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

ЛЕХАН ЮРІЙ ІГОРОВИЧ

УДК 004.4
УДК 004.85

ОБРОБКА ТА АНАЛІЗ ТЕКСТОВОЇ ІНФОРМАЦІЇ МЕТОДАМИ МАШИННОГО НАВЧАННЯ

121 «Інженерія програмного забезпечення»

Автореферат

дипломної роботи на здобуття освітнього ступеня «магістр»

Тернопіль 2018

Проект виконано на кафедрі програмної інженерії Тернопільського національного технічного університету імені Івана Пулюя.

Керівник проекту: доктор технічних наук, професор
Пастух Олег Анатолійович,
Тернопільський національний технічний університет імені Івана Пулюя

Рецинзент: доктор технічних наук, професор
Лупенко Сергій Анатолійович,
Тернопільський національний технічний університет імені Івана Пулюя

Захист відбудеться 27 грудня 2018 р. о $9^{.30}$ годині на засіданні екзаменаційної комісії №31 у Тернопільському національному технічному університеті імені Івана Пулюя за адресою: 46001, м. Тернопіль, вул. Руська, 56, навчальний корпус №1, ауд. 101.

ЗАГАЛЬНА ХАРАКТЕРИСТИКА ПРОЕКТУ

Актуальність теми проекту. Обробці природномовних текстів приділяється значна увага в розвинутих країнах Європи та США, свідченням цього є виділення величезних коштів на розробку лінгвістичного програмного забезпечення. Велику кількість науково-дослідних програм спрямовано на розвиток лінгвістичних інформаційних систем.

Вигоду застосування аналізу тональності, важко переоцінити. Аналіз тональності дозволяє отримувати корисну інформацію з наборів неструктурованих даних. Сфера використання сентиментального аналізу обмежується лише фантазією автора, і економічною доцільністю застосування даного методу. Наприклад при взаємодії з кінцевими користувачами, цей підхід можна застосовувати для модерації коментарів чи аналізу ставлення клієнтів щодо певного бренду, події або явища.

Мета проекту. Створення моделі для аналізу тональності текстової інформації.

Об'єкт, методи та джерела дослідження. Використання методів машинного навчання для класифікації текстової інформації.

Наукова новизна отриманих результатів:

- досліджено способи розробки моделей для класифікації текстової інформації;
- проведено порівняння алгоритмів класифікації;
- проаналізовано існуючі системи для аналізу тональності тексту;
- проаналізовано способи представлення даних.

Практичне значення отриманих результатів. Розроблена система дозволяє проводити аналіз тональності текстової інформації.

Апробація. Окремі результати роботи були представленні на VII науково-практичній конференції молодих учених та студентів «Актуальні задачі сучасних технологій» 28-29 листопада 2018 року (Тернопіль, Україна)

Структура проекту. Проект складається з розрахунково-пояснювальної записки та графічної частини. Розрахунково-пояснювальна записка складається з вступу, 5 частин, висновків, переліку посилань. Обсяг проекту: розрахунково- пояснювальна записка – 110 арк. формату А4, графічна частина – 12 слайдів.

ОСНОВНИЙ ЗМІСТ ПРОЕКТУ

У вступі проведено аналіз актуальності та мети проекту, поставлено задачі дослідження, наведена наукова новизна та практичне значення одержаних результатів.

В розділі **“Аналіз предметної області”** проводиться огляд літературних джерел з тематики машинного навчання та його застосування у різних сферах людської діяльності; наведено приклади застосування аналізу тональності текстової інформації, здійснено огляд існуючих рішень, обґрунтовано необхідність створення сервісу.

В розділі **“Методи та засоби вирішення проблеми”** проаналізовано та сформульовано вимоги до кінцевої системи; описано методологію проектування системи та здійснено підбір та опис методів реалізації системи, що включають методи попередньої обробки та алгоритми вирішення задачі класифікації.

В розділі **“Розробка та тестування системи”** здійснено підбір технологій реалізації проекту, що включає вибір мови програмування, середовища розробки, бібліотек та фреймворків для обробки даних та побудови моделей машинного навчання; описано варіанти представлення даних; здійснено аналіз, та підготовку вхідних даних; подано деталі реалізації основних компонент системи; описано процес навчання моделей та методологію їх тестування; описано процес інтеграції моделей.

В розділі “**Обґрунтування економічної ефективності**” проведено розрахунок норм часу на виконання дипломного проекту, витрат на електроенергію, суму амортизаційних відрахувань та ціну дослідження. Також визначено витрати на оплату праці, відрахування на соціальні заходи та економічну ефективність і термін окупності капітальних вкладень.

В розділі **“Охорона праці та безпека в надзвичайних ситуаціях”** розглянуто питання облаштування робочих місць для користувачів ПК з урахуванням оптимальних параметрів освітлення, мікроклімату, шуму та вібрації, забезпечення захисту від дії рентгенівського та електромагнітного, ультрафіолетового та інфрачервоного випромінювання. Описано можливості використання комп’ютерної

техніки для оцінки можливої обстановки у випадку надзвичайних ситуації та подано функціональні заходи у сфері державного регулювання та контролю захисту населення і територій.

У загальних висновках щодо дипломного проекту описано результати проектування та розробки системи для аналізу тональності текстової інформації. В додатках до пояснювальної записки наведено зразки програмного коду, а також скриптів для розгортання системи. Додано диск з програмним забезпеченням та пояснювальною запискою до розробки.

ВИСНОВКИ

Даний проект присвячено створенню моделі машинного навчання для аналізу тональності текстової інформації. Проект розпочався з введення в предметну область, що включає ознайомлення з цілями та завданнями для яких застосовується інтелектуальний аналіз тексту. Іншим напрямом аналізу був огляд попередніх досліджень та алгоритмів в області машинного навчання. Також було здійснено аналіз існуючих рішень в контексті аналізу тональності текстової інформації.

В процесі розробки було здійснено аналіз методів представлення даних, описано процес підготовки даних для побудови моделей. Для побудови моделей були використані підходи, що базуються на n-грамному та послідовному представленні даних. В процесі розробки здійснено порівняльний аналіз моделей машинного навчання на основі наступних алгоритмів: наївний баєсів класифікатор, багатошаровий перцептрон, лінійна регресія, метод опорних векторів, AdaBoost, Random Forest та рекурентна нейронна мережа. Для оцінки якості моделі використано гармонійну середню (F1) між точністю (precision) та відгуком (recall) системи з застосуванням перехресної валідації.

В результаті найкращу оцінку для бінарної класифікації отримала модель побудована на основі лінійної регресії, що набрала 0.89. Найкращу для оцінку серед моделей на основі нейронних мереж досягнуто за допомогою рекурентної мережі, точність якої становить 0.98. Результати свідчать про однозначну можливість застосування машинного навчання для сентиментальної класифікації текстової інформації.

СПИСОК ОПУБЛІКОВАНИХ АВТОРОМ ПРАЦЬ ЗА ТЕМОЮ РОБОТИ

1. Лехан Ю.І. Пастух О.А. Обробка та аналіз текстової інформації методами машинного навчання // Тези доповіді на VII науково-практичній конференції молодих учених та студентів «Актуальні задачі сучасних технологій». – Тернопіль, ТНТУ, 2018.

АНОТАЦІЯ

Дипломний проект присвячений розробці моделі для аналізу тональності текстової інформації засобами машинного навчання.

Об’єктом дослідження є використання методів машинного навчання для класифікації текстової інформації.

Предмет дослідження: методи аналізу і обробки текстової інформації та засоби машинного навчання, призначені для забезпечення автоматичної класифікації.

Мета роботи: створення моделі для аналізу тональності текстової інформації.

Система розроблена з допомогою мови програмування Python та з застосуванням бібліотек Keras, NumPy, SciPy, Scikit-Learn, TensorFlow. Система надає API для здійснення аналізу тональності текстової інформації.

Модель побудована на основі лінійної регресії досягла оцінки у 0.89 за критерієм гармонійної середньої між точністю та відкликом (F1) при здійсненні бінарної класифікації. Модель побудована на основі рекурентної нейронної мережі змогла отримати 0.98 за тим же критерієм.

Ключові слова: МАШИННЕ НАВЧАННЯ, НЕЙРОННІ МЕРЕЖІ, ДЕРЕВА ПРИЙНЯТТЯ РІШЕНЬ, ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ТЕКСТУ.

ABSTRACT

Diploma work on theme «Processing and analysis of textual information by methods of machine learning» by student Yurii Lekhan. – Ternopil Ivan Pul'uj National Technical University, Faculty of Computer Information Systems and Software Engineering, Software engineering department, group SPm-61 // Ternopil, 2018.

Pages. – 110, pictures. – 25, tables. – 5, slides – 12, add. – 4

The purpose of the diploma is to create a model which provides maximum efficiency for automatic sentiment analysis of text content. Methods and tools used to develop the system: the Python programming language and its libraries, the development environment of the PyCharm IDE, the development environment, and the Agile methodology of software development.

The result of the work is the application which provides API for sentiment text analysis.

Keywords: TEXT ANALYSIS, MACHINE LEARNING, TF-IDF, RECURRENT NEURAL NETS, PYTHON.