

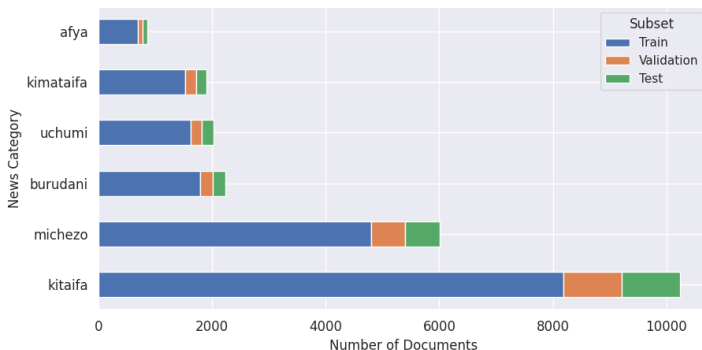
Graph Convolutional Network for Swahili News Classification

Alexandros Kastanos and Tyler Martin

April 2021

Introduction

- Swahili is an under-represented language in NLP research
 - Annotated datasets and accessible benchmarks
 - Techniques developed for high-resource languages may not transfer to a low-resource context
 - Purpose-built tools and libraries
- Zenodo Swahili News Classification Dataset [David, 2020]



Semi-Supervised Swahili News Classification

- Semi-supervised context is applicable in low-resource NLP
 - label sparsity
- Key features and contributions:
 - Set of accessible benchmarks
 - First application of Graph Neural Networks (GNNs) on an African language
 - Memory efficient variant of Text Graph Convolutional Network [Yao, 2019]

Baselines

- Traditional NLP benchmarks
 - TF-IDF
 - Counts
 - Averaged fastText embeddings [Bojanowski, 2017]
 - PV-DBOW [Le, 2014]
 - PV-DM [Le, 2014]

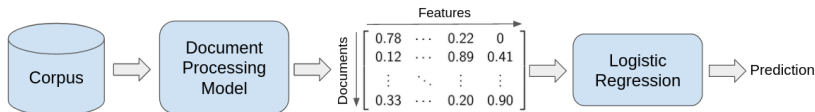
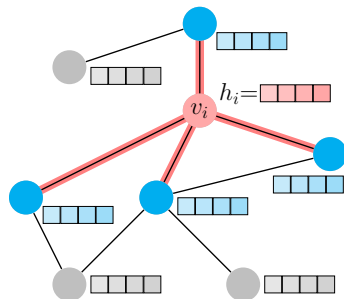


Figure: High-level overview of the baseline model pipeline.

Graph Neural Networks for low-resource NLP

- A corpus contains an implicit graph structure
 - Semantic and syntactic relationships
- GNNs for semi-supervised learning
 - Aggregate information from a neighbourhood of nodes



Semi-supervised benchmark comparison

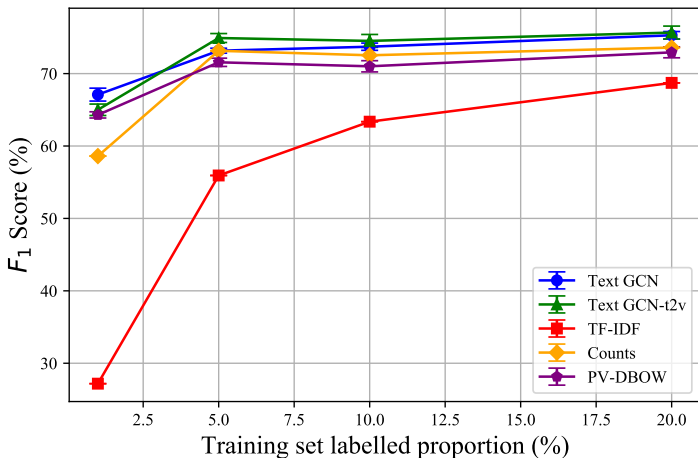
20% of the training set is labelled

Model	Accuracy (%)	Macro F ₁ (%)
TF-IDF	83.07 ± 0.00	68.72 ± 0.00
Counts	83.32 ± 0.00	73.60 ± 0.00
fastText	67.47 ± 0.00	32.41 ± 0.00
PV-DBOW	81.64 ± 0.47	72.93 ± 0.75
PV-DM	77.01 ± 0.38	67.50 ± 0.64
Text GCN	84.62 ± 0.10	75.29 ± 0.52
Text GCN-t2v	85.40 ± 0.22	75.67 ± 0.90

Table: Mean and standard deviation test set score comparison

Semi-supervised benchmark comparison

Reducing the number of labels present



Conclusion

- In our paper, we show that:
 - Text GCN for semi-supervised Swahili news classification outperforms traditional methods
 - Representing a corpus as a graph
- Future work to consider:
 - Alternative graph structures from text
 - Inductive GNN methods

References

-  Davis David (2020)
Swahili: News classification dataset.
-  Liang Yao, Chengsheng Mao, and Yuan Luo (2019)
Graph Convolutional Networks for Text Classification.
AAAI pages 7370–7377.
-  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.
-  Quoc Le and Tomas Mikolov. (2014)
Distributed representations of sentences and documents.
Proceedings of the 31st International Conference on Machine Learning, volume 32 of Proceedings of Machine Learning Research pages 1188–1196.

Find us on Github:

