/2022/02/itmconf_icacs2022_01001.pdf
14. Giorgos Orphanos, Dimitris Kalles, Thanasis Papagelis and Dimitris Christodoulakis. Decision Trees and NLP: A Case Study in POS Tagging. URL:https://faculty.washington.edu/fxia/courses/LING572/decison_tree99.pdf
15. Anshul Saini. Decision Tree Algorithm – A Complete Guide. URL:https://www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm/
16. Bernard Kurka. Natural Language Processing and Random Forest Classifier. URL:https://levelup.gitconnected.com/classifying-reddit-posts-with-natural-language-processing-and-random-forest-classifier-af2d8fa77bd3
17. Sruthi E R. Understand Random Forest Algorithms With Examples. URL:https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/
18. Pavan Vadapalli. Random Forest Vs Decision Tree: Difference Between Random Forest and Decision Tree.URL: https://www.upgrad.com/blog/random-forest-vs-decision-tree/
19. Kanish Shah, Henil Patel, Devanshi Sanghvi, Manan Shah. A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification. URL:https://www.researchgate.net/publication/339728543_A_Comparative_Analysis_of_Logistic_Regression_Random_Forest_and_KNN_Models_for_the_Text_Classification
20. Adam Shafi. Random Forest Classification with Scikit-Learn. URL:https://www.datacamp.com/tutorial/random-forests-classifier-python