Uluslararası İleri Doğa Bilimleri ve Mühendislik Araştırmaları Dergisi
Sayı 7, S. 151-155, 6, 2023

Araştırma Makalesi

IJANSER
https://as-proceeding.com/index.php/ijanser
ISSN: 2980-0811

International Journal of Advanced Natural Sciences and Engineering Researches
Volume 7, pp. 151-155, 6, 2023

Research Article

DEFENSE AGAINST WHITE BOX ADVERSARIAL ATTACKS IN ARABIC NATURAL LANGUAGE PROCESSING (ANLP)

MOKHTAR ALSHEKH1[*], KÖKSAL ERENTÜRK[2]

[1] Ataturk University, Department of Computer Engineering, Erzurum, Türkiye
[2] Ataturk University, Department of Computer Engineering, Erzurum, Türkiye

[*]*(mokhtaralshekh94@gmail.com)*

(Received: 05 July 2023, Accepted: 24 July 2023)

(5th International Conference on Applied Engineering and Natural Sciences ICAENS 2023, July 10 - 12, 2023)

ATIF/REFERENCE: Alshekhı, M. & Erentürk, K. (2023). DEFENSE AGAINST WHITE BOX ADVERSARIAL ATTACKS IN ARABIC NATURAL LANGUAGE PROCESSING (ANLP). *International Journal of Advanced Natural Sciences and Engineering Researches*, 7(6), 151-155.

Abstract – Adversarial attacks are among the biggest threats that affect the accuracy of classifiers in machine learning systems. This type of attacks tricks the classification model and make it perform false predictions by providing noised data that only human can detect that noise. The risk of attacks is high in natural language processing applications because most of the data collected in this case is taken from social networking sites that do not impose any restrictions on users when writing comments, which allows the attack to be created (either intentionally or unintentionally) easily and simply affecting the level of accuracy of the model. In this paper, The MLP model was used for the sentiment analysis of the texts taken from the tweets, the effect of applying a white-box adversarial attack on this classifier was studied and a technique was proposed to protect it from the attack.
After applying the proposed methodology, we found that the adversarial attack decreases the accuracy of the classifier from 55.17% to 11.11%, and after applying the proposed defense technique, this contributed to an increase in the accuracy of the classifier up to 77.77%, and therefore the proposed plan can be adopted in the face of the adversarial attack. Attacker determines their targets strategically and deliberately depend on vulnerabilities they have ascertained. Organization and individuals mostly try to protect themselves from one occurrence or type on an attack. Still, they have to acknowledge that the attacker may easily move focus to advanced uncovered vulnerabilities. Even if someone successfully tackles several attacks, risks remain, and the need to face threats will happen for the predictable future.

Keywords –Adversarial Attack, Multi-Layer Perceptron, Sentiment Analysis, Text Classification, Social Media, Natural Language Process.

I. INTRODUCTION.

Natural Language Processing is one of the most promising fields of research. This field contains many sub-headings that fall under it, including sentiment analysis. Sentiment analysis is a field of research used to analyze people's feelings and opinions according to data about them, such as their comments on social media, their conversations, and their spoken and written words. (M. Farhadloo, 2018) (A. Mandal, 2018)

With the development of data mining techniques and the spread of social media, this field has witnessed wide development, but it still suffers from many challenges. One of them is the use of Arabic, which is considered one of the languages that are very rich in vocabulary and synonyms (D. M. E. D. M. Hussein, 2018).

Such errors that affect the accuracy of the model are caused by the grammatical richness of the language, the use of irony, or the diversity of opinions. (Xiaoting , Chen, & Wu, 2021) All of these errors happen unintentionally, but another type of error occurs because of the adversarial attacks that are performed by attackers to trick the model through wrong (obfuscated) entries. (Diab, Zitouni, & Habash, 2017)

The types of adversarial attacks can be divided into several categories according to the attacker's ability and access to data. The first is the "white box" attack, in which the adversary can access all information about the target neural network, including its architecture, parameters, gradients, etc. An attacker can take full advantage of network information if he is careful. (Ashrafiamiri, 2021)

The second type is the "black box" attack, and in this attack, in which the attackers cannot access the structure of the neural network used. They can only enter the inputs and see the resulting outputs. (Alsmadi I. A.-N.-H., 2022)

The last type is a "semi-white (gray) box" attack, the attacker trains a model to generate attack instances as in a white box attack, after the model is created, the attacker doesn't need the target model anymore, and the attacker at this stage only uses inputs and outputs (black box). (Bose, 2018)

The emergence of white-box attacks and their great ability to influence the accuracy of the results of the model was the main motive of this research. Also, this research is not limited to defending against an attack but also focuses on ways to create the attack and study the negative impact of the attack in addition to suggesting a way to defend against the attack.

II. Materials and Method

A. TF-IDF

tf-idf algorithm is used to determine the importance of a specific word (term) in a specific document within a specific text collection, and the increase in the importance of that word is equivalent to the number of times it appears in the specific document, and it must also be proportional to the number of times the word appears in the entire specified text. (Das, Kamalanathan, & Alphonse, 2021)

The word frequency is the number of times the word appears in the document, while the inverse document frequency is the number of times the word appears throughout the text. These two parameters are the most important parameters in calculating word vectors.

It was mentioned that the frequency of the word is the number of times the word appears in the specific document relative to the number of words in it, and each document has a different percentage of occurrence, and it is defined by the following equation:

$$tf_{ij} = \frac{n_{ij}}{\sum_k n_{ij}} \quad (1)$$

Known that:

tf_{ij}: number of times the word t_i appears in the document D_j.

The frequency of the reverse document determines the number of word appearances in the entire textual collection and was calculated using the following equation:

$$idf(w) = \log\log\left(\frac{N}{df_t}\right) \quad (2)$$

Known that:

N: number of all words

W: word weight

The final equation to calculate tf-idf is as the following:

$$tf - idf = tf_{ij} * \log\left(\frac{N}{df_i}\right) \quad (3)$$

The tf-idf algorithm is one of the most important algorithms that are used in extracting features from texts, and this algorithm was specifically chosen in this research for its ease of application and ease of calculations (the complexity of the code is less), and that it is a common choice for most researchers makes it possible to compare this research with other research is fair to standardize the feature extraction technique.

B. Multi-layer perceptron

The Multi-layer Perceptron MLP network is one of the most important types of artificial neural networks, due to its strength and ability to solve many linear and nonlinear classification problems,

where the simple perceptron network with one neuron is used to solve the linear problems, while the multilayer perceptron network is used to solve the nonlinear problems that it requires more than one line to separate the two classes.

The neuron is the main structural unit in MLP, where the MLP consists of a group of neurons connected by links bearing different weights.

These neurons are arranged in layers, where the artificial neural network contains an input layer (the number of neurons equals the number of inputs) and an output layer (the number of neurons equals the number of outputs), and a set of hidden layers containing different numbers of neurons that are specified as needed, as given in Figure 1.

The output of each neuron in the artificial neural network is calculated by taking the sum of the input values' multipliers with the weights in the connections corresponding to each weight and then adding the bias as the following: (Kriesel, 2011)

$$x = \sum_i (weight_i * input_i) + bias \qquad (4)$$

The output of the neuron is the value generated by the activation function as:

$$y = activation\ function(x) \qquad (5)$$

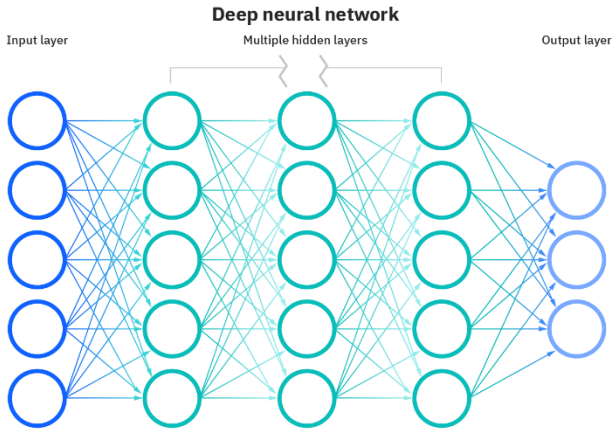

Fig 1. MLP Architecture

C. *Arabic Sentiment Twitter Corpus Dataset*

This is a dataset that researcher collected it to provide an Arabic sentiment dataset the real use is investigate deep learning approaches for Arabic sentiment analysis. (SAAD, 2020)

This dataset we collected in April 2019. The dataset is formed from 58K Arabic tweets the researcher divided those tweets into 47 tweets for training and another 11k for testing, the dataset annotated in positive and negative labels. The dataset is balanced and collected using positive and negative emoji's lexicon.

Results

Python programming language was used with a set of libraries, such as SKlearn, Pandas, NLTK, and others, to build an MLP model and train it according to the Arabic Sentiment Twitter Corpus dataset, after which the attack was performed and a method of protection was suggested.

$$Accuracy = \frac{true\ prediction d}{true\ predictions + false\ predictions} = \frac{t_p + t_n}{t_p + t_n + f_p + f_n} \qquad (6)$$

```
1 y_pred=clf.predict(X_test_tf)
```

```
1 from sklearn.metrics import accuracy_score
2 score=accuracy_score(test_data['class'], y_pred)
```

```
1 score
```

```
0.541871921182266
```

Fig 2. Accuracy Results For MLP Model Without Attack

For calculating the accuracy of the model, the accuracy score function was used, and the model accuracy without attack equals 54.17 % as shown in Figure 2.

The attack we will perform is a white box attack in which the attacker has knowledge about the mode and the data used in training and precisely we will perform an evasion attack in which from our knowledge of the model we will find weak points to perform the attack.

For example:

"أنا احب جاري"

After attack:

"انا احب من هو جاري"

The phrase is the same for the human mind, but it was different for machines. So, after making the attack, many experiments were conducted to determine the effect of the attack on the model. Figure 4 shows three cases from the experiments, as it can be noticed that the average accuracy for the model under attack equals 11.1% only.

After conducting numerous experiments and analysing the texts and observing the results of the attack and its impact and comparing them with the texts themselves, it was noted in this research that there are a group of words that do not have a real effect on the meaning for humans, but these words caused confusion on the results of the model, and therefore these words can be considered a weaknesses point of the model. Weakness points are divided into three sections: the first section is the word suffixes, the second is the antecedents and the third is in the middle of the speech.

III. DISCUSSION

The model was built and its accuracy was studied in three different cases: the first case was without applying an attack, the second case was when applying an attack and without investigation or defense means, and the third case was by applying a defensive method against attack, and the table 1. shows the accuracy of the model in the three cases, the classifier achieved an accuracy level not exceeding 55% without applying an attack or defensive method, and the accuracy level decreased to 11.11% when applying the attack, while the accuracy level reached 77.77% when applying a defensive technique, which is higher than the level of accuracy of the basic classifier without attack even, This indicates the possibility of an unintentional attack in the test dataset itself.

It can also be noted that the application of defense technology contributed to increasing the accuracy of the model in classification in all cases, even though the accuracy of the model reached 100% in some cases despite the application of the attack.

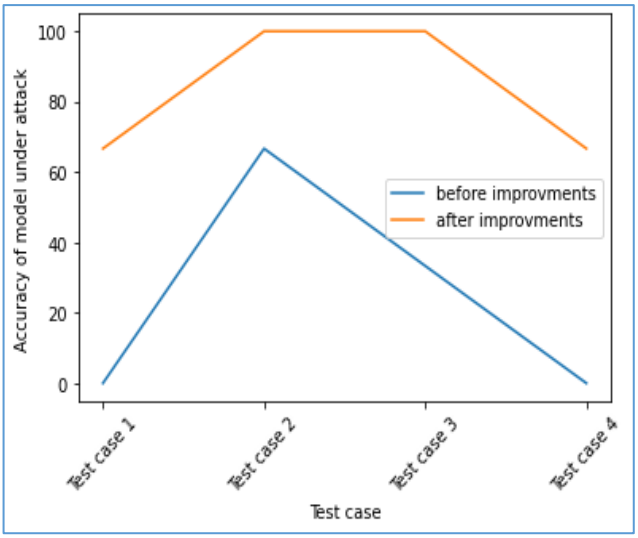

FIG 3. ACCURACY OF MODEL UNDER ATTACK

IV. CONCLUSION

After completing the study and the practical experiment and taking the results that were presented in this chapter, several conclusions were reached regarding the idea of the research. First, it can be said that the MLP model is considered good for Arabic text classification issues, as the model achieved an acceptable level of accuracy.

On the other hand, a white box adversarial attack is a serious threat to text classification systems, as the attack greatly affected classifier accuracy. Also, Speech antecedents and suffixes (such as conjunctions, definite articles, and other linguistic devices used in the Arabic language) are considered a threat to the accuracy of the compiler, who regards them as a noise signal that affects his ability to classify texts correctly.

In the end, it can be said that this research is an important contribution to the field of Arabic Natural Language Processing, and the field is still open for researchers in future studies to make more contributions, such as experimenting with new classification models different from MLP or experimenting with new datasets in Arabic. It is also important to experiment with new types of attacks and examine their impact on model accuracy.

REFERENCES

[1] Barba, J. G. (2021). Attention-based approaches for Text Analytics in Social Media and Automatic Summarization. Valencia.

[2] Giordano, E. (2021). Transferability of Adversarial Attacks: Main Influencing Factors.

[3] A. Mandal, M. K. (2018). E-R E-R E-R E-R, (Vol. 1). 2018.

[4] Abdul Kader, A. M. (2020). Adversarial Attacks on Neural Networks & Defense for it. NTNU.

[5] Alsmadi, I. A.-N.-H. (2022). Adversarial Machine Learning in Text Processing: A Literature Survey.

[6] Alsmadi, I., Ahmad, K., Nazzal, M., Alam, F., Al-Fuqaha, A., Khreishah, A., & Algosaibi, A. (2021). Adversarial Attacks and Defenses for Social Network Text Processing Applications: Techniques, Challenges, and Future Research Directions.

[7] Anantaneni, J. C. (2021). Evaluating Robustness of a CNN Architecture introduced to the Adversarial Attacks. Karlskrona, Sweden: Faculty of Computing, Blekinge Institute of Technology.

[8] Ashrafiamiri, M. (2021). An Adversarial Defense Methodology for Neural Networks based on. IRVINE: UNIVERSITY OF CALIFORNIA.

[9] Bose, A. (2018). Adversarial Attacks on Face Detectors using Neural Net based Constrained Optimization. University of Toronto.

[10] Chen, Y. (2015). Practical Adversarial Attacks Against Black Box Speech Recognition Systems and Devices. Melbourne, Florida: University of Electronic Science and Technology of China.

[11] D. M. E. D. M. Hussein, J. (2018). A survey on sentiment analysis challenges, (Vol. 30). King Saud Univ. - Eng. Sci. doi:10.1016/j.jksues.2016.04.002.

[12] Das, M., Kamalanathan, S., & Alphonse, P. (2021). A Comparative Study on TF-IDF Feature Weighting Method and its Analysis using Unstructured Dataset. Tamil Nadu, India: 5th International Conference on Computational Linguistics and Intelligent Systems.

[13] Diab, M., Zitouni, I., & Habash, N. (2017). NLP for Arabic and Related Languages (Vol. 58).

[14] F. Neri, C. A. (2014). Sentiment analysis on social media . Proc. 2012 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. . doi:10.1109/ASONAM.2012.164.

[15] G. Beigi, X. H. (2016). An Overview of Sentiment Analysis in Social Media and Its Applications in Disaster Relief BT - Sentiment Analysis and Ontology Engineering: An Environment of Computational Intelligence.

[16] Hayat, M. K. (2019). Towards Deep Learning Prospects: Insights for Social Media Analytics. (Vol. 7). IEEE Access. Retrieved from https://doi.org/10.1109/ACCESS.2019.2905101

[17] Kalaria, D. (2022). Btech thesis report on adversarial attack detection and purification of adverserially attacked images. Retrieved from http://arxiv.org/abs/2205.07859

[18] Khalil, A. (2022). Developing a Robust Def eloping a Robust Defensive System Against First Or e System Against First Order Adversarial Attacks Using Siamese Neural Network Methods. Electronic Theses and Dissertations. Retrieved from https://scholar.uwindsor.ca/etd/?utm_source=scholar.uwindsor.ca%2Fetd%2F8703&utm_medium=PDF&utm_campaign=PDFCoverPages

[19] Kong, Z. X. (2021). A Survey on Adversarial Attack in the Age of Artificial Intelligence. Wireless Communications and Mobile Computing. Retrieved from https://doi.org/10.1155/2021/4907754

[20] Kriesel, D. (2011). A Brief Introduction to Neural Networks. Retrieved from dkriesel. com: http://www.dkriesel.com/en/science/neural_networks

[21] LI, W. (2019). DEFENDING VISUAL ADVERSARIAL EXAMPLES WITH SMOOTHOUT REGULARIZATION. New Brunswick, New Jersey: Rutgers, The State University of New Jersey.

[22] M. Farhadloo, E. R. (2018, August). Fundamentals of sentiment analysis and its applications, (Vol. 639). (S. C. Intell., Ed.) doi:10.1007/978-3-319-30319-2_1.

[23] Moradi, M., & Samwald, M. (2019). Improving the robustness and accuracy of biomedical language models through adversarial training.

[24] SAAD, M. (2020). Arabic Sentiment Twitter Corpus. Retrieved from Kaggle: https://www.kaggle.com/datasets/mksaad/arabic-sentiment-twitter-corpus

[25] T. Luo, S. C. (2013). Trust-based Collective View Prediction Trust. doi:10.1007/978-1-4614-7202-5.

[26] Tsai, A. Y.-T., Yang, T., & Chen, E. (2019). Adversarial Attack on Sentiment Classification. Florence, Italy.

[27] Tsunoda, Y. (2021). On an adversarial example attack by partially monitoring and tampering input features to machine learning models; its possibilities and countermeasures. Osaka University.

[28] Xiaoting , L., Chen, L., & Wu, D. (2021). Turning Attacks into Protection: Social Media Privacy Protection Using Adversarial Attacks. SIAM.

[29] Xie, Y. W.-Y. (2022). A Word is Worth A Thousand Dollars: Adversarial Attack on Tweets Fools Stock Prediction. Conference of the North American Chapter of the Association for Computational Linguistics. Retrieved from https://github.com/yonxie/

[30] Xu, H. M. (2020). Adversarial Attacks and Defenses in Images, Graphs and Text: A Review. In International Journal of Automation and Computing (Vol. 17). Chinese Academy of Sciences.

[31] Y. Sani, A. M. (2014). An overview of neural networks use in anomaly intrusion detection systems . Retrieved from IEEE Student Conf. Res. Dev.

[32] Zago, J. G. (2021). Defense Methods for Convolutional Neural Networks Against Adversarial Attacks. Florianópolis: FEDERAL UNIVERSITY OF SANTA CATARINA.

[33] ZHANG, W. E., SHENG, Q. Z., & ALHAZMI, A. (2019). Adversarial Attacks on Deep Learning Models in Natural Language Processing: A Survey (Vol. 1). Association for Computing Machinery. Retrieved from https://doi.org/10.1145/nnnnnnn.nnnnnnn