Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20)
Doctoral Consortium Track

On Building an Interpretable Topic Modeling Approach for the Urdu Language

Zarmeen Nasim
Artificial Intelligence Lab, Institute of Business Administration (IBA), Pakistan
znasim@iba.edu.pk

Abstract

This research is an endeavor to combine deep-learning-based language modeling with classical topic modeling techniques to produce interpretable topics for a given set of documents in Urdu, a low resource language. The existing topic modeling techniques produce a collection of words, often uninterpretable, as suggested topics without integrating them into a semantically correct phrase/sentence. The proposed approach would first build an accurate Part of Speech (POS) tagger for the Urdu Language using a publicly available corpus of many million sentences. Using semantically rich feature extraction approaches including Word2Vec and BERT, the proposed approach, in the next step, would experiment with different clustering and topic modeling techniques to produce a list of potential topics for a given set of documents. Finally, this list of topics would be sent to a labeler module to produce syntactically correct phrases that will represent interpretable topics.

1 Introduction

The past decade has seen a rapid increase in the amount of online content. The primary sources of such content include blogs, news, and social networking websites. Online social networking sites not only connect people with each other but these platforms also allow users to discuss their opinions related to any political or social issue. People from different parts of the world can raise their voices in favor or against of any global socio-economic issues. Furthermore, policymakers can get useful insights into public opinion from social networks.

The humongous amount of data generated each day over the internet, however, makes it impossible for people to process the data manually to get useful insights. In the domain of natural language processing, document clustering is a technique for grouping texts that contain similar information. For instance, clustering similar tweets make it easier for humans to understand the topic on which people are sharing their opinions. Similarly, the clustering of news headlines published by various media houses helps the reader to understand different perspectives of the same news.

Even though there has been a significant advancement in the field of document clustering, automatic assignment of labels to each cluster, also known as topic modeling, is still an active area of research. The aim of topic modeling is to discover latent topics that represent the documents in the corpus. The existing statistical approaches produce a list of potential topics but, more often than not, the topics are uninterpretable as the words that constitute a topic are not organized in a syntactically and semantically correct phrase/sentence. The reported research and development are even less advanced for low resource languages, such as Urdu, which is one of the most popular languages of South Asia and is spoken by more than 300 million speakers in various parts of the world [Daud *et al*., 2017]. The few reported works [Shakeel *et al*., 2018; Rehman *et al*., 2018] for the Urdu language focused on the classical methods and have not exploited the power of deep learning either in isolation or in combination with the traditional statistical methods. With the language support provided by Twitter and various other online platforms, a vast corpus for the Urdu language is available in digital format. The availability of corpora opens research opportunities in Urdu language processing. It is worth mentioning that the current techniques developed for document clustering in the English language cannot be directly applied to the Urdu language due to the morphological differences between both languages.

The objective of this research study is to build an interpretable topic modeling framework for Urdu text. The focus will be on short-length documents such as news headlines and tweets. Topic modeling on a short-length text is challenging due to the sparsity of short text. It is expected that the framework will not only cluster similar documents but will also produce interpretable topics for each cluster.

2 Research Questions

The aim of this research study is to answer the following two research questions.

a) Can we split the collection of short-length text documents into meaningful clusters?
b) Can we assign meaningful labels to the clusters?

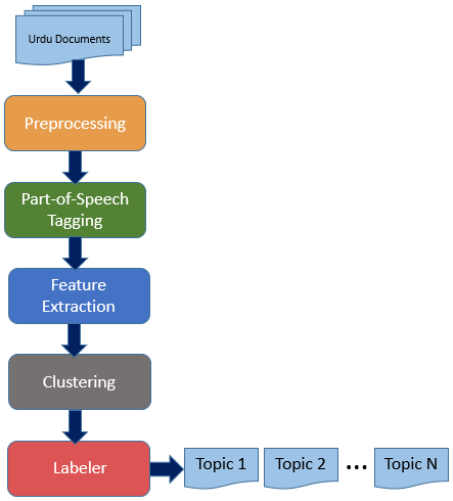

Figure 1: Proposed Methodology

3 Proposed Methodology

As discussed above, the focus of this research is to produce interpretable topics for short-length Urdu documents. Figure 1 shows the workflow of the proposed methodology. The input data will be first passed through a pre-processing module for removal of punctuation, diacritics, URLs, white spaces, and non-Urdu characters before being passed to different modules for further processing. The experiments will be performed on Urdu News headlines and Tweets dataset. Each module along with the expected set of experiments is briefly described below.

3.1 POS Tagging

Tagging documents with part-of-speech will be helpful while deciding the topic of a text document. A POS Tagger trained for the Urdu Language will be used to tag the tokens in each news headline.

3.2 Feature Extraction Module

The feature extraction module will extract relevant features from the text. In this study, several experiments will be conducted with various features which include term frequency-inverse document frequency (TF-IDF), embeddings obtained from Word2Vec models [Mikolov *et al*., 2013] and contextualized word embeddings from state-of-the-art models such as BERT [Devlin *et al.*, 2018].

3.3 Clustering

The clustering module will take a feature matrix as input and will produce semantically coherent clusters. It is expected that these clusters will represent the underlying topics that the input text contains. This research aims to perform a series of experiments to identify the method that works best with different word embeddings. We will evaluate various clustering algorithms which include K-Means and its variants, Affinity propagation and others. We will also experiment with the classical topic modeling technique using Latent Dirichlet Allocation (LDA).

3.4 Labeler

The labeler module will take the topmost representative words and phrases from each cluster and will use the closest phrase as the topic of the cluster. The closeness will be determined by computing the cosine similarity between cluster centroid embedding and phrase embeddings.

4 Future Work

This research aims at building an interpretable topic modeling framework for a low-resource language Urdu. Some preliminary experiments have been done using K-means clustering and LDA topic modeling algorithm on Urdu News headlines and Urdu Tweets datasets. These experiments have shown that the classical topic modeling technique using LDA is not producing high-quality clusters due to the sparsity of short-length text.

Currently, we are working on building a large corpus of Urdu documents for training Word2Vec models. This collection of Urdu documents will be used to fine-tune the BERT model for feature extraction purposes. In the future, we intend to propose a hybrid technique of using classical LDA-based topic modeling with neural embeddings. For labeling clusters, we aim to propose an algorithm that will make use of linguistic features along with word embeddings to construct a meaningful topic.

References

[Daud *et al.*, 2017] Ali Daud, Wahan Khan, and Dunren Che. Urdu language processing: a survey. *Artificial Intelligence Review*, 47(3): 279-311, 2017.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[Mikolov *et al.*, 2013] 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*, pages 3111–3119, 2013.

[Rehman *et al.*, 2018] Anwar U. Rehman, Zobia Rehman, Junaid Akram, Waqar Ali, Munam Ali Shah, and Muhammad Salman. Statistical Topic Modeling for Urdu Text Articles. In *Proceedings of 24th International Conference on Automation and Computing (ICAC)*, pages 1–6, September 2018. IEEE.

[Shakeel *et al*., 2018] Khadija Shakeel, Ghulam Rasool Tahir, Irsha Tehseen, and Mubashir Ali. A framework of Urdu topic modeling using latent dirichlet allocation (LDA). In *Proceedings of 8th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 117–123, January 2018. IEEE.