https://www.proconference.org/index.php/usc/article/view/usc33-00-018
DOI: 10.30888/2709-2267.2025-33-00-018

УДК 004.8:004.738.5

MULTIMODAL AUDIO ANALYSIS IN SOCIAL MEDIA: AN AI-DRIVEN APPROACH TO EMOTIONAL INSIGHT

МУЛЬТИМОДАЛЬНИЙ АНАЛІЗ АУДІО В СОЦІАЛЬНИХ МЕРЕЖАХ: ПІДХІД НА ОСНОВІ ШТУЧНОГО ІНТЕЛЕКТУ ДЛЯ РОЗУМІННЯ ЕМОЦІЙ

Tumanov O.O. / Туманов О.О.
PhD / доктор філософії
ORCID: 0000-0003-0674-0037
V.N.Karazina Kharkiv National University, Kharkiv, Maydan Svobody, 4, 61022
ХНУ ім.В.Н.Каразіна, Харків, Майдан Свободи, 4, 61022

Анотація. *У цій статті представлено підхід на основі штучного інтелекту до мультимодального аналізу аудіоконтенту в соціальних мережах, який виходить за рамки традиційного аналізу тексту. Методологія поєднує вилучення акустичних ознак з використанням глибокого навчання та моделей великих мов (ВММ) для класифікації та синтезу емоційних даних. Особлива увага приділяється аналізу відповідності між мовленням, текстом та фоновими звуками для виявлення складних емоційних станів, таких як сарказм. Описано практичне застосування методу в охороні здоров'я (моніторинг психічного здоров'я) та маркетингу, а також обговорено пов'язані етичні аспекти. На завершення, цей підхід надає дослідникам науково обґрунтовані інструменти для створення точних прогностичних моделей поведінки.*

Ключові слова: *великі мовні моделі, мультимодальний аналіз, соціальні медіа, аналіз аудіо, ШІ.*

Abstract. *This paper presents an artificial intelligence-based approach to multimodal analysis of audio content in social media that goes beyond traditional text analysis. The methodology combines acoustic feature extraction with deep learning and Large Language Models (LLMs) for classification and synthesis of emotional data. Particular attention is paid to analyzing correspondences between speech, text, and background sounds to detect complex emotional states, such as sarcasm. Practical applications of the method in healthcare (mental health monitoring) and marketing are described, and related ethical aspects are discussed. Finally, this approach provides researchers with scientifically sound tools for creating accurate predictive models of behavior.*

Keywords: *Large Language Models, Multimodal Analysis, Social Media, Audio Analysis, AI.*

Вступ.

Поширення аудіо контенту соціальних медіа (голосові повідомлення, відео-блоги, подкасти) разом із розвитком штучного інтелекту (ШІ) та вдосконаленої обробки природної мови, свідчить про те, що традиційний аналіз, що ґрунтується винятково на тексті, є недостатнім [7]. Поява великих мовних моделей (ВММ) та складних методів обробки створила нову парадигму для аналізу мультимодальних даних. Інтегруючи акустичний аналіз з цими моделями, ми можемо вилучити не лише лексичний зміст, а й додаткові ознаки, які є потужними індикаторами емоційного стану. Це виводить аналіз поведінки

аудиторії на якісно новий рівень, надаючи дослідникам доступ до раніше прихованої емоційної інформації.

Метою цього дослідження є детальний аналіз основних параметрів аудіоконтенту та методів їх вивчення. Робота зосереджена на науковому підході до аналізу, включаючи вивчення фонової музики та звуків, які користувач свідомо додає у свій контент.

Виклад основного матеріалу. На першому етапі аналізу аудіоконтенту відбувається вилучення акустичних ознак. На відміну від простої конвертації мови в текст, науковий підхід передбачає вилучення з аудіо-сигналу комплексу статистичних параметрів, які несуть інформацію про емоції. Ці ознаки можна розділити на дві основні групи (таблиця 1).

Таблиця 1 – Класифікація акустичних ознак аудіоконтенту

Тип ознак	Опис	Приклади
Низькорівневі (Low-Level Descriptors, LLDs)	Базові фізичні характеристики звукового сигналу, що вимірюються на коротких часових інтервалах (фреймах).	- **Висота тону** (Pitch, f0): основна частота, що відображає емоції (радість/гнів – високий, смуток – низький). - **Енергія/Інтенсивність** (RMSE): гучність звуку, яка вказує на силу емоцій. - **Мелчастотні кепстральні коефіцієнти (MFCC)**: набір коефіцієнтів, що імітують сприйняття звуку людським вухом. - **Швидкість мовлення**: кількість складів/слів за одиницю часу, що свідчить про стан людини (тривога чи смуток).
Високорівневі (High-Level Descriptors)	Статистичні узагальнення низькорівневих ознак, що розраховуються на рівні всього висловлювання.	- Середнє значення, дисперсія, максимум/мінімум для кожного LLD. - Коефіцієнти кореляції між різними LLDs. - Індекс гармонійності: міра «музичності» звуку

Джерело: авторська розробка на базі [1, 3, 4, 5].

Збір акустичних ознак є критичним кроком у мультимодальному аналізі, оскільки їхня якість безпосередньо впливає на точність подальшого аналізу. Ці ознаки можна розділити на дві основні групи. Низькорівневі ознаки – це основні фізичні характеристики, такі як висота тону, енергія та MFCC, які вимірюються через короткі інтервали часу. На противагу цьому, високорівневі ознаки – це статистичні узагальнення, розраховані на рівні всього висловлювання, такі як середнє значення або дисперсія низькорівневих ознак.

Наступний етап полягає у класифікації та статистичному моделюванні вилучених ознак з аудіоконтенту з використанням ШІ. Після вилучення ознак вони подаються до статистичних моделей для класифікації. Цей етап є найважливішим у нашому дослідженні.

На відміну від простої класифікації «позитивне/негативне», науковий підхід дозволяє використовувати багатовимірні моделі емоцій, такі як модель валентності-збудження-домінування (VAD) [6], де:

- Валентність (Valence): Ступінь позитивності або негативності емоції.
- Збудження (Arousal): Ступінь інтенсивності емоції.
- Домінування (Dominance): Ступінь контролю, який людина відчуває над ситуацією.

Цей підхід дозволяє представити кожну емоцію як точку в тривимірному просторі, що набагато точніше, ніж дискретні категорії.

Для проведення такого аналізу використовуються різноманітні статистичні моделі та алгоритми. Традиційні методи машинного навчання, такі як метод опорних векторів (SVM) [2], логістична регресія та k-найближчих сусідів (KNN), ефективні для класифікації на основі витягнутих ознак. Більш просунуті моделі глибокого навчання, такі як рекурентні нейронні мережі (RNN) з довгою короткочасною пам'яттю (LSTM) та трансформатори, здатні аналізувати послідовності акустичних ознак, автоматично виявляючи складні емоційні патерни та підвищуючи точність класифікації.

Нещодавні досягнення також призвели до використання великих мовних моделей (LLM), які можна додатково навчити обробляти та інтегрувати

акустичні та візуальні дані. Це дозволяє проводити більш точні класифікації та навіть генерувати емоційні резюме. Наприклад, ВММ може аналізувати транскрипцію, акустичні параметри та музичні особливості, щоб сформувати детальний опис емоційного стану користувача, наприклад: «Мовець звучить розчаровано, незважаючи на позитивний тон тексту, що вказує на саркастичний підтекст, який підкріплюється веселою, але дещо недоречною фоновою музикою».

Далі йде фаза інтеграції та синтезу даних, що керовані ШІ. Цей етап є кульмінацією наукового підходу. Емоційні дані з аудіо (наприклад, оцінки VAD) можна інтегрувати з аналізом тексту та, що найважливіше, з аналізом музики або звукових ефектів. Це дозволяє нам виявляти не лише розбіжності, але й синергію між різними модальностями.

Аналіз розбіжностей: Це включає виявлення конфліктів між емоційним змістом тексту, мовлення та фонової музики.

Аналіз синергії: Це визначає, як емоційний вираз у тексті та голосі підкріплюється фоновою музикою, підсилюючи загальне повідомлення.

Мультимодальна Fusion модель: Кількісна оцінка емоційних даних з різних модальностей значно покращується за допомогою мультимодальних моделей штучного інтелекту. Замість того, щоб покладатися виключно на статистичні коефіцієнти кореляції, така модель може вивчати складні, нелінійні зв'язки між текстом, голосом та фоновою музикою. Це особливо корисно для виявлення тонких розбіжностей та синергії. Наприклад, великі мовні моделі (ВММ) можна налаштувати для безпосереднього порівняння емоційних станів, виражених у різних модальностях, та надання оцінок ймовірності саркастичної або іронічної підтексту, ефективно автоматизуючи те, що раніше було ручним процесом аналізу розбіжностей.

Як уже зазначалося, *кількісне вимірювання розбіжності* є ключовим етапом. Це можна зробити, ввівши індекс конгруентності (C_i) (1), який розраховується як кореляція між емоційною оцінкою тексту (T_i) та емоційною оцінкою аудіо (A_i).

$$C_i = corr\ (T_i, A_i) \quad (1)$$

Низька кореляція може вказувати на такі приховані емоційні стани, як сарказм, іронія або пригнічення емоцій. Ці розбіжності можуть бути статистично змодельовані як нова змінна для подальшого аналізу. Окрім того, використовуючи методи кластерного аналізу (наприклад, k-means, DBSCAN), можна виділити групи користувачів, які часто демонструють подібну поведінку [8]. Це дозволить не лише виявити саму розбіжність, але й ідентифікувати типові поведінкові патерни в різних сегментах користувачів.

Практичні Застосування та Етичні Аспекти. Використання результатів аудіоаналізу можливе у різних сферах. Так, наприклад:

- *у маркетингу та опитуваннях:* аудіоаналіз може допомогти брендам не просто зрозуміти, що говорять їхні клієнти (наприклад, у голосових відгуках), а й як вони це говорять, виявляючи рівень задоволення, розчарування або ентузіазму.
- *у сфері охорони здоров'я:* постійне виявлення розбіжності між вираженими емоціями та фоновим звуком може бути використано як статистичний індикатор для моніторингу ментального здоров'я. Наприклад, значні та постійні розбіжності можуть вказувати на ризик депресії або тривожних станів. Така система може стати частиною прогностичної моделі в телемедицині.
- *в ігровій індустрії:* аналіз голосу гравців може допомогти іграм автоматично адаптувати складність або сюжетну лінію, реагуючи на рівень фрустрації або радості.

Етичні Міркування. Робота з аудіоданими, особливо з чутливими емоційними параметрами, вимагає суворого дотримання етичних норм:

- **Конфіденційність:** Необхідно забезпечити анонімність даних і отримати інформовану згоду користувачів на збір та аналіз їхніх голосових даних.
- **Упередженість алгоритмів:** Як і будь-яка модель ШІ, моделі емоційного аналізу можуть бути упередженими, якщо їх тренують на незбалансованих наборах даних. Це може призвести до некоректної інтерпретації емоцій у

певних групах населення.

Висновки.

У цій статті ми продемонстрували, що мультимодальний аналіз аудіоконтенту є значно більш потужним, ніж традиційний текстовий аналіз. Інтеграція низькорівневих і високорівневих акустичних ознак зі статистичними моделями та алгоритмами ШІ дозволяє отримати глибокі інсайти в емоційний стан користувача. Кульмінацією цього підходу є керований ШІ аналіз конгруентності між різними модальностями — текстом, голосом та фоновими звуками. Такий підхід відкриває нові можливості для досліджень у галузі соціальних медіа, охорони здоров'я та маркетингу, надаючи науково обґрунтовані інструменти для кількісного вимірювання складних людських емоцій. Це є особливо актуальним, адже дозволяє створити більш точні та надійні прогностичні моделі поведінки.

Література

1. Anderson, C., Carpenter, C., Kranabetter, J. et al. *A Survey of High-Level Descriptors in Sound Design: Understanding the Key Audio Aesthetics Towards Developing Generative Game Audio Engines*. 2024.

2. Awad, M., Khanna, R. *Support Vector Machines for Classification*. 2015.

3. Dandashi, A., Aljaam, J. *A Survey on Audio Content-Based Classification*. 2017. P. 408–413. DOI: 10.1109/CSCI.2017.69.

4. Garg, A., Aribi, Y., Althobaiti, T. et al. *An analysis of acoustic features for accented speech classification*. Egyptian Informatics Journal. 2025. Vol. 31. P. 100743. DOI: https://doi.org/10.1016/j.eij.2025.100743.

5. Karjadi, C., Xue, C., Cordella, C. et al. *Fusion of Low-Level Descriptors of Digital Voice Recordings for Dementia Assessment*. J Alzheimers Dis. 2023. Vol. 96, № 2. P. 507–514. DOI: 10.3233/JAD-230560.

6. Işık, Ü., Güven, A., Batbat, T. *Evaluation of Emotions from Brain Signals on 3D VAD Space via Artificial Intelligence Techniques*. Diagnostics (Basel). 2023. Vol. 13, № 13. P. 2141. DOI: 10.3390/diagnostics13132141.

7. Tumanov, O. O. *Aspects of Using Social Media in Research*. Scientific Bulletin of the National Academy of Statistics, Accounting and Audit. 2019. № 4. P. 24–29. DOI: https://doi.org/10.31767/nasoa.4.2019.03.

8. Tumanov, O. O. *Statistical methods for analyzing social media data*. Business Inform. 2020. № 2. P. 266–272. DOI: https://doi.org/10.32983/2222-4459-2020-2-266-272 .

Статтю відправлено: 19.09.2025 г.

© Туманов О.О.