Improving Arabic Text Categorization Using Transformer Training Diversification

Shammur A Chowdhury, Ahmed Abdelali, Kareem Darwish
Soon-gyo Jung, Joni Salminen, Bernard J Jansen
Qatar Computing Research Institute
{shchowdhury, aabdelali, kdarwish, sjung, jsalminen, bjansen}@hbku.edu.qa

Abstract

Automatic categorization of short texts, such as news headlines and social media posts, has many applications ranging from content analysis to recommendation systems. In this paper, we use such text categorization i.e., labeling the social media posts to categories like ‘sports’, ‘politics’, ‘human-rights’ among others, to showcase the efficacy of models across different sources and varieties of Arabic. In doing so, we show that diversifying the training data, whether by using diverse training data for the specific task (an increase of 21% macro F1) or using diverse data to pre-train a BERT model (26% macro F1), leads to overall improvements in classification effectiveness. In our work, we also introduce two new Arabic text categorization datasets, where the *first* is composed of social media posts from a popular Arabic news channel that cover Twitter, Facebook, and YouTube, and the *second* is composed of tweets from popular Arabic accounts. The posts in the former are nearly exclusively authored in modern standard Arabic (MSA), while the tweets in the latter contain both MSA and dialectal Arabic.

1 Introduction

Text classification, particularly of short texts, is an important problem in NLP and has been used in a variety of tasks in social media such as identifying people’s sentiment (Mohammad et al., 2013), emotions (Abdullah and Shaikh, 2018), interests (Keneshloo et al., 2016), stance (Mohammad et al., 2016), offensive languages (Chowdhury et al., 2020; Hassan et al., 2020) and communication styles (Mubarak et al., 2020). Text classification requires the availability of manually tagged text to train effective classification models. Due to annotation costs, adapting labeled texts from one domain to tag texts in other domains is desirable, as it would avail the need to tag in-domain data.

With the recent success of pre-trained transformer-based models (e.g. BERT), various studies have adopted such models to generate contextualized embeddings for downstream tasks like text classification, using a small amount of in-domain data. To push the state-of-the-art performance, researchers have also experimented with variation in such model size (e.g. number of layers, attention head among others), architectures (BERT *vs* RoBERTa *vs* XLNet) and training data language (mono *vs* multilingual).

From the language perspective, many studies have reported that monolingual transformer models such as BERT performs significantly better than the multilingual BERT - mBERT (Polignano et al., 2019). However, very few studies have empirically shown the effect on the performance of diversifying a BERT pre-training data with formal and informal textual contents.

Therefore in this paper, we show the effectiveness of using a transformer model (named as QARiB),[1] which is trained on a *mixture of formal news* and *informal tweets* (i.e., written in dialectal Arabic). We compare the performance of QARiB with *(i)* multilingual BERT (mBERT), which is trained using multiple languages including Arabic, and *(ii)* AraBERT, which is trained on a large corpus of Arabic news (formal text only). For the evaluation, we employed these models and trained a multiclass short text classifier using *news headlines*, and then tested it on *tweets*.

[1] `https://github.com/qcri/QARIB`

Proceedings of the Fifth Arabic Natural Language Processing Workshop, pages 226–236
Barcelona, Spain (Online), December 12, 2020

As a byproduct of this work we annotate two new Arabic text categorization data sets. The *first* is a large set of social media posts collected from multiple platform – Twitter, Facebook, and YouTube – of a popular news site, where the posts are written in Modern Standard Arabic (MSA). The *second* is a set of mostly dialectal tweets that were authored by influential Arabic accounts. Details of the annotation guideline along with the annotated datasets are made publicly available.

Text categorization is complicated due to the fact that Arabic is the lingua franca for 22 countries with MSA being used as the formal language of communication while mutually unintelligible dialects are spoken in these countries and appear in social media posts. Further, social media posts, particularly tweets, are typically short and contain platform specific features such as hashtags and user mentions, further complicating the classification.

In summary, the contributions of this paper are as follows:

- We introduce two new Arabic text categorization datasets, which cover multiple social media platforms and different variations of Arabic including MSA and dialects. We publicly released the datasets along with annotation guidelines and examples.[2]
- We showcase the efficacy of using a transformer model that is trained on a mixture of formal and informal Arabic in providing effective domain adaptation for informal text categorization using formal Arabic training data.

2 Related Studies

The task of categorizing social media posts is challenging mainly due to the absence of largely annotated data. Most of the datasets that are currently available are based on news articles.
Some publicly available datasets are:

1. SANAD[3] (AlSaleh et al., 2020): This is one of the largest collection of news article for Arabic news text classification. This multi-class dataset includes news articles that are scraped from the AlKhaleej,[4] AlArabiya,[5] and Akhbarona[6] news portals. The dataset includes approximately 195k articles belonging to 6/7 categories.
2. NADiA[7] (Elnagar et al., 2020): This is a multi-label text dataset, collected by scraping SkyNewsArabia[8] and Masrawy[9] news sites. The released dataset include $486k$ articles with 52 categories.
3. Arabic News Text (ANT) Corpus[10] (Chouigui et al., 2017): This dataset was collected from RSS feeds and contains approximately 6k articles belonging to 9 categories

Other available datasets include: Khaleej-2004 (5k articles belonging to 4 categories) (Abbas and Smaili, 2005); Watan-2004 (20k articles belonging to 6 categories) (Abbas et al., 2011); and SL-RTANew[11] (20k articles belonging to 40 categories).

To capture syntactic and semantic information about words, pre-trained word static embeddings (Turian et al., 2010; Mikolov et al., 2013; Pennington et al., 2014) were widely used in many NLP tasks. Recent research advancements led to pre-trained contextual embeddings that capture much information about words in context, leading to significant improvements for many NLP tasks such as text classification and sequence labeling (Mikolov et al., 2017; Peters et al., 2018; Devlin et al., 2018; Howard and Ruder, 2018; Lan et al., 2019; Liu et al., 2019; Yang et al., 2019).

As for Arabic, various static and contextual embeddings representation have been trained. Some popular Arabic static embeddings include: Arabic word2vec (Soliman et al., 2017), a FastText model

[2] https://github.com/shammur/Arabic_news_text_classification_datasets
[3] https://data.mendeley.com/datasets/57zpx667y9/2
[4] http://www.alkhaleej.ae/portal
[5] https://www.alarabiya.net
[6] https://www.akhbarona.com
[7] https://data.mendeley.com/datasets/hhrb7phdyx/2
[8] https://www.skynewsarabia.com
[9] https://www.masrawy.com
[10] https://github.com/antcorpus/antcorpus.data
[11] https://data.mendeley.com/datasets/322pzsdxwy/1

that is trained on Wikipedia (Bojanowski et al., 2017), and dialectal word embeddings with small and noisy corpora (Erdmann et al., 2018) and with tweets (Abdul-Mageed et al., 2018; Farha and Magdy, 2019).

As for contextual embeddings, a handful of models are available (ElJundi et al., 2019; Antoun et al., 2020; Talafha et al., 2020). The first available BERT model for Arabic was multilingual BERT (mBERT), which was pre-trained on the Wikipedia dumps of 104 languages including Arabic. However, previous studies have shown that monolingual BERT models perform significantly better than the mBERT (Polignano et al., 2019). A recent Arabic BERT model (AraBERT) (Antoun et al., 2020) was trained on Wikipedia and a large collection of Arabic news articles, with the base configuration of the BERT model. The model showed success for many Arabic NLP downstream tasks. Recently, a Multi-dialect-Arabic-BERT (Talafha et al., 2020) model was released and entailed fine-tuning AraBERT on $10M$ tweets. The model was used to improve dialect identification.

Compared to the available datasets, which are mainly based on news articles, our introduced datasets are based on social media platforms. The datasets include posts written in standard and dialectal Arabic from multiple social media platforms, and labeled with 12 news categories. We publicly released the datasets for the research community.

Unlike most of the previous studies, we empirically show the effectiveness of monolingual BERT in compare to multilingual BERT for Arabic language processing and the importance of having diversely (pre-)trained BERT model for social media post classification task.

3 Datasets

In this paper, we used three different datasets. Two of them are written in MSA and contain short social media posts and news headlines, and the third is composed of tweets, many of which are authored in dialectal Arabic.

3.1 Arabic Social Media News Dataset (ASND)

To create this dataset, we collected the posts of the official Aljazeera news channel accounts on Twitter, Facebook, and YouTube from February, 2017 to September, 2019. We randomly selected 10k posts that included approximately 6k tweets, 2k Facebook posts, and 2k YouTube video titles. We annotated all the posts using Amazon Mechanical Turk (AMT).[12] We asked the turkers to assign a label to each post from one of twelve predefined categories. These categories include: *(i) art-and-entertainment*, *(ii) business-and-economy*, *(iii) crime-war-conflict*, *(iv) education*, *(v) environment*, *(vi) health*, *(vii) human-rights-press-freedom*, *(viii) politics*, *(ix) science-and-technology*, *(x) spiritual*, *(xi) sports*, and *(xii) others*. We provided the turkers with an elaborate and detailed description of each category along with example annotations. The provided annotation guideline and examples are publicly available.[13] Each turker was asked to annotated 25 different posts. We imposed three types of checks to ensure high quality annotations. These checks were as follows:

- We provided challenged questions to annotators to ensure their Arabic language proficiency. This entailed asking them 10 different multiple choice questions such as the one shown in *Example* 1. Turkers needed to answer at least 8 out of the 10 questions correctly to qualify.
- We embedded 5 posts, for which we had gold labels, within the 25 assigned to each turker. To accept the work of a turker, (s)he had to match the gold labels of at least 4 out the 5 posts. The gold labeled posts were drawn from a set of 5 thousand news headlines that were part of the SANAD dataset (AlSaleh et al., 2020).
- Each post was assigned to 3 turkers and at least 2 out of the 3 needed to agree on a label. If they did not agree, the annotations for the post were discarded. In doing so, we discarded roughly 1.7k posts. Overall, the inter-annotator agreement, as measure by Fleiss's kappa (Falotico and Quatto, 2015), was 0.69.

[12] https://www.mturk.com

[13] https://github.com/shammur/Arabic_news_text_classification_datasets

	ASND			SANAD		AITD
Classes	Train	Test	Dev	Train	Test	Test
art-and-entertainment	345	57	29	–	–	6247
business-and-economy	161	27	14	9219	1024	12270
crime-war-conflict	889	147	76	–	–	–
education	65	11	5	–	–	498
environment	121	20	11	–	–	5010
health	157	26	13	9359	1040	9456
human-rights-press-freedom	337	56	28	–	–	19477
others	773	127	66	–	–	–
politics	3387	559	288	9352	1036	9369
science-and-technology	173	28	15	9353	1040	4936
spiritual	77	13	6	2571	276	29554
sports	191	32	16	9357	1039	18875
Total	6676	1103	567	49211	5455	115692
# Labels	12			6		10

Table 1: Distribution of train and test labels for both ASND, SANAD and AITD datasets.

The final distribution of the dataset including the train/test/dev splits are listed in Table 1.

Example 1 ماذا نسمي والد الأب؟
What we call the father of the father?
Options are:

1. العم *(The uncle)*
2. الوالد *(The father)*
3. الخال *(The maternal-uncle)*
4. الجد *(The grandfather)*

As a sanity check, we identified the most discriminating terms in the posts for each category. We compared the vocabularies of the twelve classes using the valence score (Conover et al., 2011; Mubarak and Darwish, 2019; Chowdhury et al., 2020) (ϑ) for every token x as follows:

$$\vartheta(x, L_i) = 2 * \frac{\frac{C(x|L_i)}{T_{L_i}}}{\sum_l^L C(x|L_l)} - 1 \tag{1}$$

where $C(.)$ is the frequency of the token x for a given class L_i, and T_{L_i} is the total number of tokens present in the class. The valence value $\vartheta(x)$ ranges between -1 and 1, with values closer to 1 indicating strong positive correlation and values closer to -1 indicating strong negative correlation. Table 2 lists the most frequent unigrams and bigrams with $\vartheta = 1.0$, and they seem to reflect the categories. For example, the tokens with high valence for *human-rights-press-freedom* include *justice* and *detainee*, and those for *art-and-entertainment* include *artist* and *theater*.

3.2 Single-Label Arabic News Articles Dataset (SANAD)

The SANAD dataset (AlSaleh et al., 2020) is a large collection of Arabic news articles that has been used for different Arabic NLP tasks. The collected articles were assigned one of seven categories, namely: *(i) culture*, *(ii) finance*, *(iii) medical*, *(iv) politics*, *(v) religion*, *(vi) sports*, and *(vii) technology*. As can be seen, the SANAD dataset has fewer categories than ASND dataset. This imposed some limitations on our experiments. Further, we aligned the SANAD categories to ASND categories by mapping *medical* to *health*, *religion* to *spiritual*, *technology* to *science-and-technology*, and *finance* to *business-and-economy*.[14] For our work, we only considered the headlines of the news articles, while maintaining the official train-test split. We used this dataset to: *(i)* validate the performance of the model trained using the news headlines and tested on social media data; and *(ii)* train models on both the SANAD and ASND data. Details of the dataset can be found in Table 1.

[14]For the task, we ignored the label culture from SANAD.

3.3 Arabic Influencer Twitter Dataset (AITD)

For the third dataset, a domain expert identified a list of 60 Arab influencers on Twitter, who predominantly tweet in specific categories. We performed weak annotation where we labeled all the tweets in an account by the most common tweet category therein. For example, the account "eToroAr" is the official Arabic account of a stocks trading company, and hence its tweets are assumed to belong to the *business-and-economy* category. We used the Twitter APIs to crawl the last 3,200 tweets per account.

To improve annotation, given our best classifier on the SANAD and ASND datasets, we filtered out noisy accounts where the classifier did not find at least 40% of the tweets to belong to one of the categories. As a result, for the final dataset, we retained 36 twitter accounts containing 115,692 Arabic tweets. More details of the dataset can be found in Table 1.

art-and-entertainment	**business-and-economy**	**crime-war-conflict**	**education**
مسرحية (a show)	العربية عاجل (AlArabiya Breaking News)	المتحدث العسكري (Military Spokeman)	مقالات معنى (manaa articles)
الفنانة (artist)	عاجل https (Breaking News)	باسم الحوثيين (of Huthis)	manaa net
مسرح (theater)	alarabiya	الحزام (the belt)	ترجمات معنى (maana translations)
على مسرح (on theater)	الأسهم (Stocks)	الحزام الأمني (The security belt)	ترجمات (translations)
طارق العلي (Tariq Al Ali)	أوبك (OPEC)	تل أبيب (Tel Aviv)	معنى https (maana https)
الارشيف (archive)	كورونا https (Corona)	بطائرات (with airplanes)	الفلسفة (Philosophy)
من الارشيف (from the archive)	الربع الأول (First Quarter)	الحوثيون يعلنون (The Huthis annouced)	مقابلات (Interviews)
المسلسل (TV series)	أسعار النفط (Oil Prices)	الناطق (the spokmen)	مقابلات معنى (maana Interviews)
الفنان القدير (The great artist)	في الربع (In the Quarter)	قطاع غزة (Gaza strip)	مراجعات معنى (maana reviews)
سنة العرض (The year of performance)	برميل (Barrel)	قوات الاحتلال (Occupation forces)	مراجعات (reviews)
environment	**health**	**human-rights-press-freedom**	**politics**
ثادق (Thadiq)	أطعمة (Food)	المرصد (The Observartory)	حراك (movement)
الاشجار (trees)	القلب (heart)	المرصد السوري (The Syrian Observatory)	الداشر (Al Dasher)
وزارة البيئة (Environment ministry)	الصحة ارتفاع (heart elevation)	معتقلي (detainees)	١٥ سبتمبر (05-Sep)
ثادق الوطني (Thadiq National)	مصابي (patiants)	السلطات السعودية (Saudi officials)	حراك ١٥ (movement 15)
منتزه (park)	وزيرة الصحة (Minister of health)	معتقلي الرأي (Opinion detainees)	غانم (Ghanem)
الغاف (Al Ghaf)	وخروجهم (with their exit)	التعسفي (arbitrary)	الدب (Bear)
عبيثران (Wormwood)	وخروجهم من (with their exit from)	المعتقل (detainnee)	الدب الداشر (The dasher bear)
الرعي (grazing)	كوفيد ١٩ (COVID-19)	القسط (justice)	ال سعود (Al Saud)
بثادق (in Thadiq)	المستجد كوفيد (novel COVID)	الاعتقال التعسفي (Arbitrary detention)	السناب (Snap)
استزراع (Farming)	حالات الشفاء (recovering cases)	تأكد لنا (was confirmed)	
science-and-technology	**spiritual**	**sports**	**others**
تي (T)	صل (PBUH)	الدوري (Championship)	شكراً على (Thanks for)
آي تي (IT)	تعالى (Almighty)	الدوري مع (Championship with)	عزيزي (Dear)
إم آي (MI)	القديس (Saint)	مع وليد (with Waleed)	صديقنا (Our friend)
ريفيو (Review)	قال الله (Allah said)	SBC	نشكرك (We thank you)
تكنولوجي (Technology)	بك من (from you with)	وليد (Waleed)	🌹
تي تكنولوجي (T Technology)	الشعراوي (Shaarawi)	مع وليد (with Waleed)	تفاعلك (Your reaction)
تكنولوجي ريفيو (Technology review)	الله تعالى (Allah Almighty)	انفوجرافيك (Infographic)	:)
utm	إله (God)	انفوجرافيك الهلال (Hilal Infographic)	😊
هذا المقال (this article)	لا إله (No God)	كبير آسيا (Largest Asia)	على تفاعلك (for your reaction)
سامسونغ (Samsung)	سورة (Chapter)	وليد الفراج (Waleed Al Farraj)	طبعاً (of course)

Table 2: List of most frequent uni- and bi-grams units per class with $\vartheta = 1.0$.

4 Experimental Setup

4.1 Experiments

We conducted a large battery of experiments where we trained our models on SANAD or ASND individually or in combination. We used the training/test splits of the datasets. Since there is a mismatch between the number of categories between both sets, when training or testing on *SANAD alone*, we restricted our evaluation to the *six categories*. When training using *ASND alone* or in *combination with SANAD*, we used all *12 categories*.

4.2 Models

As stated earlier, we wanted to see the effectiveness of using a transformer model that is trained on a mixture of formal and informal text in comparison to other models that are trained exclusively on formal text. We also compared the transformer models to a baseline that uses an SVM classifier that is trained on character and word n-grams.

4.2.1 Support Vector Machines (SVM)

SVM (Platt, 1998) has been shown to work well for a variety of text classification tasks (Lewis, 2001; Mubarak et al., 2020). To train the baseline classifiers, we used a combination of character n-grams and word n-grams. We used a bag-of-model tf-idf weighting. For character n-grams, we varied n between 1 and 8, and we varied n for words between 1 and 5.

4.2.2 Pre-trained Bidirectional Encoder Representations from Transformers (BERT) Models

We experimented with three different BERT models as follows:

Multilingual BERT: mBERT (formal text): The model is pre-trained using a masked language modeling (MLM) objective using Wikipedia articles for 104 languages including Arabic. We used the case sensitive base model (Devlin et al., 2018).

Arabic BERT: AraBERT (formal text): This model is pre-trained on a collection of publicly available corpora including Arabic Wikipedia, the $1.5B$ words Arabic Corpus (El-Khair, 2016), the OSIAN Corpus (Zeroual et al., 2019), Assafir news articles, and 4 other manually crawled news websites (Al-Akhbar, Annahar, AL-Ahram, AL-Wafd) from the Wayback Machine. The final model is trained on approximately 70M sentences containing roughly 3B Arabic tokens (Antoun et al., 2020).

Arabic BERT: QARiB (mixed style text): This model is trained on the Arabic GigaWord corpus,[15] Abulkhair Arabic Corpus (El-Khair, 2016) , and OpenSubtitles (Lison and Tiedemann, 2016) in addition to 50 million tweets that were collected by issuing the query "lang:ar" against Twitter API. The final training corpus contains 120M sentences and tweets composed of 2.7B Arabic words.

Downstream Task Design: For the downstream tasks, we fine-tuned the aforementioned BERT models for our classification task using a learning rate of $2e-5$ with a batch size of 64 and 3 epochs. For the training, we restricted the maximum input length to 128 tokens, with no extra preprocessing of the data.

4.3 Evaluation

To asses the categorization effectiveness we used Macro F1, which is computed by average the F1 of all labels. We also report on precision n with values of n equal to 1, 2, and 3.

5 Results and Discussion

Table 3 summarizes the results of the experiments. Whenever SANAD is used alone for training or testing, the results are reported on 6 categories. When testing on SANAD while training on ASND or SANAD+ASND, we used the subset of ASND posts that contain the 6 categories. In all other cases, the training and testing were done on 12 categories. From looking at the results, we can observe the following:

Mismatch in style leads to lower results. Though both SANAD and ASND use formal text, training on one and testing on the other produces significantly lower results compared to training and testing on SANAD alone. We suspect that this due to the difference in style between news headlines and social media posts, where the latter contain platform specific features such as hashtags and mentions. The drop in the effectiveness may indicate the inability of the models to generalize well when the style changes.

Combining dataset of different styles helps. As the results show, training using ASND+SANAD performed on SANAD at par to training on SANAD. Further combining both training sets led to improved results for all models when testing on AITD. The results of training on both and testing on ASND yielded mixed results, where results of using BERT models improved significantly or matched the results of training on ASND alone. This was not the case for SVM, where the results dropped noticeably.

[15] `https://catalog.ldc.upenn.edu/LDC2011T11`

Test Set		Training Set SANAD	ASND	ASND+SANAD
SANAD	SVM	93	55	93
	mBERT	93	55	93
	AraBERT	94	60	94
	QARiB	94	81	94
ASND	SVM	64	71	66
	mBERT	61	51	70
	AraBERT	68	51	72
	QARiB	67	77	76
AITD	mBERT	–	37	51
	AraBERT	–	38	54
	QARiB	–	57	60

Table 3: Results (Macro-F1) of training using SANAD, ASND, and ASND+SANAD and testing on SANAD, ASND, and AITD. When training or testing alone, only 6 categories are used (grey cells). For all other cases, 12 categories are used (white cells).

A BERT model trained on a mixture of formal and informal data has much better generalization power compared to BERT models that are trained on formal text only. This observation is apparent across all experiments where we conducted cross-dataset training and testing and there was a mismatch in style or language variety between them. For example, when training on ASND and testing on SANAD, QARiB results were 21 points better than using AraBERT (F1 of 81 compared to 60). A similar result is observed when training and testing on social media posts. When training on ASND or ASND+SANAD and testing on ASND, QARiB led to significant improvements over mBERT and AraBERT. Results of testing on AITD show the same trend. This indicates the importance of a trained BERT model with mixed style data for effective domain adaptation. To further understand the impact of having pre-existing style/domain knowledge, we analyzed the difference in per class predictions. From the Figure 1, we observe that QARiB performed significantly better for the majority of the classes. Table 4 reports on the precision n when testing on the ASND and AITD datasets. The results show that using QARiB was far more likely, compared to AraBERT, to rank the proper category at the top. For example, when training and testing on ASND, P@2 was 90% when using QARiB compare to 69% when using AraBERT. The same was consistent regardless of which dataset is used for training (e.g., ASND+SANAD) or testing (e.g., ASND or AITD). This further reflects the efficacy of pre-training BERT on mixed data.

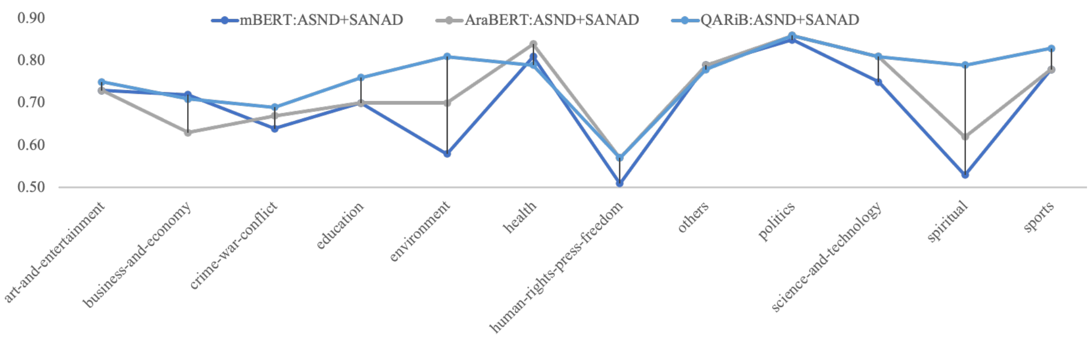

Figure 1: Class-wise F-measure performance on ASND test set using mBERT:ASND+SANAD, AraBERT:ASND+SANAD and QARiB:ASND+SANAD models.

Discussion

For error analysis we studied the confusion between categories by our best model, namely the QARiB:ASND+SANAD model. For the ASND test set, we noticed that the *politics* category is frequently confused with *crime-war-conflict* (28%) and *human-rights-press-freedom* (20%) – see Figure

Test Sets	Models	Training Sets					
		ASND			ASND+SANAD		
		P@1	P@2	P@3	P@1	P@2	P@3
ASND	mBERT	62	77	83	73	85	90
	AraBERT	56	69	73	72	87	91
	QARiB	77	90	94	76	90	94
AITD	mBERT	46	57	73	59	70	77
	AraBERT	49	71	85	64	74	79
	QARiB	68	77	83	69	79	86

Table 4: Precision n for models trained on ASND and ASND+SANAD and tested on ASND and AITD

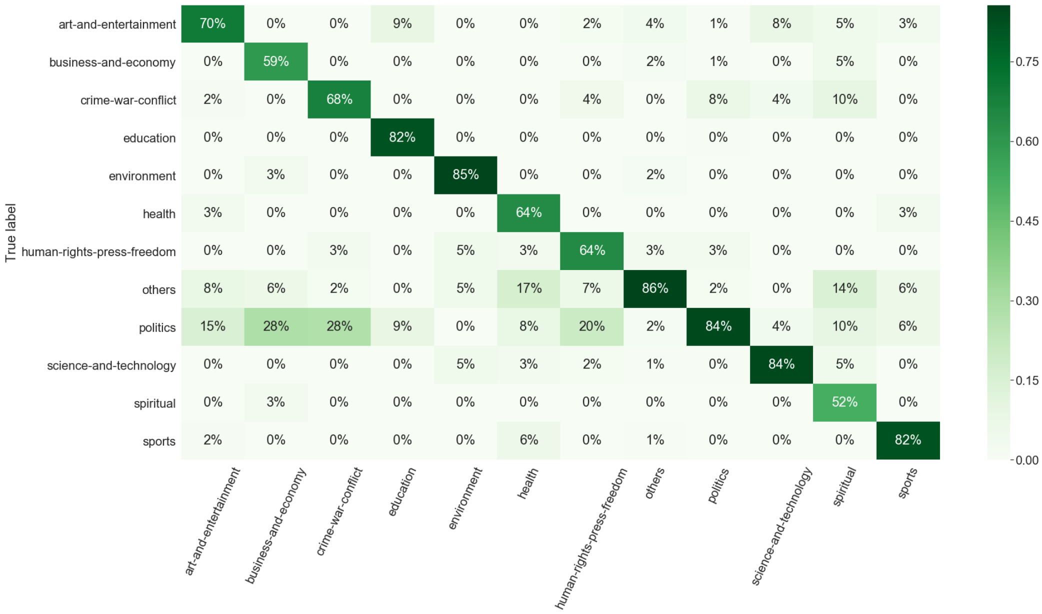

Figure 2: Confusion Matrix for the model QARiB:ASND+SANAD when tested on ASND test set.

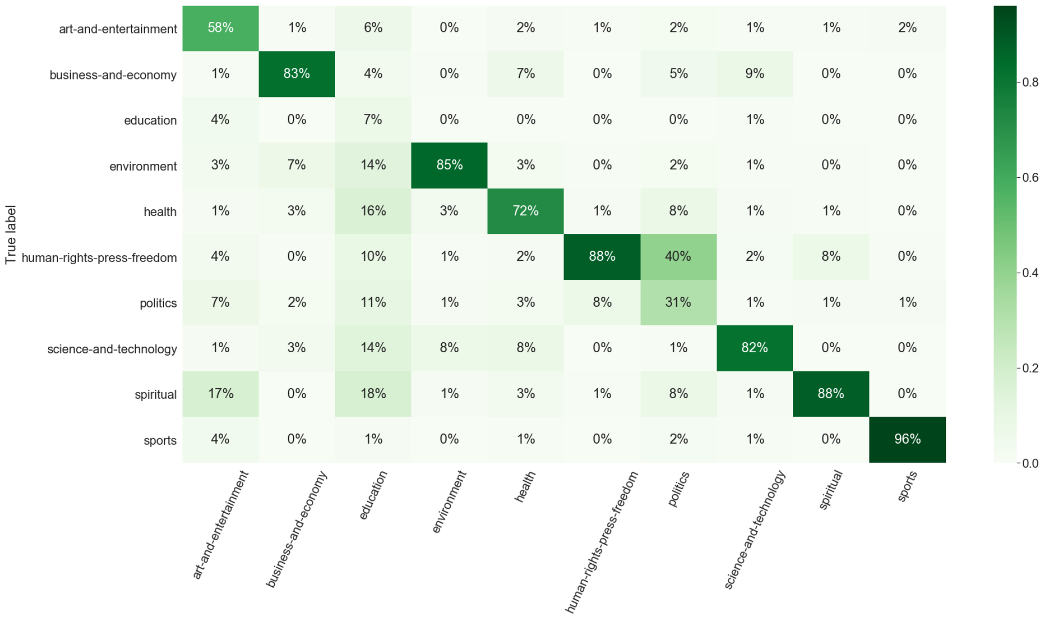

Figure 3: Confusion Matrix for the model QARiB:ASND+SANAD when tested on AITD test set. For AITD , we only considered 10 labels for our evaluation.

2. We observed a similar pattern when testing on AITD, where approximately 40% of *human-rights-press-freedom* tweets were misclassified as *politics* (see Figure 3). Further on AITD, approximately 14% and 8% of *science-and-technology* tweets were misclassified as *education* and *health* respectively. This reflects the contextual closeness of these categories and it might be beneficial to design a hierarchical ontology for such news categorization.

The key observation in this work is that diversifying the training set to cover different genres is important in improving the predictive power of models. Diversification can happen in two ways. One way is to diversify the training data for the specific task. For example, using models trained on SANAD, they were effective on the SANAD test set but yielded sub-optimal results on social media posts, though the posts were written in MSA. Training with both SANAD and ASND together significantly improved results. The second way is to diversify the training data for the pre-trained models such as BERT. As the results show, QARiB, which is trained on both formal text (news) and informal text (tweets) performed at par with AraBERT on the SANAD news headline dataset and significantly outperformed AraBERT on ASND and AITD. As evident by results on ASND, using mixed training data for BERT not only captures linguistic features (MSA vs. dialects) but also capture peculiarities of different platforms such as Twitter.

6 Conclusion

In this paper, we investigated the effect of pre-training a BERT model on a mixture of formal and informal text on text categorization compared to BERT models that were trained exclusively on formal text. We show that the former has greater generalization power, compared to the latter, and is able to significantly classify texts from different sources, such as news headlines and social media posts, and different varieties of Arabic, namely MSA and dialectal Arabic. We also introduced two new Arabic multi-class short text datasets. The first contains social media posts from the official Twitter, Facebook, and YouTube accounts for Aljazeera news channel. Though they are social media posts, they are written in MSA. The second dataset contains tweets from popular Twitter accounts, with large portion of their tweets being authored in dialectal Arabic. The key observation in our work is that diversifying the training data, whether by using diverse training data for a specific task or using diverse data to pre-train a BERT model, leads to overall improvements in classification effectiveness.

References

Mourad Abbas and Kamel Smaili. 2005. Comparison of topic identification methods for arabic language. In *Proceedings of International Conference on Recent Advances in Natural Language Processing, RANLP*, pages 14–17.

Mourad Abbas, Kamel Smaïli, and Daoud Berkani. 2011. Evaluation of topic identification methods on arabic corpora. *J. Digit. Inf. Manag.*, 9(5):185–192.

Muhammad Abdul-Mageed, Hassan Alhuzali, and Mohamed Elaraby. 2018. You tweet what you speak: A city-level dataset of arabic dialects. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Malak Abdullah and Samira Shaikh. 2018. Teamuncc at semeval-2018 task 1: Emotion detection in english and arabic tweets using deep learning. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 350–357.

Deem AlSaleh, Mashael Bin AlAmir, and Souad Larabi-Marie-Sainte. 2020. Snad arabic dataset for deep learning. In *Proceedings of SAI Intelligent Systems Conference*, pages 630–640. Springer.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 9.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Amina Chouigui, Oussama Ben Khiroun, and Bilel Elayeb. 2017. Ant corpus: an arabic news text collection for textual classification. In *2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)*, pages 135–142. IEEE.

Shammur Absar Chowdhury, Hamdy Mubarak, Ahmed Abdelali, Soon-gyo Jung, Bernard J Jansen, and Joni Salminen. 2020. A multi-platform arabic news comment dataset for offensive language detection. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6203–6212.

Michael D Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. 2011. Political polarization on twitter. In *Fifth international AAAI conference on weblogs and social media*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ibrahim Abu El-Khair. 2016. 1.5 billion words arabic corpus. *arXiv preprint arXiv:1611.04033*.

Obeida ElJundi, Wissam Antoun, Nour El Droubi, Hazem Hajj, Wassim El-Hajj, and Khaled Shaban. 2019. hulmona: The universal language model in arabic. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 68–77.

Ashraf Elnagar, Ridhwan Al-Debsi, and Omar Einea. 2020. Arabic text classification using deep learning models. *Information Processing & Management*, 57(1):102121.

Alexander Erdmann, Nasser Zalmout, and Nizar Habash. 2018. Addressing noise in multidialectal word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 558–565.

Rosa Falotico and Piero Quatto. 2015. Fleiss' kappa statistic without paradoxes. *Quality & Quantity*, 49(2):463–470.

Ibrahim Abu Farha and Walid Magdy. 2019. Mazajak: An online arabic sentiment analyser. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 192–198.

Sabit Hassan, Younes Samih, Hamdy Mubarak, Ahmed Abdelali, Ammar Rashed, and Shammur Absar Chowdhury. 2020. Alt submission for osact shared task on offensive language detection. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 61–65.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Yaser Keneshloo, Shuguang Wang, Eui-Hong Han, and Naren Ramakrishnan. 2016. Predicting the popularity of news articles. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pages 441–449. SIAM.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

David D Lewis. 2001. Applying support vector machines to the trec-2001 batch filtering and routing tasks. In *TREC*.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2017. Advances in pre-training distributed word representations. *arXiv preprint arXiv:1712.09405*.

Saif M Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. *arXiv preprint arXiv:1308.6242*.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, SemEval '16, pages 31–41, San Diego, California.

Hamdy Mubarak and Kareem Darwish. 2019. Arabic offensive language classification on twitter. In *International Conference on Social Informatics*, pages 269–276. Springer.

Hamdy Mubarak, Ammar Rashed, Kareem Darwish, Younes Samih, and Ahmed Abdelali. 2020. Arabic offensive language on twitter: Analysis and experiments. *arXiv preprint arXiv:2004.02192*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

John Platt. 1998. Sequential minimal optimization: A fast algorithm for training support vector machines.

Marco Polignano, Pierpaolo Basile, Marco de Gemmis, Giovanni Semeraro, and Valerio Basile. 2019. Alberto: Italian bert language understanding model for nlp challenging tasks based on tweets. In *CLiC-it*.

Abu Bakr Soliman, Kareem Eissa, and Samhaa R El-Beltagy. 2017. Aravec: A set of arabic word embedding models for use in arabic nlp. *Procedia Computer Science*, 117:256–265.

Bashar Talafha, Mohammad Ali, Muhy Eddin Za'ter, Haitham Seelawi, Ibraheem Tuffaha, Mostafa Samir, Wael Farhan, and Hussein T Al-Natsheh. 2020. Multi-dialect arabic bert for country-level dialect identification. *arXiv preprint arXiv:2007.05612*.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 384–394.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5753–5763.

Imad Zeroual, Dirk Goldhahn, Thomas Eckart, and Abdelhak Lakhouaja. 2019. Osian: Open source international arabic news corpus-preparation and integration into the clarin-infrastructure. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 175–182.