Міністерство освіти і науки України
Національний технічний університет
«Дніпровська політехніка»

Факультет інформаційних технологій
(факультет)

Кафедра системного аналізу і управління
(повна назва)

ПОЯСНЮВАЛЬНА ЗАПИСКА
Кваліфікаційної роботи ОКР *Магістра*
(назва освітньо-кваліфікаційного рівня)

студента Симонець Галини Василівни
академічної групи 124м-19-1
напряму підготовки: *124 Системний аналіз*

на тему: *«Застосування алгоритмів машинного навчання для обробки коментарів відеохостингу»*

Керівники	Прізвище, ініціали	Оцінка за шкалою		Підпис
		рейтинговою	інституційною	
Кваліфікаційної роботи	*К.ф-м.н.,доц. Коряшкіна Л.С.*			
розділів:				
Інформаційно-аналітичний	*К.ф-м.н.,доц. Коряшкіна Л.С.*			
Спеціальний	*К.ф-м.н.,доц. Коряшкіна Л.С.*			
Рецензент	*К.т.н.,доц. Шедловський І.А*			
Нормоконтроль	*доц. Малієнко А. В.*			

Дніпро
2020

РЕФЕРАТ

Пояснювальна записка 67 с., 14 малюнків, 10 таблиць, 2 додатка, 22 джерел.

Об'єкт дослідження: відеохостинг, що надає користувачам послуги зберігання, доставки та показу відео – «YouTube».

Предмет дослідження: коментарі під відео на «YouTube» з метою видобування з них знань завдяки віднесенню їх до різних класів.

Мета дослідження: виявлення токсичних коментарів на відеохостінгу "Youtube" шляхом класифікації неструктурованого тексту за допомогою комбінації методів машинного навчання.

Методи дослідження: методи машинного навчання для очищення, нормалізування, представлення текстових даних у вигляді прийнятним для обробки на ЕОМ. Класифікатор логістичної регресії, метод класифікації за допомогою лінійних опорних векторів без та з методом навчання – стохастичний градієнтний спуск, класифікатор «Випадковий ліс» та класифікатор з посиленням градієнта. Алгоритм оцінки роботи класифікаторів, що включає використання методів підрахунку матриці помилок, точності, повноти та Ф-міри для оцінки моделей. Для більш генералізованої оцінки використано метод перехресної перевірки. Мова програмування Python.

Економічна ефективність: очікується позитивною завдяки розробці програмних модулів, які дозволяють автоматизувати процес класифікації коментарів.

В *інформаційно-аналітичному розділі* розглянуто основні базові підходи до обробки тексту і виділено стратегії та їх алгоритми, які можна використати для опрацювання текстових коментарів під відео у «Youtube». Окрім того, описано можливості мови програмування Python.

В *спеціальному розділі* розроблено модуль з вивантаження коментарів з під відео на «Youtube», та програмний комплекс для їх класифікації кількома методами. Проведено аналіз точності моделей за яким обрано найкращі.

Практична цінність отриманих у роботі результатів полягає в оптимізації(спрощені) процесу аналізу коментарів.

Ключові слова: ОБРОБКА ПРИРОДНОЇ МОВИ, НЕСТРУКТОРОВАНІ ДАНІ, КОМЕНТАРІ, КЛАСИФІКАЦІЯ, ЛОГІСТИЧНА РЕГРЕСІЯ, МЕТОД ОПОРНИХ ВЕКТОРІВ, СТОХАСТИЧНИЙ ГРАДІЄНТНИЙ СПУСК, ВИПАДКОВИЙ ЛІС, ПОСИЛЕННЯ ГРАДІЄНТА, МАТРИЦЯ ПОМИЛОК, ПЕРЕХРЕСНА ПЕРЕВІРКА.

ABSTRACT

Explanatory note 67 p., 14 figures, 10 tables, 2 annexes, 22 sources.

The object of research is video hosting that provides users with services for storing, delivering and displaying videos - "YouTube".

The subjects of research: comments on YouTube videos in order to extract knowledge from them by assigning them to different classes.

The purpose of the research: identifying toxic comments on the video hosting "Youtube" by classifying unstructured text using a combination of machine learning methods.

Research methods: machine learning methods for cleaning, normalizing, presenting textual data in a form acceptable for processing on a computer. Logistic regression classifier, linear support vector classification method with and without training method - stochastic gradient descent, "Random forest" classifier and classifier with gradient enhancement. Algorithm for evaluating the work of classifiers, including the use of methods for calculating the matrix of errors, accuracy, completeness and F-measure for evaluating models. For a more generalized assessment, a cross-validation method was used. Python programming language.

Cost-effectiveness: expected to be positive through the development of software modules that automate the comment classification process.

In the information and analytical section, the main basic approaches to text processing are considered and strategies and their algorithms are highlighted that can be used to process text comments for a video in "Youtube". In addition, the capabilities of the Python programming language are described.

In a special section, a module has been developed for uploading comments from a video to "Youtube" and a software package for their classification by several methods. The analysis of the accuracy of the models for which the best are selected is carried out.

The practical value of the results obtained in the work lies in the optimization of the comment analysis process.

Keywords: NATURAL LANGUAGE, NON-STRUCTURED DATA, COMMENTS, CLASSIFICATION, LOGISTIC REGRESSION, SUPPORT VECTOR MACHINE, STOCHASTIC GRADIENT DESCENT, RANDOM FOREST, GRADIENT BOOSTING, CONFUSION MATRIX, CROSS-VALIDATION.