# МЕТОДОЛОГИЧНИ ПОДХОДИ В МУЛТИМОДАЛНИЯ СЕНТИМЕНТ АНАЛИЗ

## Methodological Approaches in Multimodal Sentiment Analysis

**Станимира Йорданова**[1]
*e-mail: syordanova@unwe.bg*[1]

**Абстракт**

*Мултимодалният сентимент анализ се утвърждава като изследователска област, която обхваща обработката на естествен език, компютърното зрение и обработката на реч. За разлика от традиционните подходи, които разчитат единствено на текст, мултимодалните подходи интегрират хетерогенни източници на информация като език, визуални сигнали и аудио характеристики, за давсе улови сложността на човешките емоции и мнения. Докладът представя преглед на набори от данни за мултимодалния сентимент анализ и извежда методологични архитектурни подходи, използвани в анализа. В докладът се обсъждат важните предизвикателства като синхронизацията между модалностите, дисбалансът в данните и зависимостта от контекста. Докладът има за цел да допринесе за развитието на устойчиви, мащабируеми и контекстно осъзнати модели за приложение в реални среди на сентимент анализ.*

**Abstract**

*Multimodal sentiment analysis is a research area that encompasses natural language processing, computer vision, and speech processing. Unlike traditional approaches that rely solely on text, multimodal methods integrate heterogeneous sources of information such as language, visual, and audio features to capture the complexity of human emotions and opinions. The paper presents popular datasets and methodological approaches used in the multimodal sentiment analysis architecture. Challenges such as synchronization between modalities, data imbalances, and contextual dependency are discussed. The paper aims to contribute to the development of more sustainable, scalable, and context-aware models for application in real-world sentiment analytics environments.*

**Ключови думи:** multimodal sentiment analysis, NLP, aspect-based sentiment analysis

## Въведение

Сентимент анализът е задача в обработката на естествен език (НЛП), която има за цел да се извлекат човешките емоции и мнения в комуникацията от текстови източници като отзиви за продукти и публикации в социалните медии. Тъй като общуването между хората е сложен и динамичен процес, който съчетава вербални и невербални знаци, традиционните системи за сентимент анализ, използвайки само текстови източници, често не успяват да уловят нюансите на емоциите, които се изразяват в общуването чрез взаимодействие на езиково съдържание, интонация на речта и изражения на лицето [1]. Тези ограничения неизменно водят до необходимостта от развитие на мултимодалния сентимент анализ, който има за цел да интегрира субективната информация от различните източници като текст, разговори и видеоклипове. Източниците на комуникация са богати на текстови, аудио и визуални данни и предоставят възможност за правилно определяне и извличане на мненията (чувства и емоции), изразени от хората за обекти и техните характеристики. Сложността на мултимодалния сентимент анализ произтича от необходимостта не само да се идентифицират конкретни обекти или характеристики (аспекти) в съдържанието, но и точно да се определят настроенията и мненията,

[1] Гл. ас. д-р към катедра ИТК, УНСС, email: syordanova@unwe.bg

свързани с всеки от тях, като се използват контекстуални знаци от текстова, визуална и аудио информация, както и да се постигне точно разбиране в нюансите на настроенията и мненията, осигурявайки цялостен поглед върху човешката комуникация. Ефективното извличане на настроенията и мненията чрез мултимодален сентимен анализ зависи до голяма степен от висококачествени, анотирани мултимодални набори от данни, като обработката им поставя предизвикателства пред способността на изкуствения интелект да интерпретира човешка комуникация.

Докладът има за цел да представи методологичните подходи в мултимодалния сентимент анализ от гледна точка на задачите в извличането на настроения и мнения от различни типове източници. В доклада се изследват характеристиките на необходимите данни и изискванията към тях, като се подчертава значението на процеса на анотиране на данни и постигането на представителност, количество и качество на данните. Представена е логическа архитектура на мултимодалния сентимент анализ, изследвани са методологични рамки за мултимодален сентимент анализ и са изведени предизвикателства, свързани със синхронизацията между модалностите, дисбаланса в данните, зависимостта от контекста и изискванията към изчислителната ефективност.

**Мултимодален сентимент анализ**

Традиционният подход за сентимент анализ, който има за цел да извлича мнение от текстов източник, класифицира мнението на три различни нива – *документ, изречение и аспект на обекта*. За разлика от сентимент анализа, който може да класифицира цял документ или изречение като положително, отрицателно или неутрално, мултимодалният сентимент анализ на ниво аспект цели идентифициране на конкретни обекти или характеристики на обекти (аспекти) в съдържанието и след това да определи точната полярност на мнението (напр. положителна, отрицателна или неутрална), свързана с извлечените конкретни аспекти. Този детайлен анализ позволява много по-богато и приложимо разбиране на мненията и емоциите, защото базира анализа на информация от различни типове източници.

Развитието на методологичните подходи за мултимодален сентимент анализ претърява еволюция, насочена към осъвършенстване на методите и алгоритмите за извличане на структура от различните типове данни с цел точното определяне на мнението. Ссентимент анализът преминава през няколко етапа, които се влияят от нарастването на данните в интернет пространството, развитието на методите за подготовка на данните и машинно обучение, появата на съвремени методи за класификация и развитието на изкуствения интелект. В ранния етап на своето развитие (преди 2010г.), сентимент анализът е фокусиран върху определяне на полярността на мнението на ниво документ, след което преминава към сентимент анализ на ниво изречение и ниво аспект. В този период се прилагат класически методи от машинно обучение, методи, използващи лексикони, както и хибридни методи. С появата на вграждането на думи и дълбокото обучение, се развиват методологичните подходи за откриване и извличане на аспекти и определяне на мнението по отношение на извлечените аспекти. След 2010г. се достига до признанието, че текстът сам по себе си е недостатъчен за сентимент анализ. Изображенията, аудио и видеото могат да добавят богат контекст, което подтиква появата на техники за мултимодален синтез или интегриране на данни за сентимент анализ от различни типове източници. [1]. През 2020г. напредъкът в развитието на трансформър архитектурите, обучени модели като BERT, RoBERTa и мултимодални трансформар архитектури (VisualBERT), прави възможно свързването на текст с изображение и извличане на съответствие на аспектите от текста с визуалните знаци от изображенията. След 2023г., въвеждането на големите езикови модели прави възможна интеграция им с мултимодални източници (текст, изображение, аудио) и дава насока на научната дейност към мултимодален сентимент анализ на ниво аспект, използващ промптове и инструкции към генеративните езиковите модели. Съвременните тенденции в развитието на мултимодалния сентимент анализ са насочени към използване на големи езикови модели за обяснимост, липса на езикови, звукови ресурси и изображения, както и приложения за конкретни области (напр. здравеопазване, електронна търговия, мониторинг на социалните медии).

**Източници на данни за мултимодален сентимент анализ**

B. Liu през 2012г. [2], определя мнението (sentiment, opinion) като съвкупност от четири елемента *обект/аспект на мнение*, *изразено чувство (feeling) за обекта/аспект на обекта*, *автор на мнението* и *датата, когато е изразено мнението*. В комуникацията си, човекът може да изразява *емоции* и *чувства.* От психологическа гледна точка *емоциите* обикновено са краткосрочни състояния, които се задействат от конкретни събития и включват последователност от когнитивни оценки, телесни реакции и изражения като движения на лицето или гласови знаци. Например моментен поглед на изненада или мимолетно изразяване на гняв. *Чувството* се определя като по-дълбоко, по-трайно разположение или установено мнение, често описвано като позиция към конкретна същност или концепция. Положителното чувство на човек към обект (например продукт или политическа партия) е гледна точка, която може да бъде идентифицирана чрез последователни афективни взаимодействия с обекта. Докато емоциите могат да бъдат преходни, чувствата са по-стабилни и могат да бъдат проследени чрез повтарящи се емоционални реакции. Това разграничение е важно, тъй като системите за мултимодален сентимент анализ могат да бъдат проектирани да идентифицират краткосрочни емоционални състояния или дългосрочни чувства (мнения).

Източниците на данни за мултимодален сентимент анализ са ***текст, изображение и аудио***.

- *Текстовите източници* предоставят съдържание и контекстуална информация, която е в основата на идентифицирането на аспекти и мнение. Въпреки това, в текста може да бъде изразено двусмислие, сарказъм или ирония, които често се пропускат в моделите за сентимент анализ от текстови източник.
- *Аудио източниците* предоставят информация за знаци, които липсват в писмения текст като емоционален тон, предаван чрез речеви характеристики - височина, ритъм и сила на звука. Тонът на говорещия може да подсили или да противоречи на текстовото съдържание, осигурявайки важна информация за точно класифициране на мненията.
- *Визуалните източници* предлагат невербална информация от изражения на лицето, жестовете и езика на тялото, която подпомага анализирането на човешкото състояние и поведение.

Обучението на ефективни модели за мултимодален сентимент анализ изисква разнообразни и представителни набори от данни. Висококачествените данни са от съществено значение за точното и ефективно обучение на моделите, тъй като позволяват на моделите да обобщават добре нови данни, като гарантират тяхната приложимост в реални сценарии. Данни с ниско качество, които могат да съдържат неточности, шум или несъответствия, могат да доведат до ненадеждни и по-малко точни прогнозни модели, които се учат от тези недостатъци. Некачествените данни поставят предизвикателства в обработката на данните и в обучението, свързани с шум, липсващи данни или отклонения, въведени по време на събирането на данни или анотацията. Придобиването на голям обем висококачествени, анотирани мултимодални данни е основна пречка за обучение на стабилни модели за сентимент класификация, защото извличането и обработката на мултимодални данни са свързанни с различни ограничения, включително проблеми с неприкосновеността на личния живот, строги регулаторни рамки и значителните разходи и специализиран опит за последователна анотация на данните.

За целите на изследователската дейност различни набори от данни са извлечени, анотирани и предоставени за използване, като се превръщат във важни еталони за сравнение в областта на мултимодалния аспектно-базиран анализ на мнения [3]. В таблица 1 е направено сравнение на набори от данни, извлечени и използвани за мултимодален сентимент анализ.

**Таблица 3:** Сравнение на набори от данни за мултимодален сентимент анализ

| Набор от данни | Година | Обем | Модалности | Задача/Област на приложение |
|---|---|---|---|---|
| **MOUD** [4] | 2013 | 498 изказвания | Текст, аудио, видео | Полярност на мнения Отзиви за продукти, испански език |
| **Twitter-2015** [5] | 2015 | 5,338 туитове | Текст, изображение | Мнение за аспект (3 класа), Туитър дани, английски език |
| **Twitter-2017** [5] | 2017 | 5,972 туитове | Текст, изображение | Мнение за аспект (3 класа), Туитър данни, английски език |
| **CMU-MOSI** (Multimodal Opinion-Level Sentiment Intensity) [6] | 2016 | ~2199 сегмента на мнението | Текст, аудио, видео | Интензивност на настроенията YouTube отзиви |
| **CMU-MOSEI** (Multimodal Opinion Sentiment and Emotion Intensity) [7] | 2018 | ~23,453 изречения | Текст, аудио, видео | Мнение и емоции Видео реч |
| **MELD** (Multimodal EmotionLines Dataset) [8] | 2019 | ~13,000 изказвания | Текст, аудио, видео | Емоции и чувства Диалози (ТВ предаване "Приятели") |
| **Multi-ZOL** [9] | ~2019 | 5,288 отзиви | Текст, изображение | Мнения за аспект (1–10) (6 аспекта) Отзиви за телефони |
| **TumEmo** [10] | ~2020/ 2021 | ~190,000 примера | Текст, изображение | Класификация на емоциите, Tumblr данни (изображение и надпис) |
| **MASAD** (Multimodal Aspect-Based Sentiment Analysis dataset) [11] | 2021 | ~38,000 двойки изображение–текст | Текст и изображение | Мултимодален аспектно-базиран сентимент анализ с 57 категории на аспекти. Множество домейни |
| **CH-SIMS v2.0** [12] | 2022 | ~4,406 маркирани сегменти | Текст, аудио, видео | Мултимодален сентимент анализ, с допълнителни необозначени данни (~10,161) за невербалните знаци. Видеоклипове за подчертаване на акустичните/визуалните знаци. |
| **MuSe 2022** (Passau-SFCH, Hume-Reaction, Ulm-TSST) [13] | 2022 | Множество набори от данни | Текст, аудио, видео | Множество задачи: откриване на хумор, емоционални реакции, стрес, |

| | | | | |
|---|---|---|---|---|
| | | | | спонтанно поведение, емоции/чувства. |
| **MSED** (Multi-modal Sentiment & Emotion Desire dataset) [14] | ~2022 | 9,190 двойки текст-изображение от социалните медии | Текст и изображение | Идентификация на емоции, желания, мнения. Социални медии |
| **M-EMD/ MEMD-ABSA** [15] | 2023 | ~20 000 изречения от отзиви; ~30 000 анотирани с явни и имплицитни аспекти/мнения в 5 области. | Текст | Извличане на мнения за аспекти, разглежда явни и имплицитни аспекти. Сравняване на приноса на модалностите. |
| **Multimodal Dataset for Sentiment Analysis and Classification** [16] | 2024 | 600 видеоклипа от различни източници; етикетирани за 7 емоции. | Текст, видео, аудио | Класификация на емоциите, изразени чрез видео, английски език |
| **UniC** (Dataset for emotion analysis of videos with unimodal and multimodal labels [17] | 2025 | Данни за емоции както с едномодални, така и с мултимодални етикети. | Текст, аудио, видео | Моделиране на емоции и чуства, поддържа едномодални/мултимодални етикети. YouTube данни, сравнение между мултимодалният анализ и едномодалния анализ. |
| **EmotionTalk** [18] | 2025 | 19 250 изказвания (≈23,6 часа реч)) | Текст, аудио, видео | 7 категории емоции, китайски език, богата анотация. Полезно за диалог/емоция, сентимент, стил. |
| **EmoSign** [19] | 2025 | 200 видеоклипа (жестомимичен език) с етикети за сантименти и емоции; плюс анотации на емоционални знаци. | Текст, видео | Класификация на чувства и емоции, разбиране в контекста на жестомимичен език, информация за лицето и жестовете. |
| **M-ABSA (Multilingual ABSA)** [20] | 2025 | Голям набор от данни, обхващащ 7 домейна, 21 езика | Текст | Анализ на мнения на ниво аспект. Фокус върху извличането на аспектен термин, категория и полярност. Многоезичен анализ на мнения на ниво аспект. |

Новите набори от данни се отличават с богати, нюансирани анотации като адресират недостатъчно проучени аспекти на сентимент анализа: невербалните знаци; липсващи модалности на повече езици (вкл. жестомимични езици); как чувствата и емоциите взаимодействат с желанията или стила на речта. Използването на такива публично достъпни набори от данни за сентимент анализ позволяват на изследователите да подобряват методите на анализ и да сравняват модели за класификация.

Основните предизвикателствата пред системите за мултимодален сентимент анализ са свързани със събиране на данни от източници и с ефективно им интегриране, с цел да улавят сложните връзки, които определят човешкото изразяване. Настоящите изследвания в областта [21] [22] показват, че все още данните от текстовите източници заемат централно място в системите за мултимодален сентимент анализ, а аудио и визуалните данни се използват за разрешаване на неясноти или добавяне на емоционална дълбочина в определянето на мнението. Затова съвременатa насока на научните изследвания е да се предлагат архитектури, които използват данните от трите вида източници, като всеки източник допринася еднакво с уникална информация, която подобрява цялостното разбиране на емоционалното състояние или изразеното мнение.

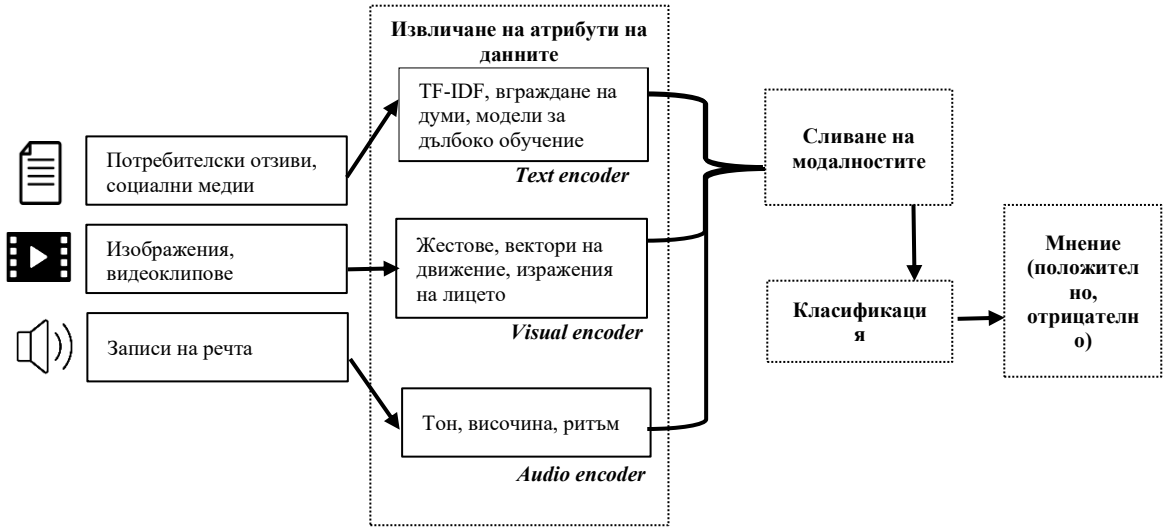
**Архитектурни подходи за мултимодален сентимент анализ**

Процесът на мултимодален сентимент анализ включва няколко взаимосвързани подзадачи, които допринасят за цялостното разбиране на настроенията/мненията в мултимодалното съдържание, позволявайки специализирана обработка и обучение.

- Първата подзадача е *извличане на аспекти от различните типове данни*. Основната цел е точно да се идентифицират и извлекат всички релевантни аспектни термини, независимо дали са изрично посочени или имплицитно загатнати от текстовите данни. Релевантните аспектни термини са конкретните обекти и техните характеристики, за които се изразява чувство или мнение.
- Следващата подзадача е *мултимодалната аспектно-ориентирана класификация* на мнението, която се определя полярността на мнението (напр. положителна, отрицателна, неутрална), свързана с всеки идентифициран аспектен термин, като използва информация, получена както от текстовите, така и от другите два източника. Важно е да се отбележи, че мнение към един аспект, може да се различава значително от цялостното мнение, изразено в публикация или документ.

При мултимодалния сентимент анализ, извличането на аспектни термини и класификацията на мненията към тях са интегрирани в единна рамка и обхваща извличане на двойки аспект-мнение, от които моделите научават и използват присъщите зависимости между тях. Този многозадачен подход води до стабилни и последователни модели и всеобхватна и точна система за сентимент анализ.

Архитектурата на система за мултимодален сентимент анализ интегрира данни от три източника (текст, аудио и изображения) и обхваща пет слоя – извличане на данни от източниците, обработка на данните чрез извличане на атрибути от данните, кодиране на данните, сливане на данните и класификация с цел определяне на чувството/мнението (фиг. 1).




**Фигура 1**: Архитектура на система за мултимодален сентимент анализ

Входните данни са сурови, необработени текстови данни, изображнения и гласови записи. Текстовите данни предоставят субективна информация като може да съдържат споменати аспекти на обекта. Изображенията (свързани снимки или продуктови изображения) или поредица от видео кадри, които съдържат визуални доказателства за обекта или аспекта на обекта (например снимка на камера или екран) или изражения на лицето, движения на главата и жестове на тялото. Гласови записи (например отзиви на клиенти, подкасти или разговори) трябва да включват емоционални характеристики (тон, интонация, темпо на речта).

В слоя *Извличане на атрибути на данните*, суровите данни се изчистват, подготвят (разпознаване на обекти в изображенията, извличане на аудио характеристики) и трансформират в смислени структури (вектори), които улавят значението на текста, визуалните елементи и емоциите в гласа. Добрата обработка и трансформация на данните осигурява ефективност и високо качество на моделите за класификация на чувствата/мненията. В процеса на извличане на атрибути на данните се прилагат енкодери, които са специализирани инструменти, предназначени да превеждат разбираеми за човека данни (език, звук, картина) в машинно разбираеми числови представяния, които се използват от приложения с изкуствен интелект.

- Подходите за подготовка и трансформация на *текстовите данни* могат да бъдат традиционни като Bag-of-Words, TF-IDF, използване на речници (VADER[1][23]) или контекстуализирани вграждане на думи от модели за дълбоко обучение като BERT, RoBERTa, които улавят семантично значение и контекст. Входното изречение се разделя на по-малки единици (токени). Всеки токен се преобразува в предварително обучен вектор с помощта на таблица за вграждане. Векторите са проектирани така, че думите със сходни значения да са близо една до друга във векторното пространство. Енкодер моделите обработват последователността от векторите чрез механизъм, наречен Внимание (Attention), за да се претегли важността на всяка дума спрямо всички останали. Енкодерът произвежда един вектор за цялата входна последователност.
- Извличането на атрибути от *видео и изображения* се извършва чрез видео енкодер, който взима поредица от видео кадри (изображения) и придружаващото ги аудио и създава представяне, което улавя както пространствената (в рамките на кадър), така и времевата (през кадрите) информация. Видео енкодерите често комбинират техники от обработка на изображения и текст. Извличането на пространствени характеристики като обекти и сцени от всеки отделен кадър, се извършва чрез конволюционна невронна мрежа (CNN). Извличане на времеви характеристики са важни за разбиране на движението и историята. Последователността от векторите на атрибутите от всеки кадър се подава в модел, който разбира последователностите - повтарящи се невронни мрежи (RNN) и трансформари (с

[1] VADER (Valence Aware Dictionary and sEntiment Reasoner)

механизъм на внимание). Видео трансформърът прилага самовнимание във всички кадри, което му позволява да разбира дългосрочни зависимости и взаимоотношения.

- В извличането на атрибути от *аудио записи* се прилага аудио енкодер, който преобразува необработени аудио сигнали в смислено представяне. Вълната често се преобразува в спектрограма (визуално представяне на спектъра от честоти във времето), след което се извличат характеристики, идентифицират се фонеми и други акустични елементи. Аудио сигналът е последователност, така че модели като RNN или трансформъри могат да го обработват директно, за да уловят времевия контекст, което е от решаващо значение за разбирането на речта и музиката.

*Мултимодално сливане*, следващият слой в архитектурата, е най-изследваният компонент в системите за мултимодален сентимент анализ. Сливането на модалностите има за цел да комбинира отделните представяния на модалностите [T, A, V] в съвместно представяне Z, което ефективно улавя както интрамодалните, така и кросмодалните взаимодействия [24]. Z е единен векторен изход, който обхваща емоционалната информация от всички участващи модалности (текст, аудио, изображение).

В методологичните подходи преобладават невронни архитектури, които могат да бъдат категоризирани в три поколения: (1) обучение за унимодално представяне с късно сливане, (2) обучение за съвместно представяне с кросмодални взаимодействия и (3) съвременни архитектури за сливане, използващи предварително обучени модели.

В първоначалните подходи в мултимодалното обучение доминира методологичният подход на късното сливане. Всяка модалност (напр. текст, изображение) се обработва от собствена, независима невронна мрежа (напр. CNN за изображения, RNN/LSTM за текст или аудио), които извличат унимодални характеристики. [25]. Сливането на извлечената информация - вектори на характеристиките [T; A; V] - се извършва в последния етап, обикновено чрез конкатениране на получените унимодални представяния, последвано от класификатор. Подходът позволява използването на добре разработени унимодални модели, но не успява да улови сложните корелации и взаимодействия между модалностите (напр. разпознаването на сарказъм), тъй като взаимодействието е минимално и се случва твърде късно в процеса.

Второто поколение архитектури се фокусира върху преодоляването на ограниченията на късното сливане чрез въвеждане на по-ранни и експлицитни механизми за взаимодействие между модалностите. Целта е да се научи съвместно представяне (joint representation) на данните като се използват трансформър-базирани модели (като BERT) и механизми за кръстосано внимание (Cross-Attention). Механизъм на кръстосаното внимание позволява на една модалност (напр. текст) да се обвърже с най-релевантните характеристики от друга модалност (напр. изображение) и обратно, създавайки силни кросмодални връзки между тях [26]. Например, думата "щастлив" в текста може да се обвърже с усмихнато лице във визуалния поток и развълнуван тон в аудио потока. При този модулен подход всяка модалност има свой енкодер, но между тях се вмъкват слоеве за експлицитно сливане и взаимодействие [27].

Най-съвременните мултимодални архитектури се характеризират с мащабно предварително обучение върху огромни масиви от мултимодални данни и стремеж към унифицирано представяне и целево сливане. Използват се модели, базирани на трансформър архитектури, обучени по методи на самоконтролирано обучение върху голям обем двойки данни (напр. изображение-текст). Тези архитектури често постигат ранно сливане като превръщат различните модалности в единен поток от токени още на входа на модела, позволявайки на един трансформър енкодер да обработва всички модалности съвместно. Мултимодални големи езикови модели използват предварително обучени големи езикови модели (LLMs) и енкодери за изображения/аудио, свързани чрез интерфейс за модалности. Те имат способността да приемат, да разсъждават и да извеждат мултимодална информация, което ги прави подходящи за по-сложен емоционален анализ, включващ разсъждение. Най-новите архитектурни подходи се насочват към пълна интеграция и унификация, използвайки предимствата на предварително обучените модели и сложните механизми за внимание, за да постигнат разбиране, което е по-близко до човешкото възприятие.

След мултимодалното сливане, класификацията на мнение/чувство се извършва, като обединеното представяне Z на всички модалности се подава към слой за класификация. При късното сливане векторът е агрегирана форма на индивидуалните изходи от класификаторите на всяка модалност. При ранното сливане векторът е получен директно от последния скрит слой на общия трансформър модел или след слоя за кросмодално внимание. Обединеният векторен изход Z се подава на един или повече напълно свързани слоя, които изпълняват задачата по класификация. По време на обучението, класификационният резултат се сравнява с реалния клас, използвайки функция на загуба (Loss Function), която ръководи актуализирането на теглата на цялата мултимодална архитектура. Чрез минимизиране на тази загуба, моделът се научава да създава обединени представяния, които най-добре разграничават класовете, като същевременно използват допълваща информация от всички модалности.

**Предизвикателства пред мултимодалния сентимент анализ**

Въпреки значителния прогрес в архитектурните и методологични подходи, извличането на мнения от различни типове източници среща предизвикателства, които произтичат от присъщата сложност на човешкия език, разнообразната природа на мултимодалните данни и субективната същност на чувствата.

Едно от основните предизвикателства е семантичната сложност. Изреченията често съдържат множество аспекти, всеки от които потенциално се отнася до различни обекти или понятия, както в текста, така и в придружаващо изображение. Освен това тези различни аспекти и съответните им области на изображението могат да носят различни чувства. Точното улавяне на нюансирани чувства в една двойка текст-изображение изисква моделът да разграничава и определя чувствата на детайлно ниво, откривайки сложните семантичнин връзки.

Разбирането на чувствата и контекстуалните вариации се моделира трудно поради двусмисленост на езика, при която една дума може да притежава множество значения в зависимост от контекста си. Двусмислеността води до напълно различни интерпретации на чувствата. Културните и контекстуалните различия също значително влияят върху начина, по който настроенията се изразяват и интерпретират, което изисква алгоритмите да бъдат обучени на различни набори от данни, за да се преодолеят тези вариации.

Друг проблем, който често съпътства данните е свързан с несъответствие между текста и изображението, придружаващо текста. Например изображението не съдържа обектите/аспектите, посочени в текста. Такова несъответствие може сериозно да влоши качеството на модела, като доведе до неточно крос-модално подравняване. Затова съвременните модели са насочвани към идентифициране и директно справяне със специфични, често срещани несъответствия в мултимодалните данни от реалния свят.

Променливостта на данните, субективността и пристрастията създават значителни препятствия пред обучението на стабилни модели, които изискват разнообразни и представителни набори от данни. Процесът на анотиране на данни включва лична преценка, която може да доведе до субективност и пристрастия. В тази връзка е необходимо установяване на ясни правила за анотация и стратегии за управление на пристрастията с цел да се гарантира създаването на надеждни анотирани набори от данни.

**Заключение**

Въпреки значителния напредък, постигнат чрез дълбокото обучение и трансформър-базираните архитектури, мултимодалният сентимент анализ все още се сблъсква с важни предизвикателства, свързани с интеграция на данните от различните източници, синхронизацията, интерпретируемостта и недостига на анотирани данни. Съществуващите модели често показват ограничена способност за обобщаване между различни домейни, езици и културни контексти поради пристрастия и малки набори от данни. Преодоляването на тези проблеми изисква разработването на прозрачни и адаптивни мултимодални архитектури,

способни да работят с непълни или данни със шумове и да предоставят интерпретируеми резултати.

Бъдещите изследвания е необходимо да бъдат фокусирани върху разширяване на многоезичните ресурси, нискоресурсното обучение и етичната употреба на данни, за да се постигнат по-надеждни, обясними и контекстуално чувствителни системи за анализ на мнения.

## Литература / References


1. Ankita Gandhi, Kinjal Adhvaryu, Soujanya Poria, Erik Cambria, Amir Hussain (2023). Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions, Information Fusion, Volume 91, 2023, Pages 424-444,ISSN 1566-2535, https://doi.org/10.1016/j.inffus.2022.09.025
2. Liu, B. (2012). Sentiment Analysis and Opinion Mining. United States: Morgan & Claypool.
3. Yang, H., Zhao, Y., Wu, Y., Wang, S., Zheng, T., Zhang, H., ... & Qin, B. (2024). Large language models meet text-centric multimodal sentiment analysis: A survey. *arXiv preprint arXiv:2406.08068*.
4. Pérez-Rosas, V., Mihalcea, R., & Narayan, S. (2013). *Utterance-level Multimodal Sentiment Analysis* (MOUD dataset). (paper & dataset page), https://aclanthology.org/P13-1096.pdf
5. Yu, J., & Jiang, J. (2019). Adapting BERT for Target-Oriented Multimodal Sentiment Classification. *International Joint Conference on Artificial Intelligence*, pp. 5408-5414
6. Zadeh, A., Zellers, R., Pincus, E., & Morency, L. P. (2016). Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*.
7. AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, Melbourne, Australia. Association for Computational Linguistics.
8. Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.
9. Xu, N., Mao, W., & Chen, G. (2019). Multi-Interactive Memory Network for Aspect Based Multimodal Sentiment Analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 371-378. https://doi.org/10.1609/aaai.v33i01.3301371
10. Xiaocui, Yang & Feng, Shi & Wang, Daling & Zhang, Yifei. (2020). Image-Text Multimodal Emotion Classification via Multi-View Attentional Network. IEEE Transactions on Multimedia. PP. 1-1. 10.1109/TMM.2020.3035277.
11. Jie Zhou, Jiabao Zhao, Jimmy Xiangji Huang, Qinmin Vivian Hu, Liang He, MASAD: A large-scale dataset for multimodal aspect-based sentiment analysis, Neurocomputing, Volume 455, 2021, Pages 47-58, ISSN 0925-2312, https://doi.org/10.1016/j.neucom.2021.05.040.
12. Liu, Y., Yuan, Z., Mao, H., Liang, Z., Yang, W., Qiu, Y., ... & Gao, K. (2022, November). Make acoustic and visual cues matter: Ch-sims v2. 0 dataset and av-mixup consistent module. In *Proceedings of the 2022 international conference on multimodal interaction* (pp. 247-258).

13. Christ, L., Amiriparian, S., Baird, A., Tzirakis, P., Kathan, A., Müller, N., ... & Schuller, B. W. (2022, October). The muse 2022 multimodal sentiment analysis challenge: humor, emotional reactions, and stress. In *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge* (pp. 5-14).
14. Ao Jia, Yu He, Yazhou Zhang, Sagar Uprety, Dawei Song, and Christina Lioma. 2022. Beyond Emotion: A Multi-Modal Dataset for Human Desire Understanding. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1512–1522, Seattle, United States. Association for Computational Linguistics.
15. Cai, H., Song, N., Wang, Z., Xie, Q., Zhao, Q., Li, K., ... & Xia, R. (2025). Memd-absa: A multi-element multi-domain dataset for aspect-based sentiment analysis. *Language Resources and Evaluation*, 1-29.
16. MALI, SHANKAR (2024), “Multimodal Dataset for Sentiment Analysis and Classification”, Mendeley Data, V2, doi: 10.17632/b7z68nykxt.2
17. Du, Q., Labat, S., Demeester, T. *et al.* UniC: a dataset for emotion analysis of videos with multimodal and unimodal labels. (2025) *Lang Resources & Evaluation* **59**, 2857–2892. https://doi.org/10.1007/s10579-025-09837-0
18. Sun, H., Wang, X., Zhao, J., Zhao, S., Zhou, J., Wang, H., ... & Qin, Y. (2025). EmotionTalk: An Interactive Chinese Multimodal Emotion Dataset With Rich Annotations. *arXiv preprint arXiv:2505.23018*.
19. Chua, P., Fang, C. M., Ohkawa, T., Kushalnagar, R., Nanayakkara, S., & Maes, P. (2025). EmoSign: A Multimodal Dataset for Understanding Emotions in American Sign Language. *arXiv preprint arXiv:2505.17090*.
20. Wu, C., Ma, B., Liu, Y., Zhang, Z., Deng, N., Li, Y., ... & Plank, B. (2025). M-ABSA: A Multilingual Dataset for Aspect-Based Sentiment Analysis. *arXiv preprint arXiv:2502.11824*.
21. Zhang,H. (2024). A comprehensive survey on multimodal sentiment analysis: Techniques, models, and applications. Advances in Engineering Innovation,12,47-52. https://www.ewadirect.com/journal/aei/article/view/16349#
22. Tianyu Zhao, Ling-ang Meng, Dawei Song (2024). Multimodal Aspect-Based Sentiment Analysis: A survey of tasks, methods, challenges and future directions, Information Fusion, Volume 112, 2024, 102552, ISSN 1566-2535, https://doi.org/10.1016/j.inffus.2024.102552
23. Barik, K., Misra, S. Analysis of customer reviews with an improved VADER lexicon classifier. J Big Data 11, 10 (2024). https://doi.org/10.1186/s40537-023-00861-x
24. Lai, S., Hu, X., Xu, H., Ren, Z., & Liu, Z. (2023). Multimodal sentiment analysis: A survey. *Displays*, *80*, 102563. https://arxiv.org/abs/2305.07611
25. Jing Gao, Peng Li, Zhikui Chen, Jianing Zhang; A Survey on Deep Learning for Multimodal Data Fusion. *Neural Comput* 2020; 32 (5): 829–864. doi: https://doi.org/10.1162/neco_a_01273
26. Mowbray, T. (2025). A Survey of Deep Learning Architectures in Modern Machine Learning Systems: From CNNs to Transformers. https://doi.org/10.5281/zenodo.17035434
27. Wadekar, S. N., Chaurasia, A., Chadha, A., & Culurciello, E. (2024). The evolution of multimodal model architectures. *arXiv preprint arXiv:2405.17927*.