LREC-COLING 2024

RoBERTa Low Resource Fine Tuning for Sentiment Analysis in Albanian

Krenare Pireva Nuci*1 , Paul Landes 2, Barbara Di Eugenio 2

1 University of Prishtina,

2 University of Illinois Chicago

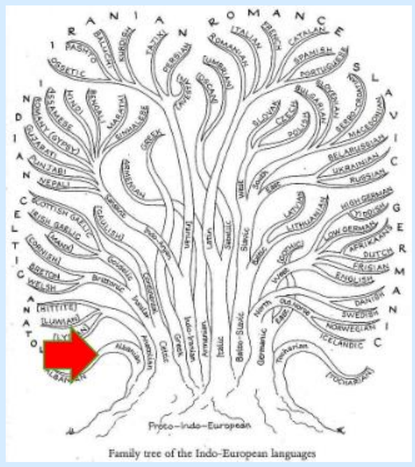
krenare.nuci@uni-pr.edu, {plande2,bdieugen}@uic.edu

1. Context

- Quality assurance: important iterative process in education
- Need to gather students feedback for quality assurance purposes
- Use NLP and DL approaches to automatically assess student feedback
- **Albanian:** challenges of low resource languages

Albanian language

- Albanian as an Indo-European Language with no close relation to any other languge
- It has 36 letters, and is rich of polisemantic terms
- Developing lingustic resources that aid in the classification of emotions and sentiment is challenging



Family tree of the Indo-European languages

2. Aims

- Determine how pretraining low resource language models, such as Albanian, affects downstream fine-tuning.
- Automatic methods to relate students' emotional states and opinions to their learning on specific educational topics in Albanian.
- Comparison between these methods with English trained models to assess feasibility of sentiment analysis task in Albanian

3. Dataset

Two datasets were created:

- one for pretraining Albanian embeddings
- for fine-tuning a model for the sentiment analysis task.

• EduSenti includes 1,163 students' feedback in Albanian and 624 students feedback in Albanian and English

• Annotated by 2 independent students

The dataset annotations include:

sentiment:	positive, neutral, and negative
emotion:	fear, sadness, anger, surprise, joy, and love
aspect:	course, professor, project, evaluation, institution, online learning, and general purposes

Example of annotated data:

Aspect	Emotion	Sentiment	Text	Lang
subject	joy	positive	Overall, I am very pleased with the way this course was conducted and I hope to continue at this pace in the other semesters as well.	en
			Në përgjithësi, jam shumë i kënaqur me mënyrën që ishte zhvilluar ky kurs dhe shpresoj që të vazhdoj me këtë ritëm edhe në semestrat tjerë	sq

4. Methods

Methods fall into two phases:

Pretraining:

(i) curation of Albanian corpus of text for pretraining embeddings,

(ii) pretraining Albanian embeddings from existing multi-language checkpoints

Fine-tuning:

(i) train new English and Albanian classification models on the annotated corpus, EduSenti sentiment dataset,

(ii) compare fine-tuned model across embeddings Models include BERT (M)ulti(L)ingual, our trained (XML-R)oBERTa (ALB)anian embeddings, and the last XLM-RoBERTa Base checkpoint

5. Albanian Large Aggregated Corpus

Sources of Albanian corpus with sentence count

Corpus	Count	Source
Oscar	1,340,766	Suárez et al.
WikiMatrix	640,955	Schwenk et al.
OpenSubtitles	222,757	Lison and Tiedemann
CCAligned	200,525	El-Kishky et al.
SETIMES	194,059	Tiedemann
QED	11,333	Abdelali et al.
TED2020	7,546	Reimers and Gurevych
GNOME	4,995	Tiedemann
Ubuntu	1,051	Tiedemann
Tatoeba	990	Tiedemann
GlobalVoices	491	Tiedemann

Pretrained Albanian corpus size

Description	*Count*
Sentences	**3, 984, 705**
Tokens	**121, 794, 474**
Characters	**647, 922, 859**

6. Results

Language	Model	mF1	mP	mR	MF1	MP	MR	WF1	WP	WR
English	BERT ML	68.75	68.75	68.75	47.29	50.32	48.52	66.60	66.36	68.75
English	BERT ML+E+T	70.31	70.31	70.31	27.52	23.44	33.33	58.06	49.44	70.31
English	fastText 300D	75.00	75.00	75.00	53.58	61.65	54.07	71.54	72.08	75.00
English	GLoVE 50D	**76.56**	76.56	76.56	**57.52**	67.66	55.19	**73.80**	74.85	**76.56**
Albanian	XLM-R ALB+E+T	57.63	57.63	57.63	26.79	28.64	31.98	46.75	42.77	57.63
Albanian	XLM-R ALB	60.17	60.17	60.17	25.04	20.40	32.42	46.48	37.87	60.17
Albanian	BERT ML	68.64	68.64	68.64	53.90	63.91	51.23	65.06	66.90	68.64
Albanian	XLM-RoBERTa Base	**73.73**	73.73	73.73	**61.07**	64.57	60.49	**71.90**	71.85	**73.73**

Sentiment model results with:

- (m)icro,
- (M)acro,
- (W)eighted F1,
- precision and
- recall.

7. Conclusion

- Created EduSenti, a large aggregated Albanian text corpus and an Albanian-English sentiment corpus that includes aspect, emotion and sentiment annotations; Available @ https://github.com/uic-nlp-lab/edusenti
- Compared multilingual models' original checkpoints with Albanian pretrained embeddings, trained fine-tuned sentiment analysis models, and reported their performance at the result section (see above).
- The Albanian language models show competitive performance with multi-language XLM-RoBERTa model (Conneau et al., 2020)
- The fine-tuned model trained from the XLMRoBERTa checkpoint speak to the feasibility of modeling the Albanian language

8. Selected References

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a Cleaner Document-Oriented Multilingual Crawled Corpus.In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 4344–4355. European Language Resources Association.

Julien Abadji, Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2021. Ungoliant: An optimized pipeline for the generation of a very large-scale multilingual web corpus. In CMLC 2021-9th Workshop on Challenges in the Management of Large Corpora.

Ahmed Abdelali, Francisco Guzman, Hassan Sajjad, and Stephan Vogel. 2014. The amara corpus: Building parallel language resources for the educational domain. In LREC, volume 14, pages 1044–1054.

Utku Umur Acikalin, Benan Bardak, and Mucahid Kutlu. 2020. Turkish sentiment analysis using bert. In 2020 28th Signal Processing and Communications Applications Conference (SIU), pages 1–4. IEEE.

Jordi Armengol-Estapé, Casimiro Pio Carrino, Carlos Rodriguez-Penagos, Ona de Gibert Bonet, Carme Armentano-Oller, Aitor Gonzalez-Agirre, Maite Melero, and Marta Villegas. Are Multilingual Models the Best Choice for Moderately Under-resourced Languages? A Comprehensive Assessment for Catalan. In Findings of the Association for Computational Linguistics: ACLIJCNLP 2021, pages 4933–4946. Association for Computational Linguistics.

Marenglen Biba and Mersida Mane. 2014. Sentiment analysis through machine learning: an experimental evaluation for albanian. In Recent Advances in Intelligent Informatics: Proceedings of the Second International Symposium on Intelligent Informatics (ISI'13), August 23-24 2013,Mysore, India, pages 195–203. Springer.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. Transactions of the association for computational linguistics, 5:135–146.

Priyavrat Chauhan, Nonita Sharma, and Geeta Sikka. 2021. The emergence of social media data and sentiment analysis in election prediction. Journal of Ambient Intelligence and Humanized Computing, 12:2601–2627.

Yanqing Chen and Steven Skiena. 2014. Building sentiment lexicons for all major languages. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 383–389.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In Proceedings of the 58t Annual Meeting of the Association for Computational Linguistics, pages 8440–8451. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186.

Bertan Karahoda, Krenare Pireva, and Ali Shariq Imran. 2016. Mel frequency cepstral coefficients based similar albanian phonemes recognition. In Human Interface and the Management of Information: Information, Design and Interaction: 18th International Conference, HCI International 2016 Toronto, Canada, July 17-22, 2016, Proceedings, Part I 18, pages 491–500. Springer.

Zenun Kastrati, Fisnik Dalipi, Ali Shariq Imran, Krenare Pireva Nuci, and Mudasir Ahmad Wani. 2021. Sentiment analysis of students' feedback with nlp and deep learning: A systematic mapping study. Applied Sciences, 11(9):3986.