Evaluating the Impact of ASR Transcription Quality on Topic Modeling for Setswana

Nerina Borchard **Rueben van der Westhuizen** **Zion van Wyk**
University of Pretoria
u21537144, u21434809, u21655325

Project Repository:

https://github.com/Evaluating-ASR-Transcription-Topic-Modeling-for-Setswana

Abstract

This study investigates the impact of Automatic Speech Recognition (ASR) transcription quality on topic modeling for Setswana, a low-resource South African language. Leveraging ASR-generated transcriptions, the research evaluates whether text enhancement techniques can improve the coherence and quality of extracted topics. Using data from Setswana podcasts and a fine-tuned ASR model, three topic modeling approaches: Non-negative Matrix Factorization (NMF), Latent Dirichlet Allocation (LDA), and BERTopic were applied and evaluated through UMass and NPMI coherence metrics. The findings reveal that transcription quality significantly affects topic coherence, with NMF emerging as the most effective model. Despite resource constraints, including limited native language tools and dependence on free ASR software, enhancements such as sentence segmentation and stop-word removal showed potential for improving outcomes. This research underscores the importance of investing in robust ASR models and NLP resources for underrepresented languages, enabling more inclusive language technologies.

1 Introduction

Topic modelling is a well-established text analysis technique used to uncover hidden thematic structures in large collections of documents[29]. By identifying patterns of word co-occurrence, topic models aid in categorization of text into coherent themes. In parallel, Automatic Speech Recognition (ASR) systems play a role in unlocking the value of audio content. These systems convert spoken language to text for several use-cases, like transcription, translation, and voice-command interfaces.[28]

However, applying these technologies to low-resource languages (LRLs) remain a significant challenge due to the scarcity of annotated datasets. South Africa is a majority LRL-speaking country, making this issue particularly urgent, with 9 of the 11 languages being low-resourced.[19] Consequently, tools for speech recognition and text analysis[18] are disproportionately underdeveloped for the linguistic minority. While there has been increasing interest in ASR and topic modelling individually, little work has explored the intersection of these two areas, particularly in South African context. Existing works will be discussed in 2 below.

This research aims to bridge the gap between ASR and topic modelling by evaluating how well topic modelling techniques perform on ASR-generated transcriptions of LRL speech data, and what can be done to improve them. The focus will be on Setswana, one of South Africa's official languages, as a case study. The project investigates the extent to which existing topic modelling techniques can extract coherent themes from ASR-generated text and whether enhancements to the raw transcriptions can significantly improve the quality of topic modelling output.

The scope of this research is limited to a single language and a dataset of spoken Setswana podcast content. Two key research questions will be addressed:

1. Can existing topic modelling techniques be effectively applied to ASR-generated transcriptions in Setswana?

2. What combination of enhancement strategies, such as sentence segmentation, most significantly improve topic coherence?

In order to achieve this, the effectiveness of performing topic modelling on the raw transcriptions from an ASR model will first be explored. The transcriptions will then be enhanced using a set of chosen strategies and the topic modelling process will be rerun. From there, we will measure any improvements in topic coherence and interpretability. By systematically comparing the impact of these enhancement methods, this research contributes to both improving NLP processes for South African languages and to deepening the understanding of how ASR and topic modelling can be better integrated. The findings may contribute to expanding access to speech and language technologies for linguistically underrepresented communities.

2 Background

2.1 Introduction

This research aims to explore the potential for topic modelling on ASR transcriptions and what available options exist for text enhancement and topic modelling improvement. In this literature survey, we will explore the existing research in topic modelling, ASR systems, the overlap between the two, as well as any gaps that occur within this research.

2.2 Topic Modelling

As previously mentioned, topic modelling is a machine learning technique that assigns topics to documents in a given corpus. It has applications in a variety of domains such as sociology[5], political science[9], and literary studies[13] among many others.

The most frequently employed method of approaching topic modelling is Latent Dirichlet allocation (LDA). In short, it proposes that "documents are represented as random mixtures over latent topics, where each topic is characterised by a distribution over words"[2]. Another important method is BERTopic, which converts documents to an embedding representation, reduces the dimensionality and then clusters them into topic representations.[10] Others include Non-Negative Matrix Factorisation (NMF)[8], the Author-Topic model[11] and Dynamic Topic Models which discover topics over time as seen in [1].

To evaluate the results of topic modelling, two methods have been debated: automated and human-centred.[12] Models were initially evaluated using held-out perplexity. Perplexity quanitifies the "uncertainty" of a model when it predicts the next token in a sequence. However, this method was proven to disagree with human interpretation.[4] Automated coherence then became the preferred method of evaluation for majority of researchers, yet is now almost solely used without human evaluation. The most popular coherence metrics found to be used are Normalised Pointwise Mutual Information (NPMI) and UMass [16]. NPMI is a statistical measure[3] to calculate the association between two words in a corpus. UMass measures "the log conditional probability between ordered word-pair in a topic".[21]

2.3 Topic Modelling on ASR Transcriptions

Research on topic modelling in South Africa often involves extracting public sentiment from some text data. This can be seen in [23] and [14] in which tweets on Covid-19 were used to model societal attitudes and discussions happening in the pandemic context. [22] uses news articles to provide insights into the South African elections.

To date, there is no available work covering topic modelling on ASR transcriptions in South African languages. However, there are studies in other LRLs, like [7] which attempts to enhance topic modelling on ASR-generated Danish text by injecting noise into the transcriptions. [17] attempts topic classification in LRLs on transcriptions using two tokenization systems. Both studies showed a significant improvement in the results after experiments and serve as inspiration for the pipeline we would like to build.

2.4 Gaps in the Literature

Given the research pipeline we would like to build, the following gaps were identified:

- Data is in a high-resourced language. Many papers that are related to the ASR-topic modelling overlap we like to investigate are in English, as seen in [15] and [25].

- There is a focus on audio enhancements for ASR transcriptions instead of text enhancements. [27]

- There is a focus on supervised topic classification or segmentation instead of unsupervised topic modelling, as seen in [6].

3 Methodology

3.1 Data Selection

The dataset selected for this research is the Setswana podcast data on Covid-19 acquired by the Data Science for Social Impact (DSFSI) lab. This dataset was chosen for two reasons. First, this is a dataset that hasn't been worked with yet and has no annotations. The team would like to explore its potential. Secondly, there are many native Setswana speakers in the research team's immediate surroundings to aid in the qualitative evaluation of the research.

3.2 Preprocessing Steps

The dataset consists of 59 episodes, all of varying lengths. See 2 for the duration of each episode. All files are in .mp3 format.

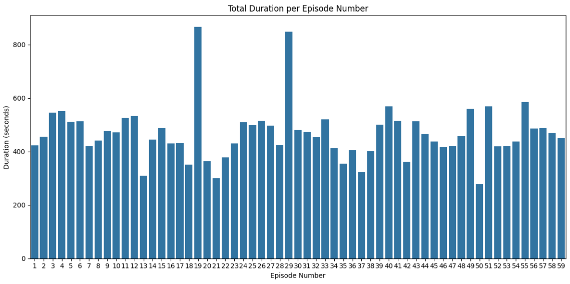

Figure 1: Bar graph representing total duration (s) per episode

Each episode, on average, was between 400 and 500 seconds long, making the audio files tedious to process. To make the data easier for processing, the audio was split into 30-second chunks using python's pydub library. This was to ensure that the transcriptions are easier to process. The audio was also resampled to 16 kHz, the industry standard for ASR models.

Audio analysis on features like spectrograms would not largely impact our results, as the goal of the experiments is to measure the impact of NLP techniques to improve the raw ASR transcriptions. Only basic audio preprocessing was implemented for this reason. Analysis on the text will happen as the experiment is conducted.

3.3 Enhancing the Transcriptions

3.3.1 Improving Word Separation

During preliminary preprocessing, we noticed that the ASR output frequently concatenated distinct Tswana morphemes, often due to misrecognized spaces or phoneme merges, resulting in strings like "konkabokamooso" rather than the correct "konka bokamoso." To automatically correct these cases, we built a Maximum Matching segmenter driven by a curated Setswana lexicon. First, we loaded a high-coverage wordlist (skipping metadata headers) and normalized all entries to lowercase. Then, for each transcription string, our algorithm scans from the current position to the end of the line, greedily matching the longest substring found in the lexicon; if no match is found, it emits a single character as a fallback. This process not only splits misjoined tokens but also preserves legitimate singletons and out of vocabulary items. In practice, applying this segmentation reduced the incidence of fused tokens, significantly improving downstream token counts and coherence in our topic modeling experiments.

3.3.2 Splitting sentences

To further normalize our transcriptions, we implemented a sentence-splitting module that breaks long, segmented strings into smaller "sentence" units. We leverage the PuoBERTa-POS tokenizer and model (dsfsi/PuoBERTa-POS) to obtain subword tokens, then apply two simple heuristics:

1. **Length threshold:** whenever the current token buffer reaches 30 tokens, we flush it as one sentence.

2. **Conjunction boundary:** if a token matches a high-frequency Tswana connector (e.g., *ke, mme, fa, jalo, le, ya, a, go, e*), we treat it as an end marker and start a new sentence thereafter.

In practice, each input string is first checked for emptiness, then tokenized into subwords. We accumulate tokens until either rule fires, convert the buffer back to text and repeat. Any leftover tokens form a final sentence. When applied to our segmented CSV, this approach reduces per unit complexity for downstream syntactic and semantic analyses, and aligns better with the POS model's expected context window.

3.3.3 Normalization Pipeline

To standardize our Setswana transcriptions for downstream analysis, we implemented a two stage normalization pipeline comprising, first, morphological lemmatization and second stop-word filtering:

1. **Lemmatization:** We developed a rule-based verb lemmatizer tailored to Setswana morphology. Starting from each token, the lemmatizer repeatedly applies a sequence of suffix stripping and replacement rules (e.g. perfect, passive, causative, applicative, reciprocal, neuter passive, and iterative forms), as well as prefix adjustments for reflexive and object markers.

Exception tables prevent over stemming of irregular forms. On a held out test set of 200 verb forms, this component improves accuracy in recovering correct dictionary lemmas, drastically reducing form-variant proliferation in our vocabulary.

2. **Stop-word Removal:** We then removed the top 100 most frequent terms identified via a TF-IDF based heuristic on our combined Setswana corpus(Leipzig + Autshumato) using a curated stop-word list. After filtering, the average document length dropped as expected, while topic coherence in our BERTopic experiments improved (UMass measure), confirming that noise reduction benefits downstream modeling.

By integrating these two modules into a single preprocessing script, we ensure that each transcription is first reduced to its canonical verb forms and then stripped of non informative tokens, yielding a cleaner, more semantically focused input for subsequent topic modeling.

4 Experiments and Results

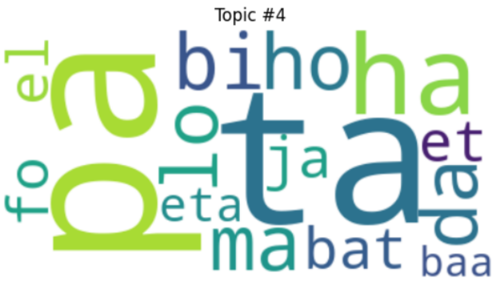

Figure 2: Wordcloud generated from NMF

The raw transcriptions were first translated and evaluated by LDA, NMF, and BERTopic. These transcriptions then underwent our enhancement and normalisation process as described in 3. The primary metrics used to evaluate model performance were Coherence (UMass), which measures topic distinctiveness, and Coherence (NPMI), which assesses word co-occurrence and semantic correlation.

UMass Scores of Transcriptions			
Time	LDA	NMF	BERTopic
Before	-1.9378	-1.1289	-1.5055
After	-1.3313	-0.5773	-0.7556

NPMI Scores of Transcriptions			
Time	LDA	NMF	BERTopic
Before	-0.0341	-0.0176	-0.0381
After	-0.0200	0.0079	-0.0220

While all models showed some improvement after preprocessing, NMF consistently outperformed the others across both coherence metrics.

5 Reflections and Discussion

The improvement between the raw and processed transcriptions is minimal, but still shows that our methodology is effective. Many challenges were faced. The most notable are those of:

- The language barrier: None of the research team members spoke Setswana. Although native speakers were consulted, the challenges presented would have been reduced if a native speaker was on the team.

- A lack of Setswana NLP resources: Although many Setswana corpora were available, there were not many tools like word lists or lemmatizers readily available. This meant the team had to create their own resources when resources already existed, as seen in [20], [24], and [26]. Much trial and error took place.

6 Conclusion

The quality of transcriptions has a significant impact on the effectiveness of topic modelling. While preprocessing and normalization led to moderate improvements (particularly in NMF, which achieved the best coherence scores) overall performance remained constrained by the limitations of free ASR tools, especially when applied to underrepresented languages like Tswana. These tools often produced inconsistent or unclear outputs, affecting both coherence and interpretability.

Despite these challenges, several meaningful and socially relevant topics still emerged. Themes such as “Be Not Afraid” and “Thank you for the good results” reflect public sentiment, uncertainty, and community narratives during the pandemic. These topics offer valuable qualitative insight, even when coherence metrics remained low.

In summary, while the quantitative performance highlights areas for improvement, the qualitative outcomes demonstrate the potential of topic modeling in low-resource settings. Future work should prioritize enhancing transcription accuracy through custom ASR models and language-specific preprocessing, enabling more reliable extraction of meaningful themes from spoken content.

References

[1] Rubayyi Alghamdi and Khalid Alfalqi. 2015. A survey of topic modeling in text mining. *Int. J. Adv. Comput. Sci. Appl.(IJACSA)*, 6(1).

[2] DM Blei, AY Ng, and MI Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3.

[3] Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, 30:31–40.

[4] Jonathan Chang, Sean Gerrish, Chong Wang, Jordan Boyd-graber, and David Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc.

[5] Paul DiMaggio, Manish Nag, and David Blei. 2013. Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of u.s. government arts funding. *Poetics*, 41(6):570–606. Topic Models and the Cultural Sciences.

[6] Zexian Dong, Jia Liu, and Wei-Qiang Zhang. 2019. End-to-end topic classification without asr. In *2019 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, pages 1–5. IEEE.

[7] Raluca Alexandra Fetic, Mikkel Jordahn, Lucas Chaves Lima, Rasmus Arpe Fogh Egebæk, Martin Carsten Nielsen, Benjamin Biering, and Lars Kai Hansen. 2021. Topic model robustness to automatic speech recognition errors in podcast transcripts. *arXiv preprint arXiv:2109.12306*.

[8] Cédric Févotte and Jérôme Idier. 2011. Algorithms for nonnegative matrix factorization with the β-divergence. *Neural computation*, 23(9):2421–2456.

[9] Justin Grimmer. 2010. A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Political Analysis*, 18(1):1–35.

[10] Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *Preprint*, arXiv:2203.05794.

[11] Liangjie Hong and Brian D Davison. 2010. Empirical study of topic modeling in twitter. In *Proceedings of the first workshop on social media analytics*, pages 80–88.

[12] Alexander Hoyle, Pranav Goel, Andrew Hian-Cheong, Denis Peskov, Jordan Boyd-Graber, and Philip Resnik. 2021. Is automated topic model evaluation broken? the incoherence of coherence. In *Advances in Neural Information Processing Systems*, volume 34, pages 2018–2033. Curran Associates, Inc.

[13] Matthew L. Jockers and David Mimno. 2013. Significant themes in 19th-century literature. *Poetics*, 41(6):750–769. Topic Models and the Cultural Sciences.

[14] Temitope Kekere, Vukosi Marivate, and Marié Hattingh. 2023. Exploring covid-19 public perceptions in south africa through sentiment analysis and topic modelling of twitter posts. *The African Journal of Information and Communication*, 2023(31).

[15] Neil Kleynhans. 2014. An investigation into spoken audio topic identification using the fisher corpus.

[16] Jia Peng Lim and Hady Lauw. 2023. Large-scale correlation analysis of automated metrics for topic models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13874–13898, Toronto, Canada. Association for Computational Linguistics.

[17] Chunxi Liu, Jan Trmal, Matthew Wiesner, Craig Harman, and Sanjeev Khudanpur. 2017. Topic identification for speech without asr. *arXiv preprint arXiv:1703.07476*.

[18] Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *arXiv preprint arXiv:2006.07264*.

[19] Franco Mak, Avashna Govender, and Jaco Badenhorst. 2024. Exploring asr fine-tuning on limited domain-specific data for low-resource languages. *Journal of the Digital Humanities Association of Southern Africa*, 5(1).

[20] GA Malema, NP Motlogelwa, and M Lefoane. A rule-based setswana verb lemmatizer.

[21] David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 262–272, Edinburgh, Scotland, UK. Association for Computational Linguistics.

[22] Avashlin Moodley and Vukosi Marivate. 2019. Topic modelling of news articles for two consecutive elections in south africa. In *2019 6th International Conference on Soft Computing & Machine Intelligence (ISCMI)*, pages 131–136. IEEE.

[23] Murimo Bethel Mutanga and Abdultaofeek Abayomi. 2022. Tweeting on covid-19 pandemic in south africa: Lda-based topic modelling approach. *African Journal of Science, Technology, Innovation and Development*, 14(1):163–172.

[24] Thapelo J Otlogetswe. 2013. Introducing tlhalosi ya medi ya setswana: The design and compilation of a monolingual setswana dictionary. *Lexikos*, 23:532–547.

[25] Matthew Purver, Konrad P Körding, Thomas L Griffiths, and Joshua B Tenenbaum. 2006. Unsupervised topic modelling for multi-party spoken discourse. In *Proceedings of the 21st international conference on computational linguistics and 44th annual meeting of the Association for Computational Linguistics*, pages 17–24.

[26] Fannie Sebolela and 1 others. 2009. *The compilation of corpus-based Setswana dictionaries*. Ph.D. thesis, University of Pretoria.

[27] Michael Stadtschnitzer, Joachim Koehler, and Daniel Stein. 2014. Improving automatic speech recognition for effective topic segmentation. *Proc. DAGA-40. Jahrestagung für Akustik, Oldenburg, Germany*.

[28] Mike Wald and Keith Bain. 2008. Universal access to communication and learning: the role of automatic speech recognition. *Universal Access in the Information Society*, 6:435–447.

[29] He Zhao, Dinh Phung, Viet Huynh, Yuan Jin, Lan Du, and Wray Buntine. 2021. Topic modelling meets deep neural networks: A survey. *Preprint*, arXiv:2103.00498.