AJEMA GLOBAL RESEARCH NETWORK

AMERICAN Journal of Engineering, Mechanics and Architecture

Volume 2, Issue 9, 2024 ISSN (E): 2993-2637

Top Accurate Models for Handling Complex Arabic Linguistic Structures

Murteza Hanoon Tuama, Wahhab Muslim mashloosh, Yasir Mahmood Younus

Department of Computer Techniques Engineering, Imam Al-Kadhum University College, Baghdad, Iraq

Abstract: Arabic is a language rich enough in morphology, morphology but poor in corpus and syntax compare to English which makes the other Arabic Applications in fields such as Arabic Natural Language Processing (NLP) include Question Answering, Named Entity Recognition (NER) and Sentiment Analysis (SA). (QA). Nevertheless, the recent developments of transformer-based models have shown that language-specific BERT models, pre-trained on large corpora, achieve an overall better performance in Arabic comprehension. * They represent the new state of the art, providing excellent results on diverse NLP tasks. We introduce AraBERT in this work - a BERT model specifically constructed for Arabic, where we strive to bring BERT's success to Arabic language similar to as achieved for English. We assess AraBERT against the company Google's multilingual BERT and other cutting-edge techniques. The study found that the newly built AraBERT surpasses the most Arabic NLP.

Keywords: Natural Language Processing , Arabic language , Sentiment Analysis, AraBERT.

INTRODUCTION

Text data mining is defined as the process of extracting data from texts, in other words, extracting information and patterns from text data that is usually unstructured data. It is itself done through natural language processing and includes several applications, including (answering questions, named entity recognition and sentiment analysis). but these models did not include context in the embeddings they created for each word. Due to this constraint, contextualized representations were added using models like ELMO. (Peters et al., 2018). Recently, there was a lot of interest in applying transfer learning to improve downstream NLP/NLU tasks. It does this by a small set of cases used to adapt large pre-trained language models. This approach has increased our ability to do the job! The primary benefit of using language models that have been pre-trained in an unsupervised fashion — a type of self-supervised learning. But it's important to know the limitations of this method. One disadvantage is the need for large corpora for the pre-training phase. Furthermore, the computational cost is also huge since the state-of-the-art models require over 500 TPUs or GPUs for weeks (Conneau et al., 2019; Raffel et al., 2019). Due to these issues, such models are limited to English and some other languages. To address this, multilingual models are taught to learn over 100 languages simultaneously, however due to limited vocabulary and scarce data, they still fall below monolingual models in performance. Some languages, particularly those with comparable vocabulary and structure, can benefit from common representations (Conneau et al., 2019) though clearly fewer than for languages with a more distant relationship. Although Arabic is difficult to represent morphologically and syntactically, this study focuses on the process of Arabic language processing through from the BERT Transformer model (Devlin et al, 2018). We evaluate ARABERT through the three tasks we mentioned earlier (answering questions,

recognizing named entities, and finally sentiment analysis). The results show that the evaluation we conducted demonstrated the superiority of ARABERT over similar baselines, including the monolingual and multilingual approaches.

We explore two datasets for downstream tasks: Dialectal Arabic (DA) and Modern Standard Arabic (MSA). Our input may be summed up as follows:

- A method for applying the BERT model to three downstream NLU tasks: Called sentiment analysis, question answering, and entity recognition;
- ARABERT is applied to these tasks using a large Arabic corpus.
- ARABERT is accessible to the general public on well-known NLP libraries.

In this section, We explain how we organized the article. We reviewed previous work on the linguistic representation of Arabic and English in Section 2. In the following section (3), the ARABERT methodology is discussed. In Section (4), we describe the dataset used. In Section 5, we detail the experimental design and results. In Section (6), we include the conclusion, recommendations, and future work.

RELATED WORKS

Put word inclusions meaningful word vectors together microlc et al word2vec (2013). Most of the further work since then focused on a few refinements of word2vec, such as GloVe (Pennington et al., 2014) and Fast Text (Mikolov et al., 2017). These early models were significantly better than earlier methods, but they still lacked context e.g. ELMO (Peters et al. 2018). Learned representations of words and phrases, and larger, more structured representations, as the result of performance increases across a wide variety of tasks. Many language understanding models have emerged since then, including ALBERT (Lan et al 2019), T5, and more. These are all pre-trained and fine-tuned given state-of-the-art unsupervised models like T5 (Raffel et al., 2019), RoBERTa (Liu et al., 2019), XLNet (Yang et al., 2019), BERT (Devlin et al., 2018) and ULMFit (Howard and Ruder, 2018). These models enhance performance using various correlation methods, altered model topologies, and expanded training corpora.

Isolated Arabic embeddings nLP researchers tried to adapt the success of word2vec (Mikolov et al., 2013) to produce language-specific embeddings. Now tried by Soliman et al. Bojanowski et al. Word2vec (2017) was trained over Wikipedia data in Arabic but the Fasttext model (2017) performed better than word2vec. While (Abu Farha and Magdy, 2019; AbdulMageed et al., 2018) trained word embeddings for Arabic on 210 million and above tweets, (Erdmann et al., 2018) proposed a way to learn multi-dialect word embeddings using a relatively small noisy corpus in order to reduce dialectal heterogeneity in Arabic.

Contextualized representations in Arabic A multilingual BERT model, produced by Google (Devlin et al., 2018), was trained to cover over 100 of the current spoken languages, yielding good performance across the majority of them and especially for the nonEnglish ones. Yet, to pre-train monolingual BERT is shown to be superior than multilingual BERT for non-English languages (Martin et al., 2019; de Vries et al., 2019) and for Italian BERT Alberto (Polignano et al., 2019). ElJundi et al. Reported (2019) that Arabic-specific contextualized representations models outdo BERT in English natural language processing experiment, but hULMonA uses ULMfit structur.

ARABERT

Methods In order to improve the performance on multiple Arabic natural language understanding tasks, we introduced an Arabic language representation model in this research. ARABERT is derived from the stacked Bidirectional Transformer Encoder (BERT) model (Devlin et al., 2018). This model underpins most state-of-the-art results across many NLP tasks in many languages. We use BERTbase125: 110 M parameters, 12 attention heads, 12 encoder blocks, 768 hidden dimensions, 512 maximum sequence length. وروجعنا preprocessing أكثر في ما

6 قبل مرحلة التعلم المسبق لتحسين إرضاء النموذج باللغة العربية Training Overview Here we present the pre-training setup, the ARABERT pre-training corpus, the fine-tuning strategy, and the advised Arabic preprocessing.

Pre-training Setup

We use full word mask and Masked Language Modeling (MLM) task, which is the original BERT pre-training target task that, in MLM, replaces 15% of the N total input tokens with the [MASK]. 10% is a random token, 10% is [MASK] token and 80% will be [MASK] token positioned where the original token was. Thus, instead of predicting the individual tokens, the model will predict the entire word since we also apply wholeword masking, thus creating an even more challenging pre-training. Neither do we, take into consideration the relationships whenever there's a pair of phrases to those phrases, so we allow The approach uses Next Sentence Prediction (NSP) to identify the association between two phrases. process, that is really useful for many of the language understanding duties, for instance Question Answering.

Pre-training Dataset

Zhu et al. Based on (2015) the first BERT was trained on 3.3 billion words from English Wikipedia and Book Corpus. There are fewer stories in the arabic wikipedial page dumps than in the english ones, we scraped stories from arabic sources ourselves. Furthermore, we used the following two publicly available Arabic corpora: (1) an open-source international Arabic corpus involved news text (Zeroual et al., 2019), a modern corpus that included up to 5M eight countries' worth of articles from ten news sources, and (2) a 1.5 billion word Arabic language corpus (El-Khair, 2016). It contains 3.5 million articles (about 1 billion tokens) from 31 news sources across 24 Arab nations. The final pretraining dataset size, after stripping out all the dialogue that fails to add any information, is 70 million individual sentences, or roughly 24 terabytes worth of text. Given the diversity of media outlets coverage across different Arab regions, this dataset can be a representative sample for publication on many topics discussed in the Arab world. We would like to note that, as a matter of principled practice, to prevent the loss of information in Latin–character phrases, it is general consideration to refer to well-known entities, scientific terms and tools with their original writing.

Segmentation of Sub-Word

Units Due to its complex concatenative structure, The lexical sparsity of Arabic is widely acknowledged (Al-Sallab et al., 2017). For example, the same word can have different forms, but still has the same meaning. For example, he says, the English equivalent of "the," "ال - Al," the definite article that is always attached to other nouns but never forms part of the word. As a result, we will see the tokens very much in duplicates when using a BERT compatible tokenization — once with the "Al-" prefix and once without. Analyses such as this type of review require that every individual term, e.g., "كتاب- kitAb" and "الكتاب-AlkitAb", be included in the lexicon, causing a lot of duplication. To overcome this problem, we first reduce the words to stems, prefixes, and suffixes using Farasa (Abdelali et al., 2016). For example, language - Alloga becomes: +لغ+ة - Al+ log +a After that we train a Sentence Piece; an unsupervised using segmented pre-training dataset text tokenizer and detokenizer (Kudo, 2018) in unigram mode. Thus, we had a subword vocabulary of approximately 60K tokens. A new Sentence Piece was then trained on the nonsegmented text to produce ARABERTv0. 1, a second version of the proposed tokenization method that does not require segmentation, to test its performance. There was a final vocabulary size of 64k tokens + 4K tokens for potentially pre-training.

Optimizing Sequence Categorization

The final hidden state of the initial token ([CLS]) appended to the sentences is used to fine-tune AraBERT for the sequence classification task. After that, we apply a second feed-forward layer and use the traditional Softmax over the output to acquire a distribution among the target classes.

To refine the log-likelihood of the correct class during fine-tuning, the classifier and pre-trained model weights were used. Each token in the sentence is labeled according to the NER task for named entity recognition in IOB2 format (Ratnaparkhi, 1998). The words in the sentence are represented in the following order: First, the first word is represented by the tag "B", and in the same way, the second word of the sentence is represented by the tag "I", while the non-entity word is represented by the tag "O". Based on the above, it becomes clear why we treat the system as a multi-class classification problem so that we can label the tokens using multiple data classification techniques. In addition, we use the AraBERT segmentation tool to pass the first subclass of each single word to the model. In order to find the answer from the text, the QA model needs to use the question and answer. This is done by predicting a start symbol and the end symbol that usually follows it, with the "end" symbol being required to appear first. The final embedding is produced during training for each symbol. The softmax layer takes the resulting embeddings and the dot product of the classifier to produce a probability distribution for each symbol. The same process is repeated for the end symbol and the symbol with the maximum probability of becoming a start symbol is determined.

EVALUATION

Sentiment analysis, named entity recognition, and question answering are the three downstream tasks in Arabic language understanding that we used to test ARABERT. We benchmarked ARABERT against the multilingual BERT version and against other state-of-the-art performance on each job.

Sentiment Analysis

The following Arabic sentiment datasets, which span several genres, domains, and dialects, were used to assess ARABERT.

- HARD: Elnagar et al. (2018)'s Hotel Arabic evaluations Dataset includes 93,700 hotel evaluations written in dialectal Arabic in addition to Modern Standard Arabic (MSA). Reviews are categorized as either favorable or negative, with a rating of 1 or 2, a rating of 4 or 5, and a value of 3 for neutral reviews that were disregarded.
- ASTD: Nabil et al. (2015) compiled 10,000 tweets in both MSA and Egyptian dialect from the Arabic Sentiment Twitter Dataset. We conducted our tests using the ASTD-B, or balanced form of the dataset.
- ArSenTD-Lev: The Arabic Sentiment tweets are usually generated from a dataset from multiple sites and include multiple topics, which adds difficulty in using them. Among these the dataset available on Twitter, or what is now called the X platform. It contains 4,000 Arabic sentiment datasets listed in the Levantine dialect and annotated on the sentiment, topic, and emotional goal (Baly et al., 2018).
- LABR: 63,000 Arabic-language book evaluations may be found in the Large-scale Arabic Book evaluations dataset (Aly and Atiya, 2013). Ratings for the reviews range from 1 to 5. The imbalanced two-class dataset, where evaluations with ratings of 1 or 2 are regarded as unfavorable and those with ratings of 4 or 5 as positive, is used as a baseline for our model.
- AJGT: There are 1,800 tweets written in Jordanian dialect in the Arabic Jordanian General Tweets dataset (Alomari et al., 2017). Positive or negative annotations were manually made to the tweets.

Baselines: Sentiment analysis is a well-known NLP task in Arabic. Earlier approaches were based on sentiment lexicons such as ArSenL (Badaro et al., 2014), a comprehensive list of MSA terms that was built through the merged use of the SentiWordNet in English and the Arabic WordNet. Several approaches have proposed arabicspecific processing for recurrent and recursive neural networks (Al Sallab et al., 2015; Al-Sallab et al., 2017; Baly et al., 2017). One approach is to employ pre-trained word embeddings for the training of Convolutional Neural

Networks (CNN (Dahou et al., 2019a)). Abu Farha & Magdy (2019) proposed a hybrid model using LSTMs for sequence and contextual interpretation and using CNNs for feature extraction. Following the ULMfit architecture (Howard & Ruder, 2018), ElJundi et al. (2019) proposed hULMonA, a state-of-the-art Arabic language model. We contrast the outputs of hULMonA and ARABERT.

Named Entity Recognition

Such tasks might involve locating and retrieving entities referred to in text. The task is then treated as a categorical (or labeling)) problem in a word level scale, where defined categories correspond to objects like: people, places, organizations, events and time. We evaluate on the Arabic NER corpus (ANERcorp) (Benajiba and Rosso, 2007). It has a total of 16.5K entity references and classifies them into 4 categories : person (39%), organization (30.4%), place (20.6%) and miscellaneous (10%)

Baseline Methods: The CoNLL 2003 (Sang and De Meulder, 2003) dataset has been a hub for progress in the NER problem in English. Conditional Random Fields (CRF) (Lafferty et al., 2001) was the first approach that employed such an architecture for the task of named-entity recognition (NER). After that CRFs, alongside with Bi-LSTM models, were used, and interesting results appeared (Huang et al., 2015; Lample et al., 2016). Further improvements were observed by using contextualized embeddings into Bi-LSTM-CRF structures (Peters et al., 2018). Our Results in a Nutshell: Large pre-trained transformers where a small gain and, we have established the current best performance (Devlin et al., 2018). For Arabic, we then compare the results of ARABERT with BERT multilingual and the Bi-LSTM-CRF baseline, which established the previous state-ofthe-art performance (El Bazi and Laachfoubi, 2019).

Responding to Inquiries

One of artificial intelligence's goals is open-domain question answering (QA), which can be accomplished by employing knowledge acquisition and natural language understanding (Kwiatkowski et al. Some small datasets such as Stanford Question Answering Dataset (SQuAD) are available which enriched English QA research (Rajpurkar et al., 2016). Nevertheless, the lack of large datasets and the specific challenges accompanying Arabic QA have obstructed research in this space. These difficulties include:

- Name spelling variations (e.g., Syria can be spelled as "سوريا - sOriyA" or "سوريةf - sOriyT" in Arabic).
- Renaming (for example: "عبد العزيز - AbdulAzIz" appears in the query, but "عبد العزيز - Abdul AzIz" appears in the response).
- The dual form "المثنى," which can take on many forms (for example, "قلمان" - "qalamAn" or "قلمين" - "qalamyn" - "two pencils").
- The grammatical gender variation: all nouns, both animate and inanimate items are divided into two genders: masculine and feminine. (ex: "كبير - "kabIr" and " كبيرة" - "kabIrT".

  - To assess ARABERT, we used the Arabic Reading Comprehension Dataset (ARCD) (Mozannar et al., 2019). The job that this job entails is to identify the span of an answer in a document given a query. ARCD consists of 2966 machine-translated questions and answers from SQuAD, referred to as ArabicSQuAD, as well as 1395 questions developed from wikipedia articles. We use 50% of ARCD for evaluation and the full Arabic-SQuAD dataset for training. The prior state of the art on ARCD was completed by Multilingual BERT with Baselines.

EXPERIMENTS Experimental Setup

On We used the official TensorFlow BERT implementation for our work.Once the pre-training data had been split up into TFRecords and sharded, it was stored to Google Cloud Storage.

Masked probability was 15%, duplication factor was 10, and random seed was 34. It was pre-trained for 1,250,000 steps on a TPUv2-8 pod. Using 128-symbol sequences, the first 900 steps were trained, and to speed up the training time, 512-symbol sequences were used to train the remaining steps. Pre-training was ended when downstream tasks were completed. We use the same approach as the German BERT open-source implementation (DeepsetAI, 2019). Tabular data was trained using the Adam optimizer with batch sizes of 512 and 128, for sequence lengths of 128 and 512, respectively, and a learning rate of 1e4. Over the course of 4 days, we scanned 27 epochs where all the tokens were trained. Adjustments For each job, independent fine-tuning was performed using the same setting. Due to time and computing constraints, we do not perform an exhaustive grid search for the optimal hyper-parameters. We utilize the splits supplied by the dataset's authors where accessible. and, when it was not, the usual 80 percent and 20 percent.

Results

Table 1 compares state-of-the-art findings and the multilingual BERT model (mBERT) with the experimental results of using AraBERT to numerous Arabic NLU downstream tasks.

Sentiment Analysis Table 1's results for Arabic sentiment analysis demonstrate that both AraBERT variants perform better than mBERT and other cutting-edge methods on the majority of evaluated datasets. Despite having been trained on MSA, AraBERT performed admirably on dialects that had never been encountered before.

Table 1: AraBERT's performance on downstream Arabic tasks in comparison to mBERT and earlier state-of-the-art systems

Task	Metric	prev. SOTA	mBERT	AraBERTv0.1/ v1
SA (HARD)	Acc.	95.7*	95.7	96.2 / 96.1
SA (ASTD)	Acc.	86.5*	80.1	92.2 / 92.6
SA (ArsenTD-Lev)	Acc.	52.4*	51	58.9 / 59.4
SA (AJGT)	Acc.	92.6**	83.6	93.1 / 93.8
SA (LABR)	Acc.	87.5†	83	85.9 / 86.7
NER (ANERcorp)	macro-F1	81.7††	78.4	84.2 / 81.9
QA (ARCD)	Exact Match	Mbert	34.2	30.1 / 30.6
	macro-F1		61.3	61.2 / 62.7
	Sent. Match		90	93.0 / 92.0

State-of-the-art output of the old BiLSTM-CRF model

As can be seen in Table 1, AraBERTv0. Achieving an F1 score of 84.2, AraBERT outperformed the Bi-LSTM-CRF model by 2.53 points setting a new state-of-the-art for NER on ANERcorp. Similar results to Bi-LSTM-CRF have been obtained while using tokenized prefixes and suffixes in AraBERT testing. We believe this happened because the start token (B-label) has a common connection with the suffixes. For example, the term "الجامعة" with the label B-ORG is replaced with "جامعة" ,"ال", and so on, with the corresponding labels B-ORG and I-ORG, respectively, putting incorrect first signals into the model. The performance of our multilingual BERT testing was poor compared to the baseline model, making it useless. Answering Questions While the F1-score appeared to improve based on the data presented in Table 1, the exact match scores were significantly lower. When I re-reviewed the results, I noticed that most of the wrong responses were just one or two words different from the right one, without any substantial difference in the semantics of the answer. Examples are shown in Tables 2 and 3. We also demonstrate 2% absolute improvement in sentence match score over prior state-of-the-art,

mBERT. (2) Sentence Match (SM) – Average over sentence that return the correct ground truth response

We can follow these steps to build a BiLSTM-CRF model to analyze the Arabic social discourse sentiment. Transform the data, create the model, fit the model, score the model, and plot the results. Through BiLSTM-CRF model, one of effective model in natural language processing. As with this flowchart:

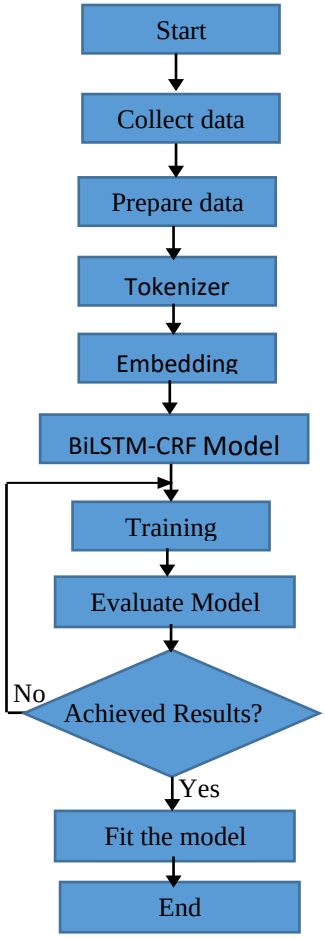

Table 2: An illustration of an incorrect result from the ARCD test set is that the preposition is the only thing that differs : "في ¯ - In".

Question	أين تاسست منظمة الامم المتحدة؟ **where was the united nations established**?
Ground Truth Prdicted Answer	In San Francisco- **في سان فرانسيسكو** San Francisco – **سان فرانسيسكو**

Table 3: An additional illustration of a false result from the ARCD test set is when the anticipated response omits "introductory" terms.

Question	ماهو النظام الخاص بدولة استراليا؟ **What is the type of government ?in Austria**
Ground Truth Prdicted Answer	Austria is a federal republic- **استراليا جمهورية فيدرالية** A federal republic – **جمهورية فيدرالية**

DISCUSSION

We found that AraBERT achieved state-of-the-art results on question answering, named entity recognition, and sentiment analysis tasks. This provides evidence for the hypothesis that a single-language pretrained language model will only outperform a multilingual model in terms of performance. Some potential reasons for that performance gain. Imagining the other is training on data going well up until October 2023. while AraBERT utilized some 24GB of data when 4.3G was used for the multilingual BERT. The second major difference is that in the multi-lingual BERT the vocabulary size used is 2k tokens while in AraBERT there were 64k tokens used. Thirdly, a much larger number of samples leads to more diverse pre-training distribution. For the last point, the NER job shows a trade-off for performance where pre-segmentation was performed before tokenization (with BERT), compared with SA and QA we can see a gain in performance because of pre-segmentation. Many agree that the Arabic language is complex, so it should be noted that the process of processing it through time step processing on the pre-training data reflects this complexity. In order to process some common prefixes that were found as a result of increasing the effective vocabulary, processing was done by deleting duplicate symbols and other unnecessary symbols, thus reducing some of the language complexity of the model, allowing for better learning. Based on the results for eight different datasets, it was shown that these factors helped facilitate the recent results for three separate challenges. The results showed that the monolingual model is better than its bilingual competitor, the model trained on text operations, especially searching in the Wikipedia encyclopedia. Like the BERT model, which supports multiple languages.

CONCLUSION

AraBERT sets a new standard for Arabic downstream tasks in several ways. It is also 300 MB smaller than multilingual BERT. We hope that our public release of the AraBERT models will lead to it becoming the new baseline for many Arab natural language processing tasks. Additionally, we believe that our work will serve as a foundation for developing and refining future models for understanding the Arabic language. We are in the process of developing an AraBERT version independent of third-party tokenizers. Additionally, we have been training models that have a deeper comprehension of the many spoken dialects of Arabic in the various Arabic-speaking countries.

REFERENCES

1. Darwish, K., Mubarak, H., Durrani, N., and Abdulali, A. (2016). Farasa: An ardent and swift Arabic segmenter. Pages 11–16 in Proceedings of the 2016 Conference of the Association for Computational Linguistics' North American Chapter: Demonstrations.
2. Elaraby, M., Alhuzali, H., and Abdul-Mageed, M. (2018). You tweet what you say: An Arabic dialect dataset at the city level. The Eleventh International Conference on Language Resources and Evaluation (LREC 2018) proceedings
3. Magdy, W., and Abu Farha, I. (2019). Mazajak: An Arabic emotion analyzer available online. In August, Florence, Italy, Proceedings of the Fourth Arabic Natural Language Processing Workshop, pp 192–198. The Computational Linguistics Association.
4. Yang, Z., Kulshreshtha, A., Nemade, G., Lu, Y., Hall, J., Fiedel, N., Thoppilan, R., Adiwardana, D., Luong, M.-T., So, D. R., and Le, Q. V. (2020). Toward an open-domain chatbot that resembles a human.
5. Hajj, H., Badaro, G., Baly, R., El-Hajj, W., Al Sallab, A., and Shaban, K. (2015). Arabesque sentiment analysis using deep learning models. Pages 9–17 in the Proceedings of the Second Workshop on Arabic Natural Language Processing.
6. El-Hajj, W., Badaro, G., Baly, R., Hajj, H., Shaban, K. B., and Al-Sallab, A. (2017). Aroma: A low-resource language recursive deep learning model for opinion mining in Arabic. ACM Transactions on Low-Resource and Asian Language Information Processing 16(4):1–20.
7. Shaalan, K., ElSherif, H. M., and Alomari, K. M. (2017). Machine learning sentiment analysis of Arabic tweets. Applied Intelligent Systems, International Conference on Industrial, and Engineering, and Other Applications, pages 602–610. Springer.
8. Atiya, A. and M. Aly (2013). LABR: A comprehensive dataset of Arabic book reviews. In Volume 2: Short Papers, Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, August, Sofia, Bulgaria, pp 494–498. The Computational Linguistics Association.
9. El-Hajj, W., Habash, N., Hajj, H., Baly, R., and Badaro, G. (2014). a comprehensive Arabic emotion vocabulary for opinion mining in Arabic. In Arabic natural language processing (ANLP) workshop proceedings, EMNLP 2014, pages 165–173.
10. Shaban, K. B., El-Hajj, W., Habash, N., Hajj, H., and Baly, R. (2017). For efficient sentiment analysis in Arabic, morphologically enhanced recursive deep models and a sentiment treebank are used. 16(4):1–21 in ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP).
11. Baly R., Khaddaj A., Hajj H., El-Hajj W., and Shaban K. B. (2018). Arsentd-lev is a multitopic corpus for target-based sentiment analysis in Arabic-Levantine tweets. OSACT 3: The Third Workshop on Open Source Arabic Corpora and Processing Tools, p. 37.
12. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzman, F., Grave, E., Ott, M., Zettle-Moyer, L., and Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. J. D. Lafferty, A. McCallum, and F. C. Pereira (2001).
13. Conditional random fields are probabilistic models for segmenting and labeling sequence data. In ICML. Martin L., Muller B., Suarez P. J. O., Dupont Y., Romary L., Eric Villemonte de la Clergerie, Seddah D., and Sagot B. (2019). Camembert, a delectable French language model.
14. Zeroual, I., D. Goldhahn, T. Eckart, and A. Lakhouaja (2019). OSIAN: Open source international Arabic news corpus created and integrated into the CLARIN infrastructure. In the Proceedings of the Fourth Arabic Natural Language Processing Workshop, pp 175-182, Florence, Italy, August. Association of Computational Linguistics.

15. Erdmann A., Zalmout N., & Habash N. (2018). Addressing noise in multilingual word embeddings. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 558-565.
16. Martin L., Muller B., Suarez P. J. O., Dupont Y., Romary L., Eric Villemonte de la Clergerie, Seddah D., and Sagot B. (2019). Camembert, a delectable French language model.
17. Kodo, T. (2018). Subword regularization is the process of improving neural network translation models by introducing several subword possibilities.
18. Rajpurkar P., Zhang J., Lopyrev K., & Liang P. (2016).Squad: Over 100,000 questions for machine understanding of text. arXiv preprint: 1606.05250.
19. Polignano, M., Basile, P., de Gemmis, M., Semeraro, G.,and Basile, V. (2019). AlBERTo: Italian BERT Language Understanding Model for NLP Challenging Tasks Based on Tweets. In Proceedings of the Sixth Italian Conference on Computati