AMWAL: Named Entity Recognition for Arabic Financial News

Muhammad S. Abdo
Indiana University
mabdo@iu.edu

Yash A. Hatekar
Indiana University
yhatekar@iu.edu

Damir Cavar
Indiana University
dcavar@iu.edu

Abstract

Financial Named Entity Recognition (NER) presents a pivotal task in extracting structured information from unstructured financial data, especially when extending its application to languages beyond English. In this paper, we present AMWAL, a named entity recognition system for Arabic financial news. Our approach centered on building a specialized corpus compiled from three major Arabic financial newspapers spanning from 2000 to 2023. Entities were extracted from this corpus using a semi-automatic process that included manual annotation and review to ensure accuracy. The total number of entities identified amounts to 17.1k tokens, distributed across 20 categories, providing a comprehensive coverage of financial entities. To standardize the identified entities, we adopt financial concepts from the Financial Industry Business Ontology (FIBO, 2020), aligning our framework with industry standards. The significance of our work lies not only in the creation of the first customized NER system for Arabic financial data but also in its potential to streamline information extraction processes in the financial domain. Our NER system achieves a Precision score of 96.08, a Recall score of 95.87, and an F1 score of 95.97, which outperforms state-of-the-art general Arabic NER systems as well as other systems for financial NER in other languages.

1 Introduction

Financial markets are characterized by their volatile dynamics and constantly changing structure. Price movements, trading volume, and market liquidity are all factors that contribute to this fluid environment. One of the major tools that were found to assist with the analysis and navigation of these complex financial landscapes is financial news. This type of news has been identified as instrumental in predicting stock price movements (Schumaker and Chen, 2009), understanding market sentiment (Devitt and Ahmad, 2007), and informing investor decisions (Alanyali et al., 2013). Additionally, the automated analysis of financial news can provide deeper insights into market dynamics, assist governments in regulating markets, and help intelligence agencies with monitoring for anomalies and unusual events (Passonneau et al., 2015).

Luckily, with the proliferation of online financial news platforms, we are now witnessing an abundance of textual data that is readily accessible for analysis. However, the majority of this data is unstructured, i.e., does not have a standardized format, which presents a significant challenge for effective analysis and interpretation.

Named Entity Recognition (NER) stands as one of the common approaches that aim at organizing such unstructured data into distinct categories, thereby facilitating identifying relations and patterns of interaction among these categories (Qu et al., 2023). Even though there have been notable advances and expansions in Arabic NER systems (Jarrar et al., 2023), the majority remain generic (i.e., detects entities for People, Organizations, Countries, etc.) rather than domain-specific, with the exception of the medical domain (Hamad and Abushaala, 2023; Nayel et al., 2023). This paper aims to address this gap by introducing AMWAL, a NER system that is designed specifically for extracting financial entities from Arabic financial news articles.

The remainder of this paper is organized as follows. Section 2 reviews works that are pertinent to building financial NER systems. Section 3 details the methodology of building AMWAL. In section 4, we report the system's results, and Section 5 is dedicated for discussion, conclusion, and potential avenues for future research.

2 Related Works

Kumar et al. (2023) is one of the few studies that

Proceedings of the Joint Workshop of the 9th FinNLP, the 6th FNP, and the 1st LLMFinLegal, pages 207–213
January 19–20, 2025.

proposed a modeling framework for financial NER using semi-structured banking transaction information from SMSs in Arabic and English. To that end, they performed student-teacher knowledge distillation by employing a pre-trained language model on English (teacher), a high-resource language, and transferring knowledge to a smaller model (student) for Arabic. They also leveraged consistency training through further fine-tuning the Arabic model on the target language using unlabeled data. Utilizing only 30 labeled examples, their model succeeded in generalizing the recognition of categories such as Merchants and Amounts in both languages. In terms of model performance, while their model achieved an F1 score of 0.9768 on the English dataset, the F1-score for the Arabic dataset was 0.6540.

Addressing limitations in existing NER resources and the scarcity of publicly available financial corpora, Jabbari et al. (2020) developed a French corpus with a custom ontology of financial concepts. The corpus focused on entities and their relationships that are pertinent to a set of identity verification guidelines called Know Your Customer (KYC). To build the corpus, they collected 1 million news articles from 40 daily French financial newspapers. Next, they compiled a list of 130 keywords featuring company names, financial interactions, currencies, etc., which were later used to randomly select 130 articles for manual annotation. Their corpus included a total of 6736 entities and 1754 relations, with varying distribution across different types. In their experiments, To test the performance of their annotated corpus in NER and relation extraction, they employed the training modules provided by SpaCy (Honnibal and Montani, 2017), which allows for custom NER training. Overall, their model achieved an F1 score of 0.73. The categories of Person and Currency exhibited the highest accuracy and recall rates, respectively. For the task of exact relation extraction and using rule-based extraction methods, they achieved a Precision score of 0.81, a Recall score of 0.34, and an F1 score of 0.49.

One of the issues that often pose challenges to NER systems is abbreviations due to their diverse forms and lack of clear distinguishing features. To address this Wang et al. (2014), developed a model specifically designed for recognizing financial abbreviations in financial Chinese news texts. Their approach leveraged domain-specific knowledge and context information in a three-step process. First, stock names were extracted as initial clues for identifying potential financial entities. This was followed by the identification of internal features such as suffix keywords, geographic terms, and adjacent words. Finally, they employed a combination of mutual information (MI), boundary information entropy (IE), and word similarity to identify potential abbreviations. This approach achieved 91.02 precision, 93.77 recall, and an F1 score of 0.92.

With regard to the available NER systems for Arabic, as mentioned above, most of the models are generic in terms of the entities they recognize. Jarrar et al. (2022) compiled an Arabic nested NER corpus, Wojood, that was manually annotated with 20 entity types and supports four layers of nesting. The overall performance of the model achieved an F1 score of 0.88. Inspired by AraBERT and BioBERT, Boudjellal et al. (2021) developed an NER model for Arabic biomedical data. The model was trained on AraBERT's original data in addition to medical Arabic literature. Their model outperformed AraBERT and BERT on the bioNER task. Similarly, Hamad and Abushaala (2023) presented a model for recognizing medical terms in Arabic text using Support Vector Machine (SVM) classification. Trained on 27 medical documents with part of speech tags, FastText, and TF-IDF embeddings, they achieved an F1 score of 77.61, which outperformed the state-of-the-art model at the time.

3 Methodology

In this section, we describe the methodology of building and training the model. First, we outline the steps followed in collecting and preprocessing the data. Then, we discuss the rationale guiding the selection of the financial entities. Finally, we talk about the training process.

3.1 Data Collection and Pre-Processing

To build a corpus for Arabic financial news, we collected a total of 26,231 articles from three major financial newspapers: 11,012 articles from Almal News, 8,106 from Al-Sharq, and 2,627 from the business section of Aljazeera newspaper. The data collected, which amounts to 9.8 million tokens, covers a time span of more than two decades from 2000 to 2023.

For data pre-processing, we employed the same steps we followed in (Hatekar and Abdo, 2023) to ensure consistency and mitigate the risk of over-

looking words because of orthographic variations. These steps included the removal of all diacritics as well as the normalization of [ؤ ئ] to [ء], [أ آ إ] to [ا], [ة] to [ه], and [ى] to [ي].

3.2 Entity selection

To avoid arbitrary selection of entities, we adopted the main entities found in the Financial Industry Business Ontology (FIBO). FIBO (EDM Council and Object Management Group, Inc., 2017) is a structured, formal ontology that provides a lexicon of vocabulary for financial concepts. Due to the spread of inconsistent terminologies in the financial domain, FIBO was created to provide a standardized way to represent financial concepts and their relationships (Bennett, 2013; Petrova et al., 2017). We also decided to supplement these entities with other entities that we deemed relevant. The entity types which we added to the model are the GEOPOLITICAL, BANK, METRIC, STOCK EXCHANGE, MEDIA, and FINANCIAL MARKET entities. Table 1 illustrates the 20 entities we chose as well as their frequencies.

While many studies opt for rule-based NER systems (Farmakiotou et al., 2000; Eftimov et al., 2017; Soomro et al., 2017), they often fail to capture entities that do not fall within these rules. For instance, if we consider the word "manufacturing", and how it is rendered in Arabic, we would have distinct unpredictable lexical and orthographic variations of the word. These variations can include translations of the word as either تصنيع or صناعة. Another option would be transliterating the word, which usually results in non-standardized, often arbitrary, variations such as مانيوفاكشرنج, مانيوفشكرنج, مانيوفاكتشرنغ, مانيوفاكتشرنج, مانيوفشكرنغ, to name but a few, and the more vowels the word has, the more variations it can exhibit. Hence, we decided to adopt a more lexical approach using corpus features.

3.3 Semi-Automated Entity Annotation

Our twofold lexical approach starts with searching the corpus for entities using the two queries below with Textometry (Heiden, 2010).

- **[Hypernym such as Hyponym]** which would extract expressions such as "بنوك مثل البنك الإسلامي الفلسطيني" *Banks such as Palestine Islamic Bank*" or "أدوات مالية مثل الأسهم" *Financial instruments such as stocks.*"
- **[Hyponym$_1$ X and Hyponym$_1$ *]** which returned search results such as "بنك القاهرة وبنك الإسكندرية" *Cairo Bank and Alexandria Bank*". Another variation of this pattern is [Hyponym X and *] as in "منتجات البترول والكيماويات" *petroleum and chemical products*"

By utilizing these queries, we are not only minimizing subjectivity, reducing effort, and streamlining entity extraction, but we are also obtaining entities that are inherently annotated. In a query such as “companies such as Apple”, the word "Apple" is already annotated as a company. Also, we only considered entities which occurred at least 5 times in the query results. Next, to ensure that we capture more entities, we performed a frequency analysis for the entities extracted with these queries. The 10 most frequent tokens (e.g., شركة *company*) and bigrams (e.g., الهيئة العامة *General Authority*) were then used as keywords for searching the corpus. In this fashion, our model captures the different variations an entity might exhibit, e.g., shortened proper noun (e.g., أوراسكوم *Orascom*), full proper noun (شركة أوراسكوم للإنشاء *Orascom Construction PLC*), or simply (e.g., شركة أوراسكوم *Orascom company*). The model also captures the non-Arabic titles of these entities should they exist in the corpus. The two search processes resulted in a total of 17185 financial entities. Finally, we performed a manual review over the entities to ensure consistency and accuracy of entities and their respective labels.

4 Experiment

The following steps were carried out to train the model. First, all TXT files were converted into a JSON format, which included the content of the article, the entity token (e.g.,الايكونوميست *the Economist*) and type (e.g., Media) as well as the start and end positions of each recognized entity in the text. Using these JSON files, we randomly allocated 80% of the data (20,984 JSON files) for training the model and 20% (5247 JSON files) for testing its performance. We performed the split at the file level to maintain consistency and to ensure that no overlapping instances exist between training and testing datasets.

For building and training the system, we leveraged SpaCy’s built-in NER toolkit. SpaCy is an

Entity	Count
CORPORATION	6840
QUANTITY OR UNIT	2406
EVENT	1417
PRODUCT OR SERVICE	1222
PERSON	1193
BANK	1185
METRIC	941
OFFICIAL	794
CITY	756
ROLE	692
GEOPOLITICAL	519
COUNTRY	436
NATIONALITY	394
GOVERNMENT ENTITY	225
TIME	217
STOCK EXCHANGE	158
FINANCIAL MARKET	130
FINANCIAL INSTRUMENT	123
CURRENCY	107
MEDIA	103
Total	**17185**

Table 1: Counts of Unique Entities

open-source library which provides tools for building different NLP applications including custom NER systems. For processing Arabic in the configuration file, SpaCy was configured to use the transformer-based model Large AraBERT (Antoun et al., 2020). The model was trained using a batch size of 50 (batch_size = 50), and to avoid overfitting, we set the dropout regularization to 0.1 (dropout= 0.1). The model was also configured to be trained with a maximum of 20,000 update steps (max_steps = 20000) and early stopping (patience = 1600). The model was then trained using a single GPU node and 64GB of memory allocation.

System	Precision	Recall	F1
AMWAL	96.08	95.87	95.97
CAMEL	91.00	91.00	91.00
WOJOOD	80.00	81.00	80.00

Table 3: Macro-Averaged Overall Performance of Models Across Systems

To evaluate the performance of the system over test data, we used Precision, Recall, and F1 scores. As table 2 illustrates, the entity types of CURRENCY, TIME, and EVENT had the overall highest precision and F1 scores, whereas CORP and PERSON were comparatively lower. Also, as Table 3 indicates, the overall performance of the model, with 96.08 Precision, 95.87 Recall, and 95.97 F1 scores outperforms other financial NER models in other languages such as Chinese (Wang et al., 2014), Turkish (Dinç, 2022), Greek (Farmakiotou et al., 2000), French (Jabbari et al., 2020), and German (Hillebrand et al., 2022). These comparisons are only meant to provide context for our model's performance rather than to serve as direct benchmarks against models in the other languages.

AMWAL demonstrates superior performance in financial NER compared to existing Arabic NER models. As shown in tables 2 and 3, AMWAL outperforms CamelBert MSA NER (Inoue et al., 2021) and Wojood FlatNER (Jarrar et al., 2022) in financial NER tasks. This improvement can be due to AMWAL's broader set of entities and labels being specifically targeted towards the financial domain. In contrast, CamelBert and Wojood are regarded as general-purpose NER models and are less specific to the financial domain.

5 Error Analysis

Evaluating our AMWAL system revealed several insights regarding its performance and limitations. Despite achieving high evaluation scores overall, specific challenges persist in the system's handling of certain entity categories, e.g., Corporation and Person. We noticed that in Corporation for instance many of the entities were not labeled correctly because several company names included categories that overlap with other categories we have such as products or services (e.g., Euromed for **Medical Industries**), nationalities (e.g., Wind **Italy**), or even temporal references. For example, ***Nissan*** shares the same spelling as the word for the month of April in Levantine Arabic "نيسان". The same issue persists with the Person category, where some individuals' names include nationalities, such as "السويدي (the Swedish). AMWAL's excellent performance also hinges on it being good at tagging seen data, which might be seen as overfitting; however, even general-purpose NER models fail at such unseen data. Thus, further analysis on training with more diverse and domain-specific data could enhance AMWAL's ability to generalize to unseen instances. Also, incorporating strategies such as expanding the training dataset to include more examples of overlapping or ambiguous categories, applying data augmentation techniques, and fine-

Entity	AMWAL			CamelBERT MSA NER			Wojood FlatNER		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
BANK	89	92	91	36	18	24	13	5	7
CITY	78	84	81	100	99	99	81	98	89
CORP	82	80	81	96	95	96	86	80	83
COUNTRY	97	97	97	88	86	87	53	69	60
CURRENCY	99	99	99	83	60	69	29	7	11
EVENT	98	98	98	96	98	97	93	94	94
FINANCIAL INSTRUMENT	97	97	97	39	12	19	0	0	0
FINANCIAL MARKET	97	92	94	0	0	0	0	0	0
GEOPOLITICAL	91	90	91	48	28	35	28	9	14
GOVERNMENT ENTITY	97	98	98	33	36	34	0	0	0
MEDIA	91	94	93	100	99	99	86	98	92
METRIC	97	92	94	28	12	17	9	2	3
NATIONALITY	97	97	97	26	23	24	0	0	0
OFFICIAL	91	86	89	67	68	68	10	34	15
ORG	85	85	85	36	20	26	0	0	0
PERSON	83	77	80	99	98	99	98	98	98
PRODUCT OR SERVICE	95	95	95	46	44	45	11	4	6
QUANTITY OR UNIT	96	96	96	51	56	53	33	6	10
ROLE	86	90	88	61	67	64	22	6	10
STOCK EXCHANGE	98	98	98	0	0	0	0	0	0
TIME	99	99	99	38	58	46	26	73	38

Table 2: Performance Metrics by Entity Type Across AMWAL, CamelBert MSA NER, and Wojood FlatNER

tuning the model with additional context-aware features could address these limitations. Additionally, employing transfer learning or leveraging external knowledge bases could help resolve ambiguities.

6 Limitations

Due to the nature of this task, i.e., recognizing entities in financial news, AMWAL may not be able to generalize over different variations of Arabic other than MSA, which means that this may limit the model's ability to generalize over other financial sources such as blogs or social media posts.

7 Conclusion

In this paper, we described the development of AMWAL, the first Arabic financial named entity recognition system. To build the model, we first created a corpus from three major Arabic financial newspapers and then used a twofold semi-automated approach to extract entities from the corpus, which we believe is adaptable to other languages that exhibit similar linguistic patterns. Further, in order to avoid arbitrary or subjective choices in selecting the entity types, we adopted financial entities from the Financial Industry Business Ontology (FIBO). We trained the model using SpaCy's custom NER pipeline and employed Arabert Large for processing the data. The evaluation results of the model on the test data showed strong performance metrics with precision at 96.08%, recall at 95.87%, and F1-score at 95.97%, outperforming financial NER systems in other languages as well as general-purpose Arabic NER systems. For future directions, we consider the following steps. First, we aim to expand the size of the corpus as well as the number of entity types. This entails restructuring the identified entities into more intricate hierarchical structures. Additionally, we are considering expanding the scope of the model to encompass not only entity types but also their interrelations, with the ultimate objective being building an Arabic financial knowledge graph that can better inform various stakeholders in the field of Finance.

8 Data Availability

We are sharing SpaCy's best model for our system as well as the SpaCy training and testing files via Github [1].

9 Acknowledgement

This research was supported in part by Lilly Endowment, Inc., through its support for the Indiana University Pervasive Technology Institute.

[1] https://github.com/Muhsabrys/AMWAL/

References

Merve Alanyali, Helen Susannah Moat, and Tobias Preis. 2013. Quantifying the relationship between financial news and the stock market. *Scientific reports*, 3(1):3578.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

Mike Bennett. 2013. The financial industry business ontology: Best practice for big data. *Journal of Banking Regulation*, 14(3):255–268.

Nada Boudjellal, Huaping Zhang, Asif Khan, Arshad Ahmad, Rashid Naseem, Jianyun Shang, and Lin Dai. 2021. Abioner: a bert-based model for arabic biomedical named-entity recognition. *Complexity*, 2021:1–6.

Ann Devitt and Khurshid Ahmad. 2007. Sentiment polarity identification in financial news: A cohesion-based approach. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 984–991.

Duygu Dinç. 2022. Financial named entity recognition for turkish news texts. Master's thesis, Middle East Technical University.

EDM Council and Object Management Group, Inc. 2017. Financial Industry Business Ontology – Indices and Indicators. OMG Document Number: formal/2017-07-01, Release Date: July 2017, Normative Reference: http://www.omg.org/spec/EDMC-FIBO/IND/1.0/.

Tome Eftimov, Barbara Koroušić Seljak, and Peter Korošec. 2017. A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations. *PloS one*, 12(6):e0179488.

Dimitra Farmakiotou, Vangelis Karkaletsis, John Koutsias, George Sigletos, Constantine D Spyropoulos, and Panagiotis Stamatopoulos. 2000. Rule-based named entity recognition for greek financial texts. In *Proceedings of the Workshop on Computational lexicography and Multimedia Dictionaries (COMLEX 2000)*, pages 75–78.

Rema Muftah Hamad and Ahmed Mohamed Abushaala. 2023. Medical named entity recognition in arabic text using svm. In *2023 IEEE 3rd International Maghreb Meeting of the Conference on Sciences and Techniques of Automatic Control and Computer Engineering (MI-STA)*, pages 200–205. IEEE.

Yash Hatekar and Muhammad Abdo. 2023. Iunadi at nadi 2023 shared task: Country-level arabic dialect classification in tweets for the shared task nadi 2023. In *Proceedings of ArabicNLP 2023*, pages 665–669.

Serge Heiden. 2010. The txm platform: Building open-source textual analysis software compatible with the tei encoding scheme. In *24th Pacific Asia conference on language, information and computation*, volume 2, pages 389–398. Institute for Digital Enhancement of Cognitive Development, Waseda University.

Lars Hillebrand, Tobias Deußer, Tim Dilmaghani, Bernd Kliem, Rüdiger Loitz, Christian Bauckhage, and Rafet Sifa. 2022. Kpi-bert: A joint named entity recognition and relation extraction model for financial reports. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 606–612. IEEE.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1):411–420.

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. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, Kyiv, Ukraine (Online). Association for Computational Linguistics.

Ali Jabbari, Olivier Sauvage, Hamada Zeine, and Hamza Chergui. 2020. A french corpus and annotation schema for named entity recognition and relation extraction of financial news. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2293–2299.

Mustafa Jarrar, Muhammad Abdul-Mageed, Mohammed Khalilia, Bashar Talafha, AbdelRahim Elmadany, Nagham Hamad, and Alaa' Omar. 2023. Wojoodner 2023: The first arabic named entity recognition shared task. *arXiv preprint arXiv:2310.16153*.

Mustafa Jarrar, Mohammed Khalilia, and Sana Ghanem. 2022. Wojood: Nested arabic named entity corpus and recognition using bert. *arXiv preprint arXiv:2205.09651*.

Sunisth Kumar, Davide Liu, and Alexandre Boulenger. 2023. Cross-lingual ner for financial transaction data in low-resource languages. *arXiv preprint arXiv:2307.08714*.

Hamada Nayel, Nourhan Marzouk, and Ahmed Elsawy. 2023. Named entity recognition for arabic medical texts using deep learning models. In *2023 Intelligent Methods, Systems, and Applications (IMSA)*, pages 281–285. IEEE.

Rebecca J Passonneau, Tifara Ramelson, and Boyi Xie. 2015. Named entity recognition from financial press releases. In *Knowledge Discovery, Knowledge Engineering and Knowledge Management: 6th International Joint Conference, IC3K 2014, Rome, Italy, October 21-24, 2014, Revised Selected Papers 6*, pages 240–254. Springer.

GG Petrova, AF Tuzovsky, and Nataliya Valerievna Aksenova. 2017. Application of the financial industry business ontology (fibo) for development of a financial organization ontology. In *Journal of Physics: Conference Series*, volume 803, page 012116. IOP Publishing.

Xiaoye Qu, Yingjie Gu, Qingrong Xia, Zechang Li, Zhefeng Wang, and Baoxing Huai. 2023. A survey on arabic named entity recognition: Past, recent advances, and future trends. *IEEE Transactions on Knowledge and Data Engineering*.

Robert P Schumaker and Hsinchun Chen. 2009. A quantitative stock prediction system based on financial news. *Information Processing & Management*, 45(5):571–583.

Pir Dino Soomro, Sanotsh Kumar, Arsalan Ali Shaikh, Hans Raj, et al. 2017. Bio-ner: biomedical named entity recognition using rule-based and statistical learners. *International Journal of Advanced Computer Science and Applications*, 8(12).

Shuwei Wang, Ruifeng Xu, Bin Liu, Lin Gui, and Yu Zhou. 2014. Financial named entity recognition based on conditional random fields and information entropy. In *2014 international conference on machine learning and cybernetics*, volume 2, pages 838–843. IEEE.