

КЛАСИФІКАЦІЯ ЖАРТІВ ЗА КАТЕГОРІЯМИ ГУМОРУ З ВИКОРИСТАННЯМ МЕТОДІВ МАШИННОГО НАВЧАННЯ

Вінницький національний технічний університет

Анотація

Дослідження присвячено задачі автоматичної класифікації жартів (анекдотів) за категорією гумору із застосуванням методів машинного навчання. Запропоновано підхід, що передбачає попередню векторизацію текстів жартів із використанням TF-IDF та балансування вибірки за допомогою SMOTE. Для розв'язання задачі класифікації застосовано логістичну регресію, яка навчалася на синтетично розширених даних. Проведено експерименти на корпусі з чотирьох категорій гумору: абсурд, каламбур, чорний гумор та сарказм. Результати класифікації демонструють високу точність за умови належної підготовки даних і правильного налаштування гіперпараметрів моделі.

Ключові слова: класифікація жартів, машинне навчання, логістична регресія, SMOTE, TF-IDF, Python.

Abstract

The study is devoted to the problem of automatic classification of jokes (anecdotes) by humour category using machine learning methods. An approach is proposed that involves pre-vectorization of joke texts using TF-IDF and sample balancing using SMOTE. To solve the classification problem, logistic regression is used, which is trained on synthetically enhanced data. Experiments were conducted on a corpus of four categories of humour: absurdity, pun, black humour and sarcasm. The classification results demonstrate high accuracy provided that the data is properly prepared and the model hyperparameters are correctly adjusted.

Keywords: joke classification, machine learning, logistic regression, SMOTE, TF-IDF, Python.

Вступ

Сучасна комп'ютерна лінгвістика активно розвивається в напрямі обробки текстів природною мовою, включаючи аналіз стилістичних особливостей текстів різних жанрів. Однією з актуальних задач у цій галузі є автоматизований аналіз гумористичних текстів, який має як теоретичне, так і прикладне значення [1] – зокрема для розпізнавання емоційного забарвлення, жанру та лінгвістичної структури повідомлень у соціальних мережах, чат-ботах або в системах рекомендацій.

Гумор як мовне явище характеризується складною семантичною та прагматичною структурою, його автоматизоване розпізнавання потребує врахування не лише лексико-семантичних ознак, а й жанрових характеристик. Актуальність визначення гумористичних ознак в природномовному контенті демонструє наявність численних сучасних підходів до автоматичного розпізнавання гумору [2]. У межах даної роботи реалізовано програмну систему, що здійснює класифікацію україномовних жартів за чотирма жанровими категоріями: абсурд, каламбур, темний гумор та саркастичний гумор. Для розв'язання цієї задачі використано методи машинного навчання, зокрема логістичну регресію, TF-IDF-векторизацію тексту, а також алгоритм SMOTE для балансування навчальної вибірки.

Метою даної роботи є розробка інтелектуальної моделі для автоматизованого аналізу гумористичних текстів українською мовою, зокрема їх класифікації за жанровими ознаками, із застосуванням сучасних методів обробки природної мови [3]. Запропонований підхід дозволяє не лише підвищити точність автоматичного визначення жанру гумору, а й створює основу для подальших досліджень у сфері стилістичного аналізу та гумористичної семантики української мови.

Результати дослідження

Для розв'язання задачі класифікації жартів за категоріями було реалізовано послідовність етапів, які охоплюють збір початкових даних, їх попередню обробку, векторизацію тексту, балансування вибірки, навчання моделі, тестування та аналіз отриманих результатів. Зокрема:

1. *Збір і підготовка вибірки даних.*

Було сформовано корпус текстів, що охоплює чотири категорії гумору: абсурдний, каламбур, темний гумор і сарказм. Для кожної категорії підготовлено два текстові файли: один для навчання (наприклад, *absurd1.txt*), інший – для тестування (наприклад, *absurd2.txt*). Кожен файл містив добірку коротких жартів (анекдотів), по одному на рядок. Обсяг тренувальної вибірки становив близько 100–200 жартів на категорію, тоді як для тестування у файлах містилось по 9 жартів на кожен категорію. Категоріям було призначено числові мітки (label): абсурд — 0, каламбур — 1, темний гумор — 2, сарказм — 3 [3].

2. *Завантаження та валідація даних.*

За допомогою функції `load_dataset` всі жарти було зчитано у структуру `DataFrame` бібліотеки *pandas* [4], де кожен рядок містив текст жарту та його відповідну числову мітку. Було проведено перевірку на наявність порожніх або некоректних значень, а також виконано первинний аналіз розподілу класів у тренувальній та тестовій вибірках [5].

3. *Векторизація тексту.*

Для перетворення текстових жартів у числові ознаки застосовано метод TF-IDF (Term Frequency — Inverse Document Frequency) за допомогою класу `TfidfVectorizer` із бібліотеки *scikit-learn* [5]. Використано параметри: видалення англійських стоп-слів, уніграми і біграми, обмеження на 5000 ознак для контролю розмірності.

4. *Балансування вибірки.*

З метою усунення дисбалансу класів застосовано метод SMOTE (Synthetic Minority Over-sampling Technique) [6], який синтезує нові приклади для менш представлених категорій, покращуючи якість навчання.

5. *Навчання моделі.*

Для класифікації застосовано логістичну регресію з регуляризацією L2 і розв'язувачем *liblinear* [7]. Модель тренувалася на збалансованій вибірці, якість оцінено за допомогою 5-кратної крос-валідації.

6. *Збереження моделі.*

Після навчання модель та векторизатор збережено у файлах `.joblib` для подальшого використання без повторного тренування.

7. *Тестування моделі.*

Модель протестовано на окремій тестовій вибірці, результати аналізовано через матрицю плутанини та класифікаційний звіт з метриками точності, повноти і F1-міри.

8. *Передбачення на нових прикладах.*

Модель також продемонструвала ефективність на нових, рандомно сформованих або введених користувачем у програму жартів, підтверджуючи здатність розпізнавати жанрові ознаки гумору.

Усі розглянуті етапи логічно структуруються у вигляді блок-схеми алгоритму (рис. 1), що візуалізує послідовність дій від завантаження даних до отримання прогнозу.

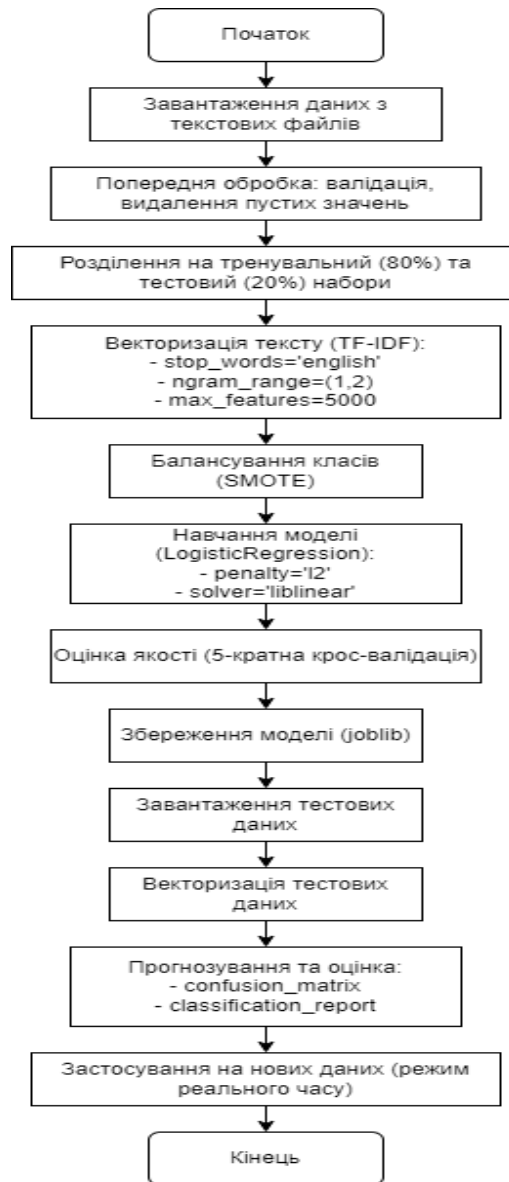


Рис.1 – Блок-схема алгоритму класифікації жартів

Для програмної імплементації запропонованого підходу було застосовано технологію Python / pandas / scikit-learn / imblearn / joblib / os. На рисунках 2 та 3 наведено приклади розпізнавання розробленою програмою жартів за типом гумору.

Жарт: Твоя скромність просто сліпить. Наче прожектор у лоб.
Категорія жарту: Саркастичний гумор

Рис. 2 – Виведення результату класифікації для внесеного користувачем жарту: визначено категорію «сарказм»

Жарт: "Усміхайся!" – казали вони. А потім ще спитали: "Чому ти такий дивний на фото з похорону?"
Категорія жарту: Абсурд

Рис.3 – Виведення результату класифікації для внесеного користувачем жарту: визначено категорію «абсурд»

Висновки

У даній роботі було розроблено модель та реалізовано програму для автоматичної класифікації жартів за жанровими ознаками, що дозволяє відносити короткі гумористичні тексти до таких категорій, як абсурд, каламбур, темний гумор і сарказм. Застосування методів машинного навчання, зокрема логістичної регресії, у поєднанні з векторизацією тексту за допомогою TF-IDF і балансування вибірки через SMOTE, продемонструвало ефективність підходу для аналізу стилістичних характеристик гумору. Побудована модель досягла стабільних показників точності, що підтверджується результатами тестування відповідної програми та класифікаційними метриками. Важливою перевагою програми є можливість застосування моделі для обробки нових, раніше не бачених прикладів у режимі реального часу. Отримані результати підтверджують, що навіть базові алгоритми машинного навчання здатні виявляти семантичні й стилістичні патерни в текстах, якщо забезпечити належну підготовку даних і збалансовану вибірку. Це відкриває перспективи для подальшого вдосконалення систем автоматичного аналізу стилю, включаючи розширення категорій гумору, використання більш складних архітектур (наприклад, трансформерів), а також дослідження культурних і лінгвістичних особливостей сприйняття жартів.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Attardo S. Humor and Cognitive linguistics. In The Routledge Handbook of cognitive linguistics 2021 Jun 3 (pp. 359-371). Routledge.
2. Шолота, В.В., Бісікало О.В. Огляд сучасних підходів до автоматичного розпізнавання гумору. Матеріали XLVIII науково-технічної конференції підрозділів ВНТУ, Вінниця, 13-15 березня 2019 р.
3. Інтелектуальна модель стилістичного аналізу гумористичних текстів [Електронний ресурс] // Архів відкритого доступу НУ «ХАІ». – 2023. – Режим доступу: <https://openarchive.nure.ua/entities/publication/d733c3c5-23b3-4b80-b732-430574b69648>, вільний.
4. McKinney W. Python for Data Analysis. 2nd Edition [Текст]. – O'Reilly Media, 2018. – 544 с.
5. Jurafsky D., Martin J.H. Speech and Language Processing. 3rd ed. – Draft [Електронний ресурс]. – 2020. – Розділ: Vector Semantics and TF-IDF. – Режим доступу: <https://web.stanford.edu/~jurafsky/slp3/>, вільний.
6. Chawla N.V., Bowyer K.W., Hall L.O., Kegelmeyer W.P. SMOTE: Synthetic Minority Over-sampling Technique [Текст] // Journal of Artificial Intelligence Research. – 2002. – Т. 16. – С. 321–357.
7. Pedregosa F. et al. Scikit-learn: Machine Learning in Python [Електронний ресурс] // Journal of Machine Learning Research. – 2011. – Т. 12. – С. 2825–2830. – Режим доступу: <https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html>, вільний.

Бісікало Олег Володимирович – д-р техн. наук, професор, завідувач кафедри АІТ, Вінницький національний технічний університет, м. Вінниця, e-mail: obisikalo@vntu.edu.ua

Урлапова Дар'я Павлівна – студентка групи ІСТ-236, факультет автоматизації та інтелектуальних інформаційних технологій, Вінницький національний технічний університет, Вінниця, e-mail: dashaurlapova@gmail.com

Телега Денис Ігорович – студент групи ІСТ-226, факультет автоматизації та інтелектуальних інформаційних технологій, Вінницький національний технічний університет, Вінниця, e-mail: denistelega09@gmail.com

Bisikalo Oleg V. – Dr.Sc. (Eng.), professor, Head of the Department of Automation and intelligent information technologies, Vinnytsia National Technical University, Vinnytsia, e-mail: obisikalo@vntu.edu.ua

Urlapova Daria P. – student of the Faculty for Intelligent information technologies and automation, Vinnytsia National Technical University, Vinnytsia, e-mail: dashaurlapova@gmail.com

Tieliega Denys I. – student of the Faculty for Intelligent information technologies and automation, Vinnytsia National Technical University, Vinnytsia, e-mail: denistelega09@gmail.com