| IEEE 2025-26 PROJECT LIST | |
|---|---|
| **Natural Language Processing** | |
| CODE | TITLE AND ABSTRACT |
| 26-ANSP-NLP-001 | **A Detailed Comparative Analysis of Automatic Neural Metrics for Machine Translation:BLEURT & BERTScore**<br>**BLEURT** is a recently introduced metric that employs BERT, a potent pre-trained language model to assess how well candidate translations compare to a reference translation in the context of machine translation outputs. While traditional metrics like BLEU rely on lexical similarities, BLEURT leverages BERT's semantic and syntactic capabilities to provide more robust evaluation through complex text representations. However, studies have shown that BERT, despite its impressive performance in natural language processing tasks can sometimes deviate from human judgment, particularly in specific syntactic and semantic scenarios. Through systematic experimental analysis at the word level, including categorization of errors such as lexical mismatches, untranslated terms, and structural inconsistencies, we investigate how BLEURT handles various translation challenges. Our study addresses three central questions: What are the strengths and weaknesses of BLEURT, how do they align with BERT's known limitations, and how does it compare with the similar automatic neural metric for machine translation, BERTSCORE? Using manually annotated datasets that emphasize different error types and linguistic phenomena, we find that BLEURT excels at identifying nuanced differences between sentences with high overlap, an area where BERTSCORE shows limitations. Our systematic experiments, provide insights for their effective application in machine translation evaluation. |
| 26-ANSP-NLP-002 | **Accelerating Multilingual Cryptocurrency Forensics: An NLP-Driven Approach for Efficient Mnemonic Identification**<br>The increasing use of cryptocurrencies in criminal activities presents significant challenges to society and the judicial system, particularly in tracking and seizing illicit digital assets. Among all relevant digital evidence, mnemonic phrases, which are critical for accessing cryptocurrency wallets, are crucial digital evidence for confiscating criminal proceeds and conducting investigations. However, traditional digital forensics tools, such as the Mnemonic Library Matching Method, lack flexibility and efficiency when handling cryptocurrency-related data. This study introduces an innovative Natural Language Processing (NLP) and deep learning approach for rapid mnemonic identification across 11 languages, including English, Spanish, and Japanese. We trained and compared four NLP deep learning models: RNN, LSTM, BiLSTM, and TextCNN, on a large-scale, real-world dataset. Our analysis reveals that the Text Convolutional Neural Network (TextCNN) model exhibits superior performance, achieving a 99.9993% accuracy rate, nearly matching the 100% accuracy of the Mnemonic Library Matching Method. Crucially, our TextCNN-driven approach processes data 40.47 times faster than the traditional method, significantly enhancing efficiency in time-sensitive forensic environments. This NLP-driven method not only maintains high accuracy while dramatically reducing processing time but also offers greater adaptability for diverse forensic needs compared to traditional techniques. By enabling more effective tracking and seizure of criminal assets, this approach aims to address the broader societal and judicial challenges posed by cryptocurrency-related criminal activities. Our research showcases the potential of NLP and deep learning in digital forensics, providing law |

| | |
|---|---|
| | enforcement with advanced tools for investigating cryptocurrency-related crimes and curbing the misuse of cryptocurrencies in illicit activities. |
| 26-ANSP-NLP-003 | **Advancing Sentiment Analysis for Low-Resource Languages Using Fine-Tuned LLMs: A Case Study of Customer Reviews in Turkish Language**<br>This study investigates the application of advanced fine-tuned Large Language Models (LLMs) for Turkish Sentiment Analysis (SA), focusing on e-commerce product reviews. Our research utilizes four open-source Turkish SA datasets: Turkish Sentiment Analysis version 1 (TRSAv1), Vitamins and Supplements Customer Review (VSCR), Turkish Sentiment Analysis Dataset (TSAD), and TR Customer Review (TRCR). While these datasets were initially labeled based on star ratings, we implemented a comprehensive relabeling process using state-of-the-art LLMs to enhance data quality. To ensure reliable annotations, we first conducted a comparative analysis of different LLMs using the Cohen's Kappa agreement metric, which led to the selection of ChatGPT-4o-mini as the best-performing model for dataset annotation. Our methodology then focuses on evaluating the SA capabilities of leading instruction-tuned LLMs through a comparative analysis of zero-shot models and Low-Rank Adaptation (LoRA) fine-tuned LlaMA-3.2-1B-IT and Gemma-2-2B-IT models. Evaluations were conducted on both in-domain and out- domain test sets derived from the original star-ratings-based labels and the newly generated GPT labels. The results demonstrate that our fine-tuned models outperformed leading commercial LLMs by 6% in both in-domain and out-domain evaluations. Notably, models fine-tuned on GPT-generated labels achieved superior performance, with in-domain and out-domain F1-scores reaching 0.912 and 0.9184, respectively. These findings underscore the transformative potential of combining LLM relabeling with LoRA fine-tuning for optimizing SA, demonstrating robust performance across diverse datasets and domains. |
| 26-ANSP-NLP-004 | **An Adapted Few-Shot Prompting Technique Using ChatGPT to Advance Low-Resource Languages Understanding**<br>The lack of annotated data in low-resource languages presents a significant challenge in natural language processing, particularly for language understanding tasks such as intent detection and slot filling. To address this, we propose a novel approach that first employs an effective cross-lingual transfer model to generate labeled data for the target language, overcoming the scarcity of labeled data in low-resource settings. The main contribution of our work lies in the second step, where we introduce an adapted few-shot prompting technique to guide ChatGPT as a large language model (LLM). In this step, a subset of the machine-generated examples is selected based on the domain of the input, ensuring that the LLM is provided with more tailored and domain-specific examples. This two-step process leads to enhanced performance in handling low-resource languages. We conduct extensive experiments on Spanish, Thai, and Persian using the Facebook-multilingual and Persian-ATIS datasets. Experimental results demonstrate that our method outperforms existing techniques for non-Latin languages, such as Thai and Persian, and matches state-of- the-art performance for Latin-based languages, such as Spanish. |
| 26-ANSP-NLP-005 | **Automated Processing of Speech Recordings for Dietary Assessment: Evaluation in the LLMIC Context**<br>Novel methods of using smartphones to collect speech records of dietary intake facilitate self-reporting of intake in free-living conditions, and may further reduce biases affecting reliable capture such as complexity and reactivity. However, the processing of speech records for dietary assessment is a time-consuming task for analysts and thus restricted by research infrastructure. Low- and Low-Middle Income Countries have been |

| | |
|---|---|
| | disproportionately affected by barriers to dietary assessment (e.g., due to varying literacy or limited research infrastructure). As such, speech records could be a promising avenue to address the research gap in Low and Low-Middle Income regions. While recent advances in speech recognition and natural language processing technology have facilitated automation of this process, no existing studies have evaluated the effectiveness of this technology when applied to an LLMIC context. To this end we adapt the methods identified in a review of the literature to this new context and evaluate their performance on a data set of Khmer speech recordings describing dietary data captured in a free-living context in Cambodia. |
| 26-ANSP-NLP-006 | **Benchmarking Open-Source Large Language Models for Sentiment and Emotion Classification in Indonesian Tweets**<br>We benchmark 22 open-source large language models (LLMs) against ChatGPT-4 and human annotators on two NLP tasks—sentiment analysis and emotion classification—for Indonesian tweets. This study contributes to NLP in a relatively low-resource language (Bahasa Indonesia) by evaluating zero-shot classification performance on a labeled tweet corpus. The dataset includes sentiment labels (Positive, Negative, Neutral) and emotion labels (Love, Happiness, Sadness, Anger, Fear). We compare model predictions to human annotations and report precision, recall, and F1-score, along with inference time analysis. ChatGPT-4 achieves the highest macro F1-score (0.84) on both tasks, slightly outperforming human annotators. The best-performing open-source models—such as LLaMA3.1_70B and Gemma2_27B—achieve over 90% of ChatGPT-4's performance, while smaller models lag behind. Notably, some mid-sized models (e.g., Phi-4 at 14B parameters) perform comparably to much larger models on select categories. However, certain classes—particularly Neutral sentiment and Fear emotion—remain challenging, with lower agreement even among human annotators. Inference time varies significantly: optimized models complete predictions in under an hour, while some large models require several days. Our findings show that state- of-the-art open models can approach closed-source LLMs like ChatGPT-4 on Indonesian classification tasks, though efficiency and consistency in edge cases remain open challenges. Future work should explore fine-tuning multilingual LLMs on Indonesian data and practical deployment strategies in real-world applications. |
| 26-ANSP-NLP-007 | **Beyond the Determiner "Al-": Expanding the Determiner Class in Arabic, and Elimination of Lexical Ambiguities by Grammars**<br>Arabic nouns can be marked for definiteness or indefiniteness. The definite article is the prefix "Al-," which confines the determiner class to a single element "Al-." This topic is generally discussed under noun inflections, such as Gender, Number, Definiteness, and Case (GNDC), in grammar textbooks. The primary aim of this paper is to expand the Arabic determiner class (DET) by incorporating additional lexical items and providing a detailed description of their syntactic context within noun phrases (NPs). In traditional grammar, these lexical items are typically classified as noun adjectives and, at best, are referred to as noun specifiers since they modify the head noun. However, these "nouns" exhibit limited inflection compared to regular adjectives and occupy a specific position within the NP sequence. Additionally, the modified head nouns are constrained in their inflectional attributes, Number, and Definiteness. Our approach is qualitative and guided by morpho-syntactic considerations. We conduct an in-depth analysis of the grammatical features of ten lexical items, focusing particularly on the dependencies between the determiner and the following noun. This analysis also addresses some semantic properties. By focusing on context-sensitive grammatical rules, the study shows how these methods can enhance precision in parsing and reduce ambiguity in |

| | |
|---|---|
| | NLP tasks, highlighting the potential for developing more refined grammar for Arabic. This work is a prototype for comprehensive studies in linguistics and NLP. |
| 26-ANSP-NLP-008 | **Code Clone Detection Techniques Based on Large Language Models**<br>Code duplication, commonly known as code cloning, is a persistent challenge in software development. While reusing code fragments boosts productivity, excessive cloning poses challenges to maintenance and elevates the risk of bugs. Therefore, integrating code clone detection into the development process is crucial. The extensive code-related knowledge inherent in Large Language Models (LLMs) renders them high-potential candidates for addressing diverse software engineering challenges. However, the effectiveness of LLMs in the specific task of code clone detection requires precise evaluation. This paper proposes an innovative methodology leveraging few-shot instruction-tuned GPT-3.5 Turbo and GPT-4 to detect code clones across all types, focusing on complex clones (Type-3 and Type-4). Unlike conventional approaches confined to specific language pairs or tasks, our method employs versatile language models, showcases generalization strengths for semantic understanding, and leverages instruction tuning with few-shot inference for task-specific adaptability in code clone detection. A conversational dataset was crafted from BigCloneBench for instruction tuning, enhancing task alignment and performance. This study evaluates the proficiency of LLMs in identifying code clones, analyzing the impact of instruction tuning, and assessing the efficiency across various clone types. Experimental results demonstrate these models achieving competitive performance against existing tools for overall and complex clone detection. Integration into an Integrated Development Environment (IDE) enables real-time detection and automated refactoring, bridging the gap between theoretical advancements and practical usability. This work highlights the potential of generalized LLMs setting a new standard in a field traditionally dominated by specialized tools and demonstrates their adaptability for complex challenges in code analysis and maintainability. |
| 26-ANSP-NLP-009 | **Comparative Analysis of Traditional and Modern NLP Techniques on the CoLA Dataset:From POS Tagging to Large Language Models**<br>The task of classifying linguistic acceptability, exemplified by the CoLA (Corpus of Linguistic Acceptability) dataset, poses unique challenges for natural language processing (NLP) models. These chal- lenges include distinguishing between subtle grammatical errors, understanding complex syntactic structures, and detecting semantic inconsistencies, all of which make the task difficult even for human annotators. In this article, we compare a range of techniques, from traditional methods such as Part-of-Speech (POS) tagging and feature extraction methods like CountVectorizer with Term Frequency-Inverse Document Frequency (TF-IDF) and N-grams, to modern embeddings such as FastText and Embeddings from Language Models (ELMo), as well as deep learning architectures like transformers and Large Language Models (LLMs). Our experiments show a clear improvement in performance as models evolve from traditional to more advanced approaches. Notably, state-of-the-art (SOTA) results were obtained by fine-tuning GPT-4o with extensive hyperparameter tuning, including experimenting with various epochs and batch sizes. This comparative analysis provides valuable insights into the relative strengths of each technique for identifying morphological, syntactic, and semantic violations, highlighting the effectiveness of LLMs in these tasks. |

| | |
|---|---|
| 26-ANSP-NLP-010 | **Disentangling Reasoning Factors for Natural Language Inference**<br>Natural Language Inference (NLI) seeks to deduce the relations of two texts: a premise and a hypothesis. These two texts may share similar or different basic contexts, while three distinct reasoning factors emerge in the inference from premise to hypothesis: entailment, neutrality, and contradiction. However, the entanglement of the reasoning factor with the basic context in the learned representation space often complicates the task of NLI models, hindering accurate classification and determination based on the reasoning factors. In this study, drawing inspiration from the successful application of disentangled variational autoencoders in other areas, we separate and extract the reasoning factor from the basic context of NLI data through latent variational inference. Meanwhile, we employ mutual information estimation when optimizing Variational AutoEncoders (VAE)-disentangled reasoning factors further. Leveraging disentanglement optimization in NLI, our proposed a Directed NLI (DNLI) model demonstrates excellent performance compared to state-of-the-art baseline models in experiments on three widely used datasets: Stanford Natural Language Inference (SNLI), Multi-genre Natural Language Inference (MNLI), and Adversarial Natural Language Inference (ANLI). It particularly achieves the best average validation scores, showing significant improvements over the second-best models. Notably, our approach effectively addresses the interpretability challenges commonly encountered in NLI methods. |
| 26-ANSP-NLP-011 | **Distilling Wisdom: A Review on Optimizing Learning From Massive Language Models**<br>In the era of Large Language Models (LLMs), Knowledge Distillation (KD) enables the transfer of capabilities from proprietary LLMs to open-source models. This survey provides a detailed discussion of the basic principles, algorithms, and implementation methods of knowledge distillation. It explores KD's impact on LLMs, emphasizing its utility in model compression, performance enhancement, and self- improvement. Through the analysis of practical examples such as DistilBERT, TinyBERT, and MobileBERT, the paper demonstrates how knowledge distillation can markedly enhance the efficiency and applicability of large language models in real-world scenarios. The discussion encompasses the varied applications of KD across multiple domains, including industrial systems, embedded systems, Natural Language Processing (NLP), multi-modal processing, and vertical domains, such as medicine, law, science, finance, and materials science. This survey outlines current KD methodologies and future research directions, highlighting its role in advancing AI technologies and fostering innovation across different sectors. |
| 26-ANSP-NLP-012 | **"Diwan": Constructing the Largest Annotated Corpus for Arabic Poetry**<br>In recent years, Arabic natural language processing has achieved remarkable advancements, particularly with the advent of large language models that have enhanced the analysis of diverse Arabic texts, including literary and artistic works such as poetry. However, these models encounter specific challenges when dealing with Arabic poetry, which is characterized by complex metrical structures, varied themes, and unique linguistic intricacies. Currently available poetry corpora are often limited in scope and depth, and typically lack the comprehensive coverage required for sophisticated computational analysis tasks. To address this gap, this paper introduces "Diwan," the largest and most precise annotated Arabic poetry dataset/corpus, designed to support scientific research and facilitate the development of AI applications in this domain. "Diwan" comprises approximately 14 million verses of poetry across 16 major categories and includes detailed annotations related to poetic genres, prosodic structures, |

| | |
|---|---|
| | thematic content, linguistic patterns, and poet-specific metadata. This dataset/corpus has been meticulously curated using advanced data collection methods, rigorous normalization and annotation protocols, and expert oversight from specialists in Arabic prosody and poetry. By leveraging intelligent annotation algorithms, Diwan serves as a foundational resource and benchmark dataset for advancing research in fields such as automatic poetry generation, metrical analysis, thematic classification, and plagiarism detection. ‘‘Diwan’’ has been compared against 4 leading corpora for Arabic poetry; this comparison proves that ‘‘Diwan’’ outperforms all of them in terms of scope, and depth. Besides, ‘‘Diwan’’ is constructed and structured in a way enables AI-powered analysis and deep-learning-based analysis to work accurately. By providing an unprecedented foundation for computational exploration of Arabic poetry, Diwan opens up new avenues and possibilities for intelligent applications and promotes the digital analysis of this rich literary heritage. |

| | |
|---|---|
| 26-ANSP-NLP-013 | **Dynamic Prompt-Based Framework for Pragmatic Quality Improvement in Business Process Models**<br>Business Process Modeling plays a critical role in enhancing organizational efficiency, standardization, and transparency. While prior research has extensively addressed syntactic and semantic quality dimensions, pragmatic quality, relevant to the clarity, interpretability, and comprehensibility of process elements, remains relatively underexplored. This study presents a dynamic prompt-based framework that leverages Large Language Models to automatically generate meaningful and contextually appropriate labels for BPMN elements, including tasks, gateways, and XOR split outgoing edges. The framework integrates contextual information extracted from BPMN models to construct structured prompts, ensuring that the generated labels are semantically precise and compliant with BPMN labeling conventions. The generated labels are evaluated using the all-MiniLM-L6-v2 sentence transformer model to assess semantic coherence and mutual exclusivity, while spaCy’s syntactic analysis validates verb presence in task labels. Experimental results on the BPMAI dataset, encompassing 29,810 BPMN models, confirm the framework’s effectiveness in enhancing pragmatic quality and improving overall model interpretability. |
| 26-ANSP-NLP-014 | **Improving Transformer Performance for French Clinical Notes Classification Using Mixture of Experts on a Limited Dataset**<br>Transformer-based models have shown outstanding results in natural language processing but face challenges in applications like classifying small-scale clinical texts, especially with constrained computational resources. This study presents a customized Mixture of Expert (MoE) Transformer models for classifying small-scale French clinical texts at CHU Sainte-Justine Hospital. The MoE-Transformer addresses the dual challenges of effective training with limited data and low-resource computation suitable for in-house hospital use. Despite the success of biomedical pre-trained models such as CamemBERT-bio, DrBERT, and AliBERT, their high computational demands make them impractical for many clinical settings. Our MoE-Transformer model not only outperforms DistillBERT, CamemBERT, FlauBERT, and Transformer models on the same dataset but also achieves impressive results: an accuracy of 87%, precision of 87%, recall of 85%, and F1-score of 86%. While the MoE-Transformer does not surpass the performance of biomedical pre-trained BERT models, it can be trained at least 190 times faster, offering a viable alternative for settings with limited data and computational resources. Although the MoE-Transformer addresses challenges of generalization gaps and sharp minima, demonstrating some limitations for efficient and accurate clinical text classification, this model still represents a significant advancement in the field. It is |

| | |
|---|---|
| | particularly valuable for classifying small French clinical narratives within the privacy and constraints of hospital-based computational resources. |
| 26-ANSP-NLP-015 | **Integrating Natural Language Processing With Vision Transformer for Landscape Character Identification**<br>Landscape character identification (LCI) has tradi- tionally relied on manual methods to integrate and visually inter- pret multiple layers of natural and cultural data across regions. While effective, these methods are constrained by subjectivity in manual categorization and face challenges in scalability and consistency when applied to larger regions. In this article, we propose a novel deep learning-based LCI method that leverages natural language processing (NLP) to enable greater flexibility in identifying landscape types through natural language guidance. Focusing on the Western Sichuan Plains, our approach integrates nine geographic information system-derived landscape elements with natural language descriptions that specify desired styles of landscape characteristics. The model features three key components: a transformer-based module for extracting natural language features, a vision transformer (ViT) for spatial feature extraction, and a feature pyramid network for decision-making. Through multimodal information fusion across the feature extraction and decision-making stages, natural language inputs effectively guide the prioritization of landscape attributes and control the granularity of the generated landscape character maps. Visualization experiments demonstrate the model's capability to produce accurate and detailed landscape character maps, with a particular emphasis on identifying agricultural landscapes in the Western Sichuan Plains. This study validates the potential of integrating NLP into LCI, offering a significant advancement in precision and adaptability for landscape characterization. |
| 26-ANSP-NLP-016 | **Large Language Models in Human–Robot Collaboration With Cognitive Validation Against Context-Induced Hallucinations**<br>The recent leap in Large Language Models (LLMs) has paved the way for several research ideas. LLMs are employed not only for personal use but also in professional contexts to enhance human productivity at work. A significant area of research is human-robot collaboration (HRC), which focuses on developing methodologies for effective interaction between humans and AI-enabled machines. In this regard, exploitation of LLMs appears to be a practical approach. However, these models are susceptible to several limitations, including context-induced errors, the propagation of misleading information, and hallucinations. Such deficiencies impede the seamless application of LLMs in scenarios where a high degree of accuracy is essential. To address this issue, this study introduces a dual-agent system designed to validate the responses generated by LLMs. This novel system is integrated into a framework called ‘‘CogniVera’’, which facilitates collaborative tasks involving a collaborative robot (cobot) through vocal interactions. This initiative represents a significant advancement in HRC, enabling robots to communicate vocally with human operators during assembly tasks. To evaluate the feasibility of this approach, a focused case study will be conducted, concentrating on the human-robot collaborative task of box assembly utilizing vocal communication. The outcomes of this study are anticipated to yield valuable insights into the efficacy of the proposed dual-agent system in enhancing the reliability and performance of LLMs in practical applications. |

| | |
|---|---|
| 26-ANSP-NLP-017 | **LLM-Enhanced Human–Machine Interaction for Adaptive Decision-Making in Dynamic Manufacturing Process Environments**<br>Modern production systems generate vast amounts of process data that hold valuable insights for optimizing manufacturing processes. However, production personnel often face the challenge of interpreting this information, especially when dealing with unexpected anomalies or when insights beyond standard reports are required. This challenge arises both from the complex data structures in which the data is provided, and the lack of analytical expertise. This research introduces an approach that leverages Large Language Models (LLMs) to facilitate natural language queries and flexible data visualization, allowing production personnel to interact effortlessly with complex datasets. Tested on process data from an industrial extrusion process that has been enhanced using data augmentation techniques, the proposed concept demonstrates the capability to retrieve relevant data and present tailored visualizations based on simple user prompts. The results demonstrate that LLM-driven data exploration can support production personnel and help overcome the challenges described, which arise from the complex nature of manufacturing data and the specialized domain knowledge required. Future work will concentrate on improving accuracy, robustness, and further integration of domain-specific knowledge, aiming to provide a more reliable and accessible tool for various industrial environments. |
| 26-ANSP-NLP-018 | **Machine Reading Comprehension for the Tamil Language With Translated SQuAD**<br>Machine Reading Comprehension (MRC) is a challenging task in Natural Language Processing (NLP), crucial for automated customer support, enabling chatbots and virtual assistants to accurately understand and respond to queries. It also enhances question-answering systems, benefiting educational tools, search engines, and help desks. The introduction of attention-based transformer models has significantly boosted MRC performance, especially for well-resourced languages such as English. However, MRC for low-resourced languages (LRL) remains an ongoing research area. Although Large Language Models show exceptional NLP performance, they are less effective for LRL and are expensive to train and deploy. Consequently, simpler language models that are targeted at specific tasks remain viable for these languages. This research examines high-performing language models on the Tamil MRC task, detailing the preparation of a Tamil-translated and processed SQuAD dataset to establish a standard dataset for Tamil MRC. The study analyzes the performance of multilingual transformer models on the Tamil MRC task, including Multilingual DistilBERT, Multilingual BERT, XLM-RoBERTa, MuRIL, and RemBERT. Additionally, it explores improving these models' performance by fine-tuning them with English SQuAD, Tamil SQuAD, and a newly developed Tamil Short Story (TSS) dataset for MRC. Tamil's agglutinative nature, which expresses grammatical information through suffixation, results in a high degree of word inflexion. Given this characteristic, the BERT score was chosen as the evaluation metric for MRC performance. The analysis shows that the XLM-RoBERTa model outperformed the others for Tamil MRC, achieving a BERT score of 86.29% on the TSS MRC test set. This superior performance is attributed to the model's cross-lingual learning capability and the increased number of data records used for fine-tuning. The research underscores the necessity of language-specific fine-tuning of multilingual models to enhance NLP task performance, for low-resourced languages. |

| | |
|---|---|
| 26-ANSP-NLP-019 | **Natural Language Processing: Classification of Web Texts Combined with Deep Learning**<br>With the increasing number of web texts, the classification of web texts has become an important task. In this paper, the text word vector representation method is first analyzed, and bidirectional encoder representations from transformers (BERT) are selected to extract the word vector. The bidirectional gated recurrent unit (BiGRU), convolutional neural network (CNN), and attention mechanism are combined to obtain the context and local features of the text, respectively. Experiments were carried out using the THUCNews dataset. The results showed that in the comparison between word-to-vector (Word2vec), Glove, and BERT, the BERT obtained the best classification result. In the classification of different types of text, the average accuracy and F1 value of the BERT-BGCA method reached 0.9521 and 0.9436, respectively, which were superior to other deep learning methods such as TextCNN. The results suggest that the BERT-BGCA method is effective in classifying web texts and can be applied in practice. |
| 26-ANSP-NLP-020 | **Sentiment Analysis of Twitter Data Using NLP Models: A Comprehensive Review**<br>Social media platforms, particularly Twitter, have become vital sources for understanding public sentiment due to the rapid, large-scale generation of user opinions. Sentiment analysis of Twitter data has gained significant attention as a method for comprehending public attitudes, emotional responses, and trends which proves valuable in sectors such as marketing, politics, public health, and customer services. In this paper, we present a systematic review of research conducted on sentiment analysis using natural language processing (NLP) models, with a specific focus on Twitter data. We discuss various approaches and methodologies, including machine learning, deep learning, and hybrid models with their advantages, challenges, and performance metrics. The review identifies key NLP models commonly employed, such as transformer-based architectures like BERT, GPT, etc. Additionally, this study assesses the impact of pre-processing techniques, feature extraction methods, and sentiment lexicons on the effectiveness of sentiment analysis. The findings aim to provide researchers and practitioners with a comprehensive overview of current methodologies, insights into emerging trends, and guidance for future developments in the field of sentiment analysis on Twitter data. |
| 26-ANSP-NLP-021 | **Summarization, Prediction, and Analysis of Turkish Constitutional Court Decisions With Explainable Artificial Intelligence and a Hybrid Natural Language Processing Method**<br>The use of artificial intelligence in legal analysis represents a significant transformation process in modern legal practices. The literature review showed that the number of studies on Turkish court decisions was limited, and only classification applications were developed. This study aims to summarize, predict, and analyze the Turkish Constitutional Court's decision texts. In this context, first a new, unique dataset was created. The dataset included class labels, original court decision texts, and summary court decision texts created by expert lawyers. Then, a hybrid summary model was established, combining extractive and abstract summarization to summarize long court decisions automatically. With this model, the summarization efficiency was increased, and the token limitation problem common in existing transformative models was eliminated. The model showed good performance for summarization by obtaining Rouge-1, Rouge-2, and Rouge-L scores of 0.6129, 0.5884, and 0.5891 respectively. After the summarization phase, the Legal Judgment Prediction applications were developed. Separate classification models were developed for court decision texts and |

| | |
|---|---|
| | summary court decision texts, obtained using the hybrid model. When the models were compared, the XGBoost model achieved the best performance in legal judgment prediction tasks, with an accuracy rate of 93.84% for full texts and 62.30% for summary texts. In the final stage of the study, the model results were explained using the SHapley Additive exPlanations method. The findings of this study emphasize the superiority of hybrid approaches to legal document analysis. They highlighted the vital role of explainable techniques in improving transparency and reliability in legal processes. |
| 26-ANSP-NLP-022 | **The Effectiveness of Large Language Models in Transforming Unstructured Text to Standardized Formats**<br><br>The exponential growth of unstructured text data presents a fundamental challenge in modern data management and information retrieval. While Large Language Models (LLMs) have shown remarkable capabilities in natural language processing, their potential to transform unstructured text into standardized, structured formats remains largely unexplored - a capability that could revolutionize data processing workflows across industries. This study breaks new ground by systematically evaluating LLMs' ability to convert unstructured recipe text into the structured Cooklang format. Through comprehensive testing of four models (GPT-4o, GPT-4o-mini, Llama3.3:70b, and Llama3.1:8b), an innovative evaluation approach is introduced that combines traditional metrics (WER, ROUGE-L, TER) with specialized metrics for semantic element identification. Our experiments reveal that GPT-4o with few-shot prompting achieves breakthrough performance (ROUGE-L: 0.8209, WER: 0.3509), demonstrating for the first time that LLMs can reliably transform domain-specific unstructured text into structured formats without extensive training. Although model performance generally scales with size, we uncover surprising potential in smaller models like Llama3.1:8b for optimization through targeted fine-tuning. These findings open new possibilities for automated structured data generation across various domains, from medical records to technical documentation, potentially transforming the way organizations process and utilize unstructured information. |
| 26-ANSP-NLP-023 | **Tokenizers for African Language**<br>Despite incredible development in the field of natural language processing (NLP), there has been a huge gap in the performance of NLP tasks between high-resource languages (HRLs) and low-resource languages (LRLs). African languages belong mainly to the LRLs, and one of the major contributing factors to the performance gap is tokenization, which plays a crucial role in NLP performance in general. Many recent studies on African languages often rely on multilingual tokenizers or general-purpose tokenizers, which are not optimized for African languages. This may lead to suboptimal performance in downstream NLP tasks. In this paper, we systematically analyze the performance of language-specific tokenizers for three African languages: Swahili, Hausa, and Yoruba. By experimental results on two classification tasks ( i.e. sentiment classification and news classification), we found that the language-specific tokenizers for African languages consistently outperformed other monolingual tokenizers, with performance gaps of up to 5.43% in sentiment classification and 4.58% in news classification. We also found that multilingual tokenizers generally work well if they are trained in many African languages rather than global HRLs. For instance, African multilingual tokenizers outperformed global multilingual tokenizers by an average of 1.70% in sentiment classification and 1.41% in news classification. The largest observed improvement was 2.61% in news classification using Logistic Regression (LR). Based on the results, we suggest a method for choosing tokenizers when analyzing data or developing models for African languages. |

| | |
|---|---|
| 26-ANSP-NLP-024 | **Unstructured Electronic Health Records of Dysphagic Patients Analyzed by Large Language Models**<br>**Objective:** Dysphagia is a common and complex disorder that complicates both diagnoses and treatment. Consequently, the associated electronic health records (EHR) are often unstructured and complex, posing challenges for systematic data analysis. **Methods and procedures:** In this study, we employ natural language processing (NLP) techniques and large language models (LLMs) to automatically analyze clinical narratives and extract diagnostic information from a diverse set of EHRs. Our dataset includes medical records from 486 patients, representing a group with diverse dysphagic conditions. We analyze diagnoses provided in unstructured free text that do not follow a standardized structure. We utilize clustering algorithms on the extracted diagnostic features to identify distinct groups of patients who share similar pathophysiological swallowing dysfunctions.<br>**Results:** We found that basic NLP techniques often provide limited insights due to the high variability of the data. In contrast, LLMs help to bridge the gap in understanding the nuanced medical information about dysphagia and related conditions. Although applying these advanced LLM models is not straightforward, our results demonstrate that leveraging closed-source models can effectively cluster different categories of dysphagia.<br>**Conclusion:** Our study provides therefore evidence that LLMs are highly promising in future dysphagia research.<br>**Clinical impact:** Dysphagia is a symptom associated with various diseases, though its underlying relation- ships remain unclear. This study demonstrates how analyzing large volumes of electronic health records can help clarify the causes of dysphagia and identify contributing factors. By applying natural language processing, we aim to enhance both understanding and treatment, supporting clinical staff in improving individualized care by identifying relevant patient cohorts.<br>**Clinical and Translational Impact Statement:** This study uses LLMs to efficiently preprocess unstructured EHRs, improving dysphagia diagnosis and patient clustering. It aligns with Clinical Research, enhancing diagnostic speed and enabling personalized treatment. |
| 26-ANSP-NLP-025 | **Word2State: Modeling Word Representations as States with Density Matrices**<br>Polysemy is a common phenomenon in linguistics. Quantum-inspired complex word embeddings based on Semantic Hilbert Space play an important role in natural language processing, which may accurately define a genuine probability distribution over the word space. The existing quantum-inspired works manipulate on the real-valued vectors to compose the complex-valued word embeddings, which lack direct complex-valued pre-trained word representations. Motivated by quantum-inspired complex word embeddings, we propose a complex-valued pre-trained word embedding based on density matrices, called Word2State. Unlike the existing static word embeddings, our proposed model can provide non-linear semantic composition in the form of amplitude and phase, which also defines an authentic proba- bilistic distribution. We evaluate this model on twelve datasets from the word similarity task and six datasets from the relevant downstream tasks. The experimental results on different tasks demonstrate that our proposed pre-trained word embedding can capture richer semantic information and exhibit greater flexibility in expressing uncertainty. |