DOI: https://doi.org/10.36910/6775-2524-0560-2024-54-22
УДК 004.89
Проніна Ольга Ігорівна, к.т.н., доцент
https://orcid.org/0000-0001-7085-8027
Голубець Артур Олексійович, магістр
https://orcid.org/0000-0002-9944-0430
Державний вищий навчальний заклад «Приазовський державний технічний університет», м. Дніпро/м.Маріуполь, Україна

МАТЕМАТИЧНА МОДЕЛЬ ФОРМУВАННЯ ВЕКТОРНОГО УЯВЛЕННЯ УКРАЇНОМОВНОГО ТЕКСТУ

Проніна О.І., Голубець А.О. Математична модель формування векторного уявлення україномовного тексту. У статті розглядається система яка визначає тональність тексту надісланого користувачем. Метою даної статті є побудова математичної моделі обробки текстових повідомлень, та визначення настрою в поставленому реченні. Використання даного модулю визначення тональності україномовного тексту може відігравати ключову роль у різних галузях та має велику перспективу для подальшої модернізації та розвинення в різних сферах діяльності. У цій роботі розроблено ефективну модель та датасет для обробки моделлю машинного навчання BERT. Заборонована модель та функція досягла досить високих показників в точності визначення тональності та настрою користувача. Модуль, який запропоновано на даний момент реалізована в форматі боту для соціальної мережі, яка визначає тональність у реальному часі.
Ключові слова: векторизація тексту, система визначення тональності, обробка речень, робота з Telegram.

Pronina O.I., Golubets A.O. Mathematical model of formation of vector representation of Ukrainian-language text. In the article we consider a system that determines the tone of a text sent by a user. The aim of this article is to build a mathematical model of text message processing, and determining the mood of a sentence. Using of this module for determining the tone of the Ukrainian-language text can play a key role in various industries. a key role in various industries and has great prospects for further modernisation and development in various fields of activity. In this paper, we have developed an efficient model and dataset for processing by the BERT machine learning model. The proposed model and function achieved quite high performance in accurately determining the user's tone and mood. The module that proposed is currently implemented in the format of a bot for the social network, which determines the tone in real time.
Keywords: vectorisation text vectorisation, tone detection system, sentence processing, work with Telegram.

Постановка наукової проблеми. Системи виявлення тональності та векторизації тексту взагалі має досить широке значення та потреби. Головною проблемою та актуальністю на даний момент є те що данні системи наразі не дуже розповсюдженні, тим паче для україномовного тексу. Системи виявлення тональності існують та функціонують на разі дуже добре усього на англійській мові. Для інших мов майже не існую систем, які б коректною виявляли та розпізнавали емоційний стан користувача. Для українського тексту на разі існує всього 2 аналоги бібліотек, які мають його підтримку:

– пакет обробки української мови в бібліотеці SpaCy [1], яка представляє з себе систему розпізнавання та векторизації природної мови;

– бібліотека роботи для векторизації мови StyloMetrix [2], який призначений для векторизації та роботи з польською мовою, але також має можливості для роботи з англійською та українською мовами.

Ці бібліотеку, є молодими та досить прогресивними, але роботу які вони можуть виконувати – усього векторизації тексту на українській мові. Подальша робота з цими текстом в бібліотеках не розглядається що є досить суттєвим мінусом для їх використання.

Аналіз останніх досліджень та публікацій. Формування структури даних є найголовнішим етапом для підготування, далі до обраного датасету вже приєднуються основні методи роботи. Перший метод роботи є векторизація обраного тексту який отриманий. У роботі [3] було розглянуто векторизацію тексту за допомоги двох методів направлених на обробку та векторизацію тексту Bag-of-Words та TF-IDF. Розглянуті методи були опрацьовані на невеликому переліку даних, автор зробив висновок, що векторизація тексту слугує основою NLP, дозволяючи машинам обробляти та витягувати значення з текстових даних. Використовуючи такі методи, як Bag-of-Words, TF-IDF, вбудовування слів і документів, система може перетворити необроблений текст на числове представлення, що полегшує вирішення різноманітних завдань NLP. Та надає можливості для подальшої обробки тексту у необхідному вигляді.

Для подальшої роботи для виявлення тональності на далі необхідно обрати перш за все тип якої самої тональності нам необхідно виявити. У роботі [4] автор зробив повний перелік існуючих

типів тональності для кожної з яких було зроблена деяка перевірка та визначена ефективність у обраній схемі. Також автором було запропоновано приблизний перелік та послідовність робіт яку необхідно провести для роботи.

Для токінізації автор пропонує біль широкий перелік методів на відміну від попереднього, але ці методи, а саме Bag-of-Words, TF-IDF були також додані для формування роботи та векторизації тексту.

В заключному етапі було розглянуто основи тонального аналізу, починаючи з поняття, типів, підходів та проблем тонального аналізу. Також розглянуто алгоритми аналізу тональності та кроки для створення та оцінки моделей тонального аналізу з приблизними рекомендаціями до використання та сфер діяльності.

В статті [5] було розглянуто важливість аналізу складових частин речення в ході векторизації тексту. Автор статті запропонував робити оброку та перевірку вхідних даних за допомоги методу POS (Маркування частин речення). Проробляю данні, автор брав за основу вибору вхідних даних з коментарів, підписів та постів політиків Facebook. Провівши випробування автор виявив, що використання готових пакетів і словників для аналізування прикметників може бути ефективним і економічно вигідним методом для дослідження емоційного забарвлення та суб'єктивності у текстах, уникаючи необхідності великої розміченої навчальної бази даних. Проте, важливо підходити з обережністю при використанні цих інструментів для висновків. Використання прикметників як індикаторів суб'єктивності та настрою може бути не завжди надійним у різних контекстах, і зв'язок між окремими прикметниками та емоційним забарвленням не завжди може бути універсальним. Тому перед формулюванням остаточних висновків про зміст тексту важливо перевірити і валідувати цю модель, використовуючи допомогу фахівців у галузі програмування або інтегруючи аналіз з іншими методами визначення суб'єктивності.

Для розширення роботи з моделями для розрізання тексту автор статті [6] провів порівняння роботи моделі BERT та тронформерів. В статі за допомогою практичних реалізацій автор зробив висновки що за допомоги моделі BERT системи векторизації та з трасформерами система дає більш високі показники при навчанні, та дає можливість створення систем глибокого навчання. Ця стаття лягла в основу попередньої роботи, на базі якої було проведено аналізування інших великих датасетів.

На разі велика кількість бібліотек та статей, які були розглянуті націлені включно на використання та роботу нажаль з англомовним текстом та слегновими виразами. У статті [7] було знайдено приблизне вирішення проблеми з пошуком даних для роботи з не дуже розповсюдженими мовами. Автор пропонує та розказую про навчання дата сету безпосередньо на готових текстах з різних сайтах та інформаційних ресурсах. У цій статті запропоновано модель розпізнавання емоцій на основі тексту.

Запропонована модель є поєднанням підходів глибокого та машинного навчання. Запропонований гібридний підхід використовує комбінацію трьох наборів даних, а саме: ISEAR, WASSA та набір даних Emotion-Stimulus. Запропонована модель має багато переваг, оскільки вона може працювати на багатотекстових реченнях, твітах, діалогах, ключових словах і словах лексикону емоцій, які можна легко розпізнати. Згідно з класифікатором ML, SVM дає найвищу точність - 78,97%. У методі DL модель Bi-GRU досягає найвищої точності 79,46%, а модель CNN досягає найвищого показника F1 - 80,76. Гібридна модель досягла точності 82,39, пригадування - 80,40, оцінка F1 – 81,27, а точність – 80,11%.

На тему векторизації та обробку є безліч різних статей по роботі, але всі вони більш менш розглядають роботи однакових методів роботи з текстом там методами векторизації.

Мета дослідження. Підвищенні ефективності обробки україномовного тексту за рахунок розроблених методів які будуть аналізувати тональність та емоційний настрій тексту.

Виклад основного матеріалу й обґрунтування отриманих результатів дослідження. Побудова програми базується на базовій схемі створення програм з векторизацією тексту. Основний принцип роботи складає з функцій підготовки тексту для обробки, векторизація тексту, обробка тексту для визначення тональності та передача даних до інтерфейсу, який вже видає висновок. Принцип роботи можна побачити у рисунках 1-3.

Рис. 1 – Схема роботи алгоритму визначення тональності тексту

Рис. 2 – Схема роботи векторизації тексту

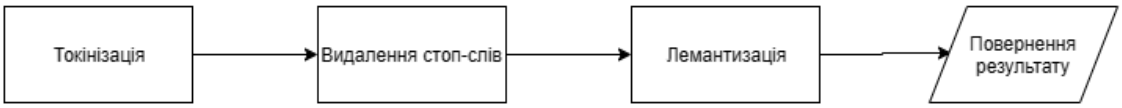

Рис. 3 – Схема попередньої обробки тексту

Після проведення попередньої підготовки в системі виконується векторизації та призначенню тексту ваги та значимості. Переведений текст до векторного вигляду, система передає всі дані до класифікатора, який виявляє який емоційний стан має кожне слово та загальна його вага. На виході користувач отримує значення до якого типу відноситься текст, позитивного або негативного та найбільше його значення.

Для роботи з текстом використовується 2 математичні моделі. По-перше математична модель роботи з текстом. В цій моделі наведена структура формування речення.

Текст повідомлення можна представити у вигляді множини речення, а речення в свою чергу складається з множини слів потужність якої дорівнює їхній кількості:

$$T = \{\{t_1, t_2, \dots, t_i, \dots, t_w\};\ w = |T|, \tag{1.1}$$

де T – це множена слів в тексті,

$t_1, t_2, \dots, t_i, \dots, t_w$ – слова тексту,

w – кількість слів в заданому текстів повідомлення.

Своєю чергою, усю множину слів тексту можна подати у вигляді об'єднання підмножин різних частин мови, при цьому кожне слово тексту може бути віднесено до однієї з цих підмножин:

$$T = \bigcup_{j=1}^{k} C_i\ ;\ t_i \in C_i;\ \ i = \overline{1, w};\ \mathrm{j} = \overline{1, \mathrm{k}}, \tag{1.2}$$

де C_i – під множина j-ої частини мови,

k – ккількість частин мови, які існують.

При подальшій роботів з цієї математичною моделюю, система отримує кожне речення окремо, далі кожне речення береться за основу та розбиваються на множину слів, які вже далі передаються до наступної моделі та визначається тональність речення.

Для визначення тональності у даній роботі важливим показником є безпосередньо кожне слово, та речення, в контексті якого базується слово.

Процес векторизації тексту та його робота працює на базі двох відомих методів – мішка слів (Bag of Words) формула 1.3 та TF-IDF формула 1.4 - 1.6, які найбільш повно проходять векторизацією заданого тексту.

$$\mathrm{BoWd} = (\mathrm{f1, d, f2, d, \dots, fN, d}), \tag{1.3}$$

де fi,d – це частота токена і у документі d,

N – кількість унікальних токенів у всій колекції документів D.

Перша за все необхідно виявити наскільки часто зустрічається. визначає, як часто певне слово зустрічається в документі. Вона розраховується за формулою 1.4.

$$\mathrm{TF(t, d)} = \frac{\text{Кількість входжень терміна t у документі d}}{\text{Загальна кількість слів у документі d}}, \tag{1.4}$$

де t – це термін який необхідно перевірити на кількість входжень,

d – документ або повідомлення яке аналізує.

Після аналізування кількості входження та формування множини термінів, які використовує поставлене речення необхідно визначити значимість цього терміну. Процес вимірює загальну значимість терміна. Високе значення IDF означає, що термін не часто зустрічається в корпусі. Це розраховується за формулою 1.5.

$$\mathrm{IDF(t, D)} = \log\left(\frac{\text{Загальна кількість документів у корпусі D}}{\text{Кількість документів,що містять термін t}}\right), \quad (1.5)$$

де t – кількість вхідних термінів, які опрацьовуються;

D – весь документ який оцінюється.

Після проведення аналізу значимості вхідних даних, перевіряємо та отримуємо показник TF-IDF шляхом отримання добутку величин. Для отримання добутку необхідно скористатися формулою 1.6.

$$\mathrm{TF - IDF(t, d, D)} = \mathrm{TF(t, d)} \times \mathrm{IDF(t, D)}, \quad (1.6)$$

де t – перелік термінів, які використовуються в аналізі,

D – повний документ або текст який аналізується,

d – речення або вираз токену,

TF – функція пошуку частих слів,

IDF – функція виявлення ваги слів.

Таким чином, TF-IDF збільшується пропорційна кількості разів, як термін зустрічається в документі, але компенсується частотою терміна в корпусі, що допомагає знизити вплив часто вживаних слів.

Для подальшої роботи та навчання моделі за мобами машинного навчання використовують методи моделі BERT. BERT (Bidirectional Encoder Representations from Transformers) використовує досить складну архітектуру, яка базується на трансформерах. Для розрахунку моделі використовувалось формули множинного механізму уваги (1.7), за допомоги якого було основана формування та навчання попередньої моделі. Сам розрахунок опирається на звичайний механізм уваги формула 1.7.

$$\mathrm{Attention(Q, K, V)} = \mathrm{softmax}\left(\frac{\mathrm{QK^2}}{\sqrt{\mathrm{d_k}}}\right)\mathrm{V}, \quad (1.7)$$

де Q – Матриця запитів до моделі,

K – Перелік ключів у матриці,

V – Значення матриці за відповідним ключем,

$\mathrm{d_k}$ – Розмірність використовуємих ключів.

Формування множинного механізму уваги базується на формулі зазначеній вище. Невідмінну від формули 1.7 множинний вибір буде приймати множену змінних які отримуються у ході формування одиничного механізму.

$$\mathrm{MultiHead(Q, K, V)} = \mathrm{Concat(head_1, head_2, \dots, head_h\ )W^O}, \quad (1.8)$$

де $\mathrm{Attention(Q, K, V)}$ – формула одиничного вибору.

BERT використовує спеціальний токен [CLS] для класифікаційних завдань та [SEP] для розділення різних сегментів вхідного тексту.

Попереднє навчання складається з декількох частин цієї моделі. По-перше включає два головних завдання: Маскований LM (Masked Language Model, MLM) та Наступне Предикативне Завдання Речення (Next Sentence Prediction, NSP).

По-друге, MLM закриває частину вхідних токенів та намагається передбачити найбільш імовірні наступні пункти, а далі за допомоги NSP намагається передбачити, чи є друге речення логічним продовженням першого.

Для тестування системи було проведено експерименти з визначенням тональності простих речень, речень з синонімами, жаргонні речення, саркастичні речення. Система показала наступні результати наведені у таблиці 1.

Таблиця 1 – Показники тестування

Тип речення	Відсоток вдалих визначень
Прості речення	80%
Речення з заміною синонімами	80%
Жаргоні речення та сленг	70%
Саркастичні	50%

При визначенні роботи було проведено також тестування на якісність та впевненість системи в оцінці. Результати наведено у рисунку 4.

На даному графіці наведено кучність аналізу 400 простих повідомлень, усі повідомлення розділяються на 3 основні групи, а саме: позитивні речення – це речення, які мають показники позитивного більше аніж 70%, негативні речення – це речення з показниками негативного настрою більше 70%. Та невизначені, дані повідомлення мають позитивний або негативний висновок, в залежності від більшого показника, але більш за все ці повідомлення вважаються не визначеними.

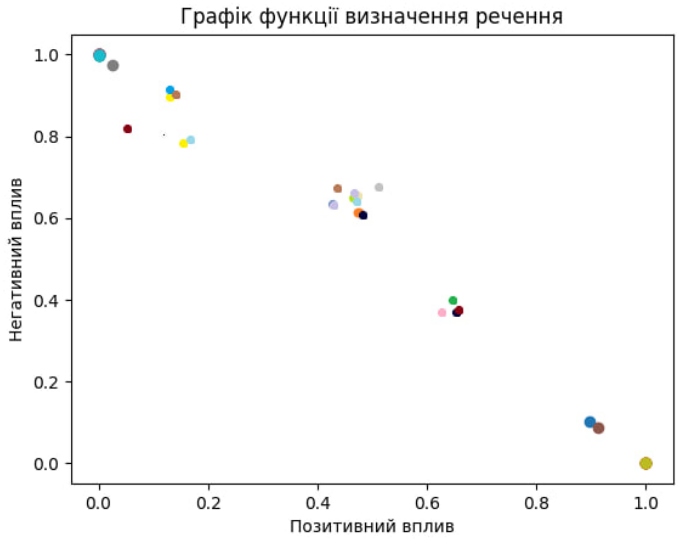

Рис. 4 – Графік кучності оцінки простих речень

Також було перевірено швидкісні можливості при навантаженні системи, які виявились досить лінійними. Вони показали, що на маленьких кількостях речень для обробки система обробляє 1 речення з середньою кількістю 15 слів. Приблизно 1 секунду. При підвищенні кількості цей показник зростає. Результати наведені на рисунку 5.

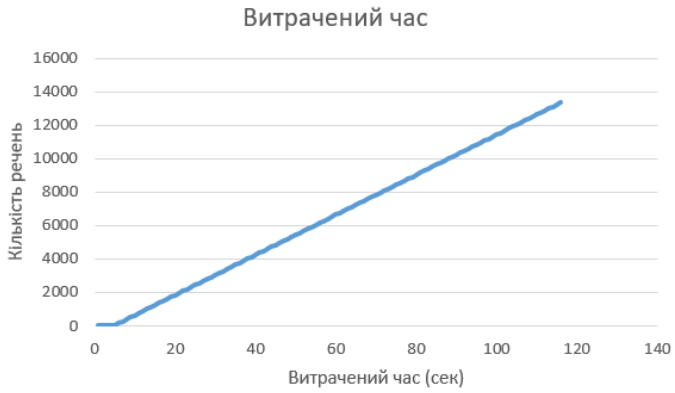

Рис. 5 – Час роботи

Аналізуючи проведені експерименти можна побачити що система працює досить вдало та система якісно може зробити аналіз тональності тексту, які отримує від користувача. Час роботи системи також досить приємний для роботи особливо на великій кількості речень.

Висновки та перспективи подальшого дослідження. В ході роботи було реалізовано математична модель, датасет та програмне забезпечення для визначення тональності повідомлення Telegram. Застосування цього модуля має досить високу практичну здатність, модуль можна використовувати для аналізу переписок з клієнтами в інтернет магазинах, аналіз коментарів на фільми, товари, послуги, аналіз закладу та багато інших сфер, у яких необхідно робити оцінку саме текстових даних. Головною перевагою є те, що бібліотека працює з кількома мовами одночасно та працює з українською мовою що є найбільш значним плюсом в роботі цього модуля.

Крім того, розробка та подальше вдосконалення цього модуля зможе знайти своє застосування у сфері штучного інтелекту, який допомагає аналітичним платформам та платформам, які роблять оцінки якості. Розроблений датасет, який можна створити на основі розробленої моделі може повністю змінити сенс програми та збільшити її сферу використання що робить його більш ефективним та гнучким для роботи та адаптування під потреби.

Головним аспектом розробки є простота та багатофункціональність, що було повністю реалізовано та дозволило підвищити ефективність модуля у рази, а реалізація API робить його досить мобільним для використання

Список бібліографічного опису

1. Industrial-StrengthNaturalLanguageProcessing. – Режим доступу: https://spacy.io/
2. Stylo Metrix: AnOpen-SourceMultilingual Toolfor RepresentingStylometric Vectors - Режим доступу: https://github.com/ZILiAT-NASK/StyloMetrix?ysclid=lqo420yxdm334651053
3. Iliev, Rumen&Dehghani, Morteza&Sagi, Eyal. (2014). Automated Text Analysisin Psychology: Methods, Applications, andFutureDevelopments. Language and Cognition. 10.1017/langcog.2014.30.
4. Yeh, Cheng-Yu&Hwang, Shaw-Hwa. (2019). Efficient Detection Approach for DTMF Signal Detection. Applied Sciences. 9. 422. 10.3390/app9030422.
5. Acheampong, Francisca&Chen, Wenyu&Nunoo-Mensah, Henry. (2020). Text-BasedEmotionDetection: Advances, ChallengesandOpportunities.
6. Sentiment Analysiswiththe Use of Transformersand BERT Conference PaperJul 2023
7. Emotion Detectionand Recognition from Text Using DeepLearning. August 2022Co mputational Intelligenceand Neuroscience 2022(45-60):1-8
8. Khrystyna Shakhovska, Iryna Dumyn, Natalia Kryvinska, MohanKrishnaKagita, "An Approachfor a Next-Word Prediction for Ukrainian Language", Wireless Communications and Mobile Computing, vol. 2021, Article ID 5886119, 9 pages, 2021
9. Maruf, Abdullah&Khanam, Fahima&Haque, Md&Masud, Zakaria. (2022). Emotion Detection from Textand Sentiment Analysis of Ukraine Russia Warusing Machine Learning Technique. International Journal of Advanced Computer Science and Applications. 13. 868-882. 10.14569/IJACSA.2022.01312101
10. Zador, Anthony&Escola, Sean&Richards, Blake&Ölveczky, Bence&Bengio, Yoshua&Boahen, Kwabena&Botvinick, Matthew&Chklovskii, Dmitri&Churchland, Anne&Clopath, Claudia&DiCarlo, James&Ganguli, Surya&Hawkins, Jeff&Körding, Konrad&Koulakov, Alexei&LeCun, Yann&Lillicrap, Timothy&Marblestone, Adam&Olshausen, Bruno&Tsao, Doris. (2023). Catalyzingnext-generationArtificial Intelligence through NeuroAI. Nature Communications. 14. 10.1038/s41467-023-37180-x.
11. Yang, Siqin&Cai, Yeyi&Xie, Wen&Jiang, Minghu. (2021). Semantic and Syntactic Processing During Comprehension: ERP Evidence From Chinese QING Structure. Frontiersin Human Neuroscience. 15. 10.3389/fnhum.2021.701923.

References

1. Industrial-StrengthNaturalLanguageProcessing. — Link: https://spacy.io/
2. Stylo Metrix: AnOpen-SourceMultilingual Toolfor RepresentingStylometric Vectors - Link: https://github.com/ZILiAT-NASK/StyloMetrix?ysclid=lqo420yxdm334651053
3. Iliev, Rumen&Dehghani, Morteza&Sagi, Eyal. (2014). Automated Text Analysisin Psychology: Methods, Applications, andFutureDevelopments. Language and Cognition. 10.1017/langcog.2014.30.
4. Yeh, Cheng-Yu&Hwang, Shaw-Hwa. (2019). Efficient Detection Approach for DTMF Signal Detection. Applied Sciences. 9. 422. 10.3390/app9030422.
5. Acheampong, Francisca&Chen, Wenyu&Nunoo-Mensah, Henry. (2020). Text-BasedEmotionDetection: Advances, ChallengesandOpportunities.
6. Sentiment Analysiswiththe Use of Transformersand BERT Conference PaperJul 2023
7. Emotion Detectionand Recognition from Text Using DeepLearning. August 2022Co mputational Intelligenceand Neuroscience 2022(45-60):1-8
8. Khrystyna Shakhovska, Iryna Dumyn, Natalia Kryvinska, MohanKrishnaKagita, "An Approachfor a Next-Word Prediction for Ukrainian Language", Wireless Communications and Mobile Computing, vol. 2021, Article ID 5886119, 9 pages, 2021

9. Maruf, Abdullah&Khanam, Fahima&Haque, Md&Masud, Zakaria. (2022). Emotion Detection from Textand Sentiment Analysis of Ukraine Russia Warusing Machine Learning Technique. International Journal of Advanced Computer Science and Applications. 13. 868-882. 10.14569/IJACSA.2022.01312101
10. Zador, Anthony&Escola, Sean&Richards, Blake&Ölveczky, Bence&Bengio, Yoshua&Boahen, Kwabena&Botvinick, Matthew&Chklovskii, Dmitri&Churchland, Anne&Clopath, Claudia&DiCarlo, James&Ganguli, Surya&Hawkins, Jeff&Körding, Konrad&Koulakov, Alexei&LeCun, Yann&Lillicrap, Timothy&Marblestone, Adam&Olshausen, Bruno&Tsao, Doris. (2023). Catalyzingnext-generationArtificial Intelligence through NeuroAI. Nature Communications. 14. 10.1038/s41467-023-37180-x.
11. Yang, Siqin&Cai, Yeyi&Xie, Wen&Jiang, Minghu. (2021). Semantic and Syntactic Processing During Comprehension: ERP Evidence From Chinese QING Structure. Frontiersin Human Neuroscience. 15. 10.3389/fnhum.2021.701923.