A Benchmark Study of Contrastive Learning for Arabic Social Meaning

Md Tawkat Islam Khondaker[†] **El Moatez Billah Nagoudi**[†] **AbdelRahim Elmadany**[†]
Muhammad Abdul-Mageed[†] **Laks V.S. Lakshmanan**

[†]Deep Learning & Natural Language Processing Group
The University of British Columbia

{tawkat@cs.,laks@cs.,muhammad.mageed@}ubc.ca

Abstract

Contrastive learning (CL) brought significant progress to various NLP tasks. Despite this progress, CL has not been applied to Arabic NLP to date. Nor is it clear how much benefits it could bring to particular classes of tasks such as those involved in Arabic social meaning (e.g., sentiment analysis, dialect identification, hate speech detection). In this work, we present a comprehensive benchmark study of state-of-the-art supervised CL methods on a wide array of Arabic social meaning tasks. Through extensive empirical analyses, we show that CL methods outperform vanilla finetuning on most tasks we consider. We also show that CL can be data efficient and quantify this efficiency. Overall, our work allows us to demonstrate the promise of CL methods, including in low-resource settings.

1 Introduction

Proliferation of social media resulted in unprecedented online user engagement. People around the world share their emotions, fears, hopes, opinions, etc. online on a daily basis (Farzindar and Inkpen 2015; Zhang and Abdul-Mageed 2022) on platforms such as Facebook and Twitter. Hence, these platforms offer excellent resources for social meaning tasks such as emotion recognition (Abdul-Mageed and Ungar 2017; Mohammad et al. 2018), irony detection (Van Hee et al. 2018), sarcasm detection (Bamman and Smith 2015), hate speech identification (Waseem and Hovy 2016), stance identification (Mohammad et al. 2016), among others. While the majority of previous social meaning studies were carried out on English, a fast-growing number of investigations focus on other languages. In this paper, we focus on Arabic.

Several works have been conducted on different Arabic social meaning tasks. Some of these focus on Modern Standard Arabic (MSA) (Abdul-Mageed et al. 2011, 2012), while others take Arabic dialects as their target (ElSahar and El-Beltagy

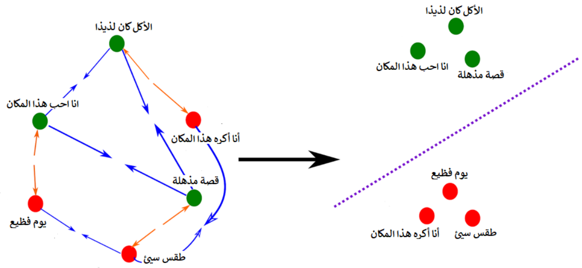


Figure 1: Visual illustration of how supervised contrastive learning works. Representations from the same class are *pulled* close to each other while representations from the different classes are *pushed* further apart.

2015; Al Sallab et al. 2015). While many works have focused on sentiment analysis, e.g., (Abdul-Mageed et al., 2012; Nabil et al., 2015; ElSahar and El-Beltagy, 2015; Al Sallab et al., 2015; Al-Moslmi et al., 2018; Al-Smadi et al., 2019; Al-Ayyoub et al., 2019; Farha and Magdy, 2019) and dialect identification (Elfardy and Diab, 2013; Zaidan and Callison-Burch, 2011, 2014; Cotterell and Callison-Burch, 2014; Zhang and Abdul-Mageed, 2019; Bouamor et al., 2018; Abdul-Mageed et al., 2020b,a, 2021b), others focused on detection of user demographics such as age and gender (Zaghouani and Charfi 2018; Rangel et al. 2019), irony detection (Karoui et al. 2017; Ghanem et al. 2019), and emotion analysis (Abdul-Mageed et al. 2016; Alhuzali et al. 2018). Our interest in the current work is improving Arabic social meaning through representation learning.

In spite of recent progress in representation learning, most work in Arabic social meaning mostly focuses on finetuning language models such as AraT5 (Nagoudi et al., 2022), CamelBERT (Inoue et al., 2021), MARBERT (Abdul-Mageed et al., 2021a), QARIB (Abdelali et al., 2021), among others. In particular, Arabic social media processing has to date ignored the emerging sub-area of contrastive learning (CL) (Hadsell et al. 2006). Given a labeled dataset, CL (Khosla et al., 2020) attempts

Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP), pages 63 - 75
December 8, 2022 ©2022 Association for Computational Linguistics

to pull representations of the same class close to each other while pushing representations of different classes further apart (Figure 1). In this work, we investigate five different supervised contrastive learning methods in the context of Arabic social meaning. To the best of our knowledge, this is the first work that provides a comprehensive study of supervised contrastive learning on a wide range of Arabic social meanings. We show that performance of CL methods can be task-dependent. We attempt to explain this performance from the perspective of task specificity (i.e., how fine-grained the labels of a given task are). We also show that contrastive learning methods generally perform better than vanilla finetuning based on cross entropy (CE). Through an extensive experimental study, we also demonstrate that CL methods outperform CE finetuning under resource-limited constraints. Our work allows us to demonstrate the promise of CL methods in general, and in low-resource settings in particular.

To summarize, we offer the following contributions:

1. We study a comprehensive set of supervised CL methods for a wide range of Arabic social meaning tasks, including abusive language and hate speech detection, emotion and sentiment analysis, and identification of demographic attributes (e.g. age, gender).

2. We show that CL-based methods outperform generic CE-based vanilla finetuning for most of the tasks. To the best of our knowledge, this is the first work that provides an extensive study of supervised CL on Arabic social meaning.

3. We empirically find that improvements CL methods result in are task-specific and attempt to understand this finding in the context of the different tasks we consider with regard to their label granularity.

4. We demonstrate that CL methods can achieve better performance under limited data constraints, emphasizing and quantifying how well these can work for low-resource settings.

2 Related Works

2.1 Arabic Social Meaning

We use the term *social meaning* (SM) to refer to meaning arising in real-world communication in social media (Thomas, 2014; Zhang et al., 2022b). SM covers tasks such as sentiment analysis (Abdul-Mageed et al., 2012; Abu Farha et al., 2021; Saleh et al., 2022; Alali et al., 2022), emotion recognition (Alhuzali et al., 2018; Mubarak et al., 2022c; Abu Shaqra et al., 2022; Mansy et al., 2022), age and gender identification (Abdul-Mageed et al., 2020c; Abbes et al., 2020; Mubarak et al., 2022b; Mansour Khoudja et al., 2022), hate-speech and offensive language detection (Elmadany et al., 2020a; Mubarak et al., 2020, 2022a; Husain and Uzuner, 2022), and sarcasm detection (Farha and Magdy, 2020; Wafa'Q et al., 2022; Abdullah et al., 2022).

Most of the recent studies are transformers-based. They directly finetune pre-trained models such as mBERT (Devlin et al., 2018), MARBERT (Abdul-Mageed et al., 2021a), and AraT5 (Nagoudi et al., 2022) on SM datasets like (Abdul-Mageed et al., 2020c; Alshehri et al., 2020; Abuzayed and Al-Khalifa, 2021; Nessir et al., 2022), using data augmentation (Elmadany et al., 2020b), ensampling (Mansy et al., 2022; Alzu'bi et al., 2022), and multi-tasks (Abdul-Mageed et al., 2020b; Shapiro et al., 2022; AlKhamissi and Diab, 2022). However, to the best of our knowledge, there is no published research studying CL on Arabic language understanding in general nor social meaning processing in paticular.

2.2 Contrastive Learning

CL aims to learn effective embedding by pulling semantically close neighbors together while pushing apart non-neighbors (Hadsell et al. 2006). CL employs a CL-based similarity objective to learn the embedding representation in the hyperspace (Chen et al., 2017; Henderson et al., 2017). In computer vision, Chen et al. (2020a) propose a framework for contrastive learning of visual representations without specialized architectures or a memory bank. Khosla et al. (2020) shows that supervised contrastive loss can outperform CL loss on ImageNet (Russakovsky et al., 2015). In NLP, similar methods have been explored in the context of sentence representation learning (Karpukhin et al., 2020; Gillick et al., 2019; Logeswaran and Lee, 2018; Zhang et al., 2022a). Among the most notable works is Gao et al. (2021) who propose unsupervised CL framework, *SimCSE*, that predicts input sentence itself by augmenting it with dropout as noise.

Recent works have been studying CL extensively for improving both semantic text similarity (STS) and text classification tasks (Meng et al. 2021; Qu et al. 2020; Qiu et al. 2021; Janson et al. 2021). Fang et al. (2020) propose back-translation as a source of positive pair for NLU tasks. Klein and Nabi (2022) argue that feature decorrelation between high and low dropout projected representations improves STS tasks. Zhou et al. (2022) design an instance weighting method to penalize false negatives and generate noise-based negatives to guarantee the uniformity of the representation space. Su et al. (2022) propose a token-aware CL method by contrasting the token from the same sequence to improve the uniformity in the embedding space. We now formally introduce these CL methods and how we employ them in our work.

3 Methods

Given a set of training examples $\{x_i, y_i\}_{i=1,\ldots,N}$ and an encoder based on a pre-trained language model (PLM), f outputs contextualized token representation of x_i,

$$H = \{\ h_{[CLS]},\ h_1,\ h_2,\ ...,\ h_{[SEP]}\ \} \quad (1)$$

Where $\boldsymbol{H}$ is the hidden representation of the final layer of the encoder.

The standard practice of finetuning PLMs passes the pooled representation $h_{[CLS]}$ of `[CLS]` to a softmax classifier to obtain the probability distribution for the set of classes $\mathbf{C}$ (Figure 2a).

$$p(y_c|h_{[CLS]}) = softmax\ (\mathbf{W}h_{[CLS]});\ \ c \in \mathbf{C} \quad (2)$$

Where $\mathbf{W} \in \mathcal{R}^{d_C \times d_h}$ are trainable parameters and d_h is hidden dimension. The model is trained with the objective of minimizing cross-entropy (CE) loss,

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\ log(p(y_{i,c}|h_{i_{[CLS]}}))^1 \quad (3)$$

3.1 Supervised Contrastive Loss (SCL)

The objective of supervised contrastive loss (Khosla et al. 2020) is to pull the representations of the same class close to each other while pushing the representations of different classes further apart. Following Gao et al. (2021), we adopt dropout-based data augmentation where for each representation h_i, we produce an equivalent dropout-based representation h_j and consider h_j as having the same label as h_i (Figure 2b). The model attempts to minimize NTXent loss (Chen et al., 2020a). The purpose of NTXent loss is to take each in-batch representation as an anchor and minimize the distance between the anchor(h_i) and the representations from the same class (P_i) while maximizing the distance between the anchor and the representation from different classes,

$$\mathcal{L}_{NTX} = \sum_{i=1}^{2N} \frac{-1}{P_i} \sum_{j \in P_i} \log \frac{e^{sim(h_i,h_j)/\tau}}{\sum_{k=1}^{2N} 1_{i \neq k} e^{sim(h_i,h_k)/\tau}} \quad (4)$$

Where τ is used to regulate the temperature. The final loss for SCL is

$$\mathcal{L}_{SCL} = (1-\lambda)\mathcal{L}_{CE} + \lambda\mathcal{L}_{NTX}$$

3.2 Contrastive Adversarial Training (CAT)

Instead of dropout-based augmentation, Pan et al. (2022) propose to generate adversarial examples applying *fast gradient sign method* (FGSM) (Goodfellow et al., 2015). Formally, *FGSM* attempts to maximize $\mathcal{L}_{CE}$ by adding a small perturbation r bounded by ϵ,

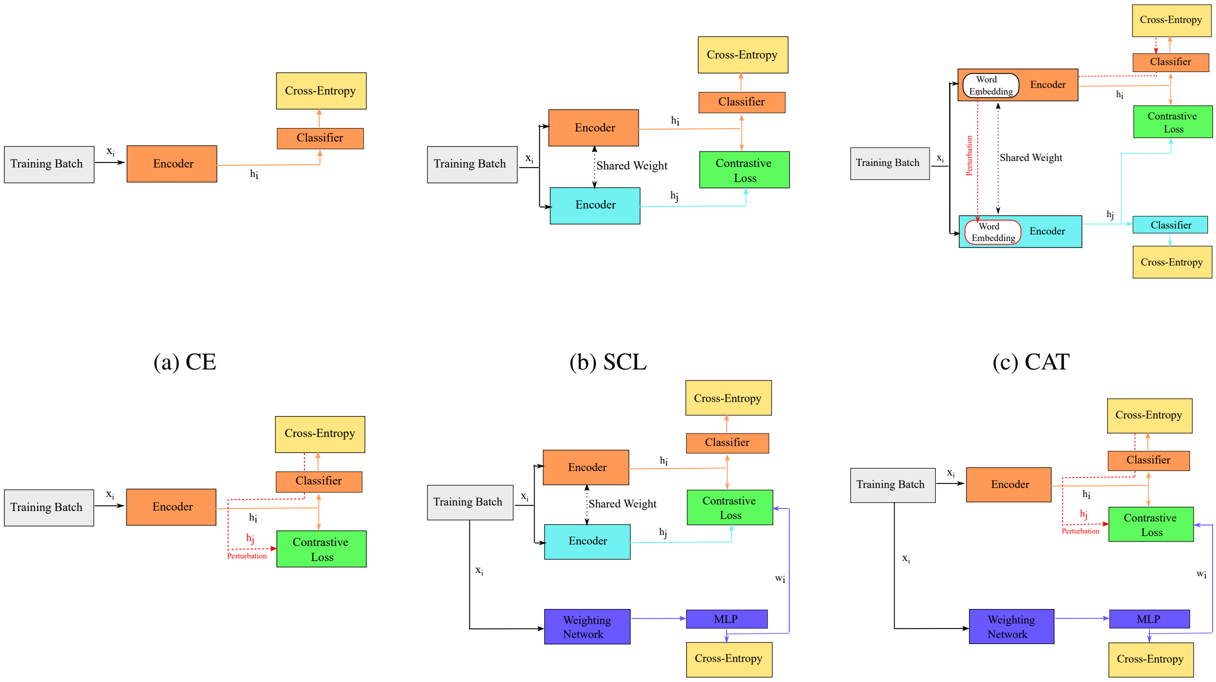
$$max\mathcal{L}_{CE} = arg\max_{r} \mathcal{L}(f(x_i + r, y_i))$$
$$s.t.\ ||r|| < \epsilon,\quad \epsilon > 0 \quad (5)$$

Goodfellow et al. (2015) approximate the perturbation r with a linear approximation around x_i and an *L2* norm constraint. However, Pan et al. (2022) propose to approximate r around the word embedding matrix $V \in \mathcal{R}^{d_V \times d_h}$ (Figure 2c), where d_V is the vocabulary size. Hence, the adversarial perturbation is computed as,

$$r = -\epsilon \frac{\nabla_V \mathcal{L}(f(x_i, y_i))}{||\nabla_V \mathcal{L}(f(x_i, y_i))||_2} \quad (6)$$

After receiving x_i, the perturbed encoder f^{V+r} outputs `[CLS]` representation h_j, which is treated as the positive pair of h_i. Both h_i and h_j are passed through a non-linear projection layer and the resulting representations are used to train the model with InfoNCE loss (Oord et al., 2018).

[1] $h_{i_{[CLS]}}$ and h_i are used interchangeably in the rest of the paper.



(a) CE (b) SCL (c) CAT

(d) TACT (e) LCL (f) TLCL

Figure 2: Illustration of supervised contrastive learning methods used in this work.

$$z_i = \mathbf{W_2} ReLU(\mathbf{W_1} h_i) \tag{7}$$

$$z_j = \mathbf{W_2} ReLU(\mathbf{W_1} h_j) \tag{8}$$

$$\mathcal{L}_{InfoNCE} = -\log \frac{e^{sim(z_i, z_j)/\tau}}{\sum_{k=1}^{2N} 1_{i \neq k} e^{sim(z_i, z_k)/\tau}} \tag{9}$$

The final loss is calculated as,

$$\mathcal{L}_{CAT} = \frac{1-\lambda}{2}(\mathcal{L}_{CE} + \mathcal{L}_{CE}^{V+r}) + \lambda \mathcal{L}_{InfoNCE}$$

3.3 Token-level Adversarial Contrastive Training (TACT)

We also study a variant of CAT where instead of perturbing the word embedding matrix V, we directly perturb the token representations h_i (Figure 2d),

$$r = -\epsilon \frac{\nabla_{h_i} \mathcal{L}(f(x_i, y_i))}{||\nabla_{h_i} \mathcal{L}(f(x_i, y_i))||_2} \tag{10}$$

$$h_j = h_i + r \tag{11}$$

Similar to CAT, we pass h_i and h_j through a non-linear projection layer and use the obtained representations to train the model to minimize InfoNCE loss (Eq. 9). We compute the final loss as,

$$\mathcal{L}_{CAT} = \frac{1-\lambda}{2}(\mathcal{L}_{CE} + \mathcal{L}_{CE}^{h+r}) + \lambda \mathcal{L}_{InfoNCE} \tag{12}$$

3.4 Label-aware Contrastive Loss (LCL)

Suresh and Ong (2021) propose to adapt contrastive loss for fine-grained classification tasks by incorporating inter-label relationships. The authors propose an additional weighting network (Figure 2e) to encode the inter-label relationships. First, both the encoder and the weighting network are optimised using cross-entropy loss ($\mathcal{L}_{CE}$), $\mathcal{L}_E$, and $\mathcal{L}_w$, respectively. The prediction probabilities obtained from the softmax layer of the weighting network are used to compute the confidence of the current sample for a given class c,

$$\mathbf{w}_{i,c} = \frac{e^{h_{i,c}}}{\sum_{k=1}^{C} e^{h_{i,k}}} \tag{13}$$

These weights are then used to train the model with NTXent loss.

$$\mathcal{L}_i = \sum_{j \in P_i} \log \frac{w_{i,y_i} \cdot e^{sim(h_i, h_j)/\tau}}{\sum_{k=1}^{2N} 1_{i \neq k} w_{i,y_k} \cdot e^{sim(h_i, h_k)/\tau}} \tag{14}$$

$$\mathcal{L}_f = \sum_{i=1}^{2N} \frac{-\mathcal{L}_i}{P_i} \tag{15}$$

Similar to Section 3.1, we use dropout-based data augmentation. Given a confusable sample, the weighting network will assign higher scores for

Dataset	Train	Dev	Test	No. of Classes
Abusive	4,677	584	585	3
Adult	33,690	5,000	5000	2
Age	5,000	5,000	5,000	3
AraNeT$_{\text{emo}}$	50,000	910	941	8
Dangerous	3,474	615	663	2
Dialect at BinaryLevel	50,000	5,000	5,000	2
Dialect at CountryLevel	50,000	5,000	5,000	21
Dialect at RegionLevel	38,271	4,450	5000	4
Gender	50,000	5,000	5,000	2
Hate Speech	6,839	1,000	2,000	2
Irony	3,621	403	805	2
Offensive	6,839	1,000	2,000	2
Sarcasm	7,593	844	2,110	2
SemEval$_{\text{emo}}$	3,376	661	1,563	4
Sentiment Analysis	49,301	4,443	4,933	3

Table 1: Statistics of datasets used in our experiments.

the classes that are more closely associated with the sample. Incorporating these high values back into the denominator of NTXent will steer the encoder toward finding more distinguishing patterns to differentiate between confusable samples. The final LCL loss is computed as follows:

$$\mathcal{L}_{LCL} = (1 - \lambda)(\mathcal{L}_E + \mathcal{L}_w) + \lambda\mathcal{L}_f \tag{16}$$

3.5 Token Adversarial LCL (TLCL)

Instead of dropout-oriented representation as an augmentation, we experiment with token adversarial representation for LCL (Figure 2f) described in Section 3.3. First, we compute the adversarial representation h_j using Eq. 10 and Eq. 11. Then, we compute NTXent loss (Eq. 14) for LCL to obtain the final token adversarial LCL loss, $\mathcal{L}_{TLCL}$. We now describe our datasets.

4 Datasets

In this section, we present the Arabic social meaning tasks and datasets used in our study. A summary of the datasets is presented in Table 1.

Abusive and Adult Content. For the abusive and adult content detection tasks, we use datasets from Mubarak et al. (2017) and Mubarak et al. (2021). These datasets consist of 1.1k and 43k tweets, respectively. For these datasets, the goal is to classify an Arabic tweet into one of the two classes in the set, i.e., *{obscene, clean}* for the abusive task, and *{adult, not-adult}* for the adult content detection task.

Age and Gender. For both tasks, we use the *Arap-Tweet* dataset (Zaghouani and Charfi, 2018) which consists of *1.3M, 160k, 160k* for the Train, Dev, and Test respecctively. The dataset covers 11 Arab regions. Zaghouani and Charfi (2018) assign age group labels from the set *{under-25, 25-to-34, above-35}* and gender from the set *{male, female}*.

Dangerous. We use the dangerous speech dataset from Alshehri et al. (2020). This dataset consists of $4,445$ manually annotated tweets labelled as either *safe* or *dangerous*.

Dialect Identification: Six datasets are used for this task: ArSarcasm$_{\text{Dia}}$ (Farha and Magdy, 2020), the Arabic Online Commentary (AOC) (Zaidan and Callison-Burch, 2014), NADI-2020 (Abdul-Mageed et al., 2020a), MADAR (Bouamor et al., 2019), QADI (Abdelali et al., 2020), and Habibi (El-Haj, 2020). The dialect identification task involves three dialect classification levels: (1) Binary-level (*MSA* vs. *DIA*), (2) Region-level (4 *regions*), and (3) Country-level (21 *countries*).

Emotion. For this task, we use two datasets: *AraNeT$_{emo}$* and *SemEval$_{emo}$*. The first one is proposed by Abdul-Mageed et al. (2020c). The dataset consists of 192K tweets labeled with the eight emotion classes from the set {*anger, anticipation, disgust, fear, joy, sadness, surprise, trust*}. *SemEval$_{emo}$* (Mohammad et al., 2018) consists of $5,603$ tweets labeled with four emotions from the set {*anger, fear, joy, sadness*}.

Offensive Language and Hate Speech. We use the dataset released by Mubarak et al. (2020) during an offensive and hate speech shared task.[2] This dataset consists of $10k$ manually annotated tweets with four tags *{offensive, not-offensive, hate, not-hate*

Irony. We use the irony identification dataset for Arabic tweets (IDAT) developed by Ghanem et al. (2019). This dataset contains $5,030$ MSA and dialectal tweets. It is labeled with *ironic* and *non-ironic* tags.

Sarcasm. We use the *ArSarcasm* dataset released by (Farha and Magdy, 2020). *ArSarcasm* contains $10,547$ tweets. The tweets are labeled with *sarcasm* and *not-sarcasm* tags.

Sentiment Analysis This task includes 19 sentiment datasets. We merge the 17 datasets benchmarked by Abdul-Mageed et al. (2021a) with two new datasets: Arabizi sentiment analysis dataset (Fourati et al., 2020) and AraCust (Almuqren and Cristea, 2021), a Saudi Telecom Tweets corpus for sentiment analysis. The data contains *190k, 6.5k, 44.2k* samples for Train, Dev and Test. The dataset is labeled with three tags from the set {*positive*, *negative*, *neutral*}.

5 Experimental Setup

We implement all the methods using MARBERT (Abdul-Mageed et al., 2021a) (`UBC-NLP/MARBERT`) from HuggingFace's Transformers library (Wolf et al., 2020), as the backbone architecture. We use MARBERT as it is reported to achieve SOTA on a wide range of Arabic language understanding tasks in Abdul-Mageed et al. (2021a). Our methods, however, can be applied to any other model. We use the same hyperparameters for all the methods to ensure fair comparisons. We set the maximum sequence length to 128 and use a batch size of 16 to train the models using Adam optimizer with a learning rate $5e-5$. The initial number of training epochs is set to 25 with an early stopping threshold of 5. For CL-based models, we set λ to 0.5 and τ to 0.3. For all the experiments, we consider the checkpoint with the best macro F_1 score on the development sets to evaluate performance on the respective test sets. To limit GPU usage during our experiments, we normalize all datasets considered by limiting the size of Train, Dev, and Test splits to *50k, 5k, 5k* samples respectively.[3]

[2] http://edinburghnlp.inf.ed.ac.uk/workshops/OSACT4/

[3] For example, for the *Age* and *Gender* datasets, Train, Dev, and Test splits have 1.3m, 160k, and 160k, respectively. So, we randomly pick 50k, 5k, and 5k samples respectively.

6 Results

As explained, we compare different methods on 15 different Arabic social media datasets involving binary and multiclass classification. We present performance of the methods in Table 2. Evidently, CL-based methods achieve better performance on majority of the tasks. On average, three out of five CL-based methods (LCL, SCL, and TACT) achieve better performance than CE-MARBERT. Overall, LCL achieves the best F_1-score averaging across all the tasks.

It is important to note that there is no unique superior method across the tasks. This shows that CL-based methods can be task-specific, depending on the nature of how they are formulated. For example, LCL performs well on multiclass datasets such as *Abusive* and $AraNeT_{emo}$, while TLCL performs well on $SemEval_{emo}$. LCL and TLCL adopt more fine-grained representations with the incorporation of the weighting network which consequently helps them distinguish confused classes. However, for *Dialect at RegionLevel*, we speculate that since the labels are already fine-grained, it is more important to improve the robustness rather than inter-label relationship. Therefore, CAT achieves best performance on this task, followed by TLCL. Similarly, on binary classification tasks such as *hate speech* and *Offensive language detection*, where a subtle semantic change in meaning can alter the labels, robust methods are expected to outperform others. Therefore, adversarial methods like CAT and TACT achieve better F_1-score.

For most of the tasks, F_1-scores obtained from different CL-methods are close to each other and the vanilla SCL achieves similar average score to the other models. This proves that although task-specific formulation may help the models to improve on a certain task, the most important factor evolves around the fundamental *minmax* nature of contrastive learning which is minimizing the distance among the representations of the same class while maximizing the distance among the representations of the different classes.

7 Analysis

7.1 Data Efficiency

To investigate how the methods perform with limited data, we train the models under different size constraints using three datasets (one binary and

	CE	SCL	CAT	TACT	LCL	TLCL
Abusive	77.15	78.09	76.48	75.69	**78.32**	75.26
Adult	88.16	**89.50**	86.54	89.13	88.85	89.48
Age	44.22	45.12	42.28	**46.45**	45.90	43.20
AraNeT$_{emo}$	62.47	61.49	59.31	57.99	62.56	**64.13**
Dangerous	61.44	63.76	67.83	66.00	65.76	**69.28**
Dialect at BinaryLevel	85.71	85.63	**86.67**	84.98	85.79	81.84
Dialect at CountryLevel	32.84	**33.63**	33.24	32.69	33.62	31.34
Dialect at RegionLevel	65.29	64.78	**65.54**	64.56	62.92	64.92
Gender	62.23	63.56	65.58	65.77	**65.90**	65.14
Hate Speech	80.91	80.00	71.06	**82.62**	81.00	75.26
Irony	**84.75**	84.30	84.72	84.18	84.29	83.43
Offensive	90.43	89.92	**91.37**	91.23	90.41	88.84
Sarcasm	70.67	71.09	72.09	74.14	**75.32**	69.40
SemEval$_{emo}$	79.25	77.22	77.08	77.85	**80.61**	78.59
Sentiment Analysis	**77.69**	77.32	76.89	76.68	75.61	74.82
Avg.	70.88	71.03	70.45	71.33	**71.79**	70.33

Table 2: Macro F1-score of the models on Arabic social media datasets. Here, *CE* = Cross-Entropy; *SCL* = Supervised Contrastive Learning; *CAT* = Contrastive Adversarial Training; *TACT* = Token-level Adversarial Contrastive Training; *LCL* = Label-aware Contrastive Loss; *TLCL* = Token Adversarial LCL.

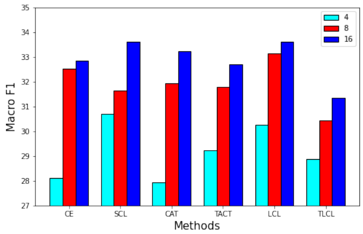
	Dialect-Country				**Dialect-Region**				**AraNeT$_{emo}$**			
	10%	**25%**	**50%**	**100%**	**10%**	**25%**	**50%**	**100%**	**10%**	**25%**	**50%**	**100%**
CE	27.78	30.5	30.91	32.84	63.09	63.16	63.59	65.29	53.85	56.73	59.18	62.47
SCL	28.49	**31.87**	**32.89**	**33.63**	63.08	63.23	63.37	64.78	54.47	58.35	58.35	61.49
CAT	26.57	30.33	32.71	33.24	**64.32**	**65.3**	**65.42**	**65.54**	54.75	54.03	55.51	59.31
TACT	27.63	29.88	32.04	32.69	63.8	64.1	64.32	64.56	53.27	**59.3**	59.18	57.99
LCL	**28.97**	30.5	31.78	33.62	63.72	64.72	65.06	62.92	**55.47**	59.25	62.21	62.56
TLCL	27.69	30.44	32.18	31.34	62.71	64.53	64.6	64.92	54.62	59.31	**62.98**	**64.13**

Table 3: Model performance on varying dataset sizes. **Bold** values represent the best performance for a particular dataset and dataset size.

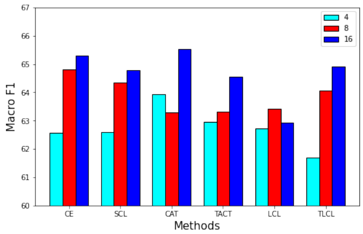
two multiclass). We present results of this set of experiments in Table 3. One interesting observation is that improvement in performance is not always monotonic with respect to data size. We believe that larger-sized training sets only aid models with test samples with idiosyncrasies and that small training sets sufficiently cover a wide range of data distributions. However, we observe that CE-MARBERT fails to outperform CL-based methods in any constraint. Specifically, for *Dialect at CountryLevel* dataset, 50% of the data is sufficient for SCL to outperform CE-MARBERT trained on the full dataset. Additionally, CAT achieves comparable performance to CE-MARBERT with 50% training data. For *Dialect at RegionLevel* dataset, only 10% training data is sufficient for CAT, TACT, and LCL to outperform CE-MARBERT with 50% training data. Moreover, CAT requires only 50% training data to outperform CE-MARBERT with full training data. Finally, for *AraNeT$_{emo}$* dataset, LCL, TACT, and TLCL with 25% training data outperform CE-MARBERT with 50% training data. TLCL with 50% data outperforms CE-MARBERT with full (i.e., 100%) training data while LCL with 50% data achieves similar performance. *This analysis shows that enhancing the representations of different classes via CL helps the model to produce more distinguishable clusters. As a result, the models require only smaller training data to project a sample to a particular class.*

7.2 Impact of Batch Size

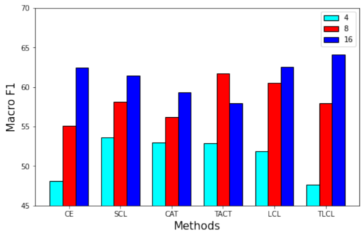
We study how batch size affects model performance. We consider batch sizes of 4, 8, 16 on three datasets, showing performance in Figure 3. We observe that, with only a few exceptions, performance of the models increases along with the increase of batch size. Larger batch sizes contain more samples from different classes, which helps the model to learn better via comparing these samples. Our



(a) Dialect at CountryLevel

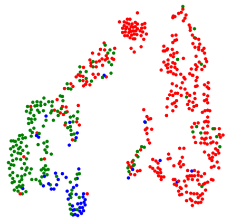


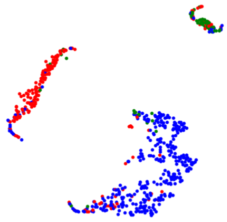
(b) Dialect at RegionLevel

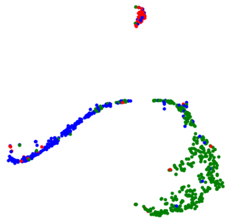


(c) AraNeT$_{\text{emo}}$

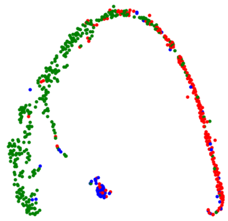
Figure 3: Ablation study on the impact of batch size on performance of the models.


(a) CE


(b) SCL


(c) CAT


(d) TACT


(e) LCL

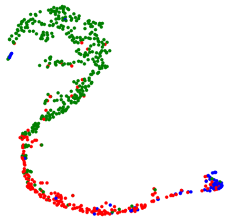

(f) TLCL

Figure 4: t-SNE representations of the validation set of *abusive* dataset (green = normal, red = abusive, blue = hate).

analysis corroborates findings of prior works such as Chen et al. (2020b), Cao et al. (2022), and Qiu et al. (2021) that propose the incorporation of a separate memory bank to hold the negative samples for comparison.

7.3 Visualization of Representations

We plot t-SNE representations of the test samples from the *Abusive* dataset in Figure 4. The representations are colored with true labels. We notice that CL-based methods cluster *normal* and *abusive* samples far from each other, unlike CE-MARBERT. Since CL attempts to maximize the distance between different classes, it helps the models produce more distinct clusters. Additionally, LCL and TLCL methods cluster *abusive* and *hate* classes better than other methods. Since, they capture inter-label relations, the methods identify confusable examples of *abusive* and *hate* better than other methods.

8 Limitations

An inherent limitation of CL methods is their reliance on hyperparameters. In particular, they are sensitive to batch size. Larger batch sizes usually yield better performance. Other hyperparameters like τ and λ can also impact performance given a specific task. Lastly, the accommodation of larger batch size comes at the cost of higher computational resources.

9 Conclusion

In this work, we study various supervised contrastive learning methods for a wide range of Arabic social meaning tasks. We show that CL-based methods outperform generic cross entropy finetuning for majority of the tasks. Through empirical investigations, we find that improvements resulting from applying CL methods are task-specific. We interpret these results vis-a-vis different downstream tasks, with a special attention to the number of classes involved in each task. Finally, we demonstrate that CL methods can achieve better performance with limited training data and hence can be employed for low-resource settings.

In the future, we plan to extend our work beyond sentence classification by experimenting on tasks such as token-classification and question-answering. Our work stands as a comprehensive investigation of applying contrastive learning to Arabic social meaning. We hope this work will trigger further investigations of CL in Arabic NLP in general.

References

Ines Abbes, Wajdi Zaghouani, Omaima El-Hardlo, and Faten Ashour. 2020. Daict: A dialectal Arabic irony corpus extracted from twitter. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6265–6271.

Ahmed Abdelali, Sabit Hassan, Hamdy Mubarak, Kareem Darwish, and Younes Samih. 2021. Pre-training bert on arabic tweets: Practical considerations. *arXiv preprint arXiv:2102.10684*.

Ahmed Abdelali, Hamdy Mubarak, Younes Samih, Sabit Hassan, and Kareem Darwish. 2020. Arabic Dialect Identification in the Wild. *Proceedings of the Sixth Arabic Natural Language Processing Workshop*.

Muhammad Abdul-Mageed, Hassan AlHuzli, and Mona Diab DuaaAbu Elhija. 2016. Dina: A multi-dialect dataset for arabic emotion analysis. In *The 2nd workshop on Arabic corpora and processing tools*, page 29.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021a. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Muhammad Abdul-Mageed, Mohammed Korayem, and Ahmed YoussefAgha. 2011. “Yes we can?": Subjectivity annotation and tagging for the health domain. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 666–671.

Muhammad Abdul-Mageed, Sandra Kübler, and Mona Diab. 2012. SAMAR: A system for subjectivity and sentiment analysis of arabic social media. In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, pages 19–28. Association for Computational Linguistics.

Muhammad Abdul-Mageed and Lyle Ungar. 2017. EmoNet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 718–728.

Muhammad Abdul-Mageed, Chiyu Zhang, Houda Bouamor, and Nizar Habash. 2020a. NADI 2020: The first nuanced Arabic dialect identification shared task. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 97–110, Barcelona, Spain (Online). Association for Computational Linguistics.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, Houda Bouamor, and Nizar Habash. 2021b. NADI 2021: The second nuanced Arabic dialect identification shared task. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 244–259, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, and Lyle Ungar. 2020b. Toward micro-dialect identification in diaglossic and code-switched environments. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5855–5876, Online. Association for Computational Linguistics.

Muhammad Abdul-Mageed, Chiyu Zhang, Azadeh Hashemi, et al. 2020c. AraNet: A deep learning toolkit for arabic social media. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 16–23.

Malak Abdullah, Dalya Alnore, Safa Swedat, Jumana Khrais, and Mahmoud Al-Ayyoub. 2022. Sarcasmdet at semeval-2022 task 6: Detecting sarcasm using pre-trained transformers in english and arabic languages. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1025–1030.

Ibrahim Abu Farha, Wajdi Zaghouani, and Walid Magdy. 2021. Overview of the WANLP 2021 shared task on sarcasm and sentiment detection in Arabic. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 296–305, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Ftoon Abu Shaqra, Rehab Duwairi, and Mahmoud Al-Ayyoub. 2022. A multi-modal deep learning system for arabic emotion recognition. *International Journal of Speech Technology*, pages 1–17.

Abeer Abuzayed and Hend Al-Khalifa. 2021. Sarcasm and sentiment detection in arabic tweets using bert-based models and data augmentation. In *Proceedings of the sixth Arabic natural language processing workshop*, pages 312–317.

Mahmoud Al-Ayyoub, Abed Allah Khamaiseh, Yaser Jararweh, and Mohammed N Al-Kabi. 2019. A comprehensive survey of arabic sentiment analysis. *Information Processing & Management*, 56(2):320–342.

Tareq Al-Moslmi, Mohammed Albared, Adel Al-Shabi, Nazlia Omar, and Salwani Abdullah. 2018. Arabic senti-lexicon: Constructing publicly available language resources for arabic sentiment analysis. *Journal of Information Science*, 44(3):345–362.

Ahmad Al Sallab, Hazem Hajj, Gilbert Badaro, Ramy Baly, Wassim El Hajj, and Khaled Bashir Shaban. 2015. Deep learning models for sentiment analysis in arabic. In *Proceedings of the second workshop on Arabic natural language processing*, pages 9–17.

Mohammad Al-Smadi, Bashar Talafha, Mahmoud Al-Ayyoub, and Yaser Jararweh. 2019. Using long short-term memory deep neural networks for aspect-based sentiment analysis of arabic reviews. *International Journal of Machine Learning and Cybernetics*, 10(8):2163–2175.

Muath Alali, Nurfadhlina Mohd Sharef, Masrah Azrifah Azmi Murad, Hazlina Hamdan, and Nor Azura Husin. 2022. Multitasking learning model based on hierarchical attention network for arabic sentiment analysis classification. *Electronics*, 11(8):1193.

Hassan Alhuzali, Muhammad Abdul-Mageed, and Lyle Ungar. 2018. Enabling deep learning of emotion with first-person seed expressions. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 25–35.

Badr AlKhamissi and Mona Diab. 2022. Meta AI at Arabic hate speech 2022: MultiTask learning with self-correction for hate speech classification. In *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, pages 186–193, Marseille, France. European Language Resources Association.

Latifah Almuqren and Alexandra Cristea. 2021. Aracust: a saudi telecom tweets corpus for sentiment analysis. *PeerJ Computer Science*, 7:e510.

Ali Alshehri, Muhammad Abdul-Mageed, et al. 2020. Understanding and detecting dangerous speech in social media. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 40–47.

Salaheddin Alzu'bi, Thiago Castro Ferreira, Lucas Pavanelli, and Mohamed Al-Badrashiny. 2022. aixplain at arabic hate speech 2022: An ensemble based approach to detecting offensive tweets.

David Bamman and Noah Smith. 2015. Contextualized sarcasm detection on twitter. In *proceedings of the international AAAI conference on web and social media*, volume 9, pages 574–577.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. The MADAR Arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Houda Bouamor, Sabit Hassan, and Nizar Habash. 2019. The madar shared task on arabic fine-grained dialect identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 199–207.

Rui Cao, Yihao Wang, Yuxin Liang, Ling Gao, Jie Zheng, Jie Ren, and Zheng Wang. 2022. Exploring the impact of negative samples of contrastive learning: A case study of sentence embedding. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3138–3152, Dublin, Ireland. Association for Computational Linguistics.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020a. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

Ting Chen, Yizhou Sun, Yue Shi, and Liangjie Hong. 2017. On sampling strategies for neural network-based collaborative filtering. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 767–776, New York, NY, USA. Association for Computing Machinery.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. 2020b. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*.

Ryan Cotterell and Chris Callison-Burch. 2014. A multi-dialect, multi-genre corpus of informal written arabic. In *LREC*, pages 241–245.

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*.

Mahmoud El-Haj. 2020. Habibi-a multi dialect multi national arabic song lyrics corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1318–1326.

Heba Elfardy and Mona T Diab. 2013. Sentence level dialect identification in arabic. In *ACL (2)*, pages 456–461.

AbdelRahim Elmadany, Chiyu Zhang, Muhammad Abdul-Mageed, and Azadeh Hashemi. 2020a. Leveraging affective bidirectional transformers for 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 102–108.

AbdelRahim Elmadany, Chiyu Zhang, Muhammad Abdul-Mageed, and Azadeh Hashemi. 2020b. Leveraging affective bidirectional transformers for offensive language detection. In *The 4th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT4), LREC*.

Hady ElSahar and Samhaa R El-Beltagy. 2015. Building large arabic multi-domain resources for sentiment analysis. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 23–34. Springer.

Hongchao Fang, Sicheng Wang, Meng Zhou, Jiayuan Ding, and Pengtao Xie. 2020. Cert: Contrastive self-supervised learning for language understanding. *arXiv preprint arXiv:2005.12766*.

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.

Ibrahim Abu Farha and Walid Magdy. 2020. From Arabic Sentiment Analysis to Sarcasm Detection: The ArSarcasm Dataset. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 32–39.

Atefeh Farzindar and Diana Inkpen. 2015. Natural language processing for social media. *Synthesis Lectures on Human Language Technologies*, 8(2):1–166.

Chayma Fourati, Abir Messaoudi, and Hatem Haddad. 2020. Tunizi: a tunisian arabizi sentiment analysis dataset. In *AfricaNLP Workshop, Putting Africa on the NLP Map. ICLR 2020, Virtual Event*, volume arXiv:3091079.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bilal Ghanem, Jihen Karoui, Farah Benamara, Véronique Moriceau, and Paolo Rosso. 2019. Idat@fire2019: Overview of the track on irony detection in arabic tweets. In *Mehta P., Rosso P., Majumder P., Mitra M. (Eds.) Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2019). CEUR Workshop Proceedings. In: CEUR-WS.org, Kolkata, India, December 12-15*.

Daniel Gillick, Sayali Kulkarni, Larry Lansing, Alessandro Presta, Jason Baldridge, Eugene Ie, and Diego Garcia-Olano. 2019. Learning dense representations for entity retrieval. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 528–537, Hong Kong, China. Association for Computational Linguistics.

Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*.

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-hsuan Sung, Laszlo Lukacs, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply.

Fatemah Husain and Ozlem Uzuner. 2022. Investigating the effect of preprocessing arabic text on offensive language and hate speech detection. *Transactions on Asian and Low-Resource Language Information Processing*, 21(4):1–20.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 92–104, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Sverker Janson, Evangelina Gogoulou, Erik Ylipää, Amaru Cuba Gyllensten, and Magnus Sahlgren. 2021. Semantic re-tuning with contrastive tension. In *International Conference on Learning Representations, 2021*.

Jihen Karoui, Farah Banamara Zitoune, and Veronique Moriceau. 2017. Soukhria: Towards an irony detection system for arabic in social media. *Procedia Computer Science*, 117:161–168.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673. Curran Associates, Inc.

Tassilo Klein and Moin Nabi. 2022. SCD: Self-contrastive decorrelation of sentence embeddings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 394–400, Dublin, Ireland. Association for Computational Linguistics.

Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. In *International Conference on Learning Representations*, volume abs/1803.02893.

Asmaa Mansour Khoudja, Mourad Loukam, and Fatma Zohra Belkredim. 2022. Towards author profiling from modern standard arabic texts: A review. In *Proceedings of Sixth International Congress on Information and Communication Technology*, pages 745–753. Springer.

Alaa Mansy, Sherine Rady, and Tarek Gharib. 2022. An ensemble deep learning approach for emotion detection in arabic tweets. *International Journal of Advanced Computer Science and Applications*, 13(4).

Yu Meng, Chenyan Xiong, Payal Bajaj, Paul Bennett, Jiawei Han, Xia Song, et al. 2021. Coco-lm: Correcting and contrasting text sequences for language model pretraining. *Advances in Neural Information Processing Systems*, 34:23102–23114.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

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-2016)*, pages 31–41.

Hamdy Mubarak, Hend Al-Khalifa, and AbdulMohsen Al-Thubaity. 2022a. Overview of osact5 shared task on arabic offensive language and hate speech detection.

Hamdy Mubarak, Shammur Absar Chowdhury, and Firoj Alam. 2022b. Arabgend: Gender analysis and inference on arabic twitter. *arXiv preprint arXiv:2203.00271*.

Hamdy Mubarak, Kareem Darwish, and Walid Magdy. 2017. Abusive language detection on Arabic social media. In *Proceedings of the First Workshop on Abusive Language Online*, pages 52–56, Vancouver, BC, Canada. Association for Computational Linguistics.

Hamdy Mubarak, Kareem Darwish, Walid Magdy, Tamer Elsayed, and Hend Al-Khalifa. 2020. Overview of OSACT4 Arabic offensive language detection shared task. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 48–52, Marseille, France. European Language Resource Association.

Hamdy Mubarak, Sabit Hassan, and Ahmed Abdelali. 2021. Adult content detection on Arabic Twitter: Analysis and experiments. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 136–144, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Hamdy Mubarak, Sabit Hassan, and Shammur Absar Chowdhury. 2022c. Emojis as anchors to detect arabic offensive language and hate speech. *arXiv preprint arXiv:2201.06723*.

Mahmoud Nabil, Mohamed Aly, and Amir F Atiya. 2015. Astd: Arabic sentiment tweets dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2515–2519.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2022. AraT5: Text-to-text transformers for Arabic language generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 628–647, Dublin, Ireland. Association for Computational Linguistics.

Mohamed Aziz Ben Nessir, Malek Rhouma, Hatem Haddad, and Chayma Fourati. 2022. icompass at arabic hate speech 2022: Detect hate speech using qrnn and transformers.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Lin Pan, Chung-Wei Hang, Avirup Sil, and Saloni Potdar. 2022. Improved text classification via contrastive adversarial training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11130–11138.

Yao Qiu, Jinchao Zhang, and Jie Zhou. 2021. Improving gradient-based adversarial training for text classification by contrastive learning and auto-encoder. *arXiv preprint arXiv:2109.06536*.

Yanru Qu, Dinghan Shen, Yelong Shen, Sandra Sajeev, Jiawei Han, and Weizhu Chen. 2020. Coda: Contrast-enhanced and diversity-promoting data augmentation for natural language understanding. *arXiv preprint arXiv:2010.08670*.

Francisco Rangel, Paolo Rosso, Anis Charfi, Wajdi Zaghouani, Bilal Ghanem, and Javier Sánchez-Junquera. 2019. Overview of the track on author profiling and deception detection in arabic. *Working Notes of FIRE 2019. CEUR-WS. org, vol. 2517*, pages 70–83.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3).

Hager Saleh, Sherif Mostafa, Abdullah Alharbi, Shaker El-Sappagh, and Tamim Alkhalifah. 2022. Heterogeneous ensemble deep learning model for enhanced arabic sentiment analysis. *Sensors*, 22(10):3707.

Ahmad Shapiro, Ayman Khalafallah, and Marwan Torki. 2022. Alexu-aic at arabic hate speech 2022: Contrast to classify. *arXiv preprint arXiv:2207.08557*.

Yixuan Su, Fangyu Liu, Zaiqiao Meng, Tian Lan, Lei Shu, Ehsan Shareghi, and Nigel Collier. 2022. TaCL: Improving BERT pre-training with token-aware contrastive learning. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2497–2507, Seattle, United States. Association for Computational Linguistics.

Varsha Suresh and Desmond Ong. 2021. Not all negatives are equal: Label-aware contrastive loss for fine-grained text classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4381–4394, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jenny A Thomas. 2014. *Meaning in interaction: An introduction to pragmatics*. Routledge.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. Semeval-2018 task 3: Irony detection in english tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 39–50.

Al-Jamal Wafa'Q, Ahmad M Mustafa, and Mostafa Z Ali. 2022. Sarcasm detection in arabic short text using deep learning. In *2022 13th International Conference on Information and Communication Systems (ICICS)*, pages 362–366. IEEE.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Wajdi Zaghouani and Anis Charfi. 2018. Arap-tweet: A large multi-dialect twitter corpus for gender, age and language variety identification. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Omar F Zaidan and Chris Callison-Burch. 2011. The arabic online commentary dataset: an annotated dataset of informal arabic with high dialectal content. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 37–41. Association for Computational Linguistics.

Omar F Zaidan and Chris Callison-Burch. 2014. Arabic dialect identification. *Computational Linguistics*, 40(1):171–202.

Chiyu Zhang and Muhammad Abdul-Mageed. 2019. No army, no navy: Bert semi-supervised learning of arabic dialects. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 279–284.

Chiyu Zhang and Muhammad Abdul-Mageed. 2022. Improving social meaning detection with pragmatic masking and surrogate fine-tuning. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 141–156.

Chiyu Zhang, Muhammad Abdul-Mageed, and Ganesh Jawahar. 2022a. Infodcl: A distantly supervised contrastive learning framework for social meaning. *arXiv preprint arXiv:2203.07648.*

Chiyu Zhang, Muhammad Abdul-Mageed, and El Moatez Billah Nagoudi. 2022b. Decay no more: A persistent twitter dataset for learning social meaning. *Workshop Proceedings of the 16th International AAAI Conference on Web and Social Media.*

Kun Zhou, Beichen Zhang, Xin Zhao, and Ji-Rong Wen. 2022. Debiased contrastive learning of unsupervised sentence representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6120–6130, Dublin, Ireland. Association for Computational Linguistics.