Engineering Science and Technology, an International Journal 43 (2023) 101433

HOSTED BY

ELSEVIER

Contents lists available at ScienceDirect

Engineering Science and Technology, an International Journal

journal homepage: www.elsevier.com/locate/jestch

Arabic sentiment analysis using GCL-based architectures and a customized regularization function

Mustafa Mhamed [a,c], Richard Sutcliffe [b,a,*], Xia Sun [a], Jun Feng [a,*], Ephrem Afele Retta [a]

[a] School of Information Science and Technology, Northwest University, Xi'an 710127, China
[b] School of Computer Science and Electronic Engineering, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK
[c] College of Information and Electrical Engineering, China Agricultural University, Beijing 100089, China

ARTICLE INFO

Article history:
Received 3 September 2022
Revised 18 April 2023
Accepted 22 April 2023
Available online 29 May 2023

Keywords:
Arabic sentiment analysis (ASA)
Natural language processing (NLP)
Custom regularization function (CRF)
Gated convolution long (GCL)

ABSTRACT

Sentiment analysis aims to extract emotions from textual data; with the proliferation of various social media platforms and the flow of data, particularly in the Arabic language, significant challenges have arisen, necessitating the development of various frameworks to handle issues. In this paper, we firstly design an architecture called Gated Convolution Long (GCL) to perform Arabic Sentiment Analysis. GCL can overcome difficulties with lengthy sequence training samples, extracting the optimal features that help improve Arabic sentiment analysis performance for binary and multiple classifications. The proposed method trains and tests in various Arabic datasets; The results are better than the baselines in all cases. GCL includes a Custom Regularization Function (CRF), which improves the performance and optimizes the validation loss. We carry out an ablation study and investigate the effect of removing CRF. CRF is shown to make a difference of up to 5.10% (2C) and 4.12% (3C). Furthermore, we study the relationship between Modern Standard Arabic and five Arabic dialects via a cross-dialect training study. Finally, we apply GCL through standard regularization (GCL+L_1, GCL+L_2, and GCL+$L_{ElasticNet}$) and our L_{new} on two big Arabic sentiment datasets; GCL+L_{new} gave the highest results (92.53%) with less performance time.

1. Introduction

Sentiment analysis is a form of natural language processing which detects the sentiment expressed in a text [1]. By 2025, it is estimated that sentiment analysis will be worth $3.8 billion [2], due to its many practical applications in business and politics. As a result, it has become a very active research field in recent years [3].

Initially, the majority of sentiment analysis research related to English text [4–9]. However, there has been a lot of interest in the Arabic language lately [10–14]. Furthermore, surveys have examined Arabic resources and strategies, in order to draw conclusions and identify difficulties associated with Arabic sentiment analysis [15–18]. This is not surprising, since Arabic is spoken by many people all over the world, is a significant language accepted by the United Nations [19], and is the fourth most popular language on the Internet [20].

The Arabic language comprises three classes, modern standard Arabic (MSA), dialect Arabic (DA), and classical Arabic (CA) [21]. MSA is used in official settings, including news reporting, educational institutions, and commercial forums. In contrast, Arabic dialects that vary from country to country are employed in casual writing, notably on social media. Classical Arabic is used in religious writings such as the Holy Qur'an and for prayer.

Deep Learning (DL) is an area of machine learning that deals with artificial neural networks, which are algorithms inspired by the structure and function of the brain [22]. Many DL techniques are now used in Arabic sentiment analysis systems. In particular, methods such as word embeddings, Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Long short-term memory (LSTM), and Hierarchical Attention Networks (HAN) have been used with great success [23–25]. However, despite data indicating that increasingly hard tasks necessitate more complex structures [17], especially in Arabic sentiment analysis, more sophisticated approaches to address classification difficulties are few. The following are important aspects of this study:

* Corresponding authors at: School of Computer Science and Electronic Engineering, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK (R. Sutcliffe), School of Information Science and Technology, Northwest University, Xi'an 710127, China (J. Feng).

E-mail addresses: rsutcl@essex.ac.uk, rsutcl@nwu.edu.cn (R. Sutcliffe), fengjun@nwu.edu.cn (J. Feng).

https://doi.org/10.1016/j.jestch.2023.101433
2215-0986/

1. We create an Arabic Sentiment Analysis (ASA) model which domain-independent. We test the proposed approach utilizing evaluations from various domains with a wide range of word relationships and assess the effectiveness of each dataset independently.
2. Utilizing the suggested strategy with thorough preprocessing techniques aids in the finest data cleaning, followed by the extraction of the best features that increase the effectiveness of the model.
3. Most earlier research lacked optimization of loss functions; our strategies, together with a customized loss function, concentrate on enhancing accuracy.
4. The proposed Custom Regularization Function (CRF) is more extensive than the standard, allowing us to optimize zone choices for our hyperparameter weights and so give the greatest features for future selection. Relative to current baselines, our technique offers the highest categorization performance and optimizes the performance through various loss functions.
5. A cross-dialect training study investigates the relationships between Modern Standard Arabic and five Arabic dialects, as follows:
   (a) An Arabic collection is chosen that only includes MSA vocabulary;
   (b) Dialect samples are selected that are readily accessible;
   (c) Baselines are generated via training and validation on MSA data;
   (d) Models are trained using the suggested approach on MSA data and validated through dialect datasets;
   (e) The results are analysed.

Regularization is a fundamental component of machine learning, especially deep learning, that allows for good generalization to unknown data even when trained on a small training set or with a poor optimization process [26]. Loss functions are significant in every predictive method because they establish a goal to measure the approach's performance. There are several types that differ according to the tasks, whether in classification or regression [27]. The parameters learned by the model are set by minimizing a given loss function.

Below are the main contributions:

- We propose a new architecture called Gated Convolution Long (GCL) for Arabic sentiment analysis. We address the issue of long training samples, extract the best features for binary and n-ary classification, and boost the effectiveness of ASA.
- We develop a custom regularization function (CRF), which helps to improve the performance of the proposed model.
- We perform an ablation study which demonstrates that the improved results are due to CRF.
- We conduct a comparison study with a standard loss function, and show that our custom regularization aids in optimizing the loss function's performance.
- We show that the proposed method offers the best classification performance relative to current baselines.
- We use the proposed method to investigate the link between sentiments in Modern Standard Arabic and those in five different Arabic dialects.
- Finally, we compare the proposed technique with the standard regularization function on very large Arabic datasets; our model incorporating CRF was more effective, performed better, and took less time.

The paper is organized as follows. Section 2 reviews previous work on sentiment analysis for Arabic. Section 3 outlines the proposed approach and model architecture. Section 4 presents our experiments, including preprocessing steps, experimental settings, baselines, results, and discussion. Finally, Section 5 is the conclusion and suggests future work.

2. Previous work

Arabic content has been significantly produced on websites and social media over the last ten years. On social media, opinions are freely expressed, making them an excellent source for trend analyses in various professional, commercial, and popular periodicals. See [54,55] for recent surveys of work in Arabic sentiment, addressing models, datasets, and results for significant modern research on the Arabic language. For contemporary deep learning methodologies and semi-supervision, refer to [56–58].

Current sentiment datasets for Arabic are shown in Table 1. As can be seen, five datasets are for Modern Standard Arabic (MSA) alone, seven are for Arabic dialects, Algerian (ALG), Jordanian (JOR), Lebanese (LEB), Levantine (LEV), Moroccan (MOR), Saudi (SAU), and Sudanese (SUD), while two combine MSA with Dialects (DIA). In this work, we will use AHSD, ArTwitter, MASC and BBN, as will be described later.

Table 2 summarizes recent research on Arabic sentiment analysis, including the dataset used, the form of Arabic (MSA or dialect), the model, and the performance result. We will now review these works, starting with those using machine learning. After this we will discuss neural network approaches.

Tabii et al. [50] applied Naïve Bayes (NB) [59], Maximum Entropy [60], and Support Vector Machines (SVM) [61] on two datasets, the Moroccan Sentiment Analysis Corpus (MSAC) [30] and SemEval-2017 [62]. Among the individual classifiers the SVM was the best, and when they used ensemble classifiers they achieved the highest accuracy (83.45%).

Table 1
Arabic sentiment datasets.

Datasets	Language	Source	Size	#Classes	Balanced
AHSD [28]	SAU	Twitter	104 K	2	N
ArTwitter [29]	JOR	Twitter	179.55 K	2	Y
MASC [30]	MOR	Google Play, Twitter, Facebook	3.27 MB	2	N
BBN [31]	LEV	Website posts	207 K	3	N
SudSenti2 [32]	SUD	Facebook, YouTube	344 K	2	Y
DzSenti [33]	ALG	Facebook	100 K	2	Y
AO [34]	MSA	Facebook	30.15 K	2	N
LD [35]	LEB	Google Maps, Zomato	31.32 K	2	N
ABD [29]	MSA	Twitter	20 K	2	Y
YT [36]	MSA	YouTube	70 K	2	N
HARD [37]	MSA	Book	54.36 MB	3	N
LABR [38]	MSA	Book	11.6 MB	3	N
ASTD [39]	MSA + DIA	Twitter	10 K	4	N
Shami-Senti [40]	DIA	Twitter	2.5 K	3	N
ArSentD-LE [41]	MSA + DIA	CrowdFlower platform	4 K	5	Y

Table 2
Previous work on Arabic sentiment analysis.

Paper	Dataset	Language	Model	Result
[42]	AHSD (2C)	SAU	CNN-LSTM	88.10%
[28]	AHSD (2C)	SAU	CNN	90.00%
[43]	AHSD (2C)	SAU	CNN + Word2Vec	92.00%
[44]	AHSD (2C)	SAU	BiLSTM	92.61%
[45]	ArTwitter (2C)	JOR	RNN	85.00%
[46]	ArTwitter (2C)	JOR	LSTM	87.27%
[44]	ArTwitter (2C)	JOR	BiLSTM	91.82%
[47]	BBN (3C)	LEV	Linear Classifier	65.31%
[48]	BBN (3C)	LEV	CNN	66.67%
[49]	BBN (3C)	LEV	CNB	71.06%
[46]	Web-crawled (2C)	MSA	CNN	85.01%
[50]	MSAC (2C)	MOR	SVM	83.45%
[36]	YouTube text (2C)	MSA	SVM	77.00%
[51]	JDT (2C)	JOR	SVM + LR	82.10%
[52]	ABD (2C)	MSA	DMNB	87.50%
[53]	SudSenti2, SudSenti3 (2C,3C)	SUD	SCM + MMA	92.75%, 84.39%
[35]	Lebanon dialect (2C)	LEB	LR	88.00%
[34]	Arabic opinions (2C)	MSA	SVM	76.33%

Afooz et al. [36] compared their Ensemble model, which included XGBoost (XG), Gradient Boosting (GB), AdaBoost (ADA) and Random Forest (RF), with machine learning classifiers on Arabic text which was collected from YouTube comments. SVM, followed by Linear Regression (LR), had the best performance accuracy (77.00%).

Atoum and Nouman [51] used SVM and NB with n-gram vector selection (bigrams, unigrams, and trigrams), on Jordanian dialect tweets (JDT). Results showed that the SVM gave the higher accuracy in all cases (bigrams 74.00%, stemmed unigrams 82.10%, trigrams 76.00%). AlSalman [52] developed a Discriminative Multinomial Bayes (DMNB) approach, then compared it with NB, SVM, K-Nearest Neighbor (KNN) [63], and Decision Trees [64] on an Arabic dataset, which consists of 2,000 Arabic tweets with two classes. DMNB had the best accuracy with 87.50%, better than the baselines. Al Omariet et al. [35] used LR [65] on the Lebanon dialect which they collected from restaurants, shops, hotels, Google, and Zomato. Their results indicated that the binary rating of negative feelings (P = 0.80, R = 0.80) is less than the positive (P = 0.88, R = 1.00). Salameh et al. [47] created a dataset of Levantine Arabic sentiment which they called the BBN Dataset, then applied their Linear systems, with a performance of 65.31%. El-Beltagy et al. [49] also used the BBN Dataset, this time applying a Complement Naïve Bayes (CNB) classifier [66], and achieving accuracy 71.06%. Al-Kabi et al. [34] applied the SVM, NB, and KNN algorithms, with three tools (SentiStrength, SocialMention, and AOPI), to a corpus of 3,015 Arabic opinions, which they collected from three main domains: Food, Sport, and Weather. SVM proved the best (76.33%). The AOPI tool was shown to be more effective than the two free online tools.

We now discuss the deep learning approaches to Arabic sentiment analysis. Alayba et al. [42] utilized a combination of CNN and LSTM with various tokens (ch5-gram-level, and word-level) on the Arabic Health Services Dataset (AHSD). The best result was 88.10%. Dahou et al. [67] used a CNN with two word embedding models, CBOW and Skip-Gram, to create vector representations. The corpus comprised 10 billion words collected from web pages. Performance was 85.01%. Al-Azani and El-Alfy [46] applied CNN [68], LSTM [69], and CNN-LSTM to analyze ArTwitter datasets; LSTM with dynamic CBOW gave the best result (87.27%). Mhamed et al. [53] presented two Sudanese Arabic sentiment datasets, one 2-Class (SudSenti2) and one 3-Class (SudSenti3). After detailed preprocessing, they applied their proposed classifier, the Sentiment Classification Model with Mean Max Average Pooling (SCM + MMA). Accuracy was 92.75% for the 2C dataset and 84.39% for the 3C. Their model showed the best performance compared to ML and NN classifiers.

Boudad et al. [48] used CNN with Skip-Gram and CBOW on the BBN dataset, and accuracy was 66.67%. Elshakankery and Ahmed [45] proposed hybrid incremental learning, which consists of two machine learning classifiers and one deep learning model. The deep learning classifier was always RNN [70], while the ML classifiers were LR and SVM. They further applied the same methods to the Mini Arabic Sentiment Tweets Dataset (MASTD) with an accuracy of 83.73%. For ArSAS [71], accuracy was 81.52%, for GS [72] accuracy was 68.09%, for the Syrian Corpus [73] accuracy was 85.28%, and for ArTwitter, it was 85.00%. Alayba et al. [28] applied NB, LR, SVM, DNNs, and CNN to the AHSD dataset. CNN gave the highest performance with accuracy 90.00%. Alayba et al. [43] enhanced the accuracy for the previous AHSD datasets by using CNN, but this time with the Word2Vec [74] model, which was constructed from a large Arabic journal corpus; accuracy was 92.00%. Recently, Elfaik et al. [44] used a Bidirectional LSTM [75] model on several datasets: ASTD [39], ArTwitter, LABR, MPQA [76], Multi-Domain [77], and AHSD. Accuracies were 79.25%, 91.82%, 80.70%, 75.85%, 89.70%, and 92.61%, respectively.

In summary, for Arabic sentiment models using ML, Tabii et al. [50], Afooz et al. [36], and Al-Kabi et al. [34] use SVM, Atoum and Nouman [51] combine SVM with LR, AlSalman [52] use DMNB, Al Omariet et al. [35] [47] use LR, and Salameh et al. [49] use CNB.

For the deep learning, Al-Azani and El-Alfy [46], Boudad et al. [48], and Alayba et al. [28,43] utilized CNN, Elshakankery and Ahmed [45] and Elfaik et al. [44] employed Bi-LSTM and RNN, and Alayba et al.[42] used CNN with LSTM.

Concerning the machine learning classifiers, we note that SVM was the most used, while LR achieved the highest performance. Generally, ML models have three problems. First, they are susceptible to noise; a small amount of incorrectly labeled samples can have a significant impact on performance. Second, selecting the perfect kernel is a difficult undertaking. Third, when the dataset is large, training is slow.

Regarding DL, CNNs were more applied, but Bi-LSTM showed the best accuracy among all the algorithms. For the CNNs, obstacles were the standard selection of the optimal architecture with normal hyperparameters for training, and poor preprocessing of the Arabic text, causing the data held in adjacent words not to be learned effectively, hence reducing the CNN's capability to select the best features for prediction. For Bi-LSTM, i.e. using two LSTM cells, one for each direction, it is costly, and it took a long time to train the Arabic context. Generally, we note that the DL models are better than the ML classifiers.

In this work, we will present an architecture called GCL, with unique regularization functions for 2C and 3C sentiment classification. Regularization (see next section) is an extra approach aimed at improving the model's generalisation, producing better results on the test set [78].

Additionally, while evaluating, we consider the loss function's accuracy and performance compared to other loss functions such as Binary-Cross-Entropy, Hinge, Poisson, and KL-divergence.

3. Proposed method

3.1. Outline

We designed a new architecture for ASA called GCL, based on various deep neural architectures with novel CRF Fig. (1). We also developed our previous preprocessing approaches [79], with different cleaning of the Arabic context (C_P). We start by choosing the Arabic sentence inputs $X_N = (x1, x2, \ldots, xn)$ and figuring out the context length, which varies from corpus to corpus. The data cleaning process $Y_M = (y1, y2, \ldots, ym)$, begins from the input by removing special characters, punctuation marks, and all diacritics (see next subsection). The proposed method works with both 2C and 3C classification. We trained and tested on the AHSD (2C, SAU), ArTwitter (2C, JOR), MASC (2C, MOR), and BBN (3C, LEV) Arabic datasets.

3.2. Text preprocessing and normalization steps

Text data created from natural language is noisy and unstructured. Text preprocessing entails putting text into a neat, standardized structure to convert it into a form suitable for further analysis and training [80]. Text preprocessing methods may be broad so that they can be used in various applications, or they can be tailored for a particular goal. For instance, the techniques used to analyze scientific articles, including equations and other mathematical symbols, may differ greatly from those used to analyze user feedback on social networking sites [81]. The preprocessing steps used (see Fig. 2) were similar to those we developed previously [79]:

- We removed special characters, punctuation marks, and all diacritics.
- We removed all digits, including dates.
- We removed repeated characters, keeping only one or two repeated characters.
- We removed any non-Arabic characters.
- We applied the tokenizer from the Keras package [82].
- We used the standard Arabic stopwords from NLTK [83].
- We carried out text normalization [79].

3.3. Input layer

All the features in a sample are represented by the initial vector sequence M ($m_1 \ldots m_i \ldots m_n$) where n is the number of features.

We use AraVec [85] to convert Arabic words into vectors. This process is based on Word2vec [74] which is an open-source tool that pre-trains word embeddings on a large data set, and which was originally used for English.

As is well known, the key idea behind word embeddings is that words with similar meanings are converted to similar vectors, which enables that similarity to be determined by vector comparison methods such as dot product or angle. They are also ideal for input to neural network models, as has been shown for a large number of NLP tasks in many different languages, including Arabic.

In the context of word embeddings, there are three interesting questions to consider, (1) how to handle word polysemy, i.e. words with many meanings, (2) how to handle metaphorical or hidden meanings, and (3) how to handle words whose meanings differ from country to country.

Concerning polysemy, a good example in Arabic is 'hib' which can mean 'love' or 'seed'. Word embeddings such as AraVec are the result of training on datasets. In cases where a word can have many meanings within the training data used, the resulting vector will reflect aspects of all these, i.e. it will maintain and reflect the ambiguity. Then, later stages of the neural network model, which uses the embeddings as inputs, will tend to select the appropriate senses through the domain specific learning process.

Turning to hidden meanings, an example is 'Asad' which literally means 'he is a lion', but which actually means 'he is a hero'. The datasets used in our experiments comprise social media exchanges, which are informal and contain many indirect uses of words and phrases, such as the description of a hero as a lion. A neural network such as the proposed model cannot solve this prob-

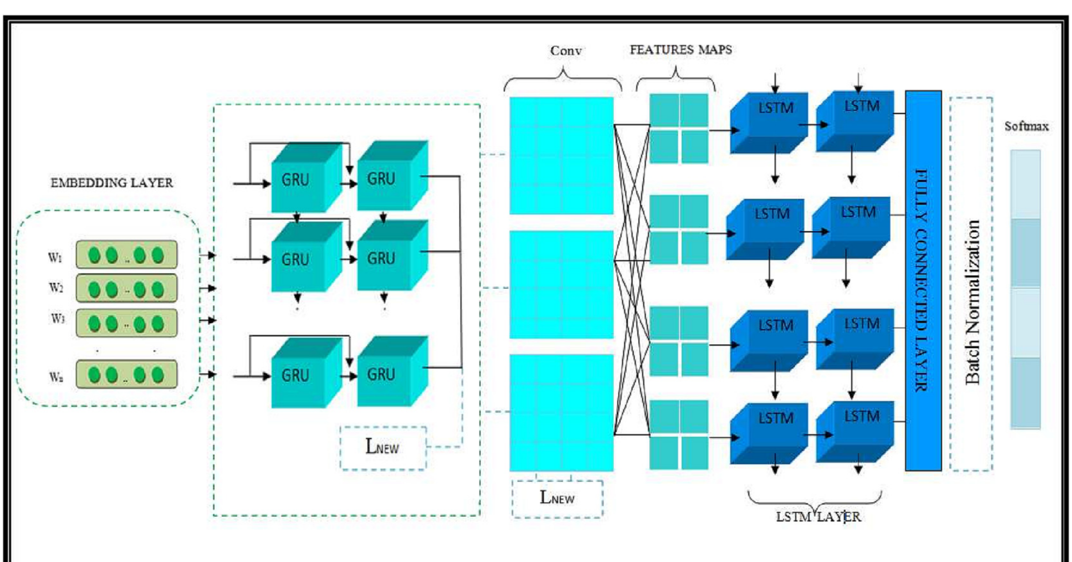

Fig. 1. GCL Model Architecture.

English language	Text	After preprocessing
And this topic is not suitable for anyone other than “Off”!!!	!!! «Off» و مو مناسب هالموضوع لحداااا غير لـ	مناسب الموضوع حد غير
Whoever visited the store will repeat the visit again.	من زار المـــــــــحل سيكرر الزيارة مرة أخـــــرى فعلاً@	زار المحل كرر الزياره مره اخري فعلا
Eat and drink, but do not be extravagant, for He does not like the extravagant.	وكُلُواْ وَاشْرَبُواْ وَلاَ تُسْرِفُواْ إِنَّهُ لاَ يُحِبُّ الْمُسْرِفِينَ	وكلو واشربوا ولا تسرفو انة لا يحب المسرفين
Very sweet, the right light at the right time, 100%.	#### حلو كتيييير الضوء المـــــــــناسب في الزمان المناسب 100%	حلو كتير الضوء المناسب الزمان المناسب
It's nice to be in restaurants, cafes, supermarkets, and ATMs in the Dead Sea.	حلو يكون في مطاعم و كفيهات و سوبرماركت و ATM بالبحر الميت	حلو كون مطاعم كفيهات سوبرماركت البحر الميت

Fig. 2. Steps for Arabic corpus data preparation.

lem directly, but it can do so implictly. For example, if a tweet refers to someone as a lion and the sentiment associated with the training data instance is positive, the model can learn that being described as a lion is a positive attribution, similar to being described as a hero. In such a way, the sentiment assigned to an unseen input can still be correct, even in cases of metaphorical language use.

Thirdly, we can find words whose meaning varies radically from country to country, depending on the Arabic dialect spoken. For example, ‘lbin’ can mean raw milk in one country and a product called ‘Laban Rayeb’ in others. In a system based on vector embeddings from AraVec, the meaning(s) associated with a word will depend on the dialects spoken in the training data used to create the embeddings. A typical dataset may indeed contain instances of different Arabic dialects. However, the proposed model does not rely on the interpretation of any single word in the input text; instead, the meanings of all words are converted to vector form and then input to the CNN model. In this way, the model can learn to overcome contradictions resulting from the incorrect interpretation of words. Thus, overall, we can see that the use of word embeddings in a neural network model can alleviate these three problems, still resulting in a sentiment analyis tool of high accuracy, while it cannot completely solve them.

After the embedding layer to vectorize the Arabic context Fig. (3), we then applied the GCL architecture, which is a GRU with CNN through LSTM. We trained the model to perform sentiment analysis on various dialects. In addition, the relationship between MSA and other Arabic dialects was explored via a cross-dialect training study.

3.4. GRU layer

As is well-known [86], a GRU has gating units that modulate information flow within the unit without providing separate memory cells. It calculates two gates, called update and reset, that control information flow through each hidden unit. It is shown in Fig. 4 and defined by the following equations:

$$r_s = \sigma(W_r X_s + U_r h_{s-1} + b_r) \quad (1)$$

$$z_s = \sigma(W_z X_s + U_z h_{s-1} + b_z) \quad (2)$$

$$\bar{h}_s = tanh(W x_s + U_{r_s \odot h_{s-1}} + b_h) \quad (3)$$

$$h_s = z_s \odot h_{s-1} + (1 - z_s) \odot \bar{h}_s \quad (4)$$

where z_s represents the update gate, W_z, U_z are weight matrices, and r_s is a reset gate. The input vector M_s, of all of these components is set to produce the now concealed state h_s and the previously hidden state h_{s-1}. The logistic sigmoid function is denoted by α, and $\odot$ is the multiplication of elements.

The update gate is determined from the current input and the preceding time phase hidden state. This gate determines how much new memory and old memory parts in the final memory can be mixed. The reset gate is measured similarly but with different sets of weights. It manages the balance between previous and new memory input. Here we applied our GRU layer with 128 filters and the custom regularization function. Our GRU layer has shown superior ability to handle lengthy Arabic context data Fig. (5) during the training, and to be faster than other approaches.

3.5. Convolutional layer

In this layer [87] we extract the local and multiple features, by using the following equation which illustrates how a filter F_i learns feature Map M^i_j:

$$M^i_j = f(V_{j:j+W-1}) \odot W^i + b^i \quad (5)$$

where W represents the matrix weight bias, $V_{j:j+W-1}$ is a token vector, $\odot$ is the convolution operation, with max pooling, or average

Fig. 3. Arabic context visualization.

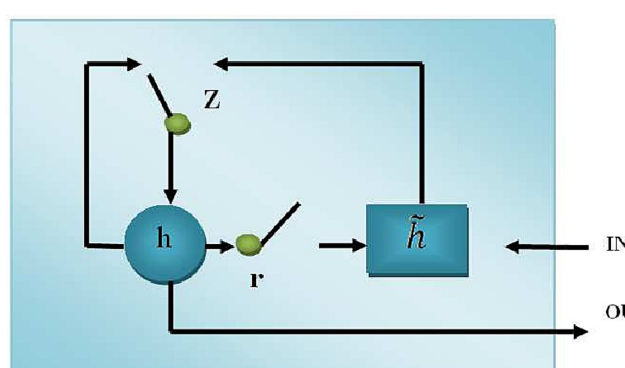

Fig. 4. GRU architecture (derived from Le [84]).

Sentiment analysis	English language	Arabic text samples	Datasets
Negative	According to my follow-up in this hospital, before entering any doctor, I do research and most of his doctors praise them, but my suggestion is that the management be changed and a special unit for licenses be established so that the follow-up is carried out first and for a hospital like its capabilities not to fall into this error.	حسب متابعتي في هذا المستشفى قبل دخولي لاي دكتور اعمل بحث واغلب اطبائه مشاد فيهم لكن مقترحي ان يتم تغيير الاداره وانشاء وحده خاصه عن التراخيص لكي يتم المتابعه اول باول ولا يقع مستشفى مثل امكانياته بهذا الخطا	AHSD
Negative	Unfortunately, some believe that the nuclear reactor will solve our economic problems, but Jordan does not have financial resources, and this means the need for a foreign partner, "as announced by the Energy Authority," who owns the majority of shares, as in Turkey, for example, and this means that this partner will own the largest part of the reactor and will not be owned by Jordan, and therefore the government will also pay the costs of electricity to the investor, and therefore the nuclear reactor will not provide you with electricity at a low price.	للأسف البعض يعتقد ان المفاعل النووي سيحل مشاكلنا الاقتصادية ولكن الأردن لا يمتلك موارد مالية وهذا يعني الحاجة لشريك أجنبي "كما اعلنت هيئة الطاقة" يملك غالبية الأسهم كما في تركيا على سبيل المثال وذلك يعني ان هذا الشريك سيمتلك الجزء الأكبر من المفاعل ولن يكون ملك للأردن وبالتالي ستقوم الحكومة ايضاً بدفع تكاليف الكهرباء للمستثمر وبالتالي المفاعل النووي لن يوفر لك الكهرباء بسعر قليل	ArTwitter
Positive	I graduated from the university with an excellent average, congratulations, and may God help you with this responsibility and honesty, but you have led it and may God bless you, top boy. Convince them to change what is under my knowledge from a new writer to anything. If you can convince them of this, it will be really something.	تخرجت من الجامعة بمعدل ممتاز، ألف مبروووك والله يعينك على هالمسؤولية والأمانة بس أنت قدها وقدوووود الله الله يا فتى القمة أقنعهم يغيرون ما تحت معرفي من كاتب جديد إلى أي شيء إذا قدرت تقنعهم بهالشي تكون فعلاً شي..	BBN
Positive	The price is excellent and the teaching is excellent, but there is no credibility because, frankly, I fell into the trap of fraud because the diploma certificate was not certified by the government agency affiliated with the institute, and they certified all certificates as separate courses with conflicting dates, as if they had not already taken a full diploma for a year, so the diploma was only certified by the same institute, and this is what made the diploma useless in relation to government jobs. I was satisfied with the courses as if they were separate circuits, and not all the certificates, but only those that do not have a conflict in the dates, because the diploma and certificates that have a conflict were rejected by the government agency.	السعر ممتاز والتدريس ممتاز ولكن لا توجد مصداقية لاني وبصراحة وقعت في فخ النصب لان شهادة الدبلوم لم تكن مصدقة من قبل الجهة الحكومية التابعة للمعهد وقد قاموا بتصديق الشهادات جميعها كدورات منفصلة مع تواريخ متضاربه وكأنها بالفعل لم تأخذ كدبلوم كامل لمدة سنة فالدبلوم لم يصدق الا من نفس المعهد وهذا ماجعل الدبلوم عديم الفائدة بنسبة للوظائف الحكومية فقط اكتفيت بالدورات وكأنها دوارات منفصلة وليس كل الشهادات أنما فقط التي لا يوجد بها تضارب في التواريخ لان تم رفض الدبلوم والشهادات التي بها تضارب من قبل الجهة الحكومية (((من جد الله يسامحهم)))	MASC

Fig. 5. Sample texts from the datasets.

pooling to pick various features from the $M_j^{i\times j}$. Here we add our Custom Regularization Function (see Section 3.8), and dropout to avoid overfitting and optimize the performance.

3.6. LSTM layer

This solves the problem of disappearing error gradients and captures long-term dependencies [89]. Three internal gates govern the flow of data to and from the memory blocks, as shown in Fig. 6 and defined as follows:

$$h_s = f(W_s.m_s + U_s.h_{s-1}) \tag{6}$$

$$f_s = \sigma(W_f.X_s + U_f.h_{s-1} + b_f) \tag{7}$$

$$i_s = \sigma(W_i.X_s + U_i.h_{s-1} + b_i) \tag{8}$$

$$o_s = \sigma(W_o.X_s + U_o.h_{s-1} + b_o) \tag{9}$$

$$c_s = f_s \circ c_{s-1} + i_s \circ tanh(W_c.X_s + U_c.h_{s-1} + b_c) \tag{10}$$

$$h_s = o_s \circ tanh(c_s) \tag{11}$$

where h_s is a regular hidden state, W_s, U_s are weight matrices, m is the input vector, $f(v)$ is a non-linear function, f_s is the forget layer, σ is a sigmoid function, Xs is a cell parameter, b is the Bias, i_s is the input layer, chosen to be tanh, and Os is an output gate.

Our LSTM layer, with various outputs, processes the data and deals effectively with data noise as well as continuous values from the preceding layer.

3.7. Regularization

Prior to the development of deep learning, regularization was utilised for decades. Simple functions have usually been used with machine learning models and statistical approaches. The regularisation did not need to be as sophisticated since the functions were less capable [90]. Classical regularizations are divided into two categories:

$$L_2 = loss + \Lambda/2M + \Sigma||w^2|| \tag{12}$$

$$L_1 = loss + \Lambda/2M + \Sigma||w|| \tag{13}$$

where L_2, L_1 are the Regularization functions, $loss$ is the loss function, Λ is the regularization parameter, M is the number of the layer, and w is the weight for the layer.

3.8. CRF

The proposed Custom Regularization Function is illustrated in Fig. 7. We start with standard Regularization. L_2 on the weight side soon forces all the values from zero.[1] It is very good. L_1 presses directly the weight to zero, and it is weak compared to L_2 [91].

When we customize our L_{new} (Fig. 7) by calculating the absolute value among the values to be zero, the average of the values will tend to zero.

To optimize zone selections for our hyperparameter weights, our proposed extension is wider than L_1 and L_2, which helps to provide the best features for future selection. The Custom Regularization Function L_{new} is defined as:

$$L_{new} = loss + \Lambda/2M + \Sigma||(w * w - w/2)|| \tag{14}$$

L_{new} helps improve the mean cost, which makes the overall errors small. In summary, the contribution of L_{new} to the proposed method is:

- It is more sensitive to the quality of the output.
- It lessens the complexity of the model.

[1] https://forum.huawei.com/enterprise/en/what-is-regularization/thread/724117-895.

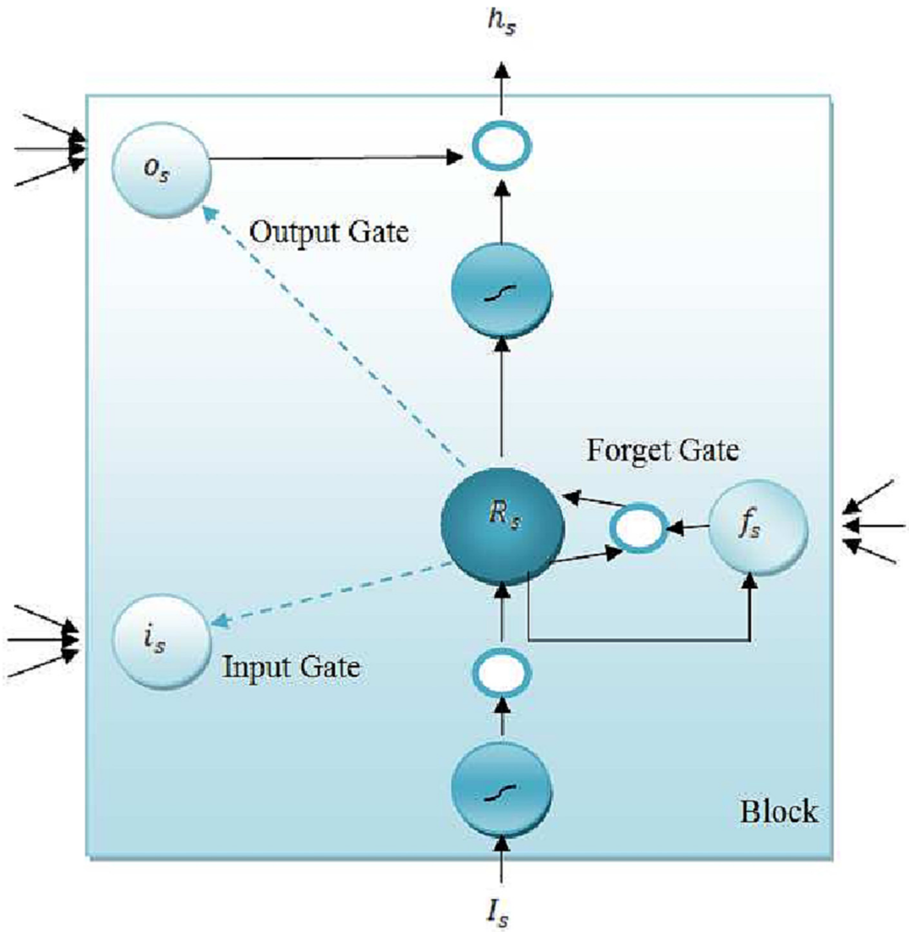

Fig. 6. LSTM architecture (derived from Yu et al. [88]).

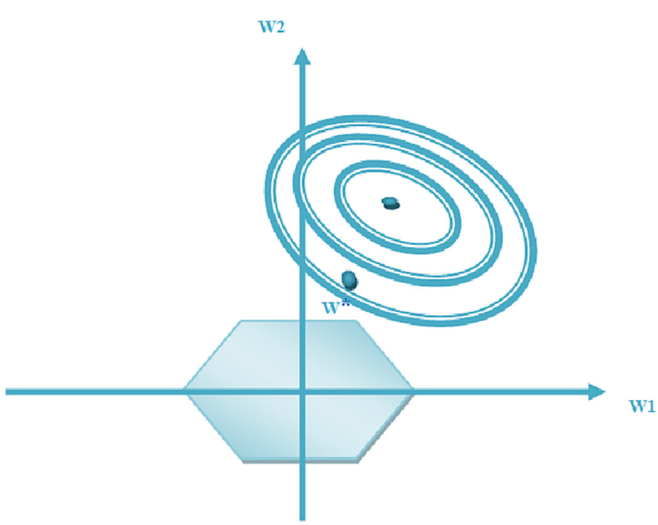

Fig. 7. Custom regularization.

3.9. GCL model architecture

The proposed architecture was shown earlier in Fig. 1. This includes an embedding layer which can turn each word into a fixed-length, predetermined-size vector using embeddings; max-features represents the number of unique words, embedding-size equals 128 or 300 with a max-len of [30, 50, or 150]; after that a gated recurrent unit with 128 filters, which can solve long sequence training issues and improve efficiency and accuracy. After that, the convolutional neural network layers with 64 filters; they are capable of using different lengths and weights of windows for the number of feature maps to be created, and can be used for both dual and multiple classifications.

Kernel size is equal to three – this is the width and height of the filter mask for the CNN layer. Padding is set to 'valid', activation is equal to ReLU. This provides nonlinearity to a system that has essentially only been doing linear computations throughout the Conv layers.

Strides is equal to one, followed by [global average pooling 1D, global max pooling 1D]. Pool size equals two, then the customized regularization function for both previous layers, which helps us to improve the performance and optimize the validation loss when we compare to the classification loss functions – see Section (4.4). After that, Dropout (0.25) and an LSTM with output [90, 80, or 50], then Flatten, then batch normalization which lessens the gradient's reliance on the parameters' original values or scales and decreases the inner variational shifting, and finally, a dense layer with a softmax or Sigmoid layer i.e. a fully connected layer to predict the output of the class from either three sentiment classes (Positive, Negative, Neutral), or two classes (Positive and Negative).

4. Experiments

4.1. Datasets

Our model is trained on the AHSD, ArTwitter, MASC, and BBN datasets (see Table 1 earlier). Table 3 shows the outline statistics. The datasets can be described as follows:

Table 3
Datasets for our experiments.

Dataset	Dialect	POS	NEG	NEU	Total
AHSD (2C)	SAU	628	1398	-	2,026
ArTwitter (2C)	JOR	1000	1000	-	2,000
MASC (2C)	MOR	4,476	2,257	-	6,733
BBN (3C)	LEV	498	575	126	1,119
DzSenti (2C)	ALG	24,932	24,932	-	49,864
LABR (3C)	MSA	42,724	8,174	12,168	63,066

The Arabic Health Services Dataset, AHSD[2] is for Saudi Arabic (SAU). It was collected from Twitter by Alayba et al. [28] and contains 2,026 tweets, with two unbalanced classes, 628 positive tweets, and 1,298 negative tweets.

ArTwitter[3] is for Jordanian Arabic (JOR). It was created manually from Twitter [29] and consists of two balanced classes, 1,000 positive tweets and 1,000 negative.

The Multi-domain Arabic Sentiment Corpus, MASC[4] is for Moroccan Arabic (MOR). It contains 8,860 ratings from various realms and dialects of Arabic [30]. The information was gathered manually from a variety of sources, including the Jeeran and Qaym websites, Google Play, Twitter, and Facebook. There are 4,476 positive tweets and 2,257 negative.

The BBN Dataset, BBN[5] is for Levantine Arabic (LEV) and consists of three classes, 498 positive, 575 negative, and 126 neutral [31]. It uses as its starting point a random set of 1,200 Levantine dialect phrases taken from the BBN Arabic-Dialect-English Parallel Text which itself consists of Levantine-English and Egyptian-English parallel texts [92].

The DzSenti corpus [33] consists of 49,864 items, 24,932 negatives, and 24,932 positives, including MSA with Algerian dialect. It is publicly available.[6]

LABR, the Large-Scale Arabic Book Review dataset,[7] was developed by Aly et al. [38] and encompasses 63,000 items in MSA, with three classes, 42,724 positives, 8,174 negatives, and 12,168 neutral.

4.2. Experimental settings

We train the model and evaluate using the following metrics. True Positives (TP) is the number of correctly classified positive Tweets, True Negatives (TN) is the number of correctly classified negative Tweets, False Positives (FP) is the number of tweets incorrectly classified as positive, and False Negatives (FN) is the number of tweets incorrectly classified as negative. After that, the following performance measures are computed [93]:

$$Accuracy = (TP + TN) \div (TP + TN + FP + FN) \quad (15)$$

$$Precision = TP \div (TP + FP) \quad (16)$$

$$Recall = TP \div (TP + FN) \quad (17)$$

$$F1 = 2 * (Precision * Recall) \div (Precision + Recall) \quad (18)$$

These measures are widely used in related work [94–96]. The following tuning and hyperparameter settings were used: Embedding size [128, 300], Pooling [2, 4, 6], Batch-size [64, 128, 164],

[2] https://bitbucket.org/a_alayba/arabic-health-services-ahs-dataset/src/master/.
[3] www.kaggle.com/lakshmi25npathi/twitter-data-set-for-arabic-sentiment-analysis.
[4] http://github.com/almoslmi/masc.
[5] https://github.com/ZarahShibli/sentiment_analysis.
[6] https://github.com/adelabdelli/DzSentiA.
[7] http://www.mohamedaly.info/datasets/labr.

Table 4
Experiment 1: Accuracy of proposed model with 2C and 3C datasets in different dialects.

Dataset	Dialect	Model	Accuracy	F1
AHSD (2C)	SAU	[42]	88.10%	-
		[28]	90.00%	-
		[43]	92.00%	-
		[44]	92.61%	86.03
		Proposed Model	**95.50%**	**95.00**
ArTwitter (2C)	JOR	[45]	85.00%	-
		[46]	87.27%	-
		[44]	91.82%	92.39
		Proposed Model	**93.88%**	**93.50**
MASC (2C)	MOR	[50]	83.45%	-
		Proposed Model	**86.64%**	**85.50**
BBN (3C)	LEV	[47]	65.31%	-
		[48]	66.67%	-
		[49]	71.06%	-
		Proposed Model	**74.92%**	**74.10**

Table 5
Experiment 2: Comparison of loss functions, as measured by loss and validation loss, when used with GCL on the four datasets.

Approach	Loss	Validation loss	Time
	AHSD		
GCL + Binary-Cross-Entropy	0.00002361	0.437	318s
GCL + Hinge	0.5025	0.5571	274s
GCL + Poisson	0.5	0.6934	335s
GCL + KL-divergence	0.00001443	0.4316	361s
GCL+ (CRF + Binary-Cross-Entropy)	0.1175	**0.3822**	289s
	ArTwitter		
GCL + Binary-Cross-Entropy	0.002649	0.4491	162s
GCL + Hinge	0.5045	0.5624	103s
GCL + Poisson	0.501	0.6868	115s
GCL + KL-divergence	0.002544	0.3896	268s
GCL+ (CRF + Binary-Cross-Entropy)	0.1304	**0.3685**	126s
	MASC		
GCL + Binary-Cross-Entropy	0.008539	0.8894	344s
GCL + Hinge	0.5119	0.6094	371s
GCL + Poisson	0.5035	0.8068	352s
GCL + KL-divergence	0.01303	0.6266	345s
GCL+ (CRF + Binary-Cross-Entropy)	0.1297	**0.5449**	338s
	BBN		
GCL + Binary-Cross-Entropy	0.02139	0.6013	271s
GCL + Hinge	0.68	0.7641	265s
GCL + Poisson	0.3406	0.5852	306s
GCL + KL-divergence	0.02217	0.8229	265s
GCL+ (CRF + Binary-Cross-Entropy)	0.1853	**0.5442**	272s

Table 6
Experiment 3: Ablation Study. Accuracy of proposed GCL model is shown with and without custom regularization function CRF.

Model	AHSD	ArTwitter	MASC	BBN
GCL with CRF	95.50%	93.88%	86.64%	74.92%
GCL without CRF	90.40%	91.77%	84.50%	70.80%

Table 7
Experiment 4(a): Cross-dialect training experiments between MSA and dialects, using proposed GCL model.

Train	Test	Accuracy
HARD (MSA)	HARD (MSA)	94.27%
	AHSD (SAU)	85.29%
	MASC (MOR)	84.20%
	BBN (LEV)	83.24%
	SudSenti2 (SUD)	80.57%
	ArTwitter (JOR)	78.98%

Table 8
Experiment 4(a): Results in terms of P, R, F, Macro Average, broken down by Positive and Negative sentiment.

Train vs. Test	Precision		Recall		F1		Macro Avg
	Positive	Negative	Positive	Negative	Positive	Negative	
HARD vs. HARD	0.93	0.93	0.95	0.95	0.93	0.95	0.94
HARD vs. AHSD	0.80	0.90	0.89	0.83	0.84	0.86	0.85
HARD vs. MASC	0.85	0.84	0.85	0.84	0.85	0.84	0.84
HARD vs. BBN	0.79	0.89	0.90	0.76	0.84	0.82	0.83
HARD vs. SudSenti2	0.78	0.83	0.84	0.77	0.81	0.80	0.81
HARD vs. ArTwitter	0.75	0.85	0.87	0.72	0.80	0.78	0.79

Kernel-size [3, 5], Number-classes [2, 3], Epoch [10, 50, 100], with Adam optimizer and 0.001 Learning Rate. For the implementation, we used the Tensorflow framework.[8]

4.3. Experiment 1: 2C and 3C sentiment classification

We applied the proposed method (GCL) to the four datasets, AHSD (2C), ArTwitter (2C), MASC (2C), and BBN (3C). Ten-fold cross-validation was used for all models, using a random 80% for each training, and the remaining 20% for testing. Results are in Table 4. The accuracy of the proposed method, GCL, was 95.50% for AHSD, 93.88% for ArTwitter, 86.64% for MASC, and 74.92% for BBN. In all cases these are higher than the previous baselines.

4.4. Experiment 2: comparison of loss functions

The loss function is the function that determines the distance between the algorithm's current outcome and the desired output. It provides a means of assessing how well the prediction mimics the data.[9] Loss functions are used to calculate the amount a model should try to reduce its error throughout learning. There are various types.[10]

We utilized the proposed method (GCL) with various loss functions: Binary-Cross-Entropy, Hinge, Poisson, and KL-divergence. We also included our customized regularization plus binary-cross-entropy (CRF + binary-cross-entropy). The four datasets from the previous experiment were used, with the same ten-fold cross-validation for all approaches. Results are in Table 5.

For AHSD, validation losses with GCL + BC, GCL + Hinge, GCL + Poisson, GCL + KL-divergence, and GCL + (CR + Binary-Cross-Entropy) were 0.437, 0.5571, 0.6934, 0.4316, and 0.3822, respectively. GCL+ (CR + Binary-Cross-Entropy) had the lowest validation loss (0.382), and the lowest time except for GCL + Hinge.

For ArTwitter, MASC and BBN, GCL + (CR + Binary-Cross-Entropy) also had the lowest validation loss (0.3685, 0.5449, 0.5442). We therefore conclude that the proposed method with customized regularization plus binary-cross-entropy was the best performing model on the four datasets.

4.5. Experiment 3: ablation study

We carried out an ablation study on the proposed method using the four datasets. Training of the proposed GCL model was done both with and without the proposed custom regulation function (CRF). We used ten-fold cross-validation and report the average results in Table 6. Accuracies of GCL + CRF using AHSD, ArTwitter, MASC, and BBN were 95.50%, 93.88%, 86.64%, and 74.92%. For GCL without CRF, these reduced to 90.40%, 91.77%, 84.50%, and 70.80% respectively, changes of −5.10%, −2.11%, −2.14%, and −4.12%. The results show that CRF improves the performance of GCL during training, for all datasets.

Table 9
Experiment 4(b): Cross-dialect training between MSA and all dialects combined, using proposed GCL model.

Train	Test	Accuracy
HARD (MSA)	MAMBS (SAU, JOR, MOR, LEV, SUD)	76.70%

4.6. Experiment 4: MSA-dialect association study

We used our model to study the Arabic sentiment association between MSA and Arabic dialects. We chose 5,000 MSA texts from HARD[11] [37] with 2,500 positive examples and 2,500 negative. For AHSD (SAU), ArTwitter (JOR), MASC (MOR), SudSenti2[12] (SUD) [53], and BBN (LEV) we took 2,000 tweets from each one to represent text samples in these Arabic dialects.

In the first part, we trained our model on HARD and then evaluated on AHSD, ArTwitter, MASC, SudSenti2, and BBN. Results are in Tables 7 and 8. As the results show, the best result is obtained by training and testing on MSA (94.27%). We can consider this our 'baseline' in comparing dialects with MSA. The highest accuracy after that is for SAU (85.29%, −8.98%), followed by MOR (84.20%, −10.07%), and LEV (83.24%, −11.03%). There is then a gap of 2.67% before we reach SUD (80.57%, −13.70%) and JOR (78.98%, −15.29%). So we can conclude that the most similar dialect to MSA is SAU and that the least similar dialects are SUD and JOR.

In the second part, we combined the five previous dialect datasets and named it the Main Arabic Multi Binary Sets (MAMBS) corpus. We then trained GCL on HARD and tested on MAMBS. Results are in Tables 9 and 10. Accuracy was 76.70%, compared to our 'baseline' figure of 94.27% from Table 7. Naturally this is lower, and indeed it is behind the lowest figure in Table 7 (JOR, 78.98%), exactly as we would expect. What this result suggests is that we can achieve a useful performance figure on different dialects when training on MSA, but to achieve high accuracy, we need to use specialized training data.

The Saudi dialect comprises seven local dialects derived from ancient Arabic, while the Moroccan dialect has words from the Spanish and French dictionaries. Certain words from Turkish and English appear in the Sudanese dialect. Also, the Lebanese dialect contains influences from Aramaic and Syriac. Since the essence of all dialects is in the MSA, some differences in vocabulary resulted in various associations with MSA in the results when using the proposed methods in the classification tasks. Please refer back to Section 3.3 for further dis-

[8] https://github.com/mustafa20999/ASA-using-GCL-based-architectures-and-CRF.
[9] https://machinelearningmastery.com/how-to-choose-loss-functions-when-training-deep-learning-neural-networks/.
[10] https://keras.io/api/losses/.
[11] https://github.com/elnagara/HARD-Arabic-Dataset.
[12] https://github.com/mustafa20999/Sudanese-Arabic-Sentiment-Datasets.

Table 10
Experiment 4(b): Results in terms of P, R, F, Macro Average, broken down by Positive and Negative sentiment.

	Precision		Recall		F1		Macro Avg
	Positive	Negative	Positive	Negative	Positive	Negative	
HARD vs. MAMBS	0.73	0.81	0.81	0.72	0.77	0.76	0.77

Table 11
Experiment 5: GCL with several regularizations on the huge LABR and DzSenti datasets.

Model	LABR	Time	DzSenti	Time
GCL+L_1	91.36	43 m 30s	86.55	1 h 2 m 18s
GCL+L_2	91.71	1 h 24 m 18s	86.92	27 m 3s
GCL+$L_{ElasticNet}$	92.35	38 m 55s	86.97	21 m 41s
GCL+L_{new}	**92.53**	**33 m 49s**	**87.26**	**20 m 27s**
Baselines	91.9% [97]	-	86.00% [17]	-

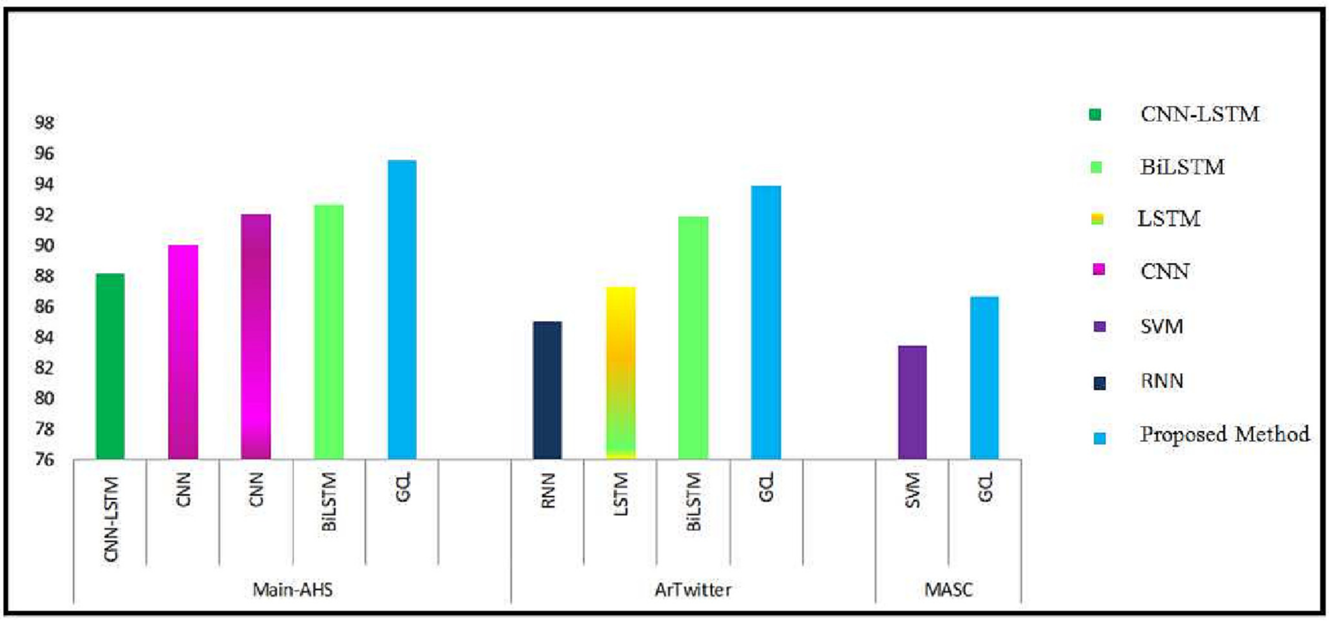

Fig. 8. Models applied to 2C Arabic sentiment datasets (GCL is proposed model).

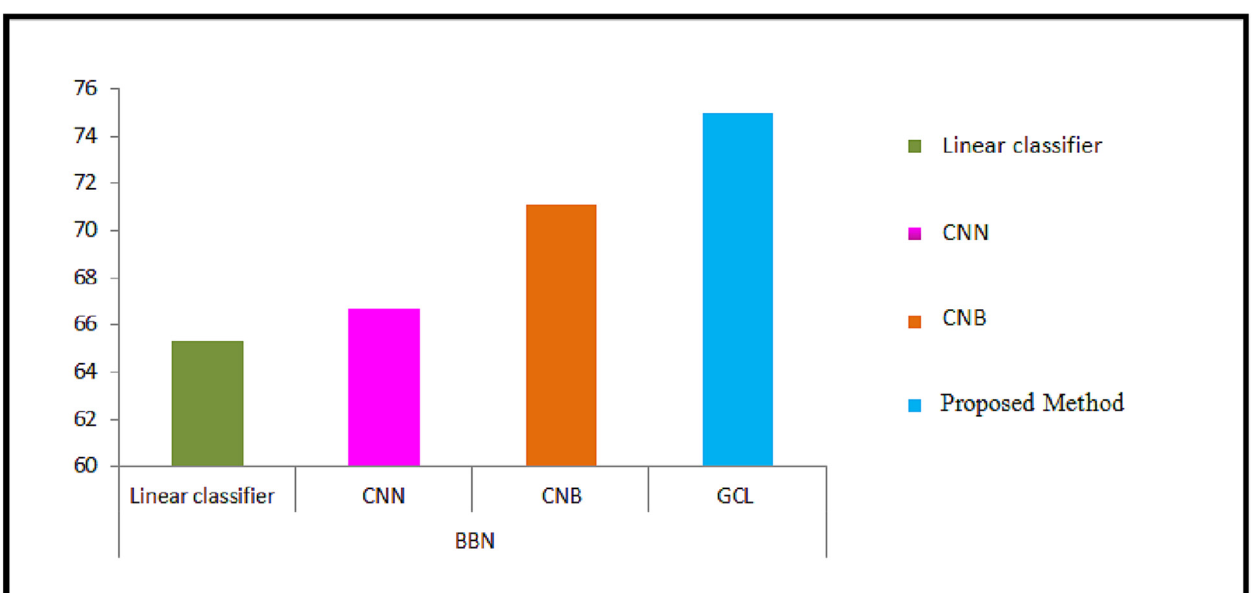

Fig. 9. Models applied to 3C Arabic sentiment datasets (GCL is proposed model).

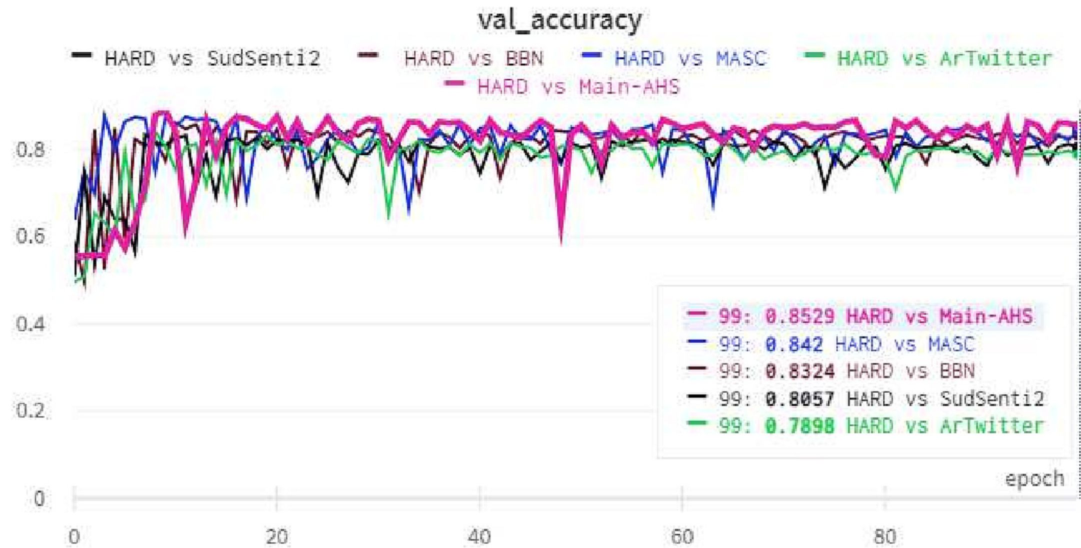

Fig. 10. Validation performance of proposed GCL model on HARD vs. Main-AHS, ArTwitter, MASC, SudSenti2, and BBN.

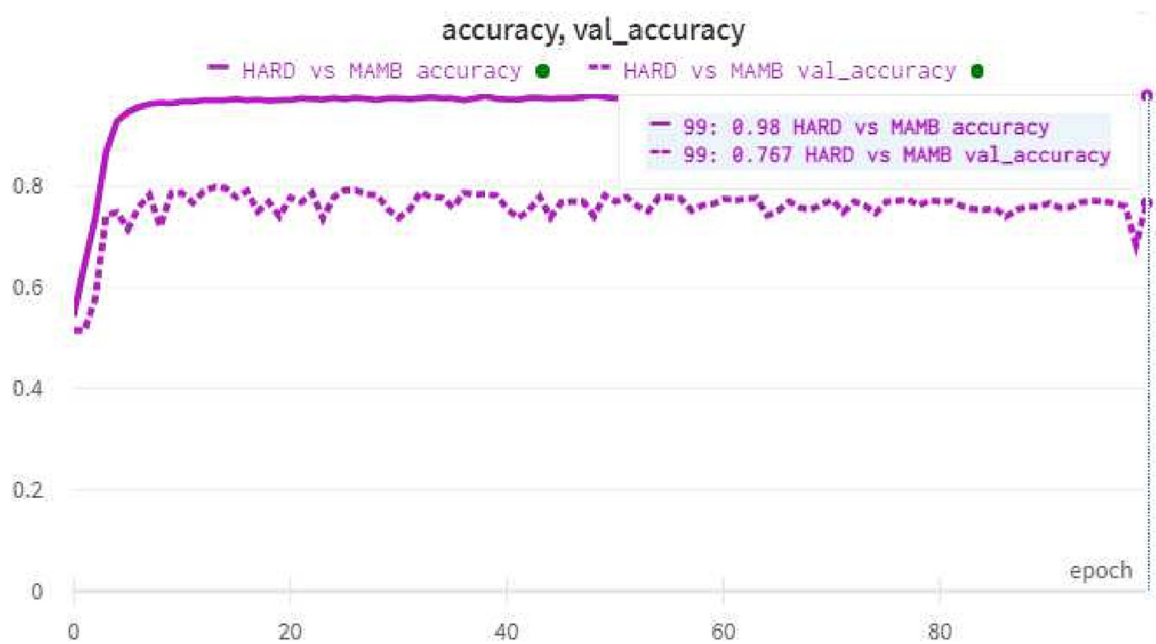

Fig. 11. Accuracy and validation accuracy on HARD vs. MAMB datasets.

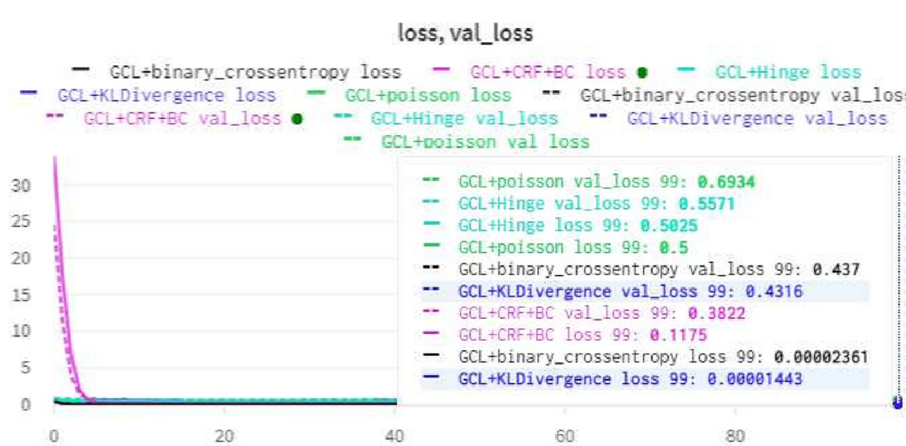

Fig. 12. Loss and validation loss for proposed GCL method on AHSD dataset (2C).

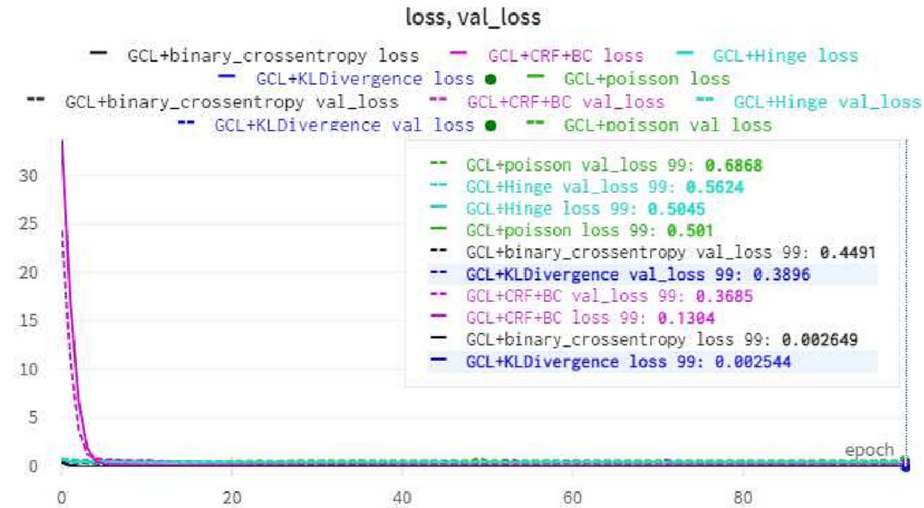

Fig. 13. Loss and validation loss for proposed GCL method on ArTwitter dataset (2C).

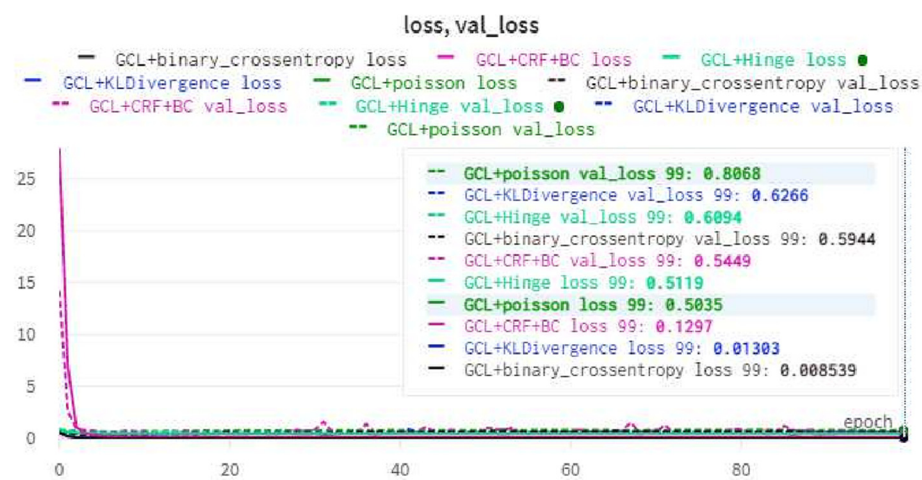

Fig. 14. Loss and validation loss for proposed GCL method on MASC dataset (2C).

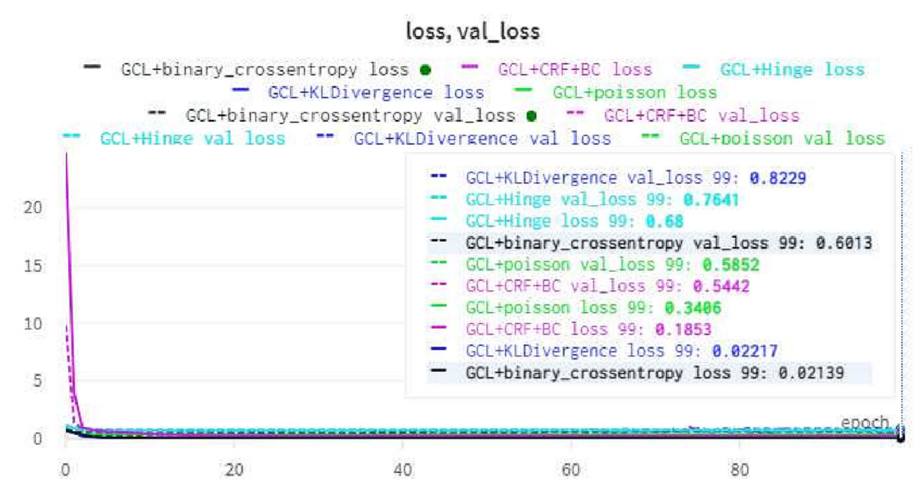

Fig. 15. Loss and validation loss for proposed GCL method on BBN dataset (2C).

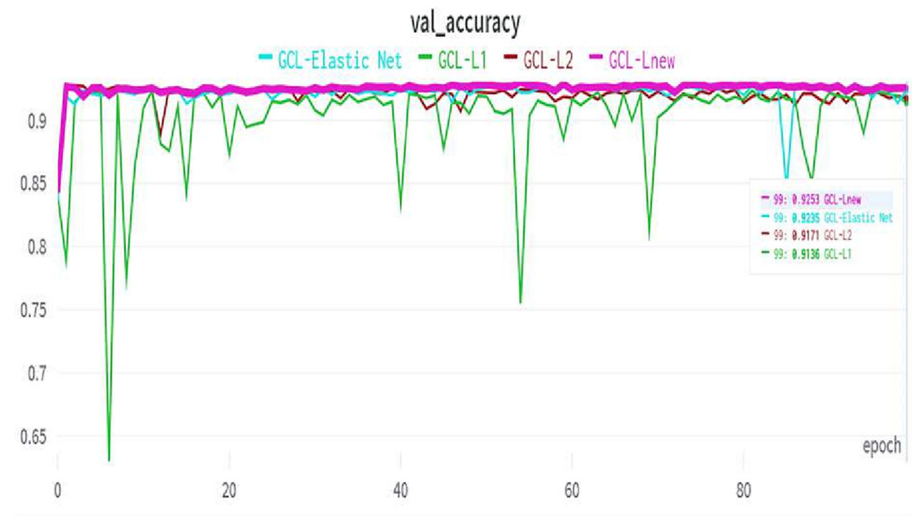

Fig. 16. Validation performance on LABR dataset.

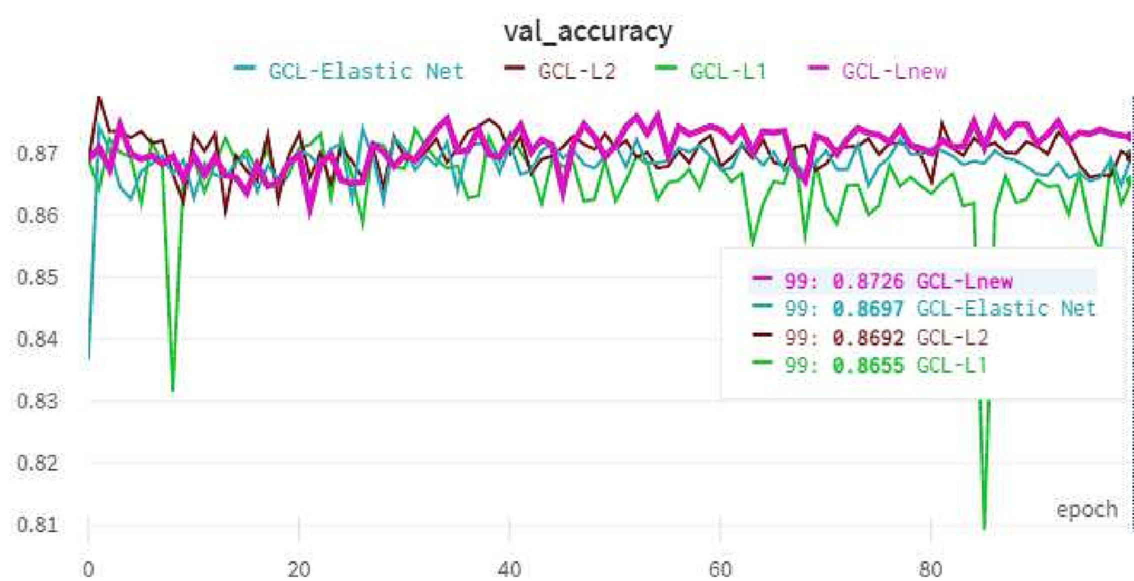

Fig. 17. Validation performance on DzSenti dataset.

cussion on NN models and the effects of polysemy, metaphor and dialects.

4.7. Experiment 5: evaluation of the proposed approach (GCL) with several regularization functions on massive arabic corpora

We used GCL with different regularizations on LABR,[13] the Large-Scale Arabic Book Review, containing 63,000 MSA items, and the DzSenti dataset, which comprises 49,864 items from social media, including both MSA and the ALG dialect (see Table 3).

When we applied GCL+L_1, GCL+L_2, GCL+$L_{ElasticNet}$,[14,15] and GCL +L_{new} on LABR, the accuracy and the times[16] were 91.36% (43 m 30s), 91.71% (1 h 24 m 18s), 92.35% (38 m 55s), and 92.53% (33 m 49s), respectively, as shown in Table 11 and Fig. 16.

On the DzSenti dataset, the performance was 86.55% (1 h 2 m 18s) with GCL+L_1, 86.92% (27 m 3s) with GCL+L_2, 86.97% (21 m 41s) with GCL+$L_{Elastic}$, and 87.26% (20 m 27s) with GCL+L_{new}, as shown in Table 11 and Fig. 17.

First, we note that GCL+L_{new} had the highest performance and used less time with both datasets and exceeded the baseline; for LABR, accuracy was 92.53% compared to 91.9% [97], and for DzSenti it was 87.26% compared to 86.00% [17].

Second, the results show that our models can efficiently handle large Arabic datasets. Also, throughout the training, the accuracy remained consistent.

4.8. Validation loss training

Figs. 8 and 9 show the performance accuracy of the proposed method and baselines on the 2C and 3C datasets. Figs. 10 and 11 show the validation performance of HARD on the individual, and grouped training. Finally, Figs. 12–15 show the loss and validation loss of the GCL model with five standard loss functions, after 100 epochs, on the AHSD, ArTwitter, MASC, and BBN datasets respec-

[13] http://www.mohamedaly.info/datasets/labr.
[14] https://github.com/christianversloot/machine-learning-articles/blob/main/how-to-use-l1-l2-and-elastic-net-regularization-with-keras.md
[15] Elastic Net (L_1+L_2).
[16] (h: hours, m: minutes, and s: seconds).

tively. GCL + (CRF + BC) gives us the best validation loss and the least execution time, when applied to the four datasets.

For the training and validation epochs, the suggested technique was stable. On different datasets, the loss and validation loss were also stable. This demonstrates that the proposed method can be used effectively within training regimes.

5. Conclusion and future work

In this study, we developed an Arabic Sentiment Analysis model called Gated Convolution Long (GCL), based on GRU, CNN, and LSTM. The model incorporates a Customized Regularization Function (CRF).

We then carried out five experiments.

First, GCL was independently trained and tested on four different datasets, AHSD (2C), ArTwitter (2C), MASC (2C), and BBN (3C). The proposed model outperformed the baselines for all datasets.

Second, we created versions of GCL with five different loss functions, Binary-Cross-Entropy, Hinge, Poisson, KL-divergence, and CRF + Binary-Cross-Entropy. These were trained against the same four datasets. CRF + Binary-Cross-Entropy had the lowest validation loss in all cases.

Third, we conducted an ablation investigation using GCL and the same four datasets. We trained both with CRF and without CRF, and tested the resulting model. The results showed that CRF improved the performance of GCL for all datasets.

Fourth, we used the proposed model to analyze the link between emotions in Modern Standard Arabic and those in five distinct Arabic dialects. First, we trained on HARD (MSA) and evaluated on AHSD (SAU), ArTwitter (JOR), MASC (MOR), SudSenti2 (SUD), and BBN (LEV). We found that the most similar dialect to MSA is SAU, and that the least similar dialects are SUD and JOR. Second, we trained on HARD and tested on all dialects together. This showed that a useful level of performance could be obtained in this way, but lower than when training and testing on a specific dialect.

Fifth, we applied GCL using standard regularizations (GCL+L_1, GCL+L_2, and GCL+$L_{ElasticNet}$) and our L_{new} on two big Arabic sentiment datasets, LABR (MSA) and DzSenti (MSA, ALG); GCL+L_{new} gave the highest results with less training time.

Future research will examine the effectiveness of the proposed approaches using a variety of datasets, such as reviews of eateries, technology, the news, and different language-specific archives.

Declaration of Competing Interest

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

Acknowledgement

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).

References

[1] M.A. El-Affendi, K. Alrajhi, A. Hussain, A novel deep learning-based multilevel parallel attention neural (mpan) model for multidomain arabic sentiment analysis, IEEE Access 9 (2021) 7508–7518.
[2] A.R. Pathak, M. Pandey, S. Rautaray, Topic-level sentiment analysis of social media data using deep learning, Appl. Soft Comput. 108 (2021) 107440.
[3] A. Elnagar, O. Einea, L. Lulu, Comparative study of sentiment classification for automated translated latin reviews into arabic, in: 2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA), IEEE, 2017, pp. 443–448.
[4] P. Koratamaddi, K. Wadhwani, M. Gupta, S.G. Sanjeevi, Market sentiment-aware deep reinforcement learning approach for stock portfolio allocation, Eng. Sci. Technol. Int. J. 24 (4) (2021) 848–859.
[5] M.M. Agüero-Torales, J.I.A. Salas, A.G. López-Herrera, Deep learning and multilingual sentiment analysis on social media data: An overview, Appl. Soft Comput. 107373 (2021).
[6] W. Li, L. Zhu, Y. Shi, K. Guo, E. Cambria, User reviews: Sentiment analysis using lexicon integrated two-channel cnn–lstm family models, Appl. Soft Comput. 94 (2020) 106435.
[7] K. Chakraborty, S. Bhatia, S. Bhattacharyya, J. Platos, R. Bag, A.E. Hassanien, Sentiment analysis of covid-19 tweets by deep learning classifiers-a study to show how popularity is affecting accuracy in social media, Appl. Soft Comput. 97 (2020) 106754.
[8] A. Onan, S. Korukoglu, H. Bulut, Lda-based topic modelling in text sentiment classification: An empirical analysis, Int. J. Comput. Linguistics Appl. 7 (1) (2016) 101–119.
[9] A. Onan, Sentiment analysis on twitter based on ensemble of psychological and linguistic feature sets, Balkan J. Electr. Comput. Eng. 6 (2) (2018) 69–77.
[10] A. Diwali, K. Dashtipour, K. Saeedi, M. Gogate, E. Cambria, A. Hussain, Arabic sentiment analysis using dependency-based rules and deep neural networks, Appl. Soft Comput. 127 (2022) 109377.
[11] A.S. Mohammad, M.M. Hammad, A. Sa'ad, A.T. Saja, E. Cambria, Gated recurrent unit with multilingual universal sentence encoder for arabic aspect-based sentiment analysis, Knowl.-Based Syst. 107540 (2021).
[12] A. Alwehaibi, M. Bikdash, M. Albogmi, K. Roy, A study of the performance of embedding methods for arabic short-text sentiment analysis using deep learning approaches, J. King Saud Univ.-Comput. Inf. Sci. 34 (8) (2022) 6140–6149.
[13] M. Al-Ayyoub, A.A. Khamaiseh, Y. Jararweh, M.N. Al-Kabi, A comprehensive survey of arabic sentiment analysis, Inf. Process. Manage. 56 (2) (2019) 320–342.
[14] M. Alassaf, A.M. Qamar, Improving sentiment analysis of arabic tweets by one-way anova, J. King Saud Univ.-Comput. Inf. Sci.
[15] O. Oueslati, E. Cambria, M.B. HajHmida, H. Ounelli, A review of sentiment analysis research in arabic language, Future Gener. Comput. Syst. 112 (2020) 408–430.
[16] Q.A. Xu, V. Chang, C. Jayne, A systematic review of social media-based sentiment analysis: Emerging trends and challenges, Decis. Anal. J. (2022) 100073.
[17] A.B. Nassif, A. Elnagar, I. Shahin, S. Henno, Deep learning for arabic subjective sentiment analysis: Challenges and research opportunities, Appl. Soft Comput. 98 (2021) 106836.
[18] T.H. Alwaneen, A.M. Azmi, H.A. Aboalsamh, E. Cambria, A. Hussain, Arabic question answering system: a survey, Artif. Intell. Rev. 55 (1) (2022) 207–253.
[19] N. Habash, O. Rambow, G.A. Kiraz, Morphological analysis and generation for arabic dialects, in: Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages, 2005, pp. 17–24.
[20] N. Boudad, R. Faizi, R. Thami, R. Chiheb, Sentiment analysis in arabic: a review of the literature, Ain Shams Eng. J. 9 (4) (2018) 2479–2490.
[21] S. Boukil, M. Biniz, F. El Adnani, L. Cherrat, A.E. El Moutaouakkil, Arabic text classification using deep learning technics, Int. J. Grid Distrib. Comput. 11 (9) (2018) 103–114.
[22] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.
[23] L. Zhang, S. Wang, B. Liu, Deep learning for sentiment analysis: A survey, Wiley Interdisciplinary Reviews, Data Min. Knowl. Disc. 8 (4) (2018) e1253.
[24] A.M. Alayba, V. Palade, Leveraging arabic sentiment classification using an enhanced cnn-lstm approach and effective arabic text preparation, J. King Saud Univ.-Comput. Inf. Sci.
[25] B. Brahimi, M. Touahria, A. Tari, Improving sentiment analysis in arabic: A combined approach, J. King Saud Univ.-Comput. Inf. Sci. 33 (10) (2021) 1242–1250.
[26] D. Warde-Farley, I. Goodfellow, 11 adversarial perturbations of deep neural networks, Perturbations, Optimization, and Statistics 311.
[27] I. Goodfellow, Y. Bengio, A. Courville, Deep learning (adaptive computation and machine learning series), Cambridge Massachusetts (2017) 321–359.
[28] A.M. Alayba, V. Palade, M. England, R. Iqbal, Arabic language sentiment analysis on health services, in: 2017 1st international workshop on arabic script analysis and recognition (asar), IEEE, 2017, pp. 114–118.
[29] N.A. Abdulla, N.A. Ahmed, M.A. Shehab, M. Al-Ayyoub, Arabic sentiment analysis: Lexicon-based and corpus-based, in: 2013 IEEE Jordan conference on applied electrical engineering and computing technologies (AEECT), IEEE, 2013, pp. 1–6.
[30] T. Al-Moslmi, M. Albared, A. Al-Shabi, N. Omar, S. Abdullah, Arabic senti-lexicon: Constructing publicly available language resources for arabic sentiment analysis, J. Inf. Sci. 44 (3) (2018) 345–362.
[31] S.M. Mohammad, M. Salameh, S. Kiritchenko, How translation alters sentiment, J. Artif. Intell. Res. 55 (2016) 95–130.
[32] SudSenti, Two large sudanese arabic sentiment datasets, https://github.com/mustafa20999/Sudanese-Arabic-Sentiment-Datasets, accessed: 2022-02-10 (2021).

[33] A. Abdelli, F. Guerrouf, O. Tibermacine, B. Abdelli, Sentiment analysis of arabic algerian dialect using a supervised method, in: 2019 International Conference on Intelligent Systems and Advanced Computing Sciences (ISACS), IEEE, 2019, pp. 1–6.
[34] M. Al-Kabi, I. Alsmadi, R.T. Khasawneh, H. Wahsheh, Evaluating social context in arabic opinion mining., Int. Arab, J. Inf. Technol. 15 (6) (2018) 974–982.
[35] M. Al Omari, M. Al-Hajj, N. Hammami, A. Sabra, Sentiment classifier: Logistic regression for arabic services' reviews in lebanon, in: 2019 international conference on computer and information sciences (iccis), IEEE, 2019, pp. 1–5.
[36] W.M. Yafooz, E. Hizam, W. Alromema, Arabic sentiment analysis on chewing khat leaves using machine learning and ensemble methods, Eng. Technol. Appl. Sci. Res. 11 (2) (2021) 6845–6848.
[37] A. Elnagar, Y.S. Khalifa, A. Einea, Hotel arabic-reviews dataset construction for sentiment analysis applications, in: Intelligent Natural Language Processing: Trends and Applications, Springer, 2018, pp. 35–52.
[38] M. Aly, A. Atiya, Labr: A large scale arabic book reviews dataset, in: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2013, pp. 494–498.
[39] M. Nabil, M. Aly, A. Atiya, Astd: Arabic sentiment tweets dataset, in: Proceedings of the 2015 conference on empirical methods in natural language processing, 2015, pp. 2515–2519.
[40] K. Abu Kwaik, M.K. Saad, S. Chatzikyriakidis, S. Dobnik, Shami: A corpus of levantine arabic dialects, in: Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018), 2018.
[41] R. Baly, A. Khaddaj, H. Hajj, W. El-Hajj, K.B. Shaban, Arsentd-lev: A multi-topic corpus for target-based sentiment analysis in arabic levantine tweets, arXiv preprint arXiv:1906.01830.
[42] A.M. Alayba, V. Palade, M. England, R. Iqbal, A combined cnn and lstm model for arabic sentiment analysis, in: International cross-domain conference for machine learning and knowledge extraction, Springer, 2018, pp. 179–191.
[43] A.M. Alayba, V. Palade, M. England, R. Iqbal, Improving sentiment analysis in arabic using word representation, in: 2018 IEEE 2nd International Workshop on Arabic and Derived Script Analysis and Recognition (ASAR), IEEE, 2018, pp. 13–18.
[44] H. Elfaik et al., Deep bidirectional lstm network learning-based sentiment analysis for arabic text, J. Intell. Syst. 30 (1) (2021) 395–412.
[45] K. Elshakankery, M.F. Ahmed, Hilatsa: A hybrid incremental learning approach for arabic tweets sentiment analysis, Egypt. Inf. J. 20 (3) (2019) 163–171.
[46] S. Al-Azani, E.S.M. El-Alfy, Hybrid deep learning for sentiment polarity determination of arabic microblogs, in: International Conference on Neural Information Processing, Springer, 2017, pp. 491–500.
[47] M. Salameh, S. Mohammad, S. Kiritchenko, 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.
[48] N. Boudad, S. Ezzahid, R. Faizi, R.O.H. Thami, Exploring the use of word embedding and deep learning in arabic sentiment analysis, in: International Conference on Advanced Intelligent Systems for Sustainable Development, Springer, 2019, pp. 243–253.
[49] S.R. El-Beltagy, T. Khalil, A. Halaby, M. Hammad, Combining lexical features and a supervised learning approach for arabic sentiment analysis, in: International Conference on Intelligent Text Processing and Computational Linguistics, Springer, 2016, pp. 307–319.
[50] Y. Tabii, M. Lazaar, M. Al Achhab, N. Enneya, Big Data, Cloud and Applications: Third International Conference, BDCA 2018, Kenitra, Morocco, April 4–5, 2018, Revised Selected Papers, vol. 872, Springer, 2018.
[51] J.O. Atoum, M. Nouman, Sentiment analysis of arabic jordanian dialect tweets, Int. J. Adv. Comput. Sci. Appl. 10 (2) (2019) 256–262.
[52] H. AlSalman, An improved approach for sentiment analysis of arabic tweets in twitter social media, in: 2020 3rd International Conference on Computer Applications & Information Security (ICCAIS), IEEE, 2020, pp. 1–4.
[53] M. Mhamed, R. Sutcliffe, X. Sun, J. Feng, E. Almekhlafi, E.A. Retta, A deep cnn architecture with novel pooling layer applied to two sudanese arabic sentiment datasets, arXiv preprint arXiv:2201.12664.
[54] R. Bensoltane, T. Zaki, Aspect-based sentiment analysis: an overview in the use of arabic language, Artif. Intell. Rev. (2022) 1–39.
[55] S.O. Alhumoud, A.A. Al Wazrah, Arabic sentiment analysis using recurrent neural networks: a review, Artif. Intell. Rev. 55 (1) (2022) 707–748.
[56] A. Al-Hashedi, B. Al-Fuhaidi, A.M. Mohsen, Y. Ali, H.A. Gamal Al-Kaf, W. Al-Sorori, N. Maqtary, Ensemble classifiers for arabic sentiment analysis of social network (twitter data) towards covid-19-related conspiracy theories, Appl. Comput. Intell. Soft Comput. (2022).
[57] A. Al-Laith, M. Shahbaz, H.F. Alaskar, A. Rehmat, Arasencorpus: A semi-supervised approach for sentiment annotation of a large arabic text corpus, Appl. Sci. 11 (5) (2021) 2434.
[58] M. Hadwan, M. Al-Hagery, M. Al-Sarem, F. Saeed, Arabic sentiment analysis of users' opinions of governmental mobile applications, Comput. Mater. Continua 72 (3) (2022) 4675–4689.
[59] M.A. Saloot, N. Idris, R. Mahmud, S. Ja'afar, D. Thorleuchter, A. Gani, Hadith data mining and classification: a comparative analysis, Artif. Intell. Rev. 46 (1) (2016) 113–128.
[60] A.M. El-Halees, Arabic text classification using maximum entropy, IUG J. Nat. Stud. 15 (1).
[61] Q. Ye, Z. Zhang, R. Law, Sentiment classification of online reviews to travel destinations by supervised machine learning approaches, Expert Syst. Appl. 36 (3) (2009) 6527–6535.
[62] S. Rosenthal, N. Farra, P. Nakov, Semeval-2017 task 4: Sentiment analysis in twitter, in: Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017), 2017, pp. 502–518.
[63] D. Wettschereck, T.G. Dietterich, Locally adaptive nearest neighbor algorithms, Adv. Neural Inf. Process. Syst. (1994) 184.
[64] A. Priyam, G. Abhijeeta, A. Rathee, S. Srivastava, Comparative analysis of decision tree classification algorithms, Int. J. Curr. Eng. Technol. 3 (2) (2013) 334–337.
[65] D.W. Hosmer Jr, S. Lemeshow, R.X. Sturdivant, Applied logistic regression, vol. 398, John Wiley & Sons, 2013.
[66] J.D. Rennie, L. Shih, J. Teevan, D.R. Karger, Tackling the poor assumptions of naive bayes text classifiers, in: Proceedings of the 20th international conference on machine learning (ICML-03), 2003, pp. 616–623.
[67] A. Dahou, S. Xiong, J. Zhou, M.H. Haddoud, P. Duan, Word embeddings and convolutional neural network for arabic sentiment classification, in: Proceedings of coling 2016, the 26th international conference on computational linguistics: Technical papers, 2016, pp. 2418–2427.
[68] Y. Zhang, B. Wallace, A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification, arXiv preprint arXiv:1510.03820.
[69] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[70] P.J. Werbos, Backpropagation through time: what it does and how to do it, Proc. IEEE 78 (10) (1990) 1550–1560.
[71] A. Elmadany, H. Mubarak, W. Magdy, Arsas: An arabic speech-act and sentiment corpus of tweets, OSACT 3 (2018) 20.
[72] E. Refaee, V. Rieser, An arabic twitter corpus for subjectivity and sentiment analysis., in: LREC, 2014, pp. 2268–2273.
[73] K. Elshakankery, M.F. Ahmed, Egyptian informatics journal.
[74] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in neural information processing systems, 2013, pp. 3111–3119.
[75] M. Liwicki, A. Graves, S. Fernàndez, H. Bunke, J. Schmidhuber, A novel approach to on-line handwriting recognition based on bidirectional long short-term memory networks, in: Proceedings of the 9th International Conference on Document Analysis and Recognition, ICDAR 2007, 2007.
[76] V. Stoyanov, C. Cardie, J. Wiebe, Multi-perspective question answering using the opqa corpus, in: Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, 2005, pp. 923–930.
[77] H. ElSahar, S.R. El-Beltagy, Building large arabic multi-domain resources for sentiment analysis, in: International Conference on Intelligent Text Processing and Computational Linguistics, Springer, 2015, pp. 23–34.
[78] J. Kukačka, V. Golkov, D. Cremers, Regularization for deep learning: A taxonomy, arXiv preprint arXiv:1710.10686.
[79] M. Mhamed, R. Sutcliffe, X. Sun, J. Feng, E. Almekhlafi, E.A. Retta, Improving arabic sentiment analysis using cnn-based architectures and text preprocessing, Comput. Intell. Neurosci. (2021).
[80] Z. Rahimi, M.M. Homayounpour, The impact of preprocessing on word embedding quality: a comparative study, Language Resour. Eval. (2022) 1–35.
[81] L. Hickman, S. Thapa, L. Tay, M. Cao, P. Srinivasan, Text preprocessing for text mining in organizational research: Review and recommendations, Organizational Res. Methods 25 (1) (2022) 114–146.
[82] M. Hammad, M. Al-awadi, Sentiment analysis for arabic reviews in social networks using machine learning, in: Information technology: new generations, Springer, 2016, pp. 131–139.
[83] C.D. Manning, M. Surdeanu, J. Bauer, J.R. Finkel, S. Bethard, D. McClosky, The stanford corenlp natural language processing toolkit, in: Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations, 2014, pp. 55–60.
[84] N.Q.K. Le, Fertility-gru: identifying fertility-related proteins by incorporating deep-gated recurrent units and original position-specific scoring matrix profiles, J. Proteome Res. 18 (9) (2019) 3503–3511.
[85] A.B. Soliman, K. Eissa, S.R. El-Beltagy, Aravec: A set of arabic word embedding models for use in arabic nlp, Proc. Comput. Sci. 117 (2017) 256–265.
[86] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using rnn encoder-decoder for statistical machine translation, arXiv preprint arXiv:1406.1078.
[87] P. Yann LeCun, P. Haffner, L. Bottou, Object recognition with gradient-based learning, Red Bank NJ: AT&T Shannon Lab.
[88] Y. Yu, X. Si, C. Hu, J. Zhang, A review of recurrent neural networks: Lstm cells and network architectures, Neural Comput. 31 (7) (2019) 1235–1270.
[89] S. Hochreiter, The vanishing gradient problem during learning recurrent neural nets and problem solutions, Int. J. Uncertainty Fuzziness Knowl.-Based Syst. 6 (02) (1998) 107–116.
[90] I. Goodfellow, Y. Bengio, A. Courville, Regularization for deep learning, Deep Learn. (2016) 216–261.
[91] A. Ahrens, C.B. Hansen, M.E. Schaffer, lassopack: Model selection and prediction with regularized regression in stata, Stata J. 20 (1) (2020) 176–235.
[92] R. Zbib, E. Malchiodi, J. Devlin, D. Stallard, S. Matsoukas, R. Schwartz, J. Makhoul, O. Zaidan, C. Callison-Burch, Machine translation of arabic dialects, in: Proceedings of the 2012 conference of the north american chapter of the association for computational linguistics: Human language technologies, 2012, pp. 49–59.

[93] F. Rustam, M. Khalid, W. Aslam, V. Rupapara, A. Mehmood, G.S. Choi, A performance comparison of supervised machine learning models for covid-19 tweets sentiment analysis, Plos one 16 (2) (2021) e0245909.
[94] T. Srivastava, Important model evaluation metrics for machine learning everyone should know, Commonly Used Machine Learning Algorithms: Data Science 2020.
[95] R. Rawat, V. Mahor, S. Chirgaiya, R.N. Shaw, A. Ghosh, Sentiment analysis at online social network for cyber-malicious post reviews using machine learning techniques, Computationally intelligent systems and their applications (2021) 113–130.
[96] J. Thomas, S. McDonald, A. Noel-Storr, I. Shemilt, J. Elliott, C. Mavergames, I.J. Marshall, Machine learning reduced workload with minimal risk of missing studies: development and evaluation of a randomized controlled trial classifier for cochrane reviews, J. Clin. Epidemiol. 133 (2021) 140–151.
[97] A. Barhoumi, N. Camelin, C. Aloulou, Y. Estève, L. Hadrich Belguith, An empirical evaluation of arabic-specific embeddings for sentiment analysis, in: International Conference on Arabic Language Processing, Springer, 2019, pp. 34–48.