



P-ISSN 2355-2794  
E-ISSN 2461-0275

# Computational Linguistics and Natural Language Processing Techniques for Semantic Field Extraction in Arabic Online News

Maulana Ihsan Ahmad\*  
Moh. Kanif Anwari

<sup>1</sup>Master's Program of Arabic Language and Literature, Faculty of Adab and Cultural Sciences, Universitas Islam Negeri Sunan Kalijaga, Yogyakarta 55281, INDONESIA

## Abstract

*The research aimed to extract semantic fields from Arabic online news and advance Natural Language Processing (NLP) applications in understanding and managing news information effectively. It provides a comprehensive approach to processing and analyzing large volumes of Arabic news data by integrating semantic field analysis, NLP, and computational linguistics. Using quantitative methods, Arabic news articles were collected and processed with Python, a popular programming language in data analysis, and applied various NLP techniques and machine learning models to accurately extract semantic fields. The primary objective was to evaluate the effectiveness of different classification models in categorizing Arabic news and to identify the most suitable model for semantic field extraction. The research evaluated five classification models: Naive Bayes, Support Vector Machine (SVM), Logistic Regression, Random Forest, and Gradient Boosting. Among these, SVM achieves the highest overall accuracy of 90%. Specifically, SVM demonstrated exceptional performance in categorizing sports-related news, with a 99% probability and an F1-Score of 98%. However, it faced challenges in categorizing health and science news, achieving a lower F1-Score of 79%. Overall, the study demonstrated the effectiveness of computational methods, particularly SVM, in classifying Arabic news and extracting semantic fields, thereby advancing NLP and computational linguistics. The findings highlighted the potential of SVM for accurate*

---

\* Corresponding author, email: [23201011009@student.uin-suka.ac.id](mailto:23201011009@student.uin-suka.ac.id)

**Citation in APA style:** Ahmad, M. I., & Anwari, M. K. (2024). Computational linguistics and natural language processing techniques for semantic field extraction in Arabic online news. *Studies in English Language and Education*, 11(3), 1686-1709.

Received March 31, 2023; Revised July 15, 2024; Accepted August April 28, 2024; Published Online September 30, 2024

<https://doi.org/10.24815/siele.v11i3.38090>

*news analysis and the need for further enhancement of NLP techniques to address multilingual and domain-specific challenges.*

**Keywords:** Computational linguistics, natural language processing, semantic field.

## 1. INTRODUCTION

The digital landscape is evolving rapidly and the proliferation of information accessible via online platforms continues unabated, encompassing a wide array of online news sources, including those in Arabic. Online news has emerged as a principal reservoir of information for a diverse global audience. Serving as a conduit for contemporaneous events of public interest, news holds considerable sway, particularly concerning matters impacting public figures or events directly affecting readers (Gordon, 2018). However, amidst this exponential growth, the significance of online news in the sphere of natural language processing (NLP) cannot be overstated, owing to its wealth of essential information (Pennington et al., 2014).

One notable aspect of NLP is automatic semantic field extraction, a process entailing the navigation of a complex web of word associations based on factors such as similarity, context, and connections to specific lexemes (Fauziyah, 2023). In Arabic news, NLP poses heightened challenges due to its complex and unique structure, as well as its richness in meaning (Al-Sarayreh et al., 2023). For instance, a sentence such as “خسر منتخب قطر أمام نظيره الاسترالي بنتيجة ٣ / ٠ صفر في ختام مباريات دور ال١٦ لبطولة آسيا للرجال” as “Qatar team lost to the Australian counterpart with a score of 3-0 in the closing match of the round of 16 of the Asian men’s volleyball tournament currently held in the Emirate of Dubai, United Arab Emirates” can be situated within the semantic field ‘الرياضة’ ‘sports’ due to word associations like ‘خسر’ ‘lost’, ‘منتخب’ ‘team’, ‘بنتيجة’ ‘score’, ‘نظير’ ‘counterpart’, ‘مباريات’ ‘matches’, and ‘الكرة’ ‘ball’.

The Arabic language presents several specific challenges for NLP, particularly for non-Arabic speakers. First, Arabic is a highly inflected language, meaning that words can take many different forms depending on their role in a sentence. This morphological complexity includes prefixes, suffixes, and infixes that modify the root of the word, making it challenging to identify and analyze semantic fields (Issa, 2023). Second, Arabic script is written from right to left and often omits short vowels, leading to ambiguities that require contextual understanding to resolve (Saiegh-Haddad & Henkin-Roitfarb, 2014). Third, the language has a rich system of derivational morphology, where new words are created from root patterns, adding another layer of complexity (Ryding, 2014). These linguistic characteristics necessitate advanced techniques and careful consideration when developing NLP models for Arabic semantic field extraction.

The importance of discerning thematic content within a text has been well-established, particularly through the utilization of semantic fields. This holds not only within sports but also across a spectrum of other domains such as politics, economy, crime, technology, science, culture, social issues, national affairs, and health. Consequently, the imperative development of NLP methodologies for automatically extracting semantic fields from texts becomes evident. NLP, at its core, integrates

computational linguistic principles and artificial intelligence to comprehend, analyze, and generate text based on natural language (Nerbonne, 2022). In essence, NLP seeks to address inquiries about “how computers can be utilized to understand and manipulate natural language text” across a multitude of tasks (Torregrosa et al., 2023). Given that the majority of internet content is text-based, NLP emerges as a profoundly relevant and effective approach.

Recent strides in NLP have focused on enhancing the comprehension, analysis, and processing of texts in various languages, including Arabic. NLP, a branch of artificial intelligence, deals with the interaction between computers and human language and includes applications like machine translation, sentiment analysis, text summarization, and information retrieval. While NLP techniques have been widely applied to languages such as English, Chinese, and Spanish, there is a growing focus on Arabic due to its linguistic complexity and increasing online content. Enhancing NLP capabilities for Arabic not only contributes to the global field but also addresses the specific needs and challenges of Arabic speakers. Research has refined text classification methodologies through deep learning and machine learning techniques (Habeeb et al., 2023; Wahdan et al., 2024) and tailored sentiment analysis for Arabic contexts (Abdelfattah et al., 2023; Al-Harbi et al., 2023; Aoumeur et al., 2023; Berrimi et al., 2023; Mhamed et al., 2023). Additionally, extensive exploration into machine translation and text summarization (Çetinel, 2022; Imam & Masama, 2023; Khanjari, 2023) has advanced NLP-related research within Arabic linguistics. According to the above conception, extracting semantic fields from Arabic online news texts is also a form of extensive exploration, which makes a significant contribution to academic discourse.

Despite the advancements in NLP, there is a notable gap in applying these techniques to Arabic news content, particularly concerning the extraction and analysis of semantic fields. This study aims to address this gap by exploring various NLP methods and machine learning models to determine their effectiveness in extracting semantic fields from Arabic news articles. The primary research questions guiding this study are:

1. What are the most effective NLP techniques for extracting semantic fields from Arabic online news?
2. How do these techniques perform across different news categories?

To achieve this, the research has three main objectives, namely to develop a methodological framework for the automated extraction of semantic fields from Arabic online news, to evaluate the performance of different NLP techniques and machine learning models in this context, and to identify the most suitable model for accurately categorizing and analyzing Arabic news content. By achieving these objectives, this research seeks to advance the field of NLP for Arabic language processing and provide valuable insights into the automated analysis of online news media.

By developing NLP methods to extract semantic fields from Arabic online news, this research aims to broaden the scope of NLP across various contexts, encompassing computational linguistic studies, text analysis, and practical applications such as news classification. Ultimately, it seeks to offer theoretical contributions to the field of NLP by developing a novel methodological framework specifically tailored to address the unique linguistic challenges of Arabic. The research also holds significant practical implications in comprehending and managing news information in Arabic by

improving the accuracy and efficiency of classifying Arabic news articles into relevant categories, enabling news agencies and information retrieval systems to better manage and curate content.

## **2. LITERATURE REVIEW**

### **2.1 Semantic Field**

Research on semantic fields or domains has become a central focus in linguistics, elucidating conceptual territories that interconnect words or expressions with shared meanings or concepts. Semantic fields help linguists and researchers understand how language encapsulates cultural and contextual meanings, revealing the intricacies of human communication (Gärdenfors, 2017). Numerous scholarly inquiries have shed light on various facets within this domain, emphasizing its profound implications for linguistic and cultural comprehension.

One significant area of research is the examination of how semantic fields are influenced by major events and cultural contexts. For instance, Syafroni (2021) explored the semantic field of education during the COVID-19 pandemic, providing insights into how the pandemic influenced educational terminology and concepts. This research highlighted how significant events shaped the semantic landscape of a particular field.

Similarly, understanding specific terms within various cultural and linguistic contexts provides deeper insights into semantic fields. Yuliana (2021) examined the semantic fields within Meno-Meni dialect of the Sasak language by focusing on the term *terjatuh* 'falling'. This study illustrated how specific words carry different connotations and meanings within cultural contexts. In a similar vein, Sutarsih (2020) analyzed the semantic nuances of *membesar* 'to grow' in Javanese, uncovering divergences in meaning across various cultural settings. These studies illustrate the importance of understanding semantic fields in different linguistic and cultural contexts.

Further research has focused on the influence of cultural context on lexical semantics. Patmasada et al. (2020) and Sumarna et al. (2020) investigated into the semantic relations of verbs within the Kapuas Hulu dialect of Malay and the weaving practices of the Sambas Malay community, respectively. Their research highlighted the influence of cultural context on lexical semantics. Hilmi et al. (2022) explored the varied meanings of words within Sasak dialects, while Isnaeni & Riana (2022) investigated dialectometric isolec of Komering and Lampung, further emphasizing the diversity and richness of semantic fields within different linguistic environments.

In poetic and literary studies, Nor and Hussain (2020) explored the conceptualization of space in language through Malay pantoum and Irbah (2022) probed associative meanings in poetry anthologies. These studies highlight the relationship interplay between language, culture, and literary expression in shaping semantic fields.

Comparative analyses offer further insights into semantic fields across different languages and cultures. Bekirov and Rublyova (2019) and Georgakopoulos and Polis (2022) examined disparities and congruences in idiom usage and methodologies related to semantic mapping. These comparative perspectives highlight the universal

and unique aspects of semantic fields in various linguistic contexts. [Nikolina \(2023\)](#) and [Serrano-Niza \(2018\)](#) analyzed semantic structures in Arabic literary texts and lexicography, respectively. [Denisenko and Kalinina \(2020\)](#) and [Sukiman \(2015\)](#) investigated the verbal expression of meaning and metaphors in Arabic, while [Arianto \(2018\)](#) focused on semantic fields within Arabic poetry.

This body of research stresses the imperative of exploring semantic fields to foster a nuanced understanding of the Arabic language. It forms a strong foundation for further inquiry into semantic field extraction within Arabic-language online news, utilizing NLP methodologies and linguistic computing. Through NLP approaches, researchers can develop sophisticated techniques for automated extraction and analysis of semantic fields in Arabic-language online news, advancing our comprehension of language and meaning in digital contexts.

## 2.2 Natural Language Processing (NLP)

Natural Language Processing (NLP) is an interdisciplinary field that combines computational linguistics and artificial intelligence to enable computers to understand, interpret, and generate human language ([Bolshakov & Gelbukh, 2004](#)). This rapidly advancing field covers a wide range of applications, including text processing, sentiment analysis, medical research, and cultural studies. NLP techniques are crucial for transforming unstructured text data into a structured format suitable for analysis such as:

- case folding to convert all characters in the text to lowercase to ensure uniformity and reduce the complexity of text analysis.
- tokenization to split the text into individual units, such as words or phrases, called tokens. Tokenizers like “Farasa” are specifically designed for Arabic text, helping to accurately segment words and handle language-specific challenges.
- stop word removal to eliminate common words that carry little informational value (e.g., “و”, “في”) to focus on more meaningful terms.
- text pre-processing involves cleaning the text by removing non-alphanumeric characters and numbers to prepare it for further analysis.

Recent studies highlight key advancements and challenges in these diverse areas, illustrating the dynamic evolution and broad impact of NLP.

Sentiment analysis is a critical application of NLP, helping to assess public opinion and emotional responses from textual data. [Khurana et al. \(2023\)](#) outlined four phases of NLP development, stressing its importance in sentiment analysis and enhancing understanding of medical content. [Pellicer et al. \(2023\)](#) emphasized data augmentation techniques to improve NLP data quality, a crucial factor in effective sentiment analysis. [Drikvandi and Lawal \(2023\)](#) introduced sparse principal component analysis to address the challenge of text data sparsity, thereby improving the accuracy and reliability of sentiment analysis models.

NLP has made significant strides in medical research, particularly in understanding and processing medical texts. [Müller et al. \(2023\)](#) demonstrated NLP role in pandemic contexts with COVID-Twitter-BERT, an adaptation that helped track and analyze public health information during the COVID-19 pandemic. This highlights the potential of NLP in managing and interpreting large volumes of medical data for better healthcare outcomes.

NLP also contributes to cultural studies by analyzing and interpreting literary and historical texts. Bashir et al. (2023) applied NLP to Quranic research and pioneered prompt-based learning to enhance the understanding of religious texts. Mahmoud and Negied (2023) proposed deep learning techniques for analyzing Arabic poetry, illustrating NLP capability to preserve and interpret cultural heritage through advanced computational methods.

Advancements in machine learning and computational techniques are driving the evolution of NLP. Wu et al. (2023) explored the application of graph neural networks in NLP, highlighting innovative approaches to text processing. Treviso et al. (2023) discussed resource-efficient NLP methods, addressing the need for sustainable and scalable NLP solutions. Arkhangelskaya and Nikolenko (2023) reviewed deep learning architectures, emphasizing their role in enhancing NLP capabilities.

NLP impact spans various industries, providing tailored solutions for specific needs. Pauzi and Capiluppi (2023) examined NLP applications in software traces, improving software maintenance and development processes. Boulieris et al. (2023) introduced the fraudNLP dataset, which aims to enhance fraud detection systems through advanced NLP techniques.

Arabic NLP has seen significant progress with researchers focusing on language-specific challenges and solutions. Ghamdi et al. (2023) updated BERT for Arabic dependency parsing, enhancing syntactic analysis capabilities for the Arabic language. Bourahouat et al. (2023) reviewed Arabic NLP (ANLP) techniques, providing a comprehensive overview of recent advancements. Moussa and Mourhir (2023) introduced a Moroccan dialect entity recognition dataset, while Altamimi and Alayba (2023) released a modern Arabic news dataset, contributing valuable resources for further research. Larabi-Marie-Sainte et al. (2023) developed a clustering model for Arabic texts and Jamaledyn et al. (2023) proposed methods for Arabic news classification. Rahma et al. (2023) reviewed techniques for sarcasm detection in Arabic, highlighting the unique challenges and solutions in this area. These studies illustrate NLP dynamic evolution and broad impact across industries and disciplines. The field continues to promise further innovations in linguistic understanding and application as it advances.

### **2.3 Computational Linguistics**

Computational linguistics is the study of using computational methods to process and analyze human language. This interdisciplinary field bridges computer science and linguistics, providing tools and techniques for understanding and generating human language (Mehdivi, 2008). Its applications span across various domains such as health, education, crime analysis, social media sentiment, and information management. This section categorizes recent advancements in computational linguistics to highlight its diverse applications and impact. Thus, computational linguistics has a broad impact across various fields, enhancing understanding and management of language-related challenges in health, education, crime analysis, social media, and information management through computational methods.

Computational linguistics plays a crucial role in health and medical research. Giarnieri and Scardapane (2023) discussed AI advancements in cytology, the study of cells, showcasing the integration of language technologies for medical diagnosis. Their study highlights how computational methods can enhance the accuracy and efficiency

of diagnosing diseases by analyzing medical texts and records. Additionally, [Alliheibi et al. \(2021\)](#) analyzed Saudi citizens' COVID-19 vaccine sentiments on Twitter, demonstrating the technological role in health sentiment analysis. These studies underline the potential of computational linguistics in improving healthcare outcomes through better data analysis and understanding of public health sentiments.

In the field of education, computational linguistics helps improve teaching and learning processes by analyzing communication patterns. [Shin et al. \(2023\)](#) analyzed pedagogical discourse markers in online algebra learning, revealing the impacts of instructor-student communication on learning outcomes. Their research provides insights into how computational methods can be used to enhance educational interactions and support effective teaching strategies. [Martin et al. \(2023\)](#) longitudinally studied pragmatic language in neurodevelopmental disorders, further demonstrating the importance of computational analysis in educational and developmental contexts.

Computational linguistics is also applied in crime analysis and law enforcement. [Alruily et al. \(2014\)](#) developed a crime profile system for Arabic, employing computational linguistics to analyze crime data. This system helps in profiling and understanding criminal activities through the analysis of linguistic patterns in crime reports and other related texts. The integration of computational methods in crime analysis aids law enforcement agencies in making data-driven decisions and improving public safety.

Social media sentiment analysis is a prominent application of computational linguistics, where researchers analyze public opinion and emotional responses expressed on social platforms. [Heaton et al. \(2023\)](#) and [Davani et al. \(2023\)](#) explored social media sentiment analysis, addressing issues from A-level exam controversy to hate speech detection. These studies highlight the importance of understanding public sentiment in real time, enabling organizations and policymakers to respond appropriately to public concerns and issues.

In information management, computational linguistics aids in analyzing and crafting effective policies. [Gürsakal et al. \(2023\)](#) and [Kim et al. \(2023\)](#) applied high-frequency word analysis in scientific articles and investment risk disclosures, respectively. Their work demonstrates how computational methods can enhance the clarity and effectiveness of information dissemination. [Zhang et al. \(2023\)](#) highlighted the gap in privacy understanding, advocating for the use of computational linguistics in creating effective privacy policies. These studies demonstrate the role of computational linguistics in improving information management and ensuring that policies are well-understood and effectively communicated.

### 3. METHODS

#### 3.1 Research Design

This quantitative research employed both an exploratory and experimental design to investigate the potential integration of semantic field linguistic analysis, natural language processing (NLP) techniques, and computational linguistic methods.

An exploratory design was used to examine the fundamental aspects of the research problem and identify new variables, patterns, and relationships (Zahrah & Octora, 2023). While the experimental design involved controlled testing to determine the cause-and-effect relationships between variables (Fatihah et al., 2023). The combination of these designs allowed for a thorough investigation from both exploratory and confirmatory perspectives. This integrated approach was highly relevant to the research as it ensured a comprehensive analysis of the theoretical and practical aspects of semantic field extraction in Arabic online news.

### 3.2 Data Source

The data source for this research comprised 152.793 online news articles from six sources: [alsumaria.tv](http://alsumaria.tv), [assahraa.ma](http://assahraa.ma), [al-sharq.com](http://al-sharq.com), [addustour.com](http://addustour.com), [akhbarelsyaha.com](http://akhbarelsyaha.com), and [sarayanews.com](http://sarayanews.com). These sources were carefully selected for their diverse geographical and topical coverage, representing different regions and perspectives within the Arab world such as Iraq, Morocco, Qatar, and Jordan. This diversity was essential for providing a comprehensive dataset that reflected a wide range of semantic fields and linguistic usage across various contexts.

These articles spanned 11 categories, including أمن 'security', اقتصاد 'economy', تكنولوجيا 'technology', ثقافة 'culture', دوليات 'international', رياضة 'sports', سياحة 'travel', سياسة 'politics', صحة 'health', علم 'science', and مجتمع 'society', with each category represented by 10.000 to 20.000 articles, totaling 42.661.503 words (754.584 in Arabic, 74.727 non-Arabic words, and 41.832.192 stopwords).

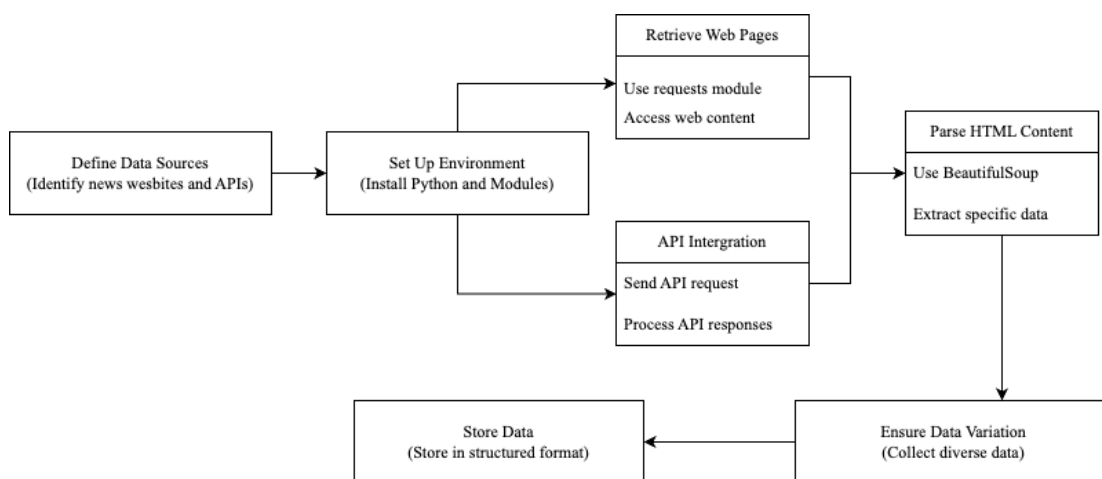
The selection of these categories ensured a comprehensive coverage of various domains, which is crucial for analyzing the semantic fields in different contexts. This extensive dataset allowed for a detailed examination of how semantic fields were constructed and utilized in Arabic-language online news, addressing the research questions related to semantic field extraction and the application of NLP techniques in this context.

### 3.3 Data Collection

Data collection involved web scraping techniques or using application programming interfaces (APIs) for automatic and structured extraction of Arabic news data. The implementation was done using Python, a popular programming language known for its versatility and ease of use in data analysis. The process utilized several key modules such as *requests* to allow the retrieval of web pages and their content, *sys* to provide access to some variables used or maintained by the Python interpreter and to functions that interact strongly with the interpreter, and *BeautifulSoup* library to enable the extraction of specific data from web pages. Ensuring adequate variation in the collected data was crucial for effective model training and testing.

The data collection process was methodically structured to ensure a comprehensive extraction of Arabic news data. First, the data sources were defined by identifying relevant news websites and APIs. Next, the environment setup involved installing Python, a popular programming language known for its versatility and ease of use in data analysis and necessary modules, such as *requests*, *sys*, and *BeautifulSoup*. The *sys* module played a supportive role by providing access to command-line arguments for configuring the script, managing system-specific

parameters, and handling script execution errors. The process then progressed to retrieving web pages where the *requests* module was used to fetch web content. This was followed by parsing HTML content using the *BeautifulSoup* library to extract specific data elements from the retrieved HTML. Concurrently, API Integration was performed by sending requests to APIs and processing the responses to gather structured data. To ensure the collected data had sufficient diversity, data variation was emphasized, which involved collecting data from a variety of sources and categories. Finally, the collected data was stored in a structured format, such as CSV or a database, to facilitate easy access and subsequent analysis. The detailed steps of this process can be seen in Figure 1 below.



**Figure 1.** Flowchart of data collection.

### 3.4 Data Analysis

In this research, several key steps were followed for data analysis. Firstly, the text underwent pre-processing, which involved removing non-alphanumeric characters, numbers, and stop words, and segmenting the text into segmented words using a word tokenizer like “Farasa”. Additionally, NLP techniques such as case folding, tokenization, and stop word removal were applied to structure the text for analysis. Subsequently, raw data was vectorized with TF-IDF unigrams, and models including naive bayes, SVM, logistic regression, random forest, and gradient boosting were trained using 80% training data, 5% testing data, and 15% validation data. Model evaluation focused on accuracy, precision, recall, and F1-score, with an emphasis on accuracy in predicting document categories. Finally, the classification model was used to predict news categories, and semantic field extraction was performed to identify information within each news item.

## 4. RESULTS

### 4.1 Classification Model Performance

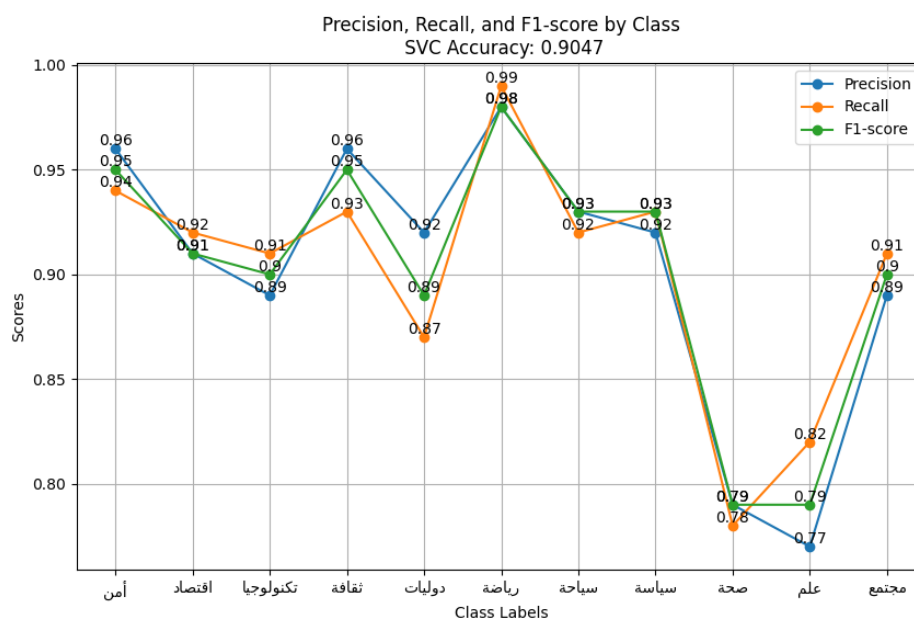
In the analysis, we evaluate the performance of five different classification models for classifying online Arabic news. The models tested were Naive Bayes,

Support Vector Machine (SVM), Logistic Regression, Random Forest, and Gradient Boosting. Performance evaluation employs standard metrics, including accuracy, precision, recall, and F1-score as presented in Table 1.

**Table 1.** Evaluation of news classification models.

| Model               | Accuracy | Precision | Recall | F1-score | Execution time (s) |
|---------------------|----------|-----------|--------|----------|--------------------|
| Naive Bayes         | 0.8319   | 0.8414    | 0.8319 | 0.8308   | 0.438164           |
| SVM                 | 0.9047   | 0.9054    | 0.9047 | 0.9049   | 875.475578         |
| Logistic Regression | 0.8962   | 0.8978    | 0.8962 | 0.8966   | 0.396247           |
| Random Forest       | 0.8668   | 0.8714    | 0.8668 | 0.8669   | 2.876709           |
| Gradient Boosting   | 0.8673   | 0.8605    | 0.8573 | 0.8581   | 0.340603           |

The evaluation results in Table 1 reveal that SVM achieved the highest accuracy rate of 90.47%, followed closely by logistic regression with 89.62%, Gradient Boosting at 86.73%, Random Forest with 86.68%, and Naive Bayes with 83.19%. Accuracy serves as an important metric that reflects the overall correctness of the model's predictions. In terms of precision, SVM also led with 90.54%, followed by Logistic Regression with 89.78%, Random Forest with 87.14%, Gradient Boosting with 86.05%, and Naive Bayes with 84.14%. Precision indicates how accurately the model classifies positive data. For recall, SVM again ranked highest at 90.47%, with Logistic Regression at 89.62%, Random Forest at 86.68%, Gradient Boosting at 85.73%, and Naive Bayes at 83.19%. Recall is critical for measuring the model's ability to correctly identify positive data points. Similarly, SVM led in F1 score with a value of 90.49%, followed by Logistic Regression at 89.66%, Random Forest at 86.69%, Gradient Boosting at 85.81%, and Naive Bayes at 83.08%. The F1 score combines both precision and recall, providing an overall performance assessment of each model.



**Figure 2.** Comparison graph of precision, recall, and F1-score per category in SVM model.

When examining execution time, Gradient Boosting was the fastest, completing in 0.34 seconds, while Logistic Regression followed with 0.396 seconds, Naive Bayes with 0.438 seconds, Random Forest with 2.877 seconds, and SVM with the longest execution time at 875.476 seconds. Execution time is a critical factor for deploying models, particularly in real-time or resource-limited environments. In summary, while SVM demonstrated superior performance in terms of accuracy, precision, recall, and F1 score, Gradient Boosting excelled in execution speed. The choice of a model should, therefore, depend on the specific requirements of the application and the relevant evaluation criteria. For the semantic field analysis, SVM was selected due to its high accuracy. Figure 2 illustrates the classification details using the SVM model, which shows the highest F1 score of 0.98 (98%) in the sports category.

## 4.2 Semantic Field Extraction

In this section, we present the successful extraction of relevant meaning from Arabic online news articles. The choice of the SVM model for classification, based on its superior accuracy, enabled us to delve into specific examples. We focused on news from the sports category, exhibiting the highest F1-score among the 11 categories, as depicted in Figure 1.

The process of semantic field extraction revealed deeper contextual meaning within the text. Consider the following sports news:

N1:

فوز الوكرة والسد في تصفيات كأس سمو الأمير للكرة الطائرة فاز فريقا الوكرة والسد على الخور والشمال تواليًا، في منافسات الجولة الأولى من التصفيات التمهيدية المؤهلة لبطولة كأس سمو الأمير المفدى للموسم الرياضي (2023 - 2024)، المقامة في الفترة من 24 إلى 29 مارس الجاري. وجاء فوز الوكرة على الخور بثلاثة أشواط مقابل شوط واحد، بواقع نتائج 25-21، و26-24، و25-15، و20-25، ليضع في رصيده أول ثلاث نقاط.

The victories of Al-Wakrah and Al-Sadd in the Emir Cup volleyball qualifiers. The Al-Wakrah and Al-Sadd teams successively defeated Al-Khor and Al-Shamal in the first round of qualifying for the Emir Cup championship of the sports season (2023-2024), which took place from March 24th to 29th. Al-Wakrah defeated Al-Khor with a score of three sets to one, with set scores of 21-25, 26-24, 25-15, and 25-20, thus earning the first three points in the standings. (Al-Sharq, 2024)

Manual semantic field extraction identified several key terms related to volleyball matches. For example, words such as 'cup' and 'volleyball' clearly associate the text with sports events. Team names like 'Al-Wakrah' and 'Al-Khor' were also identified, indicating specific teams involved. Other terms, such as 'qualifiers' and 'victory', further emphasized the competitive nature of the event. The term 'sets' provided insights into the structure of volleyball matches, suggesting that the text fell under the sports category.

However, NLP and computational linguistics take a more automated approach to this semantic field extraction. This process includes several steps like punctuation and number removal, tokenization, and stop-word removal. These steps, illustrated below, show how the text was prepared for vectorization. Table 2 provides a comparison between the original and cleaned text, demonstrating how numbers and punctuation were systematically removed (in red ink) for improved analysis.

**Table 2.** Numbers and punctuation removal.

| Original text (Arabic)                                                                                                                                                                                                                                                                                                                                                      | Meaning                                                                                                                                                                                                                                                                                                                                                                                                                                                                     | Numbers and punctuation removal                                                                                                                                                                                                                                                                                                                     | Meaning                                                                                                                                                                                                                                                                                                                                                                                                                         |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| فوز الوكرة والسد في تصفيات كأس سمو الأمير فاز فريقا للكرة الطائرة الوكرة والسد على الخور والشمال تواليا في منافسات الجولة الأولى من التصفيات التمهيدية المؤهلة لبطولة كأس سمو الأمير المفدى للموسم الرياضي (2023 - 2024)، المقامة في الفترة من 24 إلى وجاء 29 مارس الجاري. فوز الوكرة على الخور بثلاثة أشواط مقابل شوط واحد، بواقع نتائج و و و ليضع في رصيده أول ثلاث نقاط. | The victories of Al-Wakrah and Al-Sadd in the Emir Cup volleyball qualifiers. The Al-Wakrah and Al-Sadd teams successively defeated Al-Khor and Al-Shamal in the first round of qualifying for the Emir Cup championship of the sports season (2023-2024), which took place from March 24th to 29th. Al-Wakrah defeated Al-Khor with a score of three sets to one, with set scores of 21-25, 26-24, 25-15, and 25-20, thus earning the first three points in the standings. | فوز الوكرة والسد في تصفيات كأس سمو الأمير فاز فريقا للكرة الطائرة الوكرة والسد على الخور والشمال تواليا في منافسات الجولة الأولى من التصفيات التمهيدية المؤهلة لبطولة كأس سمو الأمير المفدى للموسم الرياضي المقامة في الفترة وجاء من إلى مارس الجاري فوز الوكرة على الخور بثلاثة أشواط مقابل شوط واحد بواقع نتائج و و و ليضع في رصيده أول ثلاث نقاط | The victories of Al-Wakrah and Al-Sadd in the Emir Cup volleyball qualifiers. The Al-Wakrah and Al-Sadd teams successively defeated Al-Khor and Al-Shamal in the first round of qualifying for the Emir Cup championship of the sports season, which took place from March to. Al-Wakrah defeated Al-Khor with a score of three sets to one, with set scores of , , and , thus earning the first three points in the standings. |

Once the text is cleaned, the next step is tokenization. Tokenization involves splitting the cleaned text into individual words or tokens, which allows the model to analyze each component separately. As shown in Table 3, the cleaned text is broken down into individual tokens, which are then used in further NLP steps.

**Table 3.** Tokenization process.

| Cleaned text                                                                                                                                                                                                                                                                                                                                        | Meaning                                                                                                                                                                                                                                                                                                                                                                                                                         | Tokenization                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                | Meaning                                                                                                                                                                                                                                                                                                                                                                                                                     |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| فوز الوكرة والسد في تصفيات كأس سمو الأمير فاز فريقا للكرة الطائرة الوكرة والسد على الخور والشمال تواليا في منافسات الجولة الأولى من التصفيات التمهيدية المؤهلة لبطولة كأس سمو الأمير المفدى للموسم الرياضي المقامة في الفترة وجاء من إلى مارس الجاري فوز الوكرة على الخور بثلاثة أشواط مقابل شوط واحد بواقع نتائج و و و ليضع في رصيده أول ثلاث نقاط | The victories of Al-Wakrah and Al-Sadd in the Emir Cup volleyball qualifiers. The Al-Wakrah and Al-Sadd teams successively defeated Al-Khor and Al-Shamal in the first round of qualifying for the Emir Cup championship of the sports season, which took place from March to. Al-Wakrah defeated Al-Khor with a score of three sets to one, with set scores of , , and , thus earning the first three points in the standings. | [‘فوز’، ‘الوكرة’، ‘والسد’، ‘في’، ‘تصفيات’، ‘كأس’، ‘سمو’، ‘الأمير’، ‘للكرة’، ‘الطائرة’، ‘فاز’، ‘فريقا’، ‘الوكرة’، ‘والسد’، ‘على’، ‘الخور’، ‘والشمال’، ‘تواليا’، ‘في’، ‘منافسات’، ‘الجولة’، ‘الأولى’، ‘من’، ‘التصفيات’، ‘التمهيدية’، ‘المؤهلة’، ‘لبطولة’، ‘كأس’، ‘سمو’، ‘الأمير’، ‘المفدى’، ‘للموسم’، ‘الرياضي’، ‘المقامة’، ‘في’، ‘الفترة’، ‘من’، ‘إلى’، ‘مارس’، ‘الجاري’، ‘وجاء’، ‘فوز’، ‘الوكرة’، ‘على’، ‘الخور’، ‘بثلاثة’، ‘أشواط’، ‘مقابل’، ‘شوط’، ‘واحد’، ‘بواقع’، ‘نتائج’، ‘و’، ‘و’، ‘و’، ‘ليضع’، ‘في’، ‘رصيده’، ‘أول’، ‘ثلاث’، ‘نقاط’] | [‘win’، ‘Al-Wakrah’، ‘Al-Sadd’، ‘in’، ‘qualifiers’، ‘cup’، ‘Highness’، ‘prince’، ‘for volleyball’، ‘won’، ‘teams’، ‘Al-Wakrah’، ‘Al-Sadd’، ‘on’، ‘Al-Khor’، ‘Al-Shamal’، ‘respectively’، ‘in’، ‘competitions’، ‘round’، ‘first’، ‘from’، ‘qualifiers’، ‘preliminary’، ‘qualifying’، ‘championship’، ‘cup’، ‘Highness’، ‘prince’، ‘honorable’، ‘season’، ‘sports’، ‘held’، ‘in’، ‘period’، ‘from’، ‘to’، ‘March’، ‘current’، |

|  |  |  |                                                                                                                                                                                    |
|--|--|--|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|  |  |  | ‘came’, ‘win’, ‘Al-Wakrah’, ‘on’, ‘Al-Khor’, ‘in three’, ‘sets’, ‘versus’, ‘one’, ‘set’, ‘results’, ‘and’, ‘and’, ‘and’, ‘place’, ‘in’, ‘its account’, ‘first’, ‘three’, ‘points’] |
|--|--|--|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

Stop word removal follows tokenization. Stop words are common words that do not contribute significant meaning, such as *في* ‘and’ or *على* ‘on’. Removing these words helps focus the analysis on key terms. In Table 4, you can see the result of stop word removal applied to the tokenized text.

**Table 4.** Stopword removal.

| Tokenization                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 | Meaning                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        | Filtered text                                                                                                                                                                                                                                                                          | Meaning                                                                                                                                                                                                                                                                                                                                              |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| [‘فوز’, ‘الوكرة’, ‘والسد’, ‘في’, ‘تصفيات’, ‘كأس’, ‘سمو’, ‘الأمير’, ‘اللكرة’, ‘الطائرة’, ‘فاز’, ‘فريقا’, ‘الوكرة’, ‘والسد’, ‘على’, ‘الخور’, ‘والشمال’, ‘تواليا’, ‘في’, ‘منافسات’, ‘الجولة’, ‘الأولى’, ‘من’, ‘التصفيات’, ‘التمهيدية’, ‘المؤهلة’, ‘لبطولة’, ‘كأس’, ‘سمو’, ‘الأمير’, ‘المفدى’, ‘للموسم’, ‘الرياضي’, ‘المقامة’, ‘في’, ‘الفترة’, ‘من’, ‘إلى’, ‘مارس’, ‘الجاري’, ‘وجاء’, ‘فوز’, ‘الوكرة’, ‘على’, ‘الخور’, ‘بثلاثة’, ‘أشواط’, ‘مقابل’, ‘شوط’, ‘واحد’, ‘بواقع’, ‘نتائج’, ‘و’, ‘و’, ‘و’, ‘ليضع’, ‘في’, ‘رصيده’, ‘أول’, ‘ثلاث’, ‘نقاط’] | [‘win’, ‘Al-Wakrah’, ‘Al-Sadd’, ‘in’, ‘qualifiers’, ‘cup’, ‘Highness’, ‘prince’, ‘for volleyball’, ‘won’, ‘teams’, ‘Al-Wakrah’, ‘Al-Sadd’, ‘on’, ‘Al-Khor’, ‘Al-Shamal’, ‘respectively’, ‘in’, ‘competitions’, ‘round’, ‘first’, ‘from’, ‘qualifiers’, ‘preliminary’, ‘qualifying’, ‘championship’, ‘cup’, ‘Highness’, ‘prince’, ‘honorable’, ‘season’, ‘sports’, ‘held’, ‘in’, ‘period’, ‘from’, ‘to’, ‘March’, ‘current’, ‘came’, ‘win’, ‘Al-Wakrah’, ‘on’, ‘Al-Khor’, ‘in three’, ‘sets’, ‘versus’, ‘one’, ‘set’, ‘results’, ‘and’, ‘and’, ‘and’, ‘place’, ‘in’, ‘its account’, ‘first’, ‘three’, ‘points’] | فوز الوكرة والسد تصفيات كأس سمو الأمير للكرة الطائرة فاز فريقا الوكرة والسد الخور والشمال تواليا منافسات الجولة الأولى التصفيات التمهيدية المؤهلة لبطولة كأس سمو الأمير المفدى للموسم الرياضي المقامة الفترة الجاري وجاء فوز الوكرة الخور بثلاثة أشواط شوط بواقع نتائج ليضع رصيده نقاط | win Al-Wakrah Al-Sadd qualifiers cup Highness prince for volleyball won teams Al-Wakrah Al-Sadd Al-Khor Al-Shamal respectively competitions round first qualifiers preliminary qualifying championship cup Highness prince honorable season sports held period current came win Al-Wakrah Al-Khor in three sets set results place its account points |

After pre-processing the text, vectorization is applied, which converts the text into numerical form that machine learning models can understand. The vectorization results for the filtered text are presented in Table 5, showing the importance of each word by assigning it a weight.

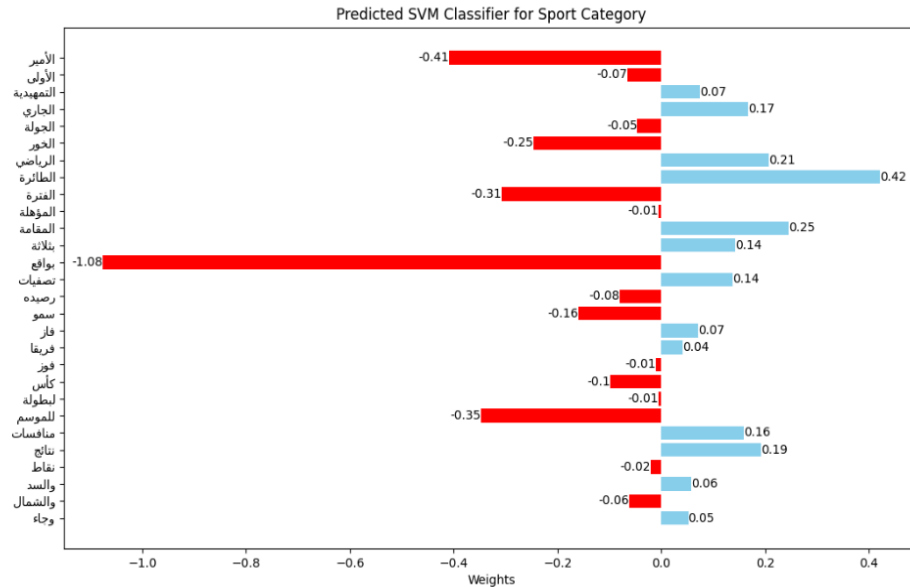
**Table 5.** Text vectorization.

| Filtered text                                                                                                                                                                                                                                                                                                                             | Meaning                                                                                                                                                                                                                                                                                                                                                                                          | Vectorization                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          | Meaning                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| فوز الوكرة والسد<br>تصفيات كأس سمو<br>الأمير للكرة الطائرة<br>فاز فريقا الوكرة<br>والسد الخور<br>والشمال تواليا<br>منافسات الجولة<br>الأولى التصفيات<br>التمهيدية المؤهلة<br>لبطولة كأس سمو<br>الأمير المفدى<br>للموسم الرياضي<br>المقامة الفترة<br>الجاري وجاء فوز<br>الوكرة الخور بثلاثة<br>أشواط شوط بواقع<br>نتائج ليضع رصيده<br>نقاط | win Al-Wakrah Al-Sadd<br>qualifiers cup Highness<br>prince for volleyball won<br>teams Al-Wakrah Al-<br>Sadd Al-Khor Al-<br>Shamal respectively<br>competitions round first<br>qualifiers preliminary<br>qualifying championship<br>cup Highness prince<br>honorable season sports<br>held period current came<br>win Al-Wakrah Al-Khor<br>in three sets set results<br>place its account points | (0.2322, "فوز"),<br>(0.3965, "السد"),<br>(0.3114, "الوكرة"),<br>(0.1380, "تصفيات"),<br>(0.1541, "كأس"),<br>(0.2602, "سمو"),<br>(0.2134, "الأمير"),<br>(0.1462, "الكرة"),<br>(0.1124, "الطائرة"),<br>(0.1182, "فاز"),<br>(0.1344, "فريقا"),<br>(0.2755, "السد"),<br>(0.1694, "الشمال"),<br>(0.1699, "تواليا"),<br>(0.1052, "منافسات"),<br>(0.1011, "الجولة"),<br>(0.0707, "الأولى"),<br>(0.0893, "التصفيات"),<br>(0.1722, "التمهيدية"),<br>(0.1330, "المؤهلة"),<br>(0.1205, "البطولة"),<br>(0.1297, "المفدى"),<br>(0.1261, "للموسم"),<br>(0.1148, "الرياضي"),<br>(0.0837, "المقامة"),<br>(0.1349, "الفترة"),<br>(0.0793, "الجاري"),<br>(0.1569, "جاء"),<br>(0.1183, "فوز"),<br>(0.1309, "بثلاثة"),<br>(0.0927, "أشواط"),<br>(0.1563, "شوط"),<br>(0.1142, "بواقع"),<br>(0.1832, "نتائج"),<br>(0.1301, "ليضع"),<br>(0.1027, "رصيده"),<br>(0.1301, "نقاط") | [("win", 0.2322), ("Al-<br>Wakrah", 0.3965), ("Al-<br>Sadd", 0.3114),<br>("qualifiers", 0.1380),<br>("cup", 0.1541),<br>("Highness", 0.2602),<br>("prince", 0.2134), ("for<br>volleyball", 0.1462),<br>("volleyball", 0.1124),<br>("won", 0.1182),<br>("teams", 0.1344), ("Al-<br>Khor", 0.2755),<br>("Al-Shamal", 0.1694),<br>("respectively", 0.1699),<br>("competitions",<br>0.1052),<br>("round", 0.1011),<br>("first", 0.0707),<br>("qualifiers", 0.0893),<br>("preliminary", 0.1722),<br>("qualifying", 0.1330),<br>("championship",<br>0.1205),<br>("honorable", 0.1297),<br>("season", 0.1261),<br>("sports", 0.1148),<br>("held", 0.1349),<br>("period", 0.0793),<br>("current", 0.0837),<br>("came", 0.1183), ("in<br>three", 0.1309), ("sets",<br>0.1569),<br>("set", 0.1563), ("with<br>results", 0.1142),<br>("results", 0.0927),<br>("place", 0.1832), ("its<br>account", 0.1301),<br>("points", 0.1027)] |

Before presenting Figure 3, it is important to understand how the SVM model determines the relevance of specific words in classifying text under the sports category. The SVM model assigns different coefficients to each word based on its contribution to the classification. Words with positive coefficients strongly support the classification of text into the sports category, while words with negative coefficients suggest a weaker or no association, potentially aligning the text with other categories.

Figure 3 demonstrates the model's identification of relevant words within the semantic field of sports. A total of 33 keywords from the original 71 words are highlighted, with each word contributing either positively or negatively to the final classification. Positive word weights, such as 'volleyball' and 'sports', indicate a strong connection to the sports category, while negative word weights suggest minimal relevance to sports.

Figure 3 illustrates how individual word coefficients influence the SVM model's classification process for Arabic news texts. Detailed information on specific word coefficients and their impact can be found in Tables 6 and 7.



**Figure 3.** Sports keyword coefficients with SVM classification.

Table 6 lists the words with positive coefficients, meaning these words strongly support the classification of the text as related to sports. Words such as الطائرة 'volleyball' and الرياضي 'sports' have the highest positive weights, indicating that their presence in the text makes it more likely to be categorized in the sports domain. Other words like التمهيدية 'preliminary', تصفيات 'qualifiers', and فاز 'victory' are also key indicators that the text is sports-related. The total weight for all positive words is 1.9669, reflecting the contribution of these terms toward reinforcing the sports classification.

**Table 6.** Positive weighted semantic field.

| No.          | Word      | Meaning      | Weight                   |
|--------------|-----------|--------------|--------------------------|
| 1            | التمهيدية | Preliminary  | 0.0745                   |
| 2            | الجاري    | Current      | 0.1675                   |
| 3            | الرياضي   | Sports       | 0.2063                   |
| 4            | الطائرة   | Volleyball   | 0.4208                   |
| 5            | المقامة   | Held         | 0.2452                   |
| 6            | بثلاثة    | By three     | 0.1421                   |
| 7            | تصفيات    | Qualifiers   | 0.1374                   |
| 8            | فاز       | Won          | 0.0710                   |
| 9            | فريقا     | Teams        | 0.0414                   |
| 10           | منافسات   | Competitions | 0.1595                   |
| 11           | نتائج     | Results      | 0.1915                   |
| 12           | والسد     | And Al-Sadd  | 0.0582                   |
| 13           | وجاء      | And came     | 0.0517                   |
| <b>Total</b> |           |              | <b>1.966930996802351</b> |

Table 7 presents words that have negative coefficients, suggesting that these words oppose the classification of the text as sports-related. For instance, words like الأمير ‘prince’, الفترة ‘period’, and بواقع ‘by a score of’ have the highest negative weights. These terms are more commonly associated with non-sports-related contexts, which decreases the likelihood that a text containing them will be classified under sports. The total negative weight is -2.9408, which highlights how much these terms detract from the sports classification.

**Table 7.** Negative weighted semantic field.

| No.          | Word    | Meaning                   | Weight                   |
|--------------|---------|---------------------------|--------------------------|
| 1            | الأمير  | The prince                | -0.4089                  |
| 2            | الأولى  | The first                 | -0.0653                  |
| 3            | الجولة  | The round                 | -0.0464                  |
| 4            | الخور   | Al-Khor (team name)       | -0.2463                  |
| 5            | الفترة  | The period                | -0.3073                  |
| 6            | المؤهلة | Qualifying                | -0.0054                  |
| 7            | بواقع   | By a score of             | -1.0761                  |
| 8            | رصيده   | His score/balance         | -0.0802                  |
| 9            | سمو     | Highness                  | -0.1605                  |
| 10           | فوز     | Victory                   | -0.0111                  |
| 11           | كأس     | Cup                       | -0.0978                  |
| 12           | لبطولة  | For the tournament        | -0.0061                  |
| 13           | للموسم  | For the season            | -0.3474                  |
| 14           | نقاط    | Points                    | -0.0205                  |
| 15           | والشمال | And Al-Shamal (team name) | -0.0616                  |
| <b>Total</b> |         |                           | <b>-2.94081187554341</b> |

Table 8 shows the probability of the text being classified into different categories. The رياضة ‘sports’ category has the highest probability at 99.9967%, strongly confirming the text as sports-related. Other categories, such as علم ‘science’ and سياحة ‘travel’, have much lower probabilities, despite having relatively high scores in comparison to non-sports categories. The سياسة ‘politics’ category has the lowest probability at 42.87%, indicating that it is the least likely category for this text.

**Table 8.** Classification probability of each category.

| No. | Category  | Meaning       | Score (%)         |
|-----|-----------|---------------|-------------------|
| 1   | رياضة     | Sports        | 99.99667356597296 |
| 2   | علم       | Science       | 99.97335216579244 |
| 3   | سياحة     | Travel        | 99.97325857961826 |
| 4   | مجتمع     | Society       | 99.9238881466324  |
| 5   | دوليات    | International | 99.92238308126784 |
| 6   | صحة       | Health        | 99.1689863875924  |
| 7   | ثقافة     | Culture       | 97.75598450119771 |
| 8   | تكنولوجيا | Technology    | 85.21249488310567 |
| 9   | أمن       | Security      | 85.09816653644927 |
| 10  | اقتصاد    | Economy       | 85.01391919583232 |
| 11  | سياسة     | Politics      | 42.87280693212932 |

The process of semantic field extraction, traditionally performed by humans, differs significantly from the method employed by computers to determine the most suitable category, as is conventionally done by humans. Humans can swiftly and accurately classify text categories with a single glance, akin to utilizing the skimming

technique (Meidina & Kasmawati, 2022). By discerning the semantic field within a text, without reading the entire passage, the text category can be promptly identified. However, this process is less straightforward for computers, which require specific steps to mimic human thought processes and ultimately generate human-like understanding. The NLP method is essential for computers to emulate human cognition (Sulistyo et al., 2021). These steps typically involve case folding, tokenization, and the removal of stopwords, as utilized in this study (Bourequat & Mourad, 2021). Fundamentally, this method represents a collaboration between linguistic, computer science, and artificial intelligence disciplines.

As illustrated in the preceding example text, humans intuitively categorize the word كأس 'cup' within the semantic field of sports, as sports tournaments commonly award cups as a form of recognition to top-performing teams or individuals. However, a computer may hold a different assumption. While the computer recognizes the word as part of the sports-related semantic field, it may assign a negative weight, potentially hindering the text classification as a sports-related article. The computer's interpretation of كأس may extend beyond sports contexts, considering it in broader competition or awards settings. This interpretation is often influenced by the training data used. If the training data includes instances of كأس appearing in non-sports-related contexts or alongside words unrelated to sports, the SVM model may adjust its weights accordingly based on the information gleaned from the training data.

## 5. DISCUSSION

The present study successfully extracted semantic fields from Arabic online news using various NLP techniques and machine learning models. The findings contribute to a deeper understanding of Arabic NLP applications and highlight several important points in the context of existing literature. The effectiveness of the SVM model, which achieved the highest overall accuracy of 90%, aligns with previous research on the utility of machine learning models for tasks involving large volumes of text data (Bourahouat et al., 2023; Ghamdi et al., 2023).

Moreover, the study builds upon previous research showing how major events influence semantic fields in specific domains. For example, Syafroni (2021) demonstrated how the COVID-19 pandemic shaped educational terminology, providing a broader framework for understanding how large-scale events impact language. Similarly, Yuliana (2021) and Sutarsih (2020) illustrated how cultural and linguistic contexts contribute to the shaping of semantic fields, supporting the importance of context-specific models in NLP applications like news categorization.

The challenges faced in categorizing health and science news in our research, where the SVM achieved a lower F1-score of 79%, are consistent with the complexities identified by Elarnaoty (2012) in opinion extraction. The richness of meaning and specialized terminology in these fields, as noted by Hilmi et al. (2022), requires more refined techniques and potentially hybrid models to improve classification accuracy. This supports the findings of Patmasada et al. (2020) and Sumarna et al. (2020), who emphasized the influence of linguistic and cultural factors on semantic fields.

Additionally, this study aligns with recent advancements in Arabic NLP, particularly the development of language-specific models, such as those outlined by

Ghamdi et al. (2023). The application of SVM for extracting semantic fields further supports the findings of Bourahouat et al. (2023), who reviewed Arabic NLP techniques, highlighting the effectiveness of tailored models for complex linguistic tasks. Moreover, the creation of datasets specific to Arabic dialects, such as the one introduced by Moussa and Mourhir (2023), underpins the need for continuous refinement of NLP models to handle the rich diversity of the Arabic language.

Furthermore, our study contributes to the literature on computational linguistics by demonstrating the applicability of NLP techniques in a variety of fields, including health, education, and social media analysis, as noted by Giarnieri and Scardapane (2023) and Shin et al. (2023). By employing pre-processing techniques like case folding and tokenization, this study affirms the importance of text normalization, aligning with the findings of Müller et al. (2023) in public health analysis.

The success of the SVM in this study highlights its potential as a robust tool for news classification and semantic field extraction in Arabic. This has practical implications for news agencies and information retrieval systems, which can utilize these findings to enhance content management and curation. Theoretically, this research supports the growing body of knowledge in computational linguistics and Arabic NLP, particularly in understanding the cultural and contextual nuances of semantic fields (Gärdenfors, 2017).

Finally, the study emphasizes the need for future research into hybrid models that combine SVM with other machine learning techniques to enhance performance across all categories. This recommendation is in line with Rahma et al. (2023), who found that hybrid approaches improve classification in complex domains like sarcasm detection. Expanding the dataset and incorporating more domain-specific data would further enhance the model's accuracy, particularly in categories like health and science.

## 6. CONCLUSION

This study has successfully demonstrated the efficacy of using NLP techniques and machine learning models for semantic field extraction and classification of Arabic online news articles. The support vector machine (SVM) model, which achieved the highest overall accuracy of 90%, proved to be a reliable tool for this purpose. Its strong performance in categorizing sports news highlights the model's ability to handle well-defined semantic fields, contributing valuable insights to Arabic NLP. The findings draw attention to the importance of advanced computational techniques in processing and understanding Arabic text, enhancing the accuracy of news classification systems while contributing to the broader field of Arabic NLP applications. This research supports the growing body of work that advocates for developing sophisticated NLP tools customized to non-Latin script languages, offering significant advancements in the field.

However, the study faced limitations, particularly in the categorization of health and science news, where the SVM model achieved a lower F1 score, pointing to the complexity of these domains and the need for more specialized models. The study was also constrained by the size and diversity of the dataset, which may affect the generalizability of the findings. Future research should address these limitations by exploring hybrid models that combine SVM with other machine-learning techniques

to improve performance across all news categories. Expanding the dataset and including a more diverse range of news articles would also enhance the robustness of the models. Additionally, future studies should investigate domain-specific NLP techniques to handle the linguistic intricacies of specialized fields such as health and science more effectively. By addressing these limitations and recommendations, the accuracy and efficiency of Arabic NLP applications can be further improved, paving the way for more advanced text analysis tools.

## REFERENCES

- Abdelfattah, M. F., Fakhr, M. W., & Rizka, M. A. (2023). ArSentBERT: fine-tuned bidirectional encoder representations from transformers model for Arabic sentiment classification. *Bulletin of Electrical Engineering and Informatics*, 12(2), 1196–1202. <https://doi.org/10.11591/eei.v12i2.3914>
- Al-Harbi, O., Hamed, A., & Alzoubi, M. (2023). A deep neural network optimized by a genetic algorithm to improve Arabic sentiment classification. *Ingénierie Des Systèmes d'Information*, 28(1), 67–75. <https://doi.org/10.18280/isi.280107>
- Allihei, F. M., Omar, A., & Al-Horais, N. (2021). Opinion mining of Saudi responses to COVID-19 vaccines on Twitter. *International Journal of Advanced Computer Science and Applications*, 12(6), 72–78. <https://doi.org/10.14569/IJACSA.2021.0120610>
- Alruily, M., Ayesh, A., & Zedan, H. (2014). Crime profiling for the Arabic language using computational linguistic techniques. *Information Processing & Management*, 50(2), 315–341. <https://doi.org/10.1016/j.ipm.2013.09.001>
- Al-Sarayreh, S., Mohamed, A., & Shaalan, K. (2023). Challenges and solutions for Arabic Natural Language Processing in social media. In A. E. Hassanien, D. Zheng, Z. Zhao, & Z. Fan (Eds.), *Business intelligence and information technology* (Vol. 358, pp. 293–302). Singapore: Springer. [https://doi.org/10.1007/978-981-99-3416-4\\_24](https://doi.org/10.1007/978-981-99-3416-4_24)
- Al-Sharq. (2024, March 25). *Fawz al-Wakra wal-Sadd fi Tasfiyat Ka's Amir lil-Kurah al-Tayyarah [Al-Wakrah and Al-Sadd win in the qualifiers for the Emir's Cup for Volleyball]*. <https://al-sharq.com/article/25/03/2024/تصفيات-في-والسد-الوكرة-فوز-الطائرة-للكرة-الأمير-سمو-كأس>
- Altamimi, M., & Alayba, A. M. (2023). ANAD: Arabic news article dataset. *Data in Brief*, 50(1), 1–7. <https://doi.org/10.1016/j.dib.2023.109460>
- Aoumeur, N. E., Li, Z., & Alshari, E. M. (2023). Improving the polarity of text through word2vec embedding for primary classical Arabic sentiment analysis. *Neural Processing Letters*, 55(3), 2249–2264. <https://doi.org/10.1007/s11063-022-11111-1>
- Arianto, A. K. (2018). Medan makna pembentuk metafora dalam syair Arab [The semantic field forming metaphors in Arabic poetry]. *Widyaparwa*, 46(2), 112–125.
- Arkhangelskaya, E. O., & Nikolenko, S. I. (2023). Deep learning for natural language processing: A survey. *Journal of Mathematical Sciences*, 273(4), 533–582. <https://doi.org/10.1007/s10958-023-06519-6>
- Bashir, M. H., Azmi, A. M., Nawaz, H., Zaghouani, W., Diab, M., Al-Fuqaha, A., & Qadir, J. (2023). Author correction: Arabic natural language processing for

- Qur'anic research: A systematic review. *Artificial Intelligence Review*, 56(11), 13951–13952. <https://doi.org/10.1007/s10462-023-10390-x>
- Bekirov, R. A., & Rublyova, V. V. (2019). Comparative analysis of Arabic, Russian, and English idioms of semantic field “fear.” *Tyumen State University Herald. Humanities Research. Humanitates*, 5(3), 6–21. <https://doi.org/10.21684/2411-197X-2019-5-3-6-21>
- Berrimi, M., Oussalah, M., Moussaoui, A., & Saidi, M. (2023). Attention mechanism architecture for Arabic sentiment analysis. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(4), 1–26. <https://doi.org/10.1145/3578265>
- Bolshakov, I. A., & Gelbukh, A. (Eds.). (2004). *Computational linguistics: Models, resources, applications*. Mexico: Ciencia De La Computación.
- Boulrieris, P., Pavlopoulos, J., Xenos, A., & Vassalos, V. (2023). Fraud detection with natural language processing. *Machine Learning*, 1–22. <https://doi.org/10.1007/s10994-023-06354-5>
- Bourahouat, G., Abourezq, M., & Daoudi, N. (2023). Systematic review of the Arabic natural language processing: Challenges, techniques and new trends. *Journal of Theoretical and Applied Information Technology*, 101(3), 1333–1343.
- Bourequat, W., & Mourad, H. (2021). Sentiment analysis approach for analyzing iPhone release using support vector machine. *International Journal of Advances in Data and Information Systems*, 2(1). <https://doi.org/10.25008/ijadis.v2i1.1216>
- Çetinel, H. (2022). İnsan ve muhiti: Muhit kavramı perspektifinden din eğitimi [Human and environment: Religious education from the perspective of the concept of environment]. *Muhakeme Dergisi*, 5(2), 80–93. <https://doi.org/10.33817/muhakeme.1205853>
- Davani, A. M., Atari, M., Kennedy, B., & Dehghani, M. (2023). Hate speech classifiers learn normative social stereotypes. *Transactions of the Association for Computational Linguistics*, 11. [https://doi.org/10.1162/tacl\\_a\\_00550](https://doi.org/10.1162/tacl_a_00550)
- Denisenko, V. N., & Kalinina, N. S. (2020). Lexical means of verbalization of uncertainty in the Arabic language on the example of modern best sellers. *RUDN Journal of Language Studies, Semiotics and Semantics*, 11(1), 36–47. <https://doi.org/10.22363/2313-2299-2020-11-1-36-47>
- Drikvandi, R., & Lawal, O. (2023). Sparse principal component analysis for natural language processing. *Annals of Data Science*, 10(1), 25–41. <https://doi.org/10.1007/s40745-020-00277-x>
- Elarnaoty, M. (2012). A machine learning approach for opinion holder extraction in Arabic language. *International Journal of Artificial Intelligence & Applications*, 3(2), 45–62. <https://doi.org/10.5121/ijaia.2012.3205>
- Fatihah, N., Nurul Khomsati, N., Sahda Widyadhari, T., & Nur Aziz, K. (2023). Analisis konsep fisika osilasi terkopel pada sistem pegas bermassa [Analysis of the physics concept of coupled oscillations in a mass spring system]. *Jurnal Penelitian Fisika dan Terapannya (JUPITER)*, 5(1), 10–15. <https://doi.org/10.31851/jupiter.v5i1.11777>
- Fauziyah, N. (2023). Analisis medan makna dan komponensial pada nama flora unik [Analysis of the meaning and componential fields of unique floral names]. *Nuansa Indonesia: Jurnal Ilmu Bahasa, Sastra, dan Filologi*, 25(2), 241–253.

- Gärdenfors, P. (2017). Semantic knowledge, domains of meaning and conceptual spaces. In *Knowledge and Action* (Knowledge and Space, Vol. 9, pp. 203–219). Cham: Springer. [https://doi.org/10.1007/978-3-319-44588-5\\_12](https://doi.org/10.1007/978-3-319-44588-5_12)
- Georgakopoulos, T., & Polis, S. (2022). New avenues and challenges in semantic map research (with a case study in the semantic field of emotions). *Zeitschrift Fur Sprachwissenschaft*, 41(1), 1–30. <https://doi.org/10.1515/zfs-2021-2039>
- Ghamdi, S. A., Khalifa, H. A., & Salman, A. A. (2023). Fine-tuning BERT-based pre-trained models for Arabic dependency parsing. *Applied Sciences*, 13(7), 1–22. <https://doi.org/10.3390/app13074225>
- Giarnieri, E., & Scardapane, S. (2023). Towards artificial intelligence applications in next generation cytopathology. *Biomedicines*, 11(8), 1–11. <https://doi.org/10.3390/biomedicines11082225>
- Gordon, F. (2018). The significance and impact of the media in contemporary society. In *Children, young people and the press in a transitioning society* (pp. 17–46). London: Palgrave Macmillan UK. [https://doi.org/10.1057/978-1-137-60682-2\\_2](https://doi.org/10.1057/978-1-137-60682-2_2)
- Gürsakar, N., Çelik, S., & Özdemir, S. (2023). High-frequency words have higher frequencies in Turkish social sciences article. *Quality and Quantity*, 57(2), 1865–1887. <https://doi.org/10.1007/s11135-022-01444-3>
- Habeeb, A., Otair, M. A., Abualigah, L., Alsoud, A. R., Elminaam, D. S. A., Zitar, R. A., Ezugwu, A. E., & Jia, H. (2023). Arabic text classification using modified artificial bee colony algorithm for sentiment analysis: The case of Jordanian dialect. In L. Abualigah (Ed.), *Classification applications with deep learning and machine learning technologies. Studies in computational intelligence* (Vol. 1071, pp. 243–288). Springer, Cham. [https://doi.org/10.1007/978-3-031-17576-3\\_12](https://doi.org/10.1007/978-3-031-17576-3_12)
- Heaton, D., Clos, J., Nichele, E., Fischer, J. (2023). Critical reflections on three popular computational linguistic approaches to examine Twitter discourses. *PeerJ Computer Science*, 9:(e1211). <https://doi.org/10.7717/peerj-cs.1211>
- Hilmi, H. S., Panjaitan, I. P., Wahyuni, S., & Ahmadi, A. (2022). Medan makna ‘jatuh’ dalam bahasa Sasak dialek Ngeno-Ngene [The semantic field of “fall” in the Sasak language dialect of Ngeno-Ngene]. *Sirok Bastra*, 10(2), 151–162. <https://doi.org/10.37671/sb.v10i2.374>
- Imam, A. I., & Masama, A. A. (2023). Alfikr al-mu’jami ladā Asy-Shaikh Ibn Ishaq Al-Tawardi Al-Sukuti fī dau’i al-ḥuqūl ad-dalāliyyah, kitab al-Jumū’ wa al-mufarradāt wa al-muṣanāyāt namūzajan [The comprehensive thought of Sheikh Ibn Ishaq al-Tawardi al-Sukuti on the illumination of the indicative fields, the book of collective and plural nouns as a model]. *Tasambo Journal of Language, Literature, and Culture*, 2(5), 178–187. <https://doi.org/10.36349/tjllc.2023.v02i01.022>
- Irbah, H. D. (2022). Makna asosiatif dalam atologi puisi 𐄂 (gil) [Associative meaning in the poetry anthology 𐄂 (gil)]. *Journal of Culture, Arts, Literature, and Linguistik (CaLLs)*, 6(2), 221–237.
- Isnaeni, M., & Riana, D. R. (2022). Perhitungan dialektometri isolek Komering dan Lampung: Apakah keduanya bahasa yang berbeda? [Dialectometry calculation of the isolek Komering and Lampung: Are they two different languages?]. *Prosiding Konferensi Linguistik Tahunan Atma Jaya (KOLITA)*, 20(20), 224–231. <https://doi.org/10.25170/kolita.20.3799>

- Issa, I. (2023). Morphological complexity in Arabic spelling and its implication for cognitive processing. *Journal of Psycholinguistic Research*, 52(1), 331–357. <https://doi.org/10.1007/s10936-022-09896-2>
- Jamaled dyn, I., El ayachi, R., & Biniz, M. (2023). An improved approach to Arabic news classification based on hyperparameter tuning of machine learning algorithms. *Journal of Engineering Research (Kuwait)*, 11(2), 1–7. <https://doi.org/10.1016/j.jer.2023.100061>
- Khanjari, S. (2023). Īhām, or the technique of double meaning in literature: Its theory and practice in the twelfth century. *Iranian Studies*, 56(3), 457–469. <https://doi.org/10.1017/irn.2023.22>
- Khurana, D., Koli, A., Khatter, K., & Singh, S. (2023). Natural language processing: state of the art, current trends and challenges. *Multimedia Tools and Applications*, 82(3), 3713–3744. <https://doi.org/10.1007/s11042-022-13428-4>
- Kim, D., Lim, D., & Wiley, J. A. (2023). Narrative investment-risk disclosure & REIT investment. *Journal of Real Estate Finance and Economics*, 66(2), 542–567. <https://doi.org/10.1007/s11146-021-09819-3>
- Larabi-Marie-Sainte, S., Bin Alamir, M., & Alameer, A. (2023). Arabic text clustering using self-organizing maps and grey wolf optimization. *Applied Sciences (Switzerland)*, 13(18), 1–17. <https://doi.org/10.3390/app131810168>
- Mahmoud, M. S., & Negied, N. (2023). A novel deep-learning based approach for automatic diacritization of Arabic poems using sequence-to-sequence model. *International Journal of Advanced Computer Science and Applications*, 14(1), 42–46. <https://doi.org/10.14569/IJACSA.2023.0140105>
- Martin, G. E., Lee, M., Bicknell, K., Goodkind, A., Maltman, N., & Losh, M. (2023). A longitudinal investigation of pragmatic language across contexts in autism and related neurodevelopmental conditions. *Frontiers in Neurology*, 14(1), 1–14. <https://doi.org/10.3389/fneur.2023.1155691>
- Mehdivi, U. (2008). *Taulīd al-asmā' min al-juzūr as-sulāsiyyah aṣ-ṣaḥīḥah fī al-lughah al-'Arabiyyah - muqārabah lisāniyyah ḥāsūbiyyah [The formation of names from the correct trilateral roots in the Arabic language - a computational linguistic approach]* (A. G. A. A.- 'Azm, Ed.; Vol. 1). Wiḥdah 'Ulūm Al-Lughah Al-'Arabiyyah wa Al-Mu'jāmiyyāt.
- Meidina, T., & Kasmawati, S. (2022). Teknik skimming untuk meningkatkan kemampuan membaca pemahaman pada siswa tunagrahita ringan kelas VII di SLB Negeri 1 Kabupaten Gowa Provinsi Sulawesi Selatan [Skimming techniques to improve reading comprehension skills in mild intellectually disabled seventh-grade students at SLB Negeri 1 Gowa District, South Sulawesi Province]. *Vidya Karya*, 37(1). <https://doi.org/10.20527/jvk.v37i1.12870>
- Mhamed, M., Sutcliffe, R., Sun, X., Feng, J., & Retta, E. A. (2023). Arabic sentiment analysis using GCL-based architectures and a customized regularization function. *Engineering Science and Technology, an International Journal*, 43(1), 1–16. <https://doi.org/10.1016/j.jestch.2023.101433>
- Moussa, H. N., & Mourhir, A. (2023). DarNERcorp: An annotated named entity recognition dataset in the Moroccan dialect. *Data in Brief*, 48(1), 1–5. <https://doi.org/10.1016/j.dib.2023.109234>
- Müller, M., Salathé, M., & Kummervold, P. E. (2023). COVID-Twitter-BERT: A natural language processing model to analyse COVID-19 content on Twitter.

- Frontiers in Artificial Intelligence*, 6(1), 1–6. <https://doi.org/10.3389/frai.2023.1023281>
- Nerbonne, J. (2022). Natural language processing in computer-assisted language learning. In R. Mitkov (Ed.), *The oxford handbook of computational linguistics (Oxford handbooks)* (2nd ed., pp. 634–698). USA: Oxford University Press, Inc.
- Nikolina, N. A. (2023). Key semantic fields in the comedy “The Forest” by A. N. Ostrovsky’s (to the 200th anniversary of the birth). *Russian Language at School*, 84(2), 43–48. <https://doi.org/10.30515/0131-6141-2023-84-2-43-48>
- Nor, N. H. M. N., & Hussain, E. M. (2020). Konsep ruang dalam bahasa dan pemikiran menerusi pantun Melayu [The concept of space in language and thought through Malay pantun]. *Jurnal Melayu Sedunia*, 3(1), 384–414.
- Patmasada, V., Madeten, S., & Muzammil, A. (2020). Relasi makna verba dalam bahasa Melayu dialek Kapuas Hulu [The relational meaning of verbs in the Kapuas Hulu dialect of the Malay language]. *Jurnal Pendidikan Dan Pembelajaran Khatulistiwa (JPPK)*, 9(7), 1–8.
- Pauzi, Z., & Capiluppi, A. (2023). Applications of natural language processing in software traceability: A systematic mapping study. *Journal of Systems and Software*, 198(1), 1–19. <https://doi.org/10.1016/j.jss.2023.111616>
- Pellicer, L. F. A. O., Ferreira, T. M., & Costa, A. H. R. (2023). Data augmentation techniques in natural language processing. *Applied Soft Computing*, 132(1), 1–20. <https://doi.org/10.1016/j.asoc.2022.109803>
- Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In A. Moschitti, B. Pang, & W. Daelemans (Eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543). Association for Computational Linguistics. <https://doi.org/10.3115/v1/D14-1162>
- Rahma, A., Azab, S. S., & Mohammed, A. (2023). A comprehensive survey on Arabic sarcasm detection: approaches, challenges and future trends. *IEEE Access*, 11(1), 18261–18280. <https://doi.org/10.1109/ACCESS.2023.3247427>
- Ryding, K. C. (2014). Derivational morphology: The root/pattern system. In *Arabic A Linguistic Introduction* (pp. 55–78). UK: Cambridge University Press. <https://doi.org/10.1017/CBO9781139151016.007>
- Saiegh-Haddad, E., & Henkin-Roitfarb, R. (2014). The structure of Arabic language and orthography. In *Handbook of arabic Literacy* (Literacy Studies, Vol. 9, pp. 3–28). Dordrecht: Springer. [https://doi.org/10.1007/978-94-017-8545-7\\_1](https://doi.org/10.1007/978-94-017-8545-7_1)
- Serrano-Niza, D. (2018). El color rojo según el Kitāb al-Mujaṣṣaṣ de Ibn Sīdah. Una aportación a la lexicografía árabe [The color red according to the Kitāb al-Mujaṣṣaṣ by Ibn Sīdah. A contribution to Arabic lexicography]. *Al-Qanṭara*, 39(1), 101–126. <https://doi.org/10.3989/alqantara.2018.004>
- Shin, J., Balyan, R., Banawan, M. P., Arner, T., Leite, W. L., & McNamara, D. S. (2023). Pedagogical discourse markers in online algebra learning: Unraveling instructor’s communication using natural language processing. *Computers and Education*, 205. <https://doi.org/10.1016/j.compedu.2023.104897>
- Sukiman, U. (2015). Makna figuratif senjata dalam idiom bahasa Arab (kajian semantik) [The figurative meaning of weapons in Arabic idioms (a semantic study)]. *Adabiyāt: Jurnal Bahasa dan Sastra*, 14(2), 244–265. <https://doi.org/10.14421/ajbs.2015.14204>

- Sulistyo, D., Ahda, F., & Fitria, V. A. (2021). Epistemologi dalam natural language processing [Epistemology in natural language processing]. *Jurnal Inovasi Teknologi dan Edukasi Teknik*, 1(9). <https://doi.org/10.17977/um068v1i92021p652-664>
- Sumarna, R., Amir, A., & Syahrani, A. (2020). Medan makna aktivitas mengayam masyarakat Melayu Sambas [The field of meaning of the weaving activities of the Sambas Malay community]. *Jurnal Pendidikan dan Pembelajaran Khatulistiwa (JPPK)*, 19(12), 1–8. <http://ejournal.uin-suka.ac.id/adab/Adabiyat/article/view/14204>
- Sutarsih, S. (2020). Kata Bermakna Membesar dalam Bahasa Jawa [The Meaning Words of Enlarge in Javanese]. *Jalabahasa*, 13(2), 193–210. <https://doi.org/10.36567/jalabahasa.v13i2.93>
- Syafroni, R. N. (2021). Medan makna bidang pendidikan dalam masa pandemi Covid-19 [The field of meaning of education during the Covid-19 pandemic]. *Prosiding Seminar Nasional Linguistik Dan Sastra (Semnalisa)*, 224–227.
- Torregrosa, J., Bello-Orgaz, G., Martínez-Cámara, E., Ser, J. Del, & Camacho, D. (2023). A survey on extremism analysis using natural language processing: definitions, literature review, trends and challenges. *Journal of Ambient Intelligence and Humanized Computing*, 14(8), 9869–9905. <https://doi.org/10.1007/s12652-021-03658-z>
- Treviso, M., Lee, J.-U., Ji, T., Aken, B. van, Cao, Q., Ciosici, M. R., Hassid, M., Heafield, K., Hooker, S., Raffel, C., Martins, P. H., Martins, A. F. T., Forde, J. Z., Milder, P., Simpson, E., Slonim, N., Dodge, J., Strubell, E., Balasubramanian, N., ... Schwartz, R. (2023). Efficient methods for natural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 11(1), 826–860. [https://doi.org/10.1162/tacl\\_a\\_00577](https://doi.org/10.1162/tacl_a_00577)
- Wahdan, A., Al-Emran, M., & Shaalan, K. (2024). A systematic review of Arabic text classification: areas, applications, and future directions. *Soft Computing*, 28(2), 1545–1566. <https://doi.org/10.1007/s00500-023-08384-6>
- Wu, L., Chen, Y., Shen, K., Guo, X., Gao, H., Li, S., Pei, J., & Long, B. (2023). Graph neural networks for natural language processing: A survey. *Foundations and Trends in Machine Learning*, 16(2), 119–328. <https://doi.org/10.1561/22000000096>
- Yuliana, B. T. (2021). Medan makna kata “terjatuh” dalam bahasa Sasak dialek Meno-Meni: Kajian semantik [The field of meaning of the word “fall” in the Sasak language dialect of Meno-Meni: A semantic study]. *Anafora: Jurnal Penelitian Mahasiswa Pendidikan Bahasa Dan Sastra Indonesia*, 1(1), 12–21. <https://doi.org/10.25134/ajpm.v1i1.9>
- Zahrah, Z., & Octora, T. Y. (2023). Analisis konservasi energi pada dinding kaca ruang tunggu Bandara Ahmad Yani Semarang [Energy conservation analysis of the glass walls in the waiting room of Ahmad Yani Airport Semarang]. *Jurnal Teknik Transportasi*, 2(2), 115–127. <https://doi.org/10.54324/jtt.v2i2.29>
- Zhang, J., Hua, J., Niu, N., Chen, S., Savolainen, J., & Liu, C. (2023). Exploring privacy requirements gap between developers and end users. *Information and Software Technology*, 154. <https://doi.org/10.1016/j.infsof.2022.107090>