Inferring the Topic(s) of Wikipedia Articles

Marina Zavalina, Sarthak Agarwal, Chinmay Singhal, Peeyush Jain
Center for Data Science, New York University

Abstract

Developing a multilingual topic classification model that performs better than Wikipedia's current model and scales easily to languages other than English.

Introduction

Given the variety of Wikipedia articles, it is incredibly useful to classify them into a smaller set of general categories. The current gradient-boosting *Drafttopic* model is simple and effective, but is currently implemented only for English and is difficult to scale to more languages. We present here a simpler Bag-of-Words (BoW) model that significantly improves the classification performance and also investigate alternative approaches such as LSTMs, GNNs among others, that scale better to more languages.

Data:
115K articles in English (~2% of all Wikipedia articles) and 33K articles in English, Russian and Hindi, which are aligned.

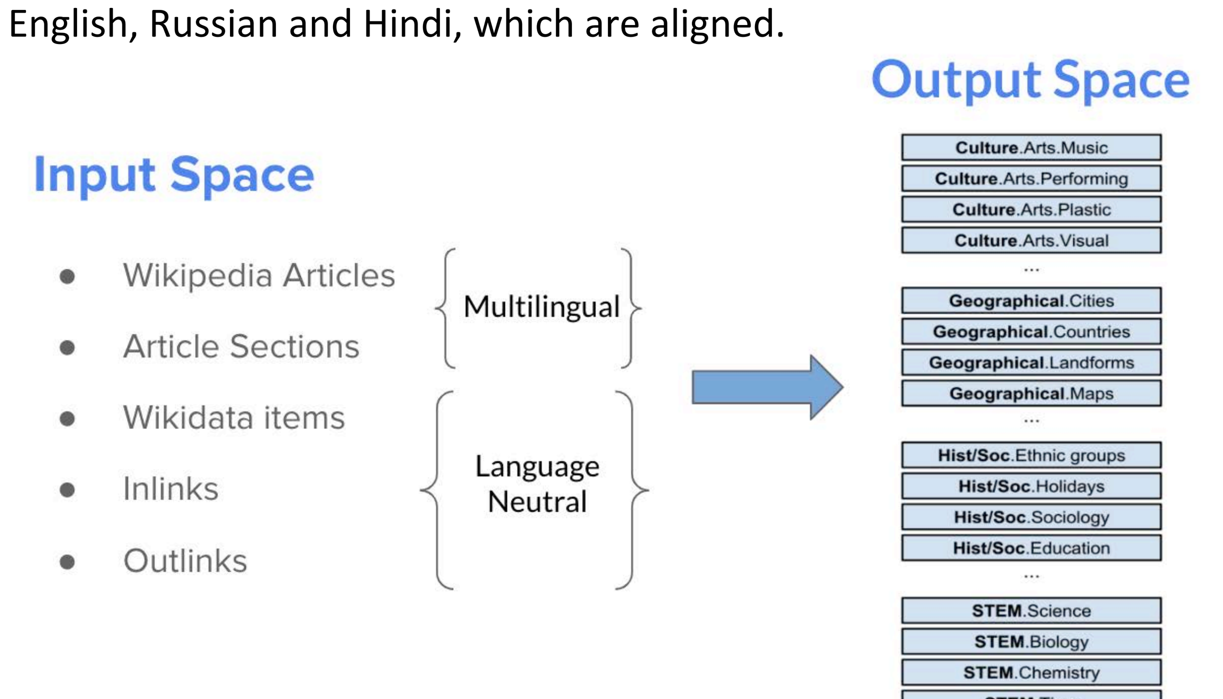

Figure 1. Data overview

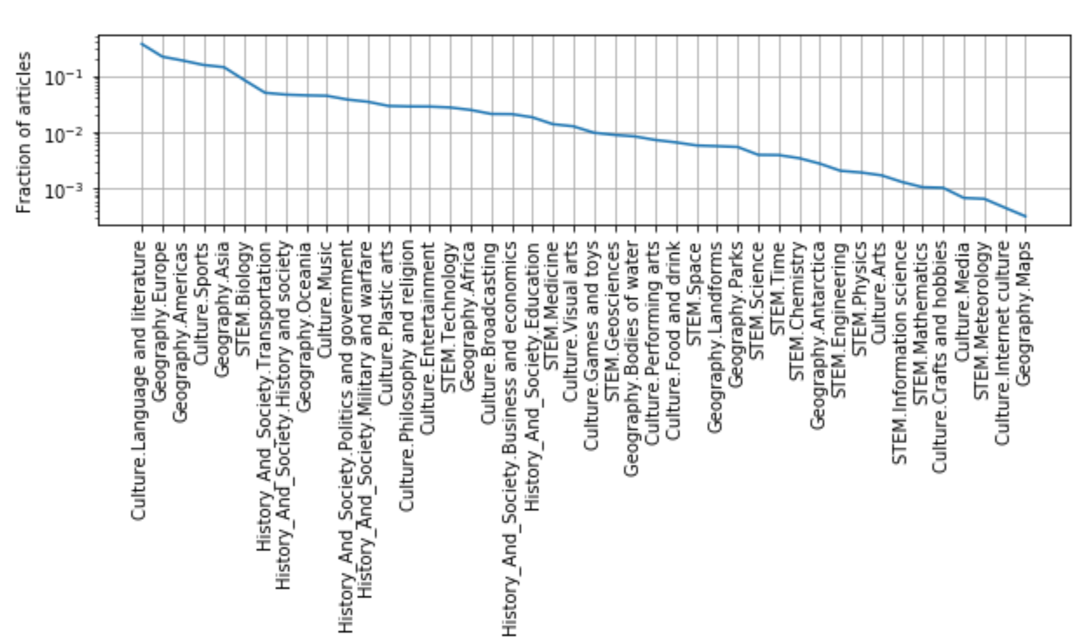

Figure 2. Distribution of categories

Methods

BoW NN Model: The model uses an unordered document representation by taking an average of the fastText embeddings for all words in an article. We pass this representation through a neural network with hidden layers, apply sigmoid at the end, and output scores for the 44 categories.

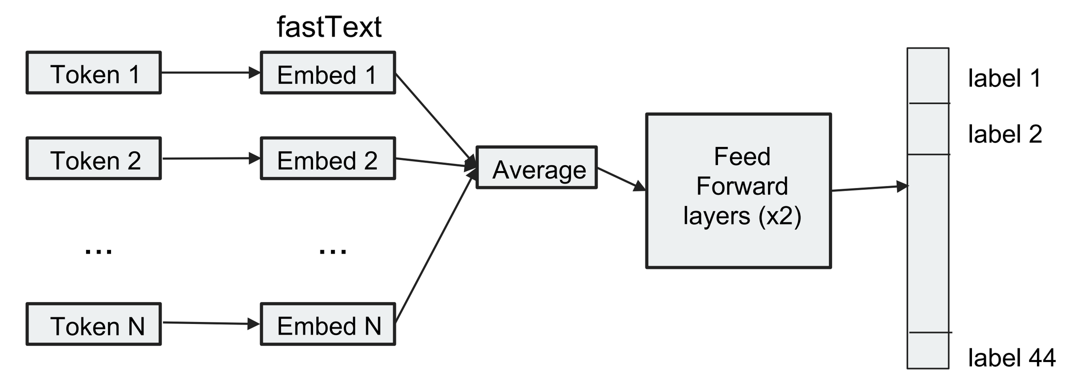

Figure 3. BoW NN model architecture

Multilingual Embeddings: For a multilingual approach, we use the best BoW model trained on English articles and plug in the aligned embeddings for other languages (Russian or Hindi for our experiments). Another approach is to train on a mix of articles in different languages.

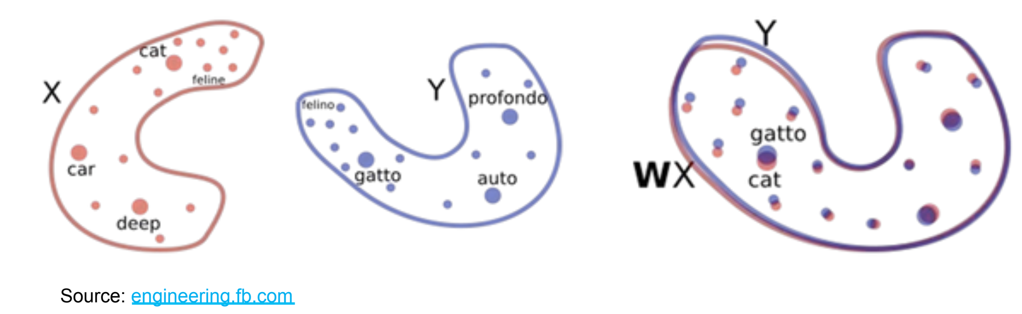

Figure 4. fastText aligned word embeddings

Alternative Text Classification Models: We explore advanced networks such as LSTM, self-attention based transformers. Another approach we experiment is to replace the self attention weights with softmax of Inverse Document Frequencies (IDF) of the words in the articles.

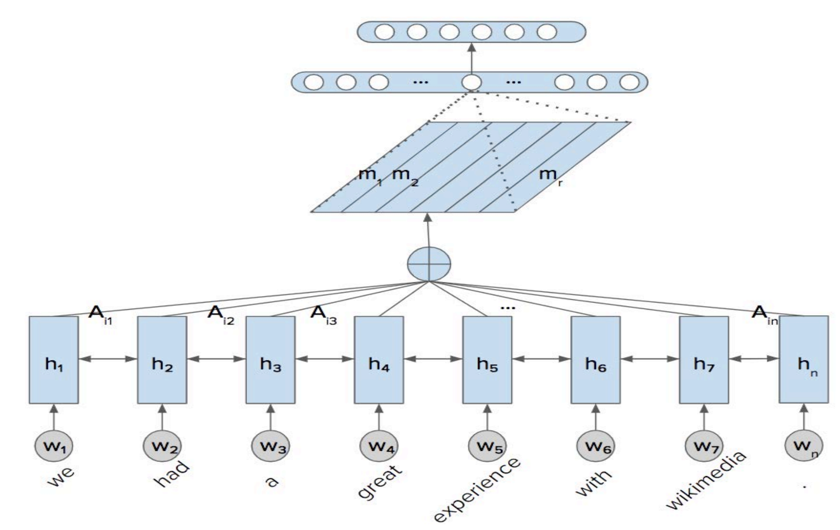

Figure 5. Self Attention LSTM Architecture

Language Independent Models: Wikipedia articles are comprised of alternate metadata like section headers and links. We evaluate the BoW NN model performance on section headers and featurized inlinks & outlinks. Additionally, we train a Graph Neural Network using network connections as features.

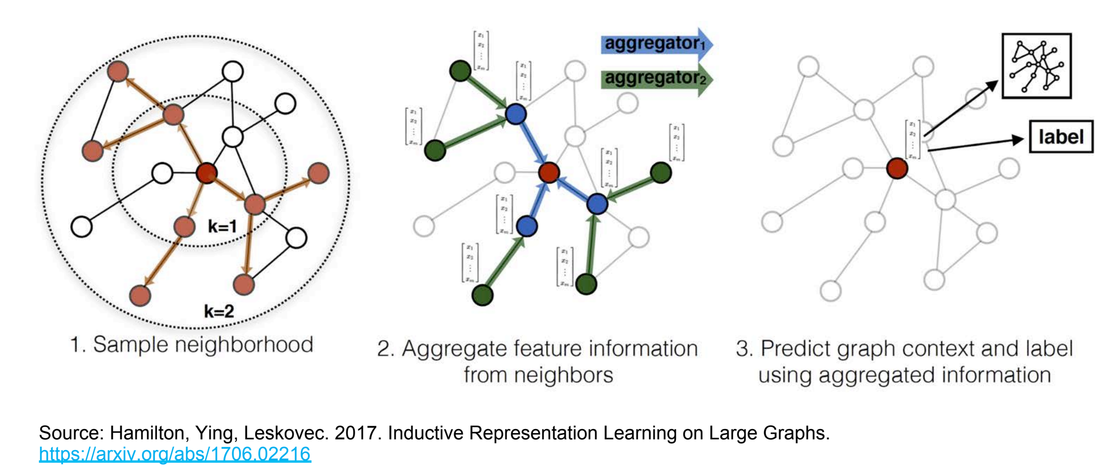

Figure 6. GraphSAGE Architecture

Transfer Learning Approach: We explore the following setting: given a trained classification model and sufficient labeled examples or just a few examples for a new category (not seen before), can we classify articles belonging to this new category accurately.

Results: Language Dependent Models

- **Neural network BoW** model improves upon existing Gradient Boosting model - **from 0.668 to 0.816 micro F1 score**.
- More advanced architectures, such as **LSTM** and **Self Attention LSTM**, don't further improve the score.
- **Inverse Document Frequency (IDF)** can be used as attention weights, reducing the number of parameters to train.

Model	Micro Precision	Micro Recall	Micro F1 Score
Existing model (drafttopic)	0.826	0.576	0.668
Bag of Words NN	0.830	**0.802**	**0.816**
Basic LSTM	**0.867**	0.754	0.807
Self Attention LSTM	0.861	0.774	**0.816**
Inverse Document Frequency LSTM	0.833	0.777	0.804
Transformer	0.841	0.766	0.801

- **Frozen model** trained on **English** articles does not perform well on a new language (**Russian**). Possible reasons can be poor alignment of multilingual fastText embeddings or model overfitting on English language.
- Micro F1 score for model trained on a new language improves when the weights are **initialized with a pre-trained English model**.
- For a given language, **multilingual** models (trained on mix of aligned English, Russian and Hindi articles, 10k each) perform comparable to **monolingual** models (trained on articles in given language, 10k).

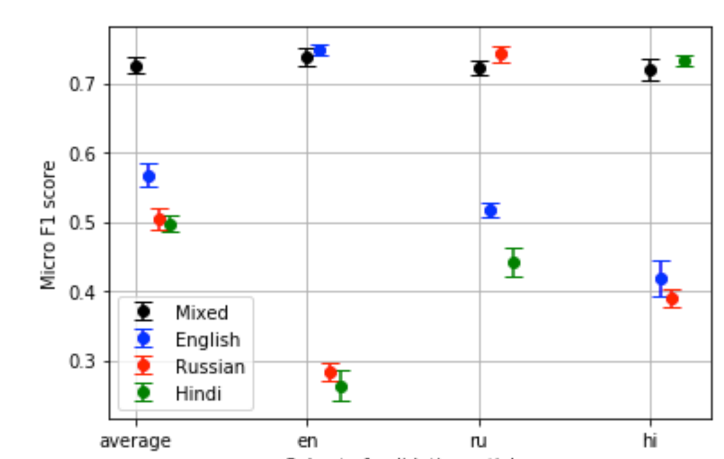

Figure 7. Performance of Multilingual vs. Monolingual BoW NN

Results: Language Independent Models

- **Links** and **section header** based approaches don't perform well in comparison to the approaches using the whole article text.
- GNN model that uses just connections between articles performs surprisingly well, **GraphSAGE** achieves **0.642 micro F1 score**.

Metadata	Micro Precision	Micro Recall	Micro F1 Score
Section Text	0.762	0.403	0.531
Wiki Links (BoW)	0.550	0.513	0.534
Wiki Links (GraphSAGE)	0.649	0.636	**0.642**

- The results suggest that **transfer learning** is indeed a feasible option to further explore once we can have a larger dataset for fine tuning.

Model	Accuracy
Feature extractor model	84.98%
Fine-tuned model	90.37&
Class embedding model	78.84%
Reference document model	77.95%

Conclusion and Future Work

- BoW NN is the best model from our experiments with a micro F1 score of 0.816. It trains much faster and has fewer parameters compared to Self-Attention LSTM model, which has a similar score.
- Graph NN model shows great potential as it achieves 0.642 micro F1 score using only links as input. This approach is scalable for all languages and we can experiment further by including node features such as text embeddings for articles.
- Our LSTM based networks do not perform as expected and we hypothesize that model performance will increase with a larger dataset.
- All categories in the dataset are not evenly balanced and we can explore weighted loss functions.

Acknowledgements

We would like to thank our amazing mentor Isaac Johnson, Wikimedia Foundation for all his help and support throughout the project and also for all the effort he put in getting us quality datasets to work with. His ideas helped a lot in giving a proper direction to our project. We would also like to thank Anastasios Noulas, New York University for his constructive feedback and all his suggestions about graph models.

References

1. *Sumit Asthana and Aaron Halfaker. 2018. With Few Eyes, All Hoaxes Are Deep.Proc. ACM Hum.-Comput. Interact. 2, CSCW, Article 21 (Nov. 2018), 18 pages. https://doi.org/10.1145/3274290*
2. *Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. Transactions of the Association for Computational Linguistics 5 (2017), 135–146.*
3. *S. J. Pan and Q. Yang. 2010. A Survey on Transfer Learning. IEEE Transactions on Knowledge and Data Engineering 22, 10 (Oct 2010), 1345–1359. https://doi.org/10.1109/TKDE.2009.191*
4. *Lin et al. 2017. A Structured Self-attentive Sentence Embedding. CoRR (Aug 2018). http://arxiv.org/abs/1703.03130*
5. *William L. Hamilton, Rex Ying and Jure Leskovec. 2017. Inductive Representation Learning on Large Graphs. CoRR, http://arxiv.org/abs/1706.02216*

github.com/mmarinated/topic-modeling