Multilingual hope speech detection: A Robust framework using transfer learning of fine-tuning RoBERTa model

3. Methods

3.1 Translation Steps:

Two corpora, one in English and one in Russian, are created by translating YouTube comments between the two languages using the Google Translator API. The translated comments are manually edited to address inconsistencies and issues and combined with the original datasets in each language.

3.2 Preprocessing Steps:

Standard text preprocessing steps are employed, including the removal of hashtags, HTML tags, mentions, punctuation, and URLs. Text is transformed to lowercase, numbers are removed, and emojis/emoticons are replaced with corresponding text representations. For English comments, misspelled words are corrected, and abbreviations are decoded. Two approaches for multilingual text classification are considered: translation-based (translating text into a universal language) and joint multilingual (combining comments from both languages).

3.3 RoBERTa Model:

The RoBERTa model, a modified version of BERT, is explored for its ability to capture the context of hope speech in multilingual text. Three variants of RoBERTa are used: XLM-RoBERTa (joint multilingual), English-RoBERTa (translation-based), and Russian-RoBERTa (translation-based). Text representation involves tokenization using RoBERTa tokenizer for English text, XLM-RoBERTa tokenizer for multilingual text, and a custom RoBERTa-Russian tokenizer for Russian text. Fine-tuning is performed using grid search to find optimal hyperparameter values, including sequence length, batch size, learning rate, weight decay, warmup ratio, hidden dropout, attention dropout, and the number of epochs.

3.3.3 Catastrophic Forgetting:

The issue of catastrophic forgetting in transfer learning is addressed by experimenting with a range of learning rates during fine-tuning.

3.3.4 Overfitting:
Overfitting is mitigated by selecting an appropriate number of epochs for training, with careful monitoring of the loss function on the validation data.
3.3.5 Text Classification:
A single layer is added on top of the RoBERTa base model for binary text classification (hope speech or not). The model is trained to optimize binary cross-entropy loss. The proposed methodology aims to leverage the strengths of the RoBERTa transformer model for multilingual hope speech detection, considering both translation-based and joint multilingual approaches. The grid search technique is employed for hyperparameter tuning, and special attention is given to addressing issues such as catastrophic forgetting and overfitting. The fine-tuned models are evaluated on labeled datasets, and the results are presented in terms of accuracy and F1-score. The methodology is implemented and tested on both English and Russian datasets, demonstrating its applicability and effectiveness.

The experimental setup involves conducting several experiments to fine-tune three versions of the RoBERTa model and compare the performance of the proposed pipeline with various baselines. The experiments are conducted on two hope speech datasets in Russian and English. The datasets are split into 80% for training and validation and 20% for testing, with further splits for training and validation. The datasets are processed manually to address unknown samples and categorize them into hope speech or not-hope class labels.

4.1 Multi-lingual Corpora (English & Russian):

4.1.1 English Dataset:

- Original Dataset: 29,744 instances (2,586 hope-speech, 27,158 not-hope)
- Train-Validation-Test Split: 21,415 training, 2,380 validation, 5,949 testing

4.1.2 Russian Dataset:

- Original Dataset: 5,425 instances (1,294 hope-speech, 4,131 not-hope)
- Train-Validation-Test Split: 3,906 training, 434 validation, 1,085 testing

4.2 Evaluation Metrics:

- Four metrics are employed: accuracy, precision, recall, and F1-score.

4.3 Comparison Methods:

4.3.1 Feature Models:

- TF-IDF Tokenizer:

    - TF-IDF is used as a weighting technique for terms in long-text datasets.
    - Formulation: TF-IDF(t,d)=tf(t,d)×idf(t)

4.3.2 Word Embedding:

- Word2Vec:
    - Skip-gram model with a window size of 100 dimensions.
    - Used for generating word embeddings.

4.3.3 RoBERTa as a Feature Model:

- RoBERTa:
    - RoBERTa is considered as a feature model and combined with ML and DL models for classification.

4.3.4 Baseline Models:

- Combinations of feature models with various ML and DL models are considered as baseline models.
    - TF-IDF + SVM, TF-IDF + RF, TF_IDF + CNN
    - Word2Vec + SVM, Word2Vec + RF, Word2Vec + CNN
    - RoBERTa + SVM, RoBERTa + RF, RoBERTa + CNN

4.4 Experimental Comparison:

- Fine-tuned RoBERTa models (English, Russian, and Multi-lingual) are compared with baseline models using the mentioned feature models and classifiers.
- Evaluation metrics (accuracy, precision, recall, and F1-score) are used to assess the performance of each model on the test datasets.

5. Results and Analysis

5.1 Joint Multi-lingual Model Analysis

5.1.1 XLM-RoBERTa Performance:
The joint multi-lingual model, XLM-RoBERTa, is fine-tuned on the combined English and Russian dataset. The experiments include varying sequence lengths, batch sizes, and learning rates. The results in Table 4 showcase that the best performance is achieved with a sequence length of 200, batch size of 32, and learning rate of 1e-5, resulting in 91.95% accuracy and 78.78% macro F1-score.

5.1.2 Comparison with Baseline Models:
Fine-tuned XLM-RoBERTa is compared against nine state-of-the-art models, including TF-IDF + SVM, TF-IDF + RF, Word2Vec + CNN, and others. The proposed model outperforms benchmarks, with significant improvements in accuracy and F1-score. For instance, the proposed model achieves a 0.66% improvement in accuracy and a 10.44% improvement in F1-score for the hope class compared to the best-performing baseline (TF-IDF + SVM).

5.2 Joint Translated-Approach Analysis

5.2.1 English-RoBERTa Performance:
Fine-tuning English-RoBERTa on the joint-translated English corpus results in superior performance. The model achieves the highest macro F1-score of 79.48% and accuracy of 92.20% with a sequence length of 200. Comparison with nine benchmarks shows substantial improvement, with the proposed framework outperforming by 6.67% in macro F1-score.
5.2.2 Russian-RoBERTa Performance:
Fine-tuning Russian-RoBERTa on the joint-translated Russian corpus yields the best performance with a macro F1-score of 80.24% and accuracy of 93.57%. The proposed model significantly outperforms benchmarks, particularly TF-IDF + SVM, with a 2.44% improvement in accuracy and a 7.36% improvement in macro F1-score.

5.3 Overall Findings:

- The proposed framework demonstrates superior performance across joint multi-lingual and joint-translated approaches.
- The joint multi-lingual model (XLM-RoBERTa) achieves a maximum accuracy of 91.95% and a macro F1-score of 78.78%.
- Joint-translated English-RoBERTa achieves an accuracy of 92.20% and a macro F1-score of 79.48%.
- Joint-translated Russian-RoBERTa achieves the highest accuracy of 93.57% and the highest macro F1-score of 80.24%.
- The proposed framework consistently outperforms nine benchmark models across various scenarios and evaluation metrics.

6. Discussions and Limitations

6.1 Key Findings:

The study makes a significant contribution to the identification of hope speech expressions in YouTube comments. It introduces a multi-lingual framework, addressing a gap in existing literature that primarily focuses on mono-lingual approaches. The proposed methodology, utilizing joint multi-lingual and translation-based techniques, proves effective for hope speech detection in both English and Russian languages. The study leverages transfer learning by fine-tuning the RoBERTa model, achieving notable results, including 94% accuracy and an 80.24% macro F1-score with the joint-translation approach. Moreover, the proposed framework outperforms nine baseline models, showcasing its practicality and potential applicability to other Multilingual Text Classification (MTC) problems.

6.2 Limitations:

Limited Language Coverage: The study focuses on English and Russian languages, and expanding the framework to include other widely used languages, both high and low-resource, could enhance its versatility.

Dataset Size: While the datasets used are not small, evaluating the framework on larger datasets would strengthen the generalizability of the findings.

Binary Classification: The study simplifies hope speech as a binary classification problem. Future work could explore subcategories of hope speech, such as Realistic-hope, Generalized-hope, and Unrealistic-hope, to provide more nuanced insights.

7. Conclusion:

The study addresses the challenge of multi-lingual hope speech detection in English and Russian languages through a novel joint multi-lingual and joint translation-based approach. Leveraging transfer learning with RoBERTa models, the proposed framework achieves promising results, outperforming baseline models. The extensive experiments demonstrate the effectiveness of the methodology, with the Russian joint-translated model attaining the highest accuracy (93.57%) and macro F1-score (80.24%). The study concludes that the proposed framework is not only applicable to hope speech detection but also holds promise for other NLP tasks.

8. Future Work:

The proposed methodology opens avenues for future research:

Extension to Other Languages: The framework can be extended to cover a broader range of languages, especially low-resource languages, expanding its applicability.

Scaling to Larger Datasets: Evaluating the framework on larger datasets will enhance the generalizability of results and validate its robustness.

Fine-grained Classification: Exploring finer classifications of hope speech, such as Realistic-hope, Generalized-hope, and Unrealistic-hope, would provide a deeper understanding of hope expressions.

Hybrid Models: Future work can explore hybrid models, combining transformer architectures with evolutionary algorithms, to address language complexity and ambiguity more effectively.

In conclusion, the proposed framework serves as a valuable tool for multi-lingual hope speech detection, and its potential can be further harnessed for diverse applications in natural language processing.