Applied Soft Computing 107 (2021) 107373

Contents lists available at ScienceDirect

Applied Soft Computing

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

Deep learning and multilingual sentiment analysis on social media data: An overview

Marvin M. Agüero-Torales [a,*], José I. Abreu Salas [b], Antonio G. López-Herrera [a]

[a] Department of Computer Science and Artificial Intelligence, University of Granada, Calle Daniel Saucedo Aranda, s/n, 18071, Granada, Spain
[b] University Institute for Computing Research, University of Alicante, Carretera de San Vicente del Raspeig s/n, Alicante, Valencia, Spain

ARTICLE INFO

Article history:
Received 12 June 2020
Received in revised form 24 February 2021
Accepted 25 March 2021
Available online 1 April 2021

MSC:
68-02
68T50
68T07
91D30

Keywords:
Sentiment analysis
Multilingual
Cross-lingual
Code-switching
Deep learning
Natural language processing (NLP)
Social media

ABSTRACT

Twenty-four studies on twenty-three distinct languages and eleven social media illustrate the steady interest in deep learning approaches for multilingual sentiment analysis of social media. We improve over previous reviews with wider coverage from 2017 to 2020 as well as a study focused on the underlying ideas and commonalities behind the different solutions to achieve multilingual sentiment analysis. Interesting findings of our research are (i) the shift of research interest to cross-lingual and code-switching approaches, (ii) the apparent stagnation of the less complex architectures derived from a backbone featuring an embedding layer, a feature extractor based on a single CNN or LSTM and a classifier, (iii) the lack of approaches tackling multilingual aspect-based sentiment analysis through deep learning, and, surprisingly, (iv) the lack of more complex architectures such as the transformers-based, despite results suggest the more difficult tasks requires more elaborated architectures.

1. Introduction

Sentiment Analysis (SA) allows us to automatically evaluate the opinion of people toward products, services, and other entities. This knowledge can help to make better decisions looking to improve key performance indicators. Besides, the massive adoption of social media such as Facebook and Twitter, platforms for e-commerce and services like Amazon, and even review-specialized sites such as Rotten Tomatoes, unleashed a vast amount of content to be analyzed. This data is naturally multilingual and multicultural thus, an analysis based on a single language may carry the risk of not capturing the overall insights [1]. Moreover, important challenges can prevent fully leverage this data. Except for a few cases, e.g., English, most languages lack well-maintained resources widely used for SA such as annotated corpus and lexicons. Second, it could be not straightforward to adapt the same SA model to different languages, for example, due to variations in word order or usage, or the noise introduced by machine translation. Also, we have code-switching content, where users express their opinions using a mixture of languages in the same sentence.

Multilingual Sentiment Analysis (MSA) is an attempt to address those issues through several strategies. For example, taking advantage of resource-rich languages to perform SA in a resource-poor language as characteristic in cross-lingual sentiment analysis. Also, developing language-independent models capable to handle SA in different languages or a code-switching setup.

There is a wide spectrum of approaches for SA, for example, [2–5], which can be relied on supervised but also in unsupervised methods that exploit sentiment lexicons, grammatical analysis, and syntactic patterns. In Sections 2.1 and 2.3 we include a panoramic of the different formulations of this task as well as the evolution of SA and MSA. More recently, deep learning (DL) approaches have become a trend leading to state-of-the-art results, with authors such as [6–8] exploring Convolutional Neural Networks, Adversarial Networks, and Recurrent Neural Networks among other models. In Section 2.2 we resume some of the advancements of deep learning for SA as an introduction for the main topic of this work, the applications of deep learning in multilingual sentiment analysis in social media.

* Corresponding author.
E-mail addresses: maguero@correo.ugr.es (M.M. Agüero-Torales), ji.abreu@ua.es (J.I. Abreu Salas), lopez-herrera@decsai.ugr.es (A.G. López-Herrera).

https://doi.org/10.1016/j.asoc.2021.107373
1568-4946/

Using the methodology detailed in Section 3 as a guideline, we curated and reviewed 24 relevant research papers. We categorized them as regard the main idea in multilingual, cross-lingual or code-switching approaches, covered in 4.1, 4.2 and 4.3 respectively. For each one, we discuss its distinctive contributions, the experimental setup, corpus, and main results. Also, Section 4.4 includes a comparative that allows a quick and broad view of the advances in the domain. This analysis drew interesting conclusions such as the few works, to date, leveraging recent developments in contextual embedding. Other main findings and conclusions are covered in Sections 5 and 6 .

As sentiment analysis and deep learning approaches have been growing as an important research field, there have been early efforts [1,9–13] to systematize the knowledge corpus in this domain, works extended lately by [14–16]. Recently, [17] examined the fundamentals of the multilingual case. However, more than seventeen works we identified introducing or exploring specific ideas for MSA have not been studied by the previous reviews. Another of our main contributions is to drive the review by the underlying hypothesis of each work, not only analyzing them as regards the type of neural network they used. This is important since the same task can be tackled by very different ideas. Also, we focused the analysis on the current three major strategies for MSA: multilingual, cross-lingual, and code-switching. This high-level view of the domain can help to unveil interesting patterns more than the type of neural network implemented. For example, the use of adversarial training to learn language-agnostic features.

2. Preliminaries

In this section, we cover the fundamental concepts in Sentiment Analysis and Multilingual Sentiment Analysis. Also, some of the antecedents about the applications of Deep Learning to this problem.

2.1. Sentiment analysis on social media

Starting from Wiebe et al. [18] work in the late 90s, there has been a surge of interest in the different setups of SA. In general, it can be done at a document, sentence, or aspect level [5] and the classification in terms of positive, negative, or neutral, but also other more fine-grained scales such as a ranking from 1 to 5. This attention over SA is closely tied to Social Media in its key role in the rise of modern SA particularly with the works of Pang et al. [19], and Turney [20], in 2002. The first used machine learning (ML) classification techniques over movie review data outperforming human-produced baselines. The second achieved an average accuracy of 74% for his recommendations based on online reviews, which used Semantic Orientation (SO) applied to unsupervised classification. Later, Pang and Lee (2008) [2], focused on the fundamentals and basic applications of SA, with a list of resources such as lexicons or datasets.

A comprehensive review that shows the maturity of SA up to 2012 can be found in the book of Liu [4]. This work covers most of the topics, definitions, research problems (e.g., opinion spam detection), types of opinions (such as explicit and implicit opinions), and classification algorithms for SA. In 2013 Feldman [21] and Cambria et al. [22] wrote about the basic techniques, key tasks, and applications as well as the evolution of the field.

Another source to take the pulse of the continuous advances in the field has been the tasks related to SA in Twitter hosted by the International Workshop on Semantic Evaluation (SemEval) from 2013 to 2017 and in 2020 for code-switching text. From the latest results, we can corroborate a shift toward the application of deep learning with 20 out of 48 systems participating in SemEval 2017 [23]. In the next section, we overview some of the recent advances in DL applied to SA without considering the multilingual task.

2.2. Deep learning on sentiment analysis task

Deep Sentiment Analysis (DSA) relies on the great potentials of DL showed for NLP tasks. Here, we briefly commented on some examples to illustrate how DL has been leveraged within the SA.

Word embeddings are used for language modeling and feature learning. They are commonly used as an input of the DL models, being Word2Vec [24] and GloVe [25] two frequently used approaches. Also, there are contextualized embeddings such as ELMo [26], which represent better the polysemy of the words. Besides using pre-trained embeddings, they can be learned to encode some specific task semantics. In the context of SA, this approach has been explored in works such as [27] and [28].

Another trending field within DL is the attention mechanism, which allows the model to non-uniformly weigh the contribution of the context when computing the output. This is another choice that is being used frequently in SA, for example, to capture the interaction between aspects and their context as in [29,30].

Also, there has been a great interest in novelty architectures for SA. One approach that has been received considerable attention when working at the document level, is the design of hierarchical models which learn a representation for sentences from its words, on top of this level, another model can learn representations for documents. Different alternatives such as Convolutional Neural Networks (CNN) or Long Short-Term Memory (LSTM) can be used at each level. Works in [31,32] and [33] are some examples of this approach.

As the last additional example of the ideas that have been explored, not only for monolingual SA but also for MSA we mention the use of adversarial learning to produce a set of domain-independent features. This is the hypothesis of works such as [34] and [35] for cross-domain SA.

The concepts discussed in this section do not exhaust the applications of DL to SA, a more complete revision can be found in [36].

2.3. Multilingual sentiment analysis

In this section, we introduce the fundamentals of multilingual sentiment analysis as well as some of the earlier applications of DL to MSA. Initially, the applications of SA have been developed basically for one language, English in most cases, but the multilingual nature of Social Media has shifted the field to a multilingual analysis. Also, advances in SA backed by DL have made it possible to include low resource languages and avoid the use of translation tools.

A frequent approach for MSA is called Cross-Language (also Cross-Lingual) Sentiment Classification [37] which relies on machine translation [38,39]. For example, [40] reported an improvement over non-DL classifiers (SVM — Support Vector Machine) translating from other languages (Hindi, Marathi, Russian, Dutch, French, German, Portuguese, Spanish, and Italian) to English to use augmented word embeddings together with a CNN model. However, the cross-language approach carries several issues and weaknesses, for example, notable discrepancies in the data distribution, potential cultural distances even in a perfect translation, hard and costly translation tasks for large corpora with issues as charges, availability, performance [37].

Another usual setup for SA is the code-switching one, also called code-mixing or code-mixed. In this case, the content to be analyzed is expressed alternating two or more languages even in a single sentence. One early approach that takes on this problem with DL is Wang et al. [41] for Chinese–English. They proposed a bilingual attention LSTM to perform SA in a corpus from Weibo.com capturing the informative words from both the bilingual and monolingual contexts. Code-mixing is frequent in social media sites such as Facebook and Twitter in countries with a large part of its population speaking more than one language, for example in India where several official and non-official languages are used. This is an active research area for Hindi–English. An early proposal [42] is based in MLP (Multi-Layer Perceptron) with word-level features for Hindi–English and Bengali–English Facebook posts. A more detailed summary of SA for Indian languages with a special focus on the code-mixed text can be found in [43].

There has been a lot of interest to systematize the advances in MSA. For example, [44] and [45], the latter reviewed the principal directions of research focusing on the development of resources and tools for multilingual subjectivity and SA and addressed both multilingual and cross-lingual methods. Singhal & Bhattacharyya (2016) [11] described some of the different approaches used in SA research. Lo et al. (2017) [1] revised various of the main approaches and tools used for MSA at the time. They identified challenges and provided several recommendations with a framework for dealing with scarce resource languages. Also, in [13] and [46] we can find revisions of the field, however, they did not delve into the use of DL in MSA.

3. Methodology

Our research methodology comprises four steps to realize our main goal: to identify the underlying ideas and commonalities behind the different solutions to achieve multilingual sentiment analysis as well as to suggest future research directions.

(i) Define the research scope: the applications of deep learning to multilingual sentiment analysis on social media from January 2017 to December 2020. We choose this timeframe because the shift toward the application of DL-based sentiment analysis happened in 2017 as showed in [23].

(ii) Article search: our search terms were, (a) deep learning AND (b) sentiment analysis AND (c.1) multilingual OR multi-language OR multilanguage; (c.2) crosslingual OR cross-lingual OR cross-language OR crosslanguage; (c.3) code-mixed OR codemixed OR code-mixing OR codemixing OR code-switching OR codeswitching; (c.4) bilingual OR bi-lingual. Search queries were run in Scopus and the Web of Science.

(iii) Article verification: the search yielded 96 studies that were examined to ensure they satisfy the following criteria. (a) must handle explicitly the multilingualism either by (a.1) training with one or more languages and evaluating with a different one or others, (a.2) train with a multilingual corpus, i.e., the same model sees text in different languages during training regardless of this being at different steps. Thus, we excluded works that separately trained and evaluated the same architecture in different languages, i.e., created one model for each language trained only with data from the given language. Candidates for deletion were verified by the three authors.

We also revised the citations from the selected studies as well those referenced by previous reviews [1,12,14–17,46–49] to identify possible candidates, applying the filters (a.1, a.2). In total, 24 publications were eligible for review.

(iv) Research analysis: for each of the selected papers we extracted data and information about (a) research characteristics, as authors, year of publication, languages covered, methodology, corpus characteristics; (b) sentiment level and categorization (binary, ternary, or fine-grained, e.g., rates $0-5$); (c) deep learning architectures and techniques, and (d) results and effectiveness of the proposal against baselines or state-of-the-art models. We also reviewed each work to identify the underlying idea to handle multilingualism, with a focus on the results that assessed the hypothesis.

4. Deep learning techniques for multilingual sentiment analysis on social media

In this section, we review a large corpus of research related to the applications of deep learning to multilingual sentiment analysis. Instead of driving the analysis by the type of architecture or techniques, we choose to organize the works by its sub-domain within MSA, i.e., multilingual, cross-lingual, or code-switching approaches. Inward each category, we proceeded chronologically to track the evolution of the field but separating the aspect-based studies since in general, they lead to very specific architectures. Also, we aim to provide a high-level perspective considering the underlying hypothesis of each work. Finally, the epigraph 4.4 aids the reader to take a glimpse of the domain as regards the models, baselines, corpus, core ideas, and languages covered. For clarity, and due to the variety of datasets and languages covered by each work, Table 1 illustrates the corpus and reports the number of tweets/sentences/documents used.

4.1. Multilingual approaches

This category groups a large set of works that aims to be language-agnostic to those seen during the training. Common goals are the design of systems capable to learn directly from multilingual unpaired content and providing predictions regardless of the source language. Across the analysis, we will use acronyms or abbreviations of common domain concepts without their definition for the sake of space.

4.1.1. Sentence-based studies

Training the same model for different languages is explored by [50]. They fit a multilayered CNN in two phases. First, they learn word embeddings from a large corpus of $300M$[1] unlabeled tweets in English, Italian, French, and German. The parameters are optimized further during the second stage, trying to infer weak labels inferred from emoticons. Finally, they fine-tuned the model using a corpus of annotated tweets. Experiments evaluated a model (FML-CNN) trained in all languages at once, a model fitted in a single language (SL-CNN), and other variations. The results showed that FML-CNN reaches slightly worse performance, about 2.45% lower F1 score, compared to SL-CNN (67.79% for Italian). However, experiments suggest that FML-CNN can handle better for code-mixed text.

Another hypothesis is to exploit character level embeddings to achieve language independence. In [37] and [51] authors describe language-agnostic translation-free architectures (Conv-Char-S, Conv-Char-R) for Twitter based on a CNN that can be trained in several languages at once. They evaluated their approach using tweets in English, Portuguese, Spanish, and German from the corpus in [52] achieving an F1 score above 72.2% [51] for the multilanguage setup. The slightly worst results for some baselines such as LSTM-Emb [53] can be a trade-off since the models have ≈ 90 times fewer parameters and use ≈ 4 times less memory.

The idea of multilanguage character embeddings is explored also by [54] but mapping each character to its UTF-8 integer code. The architecture (UniCNN) is similar to [37,51], placing a CNN after the embedding layer, with a fully connected classifier at the top. They used a subset of the Twitter corpus in [52]. The UniCNN achieved accuracy $\geq 75.45\%$ for all languages. Moreover, except for English, they outperformed models that require translation or/and tokenization such as TransCNN (Word), a similar architecture that operates at word level and translated text (79.57% for English).

[1] M: Million.

Table 1
Multilingual approaches used in DSA.

Proposed model*	Baseline model*	Proposed approach	C[1]	L[2]	Language	Corpus	Reference
CNN	translate-single-lang-CNN, multi-lang-CNN, multi-lang-without-identification-CNN, Random Forest	Trained large amounts of data in various languages (is trained for every single language), in three phases: unsupervised, distant supervised, and supervised with multi-layer CNN	T	D & M	English, Italian, French, & German	Unlabeled tweets ($300M$), weakly-supervised data ($40-60M$), and annotated tweets ($71K$)	Deriu et al. (2017) [50]
CNN	**LSTM-Emb**, Conv-Emb, Conv-Emb-Freeze, Conv-Char & SVM	Cost-effective Character-based embedding and optimized convolutions	B	D	English, German, Portuguese & Spanish	Annotated tweets ($128K$, subset of $1.6M$)	Becker et al. (2017) [51]
CNN	LSTM-Emb, Conv-Emb, Conv-Emb-Freeze, Conv-Char & SVM	Character-level embeddings with few learnable parameters	B	D & M	English, German, Portuguese & Spanish	Annotated tweets ($128K$, subset of $1.6M$)	Wehrmann et al. (2017) [37]
CNN	word-Translation-CNN, char-Translation-CNN, 1-gram-SVM, 2-gram-SVM	Transformed characters into numbers corresponding UTF-8 decimal codes	B	D	English, Polish, German, Slovak, Slovenian & Swedish	Annotated tweets ($150K$, subset of $1.6M$)	Zhang et al. (2017) [54]
CNN	**SVM**	N-gram bilingual mixed (English–French) input text source (based on a Naïve approach)	B	D & M	French, English & Greek	Labeled restaurant balanced reviews (62.6 K)	Medrouk & Pappa (2017) [55]
CNN	word-CNN, char-CNN, unicode-char-CNN, CuDNNLSTM (word and char)	Word-level & Character-level embeddings with two convolutional channels (one channel for each)	B	D	English, German, Portuguese, Spanish, Polish, Slovak, Slovenian & Swedish	Annotated tweets ($193K$, subset of $1.6M$)	Zhang et al. (2017) [56]
BiLSTM	Average Skip-gram Vectors with LR & CNN-Subword-char-LSTM	Shared parameters of twin (siamese) networks with contrastive learning	T	M	English, Hindi	English annotated tweets (114K) & Hindi–English labeled sentences of Facebook posts ($3.8K$)	Choudhary et al. (2018) [57]
CNN	random initialized CNN, CNN + GloVe/FastText/Polyglot embeddings, regular CNN concatenate standard GloVe/FastText + multilingual sentiment embeddings (VADER, SocialSent or Amazon reviews), dual-channel-CNN + GloVe/FastText incorporates random initialized, Polyglot, VADER, SocialSent or static Amazon reviews embedding	Cross-lingual graph-based propagation (transfer-learning) from a rich source language with embeddings of supervised training on Amazon reviews to a dual-channel neural architecture	B	D	English, Spanish, Dutch, German, Russian, Italian, Czech, Japanese & French	Annotated movie reviews (12.2K, Rotten Tomatoes and AlloCine), labeled reviews ($20.8K$, TripAdvisor, and Amazon Fine Food) & labeled tweets ($4.8K$)	Dong & De Melo (2018) [58]
CNN	LSTM (one-layer and two-layer), BiLSTM (one-layer and two-layer), CNN-LSTM, NB & SVM	Dictionaries of character and word indexes to produce code-mixed character and word embedding for a single NN	T	M	Bambara & French (mixed)	Labeled Facebook comments ($17K$, subset of $74K$)	Konate & Du (2018) [59]
MNB + LSTM	Subword-LSTM	Ensemble of Multinomial Naïve Bayes with 1 and 2-gram features and many-to-one stacked LSTM over 3-gram encoding of sentences	T	M	English, Hindi	Hindi–English labeled sentences from Facebook posts (3.8K)	Jhanwar & Das (2018) [60]
LSTM	**CNN**	N-gram raw corpus-based input, without any preprocessing, translation, annotation nor additional knowledge features	B	D & M	French, English & Greek	Labeled restaurant and hotel balanced reviews (91.8K)	Medrouk & Pappa (2018) [61]
BiLSTM	Doc2Vec + SVM, FastText & CNN	Embedding with only distributed representation of the text	T	M	English, Bengali, Hindi & Kannada (English mixed)	Labeled sentences from Facebook comments ($22.5K$)	Shalini et al. (2018) [62]
BiLSTM	**SVM**	Learning new word embeddings based on limited training datasets and a pre-trained DNN exploiting transfer-learning from a rich source language with labeled data	B	D	English & Greek	Labeled TripAdvisor reviews ($40K$) & annotated tweets (480)	Stavridis et al. (2018) [63]
CNN + GAN + Attention-mechanism	SVM + Word2Vec, LSTM, CNN, NSC + UPA, UPNN	Combined CNN, GAN, and user attention to learn specific and independent-language features from data with authorship information	B	D	English & Chinese	Annotated tweets ($48.1K$) & Weibo posts ($53.6K$)	Wang et al. (2018) [64]
GAN + DAN	DAN, mSDA, Machine Translation + DAN, CLD-KCNN, CLDFA-KCNN	Combined Bilingual Embedding (BWE), Deep Averaging Network (DAN) and adversarial training to learn independent-language features from a source language (English)	T, F	D	English, Chinese & Arab	English reviews from Yelp (700K), hotels reviews in Chinese ($20K$ labeled / $150K$ unlabeled) & tweets in Arab ($1.2K$ labeled)	Chen et al. (2018) [65]
BiLSTM-CNN	CNN-hierarchical-BiLSTM, CNN-hierarchical-BiLSTM-gate-mechanism, LSTM, CNN, LSTM-attention-mechanism, CNN-attention-mechanism, RCNN-LSTM, hierarchical-LSTM, LSTM-sentences-relations	Word vector representation improvement based on gate mechanism, which obtains time-series relationship of different sentences in the comments through an RCNN, and gets the local features of the specific aspects in the sentence and the long-distance dependence in the whole comment through a hierarchical attention BiLSTM	B & T	D	English, Arabic, French & Chinese	Binary (400) and ternary ($3.7K$) labeled web reviews ($4.1K$)	Liu et al. (2019) [66]
BiLSTM-CNN	1-grams + 2-grams-SVM, 1-grams + 2-grams-NB-SVM, 1-grams + 2-grams-MNB, Tf–Idf-MNB, Lexicon Lookup, Char-LSTM, Subword-LSTM, FastText & SACMT	Hybrid architecture with subword level representations for the sentences, two parallel BiLSTM as Dual Encoder (Collective Encoder for overall sentiment and Specific Encoder with attention mechanism for subwords) and linguistic features network	T	M	English, Hindi	Hindi–English labeled sentences of Facebook posts ($3.8K$)	Lal et al. (2019) [67]

(continued on next page)

Table 1 (*continued*).

Proposed model*	Baseline model*	Proposed approach	C[1]	L[2]	Language	Corpus	Reference
BiLSTM	SVM-MONO, SVM-MT, ARTEXTE-SVM-based, ARTEXTE-Ensemble, BARISTA-SVM-based, BARISTA-Ensemble, BLSE, BLSE-Ensemble & BiLSTM-MT	Low-resource language embeddings + mapping function, joined with rich-resource language embedding through k-NN refinement, BiLSTM as encoder layer, then fully connected layer with softmax for prediction	F - 1	D	English, Spanish & Catalan	English labeled tweets (33.7*K*), Spanish OpeNER (1.3*K*) & Catalan MultiBooked (1*K*)	Jabreel et al. (2019) [68]
Double LSTM	Double LSTM with several combinations of optimizers and loss functions & Subword-LSTM	Low resource + code-mixed corpus to train embeddings. Joint feature of sentences (subword + word levels), preceded by Double LSTM layer	T	M	English & Hindi (mixed)	Hindi–English labeled sentences of Facebook posts (3.8*K*)	Mukherjee (2019) [69]
LSTM-AAE-BiGRU	MT-SVM, MT-BiGRU & TL-BiGRU	Contextual word embeddings (Word2Vec+LSTM, source and target languages), AAE B	D	English, Chinese & German	Amazon labeled documents (28.9*K*) and unlabeled documents (80*K*) for each pair of language category (books, DVD, music)	Shen et al. (2020) [70]	
BiLSTM + Attention-mechanism	NB, SMO (SVM), RF, BiLSTM-CNN, Double BiLSTM, GloVe + BiLSTM-CNN, GloVe + Double BiLSTM, GloVe + Attention-based-BiLSTM, $\text{BERT}_{base-uncased}$ & $\text{BERT}_{base-domain-uncased}$	Attention mechanism to extract such words that are important to the meaning of the sentence and aggregate the representation of those informative words to form the sentence vector; a sigmoid layer is used to predict the correct label	T	M	English, Hindi & Bengali (English mixed)	English, Bengali–English, Hindi–English annotated tweets (9.2*K*, 5.5*K*, 18.4*K*) & Hindi–English labeled sentences of Facebook posts (3.8*K*)	Jamatia et al. (2020) [71]
BiLSTM	LASER-CNN, FastText-BiLSTM & fastText-CNN	Transfer learning by LASER with (low-resource) language corpus, BiLSTM, then predict the sentiment of texts in other (high-resource) language	F + 1	D	Polish, Dutch, English, French, German, Italian, Portuguese, Russian & Spanish	Online medicine, hotels, school, products reviews (8.4K for each language)	Kanclerz et al. (2020) [72]
CNN-BiLSTM	NB, BiLSTM & Subword-LSTM	Three stages classification with subword embeddings + CNN-BILSTM: first positive or not, then negative or not, and then, computed classification matrix of them	T	M	English & Kannada (mixed)	Annotated YouTube comments (10.4K)	Chundi et al. (2020) [73]
LSTM-CNN	BilBOWA + CNN, VecMap + CNN, BilBOWA + LSTM, VecMap + LSTM, BilBOWA + CNN-LSTM, VecMap + CNN-LSTM & BilBOWA + LSTM-CNN	Train bilingual embeddings (VecMap, on one high-resource and other low-resource language) and uses it on target language (low-resource), followed by a DL classifier for predict polarity	B	D	English & Persian	Binary (11*K*, Persian Digikala reviews), five categories (200*K*, English Amazon reviews)	Ghasemi et al. (2020) [74]

*__Bold__: model with best performance.

[1]Classification (**C**): **B** (Binary), **T** (Ternary), **F** (Five categories).

[2]Multilingualism Level (**L**): Document (**D**), Mix (**M**).

Multilanguage character embeddings are further developed in [56] but within an architecture (Word-Character CNN) that processes the text through two parallel CNN, one for words and the other for characters. The hypothesis is that words and character features provide complementary information. Outputs from both CNN are merged before being feed to a fully connected classifier. To achieve language independence, the embedding layer is kept trainable. They used the same Twitter corpus as in [54]. The hybrid model yields a better performance ($\geq$ 77.13%) compared to pure word/character CNN such as [8] ($\geq$ 74.64%), [37] ($\geq$ 75.41%) and their former model UniCNNs [54] ($\geq$ 75.45%) for languages already studied in [54]. Interestingly, the two romance languages considered, Spanish (69.82%) and Portuguese (72.87%), had the worst performance.

Medrouk & Pappa [55] studied a similar architecture. It comprises a stack of CNN working as a feature extractor, i.e., an encoder, followed by polling and a fully connected predictor, but in this case, working at the n-gram level. To this point, CNN seems a popular choice within the domain in contrast to LSTM. The model is feed with reviews written in French, English, and Greek without any language indication. Empirical evaluation over a mix of contents from three languages yielded an F1 score of 88%. These results reinforce the assumption that the n-gram CNN can produce language-independent features capturing the local relations between words useful for multilingual polarity analysis.

Whether CNN and LSTM variants of the embedding-feature extractor–classifier architecture need extra pre-training hassle or additional complexity to handle multilingual data is investigated by [61]. Experiments were conducted training monolingual and multilingual models, achieving accuracy over 90% for both types of networks working at the n-gram level. Moreover, the fact that the multilingual models behave as well as the monolingual ones, seems to confirm the hypothesis about their ability to extract rich features without distinction if processing single or multilingual datasets.

4.1.2. Aspect-based studies

Regardless of the promising results for SA at sentence level that achieves simpler architectures such as [61], it is not a surprise that for aspect level authors proposed more complex models.

The architecture (GRCNN-HBLSTM) proposed by [66] combines two word-level feature extractors. A BiLSTM encoding sentences that take as inputs the embeddings for the topics, the aspects, and the words. The original word embedding and a feature vector from a character CNN are combined through a gate mechanism to achieve language independence. The second encoder is a regional CNN that aims to preserve the temporal relationship between different sentences, also capturing some of the long-distance dependencies of the aspects. Both feature sets are feed to a sentence-level BiLSTM together with an attention mechanism. A softmax classifier handles the output of the last layer. In experiments using a subset of the dataset in [75] their full model yields an F1 score above 78.04% in all cases outperforming baselines such as the Hierarchical LSTM [31] ($\geq$ 78.04%). What is more important, they compared a version (CNN-HBLSTM) without the gate mechanism that achieves worsts results ($\geq$ 74.66%) and the highest variance among languages.

So far, we have reviewed the purely multilingual approaches for SA. We have contrasted very different approaches. However, at the sentence level, the common strategy is to learn features from a multilanguage set using CNN and feed a classifier module with those features. Unsurprisingly, for the aspect SA setup, authors embrace more complex architectures leveraging attention mechanisms and aspect embeddings. The next section is devoted to the cross-lingual category.

4.2. Cross-lingual approaches

We categorized as cross-lingual the proposals where the focus is to leverage resource-rich language assets to extrapolate to a low-resource target, for example, through transfer learning. This is the core of the proposal in [58] for the cross-lingual projection of sentiment embeddings. Their custom architecture (Dual-Channel CNN) has one channel which works with word embeddings to extract features through convolutions. The other channel is similar but uses word sentiment embeddings which can boost the classification. Features computed from each channel are merged before being feed to a fully connected layer. It is worth noting that for low-resource languages, the sentiment embeddings can be projected from English. They evaluated their approach for English as the source and Spanish, Dutch, German, Russian, Italian, Czech, Japanese, and French. The induced embeddings lead to better results in 7 out of the 10 trials with accuracy over 79.3%.

While not common in cross-language SA, in [63] authors explored the architecture comprised of a feature extractor (BiLSTM) feed by embeddings followed by the classifier (dense layer) for transfer learning. First, they use a large dataset to train the whole model. Afterward, they fine-tuned in a small, labeled dataset from the target language, but only the embedding layer remains trainable. They trained using TripAdvisor reviews in English for the first stage and tweets in Greek for the second, being the results very sensible to the size of this dataset (accuracy drops from 91.7% to 73.2% as the dataset shrinks from 400 to 330). As the authors note, it will be interesting to study how a different degree of syntactic similarity between languages influences results.

Next works reviewed within the cross-lingual category explored the idea of using adversarial training to learn a set of language-independent features.

In [64] the authors delve into the synergies of microblog data in different languages from the same user to extract personalized language-specific or independent features to alleviate the lack of data in some sources. The architecture has four components. First, an attention mechanism encoding users as feature vectors to propagate their individuality across the system. The second component are encoders $[\theta^1, \theta^2]$, one for each language, computing specific-language features. The third element is the language-independent encoder θ^G that is feed with sentences from both languages as well the user attention vector. Encoders are CNN over different n-gram representations of the sentences and an attention mechanism for the user-specific information. The classifier is softmax layer for each language with inputs from θ^G and θ^1 or θ^2. The last module is a Generative Adversarial Network (GA) that drives θ^G to a set of features useful for SA when combined with θ^1 or θ^2 but uncorrelated with the language of the input sentence. Experiments with Twitter and Sina Weibo compared monolingual baseline models and the proposal, trained with both languages at once. Results show an increase of the F1 score up to 2.12% respect the best monolingual ($\geq$ 79.85%).

Other work that leverages adversarial training to learn a set of language-independent but highly discriminatory features is [65]. The architecture (ADAN) uses the Deep Averaging Network (DAN) in [76] as a feature extractor with a Bilingual Word Embedding (BWE) [77] as an input layer. These features are feed to the classifier and a language discriminator acting as an adversarial driving DAN to language-independent features. Empirical evaluation shows ADAN (accuracy $\geq$ 42.49%) improves in at least 6% over a version without the adversarial module (only DAN) trained with English to predict Chinese and Arab. It suggests that the adversarial mechanism is crucial for the results.

The adversarial mechanism is also critical in [70]. They build a cross-lingual word embedding using an Adversarial Auto Encoder (AAE) [78] feed from the outputs of two LSTM, one for each of the source and the target languages. On the top of this module sits a classifier based in Bidirectional Gated Recurrent Unit (BiGRU). Evaluating using Amazon comments with English as the source and Chinese and German as the target, the model (TL-AAE-BIGRU) achieved an F1 score $\geq$ 78.13%, about 3.44% better as average than the model without AEE ($\geq$ 73.25%). This result is consistent with [64,65] who noticed the benefits of the adversarial module too.

Instead of adversarial training, in [68] they opted to create universal embeddings by the combination of embeddings from high and low resources languages. The Universal SA (UniSent), involves the pre-training of two BiLSTM on English (labeled tweets), the alignment of low-resource language embeddings to the English ones with an unsupervised and domain-adversarial approach (MUSE [79]), and the fine-tuning on the low-resource languages validation sets applying an *Universal Embedding Layer*. This layer represents a word in a low-resource language by the weighted average of the most similar words to it in the English word embedding. The embeddings are feed to the classifier, a many-to-one LSTM layer. The experiments were carried in texts from OpeNER for Spanish and MultiBooked for Catalan. UniSent achieved an F1 score $\geq$ 81.4% for the binary classification and $\geq$ 54.2%, outperforming even a version tested using translated texts ($\geq$ 74.0% and $\geq$ 40.6%).

The simpler architecture comprising LASER (Language-Agnostic SEntence Representations) toolkit[2] as a language-independent embedding, a feature extractor (CNN or BiLSTM) and the classifier is proposed by [72]. The multilingual embedding aims to drive the model to language-agnostic representations, being able to perform SA in languages different from the ones seen during training. Experiments training with Polish (F1 score 79.91%) to predict other languages seems to evidence this premise since in all cases, F1 was $\geq$ 77.96%. Also, that the setup LASER+BiLSTM is better in general.

A similar architecture is proposed by [74] but using Bilingual Bag-of-Words without Word Alignments (BilBOWA) and VecMap[3] as cross-lingual embedding. Experiments over English and Persian electronic product reviews evaluated different alternatives for the embeddings and the feature extractors (CNN, LSTM, CNN-LSTM, and LSTM-CNN). The VecMap+LSTM-CNN with dynamic embeddings, i.e., fine-tuning the embeddings with the training data, achieved the best results with an F1 score of 91.82%.

So far, we have reviewed cross-lingual approaches for SMA. Though there are very different perspectives to solve the problem, they also shared some commons ideas, as the projection of resources such as sentiment embeddings. The next section is committed to reviewing the works that addressed the code-switching setup.

4.3. Code-switching approaches

In this section, we examine the reports that tackled code-switching sentiment analysis. This setup poses challenges such as spelling variations, transliteration, informal grammar forms, and the scarcity of annotated data.

The underlying idea of [57] is to learn a sentiment feature space preserving the similarity of sentences in terms of the sentiment they convey. This enables us to measure the relatedness between code-switched content and labeled data from a resource-rich corpus. The Sentiment Analysis of Code-Mixed Text (SACMT) architecture uses a siamese BiLSTM with tri-gram embeddings as input and a fully connected layer on the top. They compared a

[2] https://github.com/facebookresearch/LASER

[3] https://github.com/artetxem/vecmap

model trained only with pairs of code-mixed text (Hindi–English) with an F1 score of 67.2%, to other trained with pairs of sentences one from English and other code-mixed (75.9%). These results suggest that in effect, the model benefits from the additional data provided by the English corpus.

Most of the research in code-mixed text has focused on the English–Hindi setup. One exception is the seminal work on code-mixed Bambara–French Facebook comments in [59]. They examined different variations of the base architecture with embeddings (at word or character level) as input, followed by a feature extractor (one of LSTM, BiLSTM, CNN, CNN-LSTM) and finally, the classifier. To mitigate the lack of pre-trained embeddings in Bambara, the model learns multilanguage embeddings from characters or words in the code-mixed corpus. The best performing model was a one-layer CNN model with an accuracy of 83.23%. The comparison between LSTM and CNN as feature extractors, where the latter-one yields better results, is coherent with a noticeable preference for this type of model within the domain.

One problem when working with code-mixed data is the noise and the small size of datasets. To alleviate this, in [60] authors propose to use n-gram embeddings instead of the subword ones suggested by [80]. Another novelty idea within the domain is the model that works as an ensemble of a Multinomial Naïve Bayes (MNB) and a recurrent neural network (LSTM or BiLSTM) classifier. They trained the MNB using both word-based 1 and 2-gram features while the neural networks models with the character 3-grams. The results, together with those reported by [80], suggest that for the LSTM network, the 3-grams representation is a better option (F1 score of 58.6%) over characters (51.1%). The subword level encoding achieved 65.2% but, the difference between the architectures can mislead the conclusions. Values for the ensemble (66.1%) show that this can be of benefit.

Authors of [62] assessed different versions of the architecture that combines sequentially one feature extractor and a classifier. The first one, a document to vector (Doc2Vec) layer whose output is feed to an SVM. The other three featured a FastText classifier, a BiLSTM, and a CNN with a softmax. The last two had a trainable word embedding layer before the NN. They curated a new corpus for Kannada–English (best results for the CNN model with an accuracy of 71.5%). Also, they evaluated using two available Hindi–English and, Bengali–English corpus [81] with the BiLSTM model achieving slightly better results, 60.22% and 72.2% respectively.

In [67] authors delve into whether to use characters, words, or subwords. Also, if projecting code-mixed text to a single feature space is a rich enough representation for SA. Their architecture (CMSA) combines three parallel feature encoders before a classifier of four dense layers. The collective encoder aims to represent the overall sentiment of the sentences. It is based on a BiLSTM network whose end states are the features. The specific encoder is also a BiLSTM, but in this case, the intermediate states are also considered features through an attention mechanism. Both encoders take as input the output of a subword level CNN. The last one is a set of hand-crafted features to augment the information supplied to the model. They evaluated the effect of removing some of the components. CMSA achieved an F1 score of 82.7%, better than the model only with the specific (80.1%), or only the collective (79.5%). It seems to support the hypothesis about the synergies of the different representations.

Another model combining different feature extractors is the proposed in [69]. Like [56], they have character-level and word-level feature encoders. The first one is inspired in [80], stacking a CNN followed by two LSTM layers. It aims to help with language independence, noise, and non-standard spelling. The other extractor stacks two LSTM layers. The concatenation of the two feature sets is the input for the classifier, a stack of two dense layers and a softmax. Their model achieved an F1 score of 66.13% over the Hindi–English corpus in [80] improving about 5.5% the baseline in [80]. As in [56], combining both types of feature extractors seems to lead to better results.

Recently in [71] authors evaluated several architectures (BiLSTM-CNN, Double BiLSTM, Attention-based), each one with and without GloVe, and BERT (Bidirectional Encoder Representations from Transformers) [82] - over tweets and Facebook comments for English–Hindi and English–Bengali. The same models were trained in a monolingual corpus to observe the effects of code-mixing. The best model for code-mixing, the Attention-based model with custom word embeddings, achieved an F1 score average of 0.66 and 0.67 for the monolingual setup. It is interesting that the performance of $\text{BERT}_{base-uncased}$, the best model for the monolingual with 0.77, decreased noticeably for the code-mixed (about 0.63).

The work [73] also relies on subword embeddings. The architecture (SAEKCS), similar to [80], includes a CNN layer on top of the embeddings to extract local dependencies. Its output is processed by a BiLSTM layer after max-pooling, to learn long-term relations. On the top, a fully connected layer acts as the classifier. They evaluated SAEKCS using Kannada–English code-switching YouTube comments with an accuracy of 77.6%. They also assessed a subword LSTM (64.8%) and a BiLSTM (55.9%), suggesting that the short-term dependencies encoded by the CNN greatly benefit the model.

Finally, we also have reviewed code-switching approaches for SMA. The more trending proposal is to use subword embeddings to allow guessing the meaning of unknown/out-of-vocabulary words. We have contrasted very different mixing languages, in the majority of cases, English mix. The next section is an overview of the deep learning implementations across the different setups.

4.4. An overview of the different deep learning implementations

Up to this point, we have reviewed a comprehensive corpus of research works that had leveraged deep learning for multilingual sentiment analysis. We had focused on the underlying hypothesis of the proposed approaches, highlighting what is common or different. Also, the results on the evaluation corpus.

In this section, we aim to make it easier for the reader to have a quick overview of the material we analyzed.

Table 1 summarizes the works we reviewed, describing the core of the best model architecture (Proposed Model), baselines, distinctive ideas (Proposed Approach), the classification categories, multilingualism level, languages, and information about the corpus.

The classification categories mainly will be divided into positive or negative (binary), also with a neutral class (ternary) or five classes.[4]

5. Discussion and future directions

In this section, we discuss the main findings of our study. We also highlight some unexplored topics that may hint at interesting directions for further research.

5.1. Languages and social media in MSA

The 24 studies we analyzed covered 23 different languages. In most cases English was the resource-rich language, except [72] and [59]. However, authors have explored synergies between different languages, as shown in Fig. 1. Concerning the social

[4] Sometimes authors remove the *neutral* class or add another, e.g., *ambivalent*.

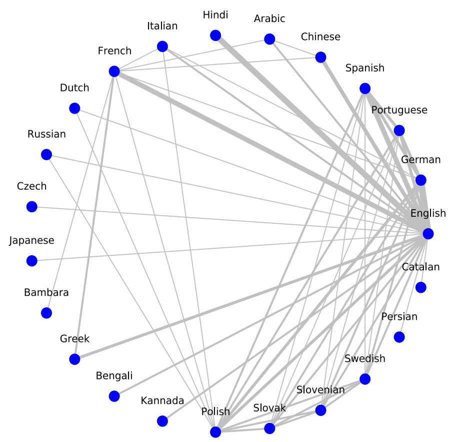

Fig. 1. Synergies between languages. A link between two languages indicates that both has been used simultaneously in a model, as source or target language, or to learn multilingual feature spaces.

media Twitter and Miscellaneous[5] account for most of the works. Fig. 2 the data de-aggregated by language and social media.

The next subsection is devoted to the analysis of the relations between the MSA setup and the architecture proposed to deal with the problem.

5.2. DL architectures for MSA

A comparison between the backbone of the different architectures suggests that in general, for multilanguage sentence-level SA, authors have explored a plainer architecture. It leverages trainable embeddings preceding the feature extractor and finally a classifier layer. Regardless, there is a wide range of alternatives in the design of the classifier, from a single BiLSTM [63] to parallel CNN [56]. Only one study focused on the aspect-level SA. So, it may be difficult to draw conclusions related to the best architectural decisions to tackle this problem. However, experiments in [66] suggest that the attention mechanism as well the aspect embedding plays a key role. Moreover, results from another work [83] show that the mere addition of attention to a simpler model such as LSTM or CNN does not lead to state-of-the-art results. This is indicative of the convenience of combining feature extractors at different levels for aspect-based SA.

For cross-lingual SA network designs are more diverse. However, two core ideas can be identified. The use of an adversarial module to drive the feature extractors to language-independent representations as in [64] or [65]. The other is leveraging trainable cross-lingual embeddings or even pre-trained ones such as Facebook LASER [72].

Code-switching backbones tend to be simpler than cross-lingual, yet more elaborated than the multilingual ones. They can include parallel encoders at a character, word, or subword levels [67,69] or implement ensemble models including deep neural networks and other classification techniques [60].

In general, a lot of effort has been devoted to compare feature extractors based on CNN, LSTM, or BiLSTM and different levels of embeddings. We analyzed the co-occurrences of the different types of networks, attention mechanisms, etc. within the same model, to visualize how authors have been using them. Edges in Fig. 3 mean that the two concepts at nodes have been used together in a model, the weight, how often this relationship has occurred. The graph shows how CNN is preferred along with hybrid architectures, where CNN + BiLSTM is the most studied with [59] and [62] reporting comparisons between models where they only changed this setting, resulting in better performance for the CNN models. Instead, results were equivalent in [61] and better for LSTM in [51]. Thus, seem to be no consensus about this subject.

5.2.1. Embedding approaches for MSA

Regardless of the domain, most authors rely on some embedding types, with a trend toward learning the embedding along the training process. There is no common opinion about whether to work at a character, word, or subword. Authors such as [51] advocate character level embeddings due to their simplicity. However, they reported slightly better results for the word level case but in [37] shows that character level could help achieve a language-independence. In [59] they compared the character and word-based embeddings, with the latter yielding better results. Also, subword embeddings reported outperforming the character ones by [67].

Unsurprisingly we can find pre-trained subword multilingual embeddings, such as [84], which can be useful for MSA and SA for low-resource languages. However, there are other alternatives to be explored such as document level [62], a combination of different levels [56,66] and even sentiment-driven embeddings [58], universal embeddings [68] or the use of some tools, such as LASER [72].

5.3. Future directions

Finally, we elaborate on the current state of research and provide a pathway for what can be done or needs to be done within the following few years.

Little-explored MSA levels. Hitherto, Aspect-Based Sentiment Analysis has not widely been addressed using multilingual deep learning approaches. As [66] suggests, tackling this problem may require more complex architectures. Moreover, it needs to be studied if current proposals can handle mixing setups such as aspect-based code-switching.

MSA setup shift across time. Fig. 4 suggest a shift of the interest from multilingual to cross-lingual and code-switching approaches. In MSA this can be explained since initially most of the works focused on the multilingual setup evaluating many variations of the same design. Researchers could perceive this path as depleted. Also, the adoption of transformer-based architectures such as BERT [82] and even multilingual models such as *Multilingual BERT*[6] allows the researchers to focus on fine-tuning the models instead of training them from scratch with a multilingual corpus as has been typical for the multilingual setup.

Multilingual representations. Multilingual embeddings and adversarial training seem to be the most common approaches to achieve multilingualism within the analyzed corpus. But there is not a common standpoint about the level of embedding to use or how a single model can encode a multilingual or language-agnostic feature space use-full for the downstream tasks. However, this debate seems to be shifted to the transformer-based architectures where different tokenizers are being considered [85]. Moreover, despite the success in training transformers in a multilingual corpus, recent studies suggest that there is a lot of room for improvement [86]. In this sense probably we will see an increased number of works studying the impact of the differences between languages and language families.

[5] Under this category, we considered works such as [66] which used the SemEval 2016 Task 5, and other corpora that aggregated text from different sources.

[6] https://github.com/google-research/bert/blob/master/multilingual.md

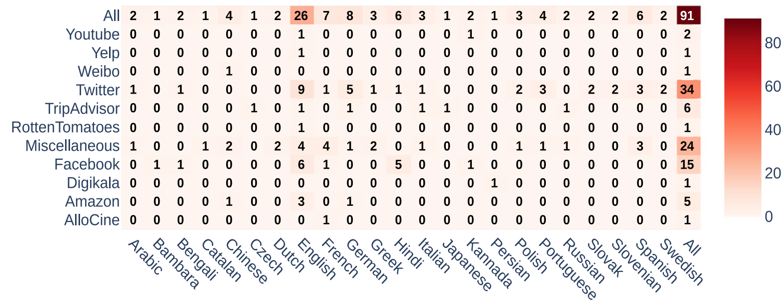

Fig. 2. Languages vs Social Media.

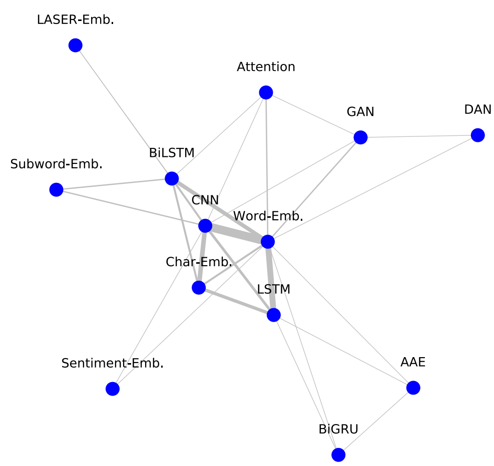

Fig. 3. Neural Network architectures and its relations, out of 24, across reviewed papers.

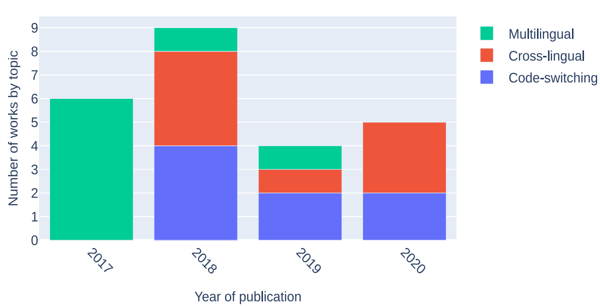

Fig. 4. MSA approaches, out of 24 papers, across fours years.

SA-specific representations. For MSA, the aforementioned architectures may handle the specifics of this domain such as the code-switching or the aspect-based setup. It will be necessary to study if it is worth to couple techniques such as attention between different levels of representation, sentiment embeddings, or adversarial learning (e.g GAN-BERT [87]).

Low-resource languages and dialects. Despite languages from different families has been studied (see Fig. 2) the coverage is far from complete. Moreover, the steady interest in sentiment analysis, the lack so far of a universal approach, and the new opportunities [88,89] would trigger the development of systems and corpora for SA in other languages. In this sense, we would see tailored solutions dealing with dialects and mixing languages. Besides India (native languages mixing with English) there are other large groups such as (a) Mexico and USA (Spanglish), (b) Brazil and its border countries, Portugal and Spain (Portuñol) (c) Paraguay (Jopara,[7] Portuñol), to mention few cases. Nevertheless, the scarce of available corpora is a challenge for tackle these code-mixing tasks. For instance, in [71] we could observe that with an English mixing MSA setup, BERT had been unable to outperform the traditional DL models. Thus, substantial progress needs still to be made.

6. Conclusions

In this work, we reviewed 24 that studied 23 and 11 different languages and sources. The observed trend evidences the steady interest in this domain, so we expect to see this direction continue.

As regards the different MSA setups, the multilingual approach seems to be decreasing in interest. However, aspect-based sentiment analysis is still an understudied domain and an open research field with a lot of scope for future works.

We highlighted the main ideas authors proposed to tackle the challenge that represents the lack of annotated data or to achieve language independent models. Despite state-of-the-art results in some cases, the simpler backbone comprising embeddings, a feature extractor, and a classifier seems to be unappropriated for more complex scenarios. Also, there are unsolved questions such as which type of embedding captures better the particulars of MSA. We hint about future research directions, for example, if ideas such as contextualized embeddings, which have proven very useful in other tasks, can further improve MSA. Finally, although studies have covered very different languages such as Arabic, Chinese, or Hindi, the world is extraordinarily rich in languages, cultures, and ways of expressing feelings. Thus, better approaches need to be assessed or developed for new scenarios.

CRediT authorship contribution statement

Marvin M. Agüero-Torales: Conceptualization, Formal analysis, Investigation, Data curation, Writing - original draft, Visualization. **José I. Abreu Salas:** Validation, Investigation, Writing - original draft, Writing - review & editing. **Antonio G. López-Herrera:** Conceptualization, Methodology, Writing - original draft, Supervision.

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.

[7] Mixing Guarani (an indigenous language) with Spanish.

Acknowledgments

This research work has been partially funded by the Generalitat Valenciana (Conselleria d'Educació, Investigació, Cultura i Esport) and the Spanish Government through the projects SIIA (PROMETEO/2018/089, PROMETEU/2018/089) and LIVING-LANG (RTI2018-094653-B-C22). We are also immensely grateful to David Vilares (Universidade da Coruña, Spain), for his recommendations provided a valuable orientation.

References

[1] S.L. Lo, E. Cambria, R. Chiong, D. Cornforth, Multilingual sentiment analysis: from formal to informal and scarce resource languages, Artif. Intell. Rev. 48 (4) (2017) 499–527.
[2] B. Pang, L. Lee, et al., Opinion mining and sentiment analysis, Found. Trends® Inf. Retr. 2 (1–2) (2008) 1–135.
[3] S. Wang, C.D. Manning, Baselines and bigrams: Simple, good sentiment and topic classification, in: Proc. of the 50th Annual Meeting of the ACL: Short Papers, Vol. 2, ACL, 2012, pp. 90–94.
[4] B. Liu, Sentiment analysis and opinion mining, Synth. Lect. Human Lang. Technol. 5 (1) (2012) 1–167.
[5] B. Liu, Sentiment Analysis: Mining Opinions, Sentiments, and Emotions, Cambridge University Press, 2015.
[6] R. Socher, A. Perelygin, J. Wu, J. Chuang, C.D. Manning, A. Ng, C. Potts, Recursive deep models for semantic compositionality over a sentiment treebank, in: Proc. of the 2013 Conf. on Empirical Methods in NLP, 2013, pp. 1631–1642.
[7] C. Dos Santos, M. Gatti, Deep convolutional neural networks for sentiment analysis of short texts, in: Proc. of COLING 2014, the 25th Int. Conf. on Computational Linguistics: Technical Papers, 2014, pp. 69–78.
[8] Y. Kim, Convolutional neural networks for sentence classification, in: Proc. of the 2014 Conf. on Empirical Methods in NLP, EMNLP, ACL, Doha, Qatar, 2014, pp. 1746–1751.
[9] D. Tang, B. Qin, T. Liu, Deep learning for sentiment analysis: successful approaches and future challenges, Wiley Interdiscip. Rev.: Data Min. Knowl. Discov. 5 (6) (2015) 292–303.
[10] L.M. Rojas-Barahona, Deep learning for sentiment analysis, Lang. Linguist. Compass 10 (12) (2016) 701–719.
[11] P. Singhal, P. Bhattacharyya, Sentiment Analysis and Deep Learning: A Survey, Center for Indian Language Technology, Indian Institute of Technology, Bombay, 2016.
[12] Q.T. Ain, M. Ali, A. Riaz, A. Noureen, M. Kamran, B. Hayat, A. Rehman, Sentiment analysis using deep learning techniques: a review, Int. J. Adv. Comput. Sci. Appl. 8 (6) (2017) 424.
[13] D. Vilares, Compositional Language Processing for Multilingual Sentiment Analysis (Ph.D. thesis), Universidade da Coruña, 2017.
[14] L.J. Zhang, S. Wang, B. Liu, Deep learning for sentiment analysis: A survey, Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 8 (2018).
[15] O. Habimana, Y. Li, R. Li, X. Gu, G. Yu, Sentiment analysis using deep learning approaches: an overview, Sci. China Inf. Sci. 63 (1) (2019) 1–36.
[16] R. Wadawadagi, V. Pagi, Sentiment analysis with deep neural networks: comparative study and performance assessment, Artif. Intell. Rev. (2020).
[17] H. Nankani, H. Dutta, H. Shrivastava, P.R. Krishna, D. Mahata, R.R. Shah, Multilingual sentiment analysis, in: Deep Learning-Based Approaches for Sentiment Analysis, Springer, 2020, pp. 193–236.
[18] J.M. Wiebe, R.F. Bruce, T.P. O'Hara, Development and use of a gold-standard data set for subjectivity classifications, in: Proc. of the 37th Annual Meeting of the ACL, 1999, pp. 246–253.
[19] B. Pang, L. Lee, S. Vaithyanathan, Thumbs up?: sentiment classification using machine learning techniques, in: Proc. of the ACL-02 Conf. on Empirical Methods in NLP, Vol. 10, ACL, 2002, pp. 79–86.
[20] P.D. Turney, Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews, in: Proc. of the 40th Annual Meeting on ACL, ACL, 2002, pp. 417–424.
[21] R. Feldman, Techniques and applications for sentiment analysis, Commun. ACM 56 (4) (2013) 82–89.
[22] E. Cambria, B. Schuller, Y. Xia, C. Havasi, New avenues in opinion mining and sentiment analysis, IEEE Intell. Syst. 28 (2) (2013) 15–21.
[23] S. Rosenthal, N. Farra, P. Nakov, SemEval-2017 task 4: Sentiment analysis in Twitter, in: Proc. of the 11th Int. Workshop on Semantic Evaluation, SemEval '17, ACL, Vancouver, Canada, 2017.
[24] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Neural and Information Processing System, NIPS, 2013.
[25] J. Pennington, R. Socher, C.D. Manning, GloVe: Global vectors for word representation, in: Empirical Methods in NLP, EMNLP, 2014, pp. 1532–1543.
[26] M.E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, in: Proc. of NAACL-HLT, 2018, pp. 2227–2237.
[27] J. Yu, J. Jiang, Learning Sentence Embeddings with Auxiliary Tasks for Cross-Domain Sentiment Classification, ACL, 2016.
[28] H. Xu, B. Liu, L. Shu, P.S. Yu, Double embeddings and CNN-based sequence labeling for aspect extraction, in: Proc. of the 56th Annual Meeting of the ACL (Vol. 2: Short Papers), ACL, 2018, pp. 592–598.
[29] B. Huang, Y. Ou, K.M. Carley, Aspect level sentiment classification with attention-over-attention neural networks, in: Int. Conf. on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation, Springer, 2018, pp. 197–206.
[30] Y. Ma, H. Peng, E. Cambria, Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive LSTM, in: Thirty-Second AAAI Conf. on Artificial Intelligence, 2018.
[31] S. Ruder, P. Ghaffari, J.G. Breslin, A hierarchical model of reviews for aspect-based sentiment analysis, in: Proc. of the 2016 Conf. on Empirical Methods in NLP, 2016, pp. 999–1005.
[32] D. Tang, B. Qin, T. Liu, Document modeling with gated recurrent neural network for sentiment classification, in: Proc. of the 2015 Conf. on Empirical Methods in NLP, 2015, pp. 1422–1432.
[33] G. Rao, W. Huang, Z. Feng, Q. Cong, LSTM with sentence representations for document-level sentiment classification, Neurocomputing 308 (2018) 49–57.
[34] Y. Ganin, V. Lempitsky, Unsupervised domain adaptation by backpropagation, in: Proc. of the 32nd Int. Conf. on Int. Conf. on Machine Learning, Vol. 37, ICML'15, JMLR.org, 2015, pp. 1180–1189.
[35] Z. Li, Y. Zhang, Y. Wei, Y. Wu, Q. Yang, End-to-end adversarial memory network for cross-domain sentiment classification, in: IJCAI, 2017, pp. 2237–2243.
[36] B. Agarwal, R. Nayak, N. Mittal, S. Patnaik, Deep Learning-Based Approaches for Sentiment Analysis, Springer, 2020.
[37] J. Wehrmann, W. Becker, H.E.L. Cagnini, R.C. Barros, A character-based convolutional neural network for language-agnostic Twitter sentiment analysis, in: 2017 Int. Joint Conf. on Neural Networks, IJCNN, 2017, pp. 2384–2391.
[38] J.T. Zhou, S.J. Pan, I.W. Tsang, Y. Yan, Hybrid heterogeneous transfer learning through deep learning, in: Twenty-Eighth AAAI Conf. on Artificial Intelligence, 2014.
[39] G. Zhou, Z. Zeng, J.X. Huang, T. He, Transfer learning for cross-lingual sentiment classification with weakly shared deep neural networks, in: Proc. of the 39th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, ACM, 2016, pp. 245–254.
[40] P. Singhal, P. Bhattacharyya, Borrow a little from your rich cousin: Using embeddings and polarities of English words for multilingual sentiment classification, in: Proc. of COLING 2016, the 26th Int. Conf. on Computational Linguistics: Technical Papers, 2016, pp. 3053–3062.
[41] Z. Wang, Y. Zhang, S. Lee, S. Li, G. Zhou, A bilingual attention network for code-switched emotion prediction, in: Proc. of COLING 2016, The 26th Int. Conf. on Computational Linguistics: Technical Papers, 2016, pp. 1624–1634.
[42] S. Ghosh, S. Ghosh, D. Das, Sentiment identification in code-mixed social media text, 2017, arXiv:abs/1707.01184.
[43] G.I. Ahmad, J. Singla, N. Nikita, Review on sentiment analysis of Indian languages with a special focus on code mixed Indian languages, in: 2019 Int. Conf. on Automation, Computational and Technology Management, ICACTM, IEEE, 2019, pp. 352–356.
[44] E. Tromp, Multilingual Sentiment Analysis on Social Media, Lap Lambert Academic Publ, 2012.
[45] C. Banea, R. Mihalcea, J. Wiebe, Multilingual sentiment and subjectivity analysis, in: Multilingual NLP, Vol. 6, Prentice-Hall New York, New York, 2011, pp. 1–19.
[46] I.S.V. Roncal, Multilingual Sentiment Analysis in Social Media (Ph.D. thesis), Universidad del País Vasco-Euskal Herriko Unibertsitatea, 2019.
[47] F. Steiner-Correa, M.I. Viedma-del Jesus, A. Lopez-Herrera, A survey of multilingual human-tagged short message datasets for sentiment analysis tasks, Soft Comput. 22 (24) (2018) 8227–8242.
[48] M.A. Abdullah, Deep Learning for Sentiment and Emotion Detection in Multilingual Contexts (ProQuest dissertations and theses, Ph.D. thesis), 2018, p. 103, Copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Last update - 2019-10-18.
[49] B. Ay Karakuş, M. Talo, İ.R. Hallaç, G. Aydin, Evaluating deep learning models for sentiment classification, Concurr. Comput.: Pract. Exper. 30 (21) (2018) e4783.
[50] J. Deriu, A. Lucchi, V. De Luca, A. Severyn, S. Müller, M. Cieliebak, T. Hofmann, M. Jaggi, Leveraging large amounts of weakly supervised data for multi-language sentiment classification, in: Proc. of the 26th Int. Conf. on World Wide Web, 2017, pp. 1045–1052.
[51] W. Becker, J. Wehrmann, H.E. Cagnini, R.C. Barros, An efficient deep neural architecture for multilingual sentiment analysis in Twitter, in: The Thirtieth Int. Flairs Conf., 2017.

[52] I. Mozetič, M. Grčar, J. Smailović, Multilingual Twitter sentiment classification: The role of human annotators, PLoS One 11 (5) (2016) e0155036.
[53] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[54] S. Zhang, X. Zhang, J. Chan, Language-independent Twitter classification using character-based convolutional networks, in: Int. Conf. on Advanced Data Mining and Applications, Springer, 2017, pp. 413–425.
[55] L. Medrouk, A. Pappa, Deep learning model for sentiment analysis in multilingual corpus, in: Int. Conf. on Neural Information Processing, Springer, 2017, pp. 205–212.
[56] S. Zhang, X. Zhang, J. Chan, A word-character convolutional neural network for language-agnostic Twitter sentiment analysis, in: Proc. of the 22nd Australasian Document Computing Symposium, ACM, 2017, p. 12.
[57] N. Choudhary, R. Singh, I. Bindlish, M. Shrivastava, Sentiment analysis of code-mixed languages leveraging resource rich languages, in: 19th Int. Conf. on Computational Linguistics and Intelligent Text Processing, CICLing-2018, 2018.
[58] X. Dong, G. De Melo, Cross-lingual propagation for deep sentiment analysis, in: Thirty-Second AAAI Conf. on Artificial Intelligence, 2018.
[59] A. Konate, R. Du, Sentiment analysis of code-mixed Bambara-French social media text using deep learning techniques, Wuhan Univ. J. Nat. Sci. 23 (3) (2018) 237–243.
[60] M. Gopal Jhanwar, A. Das, An ensemble model for sentiment analysis of Hindi-English code-mixed data, 2018, arXiv preprint arXiv:1806.04450.
[61] L. Medrouk, A. Pappa, Do deep networks really need complex modules for multilingual sentiment polarity detection and domain classification? in: 2018 Int. Joint Conf. on Neural Networks, IJCNN, IEEE, 2018, pp. 1–6.
[62] K. Shalini, H.B. Ganesh, M.A. Kumar, K. Soman, Sentiment analysis for code-mixed Indian social media text with distributed representation, in: 2018 Int. Conf. on Advances in Computing, Communications and Informatics, ICACCI, IEEE, 2018, pp. 1126–1131.
[63] K. Stavridis, G. Koloniari, E. Keramopoulos, Deriving word embeddings using multilingual transfer learning for opinion mining, in: 2018 South-Eastern European Design Automation, Computer Engineering, Computer Networks and Society Media Conf., SEEDA_CECNSM, IEEE, 2018, pp. 1–6.
[64] W. Wang, S. Feng, W. Gao, D. Wang, Y. Zhang, Personalized microblog sentiment classification via adversarial cross-lingual multi-task learning, in: Proc. of the 2018 Conf. on Empirical Methods in NLP, 2018, pp. 338–348.
[65] X. Chen, Y. Sun, B. Athiwaratkun, C. Cardie, K. Weinberger, Adversarial deep averaging networks for cross-lingual sentiment classification, in: Proc. of the 2018 Conf. on Empirical Methods in NLP, 2018, pp. 557–570.
[66] G. Liu, X. Huang, X. Liu, A. Yang, A novel aspect-based sentiment analysis network model based on multilingual hierarchy in online social network, Comput. J. (2019).
[67] Y.K. Lal, V. Kumar, M. Dhar, M. Shrivastava, P. Koehn, De-mixing sentiment from code-mixed text, in: Proc. of the 57th Annual Meeting of the ACL: Student Research Workshop, 2019, pp. 371–377.
[68] M. Jabreel, N. Maaroof, A. Valls, A. Moreno, UniSent: Universal sentiment analysis system for low-resource languages, in: CCIA, 2019, pp. 387–396.
[69] S. Mukherjee, Deep learning technique for sentiment analysis of Hindi-English code-mixed text using late fusion of character and word features, in: 2019 IEEE 16th India Council Int. Conf., INDICON, IEEE, 2019, pp. 1–4.
[70] J. Shen, X. Liao, S. Lei, Cross-lingual sentiment analysis via AAE and BiGRU, in: 2020 Asia-Pacific Conf. on Image Processing, Electronics and Computers, IPEC, IEEE, 2020, pp. 237–241.
[71] A. Jamatia, S. Swamy, B. Gambäck, A. Das, S. Debbarma, Deep learning based sentiment analysis in a code-mixed English-Hindi and English-Bengali social media corpus, Int. J. Artif. Intell. Tools 29 (5) (2020).
[72] K. Kanclerz, P. Miłkowski, J. Kocoń, Cross-lingual deep neural transfer learning in sentiment analysis, Procedia Comput. Sci. 176 (2020) 128–137.
[73] R. Chundi, V.R. Hulipalled, J. Simha, SAEKCS: Sentiment analysis for English–Kannada code switchtext using deep learning techniques, in: 2020 Int. Conf. on Smart Technologies in Computing, Electrical and Electronics, ICSTCEE, IEEE, 2020, pp. 327–331.
[74] R. Ghasemi, S.A. Ashrafi Asli, S. Momtazi, Deep Persian sentiment analysis: Cross-lingual training for low-resource languages, J. Inf. Sci. (2020) 1–14.
[75] M. Pontiki, D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, A.-S. Mohammad, M. Al-Ayyoub, Y. Zhao, B. Qin, O. De Clercq, et al., Semeval-2016 task 5: Aspect based sentiment analysis, in: Proc. of the 10th Int. Workshop on Semantic Evaluation, SemEval-2016, 2016, pp. 19–30.
[76] M. Iyyer, V. Manjunatha, J. Boyd-Graber, H. Daumé III, Deep unordered composition rivals syntactic methods for text classification, in: Proc. of the 53rd Annual Meeting of the ACL and the 7th Int. Joint Conf. on NLP (Vol. 1: Long Papers), 2015, pp. 1681–1691.
[77] W.Y. Zou, R. Socher, D. Cer, C.D. Manning, Bilingual word embeddings for phrase-based machine translation, in: Proc. of the 2013 Conf. on Empirical Methods in NLP, 2013, pp. 1393–1398.
[78] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, B. Frey, Adversarial autoencoders, 2016, arXiv:1511.05644.
[79] G. Lample, A. Conneau, M. Ranzato, L. Denoyer, H. Jégou, Word translation without parallel data, in: Int. Conf. on Learning Representations, 2018.
[80] A. Joshi, A. Prabhu, M. Shrivastava, V. Varma, Towards sub-word level compositions for sentiment analysis of Hindi-English code mixed text, in: Proc. of COLING 2016, the 26th Int. Conf. on Computational Linguistics: Technical Papers, 2016, pp. 2482–2491.
[81] B.G. Patra, D. Das, A. Das, Sentiment analysis of code-mixed Indian languages: an overview of SAIL_Code-Mixed Shared Task@ ICON-2017, 2018, arXiv preprint arXiv:1803.06745.
[82] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018, arXiv preprint arXiv:1810.04805v2.
[83] Y. Zhu, X. Gao, W. Zhang, S. Liu, Y. Zhang, A bi-directional LSTM-CNN model with attention for aspect-level text classification, Future Internet 10 (12) (2018) 116.
[84] B. Heinzerling, M. Strube, BPEmb: Tokenization-free pre-trained subword embeddings in 275 languages, in: N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, T. Tokunaga (Eds.), Proc. of the Eleventh Int. Conf. on Language Resources and Evaluation, LREC 2018, European Language Resources Association (ELRA), Miyazaki, Japan, 2018, (Conf. chair).
[85] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, Transformers: State-of-the-art NLP, in: Proc. of the 2020 Conf. on Empirical Methods in NLP: System Demonstrations, ACL, 2020, pp. 38–45, Online.
[86] T. Pires, E. Schlinger, D. Garrette, How multilingual is multilingual BERT? in: Proc. of the 57th Annual Meeting of the ACL, ACL, Florence, Italy, 2019, pp. 4996–5001.
[87] D. Croce, G. Castellucci, R. Basili, GAN-BERT: Generative adversarial learning for robust text classification with a bunch of labeled examples, in: Proc. of the 58th Annual Meeting of the ACL, ACL, 2020, pp. 2114–2119, Online.
[88] L. Yue, W. Chen, X. Li, W. Zuo, M. Yin, A survey of sentiment analysis in social media, Knowl. Inf. Syst. 60 (2) (2019) 617–663.
[89] A. Zunic, P. Corcoran, I. Spasic, Sentiment analysis in health and well-being: Systematic review, JMIR Med. Inform. 8 (1) (2020) e16023.

Marvin M. Agüero-Torales is a Ph.D. student of the Doctoral Programme in Information and Communication Technologies at the University of Granada (UGR) at the Department of Computer Science and Artificial Intelligence and, Research Engineer at the UOC (Universitat Oberta de Catalunya, IN3). Former member of the Universidade da Coruña (UDC) and BSC-CNS. His research experience and interests include (i) natural language processing; (ii) text mining; (iii) machine learning; with a particular focus on low-resource languages.

José I. Abreu Salas is a Researcher at the University Institute for Computing Research (IUII), Alicante. He has been a member of the Cuban chapter of the International Association of Pattern Recognition, and a Full-Time Professor at the University of Matanzas (UMCC), and Catholic University of the Most Holy Conception (UCSC). His research interest covers (i) data-driven solutions in natural language processing (ii) instance selection and prototype construction algorithms.

Antonio G. Lopez-Herrera is an Associate Professor in the Department of Computer Science and Artificial Intelligence at the University of Granada (UGR). He holds a degree in Computer Engineering (2003, UGR), a Ph.D. in Computer Science (2006, UGR), a degree in Library and Information Science (2007, UGR) and a Master in Information and Scientific Communication (2008, UGR). His lines of research include information access, retrieval, filtering and evaluation, recommender systems, opinion mining, sentiment analysis, and bibliometrics. He teaches at the School of Computer (UGR) and at the Faculty of Communication and Documentation (UGR).