International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 05 | May 2022 www.irjet.net p-ISSN: 2395-0072

AUTOMATIC DETECTION AND LANGUAGE IDENTIFICATION OF MULTILINGUAL DOCUMENTS

Buggidi Harshini, Matteddula Krishna Reddy, Mangalamadaka Mahema, Shaik Moula Bee, Kandula Nagalikitha, S. Nagaraju

---

ABSTRACT

Hostile correspondences have attacked web-based entertainment content. Perhaps the best answer for adapt to this issue is utilizing computational methods to segregate hostile substance. In addition, online entertainment clients are from phonetically various networks. This study intends to handle the Multilingual Offensive Language Detection (MOLD) task utilizing move learning models and the adjusting stage. We propose a successful methodology in view of the Bidirectional Encoder Representations from Transformers (BERT) that has shown extraordinary potential in catching the semantics and relevant data inside messages. The proposed framework comprises of a few phases: (1) Preprocessing, (2) Text portrayal utilizing BERT models, and (3) Classification into two classes: Offensive and non-hostile. To deal with multilingualism, we investigate various strategies, for example, the joint-multilingual and interpretation based ones. The main comprises in creating one order framework for various dialects, and the second includes the interpretation stage to change all texts into one general language then arrange them. We lead a few examinations on a bilingual dataset extricated from the Semi-managed Offensive Language Identification Dataset (SOLID). The exploratory discoveries show that the interpretation based strategy related to Arabic BERT (AraBERT) accomplishes more than 93% and 91% as far as F1-score and precision, individually.

1. INTRODUCTION

In the latest 10 years, with the ascent of an intuitive web and especially well known web-based virtual entertainment like Facebook and Twitter, there has been a remarkable addition in client created content being made available over the web. By and by any data online can show up at billions of web clients in only seconds that has prompted a positive trade of thoughts as well as brought about noxious and hostile substance over the web. Nonetheless, utilizing human arbitrators to check this hostile substance isn't any longer a viable technique. This ushers online entertainment executives to computerize the hostile language location process and oversee the substance utilizing Natural Language Processing (NLP) procedures. The Multilingual Offensive Language Detection (MOLD) task is typically demonstrated as an administered grouping issue where a framework is prepared on commented on texts containing multilingual harmful or hostile articulations. Zampieri et al., 2019, Zampieri et al., 2020 shared assignments in the International Workshop on Semantic Evaluation (SemEval) pulled in entries from in excess of 100 groups. A few chips away at hostile language distinguishing proof were proposed however just in a monolingual setting by dealing with explicit dialects like English (Zampieri et al., 2020), Arabic (Alami et al., 2020), or different dialects. In any case, a few dialects win in the overall organizations prompting multilingual variety in the text order field that can be expected in a few settings, for example, hostile language recognition, Spam separating, and so forth. Multilingual text order (MTC) is characterized as the assignment of grouping at the same time a bunch of texts written in various dialects (Arabic, English, Spanish…) and having a place with a bunch of fixed classifications across dialects. This issue is not quite the same as cross-language text arrangement (Bel et al., 2003), when an archive written in one language should be characterized in a classification framework learned in another dialect. A few methodologies exist to manage the MTC issue; the primary comprises of fostering a few monolingual classifiers where every language has a particular grouping framework (Lee et al., 2006, Amini et al., 2010, Gonalves and Quaresma, 2010). The subsequent strategy includes one grouping framework for various dialects. The essential thought is to take care of different messages with various dialects to a similar classifier then play out the preparation on a multilingual dataset. The third technique integrates the interpretation ease to design all texts to one language, then, at that point, foster one order framework (Prajapati et al., 2009, Bentaallah and Malki, 2014). In any case, no matter what the multilingual text order significance, research in this space was limited. Also, the MTC issue was never handled utilizing Bidirectional Encoder Representations from Transformers (BERT). This transformer has the capacity to figure out how to separate complex highlights from crude information consequently, accordingly attacking the normal language handling field and giving promising exhibitions (Devlin et al., 2019). Multilingual BERT has additionally pushed cutting edge on cross-lingual and multilingual comprehension tasks by

mutually pretraining huge Transformer models (Vaswani et al., 2017) on numerous dialects.

In this paper, we propose a MTC approach and devise a clever arrangement that exploits move learning strategies in the multilingual hostile language recognition field. The primary commitments to the MOLD field can be summed up as follows:

We propose a powerful way to deal with mark Arabic and English tweets into two classifications: Offensive or not hostile. To manage the multilingualism issue, the joint-multilingual procedure and the interpretation based one are investigated.

We research logical embeddings from transformers, including BERT, multilingual BERT (mBERT), and AraBERT, in particular Arabic BERT (Antoun et al., 2020), to recognize hostile language over multilingualism.

We direct a broad arrangement of analyses on the SOLID dataset to assess our technique's viability and get superior execution.

The rest of this paper is coordinated as follows. We present related work in Section 2. Area 3 portrays the proposed approach. We present trial brings about Section 4. Area 5 gives the ends and headings for future work.

2. RELATED WORK

There is a long history of multilingual text grouping and hostile language location, and we momentarily audit these two perspectives in this part.

2.1. Hostile language identification

Lately, recognizing cyberbullying, animosity, disdain discourse, harmful remarks, and hostile language in Social Media gets a lot of consideration from the specialist's local area. A few public datasets are accessible to prepare machine classifiers for those tasks. Notwithstanding, there are no standard benchmark corpora or preparing sets that can be joined to acquire more vigorous grouping frameworks. Kumar et al. (2018) introduced the report and discoveries of the common undertaking on hostility recognizable proof. The gave dataset contains 15,000 explained Facebook posts and remarks in English and Hindi. The objective was to separate between three classes: non-forceful, secretively forceful, and excessively forceful. The harmful remark grouping was an open contest at Kaggle. Different techniques were assessed for this errand on a dataset containing clients with remarks from Wikipedia. These remarks are coordinated into six classes: poisonous, extreme harmful, vulgar, danger, affront, personality disdain. Concerning disdain discourse distinguishing proof, Davidson et al. (2017) introduced a new disdain discourse identification dataset with more than 24,000 English tweets having a place with three classes: non-hostile, disdain discourse, and foulness. Mandl et al. (2019) detailed the common undertakings on hostile language distinguishing proof where three datasets were created from Twitter and Facebook and made accessible for Hindi, German, and English. Also, Zampieri et al., 2019, Zampieri et al., 2020 introduced a few hostile language recognition brings about a few dialects acquired by groups of SemEval contest.

2.2. Multilingual text grouping

Multilingual text grouping is an arising field in text arrangement. In any case, relatively few past works have been acknowledged around here. Early, Lee et al. (2006) introduced a multilingual text classification strategy utilizing the inert semantic ordering method. This technique comprises of playing out numerous monolingual methodologies on English and Chinese datasets. In another work, Prajapati et al. (2009) presented a methodology depending on the interpretation of records to all inclusive language and afterward played out the grouping. They consolidated the information utilizing WordNET to plan terms to ideas then, at that point, characterize text utilizing direct classifier Rocchio and probabilistic Naïve Bayes and K-Nearest Neighbor (KNN). Amini et al. (2010) explored MTC by consolidating two semi-directed learning methods, including co-regularization and agreement based self-preparing. They prepared different monolingual classifiers on the Reuters Corpus Volume 1 and 2 (RCV1/RCV2) containing five unique dialects: English, German, French, Italian, and Spanish. The creators approved their technique utilizing six arrangement strategies: Boost, co-regularized helping, supporting with self-preparing, Support Vector Machine (SVM) with self-preparing, co-regularization + self-preparing, and helping with full self-preparing. Bentaallah and Malki (2014) thought about two WordNet-based approaches for multilingual text classification. The primary depended on machine interpretation to straightforwardly get to WordNet and utilized a disambiguation procedure to think about just the most well-known importance of the term. While the second barred the interpretation and investigated the WordNet related with every language. Mittal and Dhyani (2015) tended to the multilingual text characterization in view of N-gram methods. They concentrated on MTC in Spanish, Italian, and English dialects. They continued by anticipating the language of a record and involved Naïve Bayes in the grouping stage. All the more as of late, Kapila and Satvika (2016) resolved the issue of MTC on

Hindi and English dialects utilizing different AI calculations, including SVM, KNN, Decision Tree, Self-Organizing Map, and Genetic Algorithms. They further developed the technique exactness by utilizing different component choice strategies.

As of late, profound brain organizations and relevant embeddings were proposed in the text grouping space for English (Lai et al., 2015, Zhou et al., 2015, Nowak et al., 2017, Devlin et al., 2019, Peters et al., 2018, Liu and Guo, 2019, and so on) and Arabic (Dahou et al., 2016, El-Alami et al., 2020, Elnagar et al., 2020, Antoun et al., 2020). Then again, a few multilingual covered language models like mBERT (Devlin et al., 2019) and XLM (Lample and Conneau, 2019) have pre-prepared enormous Transformer models (Vaswani et al., 2017) on numerous dialects. These models were investigated in cross-lingual figuring out assignments (Lample and Conneau, 2019, Conneau et al., 2019) and have demonstrated to be compelling around here.

To put it plainly, in spite of the extensive measure of work on cross-lingual text characterization, the MTC is practically disregarded and barely any examinations were proposed utilizing old style procedures like SVM and KNN. What's more, research in all out attack mode language location field had been prospected exclusively according to a monolingual point of view. Different examinations researched ongoing procedures as mBERT however just in the cross-lingual region. Another methodology is accordingly expected to concentrate on the MTC in the hostile language discovery field from profound learning angles involving promising exchange learning strategies as BERT. Hence, in this examination, we research move learning methods in the MOLD field. Our work is a moderately new region in the MTC area.

3. The proposed system

Our multilingual offensive language detection system comprises several modules, including tweet preprocessing, BERT tokenization, text representation, and tweet classification. Fig. 1 illustrates the overall flowchart of the proposed system.

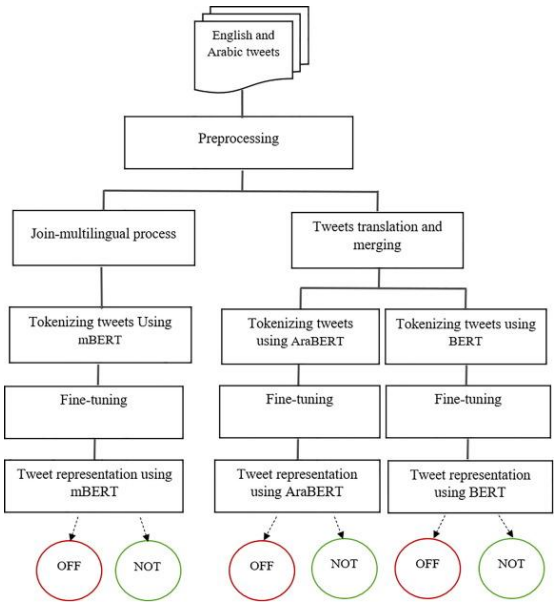

Fig. 1. The overall proposed system flowchart.

3.1. Preprocessing algorithm

The preprocessing phase consists of several steps in order to keep only pertinent information. Fig. 2 shows two examples of tweets preprocessing. We follow the upcoming steps:

1. We proceed by removing all the hashtags, HTML tags, mentions, and URLs.

2. For English text, we further replace contractions with their full forms, fix misspelled words, and convert text to lowercase.

3. We replace emojis if they exist with the text they represent since emojis or emoticons play an essential role in defining a tweet. We have two cases:

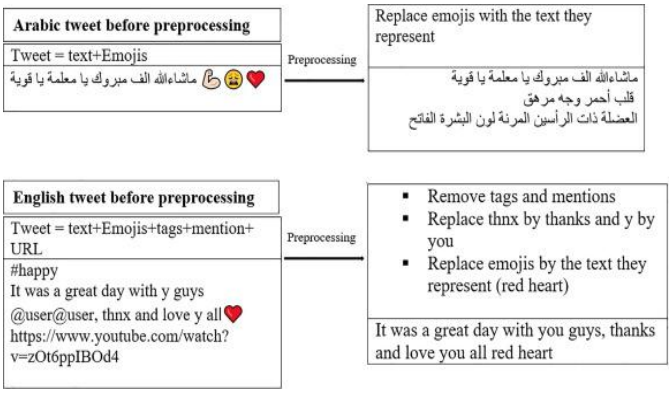

Fig. 2. Preprocessing examples for Arabic and English tweets.

- For Arabic tweets, we initially distinguish emoticons then interpret emoticons implications from English to Arabic. Subsequently, we substitute emoticons by their Arabic implications in the tweet. Fig. 2 outlines an Arabic tweet preprocessing model when emoticons exist.

- For English tweets, we supplant emoticons with their faculties assuming they exist.

After the preprocessing step, we investigate two variations (1) The joint-multilingual strategy by blending tweets without interpretation and (2) The interpretation based procedure by deciphering tweets. Fig. 3 portrays every methodology calculation.

Algorithm 1: Join-multilingual classification

```
L: number of languages;
N: number of tweets
while n ≤ L do
    while p ≤ N do
        Tweet processing;
        Tweet representation;
        Tweet prediction
    end
end
```

Algorithm 2: Join-translated classification

```
Translate tweets to one universal language;
Merge tweets in one file;
N: number of tweets
while p ≤ N do
    Tweet processing;
    Tweet representation;
    Tweet prediction
end
```

Fig. 3. The used multilingualism algorithms.

3.2. Translation

The interpretation is an interaction expecting to design multilingual tweets into one widespread language. We create two corpora as follows:

- Concerning the primary corpus, we make an interpretation of English tweets to Arabic utilizing Google Translator API. Then, we consolidate the Arabic and deciphered tweets. Thereafter, we apply the AraBERT tokenizer.

- To deliver the subsequent corpus, we make an interpretation of Arabic tweets into English in view of similar API. Then, we join the English and interpreted tweets. Then, we feed all tweets to the BERT tokenizer.

3.3. BERT tokenization

In BERT, we want a solitary vector addressing the entire information sentence to be taken care of to the classifier. Thus, the choice is that the secret condition of the main token [CLS] is taken to address the entire sentence. Then again, in the "following sentence expectation" process, BERT has to know where the main sentence closures and where the subsequent sentence starts. Subsequently, the token [SEP] is utilized. During the tokenization cycle, we first add the unique token [CLS] to the top of each tweet and the extraordinary token [SEP] among sentences and toward the end during the tokenization. For Arabic, all the word tokens are divided by the Farasa segmenter (Abdelali et al., 2016) and afterward tokenized with the AraBERT tokenizer. For English and bilingual tweets, we utilize the BERT tokenizer. At last, we map every token to a file in view of the pre-prepared BERT model jargon.

3.4. Tweet portrayal

We plan to investigate BERT in the MTC field, all the more explicitly for the multilingual hostile language recognition task. We utilize the pre-prepared models, including the uncased BERT base (L = 12-layer, H = 768, A = 110 M params), mBERT (L = 12, H = 768, A = 110 M params) and AraBERT (L = 12, H = 768, A = 110 M params), as indicated by the utilized multilingual methodology. These models pre-train profound bidirectional portrayals from the unlabeled text by mutually molding both left and right setting in all layers. Our organizations designs are worked during two stages: (1) The pre-prepared models' investigation and (2) The tweaking. During the calibrating, the BERT models are introduced utilizing the pre-prepared models. Subsequently, every one of the boundaries are calibrated in light of named information from the MOLD task.

To tweak BERT models on MOLD, as we have referenced previously, we tokenize the info texts and add [CLS] and [SEP] tokens. Then, at that point, we create for every token an information portrayal that is developed by adding the vector embeddings relating to the token, the fragment to which it has a place, and the symbolic position. Then, we feed these portrayal vectors to BERT models and tweak them. We take the last secret condition of the principal [CLS] token as the tweet portrayal. Thereafter, we standardize the acquired vectors utilizing a feed-forward layer to get the likelihood circulation over the anticipated result mark (Offensive or Not hostile).

3.5. Text order

After the calibrating stage and to recognize on the off chance that the tweet is hostile or not, we feed the tweaked tweet portrayal to a Sigmoid capacity and train

the model to enhance the double cross-entropy misfortune. During tweaking, we characterize the loads for the characterization layer as W and compute the standard grouping misfortune utilizing the last secret portrayal of [CLS]:C, i.e., logsoftmaxCW.

4. Analyses and results

4.1. The multilingual corpora

We build our own multilingual dataset that contains English and Arabic tweets since there exist no standard corpora to assess MOLD. The English dataset is removed from the semi-administered hostile language ID dataset in SemEval'2020 rivalry. This dataset was proposed by (Rosenthal et al., 2020), containing more than 9,000,000 clarified tweets following OLID's three-level scientific classification: Offensive Language Detection, Categorization of Offensive Language, and Offensive Language Target Identification. We take around 6000 English tweets from the SOLID dataset covering two classifications: Offensive and not hostile.

To develop the Arabic dataset, we remove around 7800 tweets from the dataset proposed by Mubarak et al. (2020) in SemEval'2020. The tweets are named utilizing two marks: OFF and NOT, for hostile and not hostile. The dataset measurements are outlined.

4.2. Assessment measurements

To assess the MOLD framework execution, we utilize two measurements, specifically the exactness and F1-score that are arrived at the midpoint of across the corpus classifications and characterized as follows:(1)Accuracy=TP+TNTP+TN+FP+FN(2)F1-score=TPTP+12(FP+FN)

TP (True Positive): Samples that are positive and anticipated accurately as certain;

FP (False Positive): Samples that are negative however anticipated mistakenly as sure;

FN (False Negative): Samples that are positive however anticipated inaccurately as negative;

TN (True Negative): Samples that are negative and anticipated accurately as negative.

4.3. Exploratory arrangement

We lead a bunch of trials to evaluate our framework execution. The analyses are run on the corpus portrayed in the past subsection. We split the dataset into 80% for preparing and the rest 20% for testing. We utilize Tensorflow and Keras libraries to fabricate and prepare all the BERT models. We fix the quantity of ages at 5 and the clump size at 32 for all models during the preparation stage. The most extreme info arrangement length for every one of the tweets is set at 128. We run the entire analysis stage on Google Colab. As measure assessment, two measurements are used, including Accuracy and F1. We use Adam as an analyzer in the Sigmoid layer.

4.4. Examination techniques

We think about BERT models against different brain models like Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and bidirectional RNN. For these brain organizations, the pre-handling stage is required. The important order comprises of the implanting layer, stowed away layer, and result layer. The Embedding layer comprises the primary secret layer of profound learning organizations. This layer is a grid of the size x×rwhere r is the length of the word implanting vectors (300 aspects) and x is the most extreme length of tweets that is fixed to 200 tokens. We utilize the dropout to keep away from the overfitting issue; its boundary is set to 0.5. During the preparation, we fix the quantity of ages at 10 and the bunch size at 20. The result layer uses the Sigmoid initiation work, Adam streamlining agent and the cross-entropy misfortune to anticipate tweet names. Concerning the boundaries setting, each profound brain network has its own transformation. The execution subtleties are represented in Table 2.

-

CNN

We assemble a CNN model comprising of a 1D convolutional layer with a part size of 5 and 128 channels. The following layer is the maximum pooling with default values followed by the dropout layer. At last, the result layer is answerable for influencing a class to each tweet.

-

RNN

As RNN models, we utilize Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models. Our LSTM model's progressive system comprises of one LSTM layer containing 100 units, trailed by a dropout layer. The last layer is the arrangement one that predicts the tweet classification. We save a similar engineering for the GRU model, and we change the LSTM layer by a GRU layer. This design gives the best exhibitions in the wake of attempting various models. The two models are intended to handle the issue of the Vanishing Gradient of the essential RNN.

-

Bidirectional RNN

We build a Bidirectional LSTM (BiLSTM) made out of one BiLSTM layer with 100 secret units. The yielded vectors are straightened then taken care of to the arrangement layer. While the Bidirectional GRU (BiGRU) is assembled utilizing a BiGRU layer with a similar setup of BiLSTM.

-

Model mixes

We consolidate a few models in an unexpected way. In the first place, we combine CNN and RNN layers to build CNN-LSTM and CNN-GRU. The two models involve one CNN layer followed by one RNN layer. Then, we utilize the worldwide max-pooling and the dropout layer. At long last, the expectation layer to ascribe class to tweet. Concerning the CNN-BiLSTM model, it joins CNN and Bi-LSTM organizations. The model pecking order contains one CNN layer and BiLSTM layer with 100 secret units followed by worldwide max-pooling and a dropout layer.

We evaluate the proposed approach likewise against some other exchange learning strategies including the Universal Language Model Fine-tuning model (ULMFiT) and Embeddings from Language model (ELMo), and the standard classifier SVM.

4.5. Experimental results

We have done broad analyses utilizing the transformer BERT models to examine different MTC procedures on the MOLD task. The principal set of tests is directed to examine the joint-multilingual technique utilizing the multilingual BERT. Conversely, the subsequent set concerns the interpretation based strategy assessment that is two folds by investigating AraBERT and BERT models. First and foremost, we assess this technique by applying interpretation to the English language, and also, we research a similar system utilizing interpretation to the Arabic language. To evaluate the effect of our commitment, the accompanying subsections sum up the acquired discoveries and a progression of profound examinations.

4.5.1. Joint-multilingual technique results

The main examination is intended to assess mBERT on MOLD in light of the joint-multilingual strategy Table 3 sums up the got brings about terms of precision and F1-score. First and foremost, we gauge mBERT against the traditional classifier SVM related to the component choice technique Chi-square (Khi2). To do as such, we compute the Term Frequency-Inverse Document Frequency (TFIDF) highlights lattice all in all corpus, and afterward we apply the Khi2 measurements to change the TF-IDF framework to a low space include that contains the best 1000 elements. A while later, we train a SVM in light of the spiral premise work piece on the acquired tweet portrayals to foresee the right class. The outcomes exhibit that mBERT outflanks SVM-Khi2 somewhat concerning precision. Notwithstanding, they get a similar F1 esteem. This finding isn't is business as usual since SVM stays famous regardless accomplishes exceptionally cutthroat outcomes in the text arrangement field. Also, we consider mBERT in contrast to CNN, LSTM, GRU, BiLSTM, and CNN-BiLSTM models. It tends to be seen that mBERT accomplishes 91% regarding F1 and beats the wide range of various brain models concerning F1. We have additionally looked at mBERT against the multilingual ELMo model (Peters et al., 2018), a profound contextualized word portrayal that models both punctuation and semantics attributes utilizing a profound bidirectional language model (biLM) prepared on a particular task to make embeddings. We utilize the pre-prepared multilingual ELMo (Che et al., 2018) to create tweet portrayals, and afterward we feed the got portrayals to a CNN classifier involving a 1D convolutional layer. This layer contains a channel of size 64 that is passed across the portrayal lattice to distinguish explicit elements in tweets. A while later, we apply the maximum pooling layer to downsample the approaching elements map. The result vectors are consolidated in one grid and afterward passed to a completely associated layer. Then, at that point, the dropout method is applied to lessen overfitting, and its part is set to 0.5. From Table 3, it tends to be called attention to that mBERT surpasses ELMo by practically 12% of F1. This hole can be because of the way that ELMo connects two BiLSTM in two headings of the tweets while mBERT is bidirectional.

After the mBERT assessment, we use the disarray lattice (Fig. 4) to find where the classifier neglects to foresee the right classes.

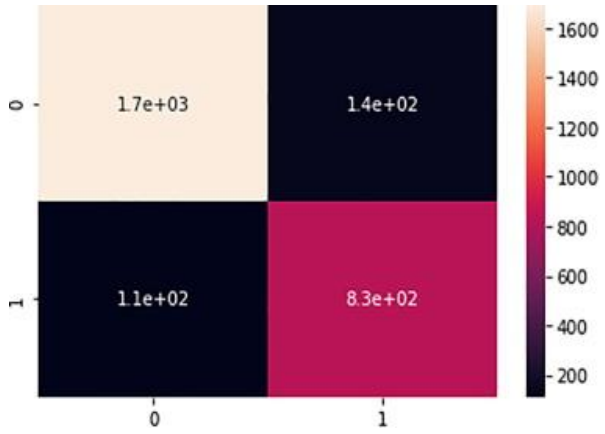

Fig. 4. Confusion matrix of the joint-multilingual text classification method (OFF“: 1, ”NOT“: 0).

4.5.2. Joint-translated monolingual method results

As we mentioned before, when we follow the joint-translated monolingual technique to deal with multilingualism, we generate two corpora using the Google Translator API. The first corpus comprises only the tweets in English. In contrast, the second one contains the tweets in Arabic. In this work, we did not use any disambiguation process to cope with ambiguity problems. The upcoming experiments concern the above method evaluation on each corpus apart.

Fig. 5 illustrates the findings of the translation-based method where we translate all tweets to English. It can be noticed that BERT achieves the best accuracy results by almost 92% of accuracy. Besides, CNN-LSTM and GRU models are worse than the other models. Moreover, CNN, CNN-LSTM, and Bi-GRU attain mostly the same accuracy ranging from 86% to 87%. The finding proves the importance of capturing the dependencies among different word combinations that can be possible in BERT but not in CNN, RNN, or their combination.

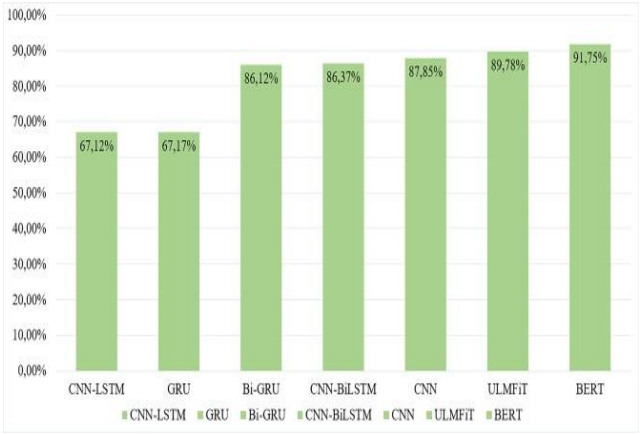

Fig. 5. Accuracy results of the joint-translated method based on the English language.

We Compare BERT further against another fine-tuned model ULMFiT, a transfer learning model. This model was proposed by (Howard and Ruder, 2018) and pretrains a language model (LM) on a large general-domain corpus and fine-tunes it. For the implementation, we follow the training scheme described by Jeremy Howard in fast.ai, taking a pretrained language model, fine-tuning it with unlabeled data, then fine-tuning the classification for the MOLD task. The obtained results show that BERT outperforms ULMFiT by 2%, thanks to the bidirectional nature of BERT that allows it to reflect more semantic in comparison with ULMFiT that captures less semantic within a tweet.

We further evaluate BERT against other BERT versions such as RoBERTa (Liu et al., 2019) and DistilBERT (Sanh et al., 2019). The accuracy and loss results are reported in Fig. 6. The obtained findings indicate that BERT outperforms RoBERTa and distilBERT in terms of accuracy. However, RoBERTa has the minimum loss value.

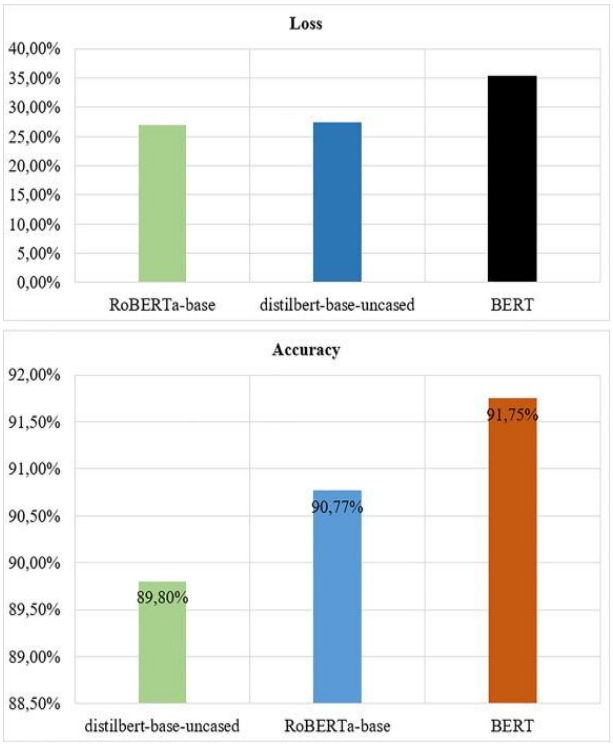

Fig. 6. Accuracy and Loss results of the joint-translated method using English language.

The third experiment is designed to evaluate AraBERT's performance in detecting multilingual offensive tweets using the translation method. In this experiment, we translate English tweets to Arabic. Then, we merge them with Arabic tweets. Table 4 describes the accuracy and F1-score results on the merged dataset. It can be observed that AraBERT beats all the deep neural networks and achieves 93% in terms of F1. Besides, it can be pointed out that the GRU model outperforms the other neural networks by a value of 86% for F1. CNN reported the same accuracy as GRU, which equals 85%. Moreover, BiLSTM achieves 84% of accuracy and F1-measure. Whereas the lowest performances are given by CNN-GRU that affords 81% in terms of F1. The Arabic UMLFiT (AraULMFiT) model is also used for comparison purposes. We fine-tune the model from (ElJundi et al., 2019) on MOLD task. We notice that the fine-tuned AraBERT model outperforms AraULMFiT by a score of 13%. This result demonstrates the advantage of using a bidirectional model (AraBERT) instead of a unidirectional model as AraULMFiT.

To go deeper in our analysis, we compare the different multilingualism techniques. Fig. 7 illustrates the obtained results. Based on the assessment, we adopt the AraBERT with the joint-translated method since this combination achieves the best F1-score by a value of 93%.

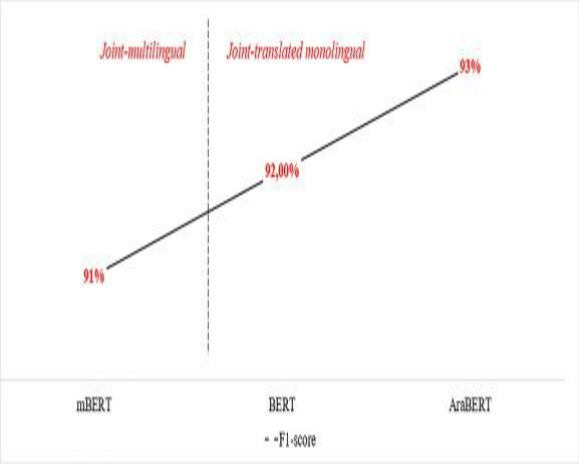

Fig.7. Joint-translated monolingual classification vs. Join-multilingual classification in terms of F1.

We have led an exhaustive set of experiments to investigate BERT models efficiency in MOLD field. The findings show the discriminating capability of the transformer BERT models to deal with the multilingual text classification in the offensive language detection field.

5. Conclusion and future work

In this paper, we presented a transfer learning-based approach to deal with multilingual text classification in the offensive language detection field. We tackled the multilingualism problem using joint-multilingual and translation-based methods. Our approach relies on the transformers BERT models, including BERT, mBERT, and AraBERT, that were fine-tuned on the multilingual offensive detection task. We followed several steps to build the MOLD system: (1) Preprocessing, (2) Tweets tokenization, (3) BERT models fine-tuning and (4) Tweets classification. We carried out an exhaustive set of experiments on a bilingual corpus extracted from the SOLID dataset. The main findings of our work confirmed that both the translation-based method and joint-multilingual one exceeded state-of-the-art methods and accomplished good F1 and accuracy scores. Importantly, our results provide evidence of BERT models robustness in the MOLD field.

Despite experimenting our approach in a bilingual context, our contribution stays operational when considering more languages. Thereby, in future work, we will boost our experiments further by integrating other languages. We also project to explore some disambiguation methods in order to tackle the ambiguity problem, when we follow the translation-based strategy, and enhance the performances. We further plan to work on other tasks such as hate speech indentation and bullying.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Abdelali et al., 2016

1. Abdelali, A., Darwish, K., Durrani, N., Mubarak, H., 2016. Farasa: A Fast and Furious Segmenter for Arabic, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations. Association for Computational Linguistics, San Diego, California, pp. 11–16. https://doi.org/10.18653/v1/n16-3003

2. H. Alami, S.O. El Alaoui, A. Benlahbib, N. En-nahnahi **LISAC FSDM-USMBA Team at SemEval-2020 Task 12: Overcoming AraBERT's pretrain-finetune discrepancy for Arabic offensive language identification**

3. Proc. Fourteenth Workshop Seman. Eval. (2020), pp. 2080-2085 View PDF CrossRefGoogle Scholar Amini et al., 2010 M.R. Amini, C. Goutte, N. Usunier **Combining coregularization and consensus-based self-training for multilingual text categorization** In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval (2010), pp. 475-482, 10.1145/1835449.1835529 View PDF View Record in ScopusGoogle Scholar

4. Antoun, W., Baly, F., Hajj, H., 2020. AraBERT: Transformer-based model for Arabic language understanding. arXiv preprint arXiv:2003.00104. Google Scholar

5. Bel, N., Koster, C. H., & Villegas, M., 2003. Cross-lingual text categorization. In International Conference on Theory and Practice of Digital Libraries, Berlin, Heidelberg, pp. 126-139. Google Scholar Bentaallah and Malki, 2014.

AUTHORS

BUGGIDI HARSHINI,
B.TECH SCHOLAR
CSE DEPARTMENT,
GATES INSTITUE OF TECHNOLOGY,
JNTU ANANTAPUR.
GOOTY.

MATTEDDULA KRISHNA REDDY
B.TECH SCHOLAR
CSE DEPARTMENT,
GATES INSTITUE OF TECHNOLOGY,
JNTU ANANTAPUR.
GOOTY.

MANGALAMADAKA MAHEMA,
B.TECH SCHOLAR
CSE DEPARTMENT,
GATES INSTITUE OF TECHNOLOGY,
JNTU ANANTAPUR.
GOOTY.

SHAIK MOULA BEE
B.TECH SCHOLAR
CSE DEPARTMENT,
GATES INSTITUE OF TECHNOLOGY,
JNTU ANANTAPUR.
GOOTY.

KANDULA NAGALIKITHA
B.TECH SCHOLAR
CSE DEPARTMENT,
GATES INSTITUE OF TECHNOLOGY,
JNTU ANANTAPUR.
GOOTY.

S. NAGARAJU
ASSISTANT PROFESSOR,
CSE DEPARTMENT,
GATES INSTITUE OF TECHNOLOGY,
JNTU ANANTAPUR.
GOOTY.