Graph Convolutional Network for Swahili News Classification

Alexandros Kastanos
Independent
alecokastanos@gmail.com

Tyler Martin
Independent
tyler.a.martin12@gmail.com

Abstract

In this work, we demonstrate the ability of Text Graph Convolutional Network (Text GCN) to surpass the performance of traditional natural language processing benchmarks on the task of semi-supervised Swahili news categorisation. Our experiments highlight the more severely label-restricted context often facing low-resourced African languages. We build on this finding by presenting a memory-efficient variant of Text GCN which replaces the naive one-hot node representation with a bag of words representation.

1 Introduction

Text classification is a fundamental application of natural language processing (NLP) in news media. From topic classification (Wang and Manning, 2012) and content moderation (Bodapati et al., 2019) to fake news detection (Wang, 2017), the impact of improving the automated classification processes of news texts has a profound influence on the daily experience of newsreaders.

Despite the importance of news categorisation and Swahili being one of the most widely spoken languages in Africa (Eberhard et al., 2021), there is a shortage of published work on text classification for Swahili. This under-representation of Swahili, as well as other low-resource languages, manifests itself in several ways including the ongoing scarcity of freely-available high-quality datasets, a shortage of accessible comparative benchmarks, and a lack of purpose-built software tools and libraries (Orife et al., 2020; Niyongabo et al., 2020; Caswell et al., 2021). Additionally, there is limited literature comparing techniques developed and regularly applied to high-resource languages in low-resource languages such as Swahili.

Our work aims to counter these challenges by contributing the following:

- We provide a set of easily accessible traditional NLP models as a benchmark for semi-supervised Swahili news classification.
- We use these benchmarks to compare against Text Graph Convolutional Network (Text GCN), a model initially developed for English. As far as we are aware, this is the first time a Graph Neural Network has been applied to text classification for any African language dataset.
- We present experiments which highlight the comparative performance of these models in a semi-supervised setup where the training set has a low proportion of labelled news documents.

2 Graph Neural Networks

Graph Neural Networks (GNNs) are a family of architectures that operate directly on irregularly structured graphs (Gori et al., 2005; Scarselli et al., 2009; Battaglia et al., 2018). The underlying mechanism of a GNN is that information is propagated through the network by each node updating its hidden state with aggregated information from a neighbourhood of nodes. This enables GNNs to generate rich representations by taking into account both the input features of the nodes and the graph structure.

2.1 Text Graph Convolutional Networks

Recognising that a corpus contains both syntactic and semantic relationships, we can represent a corpus through a graph structure. Text GCN (Yao et al., 2019) proposes a method for constructing a graph that captures the global relationships between all words and documents in the corpus. By modelling all words and documents as nodes using a one-hot feature encoding, a heterogeneous graph with a weighted adjacency matrix can be used to capture these global relationships. Edges representing word-word co-occurrences are formulated using Positive Pointwise Mutual Information (PPMI) over a fixed window size while word-document interactions are modelled using their TF-IDF value.

This graph representation is then fed into a two-layer Graph Convolutional Network (Kipf and Welling, 2017), formalised by equation 1, where Θ_0 and Θ_1 are trainable parameters, X is the input node representation, and $\widetilde{A}$ is the adjacency matrix after undergoing the *renormalisation trick*.

$$\hat{Y} = \text{softmax}\left(\tilde{A}\,\text{ReLU}\left(\tilde{A}X\Theta_0\right)\Theta_1\right) \quad (1)$$

In a semi-supervised classification setting, gradient descent can be used to train the model by calculating the cross entropy loss over the subset of labelled nodes in the training set.

3 Experiments

We use the Swahili News Classification dataset (David, 2020) which contains 23,266 news texts, each labelled as one of six possible categories. Each document is passed through a preprocessing pipeline which includes removing stop words, removing Twitter meta information, and stemming using the SALAMA Language Manager (Hurskainen, 2004, 1999). The training, validation, and test sets are generated using an 8:1:1 split.

Three traditional NLP models are used to form a comparative baseline. These are the Term Frequency Inverse Document Frequency (TF-IDF), Term Frequency Count (*Counts*), and `doc2vec` Paragraph Vector Distributed Bag of Words (PV-DBOW) (Le and Mikolov, 2014) models. Each baseline model converts a document into a feature vector, $X \in \mathcal{R}^{300}$, which is then passed to a logistic regression classifier.

We implement both the vanilla Text GCN model (Yao et al., 2019), which uses a one-hot representation for the input features of each node, and a memory-efficient variant Text GCN-t2v (`text2vec`), which uses `word2vec` and `doc2vec` representations for the word and document nodes respectively. The dimensions of the Text GCN-t2v node representations match those used for the baseline PV-DBOW model. Each experiment is repeated 5 times to obtain mean and standard deviation values[1].

Table 1 demonstrates that the two Text GCN variants outperform the three baseline models when

Model	Accuracy (%)	Macro F_1 (%)
TF-IDF	83.07 ± 0.00	68.72 ± 0.00
Counts	83.32 ± 0.00	73.60 ± 0.00
PV-DBOW	81.64 ± 0.47	72.93 ± 0.75
Text GCN	84.62 ± 0.10	75.29 ± 0.52
Text GCN-t2v	$\mathbf{85.40 \pm 0.22}$	$\mathbf{75.67 \pm 0.90}$

Table 1: Comparison of the mean and standard deviation test set accuracy and F_1 scores for all models.

20% of the training set nodes are labelled. Although the vanilla Text GCN and Text GCN-t2v perform similarly, the more compact input feature representation allows Text GCN-t2v to reduce the training time and cloud costs by factors of 5 and 20 respectively[2].

Figure 1 provides the macro F_1 scores for each model when presented with 1%, 5%, 10%, and 20% of the training set labels. This highlights that the Text GCN variants compare particularly well against the *Counts* and TF-IDF benchmarks when the proportion of training labels drops below 5%.

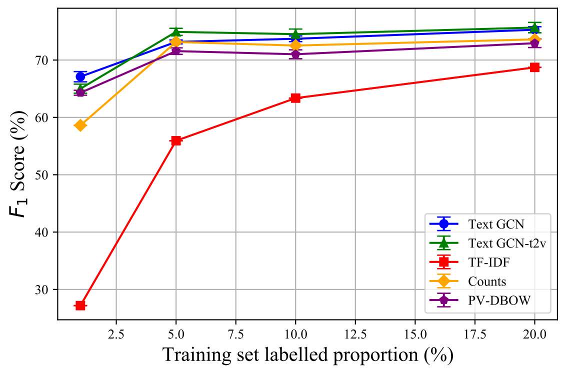

Figure 1: Test set macro F_1 scores for different labelled proportions of the training set.

4 Conclusion

This work demonstrates that Text GCN is able to outperform traditional NLP models for the task of semi-supervised Swahili news classification. Furthermore, the proposed Text GCN-t2v variant provides a meaningful reduction in memory and training cost compared to the vanilla Text GCN model without significantly sacrificing performance. Ongoing work includes a wider investigation into inductive GNN approaches, as well as alternative methods for representing a corpus as a graph.

[1]This abstract is based on Kastanos and Martin (2021). Code available at `https://github.com/alecokas/swahili-text-gcn`

[2]Pricing listed at `https://aws.amazon.com/ec2/pricing/on-demand/` as of February 2021.

References

Peter W. Battaglia, Jessica B. Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinícius Flores Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, Çaglar Gülçehre, H. Francis Song, Andrew J. Ballard, Justin Gilmer, George E. Dahl, Ashish Vaswani, Kelsey R. Allen, Charles Nash, Victoria Langston, Chris Dyer, Nicolas Heess, Daan Wierstra, Pushmeet Kohli, Matthew Botvinick, Oriol Vinyals, Yujia Li, and Razvan Pascanu. 2018. Relational inductive biases, deep learning, and graph networks. *CoRR*, abs/1806.01261.

Sravan Bodapati, Spandana Gella, Kasturi Bhattacharjee, and Yaser Al-Onaizan. 2019. Neural word decomposition models for abusive language detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 135–145, Florence, Italy. Association for Computational Linguistics.

Isaac Caswell, Julia Kreutzer, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Javier Ortiz Suárez, Iroro Orife, Kelechi Ogueji, Rubungo Andre Niyongabo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2021. Quality at a glance: An audit of web-crawled multilingual datasets.

Davis David. 2020. Swahili : News classification dataset.

David M. Eberhard, Gary F. Simons, and Charles D.Fennig. Ethnologue: Languages of the world [online]. 2021.

Marco Gori, Gabriele Monfardini, and Franco Scarselli. 2005. A new model for learning in graph domains. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 2, pages 729–734 vol. 2.

Arvi Hurskainen. 2004. Swahili language manager: A storehouse for developing multiple computational applications. *Nordic Journal of African Studies*, 13(3):363 – 397.

Avri Hurskainen. 1999. Salama: Swahili language manager. *Nordic Journal of African Studies*, 8:139–157.

Alexandros Kastanos and Tyler Martin. 2021. Graph convolutional network for swahili news classification.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1188–1196, Bejing, China. PMLR.

Rubungo Andre Niyongabo, Qu Hong, Julia Kreutzer, and Li Huang. 2020. KINNEWS and KIRNEWS: Benchmarking cross-lingual text classification for Kinyarwanda and Kirundi. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5507–5521, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Iroro Orife, Julia Kreutzer, Blessing Sibanda, Daniel Whitenack, Kathleen Siminyu, Laura Martinus, Jamiil Toure Ali, Jade Z. Abbott, Vukosi N. Marivate, Salomon Kabongo, Musie Meressa, Espoir Murhabazi, Orevaoghene Ahia, Elan Van Biljon, Arshath Ramkilowan, Adewale Akinfaderin, Alp Öktem, Wole Akin, Ghollah Kioko, Kevin Degila, Herman Kamper, Bonaventure Dossou, Chris Emezue, Kelechi Ogueji, and Abdallah Bashir. 2020. Masakhane - machine translation for africa. *CoRR*, abs/2003.11529.

Franco Scarselli, Marco Gori, Ah C. Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2009. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80.

Sida Wang and Christopher Manning. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 90–94, Jeju Island, Korea. Association for Computational Linguistics.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Graph convolutional networks for text classification. In *AAAI*, pages 7370–7377.