уряд перейде до «питань цифровізації всієї економіки», що є насправді критичним питанням розвитку самої сфери (індустрії) ІКТ. Вплив цифрових технологій перетворить державних високопосадовців – менеджерів відповідних сфер – фактично у «галузевих CIO».

Отже, внутрішній ринок ІКТ та цифровізація України є одним цілим із погляду управлінських, організаційних, інвестиційних, фінансових та інших зусиль. Гармонійний розвиток обох сфер на основі ринкових механізмів та державного «смарт-активізму» дасть змогу секторам економіки та сферам життєдіяльності протягом декількох років (замість десятиліть) здійснити гігантські кроки, бути перенесеними із застарілого середовища до сучасного та навіть надсучасного, тобто здійснити так званий цифровий стрибок.

ЛІТЕРАТУРА

1. Цифрова економіка – це реально. URL: http://chp.com.ua/ua/all-news/item/48511-tsifrova-ekonomika-tse-realno.
2. Україна 2030Е – країна з розвинутою цифровою економікою. URL: https://strategy.uifuture.org/kraina-z-rozvinutoyu-cifrovoyu-ekonomikoyu.html.
3. «Цифрова адженда України – 2020 («Цифровий порядок денний – 2020)», – ГС «ХАЙ-ТЕК ОФІС УКРАЇНА», 2016.
4. Як цифрова економіка змінить Україну. URL:https://www.epravda.com.ua/columns/2018/01/16/633057/.
5. Е-Країна 2030: Якою стане Україна за 10 років, якщо все піде гаразд. URL: https://patrioty.org.ua/ecomomic/e-kraina-2030-iakoiu-stane-ukraina-za-10-rokiv-iakshcho-vse-pide-harazd-283195.html.
6. Васьків О. М. Шевчук Ю. І. Цифрові трансформації через цифрові тренди. Фінансова політика України в умовах європейської інтеграції: зб. тез наук. доп. за матеріалами І Всеукр. наук.-практ. конф. здобувачів вищої освіти та молодих вчених. Львів: ЛНУ імені Івана Франка, 2020. Ч. 1. С. 259-261. URL: https://financial.lnu.edu.ua/wp-content/uploads/2020/04/Zbirnyk_02.2020_CH_1.pdf.

Косован О. В.
студент магістратури

Львівський національний університет імені Івана Франка
м. Львів, Україна

МАШИННЕ НАВЧАННЯ ДЛЯ АНАЛІЗУ ТОНАЛЬНОСТІ ТЕКСТУ ВІДГУКІВ КЛІЄНТІВ ОНЛАЙН МАГАЗИНІВ

Власникам інтернет-магазинів є життєво необхідним вчасно реагувати на відгуки та зауваження клієнтів. Якісний сервіс є важливим пунктом в умовах конкуренції [2, с. 93]. Тому є актуальним розробка моделі для класифікації відгуків на позитивні та негативні. Додатковою складністю

є те, що готових рішень є менше для української мови у порівнянні з англійською чи російською [3, с. 4]. Машинне навчання та деякі підходи обробки природної мови є одними з інструментів, які використовують для цього [1, с. 1].

Класифікація – це навчання з вчителем, яке вимагає набір посортованих історичних даних для моделювання. Для цього використовувалися відгуки клієнтів про інтернет-магазини. Дані було зібрано з онлайн-сервісу вибору товарів і порівняння цін – Hotline.ua. Кожен відгук включає оцінку від 0 до 10 включно та текстовий опис досвіду співпраці з певним інтернет-магазином. Відгуки, які мали оцінку 7-10, були віднесені до категорії позитивних, а ті, які були оцінені від 0 до 3, до негативних. Результатом збору даних стала база відсортованих по класах відгуків українською мовою, яка містить 3034 екземпляри (*позитивні – 1113, негативні – 1923*).

Тексти в Інтернеті зазвичай містять багато шуму та неінформативних деталей, таких як HTML-теги чи посилання. Тому необхідна попередня обробка, яка являє собою процес очищення тексту та уніфікації токенів.

Роботи, які пов’язані з аналізом тональності тексту, перед моделюванням користуються наступними підходами для очищення тексту [1, с. 3]: видалення знаків пунктуації, абзаців, чисел, перетворення всіх великих літер на малі.

Також класичними рішеннями для видалення неінформативних слів та уніфікації слів зі схожим значенням є лематизація, n-грами та інші підходи.

Лематизація – перетворення слова до його словникової форми. Цей механізм схожий зі стемінгом, але на відміну від нього, лематизація знаходить похідну форму слова, а не корінну. Тобто, слова «задоволеною», «задоволеним» будуть приведені до слова «задоволений».

N-грами – це альтернатива морфологічному розбору і видаленню стоп-слів. N-грама – це частина рядка, що складається з N символів [3, с. 264-265].

Важливим етапом перед навчанням моделей є векторизація текстового корпусу. Ефективним та популярним рішенням є TF-IDF. TF-IDF (англ. TF – *term frequency*, IDF – *inverse document frequency*) використовується для оцінки важливості слів у контексті документа, що є частиною корпусу. Вага (значущість) слова пропорційна кількості вживань цього слова у документі, і обернено пропорційна частоті вживання слова в інших документах колекції. Обчислюється TF-IDF як добуток tf (1) та idf (2) [4, с. 181].

$$tf(t,d) = \frac{n_i}{\sum_{k=0}^{m} n_k} \quad (1)$$

$$idf(t, D) = log \frac{|D|}{|(d \supset t)|} \quad (2)$$

де t – поточне слово, d – поточний документ, і n – кількість входжень поточного слова у поточний документ, D – корпус документів.

Основними показниками ефективності, які використовують для оцінки результатів класифікації, є точність (англ. *accuracy*), влучність (англ. *precision*), повнота (англ. *recall*) та F1-міра [5, с. 29].

Міра f1 базується на значеннях влучності та повноти. Precision показує відношення кількості правильно класифікованих документів, що були віднесені до певного класу, до всіх документів, що були віднесені до цього класу. Recall – це доля знайдених класифікатором документів, що відносяться до певного класу, відносно всіх документів цього класу у дата сеті [6, с. 265].

У ході розв'язання даної задачі було створено програмне забезпечення засобами мови python3 та бібліотек scikit-learn, lang-uk, nlp_uk. До всіх текстових даних було застосовано основні засоби очищення тексту та лематизацію.

У досліді використовувалися класифікатори Logistic Regression CV (*Логістична регресія*), Ridge Classifier (*Логістична регресія з нормалізацією L2*), Passive Aggressive Classifier (*Пасивно-агресивний класифікатор*), Perceptron та Multi-layer Perceptron (*багатошаровий перцептрон*). Результати кожної з моделей описані у таблиці 1.

Таблиця 1

Результати моделювання

Назва моделі	Precision	Recall	Accuracy	F1-міра
Logistic Regression CV	0.81	0.81	0.81	0.81
Ridge Classifier	0.80	0.80	0.80	0.79
Passive Aggressive Classifier	0.80	0.80	0.80	0.80
Perceptron	0.80	0.79	0.79	0.78
Multi-layer Perceptron	0.78	0.79	0.79	0.78

Отримані результати показують, що машинне навчання є одним з потенційних та ефективних інструментів для визначення тональності відгуків клієнтів. Актуальним є продовжити дослідження та збір даних

для покращення ефективності моделей. Також варто спробувати інші моделі на кшталт дерев рішень та нейронних мереж.

ЛІТЕРАТУРА

1. Babenko, Dmytro. Determining sentiment and important properties of Ukrainian-language user reviews : Master Thesis : manuscript rights / Dmytro Babenko ; Supervisor Vsevolod Dyomkin ; Ukrainian Catholic University, Department of Computer Sciences. – Lviv : [s.n.], 2020. – 35 p. : ill.
2. Hlinenko, L. K., & Daynovskyy, Y. A. (2018). State-of art and prospects of development of Ukrainian electronic commerce. Marketing and Management of Innovations, 1, 83-102. http://doi.org/10.21272/mmi.2018.1-06
3. Алгоритм токенізації та стемінгу для текстів українською мовою / А. М. Глибовець, В. В. Точицький // Наукові записки НаУКМА. Комп'ютерні науки. – 2017. – Т. 198. – С. 4-8. – Режим доступу: http://nbuv.gov.ua/UJRN/NaUKMAkn_2017_198_4
4. Метод визначення схожості новинних текстів на основі статистичної міри «term frequency-inverce document frequency» / М. О. Гранік, В. І. Месюра // Вісник Хмельницького національного університету. Технічні науки. – 2015. – № 4. – С. 180-182. – Режим доступу: http://nbuv.gov.ua/UJRN/Vchnu_tekh_2015_4_37
5. Emma Haddi, Xiaohui Liu, Yong Shi. The Role of Text Pre-processing in Sentiment Analysis. Procedia Computer Science 17 (2013) p. 26–32.
6. Аналіз впливу попередньої обробки тексту на результати текстової класифікації. Гущин І. В., Сич Д. О. // Науковий журнал «Молодий вчений» – 2018. – Х. 523. С. 264-266. – Режим доступу: http://molodyvcheny.in.ua/files/journal/2018/10/63.pdf