Semi-Supervised Topic Modeling of Albanian Tweets

Elda Pere
W266: Natural Language Processing
University of California, Berkeley
eldapere@ischool.berkeley.edu

Abstract

A majority of NLP research in the social media domain has evolved with English and a few other common languages in mind. This prevents agents with academic, political or economic interest from reaping the benefits that social media can provide from cultures speaking lesser known languages, while also keeping the latter in the dark about new products and ideas. With a novel tweet dataset we collected using research on the Albanian language, this paper aims to detect the main topics mentioned by Albanian tweets in a semi-supervised setting. It will study the performance of topic models using word embeddings trained on Word2Vec and BERT, which are then clustered using the kMeans algorithm. They are compared to each other and a Latent Dirichlet Allocation (LDA) baseline model using descriptive analysis, intrinsic measures and extrinsic metrics where hashtags are considered a ground truth label.

1. Introduction

The Albanian language is an Indo-European language spoken by a population of roughly 10 million and is considered a low-resource language in the field of natural language processing. This is evident in the little to no published research found on tasks such as part-of-speech tagging, tokenization, dependency parsing, named entity recognition, language modeling, etc. (Mati et al., 2019). With this paper, we aim to combine some of the existing literature with a novel tweet dataset to provide a comparison of topic modeling architectures for Albanian. By taking advantage of conventional algorithms and the latest transformer-based language model architectures, we will demonstrate which methods enable researchers to learn more useful topics compared to using word counts.

The data collection process consisted of using the open-source Twitter API to query tweets in Albanian between December 2015 and December 2018 that have very rarely been analyzed before due to demand and technical limitations. Since Twitter does not currently support the Albanian language, we used the paper "The 100 Most Common Words in Albanian" to filter the tweets (2010). The filtered corpus consisted of 176,123 tweets posted by 11,144 unique authors. For an in-depth exploration of this data, follow this link.

The tweets were then preprocessed by being stripped of links, hashtags and non-alphabetic characters. The Albanian **ë** and **ç** letters were replaced with "e" and "c" respectively which is often how users substitute the letters themselves for convenience. Stemming and stopword removal was done in a

rule-based method using community-sourced lists, however for more robust implementations the data can be lemmatized using a manually annotated corpus and neural models as seen in Kote et al. (2019). Instead, this paper will more thoroughly describe the performance of topic models trained on a set of different word embeddings. The models are:

a) Latent Dirichlet Allocation (LDA) to serve as a baseline model,
b) Mini-batch kMeans clustering of word embeddings trained with Word2Vec,
c) Mini-batch kMeans clustering of word embeddings pretrained from Multilingual BERT.

These are described in more detail in the Methods section. Afterwards, in the Results and Discussion section we compare the models by using the silhouette score as an intrinsic measure and the precision, recall and F1 scores as extrinsic measures where hashtag co-occurrences are used as ground truth labels. Finally, a qualitative analysis is done with a sample of top ranked Albanian keywords translated to English.

2. Background

Topic modeling is a statistical method used to discover latent topics in a collection of texts. Research on topic modeling a corpus of tweets and in different languages has been extensive, and has been supported by other techniques such as lemmatization, tokenization, language modeling and clustering.

For lemmatizing Albanian text, Kote et al. (2019) describe a robust framework building upon the work of Karanikolas (2009) in "Bootstrapping the Albanian Information Retrieval". In addition, a few GitHub repositories provide preprocessing techniques specific to the language which include lists of Albanian stopwords and suffix removal functions[1]. To tokenize the text, Wang et al. (2019) demonstrate the use of byte-level subwords being more efficient than using pure bytes in their paper "Neural Machine Translation with Byte-Level Subwords". This has gained traction and is used in this paper as well in a RoBERTa framework.

For the model architectures, we considered Latent Dirichlet Allocation, first presented in Blei et al. (2003). In later studies, the coherence score is used to evaluate LDA models and to determine the optimal number of topics to cluster, seen in Mimno et al. (2011), Steinskog et al. (2017) and Rangarajan Sridhar (2015). An additional measure of cluster quality is the silhouette score which is presented in Rajshekhar Shahapure et al. (2020), Panichella et al. (2013) and Ma et al. (2016). Topic modeling architectures are also commonly built over trained word embeddings which can be learned with Word2Vec, first introduced in "Efficient Estimation of Word Representations in Vector Space" (2013) and BERT, first introduced in "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" (2018).

A unique aspect of this paper is the semi-supervised design which is due to the use of hashtags as labels. Similar efforts have been seen in "Twitter Topic Modeling by Tweet Aggregation" (2017) which moves beyond simple LDA, increasing topic coherence.

3. Methods

[1] https://github.com/arditdine/albanian-nlp

3.1. Design

The design of this study was structured based on the type of available data. Tweet topics are not universally defined in any business setting, which led us to an unsupervised learning approach. Meanwhile, the nature of hashtags allows them to be considered as a topic label for each tweet that has at least one hashtag. The assigned hashtag label would be considered accurate in the case when the tweet (which has at least one hashtag) contains that hashtag label. We use hashtags as ground truth labels in calculating precision, recall and F1 scores, however we chose not to frame the study as a supervised learning instance because only 5% of the tweet corpus contained hashtags.

The next step was to determine the number of topics to cluster. This was done analytically, using the Elbow method. In effect, an LDA model was run with a range of topics and for each of them, we calculated the *coherence score* (Mimno et al., 2011) which has been considered a good topic indicator by Steinskog et al. (2017) and Rangarajan Sridhar (2015). As seen in Figure 1, the coherence score begins to flatten at approximately k = 30 topics.

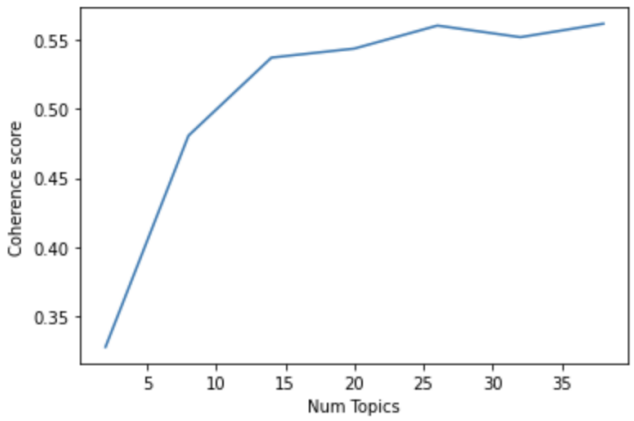

Figure 1: Elbow Method for Determining Number of Topics

3.2. LDA Model

With this predetermined number of topics, we continued to create an LDA topic model where each tweet is considered a separate document. Since LDA is known to be very susceptive to hyperparameters, these were determined to be $\alpha = 0.05$ and $\beta = 0.01$ after optimizing for coherence according to Rangarajan Sridhar (2015). On manual inspection of the topic keywords, less than 5 of the 30 topics appear to closely follow a general theme while keywords across other topics show overlap. This is further discussed in the Results and Discussion section, along with the results from the following architectures.

3.3. Word2Vec Embeddings with kMeans

In "Improving Topic Models with Latent Feature Word Representations", Nguyen et al. show that normalized pointwise mutual information scores of LDA topic models are improved when using Word2Vec word embeddings (2018). Based on their results, we chose to cluster Word2Vec[2] embeddings as the next step in our analysis. First, a tokenizer was trained with a byte-level Byte-pair encoding (BPE) tokenizer. Despite the complex grammar of the language, the tokenizer performed well on initial

[2] https://code.google.com/archive/p/word2vec/

inspection. For example, the two suffixes “ai” and “allaret” are often part of the same root (“babai” and “baballaret”) and thus are mapped to the same token. The tokens are then trained with Word2Vec and clustered using Mini Batch KMeans, which was chosen because it has shown to outperform the conventional KMeans algorithm in both accuracy and training time (Feizollah et al., 2014).

3.4. BERT Embeddings with kMeans

Both the Word2Vec/kMeans model and the LDA model have failed to include a contextual perspective of the data. This can be done by learning a transformer-based model such as BERT. After attempting to train a RoBERTa model from scratch with the tweet data, we realized other Albanian researchers could be facing the same resource constraints as we were facing. Given this, we turned to the pretrained Multilingual BERT Base model, which is trained on 104 languages including Albanian. Similar to the previous model, these embeddings were clustered using Mini Batch kMeans.

4. Results and Discussion

4.1. Extrinsic Evaluation

For each of the architectures, the tweets were assigned a hashtag that occurred most often with the topic that the tweet was clustered in. Using only the subset of tweets that had at least one hashtag – around 5%, the topic was considered correct if the assigned hashtag was actually used in the tweet by that user. This was shown by Steinskog et al. (2017) to produce a more coherent topic model than baseline models because this metadata tag used in tweets contains information that is not present in standard documents. The hashtag now serves as a ground truth metric that is used to calculate precision, recall and F1 score. These are seen in Table 1.

Model	Precision	Recall	F1
LDA	0.5655	0.5094	0.4788
Word2Vec + kMeans	0.9353	0.9239	0.9147
Multilingual BERT + kMeans	0.9028	0.9035	0.8841

Table 1: Precision, Recall and F1 scores calculated for each model. Ground truth label is considered the hashtag(s) used in the tweet. Predicted label is considered the hashtag assigned to the tweet’s topic.

The results indicate that the baseline LDA model performed significantly worse than the other two models. Clustering with Word2Vec embeddings appears to have had the best performance while clustering with Multilingual BERT falls shortly behind. The scores for LDA are approximately a 7.8% improvement over the only similar paper we found on topic modeling for Facebook reviews by Axhiu et al. (2020) who found a precision of 0.4746, recall of 0.4046 and F1 score of 0.4368 when using the most relevant topic keywords as ground truths.

4.2. Intrinsic Evaluation

We see a similar trend between Word2Vec and BERT when comparing intrinsic evaluation methods such as the silhouette score seen in Table 2. This score ranges between -1 and 1 where 1 is the best value and values near 0 indicate overlapping clusters. It was chosen as a metric for our topic clusters after studying

its use for quality analysis in Rajshekhar Shahapure et al. (2020), Panichella et al. (2013) and Ma et al. (2016).

Model	Silhouette Score
LDA	0.27
Word2Vec + kMeans	0.05
Multilingual BERT + kMeans	-0.02

Table 2: Intrinsic evaluation: Silhouette Score calculated for each model.

This metric tells a different story when comparing to the baseline LDA model in that the LDA model seems to have outperformed the other two. On manual inspection of the topic clusters, this is not necessarily the case. This demonstrates the superiority of human judgment over intrinsic evaluation methods as mentioned in Steinskog et al. (2017). The difference between the intrinsic and extrinsic measures also comes from the difference in the tweets being evaluated; the extrinsic scores are based on a small subset which is not necessarily representative of the whole corpus.

Word2Vec likely dominates over BERT embeddings in both extrinsic and intrinsic metrics because the embeddings and tokenizer were trained specifically on this corpus. Multilingual BERT was trained on 104 different languages including Albanian and used Wikipedia articles as its corpus. By their nature, Wikipedia articles are written very differently from tweets and so we would expect a custom embedding model to perform better, which we see here. When the computational resource constraints are removed, we will extend this analysis to include embeddings trained on BERT using this tweet corpus, where we expect to see improvements due to the contextualization of the embeddings.

4.3. Qualitative Evaluation

When conducting a qualitative review of the topics produced by the unsupervised frameworks, the top ranked words of each cluster showed promising results. Across each model architecture, the topics of religion, sports and politics (both internal and global) persist. Examples of this are seen in Table 3.

	LDA Model		**Word2Vec Model**		**BERT Model**	
Topic	Highest Ranking Keywords	English Translation	Highest Ranking Keywords	English Translation	Highest Ranking Keywords	English Translation
Misc.	shum, seps, fjal, ke, dite, dua, asht, them, thjesht, di	much, because, word, have, day, want, is, say, just, know	sot, kosov, shum, foto, thot, moment, fakt, lajm, histor, jasht	today, Kosovo, much, photo, says, moment, fact, news, history, outside	lajm, fund, kliko, shiku, detaj, foto, shikojen, tekst, pamj, bindun	news, latest, click, look, detail, photo, look, text, view, be convinced
Global Politics	kosov, president, serb, thac, shba, shtet, serbis, prishtin, gjykat, shqiperis	Kosovo, president, Serbia, Thaci, USA, to Serbia, Prishtine, court, to Albania	kosov, serbis, shqiperis, shba, berish, kryesor, ndodh, konsullat, soros, vesel	Kosovo, to Serbia, to Albania, USA, Berisha, main, happens, consulate, Soros, Veseli	kosov, marreveshj, mes, serbis, president, nato, maqedonis, shba, rajon, evropian	Kosovo, deal, between, to Serbia, president, NATO, to Macedonia,

						USA, region, european
Religion	allah, zemr, burr, thot, meshiroft, allahut, gruaj, kuran, imam, besim	allah, heart, man, says, may have mercy, of allah, the woman, quran, imam, faith	allah, allahut, then, namaz, madh, pare, ramazan, imam, paqj, pejgamber	allah, of allahut, said, namaz, great, first, Ramadan, imam, peace, prophet	allah, allahut, meshiroft, thot, derguar, imam, qe, ai, paqj, profet	allah, of allah, may have mercy, says, sent, imam, that, he, peace, prophet
Sports	ndeshj, ekip, mess, lojtar, shum, futboll, klub, sezon, pare, milan	match, team, Messi, player, very, football, club, season, first, Milan	ndeshj, plagos, madrid, ronaldo, lojtar, ballkan, ndar, arsenal, shqipt, skuadr	match, wound, Madrid, Ronaldo, player, Balkan, separated, Arsenal, Albanian, team	lojtar, barcelon, ka, mess, ndeshj, ronaldo, ai, madrid, trajner, milan	player, Barcelona, has, Messi, match, Ronaldo, he, Madrid, coach, Milan

Table 3: Highest ranking keywords of a sample of clusters across each model.

These terms were pulled using SciKitLearn methods for LDA and using class-based TF-IDF for Word2Vec and BERT. This was motivated by "An improved TF-IDF Algorithm in text classification" by Xu et al. (2014) where they show improved results using the class-based model. The qualitative results for the clusters mentioned (religion, politics, sports) are easily trained across each model. The remaining topics display more noise for the LDA model and less so for the other two. This can be seen in the "Miscellaneous" column in Table 3. Remaining topics can be categorized into pop culture, internal politics, crime, education, finance, cooking, and others.

5. Conclusion

In this paper, we presented a comparison of semi-supervised topic modeling frameworks on a novel dataset of Albanian tweets. Between an LDA baseline model, a MiniBatch kMeans model with Word2Vec embeddings and a MiniBatch kMeans model with pre-trained Multilingual BERT embeddings, we found that the Word2Vec model outperformed the others in terms of their extrinsic evaluation metrics, measured by using hashtags as ground truth labels.

The work presented here was inspired by the need for further research on natural language processing techniques in Albanian and other low-resource languages. We intend to extend this work with topic models using custom trained BERT embeddings as well as a hyperparameter optimization study, and hope that when we do, we will find that Twitter supports the Albanian language and that there will be more open source resources for Albanian research.

References

Axhiu, M., & Aliu, A (2020). Aspect-term extraction from Albanian reviews with topic modeling techniques. ICT Innovations 2020. ISSN 1857-7288.

Blei, David & Ng, Andrew & Jordan, Michael. (2001). Latent Dirichlet Allocation. The Journal of Machine Learning Research. 3. 601-608.

Devlin, Jacob & Chang, Ming-Wei & Lee, Kenton & Toutanova, Kristina. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Feizollah, Ali & Anuar, Nor & Salleh, Rosli & Amalina, Fairuz. (2014). Comparative Study of K-means and Mini Batch K- means Clustering Algorithms in Android Malware Detection Using Network Traffic Analysis. International Symposium on Biometrics and Security Technologies (ISBAST). 10.1109/ISBAST.2014.7013120.

Karanikolas, N. (2009). Bootstrapping the Albanian Information Retrieval. Fourth Balkan Conference in Informatics, pp. 231-235.

Kote, Nelda & Biba, Marenglen & Kanerva, Jenna & Rönnqvist, Samuel & Ginter, Filip. (2019). Morphological Tagging and Lemmatization of Albanian: A Manually Annotated Corpus and Neural Models.

Mati, Diellza & Hamiti, Mentor & Ajdari, Jaumin & Raufi, Bujar & Selimi, Besnik. (2019). Review of Natural Language Processing tasks in Albanian Language.

Mikolov, Tomas & Chen, Kai & Corrado, G.s & Dean, Jeffrey. (2013). Efficient Estimation of Word Representations in Vector Space. Proceedings of Workshop at ICLR. 2013.

Mimno, David & Wallach, Hanna & Talley, Edmund & Leenders, Miriam & Mccallum, Andrew. (2011). Optimizing Semantic Coherence in Topic Models. EMNLP 2011 - Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference. 262-272.

Nguyen, Dat Quoc & Billingsley, Richard & Du, Lan & Johnson, Mark. (2018). Improving Topic Models with Latent Feature Word Representations.

Panichella, Annibale & Dit, Bogdan & Oliveto, Rocco & Di Penta, Massimiliano & Poshynanyk, Denys & Lucia, Andrea. (2013). How to Effectively Use Topic Models for Software Engineering Tasks? An Approach Based on Genetic Algorithms. Proceedings - International Conference on Software Engineering. 10.1109/ICSE.2013.6606598.

Shahapure, Ketan Rajshekhar & Nicholas, Charles. (2020). Cluster Quality Analysis Using Silhouette Score. 747-748. 10.1109/DSAA49011.2020.00096.

Spahiu, Agim. (2010). 100 fjalet me te shpeshta ne gjuhen shqipe.

Sridhar, Vivek. (2015). Unsupervised Text Normalization Using Distributed Representations of Words and Phrases. 8-16. 10.3115/v1/W15-1502.

Steinskog, Asbjørn & Therkelsen, Jonas & Gambäck, Björn. (2017). Twitter Topic Modeling by Tweet Aggregation. In Proceedings of the 21st Nordic Conference on Computational Linguistics, pages 77–86, Gothenburg, Sweden. Association for Computational Linguistics.

Wang, C., Cho, K., & Gu, J. (2019). Neural machine translation with byte-level subwords.

Xu, Dong & Wu, Shao. (2014). An improved TFIDF Algorithm in text classification. Applied Mechanics and Materials. 651-653. 2258-2261. 10.4028/www.scientific.net/AMM.651-653.2258.