ISSN 2307–3489 (Print), ISSN 2307–6666 (Online)

Наука та прогрес транспорту. Вісник Дніпропетровського національного університету залізничного транспорту, 2020, № 6 (90)

ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНІ ТЕХНОЛОГІЇ ТА МАТЕМАТИЧНЕ МОДЕЛЮВАННЯ

УДК 004.9

Л. С. КОРЯШКІНА[1*], Г. В. СИМОНЕЦЬ[2]

[1*]Каф. «Системний аналіз і управління», Національний технічний університет «Дніпровська політехніка», пр. Д. Яворницького, 19, Дніпро, Україна, 49005, тел. +38 (095) 565 76 83, ел. пошта koriashkina.l.s@nmu.one, ORCID 0000-0001-6423-092X
[2]Каф. «Системний аналіз і управління», Національний технічний університет «Дніпровська політехніка», пр. Д. Яворницького, 19, Дніпро, Україна, 49005, тел. +38 (068) 848 42 00, ел. пошта galya.golovko.2014@gmail.com, ORCID 0000-0002-1322-7189

ЗАСТОСУВАННЯ АЛГОРИТМІВ МАШИННОГО НАВЧАННЯ ДЛЯ ОБРОБКИ КОМЕНТАРІВ ПІД НАВЧАЛЬНИМ МАТЕРІАЛОМ НА ВІДЕОХОСТИНГУ «YOUTUBE»

Мета. Автори мають на меті виявлення токсичних коментарів на відеохостингу «YouTube» під навчальним матеріалом шляхом класифікації неструктурованого тексту за допомогою комбінації методів машинного навчання. **Методика.** У роботі із зазначеним типом даних використано методи попередньої обробки для очищення, нормалізації, представлення текстових даних у вигляді, прийнятному для подальшої роботи на ЕОМ. Безпосередньо для віднесення коментарів до класу «токсичні» використано класифікатор логістичної регресії, метод класифікації за допомогою лінійних опорних векторів без та з методом навчання – стохастичним градієнтним спуском, класифікатор «випадковий ліс» та класифікатор з посиленням градієнта. З метою оцінки роботи класифікаторів використано методи підрахунку матриці помилок, точності, повноти та Ф-міри. Для більш узагальненої оцінки використано метод перехресної перевірки. Мова програмування – Python. **Результати**. На основі показників оцінки обрано найбільш результативні методи – метод опорних векторів (Linear SVM) без та з методом навчання за допомогою стохастичного градієнтного спуску. Описані технології можуть бути використані для аналізу текстових коментарів під будь-якими навчальними відео для виявлення токсичних відгуків. Також розроблений підхід може бути корисним для виявлення небажаної або навіть агресивної інформації в соціальних мережах або сервісах, де передбачені відгуки. **Наукова новизна.** У роботі використано комбінацію методів попередньої обробки специфічного виду тексту із врахуванням таких особливостей, як можливість наявності таймкодів, емоджі, посилань тощо, а також адаптовано класифікаційні методи машинного навчання для аналізу російськомовних коментарів. **Практична значимість.** Проведено оптимізацію (спрощення) процесу аналізу коментарів, необхідність якої обумовлена зростаючими обсягами текстових даних, особливо у сфері освіти через карантинні умови й перехід на дистанційну форму навчання. Обсяги навчального інтернет-контенту вже потребують автоматизації процесу обробки й аналізу відгуків із часом ця потреба тільки зростатиме.

Ключові слова: обробка природної мови; неструктуровані дані; коментарі; класифікація; логістична регресія; метод опорних векторів; стохастичний градієнтний спуск; «випадковий ліс»; посилення градієнта; перехресна перевірка

doi: https://doi.org/10.15802/stp2020/225264

ISSN 2307–3489 (Print), ISSN 2307–6666 (Online)

Наука та прогрес транспорту. Вісник Дніпропетровського національного університету залізничного транспорту, 2020, № 6 (90)

Вступ

Часто обставини вносять свої корегування в планомірний розвиток різних галузей людської діяльності. Наразі через карантинні умови зросла потреба в стрімкому інтегруванні інформаційних технологій у сферу освіти. Як результат, велику частину освітнього контенту почали оцифровувати, що породило низку коментарів. Із часом кількість коментарів тільки зростатиме, у наслідок чого їх опрацювання й аналіз можна буде здійснювати лише за допомогою різних засобів автоматизації цих процесів.

Соціальні медіа стали унікальним місцем для вільного висловлювання людьми своєї думки. Тим часом серед користувачів є групи, які зловживають цією свободою для реалізації свого токсичного мислення (образи, словесні сексуальні домагання, нецензурні висловлювання тощо). Система нагляду за ризикованою поведінкою молоді 2017 року (Центри контролю та профілактики захворювань) підрахувала, що за 12 місяців до опитування 14,9 % старшокласників зазнавали електронних знущань. Тому розумне використання науки про дані здатне сформувати більш здорове середовище для віртуальних суспільств [12].

Коментарі – це текст, і в машинному навчанні вони мають назву «неструктуровані дані». Одним із найбільших джерел потенційно цікавих і важливих неструктурованих даних є сервіс «YouTube», який наповнений невичерпною кількістю навчальних відео і коментарів до них, які швидко зростають

На хостингу наявні відео інформаційних та комерційних гігантів, які вже мають навчальний контент. Для них і не тільки аналіз коментарів потенційно може допомогти розробникам навчального матеріалу отримати інформацію про його доступність, якість викладання, а також оцінити лояльність користувачів до продуктів та аудиторії, виявити нагальні потреби в знаннях і заохочення нових клієнтів.

Мета

Основною метою цього проекту є виявлення токсичних коментарів на відеохостингу «YouTube» під навчальними матеріалами шляхом класифікації неструктурованого тексту за допомогою комбінації методів машинного навчання. Для цього насамперед потрібно проаналізувати наявні підходи й методи, які використовують під час обробки коментарів не тільки на зазначеному сервісі, але й взагалі в соціальних мережах.

Методика

У нашій роботі розглянуто основні базові підходи до обробки тексту й виділено стратегії та алгоритми, які можна використати для опрацьовування текстових коментарів під відео в «YouTube». Окрім того, розроблено сервіс із вивантаження коментарів під відео на «YouTube», комплекс програм для їх класифікації на основі комбінації декількох методів інтелектуального аналізу даних.

Сьогодні зростаючі обсяги текстових даних переходять рубіж тих, як можна обробляти вручну. Отже, розробка методів та засобів, що автоматизують і пришвидшують обробку й аналіз текстової інформації, є актуальним напрямом досліджень у галузі інформаційних технологій.

Основні надбання за цією темою. Проблемі попередньої обробки неструктурованих даних присвячені роботи [10–11]. Так, у [10] показано, що проблема великої попередньої обробки та збільшення даних може бути вирішена неявно капсульними мережами. Авторам вдалося досягти показника якості бінарної класифікації AUC у 98,46 для набору даних коментарів Kaggletoxic і продемонструвати, що він перевершує інші архітектури з великим відривом.

У [11] зазначено, що як модель попередньої підготовки ультрасучасної мовної моделі BERT (Bidirectional Encoder Reprezentations from Transformers) досягла приголомшливих результатів у багатьох завданнях із розуміння мови. У цій роботі проведено вичерпні експерименти для вивчення різних методів точного налаштування BERT на завдання класифікації тексту та запропоновано загальне рішення.

Ще у 2011 році науковці Елхам Хабірі, Джеймс Каверлі та Чіао-Фанг Хсу з технічного університету в Техасі на п’ятій міжнародній конференції з вебжурналів та соціальних медіа в Барселоні (Іспанія) презентували свій підхід до аналізу коментарів на «YouTube». Вони ви-

doi: https://doi.org/10.15802/stp2020/225264

ISSN 2307–3489 (Print), ISSN 2307–6666 (Online)

Наука та прогрес транспорту. Вісник Дніпропетровського національного університету залізничного транспорту, 2020, № 6 (90)

користали декілька прийомів для подолання часу обробки коментарів: (1) кластеризацію даних для визначення корельованих груп коментарів та (2) структуру ранжування на основі пріоритетів для автоматичного вибору інформативних коментарів, доданих користувачами. Дослідження включає кластеризацію коментарів та вибір найбільш репрезентативних серед них у кожному кластері. Запропонований метод ранжування на основі пріоритетів у поєднанні з кластеризацією на основі тематичних показників продемонстрував вищі показники порівняно з традиційними підходами (наприклад, LexRank, MEAD) [9]. Схематично означені підходи зображено на рис. 1.

Рис. 1. Підходи до аналізу коментарів, використані в роботі Елхам Хабірі

Fig. 1. Approaches to the analysis of comments used in the work of Elham Khabiri

Було виявлено, що поєднання цих двох основних характеристик дає багатообіцяючі результати. Зокрема, було оцінено запропонований алгоритм реферування коментарів для колекції відео «YouTube» та пов'язаних із ними коментарів і виявлено хорошу продуктивність порівняно з традиційними підходами реферування документів.

Майкл Чарі разом із колегами у 2014 р. опублікував статтю «Ознаки та симптоми впливу декстрометорфану з "YouTube"«, у якій за допомогою комп'ютерних технік аналізу тексту коментарів виявив, що інформація, надана коментаторами на «YouTube», корисна для розкриття токсикологічних ефектів DXM [6].

Серед використаних технік: TF–IDF – вилучення ключових слів, підрахунок частоти слів за плато, та WordNet – графічне зображення англійської мови, яке об'єднує слова з подібними значеннями в кластери, що приблизно відповідають поняттям. Графічне зображення WordNet у вказаній роботі використано для виявлення подібних відгуків про препарат.

Іван Кассельман та Майкл Генріх за допомогою вмісту контенту на «YouTube» дослідили повідомлення про використання та враження користувачів від препарату під назвою «Salvia divinorum» [4].

У статті «Ієрархічна кластеризація на основі коментарів» Елхам Хабірі, Джеймс Каверлі та Чіао-Фанг Хсу пропонують ієрархічний підхід до кластеризації коментарів, який спирається на дві ключові особливості: (1) нормалізацію термінів коментарів та вилучення ключових термінів, щоб відігнати галасливі коментарі для ефективної кластеризації; (2) компонент вставки в режимі реального часу для поступового оновлення ієрархії, заснованої на коментарях, із тим, щоб ресурси могли ефективно розміщуватися в ієрархії по мірі виникнення коментарів, без необхідності повторного генерування (потенційно) дорогої ієрархії. Тут вивчено підхід кластеризації на сайті обміну відео «YouTube» [7].

У роботі [2] запропоновано виявляти коментарі та пости, що є образливими, за допомогою методів глибокого навчання. При цьому використано набір даних токсичних коментарів Kaggle для навчання моделі та класифікації коментарів за такими категоріями: токсичні, дуже токсичні, непристойні, погрози, образи й ненависть до особистості. Застосовуючи різні методи глибокого навчання й аналізуючи результати, автори надають рекомендації щодо вибору найкращої моделі глибокого навчання для класифікації тих чи інших коментарів.

Розробці автоматизованої системи аналізу будь-якого фрагменту тексту та виявлення різних типів токсичності присвячена робота [5]. У ній використано маркований набір даних коментарів у Вікіпедії, підготовлений Jigsaw. Авторам вдалося навчити модель, забезпечуючи середню точність перевірки 98,08 % та абсолютну точність перевірки 91,64 %.

Детальний огляд інших сучасних методів машинного навчання, які застосовують для класифікації токсичності коментарів у соціальних мережах, зробив Дарко Андрочеку [3].

doi: https://doi.org/10.15802/stp2020/225264

ISSN 2307–3489 (Print), ISSN 2307–6666 (Online)

Наука та прогрес транспорту. Вісник Дніпропетровського національного університету залізничного транспорту, 2020, № 6 (90)

Класифікація. Для класифікації необхідно мати розмічені дані. Тому в роботі використано базу даних токсичних коментарів, у якій токсичні дані розмічені цифрою 1.0, а нетоксичні – 0.0. Доступ та вивантаження здійснено з відкритого сайту IT змагань Kaggle.

Попередню обробку тексту здійснено таким чином. За допомогою регулярних виразів модуля emoji розроблено дві функції для очищення тексту: функція «give_emoji_free_text» видаляє смайли, які не використовуються як дані; функція «normalize_document» виконує видалення таймкодів, посилань, пунктуації, зайвих пробілів та переходів на інший рядок, виправлення слів із літерами, що декілька разів дублюються, наприклад, «яяясно» до «ясно». Також у цій функції виконують зведення до нижнього регістра, видалення стоп-слів та токенізацію за допомогою модуля nltk.word_tokenize. Нагадаємо, токенізація – це розбиття речень на одиниці даних (слова) [1].

Наступним кроком в обробці текстових даних є стемінг або лематизація (Stemming or lemmatization). Обидва методи використовують для зменшення розмірності ознак шляхом об'єднання слів, наприклад, «бігти», «біжу», «бігли» в одне «біг» тощо. У нашій роботі застосовано останній, адже стемінг передбачає відсічення кінцівок слова, що часто може змінити його значення, особливо це стосується слів, які використовують в інтернет-словнику. Лематизація спрацьовує для відомих для неї слів, а інші залишає незмінними. Це допомагає обійти проблему неправильної заміни. Лематизацію виконано завдяки модулю StanzaLanguage, розробленому вченими Стенфордського університету.

Описані дії дають змогу зменшити розмірність векторів даних для поліпшення розуміння їх машиною [7].

Після попередньої обробки інформації проводять векторизацію ознак, або конструювання ознак, для чого використовують TfidfVectorizer.

Tfidf – feature engineering модель – це функціональна інженерна модель, яка заснована на використанні частоти ознаки в документі (реченні) (the term frequency-based feature engineering model).

Функція TfidfVectorizer від Scikit-Learn дозволяє нам безпосередньо обчислювати вектори tfidf, беручи необроблені документи як вхідні дані та обчислюючи частоти термінів у документі (в коментарі), а також обернені частоти документів (зворотній показник частоти термінів в усьому корпусі документів). Передбачено також підтримку додавання n-грамів до векторів функцій. У результаті описаної обробки замість речень отримуємо вектори ознак для кожного коментаря.

Використання класифікаторів. Безпосередньо для віднесення коментарів до класу «токсичні» використано п'ять класифікаційних моделей: класифікатор логістичної регресії, метод класифікації за допомогою лінійних опорних векторів без та з методом навчання, застосовуючи стохастичний градієнтний спуск, а також класифікатор «випадковий ліс» і класифікатор з посиленням градієнта [3, 8].

Оцінка якості класифікаторів. Для оцінки якості моделей навчання було побудовано матрицю помилок (отриману інформацію представлено на рис. 2). Для кожної моделі підраховано: кількість правильно визначених моделлю нетоксичних, помилково визначених токсичних коментарів, помилково визначених нетоксичними коментарів та правильно визначених як нетоксичні коментарів. За останнім і найважливішим для нашого аналізу показником вірно визначених токсичними коментарів, а також невеликою кількістю помилкових результатів можна встановити, що найкращими в передбаченні обох класів є методи опорних векторів. Адже вони мають по 1 139 та 1 128 вірно визначених токсичних коментарів із 1 570 наявних, так що цей показник є вищим порівняно, наприклад, із машиною підсилення градієнта. Окрім того, за допомогою методу опорних векторів 3 004 та 3 016 коментарів визначено правильно як нетоксичні (із 3 184 можливих). Для більш детального аналізу й порівняння застосованих методів класифікації використано різні метрики їх оцінки.

Так, було розраховано метрики оцінки моделей класифікаторів, такі як точності, повноти та Ф-міри, наведені на рис. 3:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}; \qquad (1)$$

doi: https://doi.org/10.15802/stp2020/225264

ISSN 2307–3489 (Print), ISSN 2307–6666 (Online)

Наука та прогрес транспорту. Вісник Дніпропетровського національного університету залізничного транспорту, 2020, № 6 (90)

$$Precision = \frac{TP}{TP + FP}; \quad (2)$$

$$Recall = \frac{TP}{TP + FN}; \quad (3)$$

$$F = 2\frac{Precision \times Recall}{Precision + Recall}. \quad (4)$$

У формулах (1) – (4) *TP* – істинно позитивні рішення; *TN* – істинно негативні рішення; *FP* – хибно-позитивні рішення; *FN* – хибно негативні рішення системи. Ці ж позначення застосовано на рис. 2.

Із рис. 1 видно, що для наших даних, найбільш придатні методи опорних векторів адже вони мають показники точності 87 %.

Оцінку класифікаційної моделі прийнято проводити шляхом перехресної перевірки (CV Score). Сутність цієї оцінки описано в інформаційному розділі. Якщо коротко, це метод оцінки аналітичної моделі та її поведінки на незалежних даних. Під час оцінювання моделі наявні дані розбивають на *k* частин. Потім на *k*–1 частинах даних проводять навчання моделі, а частину даних, що залишається, використовують для тестування. Процедуру повторюють *k* разів; у результаті кожен із *k* елементів даних використовують для тестування.

У такий спосіб отримаємо оцінку ефективності обраної моделі з найбільш рівномірним використанням наявних даних. Результати такої оцінки, коли *k* = 5, було консолідовано на рис. 4.

Чим вищий показник перехресної перевірки, тим кращою є модель. У нашому випадку – це метод опорних векторів (Linear SVM) без та з методом навчання з використанням стохастичного градієнтного спуску. Також указані методи демонструють найвищі інші метрики – точність, повноту та Ф-міру.

Після навчання, тренування та оцінки найбільш оптимальну модель використано на інших незалежних нерозмічених даних, завантажених із коментарів «YouTube».

Після того як побудовано робочу модель, останнім кроком є розгортання моделі, що, як правило, передбачає збереження моделі та необхідних залежностей, а також її розгортання як служби, API або як запущеної програми. Існують різні способи розгортання моделей машинного навчання, і це зазвичай залежить від того, як ви хочете отримати до нього доступ пізніше. У нашій роботі параметри моделі зберігаються на носії за допомогою модуля pickle, завдяки якому зручно використовувати в програмному середовищі моделі, навчені на інших даних.

	Logistic Regression	Linear SVM	Linear SVM (SGD)	Random Forest	Gradient Boosted Machines
TP	3088	3004	3016	3046	3173
FP	96	180	168	138	11
TN	944	1139	1128	712	128
FN	626	431	442	858	1442

Рис. 2. Матриці помилок для використаних моделей

Fig. 2. Error matrices for used models

Model	Logistic Regre - ssion	Linear SVM	Linear SVM (SGD)	Random Forest	Gradient Boosted Machines
Accuracy	0,85	0,87	0,87	0,79	0,69
Precision	0,86	0,87	0,87	0,80	0,76
Recall	0,85	0,87	0,87	0,79	0,69
F1 Score	0,84	0,87	0,87	0,77	0,59

Рис. 3. Розраховані метрики оцінки моделей класифікаторів

Fig. 3. Calculated metrics for evaluating classifier models

Model	Logistic Regre-ssion	Linear SVM	Linear SVM (SGD)	Random Forest	Gradient Boosted Machines
CV Score (TF)	0,82	0,86	0,86	0,78	0,69
Test Score (TF)	0,85	0,87	0,87	0,79	0,69
CV Score (TF-IDF)	0,82	0,86	0,86	0,78	0,69
Test Score (TF-IDF)	0,85	0,87	0,87	0,79	0,69

Рис. 4. Результати перехресної перевірки

Fig. 4. Cross-validation results

Приклад роботи машини опорних векторів. Після оцінки якості роботи класифікаторів виявилось, що найкращим для цієї задачі є машина опорних векторів. Завдяки збереженим під час роботи параметрам моделі є змога використати зазначений метод для класифікації коментарів під навчальним відео з назвою «Метод Крамера за 3 минуты. Решение системы линейных уравнений – bezbotvy». Під ним на мо-

doi: https://doi.org/10.15802/stp2020/225264

ISSN 2307–3489 (Print), ISSN 2307–6666 (Online)

Наука та прогрес транспорту. Вісник Дніпропетровського національного університету залізничного транспорту, 2020, № 6 (90)

мент вивантаження наявні 336 коментарів. Вибір такої невеликої кількості коментарів зумовлений тим, що для оцінки класифікатора на прикладі потрібно попередньо віднести дані до різних класів вручну. У результаті отримано 26 токсичних коментарів, що складає 8 відсотків від загальної кількості.

Коментарі під відео було вивантажено, очищено, сконструйовано з них простір ознак та подано на вхід класифікатора машини опорних векторів. На рис. 5 видно результат перетворення коментарів за рахунок очищення.

Таким чином отримано таблицю, де стовбцями є коментарі та відповідні їм класи, визначені за методом. Точність на цій вибірці складає 85 %. Частину результату можна бачити на рис. 6.

Comment	Lemma
самый тупой пример спасибо	самый тупой пример спасибо
Спасибо большое	спасибо большой
Спасибо я все понял, я первокурсник высше образование)	спасибо понять первокурсник высший образование
Х**ета	х**та
Бог *_* У меня в институте это таааааак сложно "объясняют", но тут сраз	бог институт это сложно объяснять сразу все ясно стать
СПАСИБО ВАМ, в правде я в универе не понимала эти матрицы но благ	спасибо правда универ понимать матрица благодаря ва
Ты вот если не объясняешь путем что откуда берётся, не лезь в Ютуб	объяснять путь откуда браться лезь ютуб
Спасибо большое. Я все поняла только благодаря Вам.	спасибо большой понять благодаря
Вы олень. Меня попросила написать это преподаватель.	олень просить писать это преподаватель
вот я думаю. его калькулятор не правильно решил или все ж он сам □	думать калькулятор правильно решить
класс	класс
Неуч.	неуч
А если у меня это з**тельная х**ня равно =0 тогда как мне вычислить де	это за**ительный х**ня равный вычислить дельта х у z
Х**ня не понятно	х**ня понятный
Спасибо большое очень помогли мне	спасибо большой очень помочь
Опять ты об**рался	об**раться
или б**ать начну. Пока не решил.	б**ать начать пока решить
Не, ну раз очень надо, надо разбирать. Но если ролик 20 лайков за сегод	очень разбирать ролик лайк сегодняшний вечер набрат
Спасибо. Рассказываешь лучше чем учитель. Я посоветовал одноклассни	спасибо рассказывать учитель советовать одноклассник
Спасибо!!!	спасибо
Юра. Дискриминант - это в квадратном уравнении и в моем видео тольк	юра дискриминант этот квадратный уравнение мой вид
Спасибо! А в каком уроке Вы объясняете, как получить дискриминант м	спасибо какой урок объяснять получить дискриминант
спасибо блин, вы объясняете лучше, чем все мои преподы в универе вме	спасибо блин объяснять мой препод универ вместе взят
Спасибо. А не подскажешь, что ты сейчас проходишь по матану, ангему	спасибо подсказать проходить матана ангем линал акту
Ты лучше всех рассказываешь)) честно	рассказывать честно

Рис. 5. Результат очищення коментарів

Fig. 5. Result of clearing comments

самый тупой пример спасибо	1.0
Спасибо большое	0.0
Спасибо я все понял, я первокурсник высше образование)	0.0
Х**ета	1.0
Бог * * У меня в институте это таааааак сложно "объясняют", но тут сраз	0.0
СПАСИБО ВАМ, в правде я в универе не понимала эти матрицы но благ	0.0
Ты вот если не объясняешь путем что откуда берётся, не лезь в Ютуб	1.0
Спасибо большое. Я все поняла только благодаря Вам.	0.0
Вы олень. Меня попросила написать это преподаватель.	1.0
вот я думаю. его калькулятор не правильно решил или все ж он сам 😑	0.0
класс	0.0
Неуч.	1.0

Рис. 6. Результат роботи машини опорних векторів

Fig. 6. Result of work of the support vector machine

doi: https://doi.org/10.15802/stp2020/225264

ISSN 2307–3489 (Print), ISSN 2307–6666 (Online)

Наука та прогрес транспорту. Вісник Дніпропетровського національного університету залізничного транспорту, 2020, № 6 (90)

Результати

Отже, у нашій роботі було вирішено низку прикладних задач.

По-перше, досліджено наявну практики роботи з коментарями під відео на каналі «YouTube». Робота містить огляд праць з аналізу коментарів у соціальних медіа, зокрема, із використанням ресурсів «YouTube» та приклади подібного використання для мережі «Twitter».

Окрім того, проведено оцінку якості класифікаційних моделей, побудованих на основі різних методів Data Mining із використанням показників точності, повноти, Ф-міри, матриці помилок. Для узагальнення оцінки проведено перехресну перевірку.

Загалом, ґрунтуючись на попередніх дослідженнях, із використанням технік машинного навчання під наглядом (supervised machine learning techniques) розроблено алгоритми класифікації неструктурованих текстових даних, представлених у вигляді коментарів. Головною вирішеною задачею є класифікація токсичності коментарів. Для цього використано класифікатор логістичної регресії, метод класифікації за допомогою лінійних опорних векторів без та з методом навчання – стохастичний градієнтний спуск, класифікатор «Випадковий ліс» та класифікатор із посиленням градієнта. Застосовано алгоритм оцінки роботи класифікаторів, що включає використання методів підрахунку матриці помилок, точності, повноти та Ф-міри для оцінки моделей. За всіма показниками найточнішим виявився метод опорних векторів (Linear SVM) без та з методом навчання стохастичний градієнтний спуск, у них найвищі й інші метрики оцінки, такі як точність, повнота та Ф-міра.

Наукова новизна та практична значимість

Комбінація методів попередньої обробки специфічного виду тексту, а також класифікаційних методів машинного навчання дозволила не тільки врахувати такі особливості інформації, як можливість наявності таймокодів, емоджі, посилань тощо, але й використати розроблені алгоритми для аналізу російськомовних коментарів.

Практична цінність отриманих у роботі результатів полягає насамперед у спрощенні процесу аналізу коментарів. Описані технології можуть бути використані для аналізу текстових коментарів у будь-яких соціальних мережах або сервісах, де передбачені відгуки. Аналіз коментарів, розташованих під навчальними відео, може бути корисним для розуміння доступності матеріалу, а також якості його викладання.

Висновки

Для класифікації коментарів серед розглянутих класичних методів машинного навчання доцільно використовувати метод опорних векторів без та з методом навчання стохастичний градієнтний спуск. Ці напрацювання дають змогу спростити аналіз коментарів.

Вирішені задачі попередньої обробки та поділу коментарів на «токсичні» й «нетоксичні» допоможуть сформувати доброзичливе середовище у сфері інформаційної освіти шляхом відкидання токсичних коментарів. Для коментарів під навчальними відео для школярів така практика є, безсумнівно, важливою.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Рассел М., Классен М. *Data Mining. Извлечение информации из Facebook, Twitter, LinkedIn, Instagram, GitHub*. Санкт-Петербург : Питер, 2020. 464 с.
2. Anand M., Eswari R. Classification of Abusive Comments in Social Media using Deep Learning. *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)* (Erode, 27–29 mar. 2019). Erode, India, 2019. P. 974–977. DOI: https://doi.org/10.1109/iccmc.2019.8819734
3. Andročec D. Machine learning methods for toxic comment classification : a systematic review. *Acta Univ. Sapientiae Informatica*. 2020. Vol. 12. Iss. 2. P. 205–216.

doi: https://doi.org/10.15802/stp2020/225264

4. Casselman I., Heinrich M. Novel use patterns of Salvia divinorum : unobtrusive observation using YouTube™. *Journal of Ethnopharmacology*. 2011. Vol. 138. Iss. 3. P. 662–667. DOI: https://doi.org/10.1016/j.jep.2011.07.065
5. Chakrabarty N. *A Machine Learning Approach to Comment Toxicity Classification*. Computational Intelligence in Pattern Recognition. 2020. P. 183–193. DOI: https://doi.org/10.1007/978-981-13-9042-5_16
6. Chary M., Park E. H., McKenzie A., Sun J., Manini A. F., Genes, N. Signs & symptoms of Dextromethorphan exposure from YouTube. *PLoS One*. 2014. Vol. 9. Iss. 2. P. 1–10. DOI: https://doi.org/10.1371/journal.pone.0082452
7. Hsu C.-F., Caverlee J., Khabiri E. Hierarchical comments-based clustering. *SAC'11 : Proceedings of the 2011 ACM Symposium on Applied Computing*. 2011. P. 1130–1137. DOI: https://doi.org/10.1145/1982185.1982434
8. Jiang M., Liang Y., Feng X., Fan X., Pei Z., Xue Y., Guan R. Text classification based on deep belief network and softmax regression. *Neural Computing and Applications*. 2018. Vol. 29. Iss. 1. P. 61–70. DOI: https://doi.org/10.1007/s00521-016-2401-x
9. Khabiri E., Caverlee J., Hsu C.-F. Summarizing User-Contributed Comments. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. 2011. P. 354–537.
10. Srivastava S., Khurana P., Tewari V. Identifying Aggression and Toxicity in Comments using Capsule Network. *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*. P. 98–105.
11. Sun C., Qiu X., Xu Y., Huang X. *How to Fine-Tune BERT for Text Classification?* Lecture Notes in Computer Science. Springer, Cham, 2019. P. 194–206. DOI: https://doi.org/10.1007/978-3-030-32381-3_16
12. Zaheri S. Leath J., Stroud D. Toxic Comment Classification. *SMU Data Science Review*. 2020. Vol. 3, No. 1. P. 1–13.

Л. С. КОРЯШКИНА[1*], Г. В. СИМОНЕЦ[2]

[1*]Каф. «Системный анализ и управление», Национальный технический университет «Днипровская политехника», пр. Д. Яворницкого, 19, Днипро, Украина, 49005, тел. +38 (095) 565 76 83, эл. почта koriashkina.l.s@nmu.one, ORCID 0000-0001-6423-092X
[2]Каф. «Системный анализ и управление», Национальный технический университет «Днипровская политехника», пр. Д. Яворницкого, 19, Днипро, Украина, 49005, тел. +38(068) 848 42 00, эл. почта galya.golovko.2014@gmail.com, ORCID 0000-0002-1322-7189

ПРИМЕНЕНИЕ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ОБРАБОТКИ КОММЕНТАРИЕВ ПОД УЧЕБНЫМ МАТЕРИАЛОМ НА ВИДЕОХОСТИНГЕ «YOUTUBE»

Цель. Авторы ставят целью обнаружение токсичных комментариев на видеохостинге «YouTube» под учебным материалом путем классификации неструктурированного текста с помощью комбинации методов машинного обучения. **Методика.** В работе с указанным типом данных использованы методы машинного обучения для очистки, нормализации, представления текстовых данных в виде, приемлемом для дальнейшей работы на ЭВМ. Непосредственно для отнесения комментариев к классу «токсичные» использованы классификатор логистической регрессии, метод классификации с помощью линейных опорных векторов без и с методом обучения – стохастическим градиентным спуском, классификатор «случайный лес» и классификатор с усилением градиента. С целью оценки работы классификаторов использованы методы подсчета матрицы ошибок, точности, полноты и Ф-меры. Для обобщенной оценки использован метод перекрестной проверки. Язык программирования – Python. **Результаты.** На основе показателей оценки выбраны наилучшие методы – опорных векторов (Linear SVM) без и с методом обучения с помощью стохастического градиентного спуска. Описанные технологии могут быть использованы для анализа текстовых комментариев под любыми учебными видео для обнаружения токсичных отзывов. Также разработанный подход может быть полезным для выявления нежеланной или даже агрессивной информации в социальных сетях или сервисах, где предусмотрены отзывы. **Научная новизна.** В работе использовано комбинацию методов предварительной обработки специфического вида текста с учётом таких особенностей, как возможность наличия таймокодов, эмоджи, ссылок и тому подобное, а также адаптировано классификационные методы машинного

doi: https://doi.org/10.15802/stp2020/225264

ISSN 2307–3489 (Print), ISSN 2307–6666 (Online)

Наука та прогрес транспорту. Вісник Дніпропетровського національного університету залізничного транспорту, 2020, № 6 (90)

обучения для анализа русскоязычных комментариев. **Практическая значимость.** Проведено оптимизацию (упрощение) процесса анализа комментариев, необходимость которой обусловлена растущими объемами текстовых данных, особенно в сфере образования из-за карантинных условий и перехода на дистанционную форму обучения. Объемы учебного интернет-контента уже нуждаются в автоматизации процесса обработки и анализа отзывов, а со временем эта потребность будет только расти.

Ключевые слова: обработка естественного языка; неструктурированные данные; комментарии; классификация; логистическая регрессия; метод опорных векторов; стохастический градиентный спуск; «случайный лес»; усиление градиента; перекрестная проверка

L. S. KORIASHKINA[1*], H. V. SYMONETS[2]

[1*]Dep. «System Analysis and Management», Dnipro University of Technology, D. Yavornytskoho Av., 19, Dnipro, Ukraine, 49005, tel. +38 (095) 565 76 83, e-mail koriashkina.l.s@nmu.one, ORCID 0000-0001-6423-092X
[2]Dep. «System Analysis and Management», Dnipro University of Technology, D. Yavornytskoho Av., 19, Dnipro, Ukraine, 49005, tel. +38 (068) 848 42 00, e-mail galya.golovko.2014@gmail.com, ORCID 0000-0002-1322-7189

APPLICATION OF MACHINE LEARNING ALGORITHMS FOR PROCESSING COMMENTS FROM THE YOUTUBE VIDEO HOSTING UNDER TRAINING VIDEOS

Purpose. Detecting toxic comments on YouTube video hosting under training videos by classifying unstructured text using a combination of machine learning methods. **Methodology.** To work with the specified type of data, machine learning methods were used for cleaning, normalizing, and presenting textual data in a form acceptable for processing on a computer. Directly to classify comments as "toxic", we used a logistic regression classifier, a linear support vector classification method without and with a learning method – stochastic gradient descent, a random forest classifier and a gradient enhancement classifier. In order to assess the work of the classifiers, the methods of calculating the matrix of errors, accuracy, completeness and F-measure were used. For a more generalized assessment, a cross-validation method was used. Python programming language. **Findings.** Based on the assessment indicators, the most optimal methods were selected – support vector machine (Linear SVM), without and with the training method using stochastic gradient descent. The described technologies can be used to analyze the textual comments under any training videos to detect toxic reviews. Also, the approach can be useful for identifying unwanted or even aggressive information on social networks or services where reviews are provided. **Originality**. It consists in a combination of methods for preprocessing a specific type of text, taking into account such features as the possibility of having a timecode, emoji, links, and the like, as well as in the adaptation of classification methods of machine learning for the analysis of Russian-language comments. **Practical value**. It is about optimizing (simplification) the comment analysis process. The need for this processing is due to the growing volumes of text data, especially in the field of education through quarantine conditions and the transition to distance learning. The volume of educational Internet content already needs to automate the processing and analysis of feedback, over time this need will only grow.

Keywords: natural language processing; unstructured data; comments; classification; logistic regression; support vector machine; stochastic gradient descent; random forest; strengthening the gradient; cross-validation

REFERENCES

1. Russell, M., & Klassen, M. (2020). *Data Mining. Extracting information from Facebook, Twitter, Linkedln, Instagram, GitHub*. St. Petersburg: Piter. (in Russan)
2. Anand, M., & Eswari, R. (2019). Classification of Abusive Comments in Social Media using Deep Learning. In *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 974-977). Erode, India. DOI: https://doi.org/10.1109/iccmc.2019.8819734 (in English)
3. Andročec, D. (2020). Machine learning methods for toxic comment classification: a systematic review. *Acta Univ. Sapientiae Informatica*, *12*(2), 205-216. (in English)

doi: https://doi.org/10.15802/stp2020/225264

4. Casselman, I., & Heinrich, M. (2011). Novel use patterns of Salvia divinorum: Unobtrusive observation using YouTube™. *Journal of Ethnopharmacology, 138*(3), 662-667. DOI: https://doi.org/10.1016/j.jep.2011.07.065 (in English)
5. Chakrabarty, N. (2019). *A Machine Learning Approach to Comment Toxicity Classification*. Advances in Intelligent Systems and Computing (pp. 183-193). DOI: https://doi.org/10.1007/978-981-13-9042-5_16 (in English)
6. Chary, M., Park, E. H., McKenzie, A., Sun, J., Manini, A. F., & Genes, N. (2014). Signs & Symptoms of Dextromethorphan Exposure from YouTube. *PLoS ONE, 9*(2), 1-10. DOI: https://doi.org/10.1371/journal.pone.0082452 (in English)
7. Hsu, C.-F., Caverlee, J., & Khabiri, E. (2011). Hierarchical comments-based clustering. *SAC'11: Proceedings of the 2011 ACM Symposium on Applied Computing* (pp. 1130-1137). DOI: https://doi.org/10.1145/1982185.1982434 (in English)
8. Jiang, M., Liang, Y., Feng, X., Fan, X., Pei, Z., Xue, Y., & Guan, R. (2016). Text classification based on deep belief network and softmax regression. *Neural Computing and Applications, 29*(1), 61-70. DOI: https://doi.org/10.1007/s00521-016-2401-x (in English)
9. Khabiri, E., Caverlee, J., & Hsu, C.-F. (2011). Summarizing User-Contributed Comments. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media* (pp. 354-357). (in English)
10. Srivastava, S., Khurana, P., & Tewari, V. (2018). Identifying Aggression and Toxicity in Comments using Capsule Network. *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)* (pp. 98-105). (in English)
11. Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). *How to Fine-Tune BERT for Text Classification?* In Lecture Notes in Computer Science (pp. 194-206). DOI: https://doi.org/10.1007/978-3-030-32381-3_16 (in English)
12. Zaheri, S. Leath, J., & Stroud, D. (2020). Toxic Comment Classification. *SMU Data Science Review, 3*(1), 1-13. (in English)

Надійшла до редколегії: 14.08.2020
Прийнята до друку: 14.12.2020

doi: https://doi.org/10.15802/stp2020/225264