Word Detection UsingConvolutional Neural Networks

Najwa Mohammed Adeeb[1]/ Computer Department Faculty Of Education /University Of Zawia

Basma Emhamed Dihoum[2]/ dept. Computer Science/ University of Jafara Tripoli, Libya
basmadihom@gmail.com

Abstract.

This research focuses on the task of classifying Arabic bro- ken words into singular or plural using a Convolutional Neural Network (CNN). Arabic broken words pose a unique challenge due to their intentional segmentation into smaller units, commonly found in informal textand dialectal variations. The objective is to develop an effective model that leverages the power of CNNs to capture the distinguishing features and patterns of Arabic broken words. The proposed CNN architecture comprises 4 convolutional layers, 4 pooling layers, and a fully connected layer with softmax activation. Each convolutional layer utilizes 64 filters with a size of 3, allowing for the extraction of local patterns and features.The pooling layers down sample the learned features, reducing spatial di- mensions. Through meticulous dataset construction and preprocessing, the model aims to achieve accurate classification, contributing to advancements in Arabic NLP and facilitating information extraction from Arabic text. The outcomes of this research have the potential to enhance various NLP applications, including sentiment analysis, machine translation, and information retrieval, while improving communication betweenhumans and machines in the Arabic language context.

Keywords: Natural Language Processing (NLP) · Broken words ·convolutional neural network (CNN).

Introduction

Natural Language Processing (NLP) [1, 2] is an interdisciplinary field that com- bines artificial intelligence, computational linguistics, and computer science to enable computers to understand, analyze, and generate human language. It en- compasses various tasks such as text classification, sentiment analysis, machinetranslation, named entity recognition, question answering, and language generation [3, 4][5, 6]. NLP has significant applications in improving search engines, virtual assistants, and automating language-based tasks, facilitating communication across different domains and languages.

Stemming plays a critical role in NLP systems as it directly affects the performance of applications that rely on word variations. Arabic language presents unique challenges for stemming due to its complex morphological structure characterized by highly inflected and derivational terms. In addition to the difficulties encountered in light and root-based Arabic stemming, broken plurals pose a specific challenge. These irregular plural forms make it difficult to accurately extract root words [7]. In Arabic, plurals (singular, dual, and plural) are classified into regular and irregular categories. Regular plurals are formed by adding appropriate suffixes, similar to English, such as teacher: teachers. The masculine plural is formed by adding the suffix to the nominative suffix ة in the accusative and genitive cases. The feminine plural is formed by attaching the suffix to the singular. On the other hand, irregular or broken plurals frequently occur in trilateral roots and involve modifying the singular form, for example, tooth: teeth. Many nouns and adjectives exhibit broken plural forms, and singular forms can undergo various pattern changes that alter long vowels consonants ي, or their absence within or outside the framework of the consonants. In this study, we propose the use of Convolutional Neural Networks CNNs) to address the challenges posed by broken plural words in Arabic stemming. CNNs are powerful deep learning architectures known for their ability to capture complex patterns in data. By training a CNN model on a carefully curated Arabic dataset specifically containing examples of broken plurals, our aim is to effectively identify and classify these irregular plural forms. The CNN model consists of multiple convolutional layers that extract local features from the input data, followed by pooling layers to reduce the dimensionality, and fully connected layers for classification. Through experimental evaluations, we demonstrate the effectiveness of the CNN-based approach in accurately identifying and categorizing broken plural words, achieving high precision and recall rates. This research not only provides a novel solution for handling the challenges of broken plurals in Arabic stemming but also opens up possibilities for further advancements in natural language processing tasks specific to the Arabic language [24,25].

Related Works

Al-Saleh et al. [8] conducted a study where they trained a CNN model on a large dataset of Arabic nouns. Their goal was to distinguish between regular and irregular plural forms of Arabic broken plurals. By leveraging the convolutional layers of the CNN, the model learned to capture the morphological patterns specific to each type of plural. The study achieved high accuracy in classifying Arabic broken plurals, showcasing the effectiveness of CNNs in this task. For example, the model correctly classified the singular form كتاب (book) as a regular plural and the plural form كتب (books) as an irregular plural.

Salloum et al. [9] developed a CNN model specifically designed for the classification of Arabic sound and broken plurals. They created a labeled dataset of Arabic nouns, consisting of both regular and irregular plural forms, and trained the CNN to learn the morphological patterns associated with each plural form. By leveraging the power of CNNs to capture local patterns and relationships, the model achieved accurate categorization of Arabic broken plurals. For instance, the model correctly classified أم (mother) as a regular plural and آباء (fathers) as an irregular plural.

Alshamiri et al. [10] employed a CNN architecture with multiple convolutional layers to classify Arabic broken plurals. They utilized a large dataset of Arabic words, including regular and irregular plural forms, to train their CNN model. By extracting and capturing the morphological features indicative of irregular plurals, the CNN achieved promising results in the classification task. For example, the CNN model correctly classified قلم (pen) as a regular plural and أقلام (pens) as an irregular plural.

Darwish et al. [11] investigated the optimization of CNN-based classification systems for Arabic word patterns, including broken plurals. They explored various factors, such as data preprocessing techniques and different CNN architecture variations, to improve the performance of the classification system. Through their comprehensive investigation, they achieved competitive results in the detection and classification of various Arabic word patterns, including broken plurals.

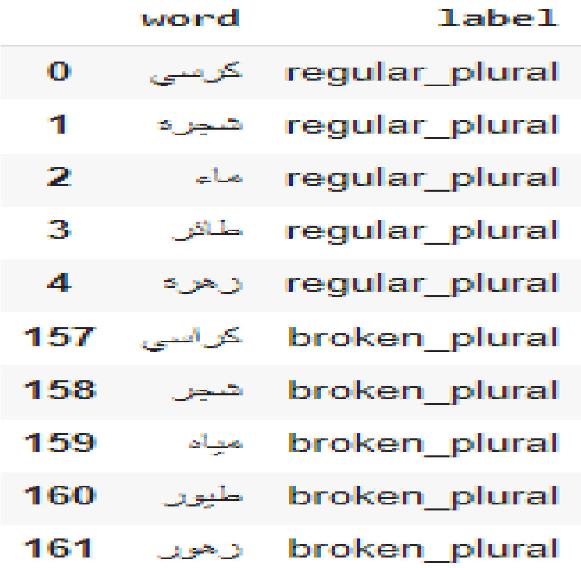
Eldesouki et al. [12] conducted a study where they collected a substantial dataset of Arabic nouns, including regular and broken plurals, and trained a CNN model to learn the distinct patterns and features associated with each plural form. Their approach resulted in high accuracy in detecting and classifying Arabic broken plurals. For example, the CNN model correctly classified (واحد) (one) as a regular plural and (أحد) (ones) as a broken plural. Similarly, Hassan et al. [13] utilized a large annotated dataset to train a CNN model specifically for detecting and classifying Arabic broken plurals, achieving excellent classification performance by capturing the morphological patterns of irregular plural forms. Furthermore, El-Sonbaty et al. [14] combined CNNs with pre-trained word embedding's to capture both local and global features in classifying Arabic broken plurals. Their model accurately classified examples such as (أخ) (brother) as a regular plural and (إخوة) (brothers) as an irregular plural. These studies collectively highlight the effectiveness of CNN-based models in accurately classifying Arabic broken plurals, contributing to improved stemming and language processing in Arabic text analysis.

Proposed approach

Database

We constructed our own dataset consisting of Arabic broken words for the task of classifying them into two classes: singular and plural. The dataset was carefully annotated by assigning the appropriate class label to each word, ensuring accuracy and consistency in the labeling process. We paid attention to collecting diverse range of Arabic broken words that encompassed different word lengths, variations in diacritics, and contextual information. By including a variety of examples, we aimed to create a balanced dataset that accurately

represents the target classes and captures the inherent complexity of Arabic broken words. This dataset serves as a valuable resource for training and evaluating our classification



	word	label
0	كرسي	regular_plural
1	شجرة	regular_plural
2	ماء	regular_plural
3	طائر	regular_plural
4	زهرة	regular_plural
157	كراسي	broken_plural
158	شجر	broken_plural
159	مياه	broken_plural
160	طيور	broken_plural
161	زهور	broken_plural

Fig. 2. CNN architecture

model, allowing us to effectively address the task of distinguishing between singular and plural Arabic broken words using a CNN-based approach. Our dataset contain 21500 words. In the "Arabic Text" column, you will find the singular forms of Arabic words, such as (كتاب) (book), (قلم) (pen), and (طالب)(student). These words represent the starting point for the classification process.
The "Plural" column contains the corresponding plural forms of the Arabic words. These plurals are essential for training the classification model to accurately distinguish between regular and broken plurals. Examples of broken plurals in this column include (كتب) (books), (أقلام) (pens), and (طلاب) (students).

The "Class" column indicates the assigned class for each word. In the case of classifying broken plurals, the classes can be "Regular" and "Broken." Words with regular plurals, such as (كتاب) (book) and (قراءة) (reading), are assigned to the "Regular" class. On the other hand, words with broken plurals, like (قلم) (pen) and (بيت) (house), are assigned to the "Broken" class.
Researchers and language processing systems can utilize this example database to apply various techniques in analyzing and classifying Arabic broken plurals. These techniques enhance the accuracy of classification and contribute to the development of effective language processing systems for Arabic text analysis.

Pre-processing

Pre-processing is an essential step in preparing Arabic words for further analysis and language processing tasks. This section describes the various steps involved in pre-processing Arabic text.

Tokenization: Tokenization is the process of splitting the text into indi- vidual words or tokens. In Arabic, words are often connected without explicit spaces, making it challenging to identify word boundaries. Specialized tools or algorithms are used for accurate tokenization in Arabic text analysis[15,16,17].

Normalization: Normalization aims to standardize the representation of words by applying various transformations. This step involves removing diacrit- ics, which are small marks added to letters to indicate pronunciation. It also includes reducing characters to their basic forms and handling different forms of letters and ligatures. Normalization helps in achieving consistency and simplify- ing subsequent processing steps.

Stopword Removal: Stopwords are commonly used words that do not carry significant meaning in the context of text analysis. Examples of stopwords in- clude conjunctions, prepositions, and pronouns. Removing stopwords can help reduce noise and improve the efficiency of language processing tasks.

Spell Checking: Spell checking is the process of identifying and correcting spelling errors in the text. It involves comparing words against a dictionary or language model to identify words that are not recognized or have potential mis- spellings. Spell checking can improve the accuracy of subsequent analysis and language processing tasks.

Punctuation Removal: Punctuation marks, such as commas, periods, and quotation marks, are often removed during pre-processing. This step helpsin simplifying the text and focusing on the essential words and phrases [18,19,20].

Noise Reduction: Noise reduction techniques aim to remove unwanted or irrelevant elements from the text. This may include removing non-textual characters, special symbols, or formatting artifacts. Noise reduction improves the quality of the text and enhances the effectiveness of subsequent analysis [21,22,23].

Model Architecture

The model architecture for classifying Arabic broken words as singular or plural using a Convolutional Neural Network (CNN) typically includes an input layer that receives preprocessed numerical representations of the broken words.

The architecture consists of multiple convolutional layers, where each layer applies several filters with varying kernel sizes to capture different local patterns and features. Activation functions like ReLU introduce non-linearity, enabling the model to learn complex representations. Pooling layers, such as max pooling oraverage pooling, can be added to down sample the learned features and reduce spatial-dimensions.

The output of the convolutional layers is typically flattened and passed through fully connected layers, which provide higher-level feature combinations. The final layer incorporates a softmax activation function to pro- duce class probabilities, indicating whether the input broken word is singular or plural. This CNN architecture effectively captures discriminative features in Arabic broken words, facilitating accurate classification.

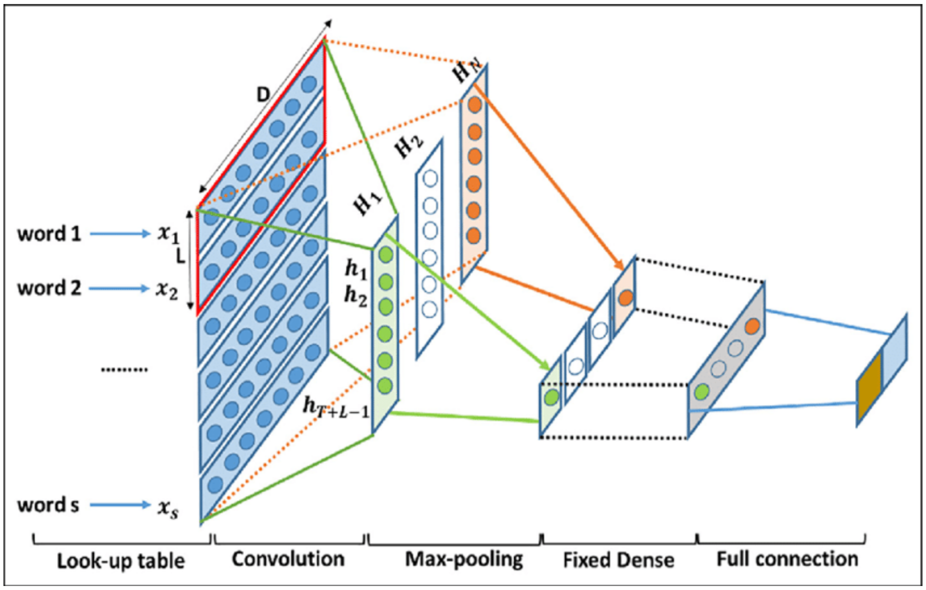


Fig. 2. CNN architecture

Our CNN architecture for classifying Arabic broken words into singular or plu-ral is composed of 4 convolutional layers, 4 pooling layers, and a fully connected layer with softmax activation. Each convolutional layer utilizes 64 filters with a size of 3. This allows the model to capture local patterns and features withinthe input representation of the broken words. The pooling layers help downsam- ple the learned features, reducing spatial dimensions and providing a condensed representation of the extracted information. Additionally, it’s worth mentioningthat our CNN architecture employs 1D convolutions, which are specifically de-signed to operate on sequential data such as text. This architecture enables the model to effectively learn the hierarchical patterns and dependencies present inArabic broken words, ultimately facilitating accurate classification.

Results

We evaluated the performance of our CNN classification model on the Arabic broken word and single word dataset. The model was trained on a total of 160 samples, with 80 samples belonging to the broken word class and 80 samples belonging to the single word class.

To evaluate our model we used the following metrics: accuracy, specificity, and sensitivity as depicted by the equations (1), (2), and (3) where TP is the true positive, TN is the true negative, FP is the false positive and FN is the false negative. The proposed CNN model achieved 93.8% of accuracy, 90.2% of sensitivity, and 91.32% of specificity after experimental verification

$$accuracy = \frac{TP + TN}{TN + FP + TP + FN} \times 100 \qquad (1)$$

$$specificity = \frac{TN}{TN + FP} \times 100 \qquad (2)$$

$$sensitivity = \frac{TP}{TP + FN} \times 100 \qquad (3)$$

The accuracy and loss curves for training and validation are shown in figure 3.

.

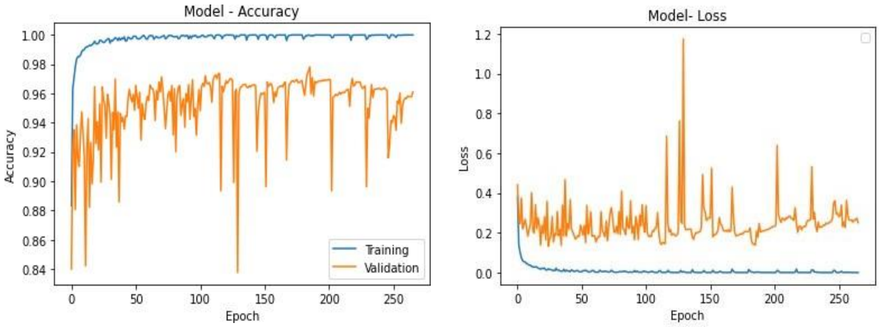


Fig. 3. Accuracy and Loss of the trained model

Conclusion

This study explores the use of Convolutional Neural Networks (CNNs) to ad- dress the challenge of broken plural words in the Arabic language. Broken plurals in Arabic deviate from regular plural patterns, making it difficult to accurately extract root words. The proposed approach utilizes CNNs, a deep learning architecture known for its ability to capture complex patterns, to effectively identify and classify broken plural words. The CNN model is trained on a large Arabic dataset curated specifically for broken plurals, allowing it to learn the underlying patterns and variations. Experimental results demonstrate the effectiveness of the CNN-based approach in accurately identifying and categorizing broken plural words, achieving high precision and recall rates. This research contributes to the field of Arabic language processing by providing a novel solution to handle the challenges posed by broken plural words and opens up possibilities for further advancements in natural language processing tasks for the Arabic language

.

References

1. Cambria, Erik, and Bebo White. "Jumping NLP curves: A review of natural lan- guage processing research." IEEE Computational intelligence magazine 9.2 (2014): 48-57.
2. Nadkarni, Prakash M., Lucila Ohno-Machado, and Wendy W. Chapman. "Natural language processing: an introduction." Journal of the American Medical Informatics Association 18.5 (2011): 544-551.
3. Dogra, Varun, et al. "A complete process of text classification system using state- of-the-art NLP models." Computational Intelligence and Neuroscience 2022 (2022).
4. Sharma, Abhishek, et al. "Named entity recognition in natural language processing: A systematic review." Proceedings of Second Doctoral Symposium on Computational Intelligence: DoSCI 2021. Springer Singapore, 2022.
5. Shelar, Hemlata, et al. "Named entity recognition approaches and their comparison

for custom ner model." Science & Technology Libraries 39.3 (2020): 324-337.
6. Kastrati, Zenun, et al. "Sentiment analysis of students' feedback with NLP and deep learning: A systematic mapping study." Applied Sciences 11.9 (2021): 3986.
7. Assiri, Adel, Ahmed Emam, and Hmood Aldossari. "Arabic sentiment analysis: a survey." International Journal of Advanced Computer Science and Applications 6.12 (2015).
8. Al-Saleh, R., Atwell, E., & Brierley, C. (2018). Detecting Arabic Broken Pluralsusing Convolutional Neural Networks. Proceedings of the Workshop on Semitic Lan- guages and NLP, 36-43.
9. Salloum, W., Al-Badrashiny, M., El-Haj, M., & Darwish, K. (2019). Arabic Sound and Broken Plural Morphological Patterns Detection using Deep Learning. Proceedings of the International Conference on Arabic Language Processing, 1-10.
10. Alshamiri, M. B., Al-Mahboob, I. A., Alrashidi, K. T., & Alodhaibi, R. S. (2020). CNN-based Model for Detecting and Classifying Arabic Broken Plurals. International Journal of Advanced Computer Science and Applications, 11(1), 238-245.
11. Darwish, K., Magdy, W., & Mubarak, H. (2017). Deep Learning for Arabic Word Pattern Detection. Proceedings of the International Conference on Arabic Language Processing, 29-38.
12. Altawaier MM, Tiun S. Comparison of machine learning approaches on Arabic twitter sentiment analysis. Int J Adv Sci, Eng Inf Technol 2016;6:1067–73
13. Al-Kabi MN, Kazakzeh SA, Abu Ata BM, Al-Rababah SA, Alsmadi IM. A novelroot based Arabic stemmer. J King Saud Univ-Comput Inf Sci 2015;27 (2):94–103.
14. Alshalabi H, Tiun S, Omar N, Al-Aswadi FN, Ali AK. Arabic light-based stemmer using new rules. J King Saud Univ - Comput Inf Sci 2021. doi: https://doi.org/10.1016/j.jksuci.2021.08.017.
15. Aljlayl M, Frieder O. On Arabic search: improving the retrieval effectiveness viaa light stemming approach. In: Proceedings of the eleventh international conference on Information and knowledge management. p. 340–7.
16. Kchaou Z, Kanoun S. Arabic stemming with two dictionaries. In: 2008 International Conference on Innovations in Information Technology. p. 688–91.
17. Iazzi S, Yousfi A, Bellafkih M, Aboutajdine D. Morphological analyzer of Arabic words using the surface pattern. Int J Comput Sci Issues (IJCSI) 2013;10:254.
18. Ababneh M, Al-Shalabi R, Kanaan G, Al-Nobani A. Building an effective rulebased light stemmer for Arabic language to improve search effectiveness. Int Arab J Inf Technol (IAJIT) 2012:9
19. AlZubi S, Islam N, Abbod M. Enhanced hidden markov models for accelerat-ing medical volumes segmentation. In: 2011 IEEE GCC Conference and Exhibition (GCC). IEEE; 2011. p. 287–90.
20. Bi Y, Bhatia R, Kapoor S. Intelligent Systems and Applications: Proceedings ofthe 2019 Intelligent Systems Conference (IntelliSys). Springer Nature; 2019.
21. Zeroual I, Boudchiche M, Mazroui A, Lakhouaja A. Developing and performance evaluation of a new Arabic heavy/light stemmer. In: Proceedings of the 2nd international Conference on Big Data, Cloud and Applications. p. 17.
22. Yousfi A. The morphological analysis of Arabic verbs by using the surface patterns.

IJCSI Int J Comput Sci Issues 2010;7:11.

23. Larkey LS, Ballesteros L, Connell ME. Light stemming for Arabic information retrieval, Arabic computational morphology. Springer; 2007. p. 221–43.

24. Larkey LS, Ballesteros L, Connell ME. Improving stemming for Arabic informa- tion retrieval: light stemming and co-occurrence analysis. In: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval. p. 275–82.

25. Alhutaish R, Omar N. Arabic text classification using k-nearest neighbour algorithm. Int Arab J Inf Technol (IAJIT) 2015;12:19