RoBERTa Low Resource Fine Tuning for Sentiment Analysis in Albanian

Krenare Pireva Nuci[*1], **Paul Landes**[*2], **Barbara Di Eugenio**[2]
Department of Computer Science, University of Prishtina[1],
Department of Computer Science, University of Illinois Chicago[2]
krenare.nuci@uni-pr.edu, {plande2,bdieugen}@uic.edu

Abstract

The education domain has been a popular area of collaboration with NLP researchers for decades. However, many recent breakthroughs, such as large transformer based language models, have provided new opportunities for solving interesting, but difficult problems. One such problem is assigning sentiment to reviews of educators' performance. We present EduSenti: a corpus of 1,163 Albanian and 624 English reviews of educational instructor's performance reviews annotated for sentiment, emotion and educational topic. In this work, we experiment with fine-tuning several language models on the EduSenti corpus and then compare with an Albanian masked language trained model from the last XLM-RoBERTa checkpoint. We show promising results baseline results, which include an F1 of 71.9 in Albanian and 73.8 in English. Our contributions are: (i) a sentiment analysis corpus in Albanian and English, (ii) a large Albanian corpus of crawled data useful for unsupervised training of language models, and (iii) the source code for our experiments.

Keywords: sentiment, low resource, large language model

1. Introduction

Quality assurance is an important component in education. It involves the systematic review of educational processes to ensure its quality over time. Traditionally the quality evaluation of the learning processes is done using quantitative methods, which is typically performed automatically using metrics such as graded assignments and test scores. However, the insight of professors and instructors and interpretation these performance statistics are fundamental in assessing students' knowledge acquisition. Furthermore, an education professional's insights could contribute to reforming policies or improve regulation at the institutional level, or even national level.

Instructors' ability to teach is also a consideration, and thus, must also be included in metrics that factor in to the overall performance of any education program. To improve the student assessment process and its impact on the development and enhancement of the quality in education in general, students should be encouraged to give their opinions in text-based form rather than just rating the processes. Limiting this expression, insofar as surveys and written feedback, has the potentially of missed opportunities to refine the education system.

Opinions expressed by students are a valuable source of information; not only for reforming policies within education institutions, but also for analyzing students' behaviour towards a course, professors and its environment. The recent pandemic has highlighted the importance of students' feedback and opinions as remote learning became more pervasive.

The utilization of deep learning (DL) to assess students' feedback has been an interest of the research community (Kastrati et al., 2021). A number of sentiment analysis models have been successful in high resource languages such as English (Talaat, 2023). However, low resource languages, such as Albanian, continue to be challenging (Sadriu et al., 2022; Itani et al., 2018). Albanian as an Indo-European language and has no close relation with any other language, and in the family of Indo-European languages, it is positioned in a distinct branch.

In this work we focus on automatic methods to assess students' emotional states and opinions to the quality of their learning process on specific educational topics in Albanian. We compare these methods with English trained models to assess the feasibility of the sentiment analysis task in a low-resource language such as Albanian.

Specifically, our goal is to determine how pretraining low resource language models, such as Albanian, affects downstream fine-tuning. Our methods include pretraining a new Albanian large language model (LLM) from multi-lingual checkpoints, and then using it to train a sentiment analysis model (Section 3). Finally we compare methods on models trained on a translated Albanian-English corpus[1] and present our results in Section 3.

[*] Indicates equal contribution.

[1] Our pretrained corpus, sentiment corpus and code are released at https://github.com/uic-nlp-lab/edusenti.

LREC-COLING 2024, pages 14146–14151

2. Related Work

Students' opinions are a valuable source of information to assess the quality of knowledge transfer. Sentiment analysis of these opinions have resulted in a good deal of recent work.

In a recently systematic mapping study Kastrati et al. (2021), show that until 2016 — 2017 researchers used sentiment analysis involving lexicon-based and dictionary-based methods (Sharma et al., 2020; Chauhan et al., 2021; Wen et al., 2014). After 2017, researchers shifted to analyzing sentiment deep learning-based models (Sadriu et al., 2022; Sharma et al., 2020). The latter approaches used non-contextual word embedding (Mikolov et al.; Bojanowski et al., 2017), BiLSTM language models (Peters et al.) and transformer architectures (Devlin et al., 2019).

Sabri et al. (2021) and Acikalin et al. (2020) tackled the sentiment analysis problem in low resource languages with pretrained BERT embeddings and translation models. The first paper applied the technique in movie and hotel reviews, whereas the second one in the social media (Tweets). The authors used a BERT fine-tuned multilingual model and compared with a the English-only BERT after machine translation.

Subsequently, Selvakumar and Lakshmanan (2022) proposed BERT based sentiment classification on two datasets: IMDB Movie Review and Amazon Fine Food Review. The author compared the BERT experimental results with eleven other commonly used ML and DL models. The accuracy of sentiment classification using BERT model reached 94% compared to common ML and DL models.

Biba and Mane (2014) used Weka (Russell and Markov, 2017), for classifying the sentiment of a political news dataset. This dataset is composed of five topics, each containing 40 positive and 40 negative sentiments. The classification was performed by using logistic regression, naive, among other algorithms.

While the of majority interest has been in English, German, Chinese, relatively little has been found in low resource languages until recently. Specifically, Albanian is found in few publications, but to the best of our knowledge, none have used sentiment analysis for students' feedback in the education domain.

Given the dearth of Albanian public datasets, Vasili et al. (2021) used an annotated Twitter dataset (Mozetič and Grčar) and sentimental lexicons dictionary by Chen and Skiena (2014) for predicting the sentiment of tweets. The authors reached the best results using LSTM based on RNN model with a F1 of 87.8 and accuracy of 79.2%.

While our work is similar, our work differs in that we created an Albanian-English annotated dataset of educational instructors' performance reviews that was annotated for sentiment, emotion and educational topic. We also experimented with fine-tuning several models on our sentiment dataset using Albanian pretrained embeddings we trained ourselves.

2.1. Dataset

Two datasets were created: one for pretraining Albanian embeddings and another for fine-tuning a model for the sentiment analysis task.

2.2. Sentiment Dataset

The sentiment dataset was collected from second and third year computer science students as during two semesters. The data was gathered from reflective papers, which included feedback of the course, professor and institution. The sentiment corpus includes 1,163 students' feedback in Albanian and 624 students feedback in Albanian and English, which were annotated by two different students as three classes: sentiment, emotion, and aspect of reviews. Each review was human translated from Albanian to English. Table 1 gives an example of the review and their annotations.

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

The annotation process consisted of several iteration processes; initially the data was preprocessed by and cleaning the text using regular expressions. Initially the Google translation API was used to translate 624 English reviews from Albanian to English. The translations were validated and revised by two annotators. The standardizing across annotators was iterative during the annotation process. Krippendorff's α coefficient (Krippendorff, 2011), was used to compute inter-annotator agreement (IAA) between the two annotators, which resulted in 0.6 for sentiment, 0.64 for emotion and 0.22 for aspect.

2.3. Albanian Large Aggregated Corpus

Because Albanian contains unique morphological and lexical characteristics, a large alphabet with 36 letters, and rich of polysemantic terms, developing linguistic resources that that aid in the classification of sentiment and emotions is challenging (Vasili et al., 2021). The intricacy increases when one

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

Table 1: Dataset example of an annotated instructor's review for aspect, emotion, sentiment.

comes across the Tosk and Gheg dialects, as well as the regional variations in accent and cultural expression (Karahoda et al., 2016; Coretta et al., 2022).

Given that Albanian is considered a low resource language, the authors set out to compile a large corpus for the purpose of training a LLM (see Section 3) for the purpose of fine-tuning the sentiment corpus. Table 3 provides the corpus statistics with almost four million sentences (647MB).

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

Table 2: Sources of the Albanian corpus with the sentence count of each.

The corpus was first constructed from multiple sources such as CCAligned dataset (El-Kishky et al., 2020) as reported in Table 2. However, the largest source was Oscar (El-Kishky et al., 2020; Suárez et al., 2020; Abadji et al., 2021, 2022; Kreutzer et al., 2022), as it provides metadata indicating language detection probabilities and the quality and level of noise in the data. Some corpora were already sentence chunked, but those that were not chunked using regular expressions on punctuation and then tokenized on white space.

This corpus is much smaller than the source as it was heavily filtered for quality. Corpora that did not include language identification was automatically tagged for language using the Python port of the well-known `langdetect`[2] library. Corpus documents containing a high portion of Albanian detected language were kept. Of those documents, only sentences detected as Albanian with token lengths between 5 and 450 were added to our corpus. Table 3 provides corpus statistics and Figure 1 shows the distribution of sentences by token length for sentences with fewer than 100 tokens.

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

Table 3: Pretrained Albanian corpus size given in number of sentences, tokens and characters.

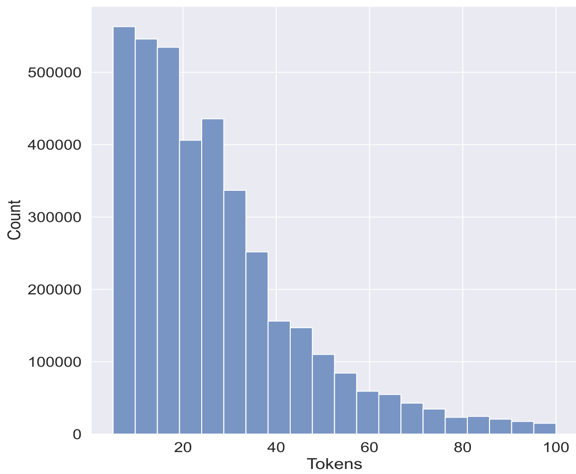

Figure 1: Albanian language corpus sentence counts by token length for sentences with fewer than 100 tokens.

3. Methods

Our methods fall into to two phases to create two kinds of models: pretraining embeddings and fine-tuned sentiment models. We first create new checkpoints from existing BERT (Devlin et al., 2019) base models on a Albanian-only language training set (Section 2.3). After these are trained, we train additional fine-tuned models on these new embeddings, but also on the same checkpoints to analyze the performance based on their trained trajectory.

More specifically, these two phases consist of:

1. Pretraining: (i) curation of Albanian corpus of text for pretraining embeddings, (ii) pretraining Albanian embeddings from existing multi-language checkpoints

[2]https://github.com/Mimino666/langdetect

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**

Table 4: Sentiment model results with (m)icro, (M)acro and (W)eighted F1, precision and recall. (E)motion and (T)opic are features added to some models. Models include BERT (M)ulti(L)ingual, our trained (XML-R)oBERTa (ALB)anian embeddings, and the last XLM-RoBERTa Base checkpoint.

2. Fine-tuning: (i) train new English and Albanian classification models on the annotated EduSenti sentiment dataset, (ii) compare fine-tuned model across embeddings

After procuring the Albanian corpus, the cased multilingual and BERT XLM-RoBERTa base (Conneau et al., 2020) checkpoints were used to train the model as they were natural choices given their training set already included Albanian. Both models used masked model training for 4 epochs with a learning rate of 3e-5.

Fine-tuned models were trained from the last checkpoints of multilingual BERT, XLM-RoBERTa and our own Albanian pretrained embeddings. The pooler output (`[CLS]`) was connected to a fully connected linear layer, which was in turn connected to the three way sentiment output (positive, negative and neutral). All were trained for 20 epochs with a learning rate of 10^{-2} that decreased a schedule of 5 epochs of no improvement.

For comparison, we also trained models using the non-contextual word vector embeddings GloVE (Pennington et al., 2014) and fastText (Bojanowski et al., 2017). As with the transformer models, a fully connected linear layer connected to the output layer, but a BiLSTM was used in place of the transformer. The Zensols Deep NLP framework (Landes et al., 2023) was used for fine-tuning model development, training, and evaluation.

4. Results

Table 4 presents the results of the fine-tuned models on the sentiment analysis task using our English and Albanian datasets. The results clearly show English favors the GloVE and fastText non-contextual word embeddings, which suggests the mixed language transformer models still do not keep up with English-only embeddings. However, the Albanian language models show competitive performance with multi-language XLM-RoBERTa model (Conneau et al., 2020). This performance is somewhat surprising given the uniqueness of Albanian and its limited representation (0.22%) in the XLM-RoBERTa training data. This contrasts with the lackluster performance of low resource languages (Catalan) with high resource language (Spanish) families (Armengol-Estapé et al.).

Surprisingly the Albanian pretrained models shows lower performance on downstream fine-tuned models. We speculate that the pretrained models performed poorly because of the small mini-batch size given GPU memory constraints. We believe additional pretraining embedding hyperparameter tuning and including next sentence training would yield significantly better results, which we leave as future work. Regardless of this model task, the fine-tuned model trained from the XLM-RoBERTa checkpoint speak to the feasibility of modeling the Albanian language.

5. Conclusion and Future Work

We have presented EduSenti, a large aggregated Albanian text corpus and an Albanian-English sentiment corpus that includes aspect, emotion and sentiment annotations. We compared multilingual models' original checkpoints with Albanian pretrained embeddings, trained fine-tuned sentiment analysis models, and reported their performance.

As far as we know, we are the first to train Albanian models for the sentiment analysis task. We believe our results motivates further work in this language with our results on the fine-tuned models. However, the fine-tuned models trained from Albanian-only embeddings clearly show there is much room for growth. Not only in terms of available datasets, but essential upstream pipeline components, such as tokenizers, still do not exist for this low-resource language.

6. Acknowledgments

This work was funded by an award by the Fulbright Scholar program and we thank them for their support.

7. Bibliographical 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: ACL-IJCNLP 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 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. Association for Computational Linguistics.

Stefano Coretta, Josiane Riverin-Coutlée, Enkeleida Kapia, and Stephen Nichols. 2022. Northern tosk albanian. *Journal of the International Phonetic Association*, pages 1–23.

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.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. CCAligned: A Massive Collection of Cross-Lingual Web-Document Pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5960–5969. Association for Computational Linguistics.

Alya Itani, Laurent Brisson, and Serge Garlatti. 2018. Understanding learner's drop-out in moocs. In *Intelligent Data Engineering and Automated Learning–IDEAL 2018: 19th International Conference, Madrid, Spain, November 21–23, 2018, Proceedings, Part I 19*, pages 233–244. Springer.

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.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, et al. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Klaus Krippendorff. 2011. Agreement and Information in the Reliability of Coding. *Communication Methods and Measures*, 5(2):93–112.

Paul Landes, Barbara Di Eugenio, and Cornelia Caragea. 2023. DeepZensols: A Deep Learning Natural Language Processing Framework for Experimentation and Reproducibility. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pages 141–146. Empirical Methods in Natural Language Processing.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

I Mozetič and M Grčar. Smailovi c j. 2016. *Multilingual Twitter sentiment classification: the role of human annotators. PLOS ONE*, 11(5):e0155036.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. *arXiv preprint arXiv:2004.09813*.

Ingrid Russell and Zdravko Markov. 2017. An introduction to the weka data mining system. In *Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education*, pages 742–742.

Nazanin Sabri, Ali Edalat, and Behnam Bahrak. 2021. Sentiment analysis of persian-english code-mixed texts. In *2021 26th International Computer Conference, Computer Society of Iran (CSICC)*, pages 1–4. IEEE.

Shpetim Sadriu, Krenare Pireva Nuci, Ali Shariq Imran, Imran Uddin, and Muhammad Sajjad. 2022. An automated approach for analysing students feedback using sentiment analysis techniques. In *Pattern Recognition and Artificial Intelligence: 5th Mediterranean Conference, MedPRAI 2021, Istanbul, Turkey, December 17–18, 2021, Proceedings*, pages 228–239. Springer.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. *arXiv preprint arXiv:1907.05791*.

B Selvakumar and B Lakshmanan. 2022. Sentimental analysis on user's reviews using bert. *Materials Today: Proceedings*, 62:4931–4935.

Sudhir Kumar Sharma, Mohit Daga, and Bhawna Gemini. 2020. Twitter sentiment analysis for brand reputation of smart phone companies in india. In *Proceedings of ICETIT 2019: Emerging Trends in Information Technology*, pages 841–852. Springer.

Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. *arXiv preprint arXiv:2006.06202*.

Amira Samy Talaat. 2023. Sentiment analysis classification system using hybrid bert models. *Journal of Big Data*, 10(1):1–18.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Roland Vasili, Endri Xhina, Ilia Ninka, and Dhori Terpo. 2021. Sentiment analysis on social media for albanian language. *Open Access Library Journal*, 8(6):1–31.

Miaomiao Wen, Diyi Yang, and Carolyn Rose. 2014. Sentiment analysis in mooc discussion forums: What does it tell us? In *Educational data mining 2014*. Citeseer.