TUNIZI attention-based sentiment analysis with token-level features

TELMINI Fyras
ENSAE DSSA
mohamed.telmini@ensae.fr

TOUZI Moahmed
ENSAE DSSA
mohamed.touzi@ensae.fr

Abstract

Sentiment prediction in textual data remains a pertinent challenge in modern natural language processing (NLP) research, particularly in the context of spontaneous spoken language and under-represented dialects. Tunizi, an Arabic dialect spoken in Tunisia, exemplifies a case involving both of these issues, characterized by its under-studied nature, irregular expressions, ambiguous syntax, and frequent code-switching with French and English. This study aims to advance the current understanding of Tunizi sentiment analysis by first introducing a novel fine-grained dataset and providing the tools for streamlined contribution. Subsequently, we develop and train an attention-based sentiment analysis model on this dataset. Lastly, we investigate the impact of incorporating fine-grained text data in training by comparing the performance of multiple versions of the attention-based sentiment analysis model, studying the potential benefits of this approach for under-represented dialects. Code for our work can be found in our Github[1] repository.

1 Background and Related Work

Arabizi dialects, as defined by (Mulki et al., 2018), represent a novel approach to writing Arabic using Roman script characters and numbers, which emerged within social media platforms across the Arab world. This writing style lacks a clear syntax and primarily relies on phonetic representations of words, resulting in a high degree of variability in the spelling of similar words. Tunizi is a variant of Arabizi that corresponds to the Tunisian spoken dialect (Dinkar* et al., 2020), adding the complexities of the Tunisian dialect, such as code-switching, to the mix. Previous studies on Tunizi sentiment analysis, such as those by (Messaoudi et al., 2020) included a significant contribution by (Fourati et al., 2020) called TUNIZI-V1, which consists of 9,210 sentences labeled as either positive or negative. Subsequent work by (Fourati et al., 2021) expanded the dataset to 100,000 sentences, labeled as positive, negative, or neutral.

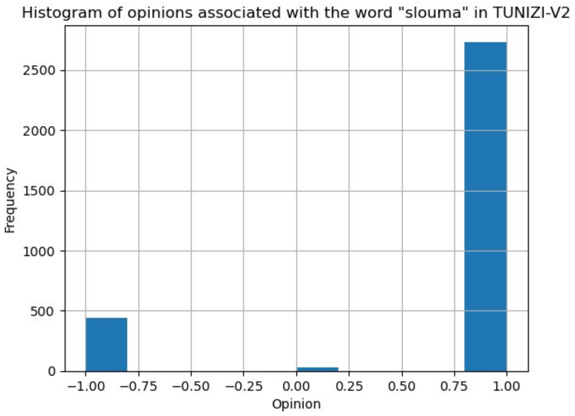

Figure 1: Histogram of opinions associated with the word ”slouma” in the TUNIZI-V2(Fourati et al., 2021) Dataset

A prevalent issue in these contributions is the absence of fine-grained representations in individual sentences, which can result in potential biases for named entities that correlate with negative opinions. This may cause opinion analysis systems to mistakenly predict a negative association with the entity itself rather than the way it is mentioned. An example of this is shown in Figure 1 where ”slouma” is a named entity heavily (if not entirely) associated to the political personality ”Slim Riahi”(Wikipedia, 2023). We suspect that this is due to how the corpus of (Fourati et al., 2021) has been constructed, specifically the usage of Facebook comment scraping from the pages of such known figures. In the first part of our work, we build upon these previous contributions by introducing a novel fine-grained dataset containing

[1] https://github.com/fyrastelmini/TunABSA

1,000 manually annotated sentences. We also provide an intuitive annotation script to facilitate easier contributions. The second part of our work draws inspiration from the multiple contributions of (Garcia et al., 2019). specifically the replication of their Bi-GRU+Attention network, which we truncate at the sentence level. We then conduct a comprehensive experimental protocol to verify the impact of pre-training this model on intermediate fine-grained labels representing the entity and polarity word associations.

2 Datasets

2.1 Fine-Grained Dataset

This work relies on a set of fine- and coarse-grained opinion annotations gathered from the ”I4D iCompass Social Media Sentiment Analysis for Tunisian Arabizi Dataset” (Zindi Africa, 2023) provided in a Zindi[2] competition. The original competition training dataset contains 70,000 sentences labeled as 1 (positive), 0 (neutral), or -1 (negative). We only kept the sentences that had a positive or negative label for our fine-grained dataset, as the neutral labels would have caused issues for the labeling process. Additionally, we extracted only the sentences containing 10 words or less to avoid extreme sentence lengths since we are only doing sentence-level opinion analysis. The dataset was then labeled using our comprehensive labeling script, which we made accessible in our GitHub repository. Since the labeling was done manually by us, we insist on the inherent experimental nature of the obtained data. We hope that further work builds upon the TUNIZI-V2 dataset (Fourati et al., 2021) by applying our labeling protocol to it. We will refer to this as the **TUNIZI token-level** dataset in the rest of this paper.

2.2 Filtered and Calibrated TUNIZI-V2 Dataset Sample

Another dataset we used consisted of a sample of 60,000 labeled sentences from the TUNIZI-V2 dataset. We filtered out the neutral-labeled sentences and those that had more than 10 words, as mentioned above. Additionally, we removed 15% of the positively-labeled sentences to obtain a more calibrated 54% positive labels and 46% negative labels split. We will refer to this as the **Sampled TUNIZI-V2** dataset in the rest of this paper.

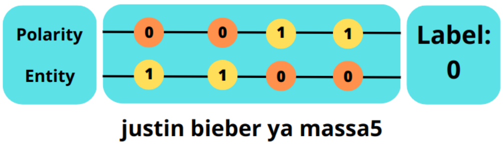

Figure 2: Structure of an annotated sentence

2.3 Labeling the fine-grained dataset

The final **TUNIZI token-level** Dataset comprises 1000 sentences with strong opinion content, each annotated at two levels of granularity as illustrated in Figure 2. The first level represents the polarity of the token, characterizing the opinion expressed at the word level, irrespective of its positive or negative nature. The second level concerns the presence of an entity, indicating the subject of the sentence. This polarity-entity separation replicates the approach used by (Garcia et al., 2019). Additionally, each sentence retains its original label from the source dataset, denoted by a value in 0,1, with 1 indicating a positive opinion and 0 indicating a negative one. We hope that future research can build upon our work and apply our labeling protocol to the (Fourati et al., 2021) TUNIZI-V2 dataset. Overall, the canonical representation of each sentence is $\mathbf{x}^{(i)} = \left(x_1^{(i)}, \ldots, x_n^{(i)}\right)$ where $\mathbf{x}^{(i)}$ is a sentence and $x_j^{(i)}$ is the j-th token within it. The target labels are canonically represented in two levels:

1. Token-level Labels: $y_{\text{Tok},j}^{(i)} = \begin{pmatrix} y_{\text{Pol},j}^{(i)} \\ y_{\text{Ent},j}^{(i)} \end{pmatrix}$

   Where $y_{\text{Tok},j}^{(i)}$ is the j-th token of the sentence i, $y_{\text{Pol},j}^{(i)}$ is value associated to its polarity, and $y_{\text{Ent},j}^{(i)}$ is associated to the entity association of it.

2. Sentence-level Label: $y_{\text{Label}}^{(i)}$ is the overall label of the sentence

The **TUNIZI token-level** dataset contains 71% Positive sentence-level labels and 29% Negative sentence-level labels. This is not ideal, but we chose to keep it as is due to the considerable effort taken in manually labelling it. We also argue that the dataset mainly serves to capture the information at the token-level rather than at the sentence-level.

[2] Zindi Africa: zindi.africa

3 Models

Our experiments are heavily based on the ones presented in (Garcia et al., 2019), which demonstrated promising results in the field of multi-modal opinion classification. We aimed to reproduce the described Bi-GRU + Self-attention architecture and training procedure, considered to be the best performing one in the paper's first conducted experiment. For that, we define two separate model architectures in order to study the effect of token-level labels on the overall performance of our model. Both models take tokenized text inputs from a pretrained BERT(Devlin et al., 2019) Tokenizer, found here[3]. The recent work by (Haddad et al., 2023) seemed to be the most appropriate tokenizer to use, but we couldn't access it. Thus we settled with the tokenizer we found on huggingface, although there was no clear indication of it's source or a proper way to cite it.

3.1 BI-GRU classifier Model

This first architecture serves as the basis for our pre-training protocol over the token-level features we extracted. This model consists of two main parts:

1. Embedding layer + Bi-directionnal GRU(Cho et al., 2014) layer: This part serves to capture the polarity-entity information present in our tokenized data. The embedding layer corresponds to a simple embedding over the vocabulary space of our BERT tokenizer.

2. Fully connected classification layers for the target label vectors: We use two fully connected layers successively for each label vector output.

A dropout layer is defined between the Embedding layer and the Bi-directionnal GRU layer in order to avoid overfitting. This network serves mainly to train both the embedding and the Bi-directionnal GRU layer in order for them to serve as initializations for an attention-based architecture. Figure 3 shows this architecture clearly. We will refer to this model as **Bi-GRU classifier** model.

3.2 BI-GRU+Attention classifier Model

This model inherits from the architecture of the previous one, by keeping both the Bi-GRU and the Embedding layer (and the dropout layer in between). After which a self-attention(Vaswani et al., 2017) Layer is added. A final fully-connected layer is also added to allow for binary classification of the labeled sentiment-analysis data. Figure 4 shows this architecture clearly.We will refer to this model as **Bi-GRU+attention** model.

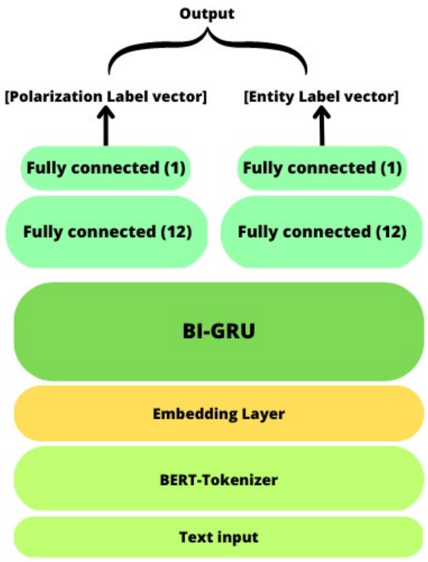

Figure 3: BI-GRU classifier Model

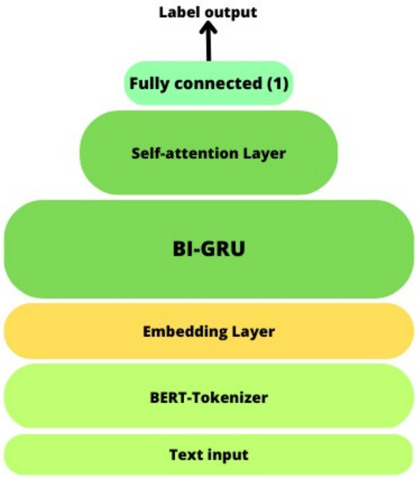

Figure 4: BI-GRU+Attention classifier Model

4 Experimental protocol

We conducted our network training in two successive phases:

1. We pre-trained the **Bi-GRU classifier** model over the token-level features from our **TU-**

[3] https://huggingface.co/ziedsb19/tunbert_zied

NIZI token-level dataset. Since this problem is analogous to a multi-class classification problem over the labels of the tokens, we opted for using **binary cross-entropy** loss. The learning rate was fixed to 10^{-3} accross all variations of this network. These training hyperparameters stayed constant over all experiments.

2. Afterwards, We train three variations of the **Bi-GRU+attention** model over our **Sampled TUNIZI-V2** Dataset. The three versions share the exact same architecture and vary only in initialization. The first version corresponds to the baseline and is trained over the labelled sentences from scratch, we call it **Baseline**. The second version loads the weights of the encoding layer and the BI-GRU layer from the **Bi-GRU classifier** model that has been pretrained, then starts training from that initialization, we call it **Pretrained**. The third version also loads those weights but freezes the two layers during training, we call it **Pretrained + Frozen**. We used binary cross-entropy loss as it outperformed all other loss functions we tried. We fixed the learning rate to 10^{-4} as most networks overfitted early during the training and a lower learning rate was needed to correctly navigate the loss landscape. In this case aswell these training hyperparameters stayed constant over all experiments.

Early stopping over validation accuracy was used in all networks. **Bi-GRU+attention** models ran for an unequal number of epochs, often between 5 and 15. The **Bi-GRU classifier** models ran for approximately 2000 epochs each before early stopping. The batch size of our networks was fixed to 256 accross all models and experiments.

On the side of the data, the train-test split ratio was 20%. Test data was generated at the start of each training sequence with stratification over the final labels enabled and was kept the same over each sequence of training.

The experiments were ran sequentially, with four variations of the BI-GRU layers sizes (8,16,32 and 64 GRU units). In each sequence, the size of the BI-GRU layer is defined. Then a corresponding instance of the **Bi-GRU classifier** network was instantiated and trained. Afterwards, three instances of the **Bi-GRU+attention** are intantiated and trained as explained above. We chose to vary the size of the BI-GRU layer because we concider it to be the main focus of our pre-training procedure. And we wanted to make sure our results arent purely cause by the BI-GRU layer sizes. We reported the **F1 scores** and **AUC**[4] for each model, as we found these scores to be more informative than simple accuracy. Our full experimental protocol is available on our repository for reproduction.

5 Results and analysis

The table1 sums up the results of the conducted experiments: The table shows us that accross all variations of models and pre-training, we fail to observe an improvement of the baseline model. This isnt counter-intuitive, as our fine-grained dataset is extremely small when compared to the 60 000 labelled sentences corpus. A definitive conclusion would the discarding of the variants with frozen pretrained layers, as this only seems to worsen the model's ability to learn. The results over the pre-training still cannot be conclusive due to the data unbalance between the pre-training on the token-level features and the training over the sentence labels. A pre-training done on a more complete corpus of token-level labelled data then on the sentence labels of that same data would be a conclusive experiment to check the validity of this approach. Due to the inherent experimental nature of the token-level labels dataset, we chose to concider this approach to be out of the scope of this paper. Nevertheless, we want to mention how the overall performance of the baseline model is quite remarkable in terms of metrics.

	Baseline		**Pretrained**		**Pretrained +Frozen**	
	F1	**AUC**	**F1**	**AUC**	**F1**	**AUC**
GRU units = 8	0.8518	0.8405	0.8434	0.8338	0.6776	0.5873
GRU units = 16	0.8476	0.8365	0.8450	0.8336	0.6755	0.5182
GRU units = 32	0.8475	0.8398	0.8456	0.8382	0.6429	0.6193
GRU units = 64	0.8487	0.8413	0.8448	0.8328	0.7008	0.4998

Table 1: Experiment results.

6 Discussion/Conclusion

This study covers an experimental protocol that studies the importance of token-level polarity and entity features in attention-based sentiment analysis. We encourage researchers to apply our method to a larger corpus of token-level labelled data to achieve greater performance gains with other

[4] AUC: Area-Under-Curve metric

model architectures. It's worth noting that while we attempted to remove bias and ambiguity from the TUNIZI data, we did not address the inherent challenge of code-switching present within it. Future work could explore approaches such as those presented in (Colombo et al., 2021a; Chapuis* et al., 2020), which propose novel loss functions that account for the problem of code-switching. Moving forward, we plan to build upon our findings and utilize the extracted emotions to improve the conditioning of sentence generation. By incorporating emotion recognition into the language generation process, we can enhance the ability of virtual conversational agents to generate more personalized and engaging responses (Mabrouk et al., 2021; Colombo, 2021; Colombo et al., 2021b; Jalalzai* et al., 2020; Colombo* et al., 2019).

References

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Hala Mulki, Hatem Haddad, and İsmail Babaoğlu. 2018. Modern trends in arabic sentiment analysis: A survey. *TAL Traitement Automatique des Langues*, 58:15.

Pierre Colombo*, Wojciech Witon*, Ashutosh Modi, James Kennedy, and Mubbasir Kapadia. 2019. Affect-driven dialog generation. *NAACL 2019*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Alexandre Garcia, Pierre Colombo, Slim Essid, Florence d'Alché Buc, and Chloé Clavel. 2019. From the token to the review: A hierarchical multimodal approach to opinion mining.

Hamid Jalalzai*, Pierre Colombo*, Chloé Clavel, Éric Gaussier, Giovanna Varni, Emmanuel Vignon, and Anne Sabourin. 2020. Heavy-tailed representations, text polarity classification & data augmentation. *NeurIPS 2020*.

Emile Chapuis*, Pierre Colombo*, Matteo Manica, Matthieu Labeau, and Chloe Clavel. 2020. Hierarchical pre-training for sequence labelling in spoken dialog. *Finding of EMNLP 2020*.

Tanvi Dinkar*, Pierre Colombo*, Matthieu Labeau, and Chloé Clavel. 2020. The importance of fillers for text representations of speech transcripts. *EMNLP 2020*.

Abir Messaoudi, Hatem Haddad, Moez Ben HajHmida, Chayma Fourati, and Abderrazak Ben Hamida. 2020. Learning word representations for tunisian sentiment analysis.

Chayma Fourati, Abir Messaoudi, and Hatem Haddad. 2020. Tunizi: a tunisian arabizi sentiment analysis dataset.

Pierre Colombo, Emile Chapuis, Matthieu Labeau, and Chloé Clavel. 2021a. Code-switched inspired losses for spoken dialog representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8320–8337, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chayma Fourati, Hatem Haddad, Abir Messaoudi, Moez BenHajhmida, Aymen Ben Elhaj Mabrouk, and Malek Naski. 2021. Introducing a large tunisian arabizi dialectal dataset for sentiment analysis. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 226–230.

Pierre Colombo. 2021. *Learning to represent and generate text using information measures*. Ph.D. thesis, (PhD thesis) Institut polytechnique de Paris.

Aymen Ben Elhaj Mabrouk, Moez Ben Haj Hmida, Chayma Fourati, Hatem Haddad, and Abir Messaoudi. 2021. A multilingual african embedding for faq chatbots. *ArXiv*, abs/2103.09185.

Pierre Colombo, Chloe Clavel, and Pablo Piantanida. 2021b. A novel estimator of mutual information for learning to disentangle textual representations. *ACL 2021*.

Zindi Africa. 2023. AI4D iCompass Social Media Sentiment Analysis for Tunisian Arabizi Competition Data. [Online; accessed 18-March-2023].

Wikipedia. 2023. Slim riahi — Wikipedia, the free encyclopedia. [Online; accessed 18-March-2023].

Hatem Haddad, Ahmed Cheikh Rouhou, Abir Messaoudi, Abir Korched, Chayma Fourati, Amel Sellami, Moez Ben Hajhmida, and Faten Ghriss. 2023. Tunbert: Pretraining bert for tunisian dialect understanding. *SN Computer Science*, 4.