AAST-NLP at ArAIEval Shared Task: Tackling Persuasion Technique and Disinformation Detection using Pre-Trained Language Models On Imbalanced Datasets

Ahmed El-Sayed[1]*, **Omar Nasr**[1]*, **Nour El-din El-madany**[2]
Intelligent System Lab
[1]Department of Computer Engineering, Arab Academy for Science and Technology
[2]Department of Electronics And Communication, Arab Academy for Science and Technology
{ahmedelsayedhabashy,omarnasr5206}@gmail.com,nourelmadany@aast.edu

Abstract

This paper presents the pipeline developed by the AAST-NLP team to address both the persuasion technique detection and disinformation detection shared tasks. The proposed system for all the tasks' sub-tasks consisted of preprocessing the data and finetuning AraBERT on the given datasets, in addition to several procedures performed for each subtask to adapt to the problems faced in it. The previously described system was used in addition to Dice loss as the loss function for sub-task 1A, which consisted of a binary classification problem. In that sub-task, the system came in eleventh place. We trained AraBERT for task 1B, which was a multi-label problem with 24 distinct labels, using binary cross-entropy to train a classifier for each label. On that sub-task, the system came in third place. We utilised AraBERT with Dice loss on both subtasks 2A and 2B, ranking second and third among the proposed models for the respective subtasks.

1 Introduction

Social media has become part and parcel of our everyday lives and a main source of information for every individual. Unfortunately, due to the nature of social media, the spread of disinformation (Alam et al., 2022a) is very relevant and causes major troubles. For example back in the COVID-19 pandemic, some researchers coined the term "infodemic" to describe the spread of false information among people during that period (Geldsetzer, 2020). Many researchers have proposed their systems to fight the spread of disinformation on social media platforms, powered by recent advances in NLP and the introduction of Large Language models including BERT (Devlin et al., 2019) which revolutionized NLP and was adapted to many tasks.

Persuasion is a type of social interaction that attempts to influence and change attitudes in an atmosphere of free choice (Perloff, 2017). Persuasion techniques are incredibly important linguistic techniques that can have massive effects on different fields and industries. An example of this is the usage of these techniques in advertising campaigns, which can lead to impressive results when it comes to changing customers attitudes and receiving their responses without imposing on them (Romanova and Smirnova, 2019). This paper tackles the various systems our team attempted for the ArAIEval 2023 shared tasks (ove). The first step was to look at some of the earlier publications from WANLP 2022 (Alam et al., 2022b), which provided a number of crucial insights that served as a foundation for our work. Related work includes the system presented in (Mubarak et al., 2023) for the identification of disinformation through samples, combined with many additional significant results as well as fine-grained disinformation labels from those samples. The following sections of the paper comprise a data section which describes the data sources and preprocessing methods applied to the data. A system section describing the pipeline, a results section, a discussion and a summary.

2 Data

In this section, we will describe the data sources and the preprocessing methods that we applied to prepare the data. We will also provide some descriptive statistics and visualizations of the data to give an overview of its characteristics and distribution.

2.1 Data Description

2.1.1 Persuasion Technique Detection

Task 1 consists of two subtasks, namely subtask 1A and 1B. The first is to determine whether the

*. equally contributed

Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP 2023), pages 565–569
December 7, 2023

tweets and paragraphs contain any persuasive techniques. The second sub-task expanded on the first by identifying the various persuasive strategies that were found in those samples.

Dataset	Train	Dev	Test
Texts	2427	259	503

Table 1: Data distribution for task 1.

The training dataset includes 2427 samples labelled as True or False, with a distribution of 1918 to 509, respectively. This indicates that the ratio of true to false cases is roughly 65.8% to 34.2%, as illustrated in Figure 1, demonstrating that the dataset had a class imbalance. This percentage was matched in the development data, which had a distribution of 202 to 57 respectively.

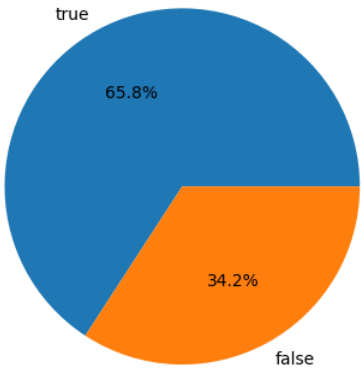

Figure 1: Data distribution for the training data for subtask 1A.

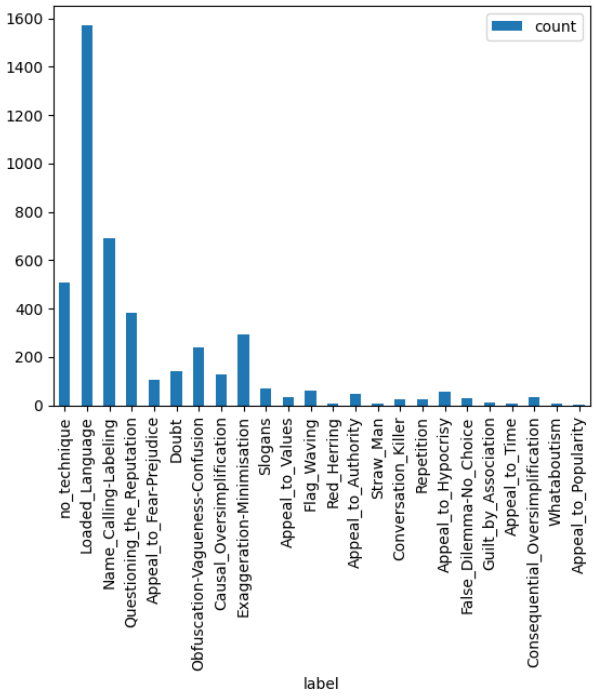

Figure 2: Subtask 1B Training Class Distribution.

The data for subtask 1B consists of samples labelled from 24 different class labels which represents the different persuasive techniques. Some of the common techniques found are "Loaded Language", "Name calling/labelling" and "Questioning the Reputation". We hypothesize that there are underlying dependencies between the techniques and correlations between different combinations which makes it a very interesting task and worthy of further exploration. As shown in Figure 2 the class distribution is severely unbalanced with underrepresented classes including "Appeal to Popularity", "Whataboutism" and several others.

2.1.2 Disinformation Detection

The objective of Task 2 comprises two subtasks. The first is classifying the samples into information and disinformation. The second involves classifying the given samples into one of four sub-classes: HS, OFF, Spam, and Rumour.

Dataset	Train	Dev	Test
Tweets	14147	2115	3729

Table 2: Data distribution for task 2A.

Dataset	Train	Dev	Test
Tweets	2648	396	876

Table 3: Data distribution for task 2B.

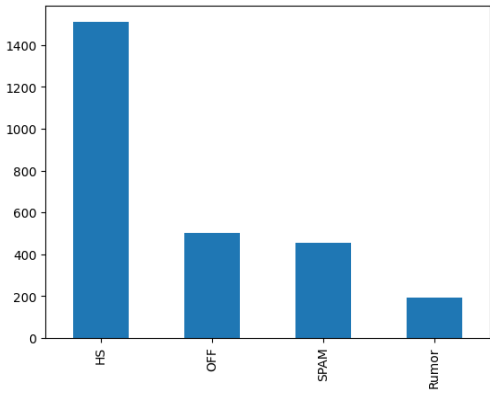

Figure 3: Data distribution for subtask 2B.

Task 2 suffers from a class imbalance problem. Figure 3 demonstrates that the rumour class is significantly underrepresented, but the HS class is significantly overrepresented. The validation set is distributed in a similar manner. Subtask 1A has a similar problem, with a distribution of 11419 to 2656 for no-disinformation to disinformation.

The validation set has a similar problem, with a distribution of 1718 to 397 for no-disinformation to disinformation.

2.2 Preprocessing

The preprocessing procedure of our data took the following steps.

- Removing Arabic stop words.
- Removing tweet related tags like LINK , RT, [مستخدم] and [فيديو].
- Applying AraBERT preprocessor, removing tashkeel, tatweel and emojis.
- Removing '_' as some tweets were ambiguously written and formatted with '_' between each letter.

3 System Description

3.1 Model Description

Our initial experiments on the conducted on the development data consisted of comparing several BERT-based models to choose from to build upon, we experimented on AraBERT (Antoun et al., 2021), MarBERT (Muhammad Abdul-Mageed and Nagoudi), ArBERT and bert-base-arabic (Safaya et al., 2020), AraBERT outperformed its peers on the 4 subtasks and was the one chosen to further experiment upon.

3.2 Addressing Class Imbalance

After inspecting the data, it was clear that one of the problems that would hinder our experiments would be the severe case of class imbalance that the provided datasets were suffering from. We experiment with three methods to mitigate the effects of imbalances in datasets. The following sections give details of each method and its corresponding effects on the results.

3.2.1 Re-Sampling

Re-sampling is the process of increasing the importance of minority classes by altering the distribution of the training datasets (Kraiem et al.). Random under sampling (RUS) consists of randomly removing datapoints from the majority class. Random oversampling (ROS) consists of randomly duplicating minority class instances. Both ROS and RUS were used to offset the data imbalance in the dataset.

3.2.2 Data Augmentation

Synthetic data was created using the NLPAUG package[1]. Contextual Word Embeddings Augmentation was used based on AraBERT, and the dataset distributions were altered to increase the importance of classes underepresented in our datasets, but one important remark was that the data created was noisy and required filtering. For example the sample shown below had it's meaning completely changed from the original sentence.

Original Data:

احساسي بقول لي ماف كورونا حتنتشر البلد.

Synthetic Data:

احساسي بقول انه فيروسات كورونا حتنتشر علي هاي البلد.

The augmented data was filtered and revised manually to check if the meaning of the new synthetic sentence matches the original sentence. Synthetic data created using this method resulted in a huge decrease in our micro-F1 score.

3.2.3 Custom Loss Functions

Several loss functions were experimented upon, initially we used Weighted Cross-Entropy loss (Ozdemir and Sonmez, 2020) for our subtasks with weights calculated via scikit (Pedregosa et al., 2011) class weight function it resulted in a slight improvement on the binary classification tasks. Although the adaptation of focal loss appeared interesting at first, it was not robust in handling the imbalance difficulties and led to overfitting. Ultimately, we conducted an experiment using Dice Loss (Li et al., 2019), a customized loss function tailored to NLP tasks based on the Sørensen–Dice coefficient (Li et al., 2019).

$$Diceloss(p, y) = 1 - \frac{2 * \sum_1^t p_i * y_i + smooth}{\sum_1^t p_i + \sum_1^t y_i + smooth} \quad (1)$$

This particular loss function led to an improvement in the F1 score for each of the corresponding tasks.

3.3 Experiment Settings

The training procedure was conducted using the Google Colab platform for training our pipeline, which has 12.68 GB of RAM, a 14.75

1. https://github.com/makcedward/nlpaug

GB NVIDIA Tesla T4 GPU, and Python language. We used ktrain's (Maiya, 2020) autofit, which applies a triangular learning rate policy (Smith, 2015). The learning rate was determined via the lr_plot function, which experiments with a range of learning rates and suggests multiple possible learning rates. The parameters set for our experiment are mentioned in the table below.

Parameter	Value
Epochs	30
Learning Rate	1e-5
Batch Size	16
Max Length	128
Optimizer	AdamW
Early Stopping Patience	5
Reduce on pleateau	2
Dice loss smoothing	1e-6

Table 4: Training parameters.

Modifications were made to adapt to the task requiremets including changing the loss function to Dice loss for binary and multiclass classification task with smoothing set to 1e-6. For the multilabel task 1B, we used a binary cross entropy loss to train 24 different classifiers each to one of the labels found in the provided dataset.

4 Results

Task	Validation	Test
1A	0.5405	0.4771
1B	0.0938	0.0868
2A	0.5173	0.5154
2B	0.2191	0.2603

Table 5: Baseline micro-f1 scores for all subtasks.

Table 5 presents the random baseline micro-f1 scores on all the respective subtasks. These micro-f1 scores were obtained through the official website of the shared task. These baselines provide a point of reference for the obtained results. The system consistently outperformed these baselines by a significant margin throughout the development process and the outline of the results of the given system is presented in the rest of this section.

Task	Training	Validation	Test
1A	0.9782	0.8301	0.7237
1B	0.8101	0.6295	0.5522
2A	0.9414	0.9031	0.9043
2B	0.9782	0.8301	0.8253

Table 6: Achieved micro-f1 scores for all subtasks.

The micro-f1 scores of the previously mentioned system, which uses AraBERT paired with task specific loss function; Dice loss for the first part of the persuasion technique detection problem and both tasks of the disinformation detection problem, and Binary Cross Entropy for the second task of persuasion technique detection labeling, are shown in Table 6. Micro-f1 was chosen as the competition's evaluation metric, and testing results were obtained once the evaluation process was completed. The results of the persuasion technique detection ranked 11th and 3rd, respectively, while the results of the misinformation detection tasks ranked 2nd and 3rd, respectively.

5 Discussion

A diverse set of limitations were encountered during the development of the aforementioned systems. Another drawback stemmed from the underlying dependencies among task 1B labels, as attempting a direct approach did not lead to optimal outcomes. The subjective labelling of tasks 1B and 2B made it difficult to leverage external data sources to further train our model. One strategy worth highlighting is the use of a CNN-BILSTM and ARABERT hybrid model (Hengle et al., 2021). However, this did not produce satisfactory results since the model appeared to overfit the training instances.With few modifications, this strategy may be viable. Furthermore, the unexpected decline in task 1A's performance necessitates further investigation and experimentation to determine the cause.

6 Summary

The proposed system based on AraBERT was detailed, and the experiments conducted were all addressed. The adaptation of dice loss boosted our performance on all of the tasks and partially addressed the issue of class imbalance yet there is a huge room for improvement. There are other intriguing future directions, like the development of a data augmentation package that supports differ-

ent data augmentation techniques. Furthering the solution to the issue of class imbalance is another intriguing path. Last but not least, the problem of the underlying dependencies and ways of tackling multilabel tasks should be inspected, and new methods should be investigated and developed in the near future. We intend to invesigate these various approaches in detail in the future since we believe there is still room for improvement in finetuning as well as experimenting with other approaches such as different hybrid model architectures and different data augmentation methods. In the future, we plan to thoroughly explore these diverse approaches because we are convinced that there is further potential for enhancing fine-tuning. This includes experimenting with alternative hybrid model architectures and various data augmentation techniques.

References

Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimiter Dimitrov, Giovanni Da San Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. 2022a. A survey on multimodal disinformation detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6625–6643, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Firoj Alam, Hamdy Mubarak, Wajdi Zaghouani, Giovanni Da San Martino, and Preslav Nakov. 2022b. Overview of the WANLP 2022 shared task on propaganda detection in Arabic. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 108–118, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2021. Arabert: Transformer-based model for arabic language understanding.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Pascal Geldsetzer. 2020. Knowledge and Perceptions of COVID-19 among the general public in the United States and the United Kingdom: a cross-sectional online survey. *Annals of Internal Medicine*, 173(2):157–160.

Amey Hengle, Atharva Kshirsagar, Shaily Desai, and Manisha Marathe. 2021. Combining context-free and contextualized representations for Arabic sarcasm detection and sentiment identification. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 357–363, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Mohamed S. Kraiem, F. Sánchez, and María N. Moreno García. Selecting the suitable resampling strategy for imbalanced data classification regarding dataset properties. an approach based on association models. (18):8546.

Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. 2019. Dice Loss for Data-imbalanced NLP Tasks. *arXiv (Cornell University)*.

Arun S. Maiya. 2020. ktrain: A low-code library for augmented machine learning. *arXiv preprint arXiv:2004.10703*.

Hamdy Mubarak, Samir Abdaljalil, Azza Nassar, and Firoj Alam. 2023. Detecting and reasoning of deleted tweets before they are posted. *arXiv preprint arXiv:2305.04927*.

AbdelRahim A. Elmadany Muhammad Abdul-Mageed and El Moatez Billah Nagoudi. Arbert & marbert: Deep bidirectional transformers for arabic.

Ozgur Ozdemir and Elena Battini Sonmez. 2020. Weighted Cross-Entropy for Unbalanced Data with Application on COVID X-ray images. *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Richard Perloff. 2017. *The dynamics of persuasion: Communication and attitudes in the 21st century*.

Irina Romanova and Irina Smirnova. 2019. Persuasive techniques in advertising. *Training Language and Culture*, 3:55–70.

Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. 2020. KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2054–2059, Barcelona (online). International Committee for Computational Linguistics.

Lauren Smith. 2015. Cyclical learning rates for training neural networks. *arXiv (Cornell University)*.