1 Sentence-Level Evidence Embedding for Claim Verification with Hierarchical Attention Networks

The authors propose an end-to-end hierarchical attention network for claim verification. This approach consists of a sentence encoder, a coherence attention module, an entailment attention module and an output layer.

The sentence encoder module consists two GRU-based recurrent neural networks. One encodes the claim, while the other encodes the evidence sentences. The assumption they make is evidence sentences should be topically coherent with the claim. They encode this knowledge through a biaffine transformation. This transformation takes care of global coherence, which addresses topical consistency of the set as a whole and local coherence, which is measure as the consistency of each sentence in relation to another sentence. To take into account the claim as well, they use a gating unit with trainable parameters.

For the entailment module, they combine the coherence attention outputs with the encoded claim using concatenation, element-wise product and element-wise difference in a linear layer. Again, they apply attention over the original sentences using these joint representations. These final representations are used in conjunction with the encoded evidence to produce the final outputs of the model.

Before fine-tuning the model, they pretrain the coherence and entailment modules. The coherence module is pretrained using a large margin objective, while the entailment module is trained using the SNLI dataset. Their model achieves better results than all the baselines (CNN, LSTM, SVM, DeClarE).

2 A Context-Aware Approach for Detecting Worth-Checking Claims in Political Debates

This paper presents a new claim-checkworthiness dataset, built from manually-annotated claims from US presidential debates. Each of the claims in these debates was annotated using information from reputable fact-checking sources. For solving this problem they employ different kinds of sentence-level, contextual and mixed features such as:

- previous state-of-the-art model features
- sentiment
- named entities
- linguistic features
- metadata
- segment sizes
- topics
- discourse features
- contradictions
- embeddings

Combined with the aforementioned features, they use a traditional machine learning approach, and a deep learning approach. They classify sentences as positive if one or more organization fact-checked a claim inside the target sentence, and instead of forming the problem as classification, instead they opt for a ranking problem. Their fully connected neural network outperforms all the baselines as well as the SVM model in mean average precision and precision at k documents. They further analyse results by comparing their system that uses all features versus the same system, but without contextual features. The findings suggest that these contextual features encode important knowledge that benefits all the metrics, especially precision at k documents metric.

3 Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking

The authors present an interesting analysis of news corpora containing satire, hoax, propaganda and trusted news articles. They noticed some often ocurring words in non-trusted news articles such as: swear words, second person pronouns, modal adverbs, action adverbs and first person pronouns. These words might also appear more often in non-checkworthy claims, although it depends on the source of the data. Their results suggest that credible sources use more comparatives, assertive words, money-related words, numbers and words similar to "hear".

Following the analysis, they provide models for predicting the reliability of news articles using four already mentioned categories. They decided to use a max-entropy classifier with L2 regularization on n-gram TF-IDF feature vectors. Their model achieves an F1 score of 0.65 where the random baseline achieves only 0.26, which is a big improvement. Besides training this model, they provide an analysis of its' highest weights. For trusted news these weights are connected to specific places or times, for satire these are hearsay words, hoax relies on divisive topics and dramatic word, while for propaganda these are dramatic abstract words such as truth or freedom and words regarding current world trends.

Finally, they try to predict the truthfulness of the data. They do this by employing a naive Bayes model, a max-entropy model, and an LSTM. As expected, the LSTM outperforms all other models when using exclusively text. Once other features are added the traditional machine learning algorithms prevail, indicating that these feature are more valuable for traditional machine learning algorithms, and that they might be redundant for the LSTM. The results reported on the test set suggest that the naive Bayes model and max-entropy classifier provide better generalization properties on the two-class problem.

4 Leveraging Commonsense Knowledge on Classifying False News and Determining Checkworthiness of Claims

This paper exploits multi-task learning by combining false news classification or check-worthy claim detection with a common sense QA task. They use an MTBERT model which is a BERT based architecture specialized for multi-task learning. For the dataset, they use posts from certain subreddits such as *TheDonald*, *fakenews* and *pol*. Besides, they use the ClaimBuster dataset for check-worthy claim detection and the CSQA dataset for common sense QA. Their experiments show that multi-task learning helps with robustness in classifying some fake news categories, it helps with classifying news from unseen publishers, and drastically improves claim check-worthiness detection.

---

5 Baselines

Four datasets were chosen for experiments:

1. CLEF2023 CheckThat! Task 1: Check-Worthiness in Multimodal and Unimodal Content
    - Subtask 1B (Multigenre) - general claim-checkworthiness dataset
2. CLEF2021 CheckThat! Task 1: Check-Worthiness Estimation
    - Subtask 1A: Check-worthiness of tweets - specific claim-checkworthiness dataset regarding COVID-19
    - also has available labels for regular claim detection, so multi-task learning might be interesting to try out
3. Fact Extraction and Verification
    - contains claims that need to be checked and a list of available evidences
    - a stratified sample of the train set will be used because of tractability - the original dataset has around 230 000 examples
4. ?

For the baselines, two models were chosen: BERT and RoBERTa. The most important standard metrics were monitored: accuracy, recall, precision, F1. The hyperparameters and the achieved metric scores can be seen in table 5. The hyperparameter choice was inspired by the original BERT and RoBERTa papers.

The performance of both baselines on each of the datasets can be seen in the figures below.

Model	Dataset	Learning rate	weight decay	batch size	Best F1 score	Best @ epoch
BERT	CT23	2e-5	1e-2	32	0.758	19
RoBERTa	CT23	2e-5	1e-2	32	0.770	4
BERT	CT21	2e-5	1e-2	32	0.701	8
RoBERTa	CT21	2e-5	1e-2	32	0.758	7
BERT	FEVER	2e-5	1e-2	8	0.827	10
RoBERTa	FEVER	2e-5	1e-2	8		
BERT		2e-5	1e-2	8		
RoBERTa		2e-5	1e-2	8		

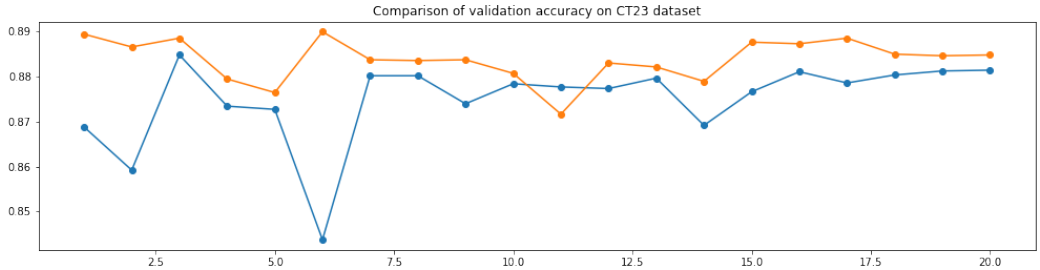

Figure 1: The accuracy of baselines over 20 epochs on the CT23 dataset

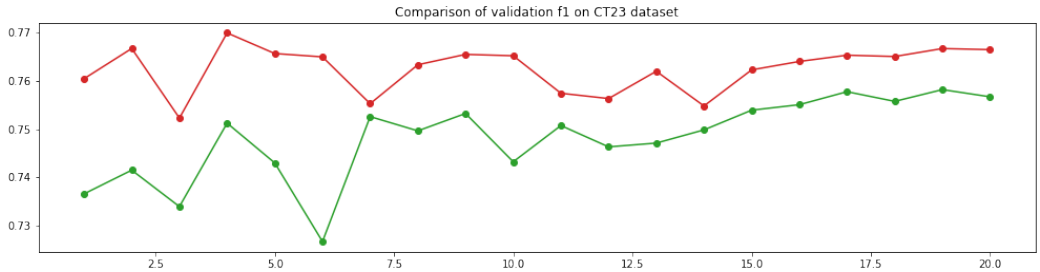

Figure 2: The F1 score of baselines over 20 epochs on the CT23 dataset

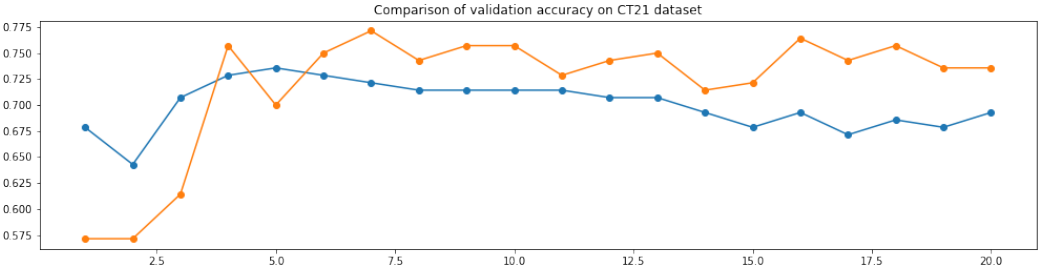

Figure 3: The accuracy of baselines over 20 epochs on the CT21 dataset

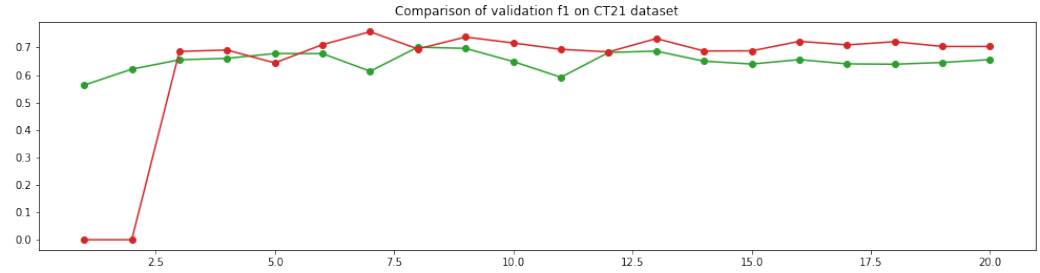

Figure 4: The F1 score of baselines over 20 epochs on the CT21 dataset

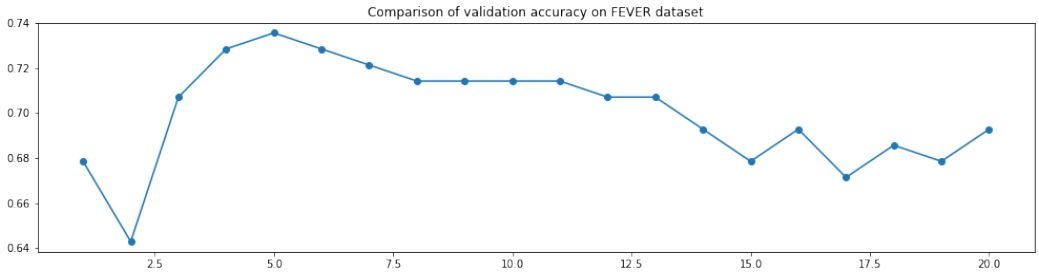

Figure 5: The accuracy of baselines over 20 epochs on the FEVER dataset

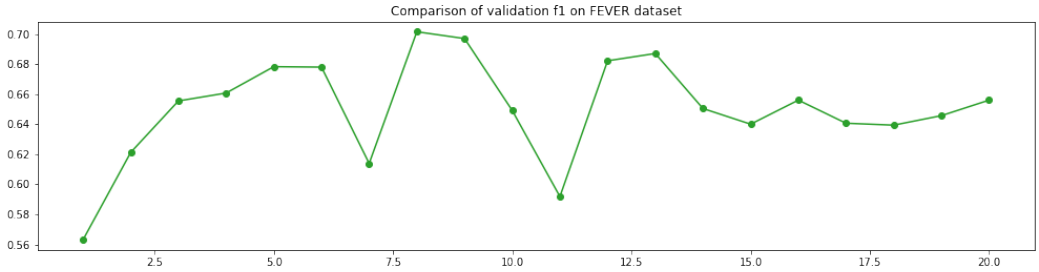

Figure 6: The F1 score of baselines over 20 epochs on the FEVER dataset