Appendix: Data Statement for Daily News - Dikgang Categorised News Corpus

Dataset name: Daily News - Dikgang Categorised News Corpus
Citations: Cite this paper.
Dataset developer(s): Vukosi Marivate (vukosi.marivate@cs.up.ac.za) and Valencia Wagner (valencia.wagner@spu.ac.za)
Data statement author(s): Vukosi Marivate
Organisation: Data Science for Social Impact Research Group
`https://dsfsi.github.io`),
Department of Computer Science, University of Pretoria, South Africa
and Sol Plaatje University

A. CURATION RATIONALE

The motivation for building this dataset was to provide one of the few annotated news categorisation datasets for Setswana. The task required identifying a high-quality Setswana news dataset, collecting the data, and then annotating leveraging the International Press Telecommunications Council (IPTC) News Categories (or codes)[19]. The identified source was the Daily News[20] (Dikgang Section) from the Botswana Government. All copyright for the news content belongs to Daily News. We collected 5000 Setswana news articles. The distribution of final categories for the dataset are shown in Figure 1.

B. LANGUAGE VARIETY

The language of this data set is Setswana (primarily from Botswana).

C. SPEAKER DEMOGRAPHIC

Setswana is a Bantu languages that is spoken in Botswana as well as several regions of South Africa [32].

D. ANNOTATOR DEMOGRAPHIC

Two annotators were used to label the news articles based on the *Daily News - Dikgang* news. Their deomographic information is shown in Table 7.

E. PROVENANCE APPENDIX

The original data is from the Daily News news service from the Botswana Government.

[19] `https://iptc.org/standards/newscodes/`

[20] `https://dailynews.gov.bw/`

Table 7. Annotator demographic

	1	**2**
Description	Annotator	Annotator
Age	30-35	20-25
Gender	Female	Male
Race/ethnicity	Black/African	Black/African
First Language(s)	Setswana	Setswana and isiXhosa
Linguistics training	Often works as a Setswana - English interpreter	Studied linguistic anthropology and works as a translator/interpreter