DOI 10.31891/2307-5732-2024-333-2-66
УДК 004.912

КУЛАЖЕНКО ВОЛОДИМИР
Державний торговельно-економічний університет
https://orcid.org/0000-0002-3535-3442
e-mail: v.kulazhenko@knute.edu.ua
МАЗУР ОЛЕКСАНДРА
Державний торговельно-економічний університет
e-mail: O.Mazur.FIT.124.20@knute.edu.ua

ДОСЛІДЖЕННЯ АЛГОРИТМУ ПОБУДОВИ МОДЕЛІ СЕНТИМЕНТ АНАЛІЗУ ПОВІДОМЛЕНЬ У СОЦІАЛЬНИХ МЕРЕЖАХ

В сучасних умовах, при необхідності постійного моніторингу настроїв суспільства, аналіз тональності повідомлень та коментарів до них дає можливість визначити чи сподобався користувачам товар, банк може дізнатись оцінку якості обслуговування з коментарів клієнтів, претенденти на виборах можуть дослідити, хто з них отримає більше голосів виборців, тощо.

Дана стаття присвячена проблемі побудови алгоритму сентимент аналізу повідомлень із соціальних мереж та його практичній реалізації засобами Python. Також, розкрито класифікація засобів проведення аналізу тональності повідомлень. Зазначено, що найбільш дієвими засобами є ті, які засновані на словниках і правилах, засоби машинного навчання та ручна обробка. Особлива увага приділена відповідним онлайн сервісам, які виконують такі задачі, наведена їх коротка характеристика.

В дослідженні були використані дані, надані сервісом YouScan, який не тільки здатний збирати потрібну для аналізу інформацію, а й здатний аналізувати україномовні тексти.

Однак, в умовах повномасштабного вторгнення та намагання агресора втрутитись у всі сфери життя, подібна інформація має бути конфіденційною. Отже, можливості витоку інформації мають бути мінімізовані. В цих умовах, слід звертатись до засобів машинного навчання, здатних працювати на локальних ресурсах. Використання аналізу вручну теж можливе, однак є неефективним економічно.

Особлива увага у роботі була приділена підготовці даних для роботи моделі, а саме: очищення повідомлень від зайвих символів, рисунків, знаків пунктуації, емодзі, тощо; токенізація тексту; стемінг отриманих векторів. В роботі було використано модель Pipeline з логістичною регресією, як основний засіб машинного навчання для вирішення задачі.

Ефективність побудованої, таким чином, моделі була перевірена на тестових даних. Були обраховані метрики для її оцінки, а саме – точність (Precision) та повнота (Recall). В результаті, було виявлено, що дана модель у 22% випадках оцінює позитивний коментар як негативний. Для усунення цього недоліку, запропоновано збільшити поріг визначення позитивної оцінки з 0,5 до 0,67.

Ключові слова: аналіз тональності повідомлень, аналіз повідомлень у соціальних мережах, машинне навчання, логістична регресія, сентимент аналіз.

KULAZHENKO VOLODYMYR, MAZUR OLEKSANDRA
State University of Trade and Economics

STUDY OF THE ALGORITHM FOR BUILDING A MODEL OF SENTIMENT ANALYSIS OF MESSAGES IN SOCIAL NETWORKS

In modern conditions, with the need to constantly monitor public sentiment, analyzing the tone of posts and comments to them makes it possible to determine whether users like a product, a bank can find out the assessment of the quality of their services from customers' feedbacks, election candidates can investigate which of them will receive more votes, etc.

This article is devoted to the problem of developing of a sentiment analysis algorithm for messages from social networks and its practical implementation using Python tools. Additionally, the classification of tools for conducting sentiment analysis of messages is disclosed. It is noted that the most effective tools are those based on dictionaries and rules, machine learning tools, and manual processing. Special attention is given to the relevant online services that perform such tasks, and a brief description of them is provided.

The data were provided by the YouScan service. This service is capable not only of collecting the necessary information for analysis but also of analyzing texts in Ukrainian.

However, in the context of a full-scale invasion and the aggressor's attempts to interfere in all spheres of life, such information should be confidential. Therefore, the possibility of information leakage should be minimized. In these circumstances, machine learning tools capable of operating on local resources should be used. The use of manual analysis is also possible, but it is not cost-effective.

Particular attention was paid to preparing data for the model, namely: cleaning messages from unnecessary characters, pictures, punctuation marks, emojis, etc.; tokenizing of the text; and stemming of the resulting vectors. In this paper, the Pipeline model with logistic regression was used as the main machine learning tool for solving the problem.

The effectiveness of the model built in this way was tested on test data. Metrics for its evaluation were calculated, namely Precision and Recall. As a result, it was found that this model evaluates a positive comment as a negative one in 22% of cases. To eliminate this drawback, it was proposed to increase the threshold for determining of a positive assessment from 0.5 to 0.67.

Keywords: social media message analysis, machine learning, logistic regression, sentiment analysis.

Постановка проблеми

В сучасному світі соціальних мереж та цифрового спілкування, вивчення настроїв повідомлень користувачів, стає надзвичайно важливим засобом для низки дисциплін, включаючи маркетинг, соціологію та психологію. Величезний об'єм генерованого користувачами контенту містить в собі ключі до розуміння громадської думки, споживацьких настроїв та емоційних станів великих груп осіб.

Настрої повідомлень користувачів несуть в собі цінну інформацію, яка може бути використана для вдосконалення продуктів, послуг та комунікаційних кампаній. Ефективний аналіз настроїв може також виявити потенційні кризові точки у людському сприйнятті, дозволяючи організаціям своєчасно реагувати на них. В контексті швидких змін суспільних настроїв та постійного потоку інформації, вміння адекватно аналізувати емоційні підтексти стає життєво важливою навичкою для забезпечення стійкості та успішності в цифровому просторі.

Найбільш поширеними сферами, де застосовується аналіз тональності тексту є моніторинг соціальних мереж, вдосконалення обслуговування клієнтів, аналіз ринку (ставлення до товарів та брендів), фінансовий аналіз, соціологічний та політичний аналіз.

Аналіз тональності повідомлень (сентимент аналіз, (англ) “sentiment analysis”) — це процес використання обробки природної мови, текстового аналізу та обчислювальної лінгвістики для ідентифікації, кількісного аналізу, дослідження емоційних нюансів та суб'єктивних оцінок у текстових даних [1].

Як правило, такий аналіз є полярним, тобто дослідник намагається визначити лише позитивну, нейтральну або негативну загальну емоційну оцінку користувачів. В деяких випадках, застосовуються більш складні і неточні техніки для визначення конкретних позитивних (радість) та негативних (смуток, гнів) емоційних станів. Такий аналіз здатний відокремити об’єктивні настрої в суспільстві, або певному ком’юніті, від суб’єктивних думок або відгуків його представників, або, навіть, керівників.

Аналіз може проводитись на різних рівнях – від аналізу окремих слів чи фраз, до цілого речення, абзацу чи документу. Звичайно, у випадку з соціальними мережами, об’єктом такого аналізу є одне повідомлення, яке може вміщати від декількох слів до десятків абзаців тексту. Це значно ускладнює проведення такого аналізу. Крім цього, на складність впливає неформальна манера спілкування користувачів у соціальних мережах, використання нішевого сленгу, іронії, сарказму, тощо.

Метою даної статті є розробка та програмна реалізація алгоритму аналізу тональності повідомлень та коментарів у соціальних мережах, використовуючи засоби машинного навчання на базі Python.

Аналіз останніх джерел

Аналіз тональності повідомлень (тексту) використовується у багатьох напрямах. Зокрема, часто його використовують у електронній комерції. Так, дослідники Almahmood R. J. K. та Tekerek A. у своїй статті [2] аналізують зв’язок між відгуками користувачів на різні товари та їх майбутніми продажами іншим покупцям за допомогою нейромережевого моделювання. Було з’ясовано, коли саме і за яких умов, вплив чужих відгуків на товар може вплинути на купівельну привабливість товару.

Аналогічну роботу, з використання згорткової нейронної мережі, було виконано у дослідженні [3]. Автори дослідили, які саме відгуки можуть позитивно або негативно вплинути на бажання купити той чи інший товар.

У роботі [4] було застосовано згорткову нейронну мережу з метою побудови плану подальшого розвитку готеля. В якості даних, дослідники використовували відгуки, які залишили клієнти на сайті.

Також, дуже часто аналіз тональності повідомлень у соціальних мережах намагаються використовувати для прогнозування цін у трейдингу. Автори дослідження [5] запропонували інформаційну систему для торгівлі в реальному часі, яка базується на аналізі повідомлень у Twitter (X). Для досягнення результату було використано декілька алгоритмів машинного навчання (логістичну регресію, машину опорних векторів та інші) на навчальній вибірці, що склала понад 200 тисяч твітів.

Ідентифікація користувачів-ботів за допомогою визначення тональності, разом з шаблонною структурою повідомлень, лягло в основу досліджень [6] та [7]. У першій роботі, досліджували коментарі користувачів під відео-записами у YouTube на теми, що стосувались COVID-19. Автори описали методику виявлення повідомлень ботів, які відрізняються від інших своєю негативною тональністю, певною структурою та частотою публікацій. Натомість, у роботі [7] досліджувався датасет з повідомленнями відомих особистостей. Були знайдені спільні особливості, які вказували на штучність написаного тексту, а саме – час публікації, практично однакова позитивна тональність, тощо.

Автори досліджень [8] та [9] зосередили свої зусилля на ідентифікації такого емоційного показника як сарказм в користувацьких повідомленнях у соціальних мережах. Дане питання має велику значимість, оскільки наявність сарказму часто змінює сенс повідомлення. В роботі порівняно ефективність використання нейронних мереж кількох типів. Також, особлива увага була приділена питанню використання хештегів. За результатом цих робіт виявилось, що хештеги сильно впливають на ймовірність виявлення сарказму, і нейронні мережі враховують їх так само ефективно, як і під час аналізу людиною.

Виклад основного матеріалу

В сучасних умовах, при необхідності постійного моніторингу настроїв суспільства, аналіз тональності (сентимент аналіз) повідомлень та коментарів до них дає можливість визначити чи сподобався користувачам товар, банк може дізнатись якість обслуговування з коментарів клієнтів, претенденти на вибори можуть дослідити кому з них ймовірно віддадуть більшу перевагу на виборах, тощо.

Існує ряд методів для класифікації тональності, які відрізняються точністю та швидкістю. До найбільш популярних на даний момент методів відносять методи, які використовують словники та відповідні правила та методи машинного навчання.

Методи, засновані на словниках, є досить ефективними і точними у своїх оцінках. Однак, вони мають значний недолік, оскільки сильно прив'язані до предметної області. Це вимагає від аналітиків регулярного перегляду бази слів, а зміна об'єкту досліджень вимагає розробки нових словників.

При використанні засобів машинного навчання, надають перевагу методам, що включають навчання з учителем, оскільки вони дають більшу точність, ніж методи без нього [10].

У даній роботі використовується модель машинного навчання з учителем, а саме – логістична регресія.

Існують, також, онлайн сервіси, здатні проводити необхідний аналіз. До найбільш поширених входять:

1. Amazon Comprehend [11] – сервіс, який використовує моделі машинного навчання для аналізу та пошуку зв'язків у тексті, а також упорядкування документів на основні обраної теми. Користувач отримує доступ до Amazon Comprehend з консолі керування AWS, інтегруючи його з даними публікацій у соціальних мережах, що зберігаються в Amazon S3.
2. Google Cloud NLP [12] – платформа аналізу тексту для підприємств, сумісна з Google Cloud Storage, використовує технологію машинного навчання Google. Функціонал включає класифікацію вмісту, аналіз настроїв, розпізнавання сутностей та аналіз синтаксису, з можливістю знаходження і позначення полів у текстах, таких як чати, електронні листи та соцмережі.
3. Chattermill [13] – інструмент аналізу тексту на базі штучного інтелекту, що збирає дані з різних джерел зворотного зв'язку. Він підтримує інтеграцію з різними додатками та використовує штучний інтелект для розуміння настроїв клієнтів та пошуку зв'язків у повідомленнях.

Однак, для роботи з українською аудиторію ці сервіси не ефективні, оскільки погано, або зовсім не опрацьовують тексти, написані українською мовою. Для досліджень в Україні краще використовувати сервіс YouScan [14]. Це платформа для моніторингу та аналітики соцмереж на базі штучного інтелекту. Сервіс дозволяє відстежувати згадування брендів, а також здійснювати аналіз та реагування на проблеми користувачів, підтримуючи такі мови, як англійська, російська та українська.

Основними функціями які надає YouScan є:

1. Моніторинг онлайн-ЗМІ, соцмереж, блогів, форумів, месенджерів та сайтів відгуків. Результати моніторингу зі звітами доступні на сайті в режимі реального часу.
2. Хмара слів, звіти тональності, демографія авторів та візуальні інсайти, які дають можливість для зручного аналізу.
3. Функція правил, яка дозволяє налаштувати роботу по розмітці згадувань. За допомогою цієї функції можна також тегувати згадування, корегувати тональність та відправляти повідомлення.

При проведенні перевірці даних, отриманих від цього сервісу, стало зрозумілим, що приблизна точність його роботи складає 70-80%, що не завжди є прийнятним. Крім того, в певний момент, сервіс перестав виконувати оцінку повідомлень у соціальній мережі Facebook.

Дана соціальна мережа є визначальною при спробах аналізу українського суспільства, тому, доводиться проводити аналіз вручну. Звісно, це не дуже ефективний підхід з економічної точки зору, але, він має свої переваги, а саме:

1. Вища точність інтерпретації отриманих результатів.
2. Конфіденційність.
3. Гнучкість адаптації в залежності від актуальних змін.
4. Знання актуального сленгу для конкретних груп населення.
5. Можливість глибшого розуміння аудиторії.

В умовах, коли керівництво компаній намагається зберегти конфіденційність, що є важливим і для країни в цілому, і, одночасно, підвищити продуктивність, варто звернути увагу на засоби, які не потребують додаткових підключень до зовнішніх систем. Таким засобом може бути застосування машинного навчання.

Перед тим як починати класифікацію тональності, дані попередньо треба проаналізувати та обробити. Мета цих процедур – зведення тексту до такого виду, який буде зрозумілим для моделей машинного навчання. Після того як дані будуть вичищені та зведені до векторів, слід переходити до навчання моделі.

Найкращим інструментарієм для проведення таких досліджень є Python та його відповідні модулі. Оскільки саме в ньому, на даний момент, реалізований повний спектр засобів машинного навчання (Scikit-learn, TensorFlow+Keras, тощо), в тому числі, і в сфері сентимент аналізу. Зокрема, використовуються такі модулі Python як Pandas, String, Matplotlib, Numpy, Re, NLTK, Collections, Langdetect, Scikit-learn (sklearn) і Emoji [15-20].

Дані, які надійшли від YouScan мають наступний вигляд – див. рис. 1.

1	Дата	Время	Заголо	Текст	Тип по	URL	Тональ	Автор	Профи	Подпи	Демог	Возрас	Источник	Место	Профи	Подпи	Тип ист	Страна	Регион
1292	03.01.202	16:00		Резников	Пост	https://w	Нейтраль	On-line UI	https://w	160000	Сообщество		facebook.com	On-line UI	https://w	160000	Соц. сеть		
1293	03.01.202	14:37		УКРАИНС	Пост	https://w	Нейтраль	ФЛОТ2017	https://w	729	Сообщество		facebook.com	ФЛОТ2017	https://w	729	Соц. сеть		
1294	03.01.202	12:36		Німеччин	Пост	https://w	Нейтраль	Сторінка	https://w	63546	Сообщество		facebook.com	Сторінка	https://w	63546	Соц. сеть		
1295	03.01.202	11:42		Anna Max	Пост	https://w	Нейтраль	Людмила	https://w	2876	Женщина		facebook.com	Людмила	https://w	2876	Соц. сеть	Украина	Одесская

Рис. 1. Приклад даних, що надає сервіс YouScan для аналізу

З повним ходом дослідження, його кодом та результатом виконання можна ознайомитись на Github за посиланням:

https://github.com/mpakeron/FB_Sentiment_Anaysis/blob/b8a29da10ce4f25d6d080bab9480b7637f353ef8/Facebook.ipynb [21].

Очищення та обробка даних відбувається у кілька кроків. А саме:

1. Виокремлення необхідних для навчання моделі даних. Як видно з рисунку вище, сервіс надає декілька стовпців даних, з яких для нас будуть важливими лише зміст повідомлення та його оцінена тональність. По останній ознаці, слід замінити значення «позитивний» та «негативний» на 0 та 1 відповідно. Мають бути видалені повідомлення, що повторюються, містять лише графіку та медіа-контент, а також ті, які написані мовами, які не будуть досліджуватись.
2. Очищення тексту. З тексту видаляються усі зайві символи та такі текстові дані, як електронна пошта, посилання, теги, телефонні номери, тощо. Весь текст приводиться до нижнього регістру. Видаляється вся пунктуація, емодзі та стоп-слова [22].

Найбільш розповсюджений список стоп-слів, тобто найбільш поширені слова та висловлювання, які не впливають на тональність тексту, можна знайти у бібліотеці NLTK (Natural Language Toolkit) або на GitHub (Tokenize UK) [23]. Однак, цей перелік може бути доповнений на розсуд аналітика.

3. Токенізація тексту. Тобто, приведення даних до векторів.
4. Стемінг токенізованого тексту. Стемінг - це процес обробки природної мови, за допомогою якого слова зводяться до їхнього кореня або "стебла". Основна ідея полягає в тому, щоб відкинути афікси (приставки, суфікси, закінчення), залишаючи лише корінь слова, який представляє його базову форму [24].

Етап попередньої обробки тексту є таким же важливим як і сама класифікація, адже від того, в якому вигляді подавати дані моделі, буде залежати її точність. Крім того, він має бути автоматизований для багаторазового використання з різними даними. Після цього потрібно розділити дані на навчальну та тестову вибірки.

В цьому дослідженні, було використано модель Pipeline (scikit-learn), яка дозволяє утворювати лінійну послідовність з методів очищення тексту та машинного навчання.

В процесі векторизації, обов'язковим є обчислення TF-IDF, тобто оцінки важливості термінів у документі або корпусі документів. *TF* (частота терміну) визначає, наскільки часто термін зустрічається у документі. Вона вираховується як відношення кількості входжень терміну до загальної кількості слів у документі. *IDF* (інвертована частота документу) визначає, наскільки інформативним є термін, шляхом обчислення зворотного відношення до кількості документів, що містять даний термін [25].

Бібліотека scikit-learn містить для цього клас TfidfTransformer. За замовчуванням, у ньому частота терміна *TF* множиться на *idf*, який обчислюється як

$$idf(t) = log\frac{1+n}{1+df(t)} + 1, \quad (1.1)$$

де *n* – це сукупна кількість документів у вибірці, *df(t)* – це кількість документів у вибірці, що містять термін *t*. Отримані вектори потім нормалізуються за Евклідовою нормою:

$$v_{norm} = \frac{v}{\sqrt{v_1^2 + v_2^2 + \cdots + v_n^2}}. \quad (1.2)$$

Однак, з огляду на швидкість роботи та зручність, ми було вирішено використовувати TfidfVectorizer. Він більш пристосований для використання у pipeline-машинах, а також, не вимагає від користувача додаткових обчислень.

В якості самої моделі було використано логістичну регресію (клас LogisticRegression у scikit-learn) [1], яка часто використовується при бінарній класифікації та передбаченні подій.

Логістична регресія надає оцінку події у бінарному форматі, тобто 1 або 0. Тобто, вона отримує ймовірнісний результат в межах між 0 та 1, та округлює його до цілого числа.

У логістичній регресії логічні перетворення застосовуються до шансів, тобто ймовірності успіху, поділеної на ймовірність невдачі. Це також відомо як логарифм шансів або натуральний логарифм шансів. , і ця логістична функція представлена такими формулами(1.3, 1.4) [15]:

$$Logit(pi) = \frac{1}{1+\exp(-pi)}, \quad (1.3)$$

$$ln\left(\frac{pi}{1-pi}\right) = Beta_0 + Beta_1{}^*X_1 + \cdots + B_k{}^*K_k \quad (1.4)$$

У цьому рівнянні логістичної регресії *Logit(pi)* є залежною змінною або змінною відповіді, а x – незалежною змінною. Бета-параметр або коефіцієнт у цій моделі зазвичай оцінюється за допомогою оцінки максимальної правдоподібності (MLE). Цей метод перевіряє різні значення бета-версії за допомогою кількох ітерацій, щоб оптимізувати для найкращого підбору шансів. Усі ці ітерації створюють функцію логарифмічної правдоподібності, а логістична регресія прагне максимізувати цю функцію, щоб знайти найкращу оцінку параметра. Як тільки оптимальний коефіцієнт (або коефіцієнти, якщо існує більше однієї незалежної змінної) знайдено, умовні ймовірності для кожного спостереження можна обчислити, зареєструвати та підсумувати разом, щоб отримати прогнозовану ймовірність. По замовчуванню, для бінарної класифікації ймовірність менше 0,5 передбачатиме 0, а ймовірність більше 0,5 передбачатиме 1.

Після навчання моделі, слід оцінити її роботу на тестових даних [21].

Попереднім результатом тестування є правильні відповіді у приблизно 80% випадках. Однак, даний результат не є остаточним. Крім того, він не дозволяє перейти від ручного аналізу до автоматизованого, оскільки відсоток успішного аналізу не є прийнятним. Для більш точного оцінювання та покращення існуючої моделі, слід вирахувати відповідні метрики, а саме precision та recall. Вони є ключовими показниками оцінки ефективності моделей машинного навчання.

Для цього використаємо нову модель, щоб попередні передбачення на цьому ж тестовому датасеті не заважали оцінці.

Точність (Precision): вимірює, яку частину позитивних випадків модель ідентифікувала правильно з усіх випадків, які вона визначила як позитивні. Формула для обчислення точності наступна [15]:

$$Precision = \frac{TP}{TP+FP} \quad (1.5)$$

де:

TP (True Positive) - кількість правильно ідентифікованих позитивних випадків,

FP (False Positive) - кількість неправильно ідентифікованих позитивних випадків.

Повнота (Recall): вимірює, яку частину позитивних випадків модель виявила з усіх дійсно позитивних випадків набору. Формула для обчислення повноти наступна [15]:

$$Recall = \frac{TP}{TP+FN} \quad (1.6)$$

де:

FN (False Negative) - кількість позитивних випадків, які були помилково ідентифіковані як негативні.

Ці метрики доповнюють одна одну, і їх оптимальне значення залежить від контексту задачі. Наприклад, у деяких випадках може бути важливіше мати високу точність, а в інших - високу повноту.

Precision даної моделі дорівнює 78%, це означає що з такою ймовірністю модель правильно оцінює повідомлення. А параметр recall показав 97%, що означає скільки відсотків поганих коментарів з наявних у тестовому датасеті модель змогла знайти. З цього можна зробити проміжний висновок, що модель працює добре з негативними коментарями та повідомленнями, але має тенденцію до помилок у визначенні позитивних.

Після створення моделі був побудований графік precision/recall прямої, щоб оцінити який поріг слід обрати для максимізації якості цієї моделі (див. рис. 2) [21].

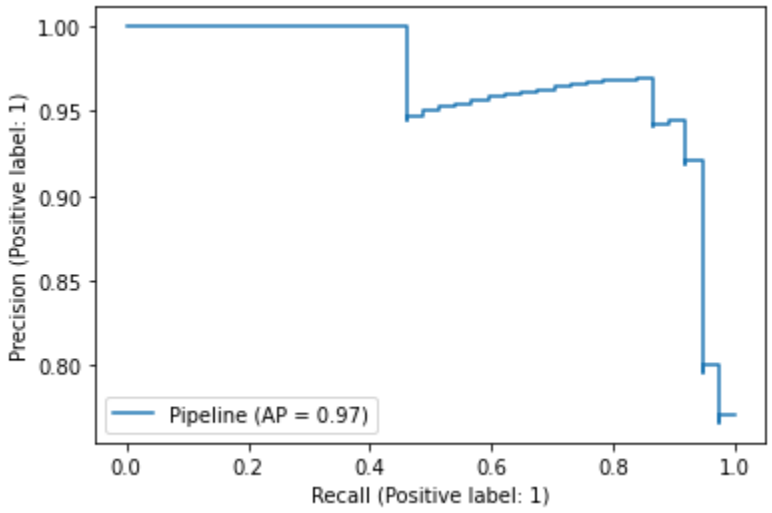

Рис. 2. Графік precision/recall моделі аналізу тональності повідомлень

Отже, при подальшому використанні даної моделі у визначенні тональності повідомлень та коментарів у соціальній мережі Facebook, слід визначити її оптимальний поріг як 0.67. Це означає, що у подальшій роботі все, що модель оцінюватиме нижче цього параметра, вона відноситиме до 0, а все що більше – до 1.

Висновки

Аналіз тональності текстових повідомлень у соціальних мережах та коментарів до них є потужним інструментом моніторингу ставлення суспільства до певних товарів, брендів, процесів або подій. Сучасні методи аналізу здатні робити це з відносно великою точністю. При цьому, в сучасних умовах ведення повномасштабної війни, варто брати до уваги безпекові виклики, і звертати увагу на інструменти, які будуть мінімально залежати від зовнішніх факторів, такі як засоби машинного навчання.

При роботі з визначенням тональності тексту засобами машинного навчання, критично важливим є робота з вхідними даними. До них входять: 1) очищення тексту від зайвих елементів, таких як: пунктуація, емодзі, електронна пошта, теги, посилання, тощо; 2) визначення та видалення стоп-слів; 3) токенізація тексту, тобто його приведення до векторів; 4) стемінг, завдяки якому залишаються лише корені слів.

Всі ці процеси значно впливають як на швидкість проведення аналізу, так і на його ефективність.

В статті, для аналізу використовувалась модель логістичної регресії. Не дивлячись на відносно непоганий результат – близько 80% правильних відповідей, було проведено аналіз моделі за допомогою додаткових метрик Precision та Recall.

Завдяки перевірці, вдалось з'ясувати, що модель помиляється при роботі з позитивними оцінками (78%). Тому варто змінити поріг поділу оцінок на позитивну та негативну з 0,5 до 0,67.

Література

1. What is Sentiment Analysis? – Режим доступу: https://aws.amazon.com/what-is/sentiment-analysis/. (дата звернення: 27.02.2024). — Назва з екрана.
2. Almahmood R. J. K. Issues and Solutions in Deep Learning-Enabled Recommendation Systems within the E-Commerce Field / R. J. K. Almahmood, A, Tekerek // Applied Sciences № 12 (21). – 2022. – P. 256–264. — Режим доступу: https://doi.org/10.3390/app122111256
3. Yin J. Y. B. Exploring Sentiment Analysis on E-Commerce Business: Lazada and Shopee / J. Y. B. Yin, N. H. M. Saad, Z. Yaacob // Tem Journal. № 11 (4). – 2022. – P. 1508–1519. — Режим доступу: https://doi.org/10.18421/TEM114-11
4. Hinduja S. Machine learning-based proactive social-sensor service for mental health monitoring using twitter data / S. Hinduja, M. Afrin, S. Mistry, A. Krishna // International journal of Information Management Data insights, № 2 (2). – 2022. – P. 103–124
5. Wang Y. Sentiment Analysis of Twitter Data / Y. Wang, J. Guo, C. Yuan, B. Li // Applied Sciences, №. 12 (8). – 2022. – P. 157–189. — Режим доступу: https://doi.org/10.3390/app122211775
6. Xie W. Emotional appeals and social support in organizational YouTube videos during COVID-19 / W. Xie, L. Damiano, C.-H. Jong // Telematics and Informatics reports. № 8 (1). – 2022. – P. 100–128
7. Abbas A. F. Bibliometrix analysis of information sharing in social media / A. F. Abbas, A. Jusoh, A. Mas'od, A. H. Alsharif, J. Ali // Cogent Business & Management, № 9 (1). – 2022. – P. 521–543. — Режим доступу: https://doi.org/10.1080/23311975.2021.2016556
8. Karyukin V. On the development of an information system for monitoring user opinion and its role for the public / V. Karyukin, G. Mutanov, Z. Mamykova, G. Nassimova, S. Torekul, Z. Sundetova, M. Negri // Journal of Big Data, No. 9 (1), – 2023. – P. 119–145. — Режим доступу: https://doi.org/10.1186/s40537-022-00660-w
9. Murphy S. T. Affect, cognition, and awareness: Affective priming with optimal and suboptimal stimulus exposures / S. T. Murphy // Journal of Personality and Social Psychology, No. 8 (3). – 2011. – P. 723–739. — Режим доступу: http://doi.org/10.1037/0022-3514.64.5.723
10. Müller A. Introduction to Machine Learning with Python. / A. Müller, Sarah Guido. Sebastopol: O'Reilly Media, Inc. – 2016. – 392 pp.
11. Amazon Comprehend – Continuously Trained Natural Language Processing. – Режим доступу: https://aws.amazon.com/ru/blogs/aws/amazon-comprehend-continuously-trained-natural-language-processing/ (дата звернення: 27.02.2024). — Назва з екрана.
12. Natural Language AI. – Режим доступу: https://cloud.google.com/natural-language. (дата звернення: 27.02.2024). — Назва з екрана.
13. Chattermill. – Режим доступу: https://www.predictiveanalyticstoday.com/chattermill/. (дата звернення: 27.02.2024). — Назва з екрана.
14. Моніторинг соцмедіа з візуальними інсайтами. – Режим доступу: https://youscan.io/. (дата звернення: 27.02.2024). — Назва з екрана.
15. Logistic regression – Режим доступу: https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression. (дата звернення: 27.02.2024). — Назва з екрана.
16. langdetect 1.0.9. – Режим доступу: https://pypi.org/project/langdetect/. (дата звернення: 27.02.2024). — Назва з екрана.
17. emoji 2.10.1. – Режим доступу: https://pypi.org/project/emoji/. (дата звернення: 27.02.2024). — Назва з екрана.
18. NLTK Documentation. – Режим доступу: https://www.nltk.org/api/nltk.tokenize.html. (дата звернення: 27.02.2024). — Назва з екрана.
19. String — Common string operations. – Режим доступу: https://docs.python.org/3/library/string.html. (дата звернення: 27.02.2024). — Назва з екрана.
20. nltk.corpus package – Режим доступу: https://www.nltk.org/api/nltk.corpus.html#nltk-corpus-package. (дата звернення: 27.02.2024). — Назва з екрана.
21. Код власної програмної реалізації алгоритму сентимент аналізу повідомлень та коментарів соціальних мереж https://github.com/mpakeron/FB_Sentiment_Anaysis/blob/b8a29da10ce4f25d6d080bab9480b7637f353ef8/Facebook.ipynb (дата звернення: 27.02.2024). — Назва з екрана.
22. Ukrainian-Stopwords. – Режим доступу: https://github.com/skupriienko/Ukrainian-Stopwords. (дата звернення: 27.02.2024). — Назва з екрана.
23. Tokenize UK. – Режим доступу: https://github.com/lang-uk/tokenize-uk. (дата звернення: 27.02.2024). — Назва з екрана.

24. Ukrainian Stemmer. – Режим доступу: https://github.com/Desklop/Uk_Stemmer.

25. sklearn.feature_extraction.text.TfidfVectorizer. – Режим доступу : https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html#sklearn.feature_extraction.text.TfidfVectorizer. (дата звернення: 27.02.2024). — Назва з екрана.

References

1. What is Sentiment Analysis? – Access mode: https://aws.amazon.com/what-is/sentiment-analysis/. (date of request: 27.02.2024). — Title from the screen.

2. Almahmood R. J. K. Issues and Solutions in Deep Learning-Enabled Recommendation Systems within the E-Commerce Field / R. J. K. Almahmood, A, Tekerek // Applied Sciences № 12 (21). – 2022. – P. 256–264. — Access mode: https://doi.org/10.3390/app122111256

3. Yin J. Y. B. Exploring Sentiment Analysis on E-Commerce Business: Lazada and Shopee / J. Y. B. Yin, N. H. M. Saad, Z. Yaacob // Tem Journal. № 11 (4). – 2022. – P. 1508–1519. — Access mode: https://doi.org/10.18421/TEM114-11

4. Hinduja S. Machine learning-based proactive social-sensor service for mental health monitoring using twitter data / S. Hinduja, M. Afrin, S. Mistry, A. Krishna // International journal of Information Management Data insights, № 2 (2). – 2022. – P. 103–124

5. Wang Y. Sentiment Analysis of Twitter Data / Y. Wang, J. Guo, C. Yuan, B. Li // Applied Sciences, №. 12 (8). – 2022. – P. 157–189. — Access mode: https://doi.org/10.3390/app122211775

6. Xie W. Emotional appeals and social support in organizational YouTube videos during COVID-19 / W. Xie, L. Damiano, C.-H. Jong // Telematics and Informatics reports. № 8 (1). – 2022. – P. 100–128

7. Abbas A. F. Bibliometrix analysis of information sharing in social media / A. F. Abbas, A. Jusoh, A. Mas'od, A. H. Alsharif, J. Ali // Cogent Business & Management, № 9 (1). – 2022. – P. 521–543. — Access mode: https://doi.org/10.1080/23311975.2021.2016556

8. Karyukin V. On the development of an information system for monitoring user opinion and its role for the public / V. Karyukin, G. Mutanov, Z. Mamykova, G. Nassimova, S. Torekul, Z. Sundetova, M. Negri // Journal of Big Data, No. 9 (1), – 2023. – P. 119–145. — Access mode: https://doi.org/10.1186/s40537-022-00660-w

9. Murphy S. T. Affect, cognition, and awareness: Affective priming with optimal and suboptimal stimulus exposures / S. T. Murphy // Journal of Personality and Social Psychology, No. 8 (3). – 2011. – P. 723–739. — Access mode: http://doi.org/10.1037/0022-3514.64.5.723

10. Müller A. Introduction to Machine Learning with Python. / A. Müller, Sarah Guido. Sebastopol: O'Reilly Media, Inc. – 2016. – 392 pp.

11. Amazon Comprehend – Continuously Trained Natural Language Processing. – Access mode: https://aws.amazon.com/ru/blogs/aws/amazon-comprehend-continuously-trained-natural-language-processing/ (date of request: 27.02.2024). — Title from the screen.

12. Natural Language AI. – Access mode: https://cloud.google.com/natural-language. (date of request: 27.02.2024). — Title from the screen.

13. Chattermill. – Access mode: https://www.predictiveanalyticstoday.com/chattermill/. (date of request: 27.02.2024). — Title from the screen.

14. Моніторинг соцмедіа з візуальними інсайтами. – Access mode: https://youscan.io/. (date of request: 27.02.2024). — Title from the screen.

15. Logistic regression – Access mode: https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression. (date of request: 27.02.2024). — Title from the screen.

16. langdetect 1.0.9. – Access mode: https://pypi.org/project/langdetect/. (date of request: 27.02.2024). — Title from the screen.

17. emoji 2.10.1. – Access mode: https://pypi.org/project/emoji/. (date of request: 27.02.2024). — Title from the screen.

18. NLTK Documentation. – Access mode: https://www.nltk.org/api/nltk.tokenize.html. (date of request: 27.02.2024). — Title from the screen.

19. String — Common string operations. – Access mode: https://docs.python.org/3/library/string.html. (date of request: 27.02.2024). — Title from the screen.

20. nltk.corpus package – Access mode: https://www.nltk.org/api/nltk.corpus.html#nltk-corpus-package. (date of request: 27.02.2024). — Title from the screen.

21. Code of the own software implementation of the sentiment analysis algorithm of messages and comments in social networks. Access mode: https://github.com/mpakeron/FB_Sentiment_Anaysis/blob/b8a29da10ce4f25d6d080bab9480b7637f353ef8/Facebook.ipynb (date of request: 27.02.2024). — Title from the screen.

22. Ukrainian-Stopwords. – Access mode: https://github.com/skupriienko/Ukrainian-Stopwords. (date of request: 27.02.2024). — Title from the screen.

23. Tokenize UK. – Access mode: https://github.com/lang-uk/tokenize-uk. (date of request: 27.02.2024). — Title from the screen.

24. Ukrainian Stemmer. – Access mode: https://github.com/Desklop/Uk_Stemmer. (date of request: 27.02.2024). — Title from the screen.

25. sklearn.feature_extraction.text.TfidfVectorizer. – Access mode: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html#sklearn.feature_extraction.text.TfidfVectorizer. (date of request: 27.02.2024). — Title from the screen.