УДК 004.89:81'322

MACHINE LEARNING AND DEEP LEARNING FOR EMOTION TONE ANALYSIS IN DIALECTAL ARABIC TEXTS

Zakzouk Sh.[1] (post-graduate student)**, Karaja N.**[2] (post-graduate student),
Scientific Supervisor – PhD, Associate Professor Avksentyeva E. Yu.[1]
[1] ITMO University
[2] ETU "LETI"
shaimaazakzouk@yandex.ru

The work was presented at KMU Congress.

Abstract
Emotion tone analysis in dialectal Arabic aims to detect fine-grained affect in informal texts, which is challenging due to Arabic's rich morphology, non-standard online orthography, Arabizi writing, and diverse regional dialects [1,2,3]. Studies show much work still targets binary sentiment or MSA, while emotion categories and dialectal varieties remain under-resourced [2,4], with Iraqi, Algerian, Moroccan, Egyptian, and Gulf dialects limited by scarce labeled corpora, mixed scripts, and inconsistent annotation [1,3]. Recent research explored ML and DL methods [3,4]; traditional ML provides strong baselines but struggles with nuanced emotions [2], whereas hybrid and transformer-based DL models achieve higher performance with proper preprocessing and embeddings [1,2,4]. The core problem is designing robust, data-efficient ML/DL pipelines generalizing across dialects and emotion categories with limited high-quality labeled data.
Keywords
emotion analysis, dialectal Arabic, deep learning, machine learning.

МАШИННОЕ ОБУЧЕНИЕ И ГЛУБОКОЕ ОБУЧЕНИЕ ДЛЯ АНАЛИЗА ЭМОЦИОНАЛЬНОГО ТОНА В ДИАЛЕКТНЫХ АРАБСКИХ ТЕКСТАХ

Закзук Ш.[1] (post-graduate student)**, Каража Н.**[2] (post-graduate student)
Научный руководитель – канд. пед. наук, доцент Авксентьева Е. Ю.[1]
[1]Университет ИТМО
[2] СПбГЭТУ «ЛЭТИ»
shaimaazakzouk@yandex.ru

Работа была представлена на Конгрессе КМУ.

Аннотация
Анализ эмоционального тона в диалектном арабском языке направлен на выявление тонких аффектов в неформальных текстах, что представляет собой сложную задачу из-за богатой морфологии арабского языка, нестандартной онлайн-орфографии, арабизского письма и разнообразия региональных диалектов [1,2,3]. Исследования показывают, что большая часть работы по-прежнему ориентирована на бинарное выражение настроения или стандартный арабский язык, в то время как категории эмоций и диалектные варианты остаются недостаточно изученными [2,4], при этом иракский, алжирский, марокканский, египетский и диалекты стран Персидского залива ограничены скудными размеченными корпусами, смешанными письменностями и непоследовательной аннотацией [1,3]. Недавние исследования посвящены методам машинного обучения и глубокого обучения [3,4]; традиционное машинное обучение обеспечивает надежные базовые показатели, но испытывает трудности с нюансированными эмоциями [2], в то время как гибридные и трансформерные модели глубокого обучения достигают более высокой производительности при надлежащей предварительной обработке и эмбеддингах [1,2,4]. Основная проблема заключается в разработке надежных, эффективных с точки зрения данных конвейеров машинного обучения/глубокого обучения, обобщающих данные

по диалектам и категориям эмоций при ограниченном количестве высококачественных размеченных данных.

Ключевые слова
анализ эмоций, диалектный арабский язык, глубокое обучение, машинное обучение,

A modern solution for emotion tone analysis in dialectal Arabic texts should integrate four components: dialect-aware preprocessing, scalable annotation, powerful representation learning, and ensemble modeling.

1. Dialect-aware preprocessing and resource construction

Studies on Iraqi, Algerian and multi-dialect Arabic confirm that preprocessing strongly impacts performance. Normalization, stop-word removal, and handling hashtags, emojis and URLs improve emotion and sentiment models [1,2,3]. For Algerian dialect, Arabizi transliteration and translation into MSA further enhance deep learning performance [3].

A cost-effective pipeline should therefore include unified character normalization, lightweight dialect identification to route texts to dialect-specific models [2], and emoji/hashtag distant supervision to inject weak emotion labels [4]. Semi-automatic corpus construction has proven effective for low-resource dialects; automatically collected data refined by manual review enabled semi-supervised deep learning to reach F1 up to 95% (intrinsic) and 89% (extrinsic) [3]. This supports building dialectal emotion corpora via distant supervision followed by targeted manual refinement.

2. Machine learning baselines and ensembles

For multi-dialect Arabic sentiment, ensembles of SVM, logistic regression, Naive Bayes, decision trees and random forests outperform individual models, showing that diverse learners capture complementary emotional cues [5].

These methods remain essential as strong baselines, competitive solutions for small datasets where SVM and random forest can surpass DL [5], and meta-learners in stacked systems combining deep representations with lightweight classifiers [5]. Consequently, an optimal pipeline integrates deep encoders with ML ensemble layers for final emotion classification.

3. Deep learning for dialectal emotion tone

Hybrid deep models show strong results for dialectal emotion analysis. A CNN - GRU model trained on Iraqi Facebook data achieved more than 91 - 92% accuracy and outperformed earlier methods [1]. For Algerian Arabic and Arabizi, semi supervised CNN and LSTM models with Word2Vec and fastText reached F1 up to 89%, confirming the efficiency of semi supervised deep learning for low resource dialects [3].

Transformer ensembles represent the current state of the art. An ensemble of CAMeLBERT, XLM RoBERTa and MARBERT achieved about 90% accuracy and surpassed standalone CNN BLSTM and classical ML models [4], while a voting ensemble of BERT models reached about 84% accuracy for Arabic tweet emotion classification compared with 76% for single models [2].

An effective pipeline should therefore use a dialect pretrained transformer as the main encoder [2,4], optionally add a light CNN BiGRU layer, train in a multilabel setup, apply semi supervised learning on unlabeled data [3], and finalize predictions with a small soft voting transformer ensemble [2].

4. Proposed research directions

Promising directions include multi task transformers for dialect identification, sentiment and emotion [2], distantly supervised emotion lexicons refined with active learning [4,5], and extending semi supervised dialect adaptation to other dialects [3]. These approaches reduce reliance on large fully annotated corpora by leveraging pretrained models and weak supervision.

The research shows that ML ensembles and deep learning, especially hybrid CNN/RNN and transformer-based models, are the most effective tools for emotion tone analysis in dialectal Arabic [1,2,3,4]. CNN-GRU networks achieve over 91% accuracy on Iraqi dialect tasks [1], while semi-supervised CNN/LSTM reach F1 up to 89% for Algerian dialect in Arabic and Arabizi [3]. Transformer ensembles outperform standalone DL and ML on multi-dialect sentiment [4], and BERT-based ensembles improve Arabic tweet emotion classification [2]. ML ensembles remain competitive for smaller or specialized corpora [5]. These systems enable real-time monitoring of public emotions, customer feedback analysis, and early detection of online emotional trends across dialects.

Literature

1. Alfarhany A.A.-R., Abdullah N.A.Z. Iraqi Sentiment and Emotion Analysis Using Deep Learning // Journal of Engineering. 2023. Vol. 29, No. 9. P. 150–165. DOI: 10.31026/j.eng.2023.09.11.
2. Hamed D., Tawfik B., Makhlouf M. Using Machine Learning to Analyze Emotions in Arabic and Dialectical Texts // Scientific Journal of King Faisal University: Basic and Applied Sciences. 2024. Vol. 25, no. 2. DOI: 10.37575/b/sci/240014.
3. Guellil I., et al. A Semi-supervised Approach for Sentiment Analysis of Arab(ic+izi) Messages: Application to the Algerian Dialect // SN Computer Science. 2021. Vol. 2, art. 118. DOI: 10.1007/s42979-021-00510-1.
4. Mansour O., Aboelela E., Talaat R., Bustami M. Transformer-based ensemble model for dialectal Arabic sentiment classification // PeerJ Computer Science. 2025. Vol. 11. Art. e2644. DOI: 10.7717/peerj-cs.2644.
5. Doughan Z., Itani S., Itani S. ArabSis: Arabic Corpus Sentiment Analysis // IEEE Access. 2025. Pp. 81083–81095. DOI: 10.1109/ACCESS.2025.3567755.