УДК 004.89
Цимбалюк Г. – ст. гр. СНнм-61
*Тернопільський національний технічний університет імені Івана Пулюя*

# ПОПЕРЕДНЯ ОБРОБКА ТЕКСТОВИХ ДАНИХ ДЛЯ ЗАДАЧ КЛАСИФІКАЦІЇ

Науковий керівник: к.т.н., доцент Никитюк В.В.

Tsymbaliuk G.
*Ternopil Ivan Puluj National Technical University*

# TEXT PREPROCESSING FOR CLASSIFICATION PROBLEMS

Supervisor: PhD, associated professor Nykytiuk V.

Ключові слова: ТОКЕНІЗАЦІЯ, ЛЕМАТИЗАЦІЯ, СТЕММІНГ, СТОП-СЛОВА, НОРМАЛІЗАЦІЯ
Key words: TOKENIZATION, LEMATIZATION, STEMING, STOP-WORDS, NORMALIZATION

Класифікація текстових даних є однією з основних задач в обробці природної мови (NLP). Вона включає в себе розподіл текстів по різних категоріях або класах на основі їх змісту. Задачі класифікації текстів широко застосовуються в таких сферах, як автоматичне сортування електронних листів (спам/не спам), категоризація статей, аналіз тональності (емоційного забарвлення) тексту, а також в машинному перекладі та виявленні тем в документах. Всі ці задачі вимагають точного розуміння тексту і правильного віднесення до відповідної категорії.

В [1] наведено наступні етапи класифікації текстових даних:

- Text sources: це перший етап, на якому текстові дані збираються з різних джерел, таких як вебсайти, документи, соціальні мережі або API.
- Corpus format: в контексті класифікації текстів визначає структуру та організацію текстових даних, що використовуються для навчання моделей.
- Text-preprocessing: включає в себе кроки попередньої обробки тексту, такі як токенізація, видалення стоп-слів, стемінг або лематизація для підготовки даних до аналізу.
- Exploratory text analysis: на цьому етапі застосовуються різні методи статистичного або лінгвістичного аналізу, щоб виявити патерни, ключові слова або теми.
- Vectorizing: текст перетворюється на числові вектори (наприклад, через TF-IDF або Word2Vec), що дозволяє моделі працювати з ним у числовому вигляді.
- Modeling: на цьому етапі застосовуються алгоритми машинного навчання для тренування моделі, яка буде класифікувати або прогнозувати на основі текстових даних.
- Assessment: перевіряється ефективність моделі на тестових даних через метрики, такі як точність, recall, F1-оцінка, щоб оцінити її здатність до класифікації.
- Vizualization - на останньому етапі результати аналізу візуалізуються через графіки, діаграми або хмарки слів, щоб зручніше зрозуміти висновки з даних.

Попередня обробка текстових даних є критично важливою на етапі класифікації, оскільки вона дозволяє зменшити шум в даних і покращити якість

навчання моделі. Попередня обробка дозволяє позбутися таких елементів, а також знижує розмірність вхідних даних, що допомагає зменшити складність моделі та покращити її ефективність. Без належної обробки навіть найкращий алгоритм класифікації може дати не точні результати.

Основні методи попередньої обробки текстових даних включають в себе токенізацію, видалення стоп-слів, стемінг і лематизацію, а також нормалізацію тексту.

Токенізація тексту — це процес розділення тексту на окремі одиниці, звані токенами, що можуть бути словами, фразами або іншими значущими елементами. Це перший і важливий етап попередньої обробки текстових даних в обробці природної мови. Токенізація дозволяє перетворити текстовий документ на набір елементів, з якими можна працювати далі, наприклад, підраховувати частоти слів або використовувати їх для моделювання.

Нормалізація тексту передбачає перетворення всіх слів в нижній регістр, видалення знаків пунктуації тощо.

Стоп-слова – це слова, які найчастіше зустрічаються в тексті і не містять жодної цінної інформації, такі як: they, there, this, where тощо. Бібліотека nltk - це звичайна бібліотека Python, яка використовується для видалення стоп-слів і включає приблизно 180 стоп-слів англійської мови.

Стемінг — це процес зведення слів до їх базових коренів (наприклад, "running" → "run"), що дозволяє зменшити варіативність словоформ.

Лематизація — це процес перетворення слів до їх початкової або лексичної форми (наприклад, "running" → "run", "better" → "good"), з врахуванням контексту та граматичних правил.

Автори [2] показали, що ефект від певного методу може значною мірою залежати від початкового набору даних і задачі класифікації. Крім того, зазначено, що іноді вплив може бути незначним, а в деяких випадках навіть призвести до погіршення результату. Також була обґрунтована необхідність попередньої перевірки набору даних на відсоток елементів, які підпадають під вплив конкретного методу.

Автори [3] покзали, що правильно підібрані методи попередньої обробки текстових даних можуть суттєво покращити якість класифікаційних моделей, проте потребують ретельного дослідження в кожній конкретній ситуації.

Проведене дослідження для задачі визначення автора тексту показало несподіваний ефект: стемінг та лематизація не сприяли підвищенню точності класифікаційних моделей, натомість зменшення простору ознак на 90% лиш незначно вплинуло на якість і точність моделей класифікації тексу за автором.

**Література.**

1. Lamba Manika & Margam, Madhusudhan. (2022). Text Pre-Processing. Text Mining for Information Professionals: An Uncharted Territory (pp.79-103) Edition: 1. Chapter: 310.1007/978-3-030-85085-2_3.
2. Orlovskyi, O., & Ostapov, S. (2020). Analysis of the text preprocessing methods influence on the destructive messages classifier. Advanced Information Systems, 4(3), 104–108. https://doi.org/10.20998/2522-9052.2020.3.14
3. Haddi, Emma & Liu, Xiaohui & Shi, Yong. (2013). The Role of Text Pre-processing in Sentiment Analysis. Procedia Computer Science. 17. 26–32. 10.1016/j.procs.2013.05.005.