*Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 2698–2705
Marseille, 11–16 May 2020


# Identifying Sentiments in Algerian Code-switched User-generated Comments

**Wafia Adouane[1], Samia Touileb[2], and Jean-Philippe Bernardy[1]**
[1] Department of Philosophy, Linguistics and Theory of Science (FLoV),
Centre for Linguistic Theory and Studies in Probability (CLASP), University of Gothenburg
[2] Language Technology Group, Department of Informatics, University of Oslo
[1] firstname.lastname@gu.se, [2]samiat@ifi.uio.no

**Abstract**

We present in this paper our work on Algerian language, an under-resourced North African colloquial Arabic variety, for which we built a comparably large corpus of more than 36,000 code-switched user-generated comments annotated for sentiments. We opted for this data domain because Algerian is a colloquial language with no existing freely available corpora. Moreover, we compiled sentiment lexicons of positive and negative unigrams and bigrams reflecting the code-switches present in the language. We compare the performance of four models on the task of identifying sentiments, and the results indicate that a CNN model trained end-to-end fits better our unedited code-switched and unbalanced data across the predefined sentiment classes. Additionally, injecting the lexicons as background knowledge to the model boosts its performance on the minority class with a gain of 10.54 points on the F-score. The results of our experiments can be used as a baseline for future research for Algerian sentiment analysis.

**Keywords:** Algerian Arabic, code-switching, user-generated data, sentiment analysis, under-resourced colloquial languages

## 1. Introduction

Sentiment Analysis (SA) is a well-established NLP task which is commonly framed as a text classification problem. SA includes various related applications, depending on the domain and its real-world use, such as product reviewing and opinion mining. It has been applied to several domains, mostly for curated monolingual data. Recently, however, the focus has been extended to new domains and settings, namely to user-generated data which reflects real-world use cases.

In this paper, we focus on identifying and classifying sentiments by analysing user-generated comments on YouTube videos. We work with comments written in Algerian Arabic, a non-standardised Arabic variety characterised by heavy code-switching between co-existing languages and language varieties mainly Modern Standard Arabic (MSA), Berber, French, and local Arabic variants.

There is a large body of work on SA for Arabic, largely for MSA and Middle Eastern Arabic varieties (for example (Rushdi-Saleh et al., 2011; Abdul-Mageed and Diab, 2012; Zbib et al., 2012; Aly and Atiya, 2013; Nabil et al., 2015; ElSahar and El-Beltagy, 2015; Salameh et al., 2015)). Nevertheless, there has been less work done for Northern African Arabic varieties, which are indeed colloquial languages, due primarily to the scarcity of written linguistic resources.

However, new social media platforms made it possible to obtain user-generated data reflecting real use of such languages. In interactive communication channels, users use speech-like languages to express themselves with spontaneous spelling for non-standardised languages. Hence this domain is potentially a useful resource for analysing the linguistic properties of this kind of unedited code-switched textual data, and therefore better understanding how these languages are naturally used.

This paper is an attempt to bridge the gap in sentiment analysis for user-generated data written in colloquial languages. As main contributions, (i) we introduce our newly built linguistic resources collected from YouTube (corpus and sentiment lexicons) for Algerian, and annotated for sentiments. To the best of our knowledge, this is the largest user-generated corpus annotated for sentiments for Algerian. (ii) We compare the performance of Support Vector Machine (SVM) and three end-to-end deep neural networks (DNNs) on identifying sentiments from a code-switched colloquial language. (iii) We try to improve the best DNN models by injecting sentiment lexicons as background knowledge and by augmenting the number of instances for minority classes in training data. All the material described in this paper will be made available on request from the authors to the research community.

In what follows, in Section 2., we briefly review related work. In Section 3., we describe our newly built linguistic resources including corpus creation and annotation, sentiment lexicon compilation along with their detailed statistics. In Section 4., we present the approaches and the models that we use to identify sentiments from comments on social media. In Section 5., we describe our experiments and discuss the results. We conclude in Section 6. with our main findings and future plans.

## 2. Related Work

The largest body of research in sentiment analysis — including its various applications— is predominantly done for English. However, recently research has expended to other languages, among others Arabic and its colloquial varieties. Traditionally, sentiment analysis was heavily based on sentiment lexicons (Turney, 2002; Hu and Liu, 2004; Taboada et al., 2011). That is to say, a sentiment of a text segment, be it a document or a sentence, was computed based on sentiment lexicon lookup by counting the number of positive or negative words, or combining them with traditional classifiers as SVMs. This approach, however, has proven to be limited because negation, intensification,

downtoning, etc., can alter the polarity of words (Taboada et al., 2011).

There have been a few attempts to overcome these limitations. For instance Taboada et al. (2011) considered classifying the polarity of documents using lexicon information and rules for negation and intensification. They reported that their approach outperformed machine learning when tested across domains. Moreover, the focus in sentiment analysis has lately shifted to the use of end-to-end learning using information from the training corpus, and sometimes relying on the use of pre-trained embeddings and incorporating sentiment knowledge into their models. For instance, Lei et al. (2018) used two sentiment lexicons and Shin et al. (2016) used extra information from six lexicons with a Convolutional Neural Networks (CNN), and reported competitive results on one of the SemEval-2016 tasks. Similarly, Barnes et al. (2019) incorporated lexicon information into a Bidirectional Long Short-Term Memory (BiLSTM) sentiment classifier using a multi-task learning framework.

Nonetheless, most approaches for sentiment analysis have not incorporated lexicons, but they heavily rely on end-to-end supervised approaches. For example, for sentence-level sentiment analysis Kim (2014) and Dahou et al. (2016) used CNN models, Qian et al. (2016) used a variation of LSTM, and Tai et al. (2015) used tree-structured LSTM.

Sentiment Analysis work for Northern African Arabic was done, among others, by Elouardighi et al. (2017) who used a combination of various features to train an SVM, random forests, and decision trees to classify MSA and Moroccan Facebook comments. At the same time, Medhaffar et al. (2017) used SVM, Binary Naive Bayes, and a Multilayer Perceptron trained on three different corpora: an MSA corpus (OCA – (Rushdi-Saleh et al., 2011)), a corpus of MSA and other Arabic dialect (LABR – (Aly and Atiya, 2013)), and TSAC (Tunisian Sentiment Analysis Corpus) corpus, a code-switched corpus, to prove that it is necessary to train classifiers on dialects to achieve good accuracy. More recently, Jerbi et al. (2019) used the code-switched corpus TSAC presented in (Medhaffar et al., 2017) for sentiment classification into the two classes positive and negative using the deep neural approaches LSTM, BiLSTM, deep-LSTM, and deep-BiLSTM, along with word embeddings. The authors reported that their approach of deep-LSTM outperformed the models presented by Medhaffar et al. (2017), and achieved and overall accuracy of 90%.

With respect to sentiment analysis for Algerian specifically, little work has been done. Mataoui et al. (2016) proposed a lexicon-based approach for SA on Facebook comments using resources translated from MSA to generate three lexicons (of keywords, negations, and intensification words), and a list of emoticons and "common" Algerian phrases used to express polarity. They reported an accuracy of 79.13%. Then Guellil et al. (2018) translated an English sentiment lexicon to Algerian while keeping and transferring the polarity of the English words. They also automatically annotated Facebook comments as positive or negative based on the constituents' polarities, using a Bag-of-Words (BoW) model and document embeddings. The authors used five different sentiment classifiers, namely SVM, Naive Bayes, Logistic Regression, Decision Trees, and Random forest and reported an F1-score of 72 for comments written solely in Arabic script, and 78 for comments written in Latin script both achieved using Logistic Regression. Likewise Soumeur et al. (2018) manually annotated a corpus of more than 25,000 comments collected from Facebook pages of 20 companies into positive, negative, and neutral classes. However, they translated all code-switched segments into Arabic words and transliterated words written in Latin script into Arabic script. They reported that a CNN model with a BoW representation achieved the best performance with an accuracy of 89.5%.

Unlike in the earlier mentioned work, in all our resources in this work, we use user-generated data without any transformation except for a simple automatic pre-processing done to reduce the size of the vocabulary. Furthermore, we take advantage of deep neural networks trained end-to-end to identify sentiments from unedited and code-switched user-generated comments written in Algerian in both Arabic and Latin scripts. Also, we experiment with two ways to improve the performance of our models, notably injecting the sentiment lexicons as background knowledge to a CNN model and augmenting the training data to overcome the problem of unbalanced data.

## 3. Linguistic Resources

### 3.1. Corpus creation and properties

To the best of our knowledge, there are no adequate freely available Algerian corpora annotated for sentiment analysis that would serve our purpose. We therefore created our own corpus. To do so, we compiled a list of 139 popular Algerian YouTube channels that span a wide range of genres from news, politics, sports, cooking, vlogs, product reviews, and TV-shows. We manually collected 50,000 comments or posts of different lengths, and removed all comments that were not written in Algerian, such as those written in Modern Standard Arabic (MSA) or another Arabic dialect such as Moroccan, Tunisian, Egyptian, etc. We also removed all comments that were entirely written in Latin script (French or Arabic written in Latin script), but we kept those written in mixed scripts –Arabic and Latin. Likewise, we kept all comments that were a mix of MSA, Berber, French and local Algerian Arabic. This decision is based on the fact that Algerian Arabic is a mixture of co-existing languages and language varieties written in non-standardised orthography both in Latin and Arabic scripts.

For example, (1) is a user-generated comment written entirely in Arabic script which displays code-switching at a word level between several varieties; exhibited in the use of the following words: MSA (عندك، الحق، عندي، و، عام، ما، حتى، مشكل – you have, right, I have, and, year, not, any, problem), French (لاماشين – washer), and local Algerian Arabic: (هاذي، ختها، وبزاف، مليحة، ملي، شريتها، معاها – this, like it, very, good, since, bought it, with it). Note that

in *formal* MSA, people might express the same meaning using other words.

(1) a. عندك الحق هاذي لاماشين عندي ختها وبزاف مليحة وعندي عام ملي شريتها وما عندي معاها حتى مشكل

b. You're right, I have the same washer and it is very good. I have had it for a year and have had no problems with it.

(2) a. لازم نبكو مع بعض ونفرحو مع بعض مالا حنا علاه نتبعو فيك غير لا ضحك نن حبيبتي نحبوك وكلش فيك نحبوه **bn courage bn continuation**.

b. We should support each other otherwise why are we following you, just to laugh! No dear, we love you and love everything in you, good luck!

Example (2) also displays code-switching, between MSA ( مع، بعض، ضحك، حبيبتي، فيك – with, each other, dear, in you), French ( نن – no) and local Algerian Arabic (لازم، نبكو، ونفرحو، مالا، حنا، علاه، نتبعو، غير، لا، نحبوك، وكلش، نحبوه – we must, we cry, we will be happy, so, we, why, follow, just, no, we love you, and everything, we love it). In addition to the code-switching at a word level, the user mixes Arabic and Latin scripts. Because local Algerian Arabic is a colloquial language with no standardised orthography, people use speech-based spelling including lots of spelling variations reflecting local or regional pronunciation. We leave the data unedited, with typos and misspelling of MSA words.

We applied the following cleaning procedure on the raw data. The encoding of Arabic characters was normalised so that equivalent characters were mapped to a single Unicode point. For example, in our data occurrences of "پ" denoting the letters "*v*" or "*p*" were substituted by "ب" (letter "*b*") because the former do not have equivalents in Arabic. All comments were anonymized, i.e. users' information was deleted manually and all mentions of people were automatically replaced by others, while keeping their context meaningful. Mentions of celebrities and political figures were kept, generic references, such as ختي (sister), خو، خويا (brother), and صاحبي، شريكي، حبيبنا، حبيبتي، حنونة (friend) were also kept. Long comments were trimmed and split where there was a clear split, both at sentence boundaries and when a user clearly expressed opinions on two different topics.

As for its statistical properties, after cleaning, our corpus consists of 36,120 unedited short colloquial comments or relatively short texts with an average length of 14.47 tokens or 74 characters, as is expected given the data source (social media). It comprises 522,890 tokens (lexical words, digits and emoticons, punctuation not included) and 78,482 unique tokens.

The corpus displays also the linguistic properties of Algerian Arabic, mentioned earlier, namely code-switching, spelling variations and spelling errors for MSA, hence increasing the data sparsity. Another property of our corpus is that it consists of discussions and sub-discussions—in essence short written multilogues. We kept the order of the comments exactly as found on the platform (YouTube) to keep a larger context. That is, a user may refer to different things at the same time in one short comment, such as comment on a video, comment on previous comments, and talk about personal experiences or something completely unrelated. Users also quote each other a lot, or quote segments from the video or from previous comments and comment on them.

### 3.2. Corpus annotation and statistics

In sentiment analysis, texts are commonly classified based on their polarity: either as positive (POS) or negative (NEG), or neutral (NEU). In our case, users comment on videos or give their feedback on something related or unrelated, they agree, disagree, give their own experience, or add new information. Therefore to be realistic and model the user-generated data at hand, we decided to add the class MIX for the cases where users combine polarity or add new information in addition of the three standard classes (POS, NEG, NEU).

More precisely, our annotation guidelines are as follows: use POS or NEG if it is understood from the comment that its generator is clearly expressing something positive or negative as in (3) and (4) respectively; use NEU if the comment does not bear any sentiment, for instance inserting a piece of information or a quote, or asking questions as in (5); otherwise use MIX for cases combining POS, NEG and/or NEU as in (6) where the user is positive about the video presenter and negative about previous comments. Two Algerian native speakers annotated the corpus described in 3.1., taking the users' perspectives into account.

(3) a. هه نموت عليك عندك عقلية فوورر

b. I love you I love your way of thinking.

(4) a. نيفو زيرو وش هاد تمسخير تزيدو توريوه

b. Low level what is this joke and you dare to show it!

(5) a. شحال فعمرك كي بديت

b. How old were you when you started?

(6) a. باين غيارين برك ياك يكرهو لي ناجح الله لا تربحهم انت طوب كملي

| | POS | NEG | MIX | NEU |
|---|---|---|---|---|
| Train | 7,330 | 5,017 | 8,357 | 5,682 |
| Dev | 1,285 | 559 | 1,187 | 581 |
| Test | 2,083 | 848 | 2,192 | 999 |
| Total | 10,698 | 6,424 | 11,736 | 7,262 |

Table 1: Distribution of comments over classes.

b. It is clear they are just jealous (of you) they hate someone who succeeds may God fail them, you are top carry on!

The final annotated corpus does not maintain balance between classes. Precisely, as shown in Table 1, we have 10,698 comments for POS, 6,424 for NEG, 11,736 for MIX, and 7,262 for NEU.

We measure the reliability of the human judgments by computing the inter-annotator agreement (IAA), corresponding to how often annotators made the same decisions (agreed) and how many times they made different decisions (disagreed). To this end, we shuffled the data and randomly selected 2,000 comments (annotated by the two annotators of the entire corpus). Their IAA computed using standard *Cohen's kappa coefficient* $\kappa$ is 0.75.

To gain more confidence on agreement, we also asked two additional Algerian native speakers to annotate the same 2,000 comments following the same guidelines as above. We calculated the inter-annotator agreement with four annotators: the two annotators of the entire data plus the two additional ones who were separately asked to annotate the previously mentioned sample of 2,000 comments. All annotators worked independently from each other after agreeing on annotation guidelines. We used two IAA measures: *Krippendorff's alpha* and *Fleiss' kappa*, and obtained the following scores: *Krippendorff's alpha* $\alpha = 0.89$, and *Fleiss' kappa* $\kappa = 0.89$. These values provide strong evidence for good agreement. The per-class tallies shown in Figure 1 indicate that annotators disagreed more on the MIX and NEU classes.

### 3.3. Sentiment lexicons

We constructed small positive and negative sentiment lexicons using the entire labelled corpus. To do so, we first identified the 2,000 terms that are the most correlated with each of the positive (POS) and the negative (NEG) classes using the chi-squared test. We looked for both unigrams and bigrams. We then curated these lists, and manually added a list of most common Algerian positive and negative words and emoticons that were not already in the lists. This resulted in a positive lexicon of 917 entries and a negative lexicon containing 647 entries. The lexicons reflect the nature of the Algerian language and contain both borrowings and code-switched entries[1].

[1]The described resources are available on request form the authors.

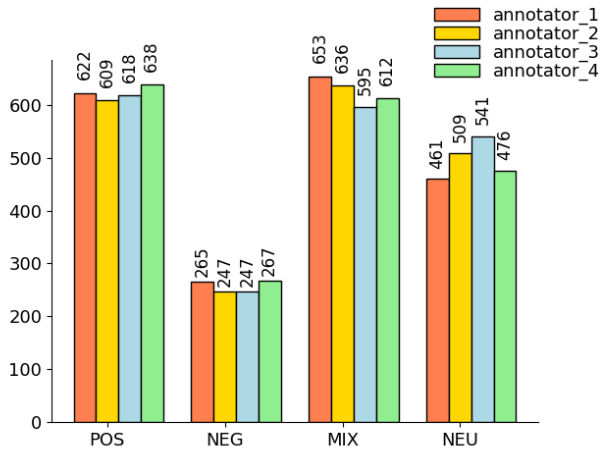



Figure 1: Per-class overview of the agreements between four native annotators.

## 4. Models

We frame the task of identifying sentiments from social media as a text classification problem. That is, given a comment (text segment of any length) predict the sentiment among the predefined classes. We compare the performance of four models: (1) a Linear Support Vector Machine (SVM) with bag-of-words representations of the data as features. Three Deep neural models with different configurations, (2) using Convolution Neural Network (CNN), (3) using Long Short-Term Memory (LSTM), and (4) using Bi-directional Long Short-Term Memory (BiLSTM) as summarised in Figure 2.

- **CNN** (in dark orange), is comprised of two passes, one creating word representations from characters, and one taking a sequence of these representations and classifying it. To construct word representations, we first use a character embedding layer mapping each of the 430 possible characters to a 50-dimensional representation. Then we use a convolution layer with filter size 3 followed by ReLu activation and max-pooling in the time domain. The sentence-level analysis is composed of two convolution layers. The first layer has 50 features and the second has 30. Both layers use a filter size of 3 with a dropout rate of 15%, followed by ReLu activation. This architecture is similar to that proposed by Adouane et al. (2019), but uses different hyper-parameters.

- **BiLSTM** (in yellow) takes word embeddings with a vocabulary size of 326,847 and embeddings dimension of 100. It consists of one BiLSTM layer with 100 units with a dropout rate of 10% followed by a global max pooling layer.

- **LSTM** (in blue) takes as input the same word embeddings as in the BiLSTM. It is composed of 2 LSTM layers where the first layer has 200 units and the second has 100 units with dropout rate of 20% between the layers.

In each of the three configurations (CNN, BiLSTM and LSTM), the final stage of the DNN is a dense layer (in

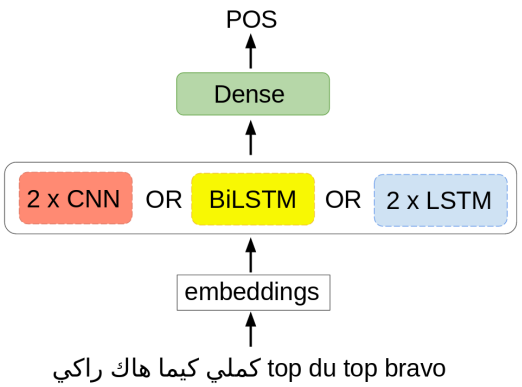



Figure 2: Model architectures. Each of the 3 DNN models takes a user comment as input and outputs a class from the set of sentiment classes POS, NEG, NEU, and MIX.

| | Train | Dev | Test | Total |
|---|---|---|---|---|
| #Comments | 26,386 | 3,612 | 6,122 | 36,120 |
| #Tokens | 367,483 | 57,874 | 97,533 | 522,890 |
| #Unique tokens | 63,117 | 16,664 | 24,274 | 78,482 |

Table 2: Corpus statistics with distribution over the 3 sets. Note that overlapping #Unique tokens between sets were counted only once, hence the total is less than the sum of the #Unique tokens of all sets.

green) with a softmax activation layer which maps their outputs to sentiment classes. All neural models are trained end-to-end for 60 epochs using a batch size of 64 for CNN and 128 for LSTM and BiLSTM, and Adam optimiser. Gradients with a norm greater than 5 are clipped.

## 5. Experiments and Results

In order to evaluate the performance of the models on identifying sentiments, we shuffle the data and randomly pick 6,122 samples as a test set, 3,612 samples as a development set and the remaining 26,386 as a train set. The hyper-parameters mentioned in Section 4. were fine-tuned on the development set. Table 1 shows the number of comments for each set by sentiment class. Table 2 gives a summary of the number of tokens and comments for each set.

As shown in Table 3, in terms of overall accuracy, the BiLSTM outperforms other models with an accuracy of 66.78%, compared to LSTM with 66.67%, CNN with 65.37% and SVM with 61.65%. In terms of macro F1, our CNN is the best performing model. Nevertheless, because the data is unbalanced, we are more interested in the performance of the models for each sentiment class.

As you can see in Figure 3a, BiLSTM (in yellow) and LSTM (in blue) are too biased to the two majority classes. They perform well on POS and MIX classes achieving a F-score of 95.77 and 95.18 respectively on POS, and 93.46 and 93.13 on MIX. However, they perform poorly on the two minority classes, namely NEG and NEU with an F-score of only 2.46 and 2.07 on NEG, and 2.41 and 4.40 on NEU.

| Model | Accuracy (%) | Macro F1 |
|---|---|---|
| BiLSTM | **66.78** | 48.53 |
| LSTM | 66.67 | 48.70 |
| CNN | 65.37 | **60.17** |
| SVM | 61.65 | 58.50 |

Table 3: Overall accuracy and macro F1 of the models.

At the same time, the CNN (in orange) performs worse than BiLSTM and LSTM with an F-score of 78.85 on POS and even less on MIX achieving only 66.29 F-score. But it performs much better on NEG and NEU with an F-score of respectively 39.11 and 56.42.

Surprisingly the SVM (in light green) performs the best for NEG achieving an F-score of 48.46, and less than the CNN on NEU with 53.60 F-score. Still the SVM performs the worst on POS and MIX with only 73.73 and 58.19 F-score respectively.

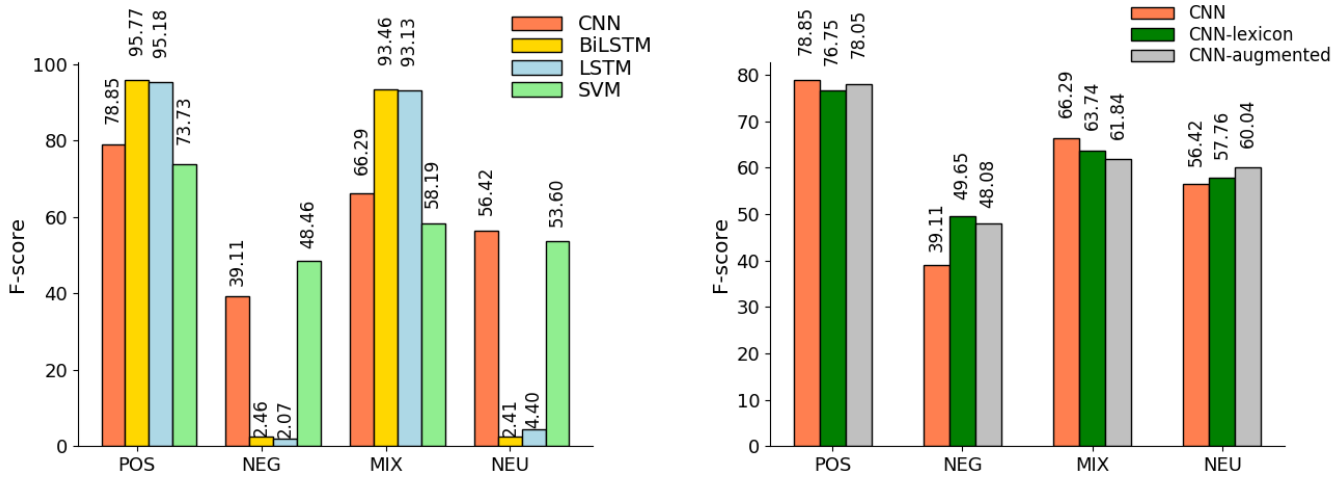
With regards to Precision and Recall, (see Appendix) the CNN achieves a reasonably good Precision and Recall on all classes along with the SVM in comparison to the LSTM and BiLSTM.

Summing up the results, the CNN outperforms LSTM-based models and SVM for all classes except for NEG where it is outperformed by the SVM. This could be explained by the fact that deep neural models —based on stacking layers with many non-linear transformations— perform better with more data as they are able to learn increasingly more abstract representations. That is, learning the underlying hierarchical structures from the data (with CNN) better fits the data at hand compared to modeling its structures sequentially (LSTM). In fact, the CNN is the only tested model that accesses information at a character level, thus potentially making it robust to data sparsity in the form of misspellings, etc. On the other hand, the Linear SVM —a binary classifier in its core— can not handle large unbalanced data with multiple classes. Unexpectedly it outperforms CNN for NEG class both in Precision and Recall (see Appendix). One possible explanation is that the number of NEG samples is not enough for the CNN to learn patterns from the noisy sparse data, and hence it is hard to extract useful features from the train data and generalise them to the test data.

In the following we experiment with other ways to improve the performance of the CNN model.

### 5.1. Injecting the sentiment lexicons to the CNN

One way to improve the CNN performance is to add information from the sentiment lexicons, described in Section 3.3., as background knowledge. To this end, we encoded the lexicons and injected them to our CNN model. As shown in Figure 3b, the lexicons boosted the performance of the CNN (referred to as CNN-lexicon in dark green) by 10.54 F-score points gain on NEG and 1.34 on NEU. Nonetheless, its F-score dropped by 2.1 points on POS and 2.55 on MIX.




(a) (b)

Figure 3: F-score of each model per sentiment class.

### 5.2. Adding augmented data to the CNN

To deal with the unbalanced data, we experimented with augmenting the number of minority classes by duplicating all NEG and NEU samples in the training set without changing the test set. We retrained the CNN-lexicon, see Section 5.1., with the augmented data and referred to it as CNN-augmented. The results in Figure 3b (in grey) indicate that the data augmentation has a positive effect on NEU with a gain of 2.28 points in F-score mainly due to boosting its Recall with 8.41 percentage points. On the other hand, its Precision on MIX increased while its Recall dropped by 8.17 points (see Appendix). The same trend is observed for NEG where Recall increased by 4.49 points, achieving the best results, Precision dropped by 5.22 points. On POS, the augmented data has a slightly positive effect both in terms of Precision and Recall. Overall, there is a trade-off between the Recall and Precision of the sentiment classes.
It is worth mentioning that we also experimented with pre-trained embeddings as opposed to learning embeddings along with a model itself (not included). For this we trained word embeddings using FastText (Joulin et al., 2016) on a large user-generated data (Adouane et al., 2019) and plugged them into the deep neural models in Section 4.. Nevertheless, the addition was not helpful, *i.e.* the difference was not clear compared to the models without the pre-trained embeddings.

### 5.3. Error analysis

Looking to the confusion matrices (not shown for concision), we found that a common confusion of the CNN models with different setups (CNN, CNN-lexicon and CNN-aug) is between MIX and POS with 349, 424 and 440 cases respectively. Whereas LSTM and BiLSTM confuse mostly between NEU and NEG with 908 and 930 cases. This is reflected in the above reported results. Confusing MIX and POS could be related to the fact that the MIX class is defined as any combination of the rest of the classes, namely NEG, NEU, and POS.

(7) a. نتي هايلا وعقليتك روعة فيك ديفو تعاودي بزاف لهدرة

b. You are gorgeous and your way of thinking is great, your flaw is that you repeat yourself a lot

(8) a. راكي فوور وموتو يا العديان كنترا فيكم

b. You are great, die you enemies, we are against you

(9) a. ديري ريجيم تنقصي تبهاي شوية

b. Follow a diet to loose some weight and you'll become a bit gorgeous

(10) a. فيديو ختامو مسك عيفتيلي قلبي

b. What a great video ending, disgusting!

(11) a. وشحال عجبتني تاع اكرهكي كرها شديدا

b. I love (when you said) I hate you so much.

We will discuss in what follows some of the misclassified examples by the CNN-lexicon model. The example in (7) is classified as POS instead of MIX. One potential solution to overcome this issue is to do a fine-grained sentiment identification at a segment level instead of comment-level.

Also it is not easy to identify the NEU class. The reason is that there are many occurrences of irony, humour, sarcasm, and metaphors in our corpus. This means that the sentiment

is sometimes not conveyed by the literal meaning of words, which are *neutral* when taken individually, but rather need a pragmatic interpretation which is only accessible when taking into account the larger cultural context. Additionally, the class can depend on the perspective taken (that is to say commenting on the content of the video, or on previous comments, or on users' (un)related personal experience). The example in (8) is misclassified as POS while human annotators classified it as MIX taking into account the user's perspective where s/he likes the video and is against the previous comments criticising it.

The example in (9) is classified as NEU by the CNN (it does not have an explicit sentiment) and MIX by human annotators with the interpretation of implicit NEG + advice. Whereas the example in (10) is NEG (with irony) but the CNN classified it as POS.

Moreover, as mentioned earlier, there are many instances of quoting, for instance referring to something in a video or mentioning something said before. Such cases could be interpreted either as POS by someone who did not see the video or read previous comments (lack of context) or as NEU by someone who knew the context. For instance, the user in example (11) liked that the video presenter said that 'she hates someone so much'. But the quote itself is a negative statement attributed to the video presenter, and thus the example is classified as NEG by the CNN.

Analysing the confusions of the CNN-lexicon shows that some of them are in line with the inter-annotator agreement in Figure 1 where the variation is more on the MIX and NEU classes.

## 6. Conclusions and Future Work

We presented in this work our new manually built linguistic resources for Algerian: a corpus which consists of more than 36,000 user-generated comments annotated for sentiments, along with sentiment lexicons of positive and negative words. We then described our models utilised to automatically identify sentiments from the user-generated comments. We discussed the performance of each model per sentiment class measured as F-score.

We found that the CNN model trained end-to-end fits better our data across the predefined sentiment classes compared to SVM, BiLSTM and LSTM models. Including the lexicons as background knowledge boosted the performance of the CNN even more on minority classes (NEG and NEU) to different extents. Moreover, analysing the confusion matrix showed that it is quite challenging to distinguish between MIX and NEU classes.

Our main contribution in this paper is building new resources for sentiment analysis from user-generated Algerian comments. We also presented experiments to identify the sentiments using our newly built resources. We believe that the results of our experiments can be used as a baseline for future research for Algerian sentiment analysis.

As future improvements, we plan to do fine-grained classification on the MIX class, i.e. identify sentiments at the segment level instead of at the comment-level. A difficulty to overcome is that a comment may be globally negative while a large segment can still be positive, as in example (10). Also being able to identify where in a comment a user has switched from being negative to positive (or vice versa) can be challenging.

Additionally, we will experiment further with multitask learning. For this, we plan to (i) jointly learn sentiment classification as main task and lexicon prediction as auxiliary task as presented in Barnes et al. (2019), and (ii) train our best performing model jointly with other tasks and investigate whether and where there will be any performance gain.

## 7. Acknowledgements

The research reported in this paper was supported by a grant from the Swedish Research Council (VR project 2014-39) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg. We are grateful to the two annotators. We also thank the anonymous reviewers for their useful comments.

## Appendix: Precision and Recall of models

| Model | POS | NEG | MIX | NEU |
|---|---|---|---|---|
| BiLSTM | **95.86** | 2.24 | **93.39** | 2.67 |
| LSTM | 95.64 | 1.92 | 92.84 | 4.75 |
| SVM | 75.59 | **55.47** | 52.91 | **57.16** |
| CNN | 78.16 | 40.87 | 65.97 | 56.08 |
| CNN-lexicon | 73.65 | 53.47 | 66.37 | 55.20 |
| CNN-augmented | 75.53 | 45.61 | 69.72 | 54.95 |

Table 4: Precision (%) of each model per sentiment class.

| Model | POS | NEG | MIX | NEU |
|---|---|---|---|---|
| BiLSTM | **95.68** | 2.71 | **93.52** | 2.20 |
| LSTM | 94.72 | 2.24 | 93.43 | 4.10 |
| SVM | 71.95 | 43.02 | 64.65 | 50.47 |
| CNN | 79.55 | 37.50 | 66.61 | 56.76 |
| CNN-lexicon | 80.12 | 46.34 | 63.74 | 57.76 |
| CNN-augmented | 80.75 | **50.83** | 55.57 | **66.17** |

Table 5: Recall (%) of each model per sentiment class.

## 8. Bibliographical References

Abdul-Mageed, M. and Diab, M. T. (2012). Awatif: A multi-genre corpus for modern standard arabic subjectivity and sentiment analysis. In *LREC*, volume 515, pages 3907–3914. Citeseer.

Adouane, W., Bernardy, J.-P., and Dobnik, S. (2019). Normalising non-standardised orthography in Algerian code-switched user-generated data. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 131–140, Hong Kong, China, November. Association for Computational Linguistics.

Aly, M. and Atiya, A. (2013). 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)*, volume 2, pages 494–498.

Barnes, J., Touileb, S., Øvrelid, L., and Velldal, E. (2019). Lexicon Information in Neural Sentiment Analysis: A

Multi-task Learning Approach. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 175–186, Turku, Finland.

Dahou, A., Xiong, S., Zhou, J., Haddoud, M. H., and Duan, P. (2016). Word embeddings and convolutional neural network for arabic sentiment classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2418–2427, Osaka, Japan.

Elouardighi, A., Maghfour, M., and Hammia, H. (2017). Collecting and processing arabic facebook comments for sentiment analysis. In *International Conference on Model and Data Engineering*, pages 262–274. Springer.

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

Guellil, I., Adeel, A., Azouaou, F., and Hussain, A. (2018). SentiALG: Automated Corpus Annotation for Algerian Sentiment Analysis. In *International Conference on Brain Inspired Cognitive Systems*, pages 557–567. Springer.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

Jerbi, M. A., Achour, H., and Souissi, E. (2019). Sentiment Analysis of Code-Switched Tunisian Dialect: Exploring RNN-Based Techniques. In *International Conference on Arabic Language Processing*, pages 122–131. Springer.

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2016). Bag of Tricks for Efficient Text Classification. *arXiv*, arXiv:1607.01759, 07.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar.

Lei, Z., Yang, Y., Yang, M., and Liu, Y. (2018). A multi-sentiment-resource enhanced attention network for sentiment classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 758–763. Association for Computational Linguistics.

Mataoui, M., Zelmati, O., and Boumechache, M. (2016). A proposed lexicon-based sentiment analysis approach for the vernacular Algerian Arabic. *Res. Comput. Sci*, 110:55–70.

Medhaffar, S., Bougares, F., Estève, Y., and Hadrich-Belguith, L. (2017). Sentiment analysis of tunisian dialects: Linguistic ressources and experiments. In *Proceedings of the third Arabic natural language processing workshop*, pages 55–61.

Nabil, M., Aly, M., and Atiya, A. (2015). Astd: Arabic sentiment tweets dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2515–2519.

Qian, Q., Huang, M., Lei, J., and Zhu, X. (2016). Linguistically regularized LSTMs for sentiment classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1679–1689, Vancouver, Canada.

Rushdi-Saleh, M., Martín-Valdivia, M. T., Ureña-López, L. A., and Perea-Ortega, J. M. (2011). Oca: Opinion corpus for arabic. *Journal of the American Society for Information Science and Technology*, 62(10):2045–2054.

Salameh, M., Mohammad, S., and Kiritchenko, S. (2015). 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*, pages 767–777.

Shin, B., Lee, T., and Choi, J. D. (2016). Lexicon integrated CNN models with attention for sentiment analysis. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 149–158, Copenhagen, Denmark.

Soumeur, A., Mokdadi, M., Guessoum, A., and Daoud, A. (2018). Sentiment Analysis of Users on Social Networks: Overcoming the Challenge of the Loose Usages of the Algerian Dialect. *Procedia computer science*, 142:26–37.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

Tai, K. S., Socher, R., and Manning, C. D. (2015). Improved Semantic Representations FromTree-Structured Long Short-Term Memory Networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1556–1566, Beijing, China.

Turney, P. D. (2002). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 417–424. Association for Computational Linguistics.

Zbib, R., Malchiodi, E., Devlin, J., Stallard, D., Matsoukas, S., Schwartz, R., Makhoul, J., Zaidan, O. F., and Callison-Burch, C. (2012). 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*, pages 49–59. Association for Computational Linguistics.