

A deep CNN architecture with novel pooling layer applied to two Sudanese Arabic sentiment data sets

Journal of Information Science
1–22

© The Author(s) 2023



Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/01655515231188341

journals.sagepub.com/home/jis



Mustafa Mhamed 

School of Information Science and Technology, Northwest University, China; College of Information and Electrical Engineering, China Agricultural University, China

Richard Sutcliffe 

School of Information Science and Technology, Northwest University, China; School of Computer Science and Electronic Engineering, University of Essex, UK

Husam Quteineh

Business and Local Government Data Research Centre, School of CSEE, University of Essex, UK

Xia Sun

School of Information Science and Technology, Northwest University, China

Eiad Almekhlafi 

School of Information Science and Technology, Northwest University, China

Ephrem Afele Retta

School of Information Science and Technology, Northwest University, China

Jun Feng

School of Information Science and Technology, Northwest University, China

Abstract

Arabic sentiment analysis has become an important research field in recent years. Initially, work focused on Modern Standard Arabic (MSA), which is the most widely used form. Since then, work has been carried out on several different dialects, including Egyptian, Levantine and Moroccan. Moreover, a number of data sets have been created to support such work. However, up until now, no work has been carried out on Sudanese Arabic, a dialect which has 32 million speakers. In this article, two new public data sets are introduced, the two-class Sudanese Sentiment Data set (SudSenti2) and the three-class Sudanese Sentiment Data set (SudSenti3). In the preparation phase, we establish a Sudanese stopword list. Furthermore, a convolutional neural network (CNN) architecture, Sentiment Convolutional MMA (SCM), is proposed, comprising five CNN layers together with a novel Mean Max Average (MMA) pooling layer, to extract the best features. This SCM model is applied to SudSenti2 and SudSenti3 and shown to be superior to the

Corresponding authors:

Richard Sutcliffe, School of Computer Science and Electronic Engineering, University of Essex, UK.

Emails: rsutcl@essex.ac.uk; rsutcl@nwu.edu.cn

Xia Sun, School of Information Science and Technology, Northwest University, Xi'an 710069, China.

Email: rainy@nwu.edu.cn

Jun Feng, School of Information Science and Technology, Northwest University, Xi'an 710069, China.

Email: fengjun@nwu.edu.cn

baseline models, with accuracies of 92.25% and 85.23% (Experiments 1 and 2). The performance of MMA is compared with Max, Avg and Min and shown to be better on SudSenti2, the Saudi Sentiment Data set and the MSA Hotel Arabic Review Data set by 1.00%, 0.83% and 0.74%, respectively (Experiment 3). Next, we conduct an ablation study to determine the contribution to performance of text normalisation and the Sudanese stopword list (Experiment 4). For normalisation, this makes a difference of 0.43% on two-class and 0.45% on three-class. For the custom stoplist, the differences are 0.82% and 0.72%, respectively. Finally, the model is compared with other deep learning classifiers, including transformer-based language models for Arabic, and shown to be comparable for SudSenti2 (Experiment 5).

Keywords

Arabic dialects; Arabic text preprocessing; convolutional neural network; neural networks; pooling layer; sentiment analysis; sentiment data set; Sudanese

1. Introduction

Sentiment analysis is an important field because it enables us to discover voices and opinions relating to topics of interest in a particular context, for example, views about political issues in elections or opinions about products or ways of providing services [1]. With the emergence of participatory web services in areas such as education and health, there is a need for sentiment analysis in identifying problems and hence upgrading quality standards. Recently, the spread of Arabic content, especially on social media, and the application of artificial intelligence and deep learning in analysing Arabic sentiments, has led researchers to delve deeper into Arabic text. Initially, this work has been carried out on Modern Standard Arabic (MSA). However, more recent work has also been concerned with the regional Arabic dialects that are often used in everyday informal communications.

The World Arabic Language Dialects map¹ indicates 21 Arabic dialects and shows the different regions of the world in which they are spoken. This can give us hints about how dialects are related to each other. Table 1 (derived from istizada.com²) shows the number of speakers for eight of the most important dialects. As can be seen, Sudanese Arabic is the fifth most widely spoken dialect, with 32 million speakers. This is why we have concentrated on Sudanese in this work.

There are a significant number of variations between dialects and MSA in terms of language:

- MSA has a dual form of short vowel, omitted in written text, in addition to the singular and plural vowel forms, for masculine and feminine [2]. Dialects often do not create such uniqueness between the sexes; instead they have an open system which is more complex than MSA, allowing the prefix and suffix to be attached to a base, and pronouns to function as indirect objects.

In the Arabic language, diacritics³ are not normally used. However, they can be found in certain contexts such as poetry, religious texts, including the Quran and Hadiths of the Prophet, and textbooks for teaching Arabic at a beginner's level.

According to Almekhlafi et al., for example, **كتب أحمد** can have three meanings in MSA when we apply different diacritics: **كَتَبَ أَحْمَدُ** means 'Ahmed wrote', **كُتِبَ أَحْمَدُ** means 'Ahmed's books'⁴ and **كُتِبَ أَحْمَدُ** means 'Ahmed was written' [3].

Similarly, Hadj et al. [4] state that diacritics are used in educational and religious literature to clarify ambiguity, modify the melody and ensure the correct interpretations.

- The Arabic language is rich in its vocabulary and due to the common use of some subvocabulary in a particular region, these words may make this subvocabulary a characteristic dialect [5]. Moreover, because of the interaction between civilisations, some vocabulary from other languages has entered Arabic [6]. For instance, words from Spanish and French are used in the Moroccan dialect, while in the Sudanese dialect, some Turkish and English words appear. Examples are shown in Table 2.
- There are differences in the conjugation of verbs, even though the root is retained [7]. For example, in MSA, the verb conjugation for the root **ل - ع - ب** is **يلعب**, but in different Arabic dialects, it can have different forms, for example, **يلعب - يلعب - يلعب**. [8]

During the spread of Arabic, different regions had their own spoken languages (e.g. Berber in North Africa and Coptic in Egypt and Sudan). When Arabic was introduced to these regions, most of the original languages fused with it, creating different dialects. Over the centuries, each of these communities underwent different transitions as they were exposed to different linguistic influences [9].

Table 1. Dialects of Arabic (derived from istizada.com).

Dialect	Areas spoken	Number of speakers
Egyptian	Egypt	64,500,000
Gulf	Bahrain, Kuwait, Oman, Qatar, Saudi Arabia, UAE	36,056,000
Sudanese	Sudan, Southern Egypt	31,940,300
Hassaniya	Mauritania, southern Morocco, south western Algeria, Western Sahara	8,842,800
Levantine	Lebanon, Jordan, Palestine, Syria	36,188,500
Maghrebi	Algeria, Libya, Morocco, Tunisia	32,608,700
Mesopotamian/Iraqi	Iraq, eastern Syria	15,655,900
Yemeni	Yemen, Somalia, Djibouti, southern Saudi Arabia	14,360,000

UAE: the United Arab Emirates.

Table 2. Dialect variation (based on Oueslati et al. [10] with Sudanese additions).

MSA word	Dialectal word	Arabizi	Country	English equivalent
حلو/جميل Jameel	حلو	7elew	Lebanon	Nice
	حلو	7ilew	Saudi	
	حلو	7low/hlow	Tunisia	
جدا Jiddan	سمح	Samh	Sudan	A lot, too much
	كثير	ktir	Lebanon	
	وايد	wayed	Emirate	
	أوي	2awi	Egypt	
	كثير	kityer	Sudan	
دراجة Darraja	برشا	barcha	Tunisia	Bicycle
	بسكلات	besklet	Tunisia	
	درافة	darraga	Egypt	
	عجلة	Ajala	Sudan	

MSA: Modern Standard Arabic.

Sudanese vocabulary is mostly inspired by MSA, but with important Greek, Turkish and English modifications to the phonology. The morphology of Sudanese words shares many features with MSA, but the method of dialectal inflexion is more complicated than MSA in some respects [8].

Table 3 illustrates the differences between MSA and Sudanese dialect by means of some examples.

First, MSA 'اليوم' ('today') corresponds to Sudanese 'الليلة' (literally 'night'). Even though the word is derived from 'night' in MSA, it nevertheless means 'day' in Sudanese. Here, the word meaning is completely reversed. Second, 'room' in MSA is 'الغرفة' while in Sudanese, it is 'الأوضة'. Many such nouns are expressed differently. Another example is the names of popular foods and beverages, such as 'coffee' (MSA 'قهوة' and Sudanese 'جينة'). Verbs can also differ; the verb 'find' in MSA is 'تجد' while in Sudanese, it is 'تشف'. Generally, we can see many differences in vocabulary as well as variations in grammar and means of expression.

Following a thorough study of such dialect differences, we have created two data sets based on social media posts, built a convolutional neural network (CNN)-based model for sentiment analysis and applied it to the data sets (Figure 1). The main contributions of this work are as follows:

- We create a two-class Sudanese sentiment data set (SudSenti2) from Facebook and YouTube.
- We build a three-class Sudanese sentiment data set (SudSenti3) from Twitter.
- We design a Sudanese stopword list and use it for text normalisation in the preprocessing phase.
- We offer free access to the data sets and analyses.⁵
- We propose a model called Sentiment Convolutional MMA (SCM) which is a five-layer CNN incorporating our Mean Max Average (MMA) pooling layer.

Table 3. Examples of differences between MSA and the Sudanese dialect.

English language	Standard Arabic	Sudanese dialect
Be careful today don't go out of the room.	كن حذراً اليوم لا تخرج من الغرفة .	اعمل حسابك الليلة ما تطلع من الأوضه .
Keep going down this road until you find a pharmacy.	إستمر في السير في هذا الطريق حتي تجد صيدلية .	أمشي دوغري لغاية تشوف اجزخانة .
Quit drinking too much coffee because it is not good for your health.	اترك شرب الكثير من القهوة لأنها غير مفيدة لصحتك .	سيب الجبنة الكثيرة ما كويسة لجسمك .

MSA: Modern Standard Arabic.

- We apply SCM to SudSenti2 and Sudesenti3, as well as to existing MSA and Saudi data sets, with good results.
- We compare the proposed MMA pooling layer to the standard pooling layer used in other works and show that it gives the best performance.
- We carry out an ablation study to demonstrate the effect of using text normalisation and the Sudanese stopword list.
- We compare SCM with other machine learning (ML), deep learning methods and Arabic transformer models such as ARBERT and MARBERT. SCM gives a high classification performance.
- Finally, we fine tune the best-performing transformer, MARBERT, with our hyperparameters, resulting in enhanced accuracy.

This article is organised as follows. Section 2 reviews previous work on sentiment analysis for Arabic. Section 3 describes the creation of the SudSenti2 and SudSenti3 data sets. Section 4 outlines the proposed model architecture. Section 5 presents our experiments, including preprocessing steps, experimental settings, baselines, results and discussion. Finally, section 6 draws conclusions and suggests future work.

2. Related work

As we have mentioned, the Arabic language is very widespread in the world and is spoken in many dialects. Some of these have already been the subject of sentiment analysis research (Table 4). Here, we discuss which dialects have been studied, what data sets were used and what sentiment analysis techniques were adopted (for a recent review of Arabic sentiment analysis, please also refer to Alharbi et al. [11]).

Tabii et al. [14] used Naïve Bayes (NB), Maximum Entropy (ME) and Support Vector Machines (SVMs) on two data sets, the Moroccan Sentiment Analysis Corpus (MSAC) [25], comprising tweets from Twitter and comments from

Table 4. Previous work on sentiment analysis for different Arabic dialects.

Paper	Arabic dialect	Split	Model	Result
Alwehaibi and Roy [12]	Saudi (2C)	80 + 20	LSTM-RNN	93.5%
Alahmary and Al-Dossari [13]	Saudi (3C)	90 + 10	CNN	86.54%
Tabii et al. [14]	Moroccan (2C)	90 + 10	Majority Voting	83.45%
Lulu and Elnagar [15]	Egyptian, Iraqi and Levantine (3C)	80 + 10 + 10	LSTM	71.4%
Al-Saqqa et al. [16]	Jordanian (2C)	90 + 10	Ensemble	93.4%
Al Omari et al. [17]	Lebanon (2C)	80 + 20	LR	89.80%
Abdelli et al. [18]	Algerian (2C)	85 + 15	SVM	0.86%
Mulki et al. [19]	Tunisian (2C)	80 + 10 + 10	Deep-LSTM	90.00%
Mulki et al. [19]	JEG, TAC and TSAC (2C)	90 + 10	Tw-StAR	82.08%
Abdel-Salam [20]	ArSarcasm	70 + 10 + 20	MHLGG	95.5%
Abdul-Mageed et al. [21]	Arabic corpus	80 + 10 + 10	Transformer	95.00%
Mhamed et al. [22]	Egyptian, MSA (2C)(n-C)	80 + 10 + 10	MCI, MC2	92.96%
Al-shaibani et al. [23]	Modern Standard Arabic	85 + 15	BiGRU	94.32%
Addi and Ezzahir [24]	Modern Standard Arabic	80 + 20	RF + SMOTE	96.00%

LSTM-RNN: Long-Short Term Memory Recurrent Neural Network; CNN: Convolutional Neural Network; LR: Logistic Regression; SVM: Support Vector Machine; Tw-StAR: Tweets-Sentiment Analysis of Arabic; MHLGG: Multi-Headed-LSTM-CNN-GRU; MCI: Multiple Classification I; MC2: Multiple Classification 2; BiGRU: Bidirectional Gated Recurrent Unit; RF + SMOTE: Random Forest + Synthetic Minority Oversampling Technique.

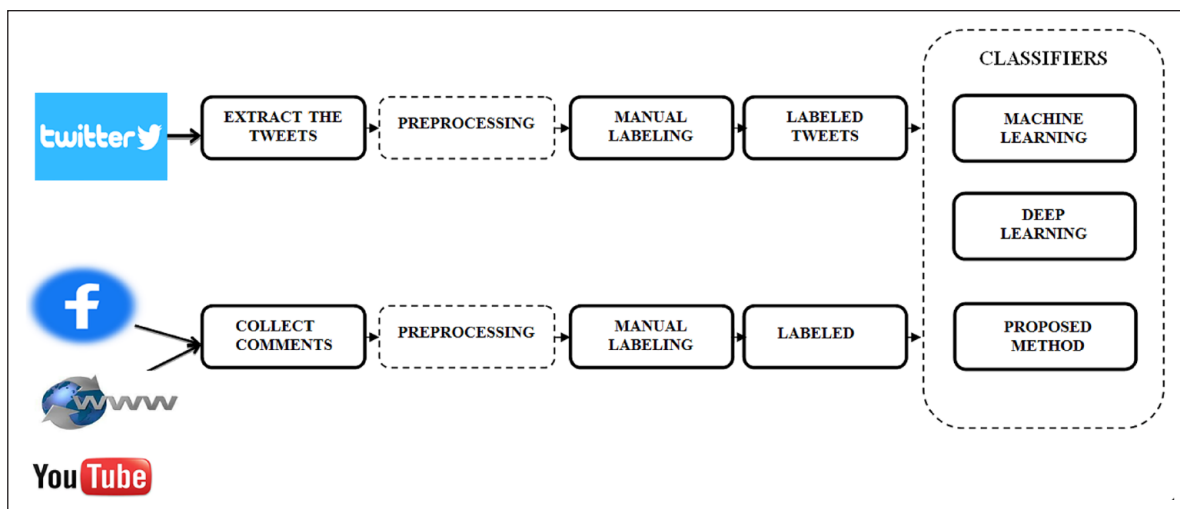


Figure 1. Overall description of work.

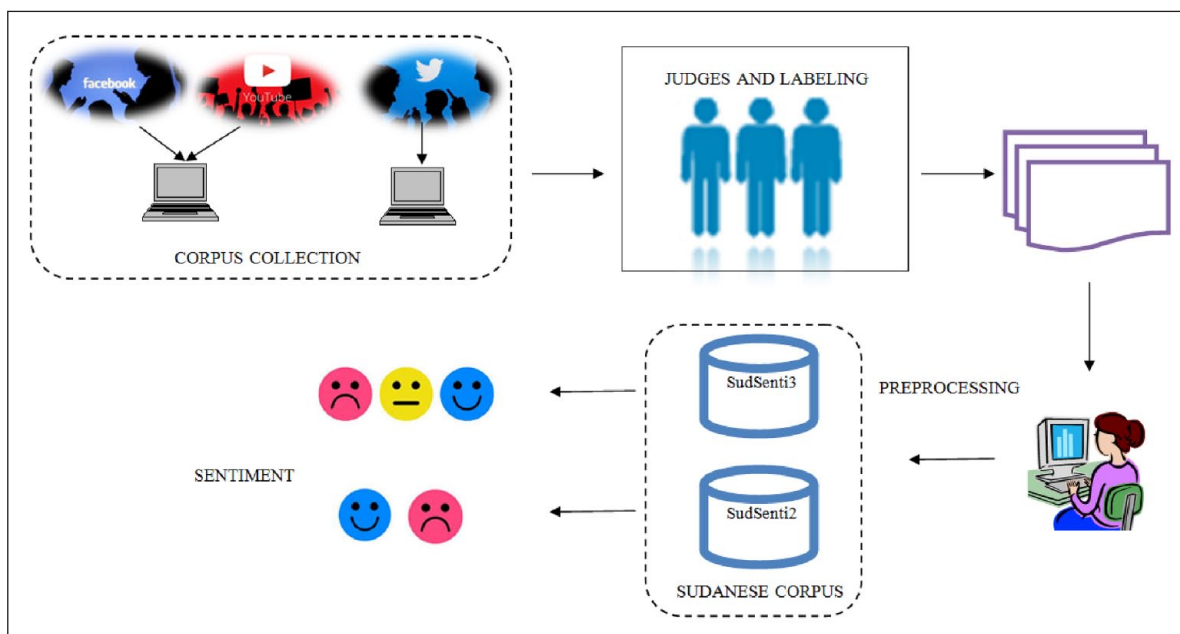


Figure 2. Overall description of data set collection.

Facebook and YouTube. SVM was the best single classifier, measured by accuracy (82.5%). The best ensemble classifier combined SVM, NB and ME classifiers with majority voting (83.45%).

Lulu and Elnagar [15] utilised Long-Short Term Memory (LSTM) [26,27], CNNs, Bidirectional Long-Short Term Memory (BLSTM) and Convolutional Long-Short Term Memory (CLSTM) on three Arabic data sets, Arabic Online Commentary (AOC) [28], Egyptian (EGP), Gulf including Iraqi (GLF) and Levantine (LEV). Results show that LSTM attained the highest accuracy (71.4%), followed by CLSTM (71.1%) and BLSTM (70.9%). It can be observed that the CNN model suffered from overfitting problems as shown by the difference between the cross-validation and test results.

Al-Saqqa et al. [16] applied NB, SVM, Decision Trees (DTs) and K-Nearest Neighbour (KNN) algorithms on four Arabic data sets – Opinion Corpus for Arabic (OCA) [29], MSA, Crawler tweets 2014 data sets [30] and the Large-scale Arabic Sentiment Analysis Data set (LABR) [31]. The aim was to determine the emotions of the Arabic text, using methods based on bigrams and voting combinations. Accuracy was 93.4%, better than the individual classifiers.

Al Omari et al. [17] used Logistic Regression (LR) with a Term Frequency–Inverted Document Frequency (TF-IDF) weighting model for feature extraction, on Arabic Services Reviews in Lebanon (ASRL) collected from Google reviews and the Zomato website in the Lebanon dialect. For positive classifications, $P=0.88$ and $R=1.00$, and for negative, $P=0.80$ and $R=0.80$. Thus, the positive result is better than the negative. Abdelli et al. [18] utilised SVM and LSTM on both Modern Arabic and the Algerian dialect [32]. The results for SVM and LSTM on the Algerian data set were 86% and 81%, respectively.

Alwehaibi and Roy [12] applied a Long-Short Term Memory Recurrent Neural Network (RNN) (LSTM-RNN) on the AraSenTi data set which comprises tweets written in MSA and Saudi dialect, manually annotated for Sentiment. Arabic word embeddings used Word2Vec, GloVe and Fasttext [33]. The LSTM-RNN model achieved 93.5% accuracy.

Alahmary and Al-Dossari [13] built another Saudi corpus, this time with three classes, produced using a semi-automatic annotation method starting with NB, followed by hand correction. They then applied SVM, LSTM, BLSTM and CNN classifiers. The highest performance was CNN (86.54%).

Jerbi et al. [34] used RNN, LSTM, BLSTM and Deep-LSTM [35] on the Tunisian Sentiment Analysis Corpus (TSAC). Deep-LSTM had the highest accuracy (90.00%). Mulki et al. [19] experimented with the effect of Named Entity Recognition (NER), using SVM and NB on four Arabic data sets, Jordanian Egyptian Gulf (JEG), Tunisian Arabic Corpus (TAC), Tunisian Election Corpus (TEC) [36] and TSAC [37]. The highest accuracy recorded was on TSAC (82.8%).

Abdel-Salam [20] applied Multi-headed-LSTM-CNN-GRU (MHLCG) and MARBERT on ArSarcasm-v2 [38]. Accuracies were 95.5% and 94.4%, respectively. MHLCG was more effective than the MARBERT model based on BERT and transformers.

Abdul-Mageed et al. [21] introduced the ARBERT and MARBERT deep bidirectional transformers and applied them to various Arabic tasks. Performance on the binary TwitterSaad⁶ was the best with an accuracy of 95.00%. The Arabic sentiment tweets data set (ASTD) (three-class) [39] and Blog Posts Sentiment Corpus (BBN) (three-class) [40] scores were 78.00% and 79.00%, respectively.

Mhamed et al. [22] presented a comprehensive new Arabic preprocessing approach, and then designed two architectures, MC1 and MC2. On the difficult ASTD data set [39], for the four-class task, accuracy was 73.17%, on three-class, it was 78.62%, and on two-class, it was 90.06%. On the large two-class Arabic Twitter Data For Sentiment (ATDFS) data set [41], their model worked effectively, with a performance of 92.96%.

Al-shaibani et al. [23] use an RNN-based approach on a data set of Arabic poems (55,440 verses and 14 m). A five-layer Bidirectional Gated Recurrent Unit (BiGRU) gave the best performance (94.32%).

Addi and Ezzahir [24] applied SVM, NB and Random Forest (RF) with two techniques – under-sampling and over-sampling. They used the Hotel Arabic Reviews Data set (HARD) – Imbalanced [42]. RF with Synthetic Minority Oversampling (SMO) gave the best accuracy (96.00%).

Here, we create two Sudanese Arabic sentiment data sets, one two-class and one three-class. After detailed preprocessing, we apply the proposed classifier SCM + MMA and compare its performance with ML, NN classifiers and Arabic transformers.

3. Data set creation

In this work, two new data sets for Sudanese are proposed. The SudSenti2 was created from Facebook and YouTube (Figure 3). The SudSenti3 was created from Twitter posts (Figure 4).

3.1. SudSenti2 data set

The following steps were carried out (Figure 2):

1. Texts were collected from Facebook⁷ and YouTube.⁸
2. All texts matching one of the following queries were downloaded, using the Orange Data Mining software:⁹ 'تنوع الثقافات المختلفة في السودان', 'هل يساعد السلام مستقبلا في التطور والتنمية ام تقاسم كعك', 'السودان بلد غني'. This resulted in 4544 matching posts.
3. Three judges were chosen to classify the posts. All were university teachers who were native speakers of Sudanese Arabic. All judges judged all posts.
4. Posts which were not considered Sudanese by at least two of the three judges were deleted.
5. Each post was then classified as negative, positive or neutral (Neutral posts were subsequently deleted.). A text is considered positive if it contains joyful, happy or amusing vocabulary, or if there is a positive emoji, or if there is more than one emotion, but the positive feeling is dominant. A text is negative if it contains negative, disappointed,

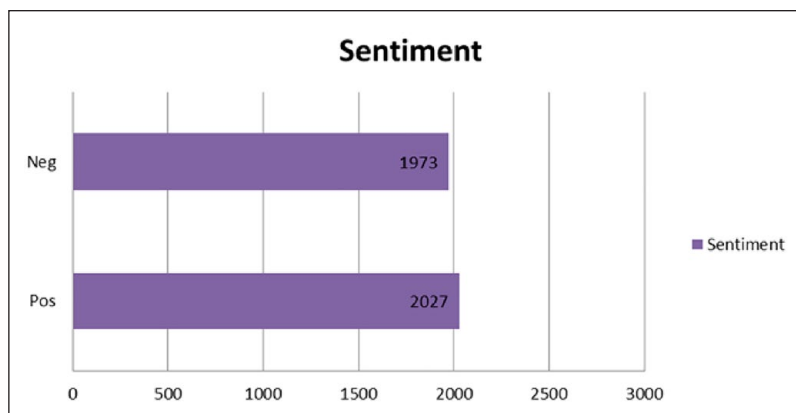


Figure 3. Total numbers of tweets for each class in SudSenti2.

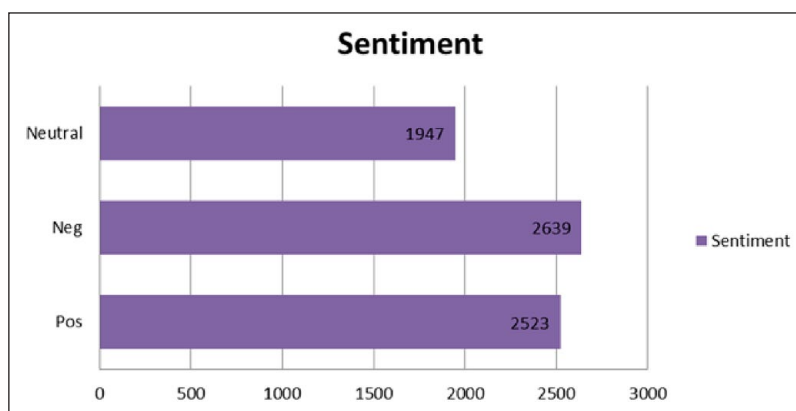


Figure 4. Total numbers of tweets for each class in SudSenti3.

sad or disturbing vocabulary, or if there is a negative emoji, or if there is more than one feeling and the negative emotion is dominant. Finally, a text is considered neutral if it not clearly positive or negative.

6. Judges worked independently. If at least two of the three judges classified a post as negative, it was judged negative sentiment and similarly for positive sentiment. Neutral posts were deleted from the collection, resulting in a two-class data set.
7. By following the above procedure, 4000 posts were selected from the original 4544. The final SudSenti2 data set contains 2027 positive posts and 1973 negative posts.

3.2. SudSenti3 data set

The following steps were carried out to produce the three-class data set:

1. All Twitter messages matching one of the four search strings listed above were downloaded using the Twitter API. This resulted in 8021 posts.
2. The same three judges classified the posts as for SudSenti2.
3. Posts not considered Sudanese by at least two of the three judges were deleted.
4. Each tweet was classified as positive, negative or neutral. SudSenti3 is thus a three-class data set.
5. Judges worked independently. Posts classified positive by at least two judges were considered positive and the same for negative and neutral. Posts where there was no majority judgement were eliminated.
6. By following the above procedure, 7109 tweets were selected from the original 8021. The resulting SudSenti3 data set contains 2523 positive posts and 2639 negative posts.

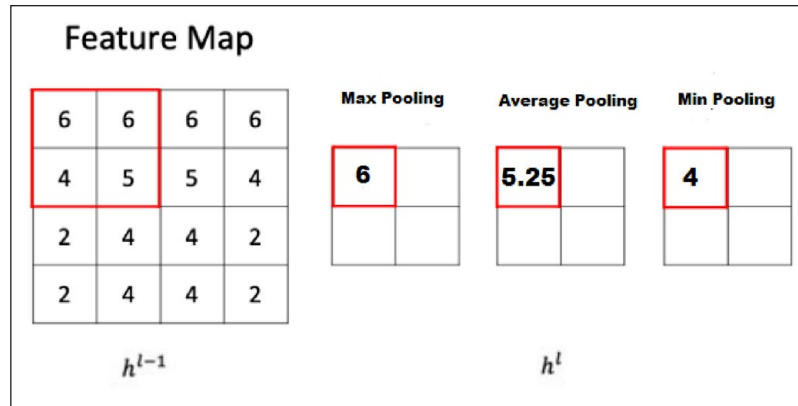


Figure 6. Standard pooling layers.

and $|P_k|$ denotes the number of activations. By collecting the outputs of all the pooling regions, the pooling feature map $E = e_1, \dots, e_k$ is obtained. We will start with a quick overview of standard pooling strategies (Figure 6).

Max pooling [47]: this takes the biggest activation in the pooling region

$$Max_K = \max_{i \in P_k} c_i, \text{ for } k = 1, \dots, K \tag{1}$$

Max pooling is ideal for extracting local characteristics from a feature map, such as edges, lines and textures.

Average pooling [48]: this calculates the mean value of activities in the pooling region

$$Avg_K = \frac{1}{|P_k|} \sum_{i \in P_k} c_i, \text{ for } k = 1, \dots, K \tag{2}$$

By smoothing the pooling region in this way, it is possible to extract global characteristics.

Min pooling [49]: this calculates the minimum value of activities in the pooling region

$$Min_K = \min_{i \in P_k} c_i, \text{ for } k = 1, \dots, K \tag{3}$$

Our proposed MMA pooling (Figures 7 and 8) calculates the mean of the max value and the average value

$$MMA_K = \frac{(\max_{i \in P_k} c_i) \left(\frac{1}{|P_k|} \sum_{i \in P_k} c_i \right)}{2} \tag{4}$$

for $k = 1, \dots, K$

MMA aims to combine the advantages of max pooling and average pooling.

4.4. Proposed approach

SCM (Figure 9) consists of an embedding layer containing max-features = num-unique-word, embedding size [128, 300] with max-len [150, 80, 50]. After that, there are four CNN layers with filters, respectively, of [512, 256, 128, 64]. Kernel-size=3, padding='valid', activation='ReLU' and strides=1. These are followed by the proposed MMA pooling function, one-dimensional (1D) pool size=2, then dense=32 and activation='ReLU', then dropout 0.5, then batch normalisation

and another dropout 0.5, then flatten and finally a softmax layer. This is fully connected to predict the sentiment between three classes (positive, negative, neutral) or two classes (positive, negative).

5. Experiments

Our experiments include four aspects (Figure 1):

1. Preprocessing the data sets and checking the steps.
2. Utilising existing ML and deep learning methods to verify performance.
3. Applying the proposed method.
4. Analysing results.

5.1. Data sets

For sentiment classification on Sudanese Arabic text, ML and deep learning models are trained using the new SudSenti2 and SudSenti3 data sets introduced in section 3. SudSenti2 consists of two classes – 2027 positive tweets and 1973 negative tweets. SudSenti3 consists of three classes – 2523 positive tweets, 2639 negative tweets and 1947 neutral tweets.

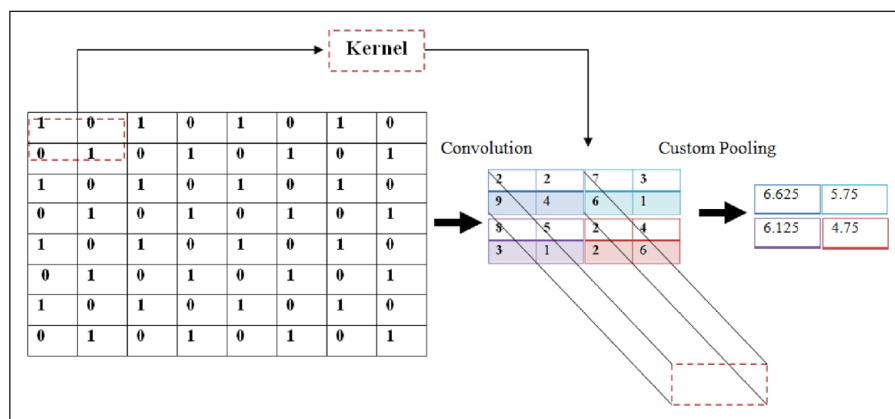


Figure 7. Mean max average (MMA).

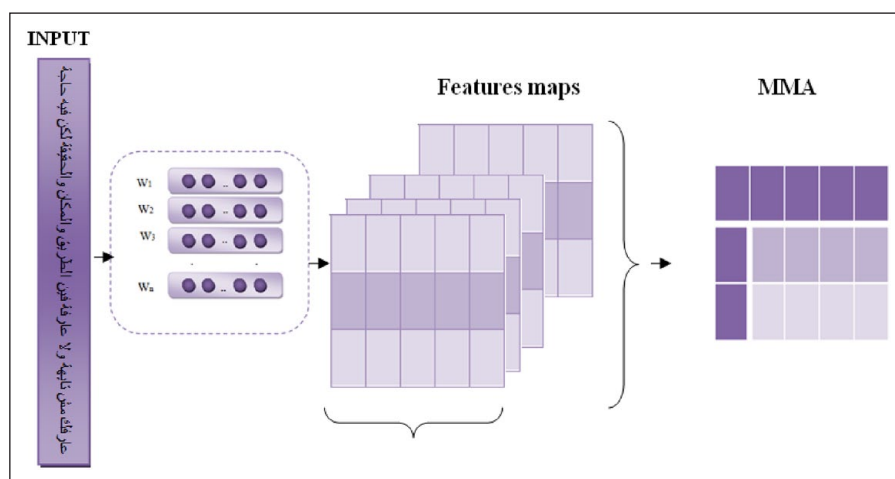


Figure 8. Proposed MMA on Arabic text.

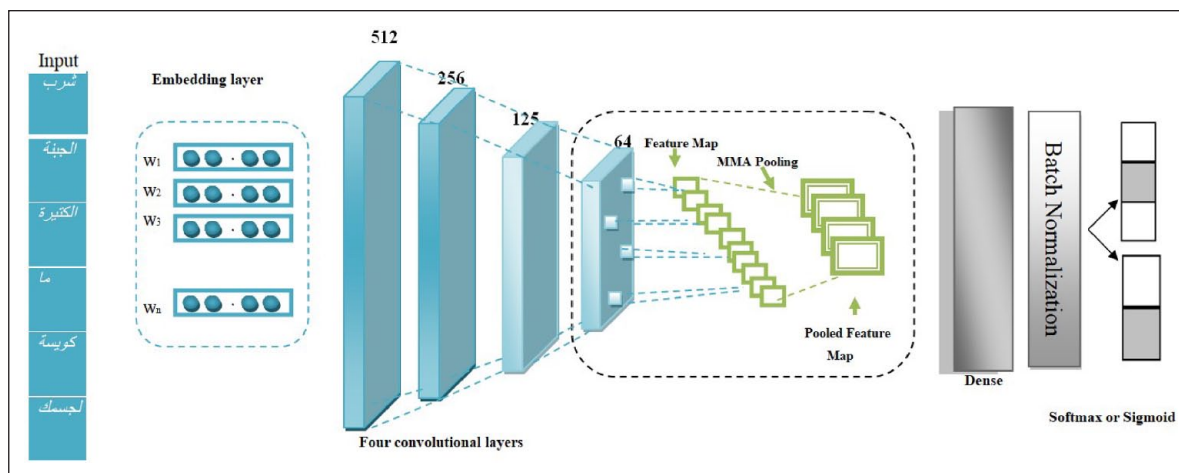


Figure 9. SCM model architecture.

For the Saudi dialect, we use the Saudi Sentiment Data set (SSD) [50].¹³ SSD consists of two classes – 2436 positive tweets and 1816 negative tweets.

For sentiment classification in MSA, the models are trained using the HARD¹⁴ produced by Elnagar et al. [42]. It is a rich data set, with more than 370,000 reviews expressed in MSA. Here, we utilised two classes – 5857 positive tweets and 6353 negative tweets.

Table 9 shows the details of the data sets.

5.2. Experimental settings

We used ML algorithms and deep learning models for training with all the Arabic sentiment data sets for two-way and three-way classifications. The ML algorithms were NB [51], LR [52], SVMs [53] and RF [54].

Deep learning models include RNN [55], CNN [56], CNN-LSTM [57]¹⁵ and the proposed method. For the SudSenti2 and SudSenti3 data sets, we split the data into 80% training, 10% validation and 10% testing.

For the SSD and HARD data sets, we applied the proposed approach and existing deep learning models. The settings for the experiments are shown in Table 10.

5.3. Experiment 1: two-way sentiment classification

The aim was to evaluate the proposed SCM + MMA model in two-way sentiment classification, working with the SudSenti2, SSD and HARD data sets. SudSenti2 was introduced in section 3. As baselines, there are four ML models (LR, RF, NB, SVM) and three NN models (RNN, CNN, CNN-LSTM). The configuration of SCM + MMA is shown in Table 10. Ten-fold cross-validation was used for all models and the average performance reported. The results are shown in Table 11.

On SudSenti2 (Sudanese dialect), the best model is SCM + MMA (accuracy 92.25%). The best ML baseline was RF (87.12%) and the best NN baseline was CNN-LSTM (89.00%).

On the SSD data set (Saudi dialect), the best model is SCM + MMA (84.02%) and the best baseline is CNN-LSTM (83.55%). Finally, on the HARD data set (MSA), the best model is again SCM + MMA (88.37%) as against the best baseline, CNN (87.06%).

In summary, the experiment showed that the proposed model performed well on two-class data sets.

5.4. Experiment 2: three-way sentiment classification

The aim was to evaluate SCM + MMA once again, this time on the new three-way Sudanese data set, SudSenti3 (section 3). Three-way classification is known to be a harder task than two-way, particularly as the neutral class can contain examples with both positive and negative aspects, a factor which may confuse the model. As baselines, there are three NN models (RNN, CNN, CNN-LSTM). The configuration of SCM + MMA was the same as in Experiment 1 (Table 10)

Table 8. Counts of words from two Sudanese wordlists which are present in the SudSenti2 and SudSenti3 data sets.

	SudSenti2	SudSenti3
Wiki*** count	52	55
Wiki total	82	82
Proportion	63.41%	67.07%
Mo3&& count	132	129
Mo3 total	248	248
Proportion	53.22%	52.01%

Table 9. Data sets for our experiments.

Data sets	Positive tweets	Negative tweets	Neutral tweets	Total
SudSenti2 (2C)	2027	1973	–	4000
SudSenti3 (3C)	2523	2639	1947	7109
SSD (2C)	2436	1816	–	4252
HARD (2C)	5857	6353	–	12,210

SudSenti2: two-class Sudanese Sentiment Data set; SudSenti3: three-class Sudanese Sentiment Data set; SSD: Saudi Sentiment Data set; HARD: Hotel Arabic Reviews Data set.

Table 10. Experimental settings.

Setting	Value(s)
Embedding size	{300}
Pooling	{2}
Batch-size	{32}
Kernel-size	{3}
Number-classes	{2, 3}
Epoch	{5, 10, 20, 50, 100, 200}
Optimiser	Adam
Learning rate	{0.001}

Table 11. Experiment I: accuracy of ML and NN sentiment classifiers on two-class data sets.

Models	Accuracy (%)		
	SudSenti2 data set (2C)	SSD data set (2C)	HARD data set (2C)
LR	86.04	–	–
RF	87.12	–	–
NB	81.45	–	–
SVM	86.23	–	–
RNN	80.75	74.03	62.86
CNN	87.75	82.49	87.06
CNN-LSTM	89.00	83.55	85.22
SCM + MMA	92.25	84.02	88.37

LR: logistic regression; RF: random forest; NB: Naïve Bayes; SVM: support vector machine; RNN: recurrent neural network; CNN: convolutional neural network; LSTM: long-short term memory.

SudSenti2 is our new two-class data set for Sudanese, created from Facebook and YouTube (section 3). SCM + MMA is the proposed model.

except that there were three outputs, not two. Once again, 10-fold cross-validation was used for all models. The results are shown in Table 12.

The best-performing model is SCM + MMA (85.23%). The best ML baseline is LR (79.37%) and the best NN baseline is CNN (83.61%).

Table 12. Experiment 2: accuracy of NN sentiment classifiers on the SudSenti3 three-class data set, created from Sudanese Twitter posts (section 3).

Models	Accuracy (%)
	SudSenti3 data set (3C)
LR	79.37
RF	78.24
NB	74.19
SVM	79.29
RNN	77.07
CNN	83.61
CNN-LSTM	81.01
SCM + MMA	85.23

LR: logistic regression; RF: random forest; NB: Naïve Bayes; SVM: support vector machine; RNN: recurrent neural network; CNN: convolutional neural network; LSTM: long-short term memory.
SCM + MMA is the proposed model.

Table 13. Experiment 3: accuracy of the SCM model with different pooling layers.

Models	Accuracy (%)		
	SudSenti2 data set (2C)	SSD data set (2C)	HARD data set (2C)
SCM + Max	90.62	84.72	89.27
SCM + Avg	91.75	84.02	87.80
SCM + Min	90.00	83.78	88.70
SCM + MMA	92.75	85.55	90.01

SCM: sentiment convolutional MMA.

The task is two-class sentiment classification, applied to the SudSenti2, SSD and HARD data sets. MMA is the proposed pooling layer.

5.5. Experiment 3: evaluation of MMA pooling

A key part of the proposed SCM + MMA model is the MMA pooling layer. The aim of this experiment, therefore, was to compare MMA with three commonly used pooling layers – Max, Avg and Min. First, in two-way classification, the performance of SCM + Max, SCM + Avg and SCM + Min was compared with SCM + MMA on SudSenti2, SSD and HARD (compare with Experiment 1). Fifteenfold cross-validation was used and the average performance reported. Results are shown in Table 13. SCM + MMA is the best-performing model on all three data sets (SudSenti2 92.75%, SSD 85.55%, HARD 90.01%). The best baseline pooling layer varies between Avg and Max by data set (SudSenti2: Avg 91.75%; SSD: Max 84.72%; HARD: Max 89.27).

Second, in three-way classification, the performance of SCM + Max, SCM + Avg and SCM + Min was compared with SCM + MMA on SudSenti3 (compare with Experiment 2). Fifteenfold cross-validation was again used. Results are shown in Table 14. Once again, SCM + MMA is the best-performing model (84.39%) with the best baseline being Max (84.11%).

In conclusion, the MMA pooling layer performs well compared with Max, Avg and Min.

5.6. Experiment 4: ablation study

We started with the proposed SCM + MMA model whose performance was reported in Table 11 (two-class) and Table 12 (three-class). First, the normalisation steps were removed and the training repeated. Second, the Sudanese stopword list was removed. In each case, 10-fold cross-validation was used, and the average performance was reported. The results are shown in Table 15 and Figure 10.

Normalisation aims to clean noise and spaces and to transform every letter into its standard form without affecting its meaning or content [43]. When we removed the normalisation steps, it affected the two-class and three-class data sets. For SudSenti2, the accuracy reduces from 92.25% to 91.82%, for SSD, it reduces from 84.02% to 83.73%, for HARD from 88.37% to 87.96% and for SudSenti3 from 85.23% to 84.78%. The differences are -0.43% , -0.29% , -0.41% and -0.45% , respectively. The results suggest that text normalisation can result in a small performance improvement.

Table 14. Experiment 3: accuracy of the SCM model with different pooling layers.

Models	Accuracy (%)
	SudSenti3 Data set (3C)
SCM + Max	84.11
SCM + Avg	83.26
SCM + Min	82.70
SCM + MMA	84.39

SCM: sentiment convolutional MMA.

The task is three-class sentiment classification, applied to the SudSenti3 data set. MMA is the proposed pooling layer.

Table 15. Experiment 4: ablation study to show the effects of text normalisation and the stopword list.

Models	Accuracy (%)			
	SudSenti2	SSD	HARD	SudSenti3
	Data set (2C)	Data set (2C)	Data set (2C)	Data set (3C)
SCM + MMA	92.25	84.02	88.37	85.23
SCM + MMA	91.82	Without normalisation 83.73	87.96	84.78
SCM + MMA	91.00	Without normalisation, stopword list 83.22	87.40	84.06

When we removed the Sudanese stopword list, it also affected the four data sets. For SudSenti2, the accuracy reduces from 91.82% to 91.00%, for SSD, it reduces from 83.73% to 83.22%, for HARD from 87.96% to 87.40% and for SudSenti3 from 84.78% to 84.06%. The differences are -0.82% , -0.51% , -0.56% and -0.72% , respectively. We also note that the differences with the Sudanese data sets were higher than for SSD and HARD. The Sudanese stopword list assists with noise reduction, which indicates that it is an important step for preprocessing.

5.7. Experiment 5: Arabic transformer models evaluated on Sudanese data sets

We trained ARBERT¹⁶ and MARBERT¹⁷ transformers on the SudSenti2 and SudSenti3 data sets. Tenfold cross-validation was used with each transformer, and the average performance was reported. Results are in Table 16 and Figure 11.

For SudSenti2, accuracies were 90.12% and 91.11%, respectively. For SudSenti3, they were 85.09% and 86.83%. MARBERT had the best performance among the transformers, so we fine-tuned it with our hyperparameters, enhancing the performance on SudSenti2 to 92.14% and on SudSenti3 to 88.44%. Recall that the performance of the proposed model SCM + MMA was 92.25% on SudSenti2 and 85.23% on SudSenti3 (Tables 10 and 11).

First, we conclude that transformers can be successfully applied to our data sets, returning excellent results. Second, MARBERT gave the best performance compared with previous ML and deep learning baselines for the Sudanese data sets in Tables 11 and 12. For SudSenti2, MARBERT + FT accuracy was 92.14% (Table 16) compared with 89.00% with CNN-LSTM (Table 11). The proposed model SCM + MMA showed comparable performance (92.25%), a good result for a CNN-based model. For SudSenti3, MARBERT + FT accuracy was 88.44% (Table 16), better than 83.61% with CNN (Table 12). SCM + MMA scored 85.23%, falling short of MARBERT + FT.

Third, MARBERT + FT was shown to make a difference of up to 1.03% (two-class) and 1.61% (three-class), which indicates the effectiveness of FT. Fourth, transformers are often better than earlier approaches to Arabic sentiment analysis. However, there are some exceptions. For example, Abdel-Salam [20] obtained a better performance than MARBERT using an MHLCG and the ArSarcasm-v2 data set. Moreover, when Abdel-Salam applied MARBERT to ASTD-3, the result was 78% which is slightly lower than Mhamed et al. [22] who obtained 78.62%.

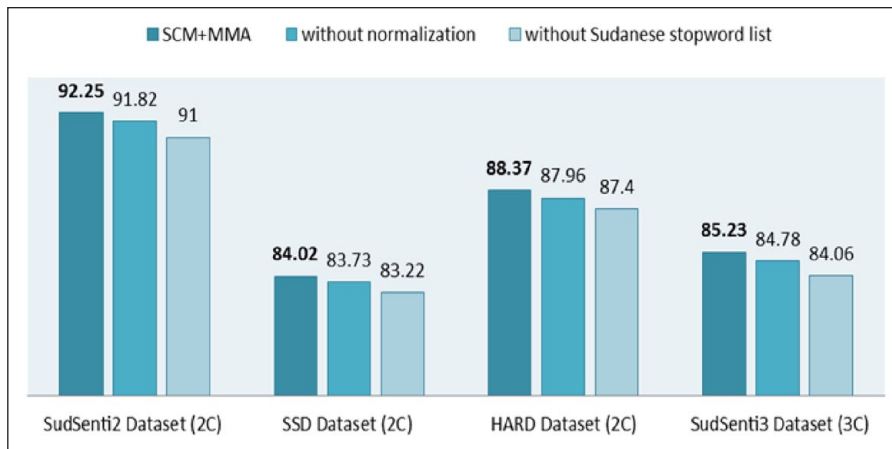


Figure 10. Ablation study.

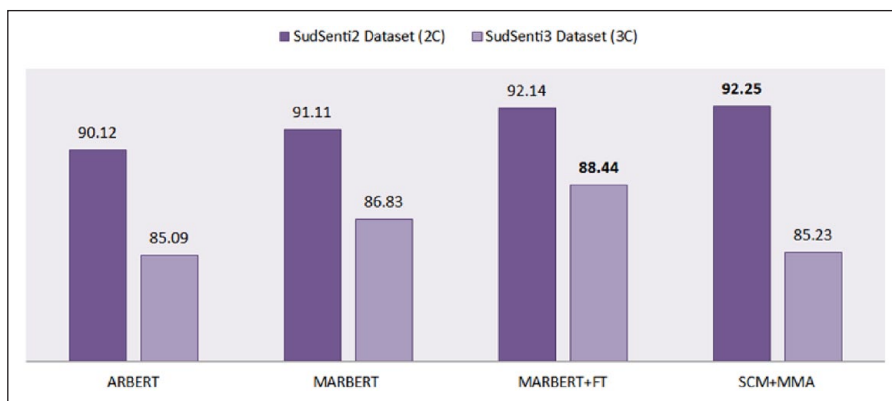


Figure 11. Accuracy of SCM + MMA and Arabic transformers on the Sudanese data sets.

Table 16. Experiment 5: accuracy of the transformer models with SudSenti2 and SudSenti3.

Models	Accuracy (%)	
	SudSenti2 data set (2C)	SudSenti3 data set (3C)
ARBERT	90.12	85.09
MARBERT	91.11	86.83
MARBERT + FT	92.14	88.44

Finally, while MSA has a substantial overlap with dialects, MSA-based transformer models do not necessarily capture dialectal nuances [58]. As a result, the proposed SCM + MMA model, which is based on CNNs, still compares well with MARBERT when applied to the SudSenti2 data set (Experiment 1 vs Experiment 5).

5.8. Accuracy during training

Figure 12 shows the accuracy and validation accuracy of the NN baseline models and the proposed method with the SudSenti2 data set. After 50 epochs, the SCM + MMA model shows the highest performance, reaching 92.25%. Figure 13 shows the same information for the SSD data set (SCM + MMA reaches 84.02%) while Figure 14 is for the HARD data set (SCM + MMA reaches 88.37%).

Figure 15 shows the accuracy and validation accuracy for the NN models and the proposed model with the SudSenti3 data set. After 50 epochs, SCM + MMA reaches 85.23%.

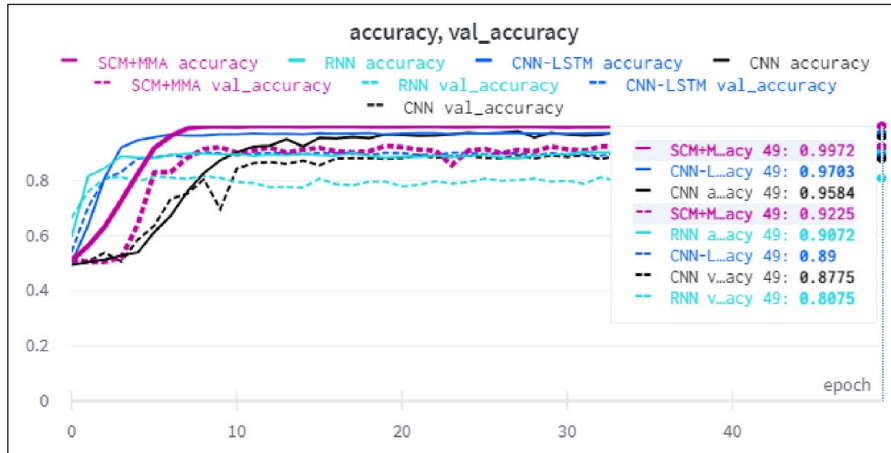


Figure 12. Accuracy and validation accuracy with the SudSenti2 data set.

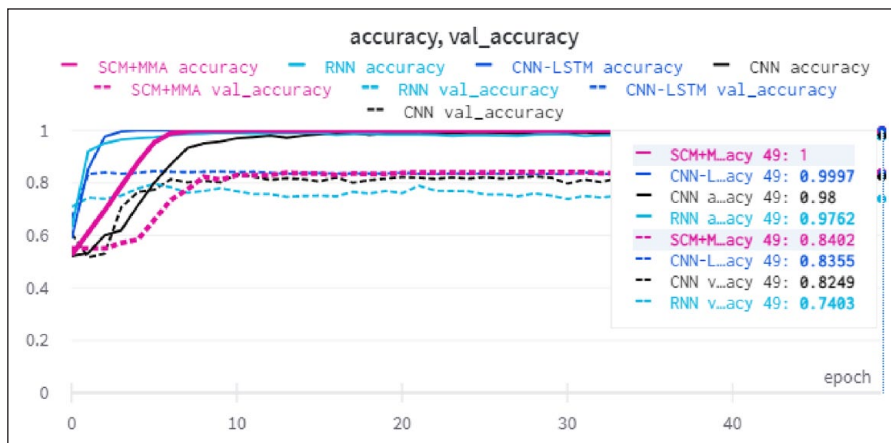


Figure 13. Accuracy and validation accuracy with the SSD data set.

For two-way classification, Figures 16–18 show the validation accuracy during training for the SudSenti2, SSD and HARD data sets. For three-way classification, Figure 19 shows the validation accuracy during training for SudSenti3. We note that the proposed method was stable over epochs for training and validation with different data sets.

6. Conclusion and future work

In this article, we first presented two new sentiment data sets for the Sudanese dialect of Arabic. SudSenti2 was collected from Facebook and YouTube, while SudSenti3 was based on Twitter tweets. Following a discussion of Arabic preprocessing methods appropriate to sentiment classification, we proposed a new model for this task, SCM. This includes four convolutional layers plus MMA, our proposed pooling layer. In two-way sentiment classification using the SudSenti2 (Sudanese), SSD (Saudi) and HARD (MSA) data sets, SCM gave good performance relative to ML and NN baselines and was comparable with Arabic transformer models. In three-way classification using SudSenti3, MARBERT + FT was the highest performing and was superior to SCM.

Concerning pooling, the proposed MMA approach was compared with Max, Avg and Min baselines and shown to perform better than them in both two-way and three-way classifications. Finally, we conducted an ablation study, which demonstrated that text normalisation and the Sudanese stopword list make small contributions to performance.

In future work, we plan to use an attention mechanism as part of a more complex deep learning method, to extract features from a huge corpus covering all Arabic sentiment dialects. We also aim to customise a new regulariser to enhance performance and optimise the loss function.

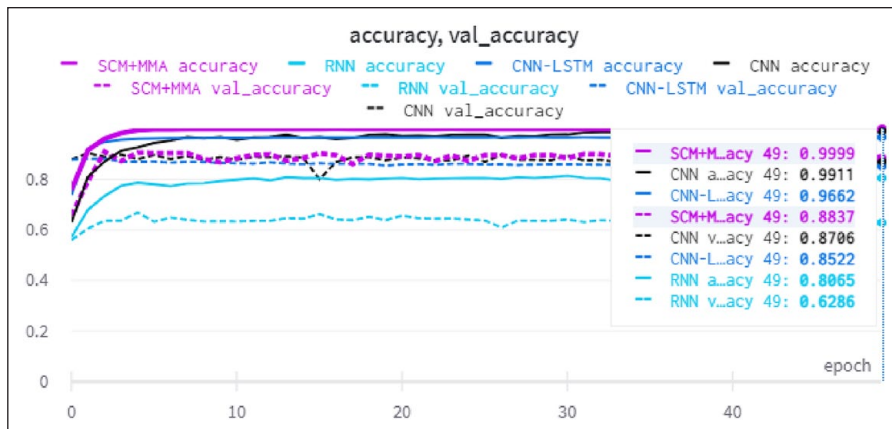


Figure 14. Accuracy and validation accuracy with the HARD data set.

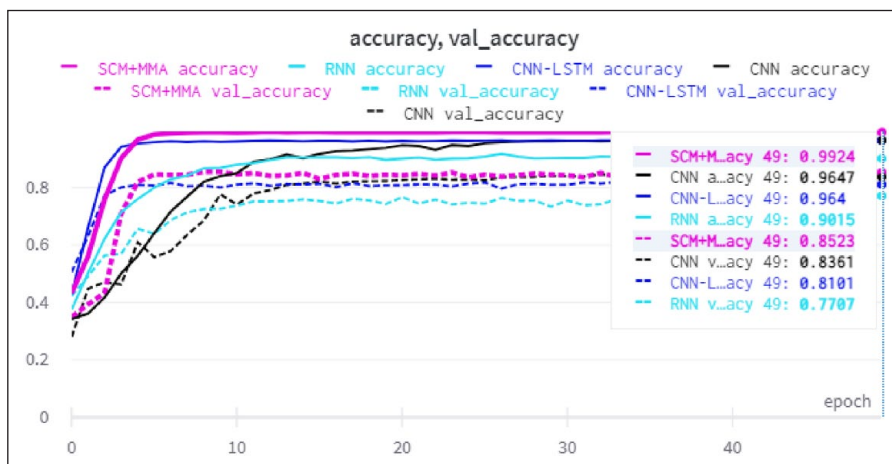


Figure 15. Accuracy and validation accuracy with the SudSenti3 data set.

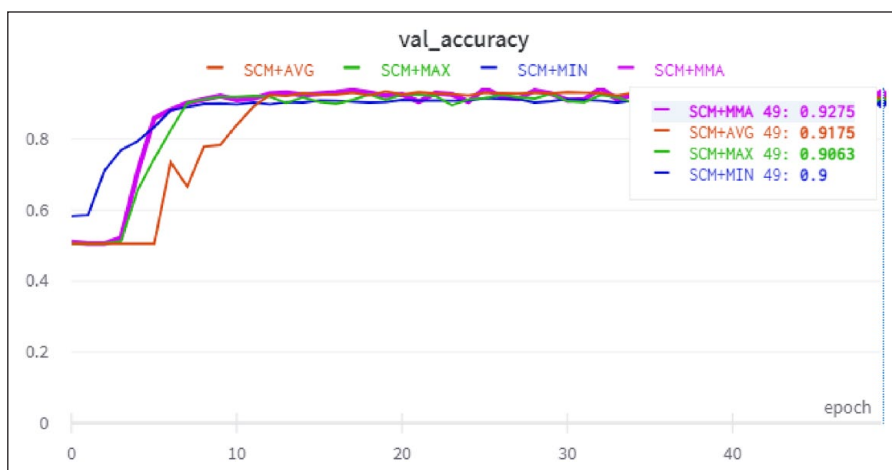


Figure 16. Validation accuracy with the SudSenti2 data set.

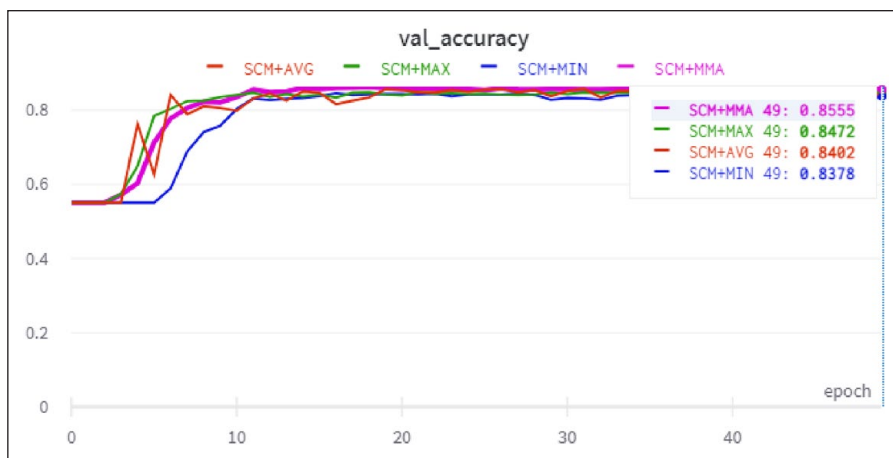


Figure 17. Validation accuracy with the SSD data set.

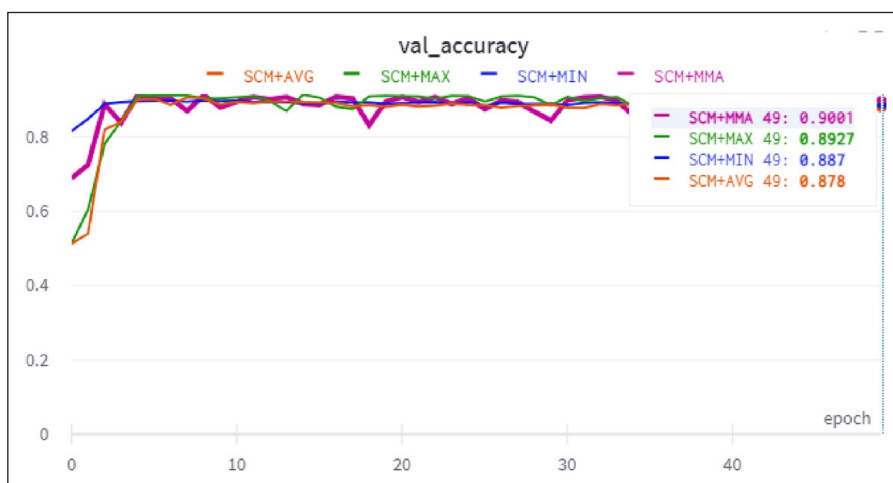


Figure 18. Validation accuracy with the HARD data set.

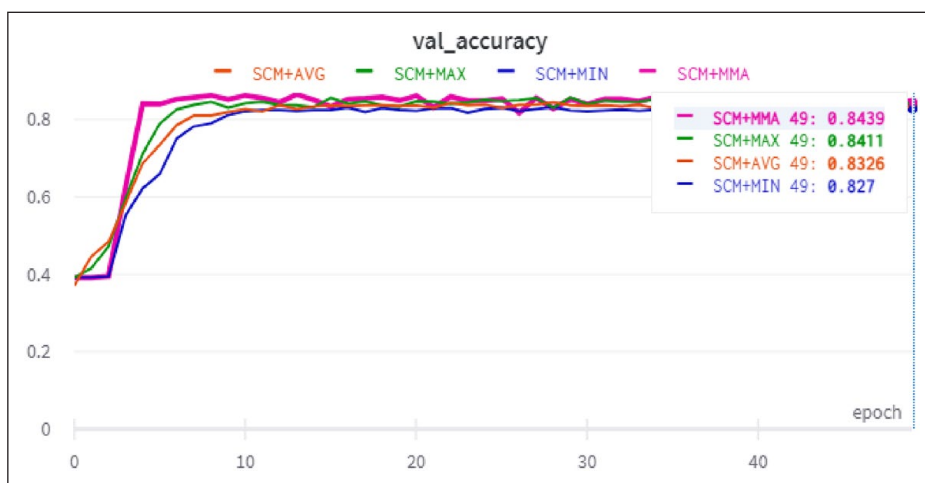


Figure 19. Validation accuracy with the SudSenti3 data set.

Author's note

Mustafa Mhamed is also affiliated to College of Information and Electrical Engineering, China Agricultural University, China.

Declaration of conflicting interests


The author(s) declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship and/or publication of this article: This research was supported by the National Natural Science Foundation of China (grant no. 61877050) and Open Project Fund of Shaanxi Province Key Lab of Satellite and Terrestrial Network Technology, Shaanxi Province Financed Projects for Scientific and Technological Activities of Overseas Students (grant no. 202160002). H.Q. acknowledges the support of the Business and Local Government Data Research Centre BLG DRC (ES/S007156/1) funded by the Economic and Social Research Council (ESRC).

ORCID iDs

Mustafa Mhamed  <https://orcid.org/0000-0002-3106-669X>

Richard Sutcliffe  <https://orcid.org/0000-0002-5549-5691>

Eiad Almekhlafi  <https://orcid.org/0000-0002-7182-9639>

Data availability

The SudSenti2 and SudSenti3 data sets are publicly available.¹⁸

Notes

1. <https://www.importanceoflanguages.com/arabic-dialects/>
2. <https://istizada.com/complete-list-of-arabic-speaking-countries-2014/>
3. https://en.wikipedia.org/wiki/Arabic_diacritics
4. We quote this example exactly from the original, although the case marker may not be correct.
5. <https://github.com/mustafa20999/Sudanese-Arabic-Sentiment-Datasets>
6. <https://www.kaggle.com/datasets/mksaad/arabic-sentiment-twitter-corpus>
7. <https://www.facebook.com/SudanLovers/>
8. <https://www.youtube.com/watch?v=h5tBHZZ4UCYt=1s>
9. <https://orangedatamining.com/>
10. <https://datareportal.com/reports/digital-2021-sudan>
11. *** https://ar.wikipedia.org/wiki/%D9%84%D9%87%D8%AC%D8%A9_%D8%B3%D9%88%D8%AF%D8%A7%D9%86%D9%8A%D8%A9
12. && <https://en.mo3jam.com/dialect/Sudanese>
13. <https://www.kaggle.com/snalyami3/arabic-sentiment-analysis-dataset-ss2030-dataset>
14. <https://github.com/elngara/HARD-Arabic-Dataset>
15. <https://www.kaggle.com/monsterspy/conv-lstm-sentiment-analysis-keras-acc-0-96>
16. <https://github.com/UBC-NLP/marbert>
17. <https://github.com/Jabalov/Arabic-Dialects-Identification/blob/main/Notebooks/MARBERT-FineTuning.ipynb>
18. <https://github.com/mustafa20999/Sudanese-Arabic-Sentiment-Datasets>

References

- [1] Liu B. Many facets of sentiment analysis. In: Cambria E, Das D and Bandyopadhyay S, et al. (eds) *A practical guide to sentiment analysis*. Cham: Springer, 2017, pp. 11–39.
- [2] Sadat F, Kazemi F and Farzindar A. Automatic identification of Arabic dialects in social media. In: *Proceedings of the first international workshop on Social media retrieval and analysis*, Gold Coast, QLD, Australia, 11 July 2014, pp. 35–40. New York: ACM.
- [3] Almekhlafi E, Moeen A-M, Zhang E, et al. A classification benchmark for Arabic alphabet phonemes with diacritics in deep neural networks. *Comput Speech Lang* 2022; 71: 101274.
- [4] Hadj Ameer MS, Moulahoum Y and Guessoum A. Restoration of Arabic diacritics using a multilevel statistical model. In: Amine A, Bellatreche L and Elberrihi Z, et al. (eds) *IFIP international conference on computer science and its applications*. Cham: Springer, 2015, pp. 181–192.
- [5] Samih Y. *Dialectal Arabic processing using deep learning*. PhD thesis, HeinrichHeine-Universität, Düsseldorf, 2017.
- [6] Zahir J. Iadd: an integrated Arabic dialect identification dataset. *Data Brief* 2022; 40: 107777.

- [7] Al-Shawakfa E, Al-Badarneh A, Shatnawi S, et al. A comparison study of some Arabic root finding algorithms. *J Am Soc Inf Sci Tec* 2010; 61(5): 1015–1024.
- [8] Hussien O, Dashtipour K and Hussain A. Comparison of sentiment analysis approaches using modern Arabic and Sudanese dialect. In: Ren J, Hussain A and Zheng J, et al. (eds) *International conference on brain inspired cognitive systems*. Cham: Springer, 2018, pp. 615–624.
- [9] Versteegh CHM. *Pidginization and creolization: the case of Arabic*, vol. 33. Amsterdam: John Benjamins Publishing, 1984.
- [10] Oueslati O, Cambria E, HajHmida MB, et al. A review of sentiment analysis research in Arabic language. *Future Gener Comp Sy* 2020; 112: 408–430.
- [11] Alharbi AI, Taileb M and Kalkatawi M. Deep learning in Arabic sentiment analysis: an overview. *J Inform Sci* 2021; 47(1): 129–140.
- [12] Alwehaibi A and Roy K. Comparison of pre-trained word vectors for Arabic text classification using deep learning approach. In: *2018 17th IEEE international conference on machine learning and applications (ICMLA)*, Orlando, FL, 17–20 December 2018, pp. 1471–1474. New York: IEEE.
- [13] Alahmary R and Al-Dossari H. A semiautomatic annotation approach for sentiment analysis. *J Inform Sci* 2023; 41: 398–410.
- [14] Tabii Y, Lazaar M, Al Achhab M, et al. Big data, cloud and applications. In: *Third international conference, BDCA 2018*, Kenitra, Morocco, 4–5 April 2018, Revised Selected Papers, vol. 872. Cham: Springer.
- [15] Lulu L and Elnagar A. Automatic Arabic dialect classification using deep learning models. *Procedia Comput Sci* 2018; 142: 262–269.
- [16] Al-Saqqa S, Obeid N and Awajan A. Sentiment analysis for Arabic text using ensemble learning. In: *2018 IEEE/ACS 15th international conference on computer systems and applications (AICCSA)*, Aqaba, Jordan, 28 October–1 November 2018, pp. 1–7. New York: IEEE.
- [17] Al Omari M, Al-Hajj M, Hammami N, et al. Sentiment classifier: logistic regression for Arabic services’ reviews in Lebanon. In: *2019 international conference on computer and information sciences (ICCIS)*, Sakaka, Saudi Arabia, 3–4 April 2019, pp. 1–5. New York: IEEE.
- [18] Abdelli A, Guerrouf F, Tibermacine O, et al. Sentiment analysis of Arabic Algerian dialect using a supervised method. In: *2019 international conference on intelligent systems and advanced computing sciences (ISACS)*, Taza, Morocco, 26–27 December 2019, pp. 1–6. New York: IEEE.
- [19] Mulki H, Haddad H, Gridach M, et al. Empirical evaluation of leveraging named entities for Arabic sentiment analysis. arXiv preprint arXiv:1904.10195, 2019.
- [20] Abdel-Salam R. WANLP 2021 shared-task: towards irony and sentiment detection in Arabic tweets using multiheaded-LSTM-CNN-GRU and MaBERT. In: *Proceedings of the Sixth Arabic natural language processing workshop*, pp. 306–311, 2021. <https://aclanthology.org/2021.wanlp-1.37/>
- [21] Abdul-Mageed M, Elmadany A and Nagoudi EMB. ARBERT & MARBERT: deep bidirectional transformers for Arabic. arXiv preprint arXiv:2101.01785, 2020.
- [22] Mhamed M, Sutcliffe R, Sun X, et al. Improving Arabic sentiment analysis using CNN-based architectures and text preprocessing. *Comput Intel Neurosc* 2021; 2021: 5538791.
- [23] Al-shaibani MS, Alyafei Z and Ahmad I. Meter classification of Arabic poems using deep bidirectional recurrent neural networks. *Pattern Recogn Lett* 2020; 136: 1–7.
- [24] Addi HA and Ezzahir R. Sampling techniques for Arabic sentiment classification: a comparative study. In: *Proceedings of the 3rd international conference on networking, information systems & security*, Marrakech, Morocco, 31 March–2 April 2020, pp. 1–6. New York: ACM.
- [25] Al-Moslmi T, Albared M, Al-Shabi A, et al. Arabic senti-lexicon: constructing publicly available language resources for Arabic sentiment analysis. *J Inform Sci* 2018; 44(3): 345–362.
- [26] Hochreiter S and Schmidhuber J. Long short-term memory. *Neural Comput* 1997; 9(8): 1735–1780.
- [27] Gers FA, Schmidhuber J and Cummins F. Continual prediction using LSTM with forget gates. In: Marinaro M and Tagliaferri R (eds) *Neural Nets WIRN Vietri-99*. Cham: Springer, 1999, pp. 133–138.
- [28] Zaidan O and Callison-Burch C. 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*, Portland, OR, 19–24 June 2011, pp. 37–41. New York: ACM.
- [29] Rushdi-Saleh M, Martín-Valdivia MT, Ureña-López LA, et al. OCA: opinion corpus for Arabic. *J Am Soc Inf Sci Tec* 2011; 62(10): 2045–2054.
- [30] Abdulla NA, Ahmed NA, Shehab MA, et al. Arabic sentiment analysis: lexicon-based and corpus-based. In: *2013 IEEE Jordan conference on applied electrical engineering and computing technologies (AEECT)*, Amman, Jordan, 3–5 December 2013, pp. 1–6. New York: IEEE.
- [31] Aly M and Atiya A. LABR: a large scale Arabic book reviews dataset. In: *Proceedings of the 51st annual meeting of the Association for Computational Linguistics*, vol. 2: Short Papers, 2013, pp. 494–498. <https://aclanthology.org/P13-2088/>
- [32] Moudjari L, Akli-Astouati K and Benamara F. An Algerian corpus and an annotation platform for opinion and emotion analysis. In: *Proceedings of The 12th language resources and evaluation conference*, pp. 1202–1210, 2020. <https://aclanthology.org/2020.lrec-1.151/>

- [33] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013.
- [34] Jerbi MA, Achour H and Souissi E. Sentiment analysis of code-switched Tunisian dialect: exploring RNN based techniques. In: Smaïli K (ed.) *International conference on Arabic language processing*. Cham: Springer, 2019, pp. 122–131.
- [35] Yilmaz T, Ergil A and İlgen B. Deep learning-based document modeling for personality detection from Turkish texts. In: Arai K, Bhatia R and Kapoor S (eds) *Proceedings of the future technologies conference*. Cham: Springer, 2019, pp. 729–736.
- [36] Sayadi K, Liwicki M, Ingold R, et al. Tunisian dialect and modern standard Arabic dataset for sentiment analysis: Tunisian election context. In: *Second international conference on Arabic computational linguistics, ACLING*, Konya, Turkey, 3rd-9th April 2016, pp. 35–53.
- [37] Medhaffar S, Bougares F, Esteve Y, et al. Sentiment analysis of Tunisian dialects: linguistic resources and experiments. In: *Third Arabic natural language processing workshop (WANLP)*, 2017, pp. 55–61. <https://aclanthology.org/W17-1307/>
- [38] Farha A, Zaghouni W and Magdy W. Overview of the WANLP 2021 shared task on sarcasm and sentiment detection in Arabic. In: *Proceedings of the sixth Arabic natural language processing workshop*, 2021, pp. 296–305. <https://aclanthology.org/2021.wanlp-1.36/>
- [39] Nabil M, Aly M and Atiya A. ASTD: Arabic sentiment tweets dataset. In: *Proceedings of the 2015 conference on empirical methods in natural language processing*, 2015, pp. 2515–2519. <https://aclanthology.org/D15-1299/>
- [40] Salameh M, Mohammad S and Kiritchenko S. Sentiment after translation: a case-study on Arabic social media posts. In: *Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: human language technologies*, 2015, pp. 767–777. <https://aclanthology.org/N15-1078/>
- [41] Alharbi R and Aljaedi A. Predicting rogue content and Arabic spammers on Twitter. *Future Internet* 2019; 11(11): 229.
- [42] Elnagar A, Khalifa YS and Einea A. Hotel Arabic-reviews dataset construction for sentiment analysis applications. In: Shaalan K, Hassanien A and Tolba F (eds) *Intelligent natural language processing: trends and applications*. Cham: Springer, 2018, pp. 35–52.
- [43] Hegazi MO, Al-Dossari Y, Al-Yahy A, et al. Preprocessing Arabic text on social media. *Heliyon* 2021; 7(2): e06191.
- [44] Sallam RM, Mousa HM and Hussein M. Improving Arabic text categorization using normalization and stemming techniques. *Int J Comput Appl* 2016; 135(2): 38–43.
- [45] Soliman AB, Eissa K and El-Beltagy SR. AraVec: a set of Arabic word embedding models for use in Arabic NLP. *Procedia Comput Sci* 2017; 117: 256–265.
- [46] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality. arXiv preprint arXiv:1310.4546, 2013.
- [47] Ranzato M, Breure Y-L, LeCun Y, et al. Sparse feature learning for deep belief networks. *Adv Neur In* 2007; 20: 1185–1192.
- [48] LeCun Y, Boser B, Denker J, et al. Handwritten digit recognition with a back-propagation network. *Adv Neur In* 1990; 2: 396–404.
- [49] Socher R, Huang E, Pennin J, et al. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. *Adv Neur In* 2011; 24. https://papers.nips.cc/paper_files/paper/2011/hash/3335881e06d4d23091389226225e17c7-Abstract.html
- [50] Alyami SN and Olatunji SO. Application of support vector machine for Arabic sentiment classification using twitter-based dataset. *J Inform Knowl Manag* 2020; 19(1): 2040018.
- [51] Mohammad AH. Arabic text classification: a review. *Modern Appl Sci* 2019; 13(5): 1–88.
- [52] Al-Turaiki I, Alshahrani M and Almutairi T. Building predictive models for MERS-COV infections using data mining techniques. *J Infect Public Heal* 2016; 9(6): 744–748.
- [53] Mohammad H, Alwada'n T and Almomani O. Arabic text categorization using support vector machine, Nave Bayes and neural network. *GSTF J Comput* 2016; 5: 108–115.
- [54] Ravi K and Ravi V. A novel automatic satire and irony detection using ensembled feature selection and data mining. *Knowl-Based Syst* 2017; 120: 15–33.
- [55] Britz D. Recurrent neural networks tutorial, part 1 –introduction to RNNs. [online]. <http://www.wildml.com/2015/09/recurrent-neuralnetworkstutorial-part-1-introduction-to-rnns/> (2015, accessed 2 May 2019).
- [56] Zhang Y and Wallace B. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. arXiv preprint arXiv:1510.03820, 2015.
- [57] Smaïli K. Arabic language processing: from theory to practice. In: *7th International Conference, ICALP 2019*, Nancy, 16–17 October 2019, vol. 1108. Cham: Springer Nature, 2019.
- [58] Abdelali A, Durrani N, Dalvi F, et al. Interpreting Arabic transformer models. arXiv preprint arXiv:2201.07434, 2022.