Functional Text Dimensions for Arabic Text Classification

Zeyd Ferhat,[1] **Abir Betka ,**[4] **Barka Riyadh,**[3] **Selma Boutiba ,**[2] **Zineddine S. Kahhoul,**[2]
Mohamed Lakhdar Tiar,[2] **Habiba Dahmani**[3]**, and Ahmed Abdelali,**[5]
[1] Laboratory of VCS, [2] Laboratory of IL3CUB, University of Biskra, Algeria
[3] Electrical Engineering Department, University of M'sila, Algeria
[4] Electrical Engineering Department, University of El-Oued, Algeria
[5] National Center for AI, SDAIA, Riyadh, KSA
{zeydferhatz, barkariyadh06}@gmail.com, abir-betka@univ-eloued.dz,
{selma.boutiba, zineddine.kahhoul, mohamedlakhdar.tiar}@univ-biskra.dz,
habiba.dahmani@univ-msila.dz, aabdelali@sdaia.gov.sa

Abstract

Text classification is of paramount importance in a wide range of applications, including information retrieval, extraction and sentiment analysis. The challenge of classifying and labelling text genres, especially in web-based corpora, has received considerable attention. The frequent absence of unambiguous genre information complicates the identification of text types. To address these issues, the Functional Text Dimensions (FTD) method has been introduced to provide a universal set of categories for text classification. This study presents the Arabic Functional Text Dimensions Corpus (AFTD Corpus), a carefully curated collection of documents for evaluating text classification in Arabic. The AFTD Corpus which we are making available to the community[1], consists of 3400 documents spanning 17 different class categories. Through a comprehensive evaluation using traditional machine learning and neural models, we assess the effectiveness of the FTD approach in the Arabic context. CAMeL-BERT, a state-of-the-art model, achieved an impressive F1 score of 0.81 on our corpus. This research highlights the potential of the FTD method for improving text classification, especially for Arabic content, and underlines the importance of robust classification models in web applications.

1 Introduction

Text classification is an essential task in Natural Language Processing (NLP) with a wide range of applications. It involves categorizing textual data into predefined labels, enabling efficient information retrieval, content filtering, sentiment analysis, media monitoring, and even critical applications such as fraud detection. However, the classification and labelling of text genres, especially in web-based corpora, presents several challenges that hinder these processes.

A prominent problem is the frequent lack of clear and explicit genre information in web corpora. Genre labels, when present, tend to be ambiguous and open to interpretation, making it difficult to reliably identify text types. Traditional corpora often have their own classification schemes, which may not be consistent with the categories commonly used in web-based texts. This inconsistency poses a challenge when integrating different sources of information.

Furthermore, the interpretation of genre labels can be subjective, leading to discrepancies between different annotators. The same text may be categorized differently by two people, leading to inconsistencies in large-scale annotation tasks. In addition, texts within the same category can have significant variations in structure, language and content, further complicating the task of genre classification.

While genre classification plays a central role in understanding linguistic features and facilitating corpus comparisons, the lack of standardized guidelines and reliable annotation practices hinders progress. Web-derived corpora often rely on the origin of the content or limited contextual clues to infer genre information, which can be unreliable and incomplete (Sharoff, 2018). This highlights the complex nature of genre classification, especially when dealing with diverse and dynamic web content.

The Functional Text Dimensions (FTD) method has been proposed to address these challenges (Sharoff, 2018; Sobirova and Cho, 2023). FTDs provide a set of universally applicable categories that offer a broad and versatile classification system. By providing a comprehensive set of guidelines, the FTD approach aims to improve the reliability and consistency of annotation practices. The inter-annotator agreement, which refers to the degree of agreement between different annotators

[1] https://huggingface.co/datasets/zeydferhat/functional_text_dimensions_for_arabic_text_classification

Proceedings of The Second Arabic Natural Language Processing Conference, pages 352–360
August 16, 2024 ©2024 Association for Computational Linguistics

for a randomly selected 350 documents per annotator reached a Krippendorff's (Krippendorff, 2006) agreement above 70% (Sharoff, 2018) . The reported agreement results suggest that the proposed categories are effective in practice. However, the existing FTD frameworks have been primarily developed and evaluated primarily with English and a few other widely used languages.

As one of the most widely spoken languages in the world, with a rich history and diverse dialects, Arabic presents unique challenges for text classification. The complexity of the Arabic language, including its varied grammatical structures, extensive vocabulary and distinct writing system, requires the development of specialised language resources and models. Although Arabic has significant differences from English and other European languages, there is a lack of comprehensive Arabic text corpora that can be used to effectively evaluate text classification models.

To fill this gap, we present the Arabic Functional Text Dimensions Corpus (AFTD Corpus) - a carefully curated collection of Arabic documents. The AFTD Corpus is designed to ensure diversity and representativeness, covering different genres and styles relevant to modern Arabic literature. By extending the FTD approach to Arabic, we aim to improve text classification performance and enable a wide range of applications in the Arabic language.

The contribution of this work is double. First, we present the AFTD corpus, a valuable resource for Arabic text classification research and development that will be open and available to the community. Second, we perform a comprehensive evaluation of state-of-the-art text classification models using the AFTD corpus, providing a benchmark for future comparisons. Through this study, we aim to demonstrate the effectiveness of the FTD approach in the Arabic context and to emphasise the importance of language-specific resources for accurate and reliable text classification.

2 Related Work

Text classification has historically been driven by the volume of content that has become available and accessible to the community over the years. The process has gone through various stages, from early manual systems to more modern digital methods.

In the late 19th century, M. Dewey (Dewey, 2004) proposed the Dewey Decimal Classification (DDC) a system that was one of the earliest attempts to classify books based on their subject matter. It used a numerical system to categorize and organize books in libraries, making it easier for librarians and users to find specific materials.

Library of Congress Classification (LCC): Developed by the Library of Congress, this system expanded on Dewey's work by providing a more comprehensive classification scheme. LCC is widely used in academic and research libraries, particularly in the United States (Dittmann et al., 2007).

With the advent of computers and the Internet, libraries moved from card catalogues to online databases. This allowed for more dynamic and flexible searching and browsing of library collections. The advent of digital libraries enabled libraries to provide access to digital content alongside physical collections. This shift required new classification systems, metadata standards and search algorithms. As the web expanded, the need to classify web content grew, giving rise to the semantic web and knowledge organisation systems. These systems aim to structure and link information on the Web in a more meaningful way, using technologies such as Resource Description Framework (RDF), Ontology Web Language (OWL) and linked data (Parker et al., 2011). Recent advances in machine learning and artificial intelligence have led to more automated methods of classifying and categorizing texts. Natural language processing (NLP) techniques are used to analyze and classify large amounts of digital content quickly and efficiently (Bishop, 2006).

Modern systems also take into account user-generated tags, reviews, and recommendations, allowing users to contribute to the classification and organization of texts and web content through social tagging and collaborative filtering (Fukumoto and Suzuki, 2002; Suchomel and Kraus, 2022). Platforms such as Goodreads, for example, rely on user reviews and tags to classify books. This improves discoverability and personalization.

Crowdsourcing and collaborative tagging systems harness the collective intelligence of users (Ramírez et al., 2019), improving classification accuracy and relevance. These systems adapt to changing user preferences and emerging trends, providing a dynamic and responsive approach to text classification. While such an approach is very efficient, its consistency and uniformity brings more challenges to data processing (Yang et al., 2021).

Arabic text classification efforts have gained momentum in recent years due to the growing amount of digital content in the Arabic language. Arabic presents unique challenges for text classification due to its complex morphology, rich inflection and multiple dialects. Early work adopted existing approaches for English and other common languages. The adoption of such approaches resulted in collections such as CLARA (Corpus Linguae Arabicae) (Zemánek, 2001), the collection of both journals and books. Later work expanded both the scope and the variety of genres (Al-sulaiti and Atwell, 2003; Alansary et al., 2007).

Efforts to build large-scale Arabic corpora and datasets have been crucial to the advancement of Arabic text classification. The Arabic Gigaword corpus (Parker et al., 2011), for example, provides an extensive collection of newswire texts labelled by the source and type of news. In addition, initiatives such as the Open Source Arabic Corpora (OSAC) (Saad and Ashour, 2010) and SANAD (Einea et al., 2019) aim to make Arabic linguistic resources more accessible to researchers and developers. Moreover, community driven projects and competitions [2], such as the Arabic Sentiment Analysis Challenge [3], have simulated innovation by providing benchmarks and encouraging collaboration among researchers. These initiatives have contributed to the development of more robust and accurate Arabic text classification systems.

3 Data and Methodology

3.1 Arabic Functional Text Dimensions Corpus (AFTD Corpus)

The Arabic corpus used in this study was carefully selected from the web to ensure its diversity and relevance to our research objectives. We conducted manual search of various websites, articles and journals from a variety of sources published in Arabic. These web searches involved translating the names and test questions of each Functional Text Dimension (FTD) listed in Table 1 into Arabic (see Table 2). Each Functional Text Dimension (FTD) category is precisely defined and characterized by its corresponding test question. For clarity of presentation, FTDs are given codes (A1, A8, etc.) and concise labels (argum, hardnews, etc.). The test question not only serves as a descriptor, but also identifies prototypical genres associated with the FTD. Consequently, the assessment of the value of a text on a given FTD takes into account its similarity to the prototypical genre specified in relation to the test question. The collective set of FTDs provides a multidimensional representation that positions each text as a point within this space (Sharoff, 2018, 2011).

For this task, eight (8) annotators—five male and three female—all of whom have at least some university education were recruited. The annotators were provided with detailed guidelines (both Arabic and English versions) and tasked with collecting relevant texts from the web. Using a shared repository, we moved on to the next category after collectively reaching the target number of texts in each category (200 documents per category). This systematic approach ensured a structured collection process aimed at achieving a balanced and comprehensive dataset across different categories.

During the text collection procedure, we retained additional meta information such as the website address, collection data of the text and the country of the source. Initially, we opted to gather 200 texts, each containing at least one paragraph illustrating at least one idea relevant to the context. For a more comprehensive descriptive analysis, we have now accumulated a total of 3,400 texts, encompassing 805,028 words in total.

In particular, Figures 2 and 3 illustrate the distribution of words across different dimensions. These figures provide insightful details about the corpus, highlighting an imbalance in the distribution of word counts across different Functional Text Dimensions (FTDs). This imbalance indicates a potential bias in the representation of certain categories, which could affect the overall analysis and interpretation of the data.

3.2 FTD Classification Experimentations

3.2.1 Data Preprocessing

Before starting the training phase, the text was tokenized using the BertTokenizer from the Hugging Face Transformers library [4]. This step ensured that the tokenization took into account the addition of padding tokens, truncation of sequences and that the maximum length of the tokenized sequence did not exceed 512 tokens. These considerations are important when preparing text data for models, particularly those with fixed input size requirements

[2] https://sina.birzeit.edu/nlu_sharedtask2024/
[3] https://kaust.link/3Nqs
[4] https://huggingface.co/

Code	Name	Description
A1	Argumentative	To what extent does the text argue to persuade the reader to support an opinion or a point of view?
A4	Fictive	To what extent is the text's content fictional?
A7	Instruct	To what extent does the text aim at teaching the reader how something works or at giving advice?
A8	Reporting	To what extent does the text appear to be an informative report of events recent at the time of writing?
A9	Legal	To what extent does the text specify a set of regulations?
A11	Personal	To what extent does the text report a first-person story?
A12	Commercial	To what extent does the text promote a product or service?
A14	Academic	To what extent does the text report academic research?
A16	Info	To what extent does the text provide reference information to define the topic of this text?
A17	Reviews	To what extent does the text evaluate a specific entity by endorsing or criticizing it?
A20	Appell	To what extent does the text request an action from the reader?
A21	Report	To what extent does the text provide a report about an event or a situation?
A3	Emotive	To what extent is the text concerned with expressing feelings or emotions?
A5	Flippant	To what extent is the text light-hearted, i.e. aimed mainly at amusing or entertaining the reader?
A15	Specialist	To what extent does the text require background knowledge or access to a reference source of a specialized subject area in order to be comprehensible?
A19	Poetic	To what extent does the author of the text pay attention to its aesthetic appearance?
A13	Propaganda	To what extent is the text intended to promote a political movement, party, religious faith, or other non-commercial cause?

Table 1: Function Text Dimension annotation guideline questions (Sharoff, 2018).

or constraints. Each model's tokenizer was used to meet its unique tokenization requirements, reflecting a tailored approach to preprocessing in line with the characteristics of each model. In addition, the dataset was split into a training set and a validation set using a 70-30 split ratio, with a random state set to 42.

3.2.2 Traditional Machine Learning Method

While the use of some traditional machine learning approaches may be considered outdated and legacy, these methods still offer significant benefits. Their efficiency and speed make them valuable tools for quickly establishing baselines for benchmarking. In addition, they help to assess the validity and consistency of the collected dataset, ensuring that the data is reliable and accurately represents the target categories. We used a range of algorithms: Logistic regression, random forest, support vector machine (SVM: linear kernel), decision tree and gradient boosting [5]. To evaluate the performance of the methods, we use the standard measures, including accuracy and F1 score. The results of these models are shown in 3)

3.3 Fine-tuning BERT Models

To fine-tune the model, we decided to use several pre-trained BERT (Bidirectional Encoder Representations from Transformers) models. We started with the BERT multilingual base model (cased) (Devlin et al., 2018) and various Arabic BERT models. The BERT multilingual base model, developed by the Google AI research team, is a versatile language representation model designed to process text in multiple languages with exceptional accuracy. The "cased" version is capable of recognising nuanced letter-casing subtleties, making it particularly powerful for multilingual natural language

[5] scikit-learn https://scikit-learn.org/

الرمز	الاسم	الوصف
A1	جدلي	إلى أي مدى يحاول النص إقناع القارئ بدعم رأي أو وجهة نظر؟
A4	خيالي	إلى أي مدى يكون محتوى النص خيالي؟
A7	تعليمي	إلى أي مدى يهدف النص إلى تعليم القارئ كيفية عمل شيء ما أو تقديم نصيحة؟
A8	إخباري	إلى أي مدى يظهر النص كتقرير عن الأحداث الجارية؟
A9	قانوني	إلى أي مدى يحدد النص مجموعة من اللوائح؟
A11	شخصي	إلى أي مدى يروي النص قصة من منظور الشخص الأول؟
A12	تجاري	إلى أي مدى يروّج النص لمنتج أو خدمة؟
A14	أكاديمي	إلى أي مدى يقدّم النص تقريراً عن بحث أكاديمي؟
A16	معلوماتي	إلى أي مدى يقدم النص معلومات مرجعية لتعريف موضوع معين؟
A17	مراجعات	إلى أي مدى يقيّم النص كيانًا معينًا من خلال تأييده أو نقده؟
A20	نداء	إلى أي مدى يطلب النص من القارئ القيام بعمل معين؟
A21	تقرير	إلى أي مدى يقدم النص تقريراً عن حدث أو موقف معين؟
A3	عاطفي	إلى أي مدى يهتم النص بالتعبير عن المشاعر أو العواطف؟
A5	هزلي	إلى أي مدى يكون النص هزلي، أي يهدف بشكل رئيسي إلى تسلية القارئ؟
A15	متخصص	إلى أي مدى يتطلب النص معرفة خلفية أو الوصول إلى مصدر مرجعي في مجال متخصص ليكون مفهوماً؟
A19	شعري	إلى أي مدى يهتم مؤلف النص بمظهره الجمالي؟
A13	دعائي	إلى أي مدى يهدف النص إلى الترويج لحركة سياسية أو حزب أو دين أو قضية غير تجارية؟

Table 2: Function Text Dimension annotation guideline questions translated into Arabic.

Model	Rec.	Pre.	Acc.	F1
Decision Tree	46.25	47.53	47.06	46.47
Gradient Boosting	64.30	68.40	64.80	64.97
Linear Kernel	68.46	69.27	68.82	68.02
Logistic Regression	73.29	73.37	73.43	**72.61**
Random Forest	60.04	66.66	60.29	59.37

Table 3: Machine learning results for AFTD collection assessment.

processing tasks. This model allows researchers and academics to work seamlessly with a wide range of languages without having to maintain separate models for each language.

We selected the Arabic models based on their suitability for the Arabic language and their effectiveness in similar classification tasks. Below is a brief description of each model used.

The selected BERT-based model was tuned with the following specific parameters: Batch size: 16 and number of training epochs: 3.

3.4 BERT models

The Multilingual BERT Base Model Cased (mBERT): It stands as a versatile language representation model, meticulously crafted by the Google AI research team (Devlin et al., 2018). This model is an important member of the extensive BERT model family, explicitly designed to provide a robust solution for handling text in multiple languages with remarkable accuracy. The "cased" variant of this model excels at recognising even the subtlest subtleties in letter casing, making it a powerful tool for those engaged in multilingual natural language processing tasks. Researchers and academics can use this model to work with a wide range of languages, eliminating the need to maintain separate models for each language.

AraBERTv0.2: Antoun et al. (Antoun et al., 2020) presented AraBERTv0.2, a carefully trained BERT base model that has been extensively pretrained on a huge corpus of 200 million sentences, amassing an impressive 77 GB of data. This corpus is an amalgamation of Modern Standard Arabic (MSA) and dialectal content, sourced primarily from Twitter data. Notable sources include Arabic Wikipedia dumps, the Arabic Corpus (El-Khair), the Open Source International Arabic News Corpus (OSIAN) (Zeroual et al., 2019), and a wealth of Arabic news content.

CAMeLBERT: Inoue et al. (Inoue et al., 2021) introduced CAMelBERT, a pre-trained language model with a unique blend of Modern Standard Arabic (MSA), Dialectal Arabic (DA) and Classical Arabic (CA). The dataset on which the model is based contains an impressive 167 GB of text data, equivalent to a staggering 17.3 billion tokens.

{ 'id': 248,
'label': 'argum',
'code': 'A1',
'text': مبادرة منير ورفاقه.. في الحركة بركة! \n'
11/8/2022 محمد عماد صابر
إبراهيم منير \n\nأثارت تصريحات نائب المرشد والقائم بالأعمال الأستاذ
إبراهيم منير الأخيرة التي تتحدث عن أننا لن نتصارع على السلطة في مصر عاصفة
من الانتقادات، بين أصوات المؤيدين وصراخ المحتجين والبيانات الهزلية لإثبات ...
'url': 'https://mubasher.aljazeera.net/opinions/2022/8/11/...',
'country': 'Qatar'
}

{ 'id': 559,
'label': 'Instruct',
'code': 'A7',
'text': تعلم الكتابة السريعة في خطوات \nتعلم الكتابة السريعة في خطوات \nأضحت الكتابة'
ضرورة لا بد منها في عالمنا اليوم، لا سيما أن معظم (touch typing) السريعة
'الأعمال والوظائف تحولت إلى الرقمية، ما يستدعي استخدام الأجهزة الإلكترونية، بدلًا ....
'url': 'https://blog.khamsat.com/touch-typing-guide/
'country': 'Inter.'
}

Figure 1: AFTD samples with additional metadata.

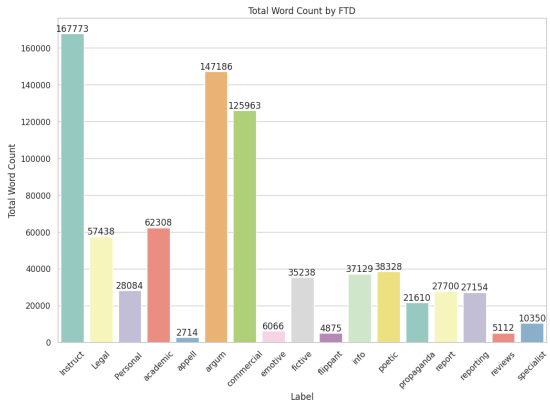

Figure 2: Total number of words per category.

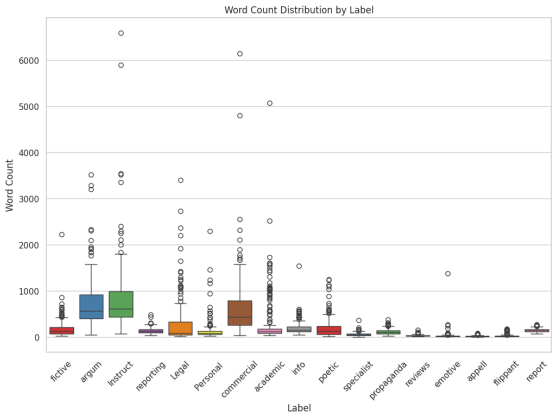

Figure 3: Distribution of word counts per document in each category.

QARiB: In their research, Abdelali et al. (Abdelali et al., 2021) trained QARiB model, which is characterized by its expanded dataset. This dataset includes 420 million tweets and 180 million text sentences, featuring a rich mixture of Modern Standard Arabic (MSA) and Dialectal Arabic (DA). Notably, the majority of the textual content is predominantly in MSA.

MARBERT: Abdul-Mageed et al. (Abdul-Mageed et al., 2021) presented MARBERT, a large-scale pre-trained masked language model with a specific focus on Dialectal Arabic (DA) and Modern Standard Arabic (MSA). The training data for this model comes from a staggering 1 billion Arabic tweets carefully selected from an extensive in-house dataset of around 6 billion tweets.

4 Results and discussion

4.1 Machine Learning Results

In the first part of our experiment, we applied different machine learning models to classify and label

Model	Loss	Acc.	F1
AraBERT	1.07	76.66	75.92
CAMeLBERT	0.67	81.37	**81.43**
MARBERT	0.79	81.68	80.62
mBERT	1.15	72.94	70.75
QARiB	0.70	80.49	80.48

Table 4: Results of BERT-based models for the assessment of AFTD outcomes.

genres in web-based corpora. The results are summarised in Table 1.

The machine learning results show that the performance of different classification models varies. Logistic regression achieved the highest accuracy at 73%, with a good balance between precision and recall. On the other hand, Decision Tree had the lowest accuracy at 47%, suggesting significant room for improvement. These results highlight the difficulty of genre classification in web-based corpora, and the significant variation in performance between models underlines the complexity of the task. Furthermore, the low accuracy of certain models, such as Decision Tree, suggests that the choice of algorithm is crucial when classifying genres in web-based content.

4.1.1 Fine-Tuning of BERT Results

In the second part of our investigation, we carried out fine-tuning of BERT-based models specifically tailored for Arabic text classification. The evaluation results are shown in Table 2.1, while the training results are shown in Table 2.2.

The fine-tuning of BERT models shows a more promising performance, especially for Arabic text classification. Models such as CAMeLBERT and MARBERT and QARiB achieved high accuracy, precision, recall and F1 scores, indicating their suitability for the task. The effect of choosing a language-specific pre-trained BERT model on performance is clearly observed, with CAMeLBERT and MARBERT and QARiB outperforming the generic mBERT model. Overall, our experiments highlight the complexity of genre classification in web-based corpora and the potential of native Arabic BERT-based models for such challenging tasks.

We perform a detailed analysis of the confusion matrix 4, for the best performing model, CAMeLBERT, which is crucial for evaluating the performance of our classification model. The confusion matrix allows us to examine how well the models' predictions match the true class labels, giving us a deeper understanding of accuracy and where the model may be making classification errors. The confusion matrix provides the following insights Academic: Our model showed an impressive ability to categorise instances under the (Academic) label, with only a negligible number of misclassifications. In particular, it was misidentified as (Specialist) 14 times and as (Info) 19 times, highlighting the common features between (Academic) and (Specialist) as well as (Info) that pose a challenge to the model. Fictitious: The model showed generally solid performance in classifying (Fictive), with only a slight misclassification of 25 instances as (Poetic). We can understand this confusion as the Fictive dimension contains poetic elements. Reporting: The (Reporting) category was occasionally misclassified as (Report) 13 times, indicating content similarities between these two categories. These results highlight the impressive accuracy of our model in classifying most categories, with only a few misclassifications detected. Investigating the reasons for these few misclassifications and further fine-tuning the model will be essential to improve overall performance. Strategies for improvement may include feature engineering and possibly the acquisition of additional labelled data to improve class separability.

5 Conclusion and Perspectives

In this paper, we presented the first Functional Text Dimensions for Arabic, a collection of 3,400 documents across 17 categories. We addressed the challenges of genre classification in web-based corpora, highlighting the complexities associated with genre labelling in Arabic. Our approach used natural language processing techniques, specifically machine learning and language models, to address these challenges.

Our research demonstrates the potential of neural models, particularly BERT-based models, to improve genre classification in web-based corpora, especially for the Arabic language. However, there is still room for improvement. We aim to expand the dataset and recruit professional annotators to improve the quality and reliability of the annotations. We will then perform a reliability analysis to ensure the integrity of the dataset.

In addition, fine-tuning the models and exploring feature engineering could significantly improve performance. The acquisition of additional labelled data would improve class separability and overall

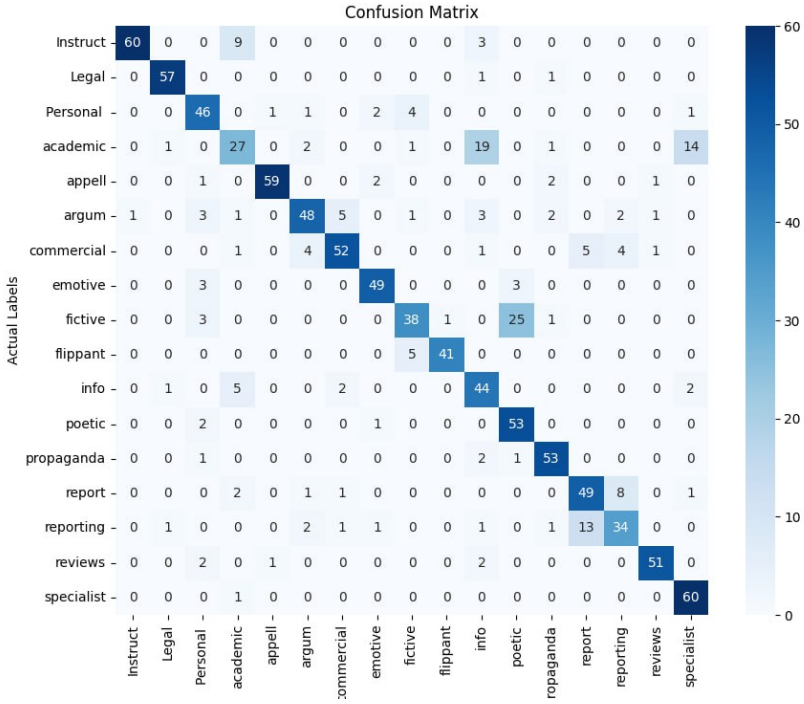

Figure 4: Confusion matrix for fine-tuned CAMeLBERT.

performance. With the emergence of Large Language Models (LLMs), we plan to explore their use both in extending the dataset and in assessing its quality, comparing their performance with traditional ML and BERT models.

References

Ahmed Abdelali, Sabit Hassan, Hamdy Mubarak, Kareem Darwish, and Younes Samih. 2021. Pre-training bert on arabic tweets: Practical considerations. *CoRR*, abs/2102.10684.

Muhammad Abdul-Mageed, AbdelRahim A. Elmadany, and El Moatez Billah Nagoudi. 2021. Arbert & marbert: Deep bidirectional transformers for arabic. In *ACL/IJCNLP (1)*, pages 7088–7105. Association for Computational Linguistics.

Latifa Al-sulaiti and Eric Atwell. 2003. The design of a corpus of contemporary arabic. In *University of Leeds School of Computing Research Series*.

Sameh Alansary, Magdy H. Nagi, and Noha Adly. 2007. Building an international corpus of arabic (ica) : Progress of compilation stage. In *7th international conference on language engineering, Cairo, Egypt*.

Wissam Antoun, Fady Baly, and Hazem M. Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. *CoRR*, abs/2003.00104.

Christopher M Bishop. 2006. *Pattern Recognition and Machine Learning*, volume 4 of *Information science and statistics*. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. Cite arxiv:1810.04805.

Melvil Dewey. 2004. *A Classification and Subject Index for Cataloguing and Arranging the Books and Pamphlets of a Library*. Project Gutenberg.

H. Dittmann, J. Hardy, and L. Musgrave. 2007. *Learn Library of Congress Classification*. Library education series. TotalRecall Publications.

Omar Einea, Ashraf Elnagar, and Ridhwan Al Debsi. 2019. Sanad: Single-label arabic news articles dataset for automatic text categorization. *Data in Brief*, 25:104076.

Fumiyo Fukumoto and Yoshimi Suzuki. 2002. Manipulating large corpora for text classification. In *EMNLP*, pages 196–203.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in arabic pre-trained language models. *CoRR*, abs/2103.06678.

Klaus Krippendorff. 2006. Reliability in Content Analysis: Some Common Misconceptions and Recommendations. *Human Communication Research*, 30(3):411–433.

Robert Parker, David Graff, Ke Chen, Junbo Kong, and Kazuaki Maeda. 2011. Arabic gigaword fifth edition ldc2011t11. Philadelphia: Linguistic Data Consortium. Web Download.

Jorge Ramírez, Marcos Baez, Fabio Casati, and Boualem Benatallah. 2019. Crowdsourced dataset to study the generation and impact of text highlighting in classification tasks. *BMC Research Notes*, 12(820).

Motaz Saad and Wesam Ashour. 2010. Osac: Open source arabic corpora. In *EEECS'10 the 6th International Symposium on Electrical and Electronics Engineering and Computer Science*, pages 118–123. European University of Lefke, Cyprus.

Serge Sharoff. 2011. In the garden and in the jungle: Comparing genres in the bnc and internet. *Genres on the web: Computational models and empirical studies*, pages 149–166.

Serge Sharoff. 2018. Functional text dimensions for the annotation of web corpora. *Corpora*, 13(1):65–95.

Firuza Urunbayevna. Sobirova and Young Soon Cho. 2023. The Essence Of Multidimensional Analysis. *European Scholar Journal*, 4(12):31–32.

Vít Suchomel and Jan Kraus. 2022. Semi-manual annotation of topics and genres in web corpora, the cheap and fast way. In *RASLAN*, pages 141–148. Tribun EU.

Keyu Yang, Yunjun Gao, Lei Liang, Song Bian, Lu Chen, and Baihua Zheng. 2021. Crowdtc: Crowd-powered learning for text classification. *ACM Trans. Knowl. Discov. Data*, 16(1).

Petr Zemánek. 2001. Clara (corpus linguae arabicae): An overview. In *EMNLP*.

Imad Zeroual, Dirk Goldhahn, Thomas Eckart, and Abdelhak Lakhouaja. 2019. Osian: Open source international arabic news corpus - preparation and integration into the clarin-infrastructure. In *WANLP@ACL 2019*, pages 175–182. Association for Computational Linguistics.