УДК 004.9

О. А. Довгань

ІТ АНАЛІЗ ТА ВИЯВЛЕННЯ ПЕРСОНАЛЬНОЇ ІНФОРМАЦІЇ СТУДЕНТІВ У ТЕКСТОВИХ ДАНИХ

Вінницький національний технічний університет

Анотація

Дана робота присвячена огляд кореляції між частинами мови й токенами персональних даних студентів у першій таблиці та штучно створеними ознаками й токенами персональних даних студентів у другій таблиці.

Ключові слова: *кореляція, токени, таблиці.*

Abstract

This work is dedicated to examining the correlation between parts of speech and tokens of personal data of students in the first table, as well as artificially created features and tokens of personal data of students in the second table.

Key words: *correlation, tokens, tables.*

Вступ

Кореляція - це статистичний показник, який описує силу та напрямок зв'язку між двома змінними. Вона вказує на те, наскільки та яким чином дві змінні змінюються разом. Коефіцієнт кореляції кількісно визначає ступінь відповідності змін в одній змінній змінам у іншій змінній. Він коливається від -1 до +1, де кореляційний коефіцієнт, близький до +1, вказує на сильну позитивну кореляцію, що означає, що зі збільшенням однієї змінної інша змінна також тенденційно збільшується. Кореляційний коефіцієнт, близький до -1, вказує на сильну негативну кореляцію, що означає, що зі збільшенням однієї змінної інша змінна тенденційно зменшується. Кореляційний коефіцієнт, близький до 0, вказує на малу або відсутню лінійну залежність між змінними [1].

У даній роботі досліджується кореляція між елементами двох таблиць.

Результати досліджень

Мета роботи - це виявлення закономірностей між персональними даними студентів у текстових даних. Текст персональних даних студентів поділяється на токени, де для токенізації тексту використовувався SpaCy English tokenizator . Даний токенізатор розбиває текст по пробілах [2], як наслідок наступний текст:”The morning is starting at the Even Green Teres tomorrow” буде токенізовано наступним чином [‘The’, ‘morning’, ‘is’, ‘starting’, ‘at’, ‘the’, 'Even’, ‘Green’, ‘Teres’, ‘tomorrow’]. І він отримає наступні значення токенів [O, O, O, O, O, O, B-STREET_ADDRESS, I-STREET_ADDRESS, I-STREET_ADDRESS, O]. Токен O можемо ігнорувати, тоді нас цікавить токени B-STREET_ADDRESS, I-STREET_ADDRESS, а саме, їх частини B-, I-, де B- початок токена, а I- це продовження або кінець токена.

На рисунку 1 зображена кореляція частин мови та спеціальних символів, що позначається на рисунку як X, з токенами персональних даних студентів [3].

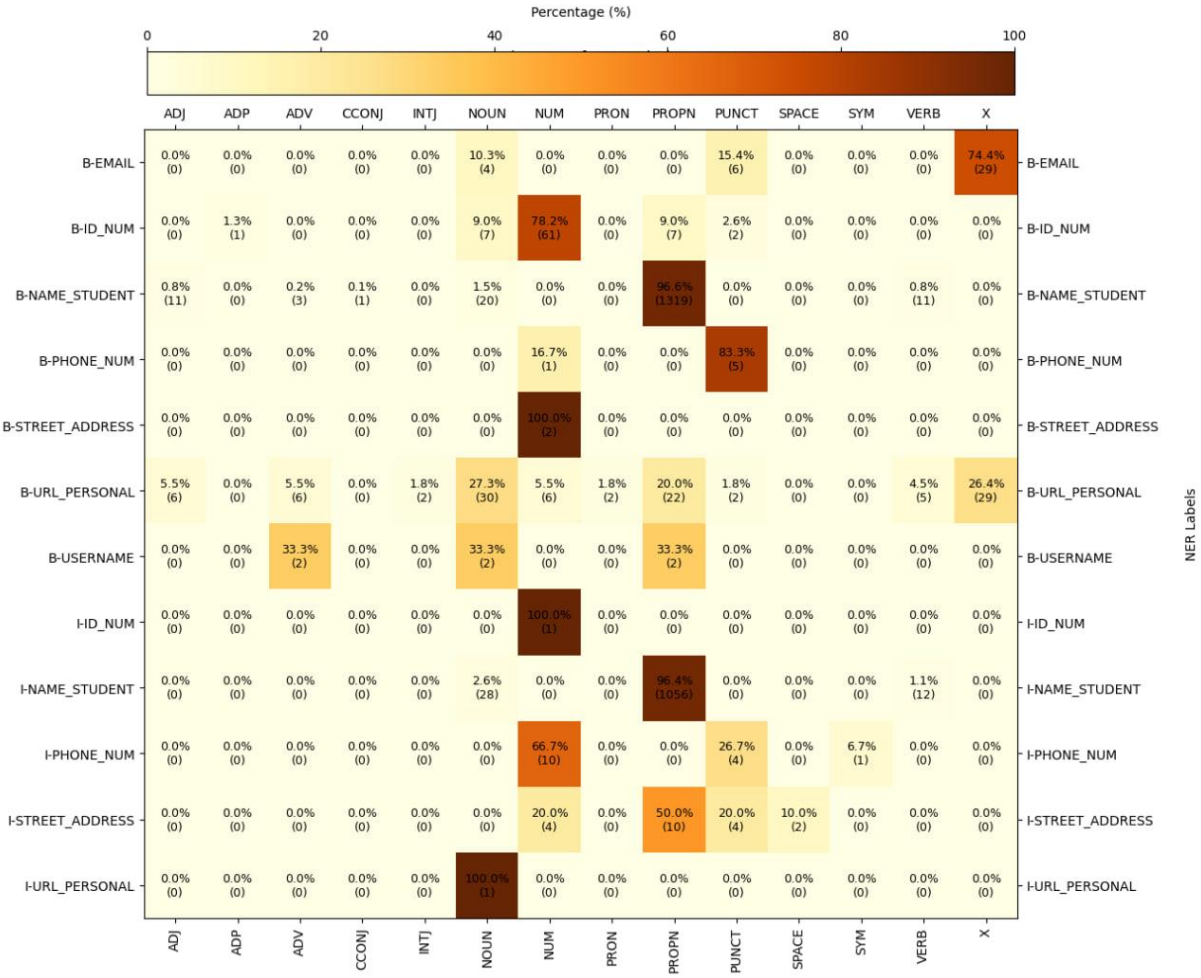

Рисунок 1 – Таблиця кореляції між частинами мови та токенами персональних даних студентів

З рисунку можемо бачити, що токен B-EMAIL істотно корелюється з спеціальними символами, до яких входить символ ‘@’. Також, даний токен корелюється з колонкою PUNCT, що свідчить про наявність пунктуації в емейлах. Наступний токен ID_NUM корелюється з колонкою NUM, NOUN, ADP, PROPN. Тут можна помітити наступну закономірність, якщо I-ID_NUM повністю корелюється з NUM та складається виключно з цифер, то B-ID_NUM частоково містить числа, а частково слова.

Токени NAME_STUDENT корелюються з NOUN та PROPN. Це свідчить про те, що імена студентів – це не лише іменники, а й власні назви. PHONE_NUM корелюються з NUM, PUNCT, SYM. Але кореляція B-PHONE_NUM і I-PHONE_NUM у колонках NUM, PUNCT обернено пропорційна, що може говорити про те, що номер починається з пунктуаційного знаку, наприклад з ‘+’.

Токени STREET_ADDRESS мають сильну кореляцію з колонкою NUM, що свідчить про те, що адреси зазвичай містять числа, де ці числа відповідають за номери будинків. Існує також суттєва кореляція з PROPN для власних назв, які відповідають назвам вулиць, назвам міст. I-STREET_ADDRESS показує помірну кореляцію з ADJ, що свідчить про те, що прикметники також використовуються для опису адрес.

Для токенів URL_PERSONAL існує помітна кореляція з колонкою NUM як для токенів типу B-, так і для токенів типу I-, що свідчить про наявність чисел у обох частинах URL-адресах; кореляція з PROPN, що свідчить про власні назви в URL-адресах. Колонка X також корелюється з URL-адресами особливо для I-URL_PERSONAL, що свідчить про символи або послідовності токенів, які не відповідають традиційним категоріям частин мови, наприклад, спеціальні символи, які зазвичай зустрічаються у URL-адресах (наприклад, косі знаки, крапки, дефіси). Для токенів типу B-PHONE_NUM та I-PHONE_NUM спостерігається сильна кореляція з колонкою NUM. Це очікуваний результат, оскільки номери телефонів складаються з послідовностей чисел.

Токен I-PHONE_NUM має невелику кореляцію з колонкою PUNCT, що може означати, що пунктуаційні знаки часто зустрічаються у межах номерних телефонів.

Токени USERNAME демонструють значну кореляцію з NOUN та NUM, що, ймовірно, відображає той факт, що імена користувачів часто містять як слова, так і числа. Також спостерігається кореляція з категорією X, аналогічна URL-адресам та адресам електронної пошти, щоб врахувати використання нестандартних символів, які типові для імен користувачів, таких як підкреслення чи крапки. Категорія частин мови PROPN також виявляється, що свідчить про включення в імена користувачів власних імен або імен.

На рисунку 2 зображена кореляція токенів персональних даних студентів з штучно створеними ознаками, кожна з яких репрезентує певну властивість токенів у тексті. До штучно створених ознак можна віднести location_in_essay, pii_length, special_chars, token_alphabetic, token_alphanumeric, token_numeric, token_other, capitalization_lowercase, capitalization_titlecase, capitalization_uppercase, position_beginning, position_middle [4].

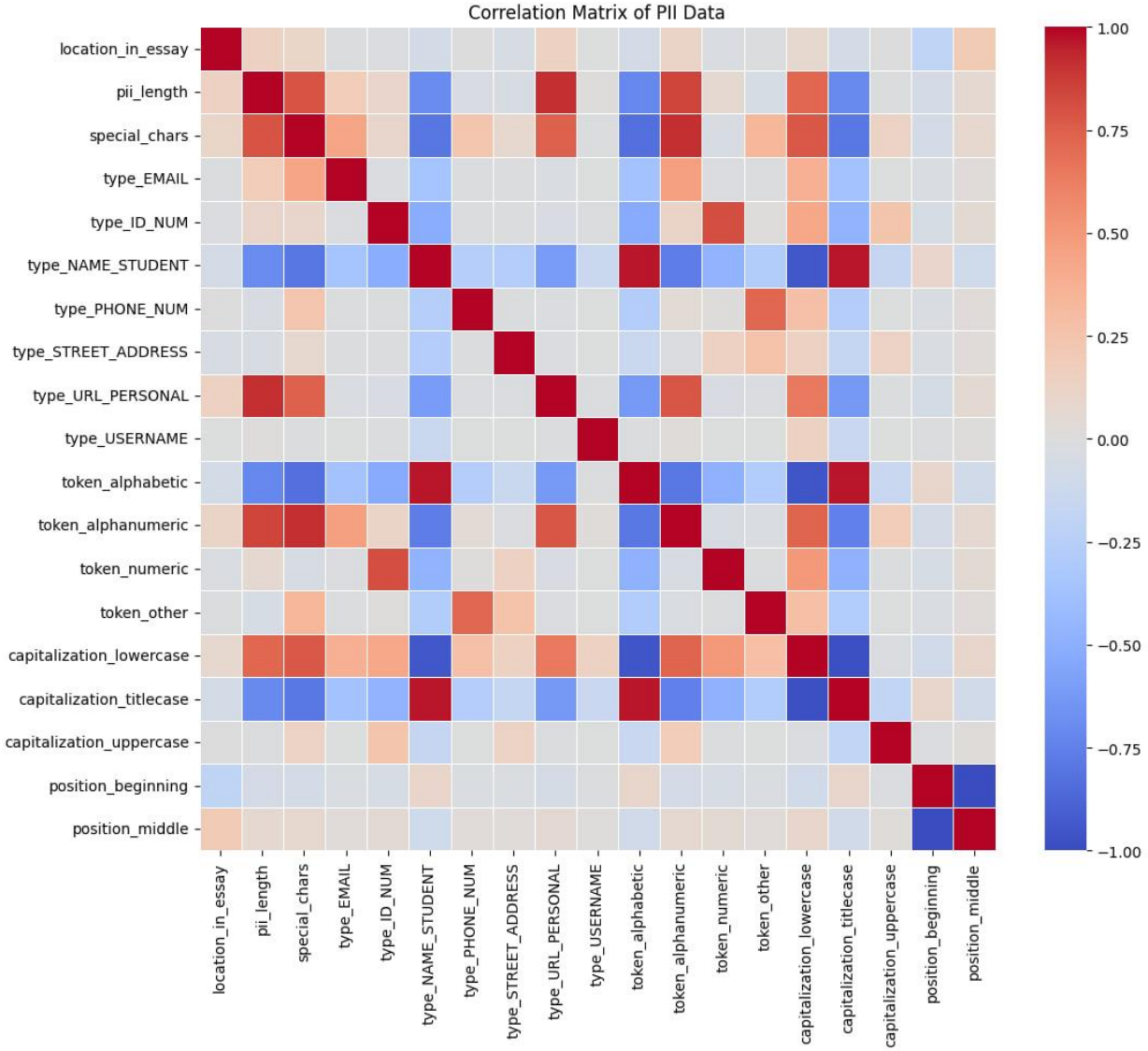

Рисунок 2 - Таблиця кореляції між штучно створеними ознаками та токенами персональних даних студентів

Таблиця 1.1 - Опис штучно створених ознак для кореляції між токенами персональних даних

Назва ознаки	**Опис ознаки**
location_in_essay	Місце в есе
pii_length	Довжина послідовності токенів, які відносяться до однієї групи
special_chars	Наявність спеціальних символів, таких як @, & і тд
token_alphabetic	Токени, які складаються виключно з букв
token_alphanumeric	Токени, які містять і букви, і цифри

token_numeric	Токени, які складаються з цифр
token_other	Токени, які не входять у групу, що позначають персональні дані студентів
capitalization_lowercase	Токен, які повністю складаються з малої літери
capitalization_uppercase	Токен, які повністю складаються з великої літери
capitalization_titlecase	Токени, які починаються з великої літери
position_beginning	Розташування токена на початку послідовності
position_middle	Розташування токена в середині послідовності

Токени type_EMAIL сильно корелюються з колонками special_chars, token_alphanumeric, capitalization_lowercase через те, що електронні адреси містять спеціальні символи, такі як "@" та ".", складаються з букв і цифр та зазвичай починаються з малої букви. Також є помірна кореляція з колонкою pii_length, це можна пояснити тим, що адреси відрізняються за довжиною, але часто, вони довші за інші PII типи.

Токени type_ID_NUM мають сильну позитивну кореляцію з колонками token_numeric і pii_length, оскільки ідентифікаційні номери зазвичай складаються з чисел та мають фіксовану довжину. Та помірну кореляцію з колонками capitalization_lowercase, capitalization_uppercase оскільки частина цих токенів складається виключно з малих літер, а частина – з великих.

Токени type_NAME_STUDENT мають сильну позитивну кореляцію з колонками token_alphabetic, capitalization_titlecase, тому що складаються виключно з букв і починаються з великої літери. Та помірну кореляцію з колонкою position_beginning, що свідчить про розташування токенів даного типу на початку есе.

Токени type_PHONE_NUM мають сильну позитивну кореляцію з колонками special_chars, token_other та з capitalization_lowercase. Кореляцію з першою колонкою можна пояснити тим, що номери складаються з спец символів, таких як дужки, знак "+", а кореляцію з третьою колонкою – всі токени цієї групи складаються з малих символів.

Токени type_URL_PERSONAL сильно корелюються з колонками special_chars, token_alphanumeric, capital_lowercase та pii_length, тому що в першому випадку містять спеціальні символи, у другому випадку є комбінацією літер, цифр і в третьому випадку не чутливі до регістру та представлені малими літерами, а в четвертому випадку – довжина токенів даної групи однакова.

Токени type_USERNAME помірна позитивна кореляція з колонкою capitalization_lowercase, оскільки нік-нейм користувачів зазвичай написаний маленькими літерами.

Висновок

У даній роботі була розглянута кореляція між частинами мови й токенами персональних даних студентів у першій таблиці та штучно створеними ознаками й токенами персональних даних студентів у другій таблиці.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Лекція 13. Кореляція. Коефіцієнт кореляції [Електронний ресурс] – Режим доступу: https://teta.at.ua/statustuka/lekcija13.pdf
2. spaCy Tokenizer [Електронний ресурс] – Режим доступу: https://www.educba.com/spacy-tokenizer/
3. PII Data Detection EDA [Електронний ресурс] – Режим доступу: https://www.kaggle.com/code/snassimr/pii-data-detection-eda?scriptVersionId=161001892&cellId=33
4. TLAL PII Data Detection EDA & Learn With Me [Електронний ресурс] – Режим доступу: https://www.kaggle.com/code/dschettler8845/tlal-pii-data-detection-eda-learn-with-me?scriptVersionId=163235580&cellId=30

Довгань Олексій Андрійович — магістр кафедри системного аналізу та інформаційних технологій, Вінницький національний технічний університет, Вінниця, e-mail: odovhan08@gmail.com

Dovhan Oleksii Andriovich - master of the department of system analysis and information technologies, Vinnytsia National Technical University, Vinnytsia, e-mail: odovhan08@gmail.com