УДК 004.912

А. О. Концевой

ЗАСТОСУВАННЯ МАШИННОГО НАВЧАННЯ ДЛЯ АНАЛІЗУ ТА КАТЕГОРИЗАЦІЇ ТЕКСТОВОЇ ІНФОРМАЦІЇ

Вінницький національний технічний університет

Анотація

Розглянуто підходи до застосування машинного навчання для аналізу та категоризації текстової інформації з соціальних мереж ті інших інтернет джерел.

Ключові слова: машинне навчання, категоризація текстової інформації, соціальні мережі, природно-мовна конструкція, NLTK.

Abstract

This paper considers approaches to the application of machine learning for the analysis and categorization of textual information from social networks and other Internet sources.

Keywords: machine learning, categorization of textual information, social networks, natural-language construction, NLTK.

Вступ

Аналіз тексту в цілому є новою галуззю вивчення. Такі галузі, як маркетинг, управління продуктами, наукові дослідження та управління, вже використовують процес аналізу та вилучення інформації з текстових даних [1]. Класифікація тексту або категоризація тексту - це діяльність по позначенню текстів природною мовою відповідними категоріями із заздалегідь визначеного набору [2].

Метою даного дослідження є створення та тестування моделі машинного навчання для категоризації текстової інформації.

Результати дослідження

В ході дослідження застосовувалися мова програмування Python, а також такі допоміжні бібліотеки як Pandas, Scikit-learn та, TextBlob.

В даному досліджені було використано набори даних (корпуси), що містить різні огляди на платформі Amazon, Yelp та IMDB. Для підготовки набору даних, дані були завантажені у фрейм даних pandas, що містить два стовпці - текст та мітку.

Наступним кроком було перетворення необроблених текстових даних у вектори об’єктів, а нові функції будуть створені за допомогою наявних бібліотек. Реалізовано вектори об`єктів з нашого набору даних.

Завершальним кроком в рамках дослідження є створення базової моделі. Необхідно розділити дані на навчальний і тестувальний набір, які дозволяють оцінити точність і перевірити, чи добре працює модель. Модель мусить добре працювати з даними, яких вона раніше не бачила, а не тільки на тестових даних.

Модель має визначати, чи рекомендує користувач товар, книгу, фільм тощо аналізуючи текст його відгуку з наявних наборів тестових даних.

В ході дослідження застосовано таку модель як логістична регресія, яка є простою, але потужною лінійною моделлю, яка є формою регресії від 0 до 1 на основі вхідного вектора ознак. Вказуючи граничне значення (за замовчуванням 0,5), для класифікації використовується регресійна модель.

На рисунку 1 зображено точність моделі логістичної регресії

```
>>> from sklearn.linear_model import LogisticRegression

>>> classifier = LogisticRegression()
>>> classifier.fit(X_train, y_train)
>>> score = classifier.score(X_test, y_test)

>>> print("Accuracy:", score)
Accuracy: 0.796
```

Рисунок 1 - Точність моделі класифікації

Логістична регресія досягла вражаючих 79,6% точності, але необхідно перевірити, як ця модель працює на інших наборах даних, які ми маємо. Точність моделі - коефіцієнт співвідношення правильно оброблених результатів о їх загальної кількості.

У цьому сценарії ми виконали та оцінили весь процес для кожного з наборів (корпусів) даних, які було обрано для даного дослідження. Для цього було створено скрипт, що автоматизує цей процес, проганяючи кожен з тестових наборів даних на тренованій моделі та виводить точність у вікно консолі.

На рисунку 2 зображено точність моделі при обробці кожного з початкових наборів даних.

```
Accuracy for yelp data: 0.7960
Accuracy for amazon data: 0.7960
Accuracy for imdb data: 0.7487
```

Рисунок 2 - Точність моделі при обробці кожного з початкових наборів даних.

Висновки

Було розглянуто особливості та підходи до класифікації текстової інформації, а також перспективи його застосування в реальному світі. Дослідження довело ефективність обраного підходу до класифікації текстової інформації, а також визначена точність тренованої моделі на декількох наборах тестових даних.

Подальшим розвитком даного дослідження є застосування нейронних мереж. Необхідно провести порівняльний аналіз існуючих моделей нейронних мереж для вибору відповідної до дослідження типу нейронної мережі.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Shami M. Text Mining: Classification, Clustering, and Applications / M. Sahami. — Chapman and Hall/CRC, 2009. — 328 с.
2. Klein E. Social Network Analysis: Methods and ApplicationsNatural Language Processing with Python: Analyzing Text with the Natural Language Toolkit / E. Klein, S. Bird. — O'Reilly Media, 2009. — 803 с.

Концевой Антон Олександрович – студент 2-го курсу аспірантури, спеціальність 126 - Інформаційні технологій та системи, факультет комп'ютерних систем та автоматики, Вінницький національний технічний університет, м. Вінниця

Науковий керівник: ***Бісікало Олег Володимирович*** – д-р техн. наук, професор, декан факультету КСА, Вінницький національний технічний університет, м. Вінниця

Kontsevoi Anton O. –2nd year graduate student, specialty 126 - Information Technology and Systems, Faculty of Computer Systems and Automation, Vinnytsia National Technical University, Vinnytsia

Supervisor: ***Bisikalo Oleh V.*** – Dr.Sc. (Eng.), Professor, Dean of the Faculty for Computer Systems and Automatic, Vinnytsia National Technical University, Vinnytsia