МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ТЕРНОПІЛЬСЬКИЙ НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
ІМЕНІ ІВАНА ПУЛЮЯ
ФАКУЛЬТЕТ КОМП'ЮТЕРНО-ІНФОРМАЦІЙНИХ СИСТЕМ І ПРОГРАМНОЇ ІНЖЕНЕРІЇ
КАФЕДРА КОМП'ЮТЕРНИХ НАУК

**СИДОР ВІКТОР ЕДУАРДОВИЧ**

УДК 519.2

# ТЕХНОЛОГІЇ МАШИННОГО НАВЧАННЯ ДЛЯ КЛАСИФІКАЦІЇ LINKEDIN-ПРОФІЛІВ IT-СПЕЦІАЛІСТІВ

Спеціальність 122 «Комп'ютерні науки»

**Автореферат**

магістерської роботи на здобуття

освітньо-кваліфікаційного рівня магістр

Тернопіль
2019

Роботу виконано на кафедрі комп'ютерних наук Тернопільського національного технічного університету імені Івана Пулюя Міністерства освіти і науки України

**Керівник роботи:** кандидат технічних наук, доцент кафедри комп'ютерних наук
**Фриз Михайло Євгенович,**
Тернопільський національний технічний університет імені Івана Пулюя,

**Рецензент:** кандидат технічних наук, доцент, проректор з науково-педагогічної роботи
**Дячук Степан Федорович,**
Тернопільський національний технічний університет імені Івана Пулюя,

Захист відбудеться 28 травня 2019 р. о $9^{.00}$ годині на засіданні екзаменаційної комісії №33 у Тернопільському національному технічному університеті імені Івана Пулюя за адресою: 46001,

## ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

**Актуальність теми роботи**. Із кожним років кількість інформації збільшується і виникає потреба у класифікації, завдяки автоматизованій класифікації текстів та інформації дозволяється забезпечити економію людського ресурсу, заощадити час та фінансові ресурси.

**Мета роботи:** дослідження та пошук рішення для текстової класифікації коротких текстів, порівняння наявних алгоритмів та визначення найкращого варіанту для вирішення проблеми класифікації LinkedIn-профілів методами машинного навчання.

**Об'єкт, методи та джерела дослідження:** Об'єктом дослідження виступає процес класифікації LinkedIn-профілів IT-спеціалістів. Предметом дослідження виступають технології машинного навчання для класифікації коротких тексів.

**Практичне значення отриманих результатів:** Створено класифікатор для класифікації LinkedIn-профілів IT-спеціалістів на основі технологій машинного навчання та визначено найкращий алгоритм для класифікації коротких текстів.

**Апробація.** Окремі результати роботи доповідались на II Міжнародній студентській науково-технічній конференції «Природничі та гуманітарні науки. Актуальні питання», Тернопіль, ТНТУ, 25 – 26 квітня 2019 р.

**Структура роботи.** Робота складається з розрахунково-пояснювальної записки та графічної частини. Розрахунково-пояснювальна записка складається з вступу, 7 частин, висновків, переліку посилань. Обсяг роботи: розрахунково-пояснювальної записки – 158 арк. формату А4, графічна частина – презентація.

## ОСНОВНИЙ ЗМІСТ РОБОТИ

**У вступі** проведено аналіз актуальності та мети роботи, поставлено задачі дослідження.

В розділі **«Інформаційні технології машинного навчання для задач класифікації тексту»**, проведено аналіз задачі класифікації профілів в LinkedIn, здійснено аналіз літературних джерел машинного навчання і видів машинного навчання та доцільність їх використання. Розглянуто текстову класифікаціюи, види текстових класифікаторів. Описано класифікатори, які використані в дипломній роботі.

В розділі **«Обґрунтування вибору інструментів та розробка класифікатора»** розроблено та описано структурну схему інформаційної системи для класифікації LinkedIn-профілів. Реалізовано основні етапи для створення класифікатора. Описано структуру програми класифікатора. Обґрунтовано вибір мови програмування та бібліотек для вирішення поставленого завдання. Створено парсер для вивантаження назв професій із соціальної мережі LinkedIn. Підготовлено дані та створено навчальну базу. Створено класифікатор, із реалізацією дев'яти алгоритмів для текстової класифікації.

В розділі **«Обґрунтування критеріїв вибору оптимального класифікатора»** здійснено порівняння за такими параметрами як: оцінка точності вибору правильної

відповіді, оцінка за часом класифікації та оцінка навчального набору для класифікатора. Здійснено вибір найкращого алгоритму за оцінкою ефективності Парето та за значенням цільової функції.

В розділі **«Спеціальна частина»** описано інтегроване середовищ розробки, описані переваги. Здійснено огляд середовищ розробки із вибором найкращого варіанту. Описано встановлення та налаштування інтегрованого середовища розробки.

В розділі **«Обґрунтування економічної ефективності»** основні техніко-економічні показники побудови інформаційної системи для класифікації LinkedIn-профілів технологіями машинного навчання

В розділі **«Екологія»** проаналізовано моніторинг атмосферного повітря актуальність та мету використання його у повсякденності. Описано статистичний аналіз тенденцій і закономірностей динаміки в екології.

В розділі **«Охорона праці та безпека в надзвичайних ситуаціях»** описано основні ергономічні вимоги до робочого місця користувача ПК. Розглянуто робоче місце користувача та описано для кожного елементу ергономічні вимоги за розмірами та формами меблів у робочому місці. Наведено спрощену модель системи «людина-комп'ютер-середовище» та розглянуті основні чинники, які шкодять здоров'ю людини при роботі за ПК.

**У загальних висновках щодо магістерської роботи** описано результати створення інформаційної системи класифікації LinkedIn-профілів технологіями машинного навчання.

**В графічній частині** приведено результати класифікації алгоритмів машинного навчання та вибір найкращого алгоритму для текстової класифікації.

## ВИСНОВКИ

В результаті виконання дипломної роботи було досліджено технології машинного навчання для створення класифікатора LinkedIn-профілів. Обґрунтовано вибір інструментів для реалізації класифікації, створено парсер, навчальну базу та класифікатор. Проведено порівняння алгоритмів для текстової класифікації, та вибрано найкращий алгоритм класифікації для коротких текстів.

## СПИСОК ОПУБЛІКОВАНИХ АВТОРОМ ПРАЦЬ ЗА ТЕМОЮ РОБОТИ

1. Сидор В. Е. Аналіз особливостей класифікаторів тексту / Сидор В. Е. Тези доповіді на II Міжнародній студентській науково-технічна конференція «Природничі та гуманітарні науки. Актуальні питання». – Тернопіль, ТНТУ, 2019. – с. 403.
2. Сидор В. Е. Обґрунтування вибору інструментів для класифікації тексту / Сидор В. Е. Тези доповіді на II Міжнародній студентській науково-технічна конференція «Природничі та гуманітарні науки. Актуальні питання». – Тернопіль, ТНТУ, 2019. – с. 403.
3. Сидор В. Е. Порівняння методів Naive Bayes, Multinomial Naive Bayes та Bernoulli Naive Bayes для задачі класифікації LinkedIn профілів / Сидор В. Е. Тези доповіді на II Міжнародній студентській науково-технічна конференція «Природничі та гуманітарні науки. Актуальні питання». – Тернопіль, ТНТУ, 2019. – с. 403.

## АНОТАЦІЯ

У дипломній роботі досліджено технології машинного навчання для вирішення проблеми класифікації LinkedIn-профілів IT-спеціалістів, алгоритмами текстової класифікації. Створено інформаційну систему, яка включає парсер, навчальний набір та класифікатор. Здійснено порівняння та вибір найкращого алгоритму класифікації для коротких текстів.

В першому розділі досліджено аналіз проблеми класифікації LinkedIn-профілів. Розглянуто технології машинного навчання, види машинного навчання. Описано текстовий класифікатор, та розглянуто дев'ять алгоритмів класифікації, які використані у дипломній роботі.

В другому розділі описано структурну схему інформаційної системи, обґрунтовано вибір мови програмування та бібліотек для реалізації поставного завдання. Створено та описано процес роботи парсера, навчальної бази, класифікатора.

В третьому розділі здійснено порівняння алгоритмів класифікації та вибрано найкращий класифікатор для класифікації LinkedIn-профілів чи коротких текстів. Вибір ґрунтувався на основі ефективності Парето та цільової функції.



**Ключові слова**: МАШИННЕ НАВЧАННЯ, КЛАСИФІКАЦІЯ, LINKEDIN ПРОФІЛІ



## ANNOTATION

In the thesis the technologies of machine learning for solving the problem of classification of the LinkedIn-profiles for IT specialists, algorithms of text classification are investigated. An information system has been created that includes a parser, a training set and a classifier. Comparison and selection of the best classification algorithm for short texts is made.

The first section analyzes the classification problem of LinkedIn-profiles. Techniques of machine learning, kinds of machine learning are considered. The text classifier is described, and nine classification algorithms used in the thesis are considered.

The second section describes the structural scheme of the information system, the choice of the language of programming and libraries for the implementation of a task is justified. The process of work of the parser, the training base, the classifier has been created and described.

The third section compares the classification algorithms and selects the best classifier for the classification of LinkedIn profiles or short texts. The choice was based on Pareto's effectiveness and the target function.



**Keywords:** MACHINE LEARNING, CLASSIFICATION, LINKEDIN PROFILES