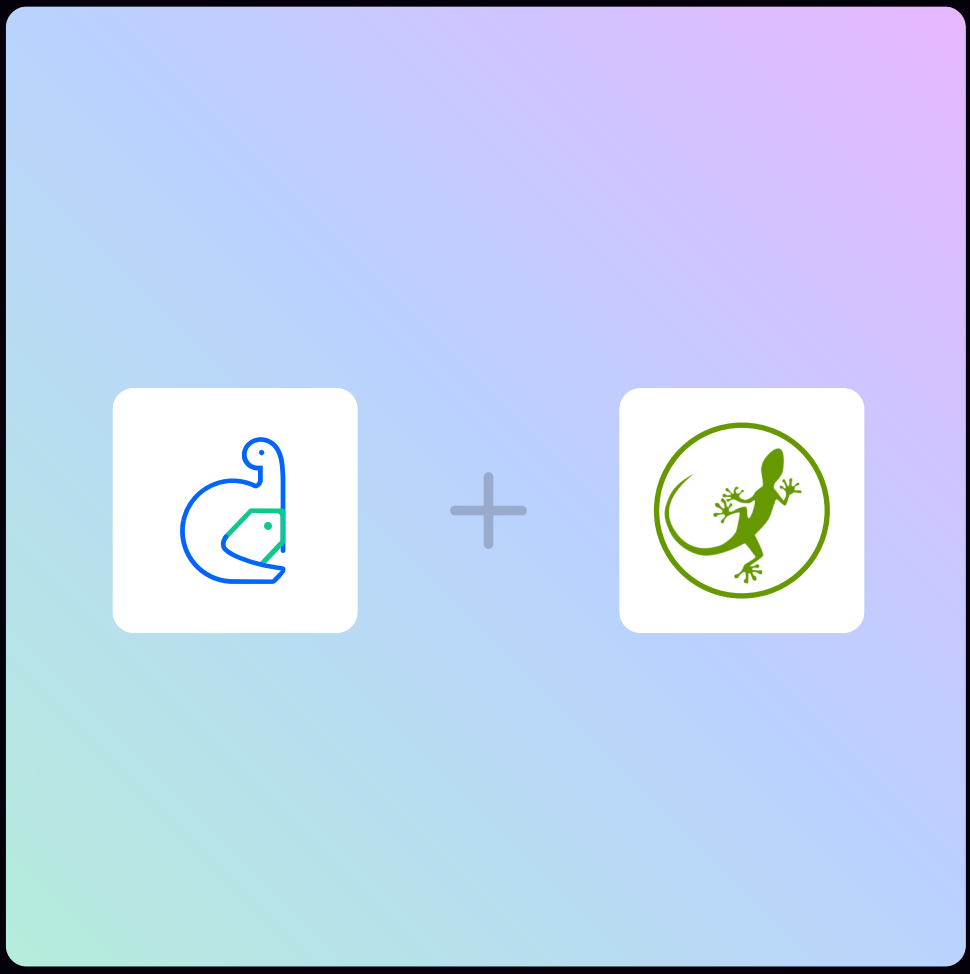


Mongabay: First Indonesian Weak Supervised Dataset - Curated by Data Programming

datasaur.ai

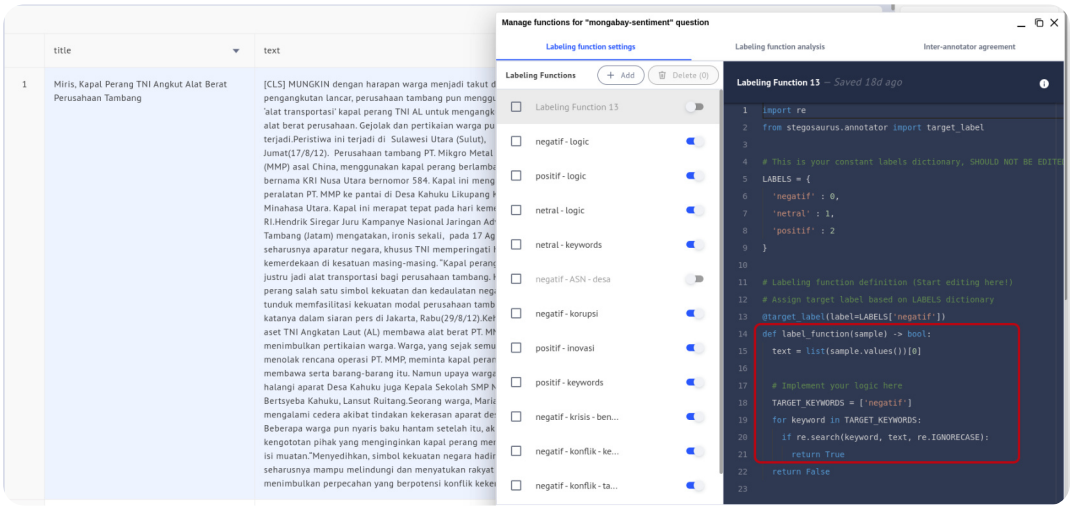
Whitepaper
November 2023

Introduction

Generating large-scale, high-quality labeled datasets for Natural Language Processing (NLP) poses a considerable challenge. Notably, when this labeling process relies on manual annotation by human evaluators, it consumes substantial time, effort, and financial resources. In response to these challenges, Datasaur introduced Data Programming, a weak supervision feature, which effectively resolves this pain point. Data Programming employs algorithms to automatically and precisely label data, significantly enhancing efficiency and proving to be up to 9.6 times more effective than manual labeling.
Recognizing the potential of Data Programming, we initiated internal research to construct a weakly curated NLP dataset sourced from an Indonesian conservation portal. The findings of this research were also featured in South East Asian Language Processing (SEALP 2023) workshop, which was part of The 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AACL 2023). Through this work, we not only validated the prowess of our data programming feature but also contributed to enhancing open-sourced Indonesian NLP datasets, fostering further developments in Indonesian NLP research.

Data Programming

We have published concise explanations and tutorials on Data Programming in our blog post. Our Data Programming feature introduces a labeling function equipped with an interactive editor that empowers users to apply their heuristics for labeling entire datasets. These heuristics can range from simple regex formulas to sophisticated OpenAI's ChatGPT. Detailed examples of pre-built labeling functions can be found in our documentation. Within our Data Programming tool, we've established a Python code template for labeling functions, which allows users to easily modify the template based on specific patterns or conditions. By default, this template employs collections of keywords as the primary logic for predicting labels.



Graphic 1: Labeling function template. The red box area contains codes that can be modified by the user.

Labeling functions, when used together with others, can give a good estimate of the true answer, even if they're noisy individually. This is possible because of the incorporation of label model calculations in the background, which aggregate all labeling functions' answers and determine the final label. Typically, these calculations depict how labeling functions agree or disagree, forming a kind of relationship graph. (Ratner et al., 2016, 2020; Alexander et al., 2022).

Our labeling functions' performance can be assessed using two types of metrics: weak supervision metrics (coverage, overlap, and conflict) and IAA (Inter-Annotator Agreement). Coverage, overlap, and conflict are standard metrics that represent relations and interactions among labeling functions. In contrast, IAA is a metric that is commonly used by data scientists to gauge how consistently multiple annotators agree on the same label for a particular category or class. In summary, coverage, overlap, and conflict illustrate the quality of labeling function collection from a label model perspective, while IAA measures the consistency of labeling functions that predict the same label.

Data Programming in Our Research

Dataset Curating

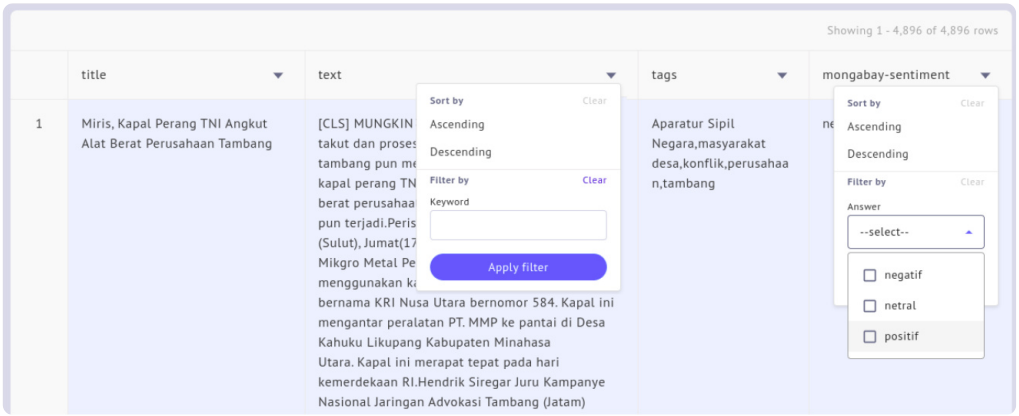
Our dataset comprises 1200 articles from the Mongabay conservation portal, spanning from 2012 to 2023. For the purpose of this research, each article was divided into 2-3 chunks, resulting in a total of 4,096 chunked articles curated for this study.

Initially, our focus was on curating a hashtag classification dataset, recognizing the significance of hashtags for content organization and searchability in the editorial realm. However, upon further exploration, we noticed that, even in a news format, each article still conveys the sentiments of its author. Consequently, we made the decision to construct a sentiment classification dataset based on the clusters of hashtags we collected.

Our curation process began with the assembly of a diverse range of popular hashtags relevant to Indonesian news and conservation topics, forming 31 labels in hashtag classification dataset. Subsequently, we categorized these hashtags into sentiment groups, including positive, neutral, and negative categories.

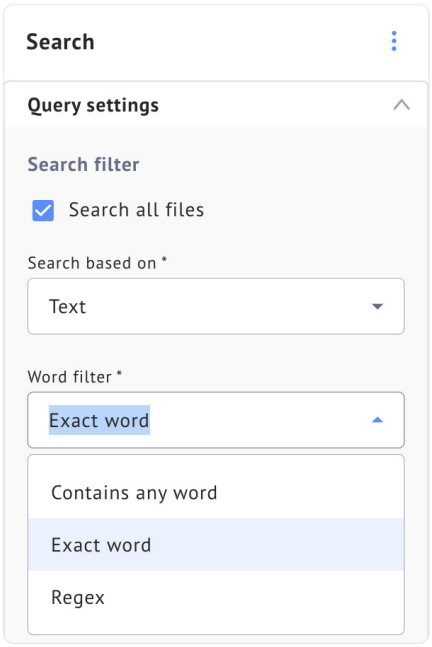
To initiate our data curation process, we commence with data exploration, with a primary emphasis on building taxonomies for hashtag classification datasets. This exploration entails the search for popular hashtags associated with Indonesian conservation topics, drawing from external sources as well as our own dataset. Within Datasaur, we provide tools and extensions that streamline the data exploration process:

- **Sort and column filter**
  In the **text** column, the user can filter rows containing particular keywords, while in the **label** column, they can select and filter targeted answer(s).

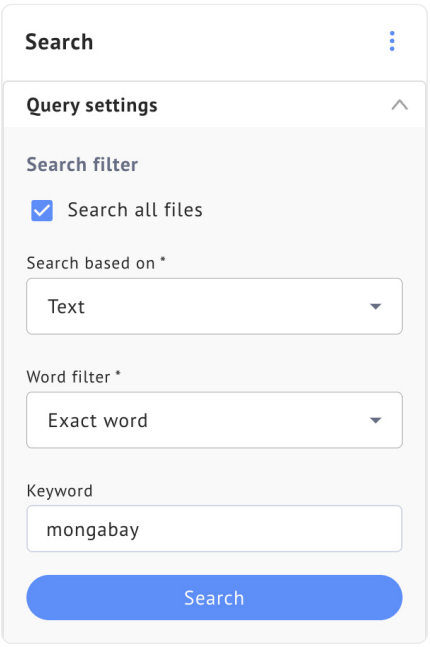


Graphic 2.1: Sort and filter columns ilustration.

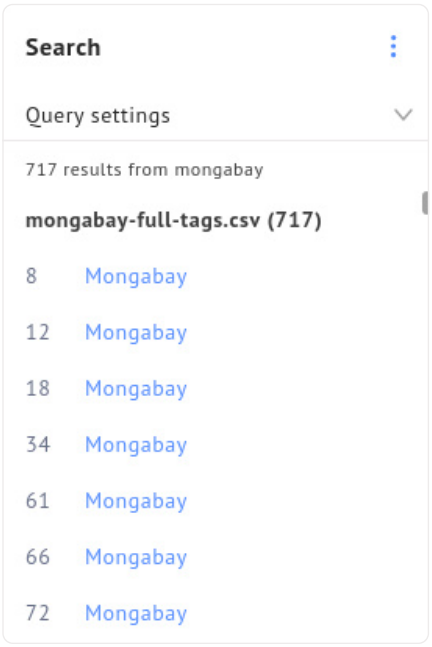
- **Search extension**
  For a broader view of specific keywords, Datasaur also has **Search extension** that allows users to see the distribution of specific keywords in **text** or **label** columns. This extension isn't limited to simple keyword searches; it also supports the use of regular expressions to match complex pattern in dataset



Graphic 2.2a: Word filter options in search extension



Graphic 2.2b: The interface of search extension

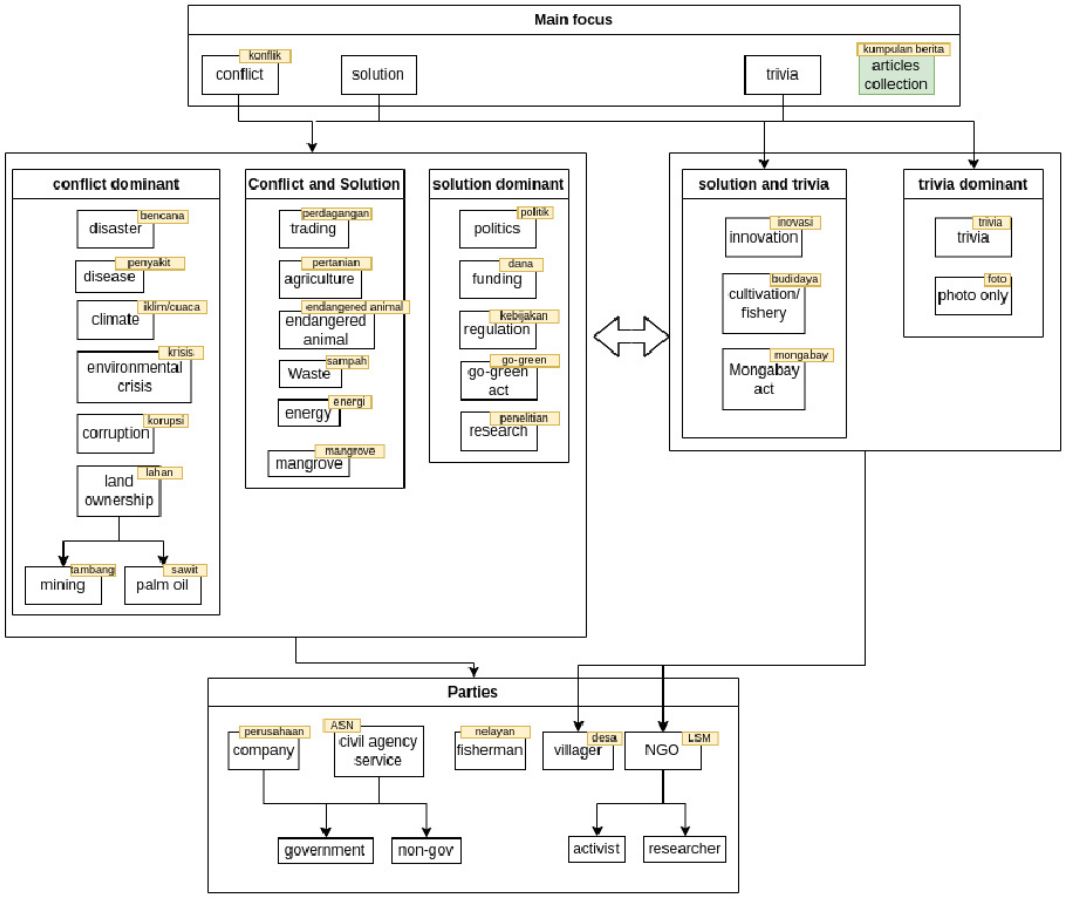


Graphic 2.2c: The search result across all rows

Following the data exploration result, we have successfully gathered a total of 31 hashtags from the dataset, which provide a structural representation of how conservation articles are composed on the Mongabay portal.

English: conflict, disaster, disease, climate, environmental crisis, corruption, land ownership, mining, palm oil, trading, agriculture, endangered animal, waste, energy, mangrove, politics, funding, regulation, go-green act, research, innovation, cultivation/ fishery, mongabay acts, trivia, photo only, articles collections, company, civil agency service, fisherman, villager, NGO

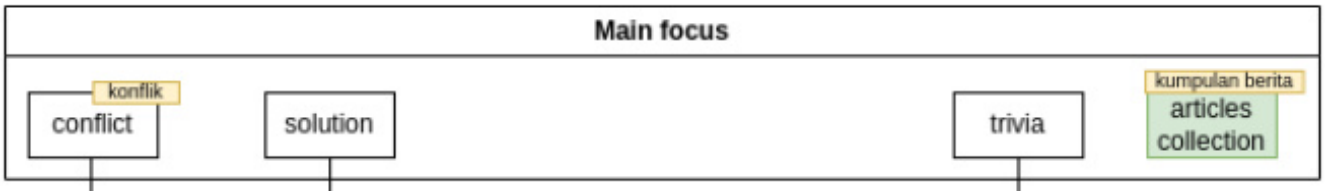
Indonesian: konflik, bencana, iklim/cuaca, krisis, korupsi, lahantambang, sawit, perdagangan, pertanian, hewan terancam punah, sampah, energi, mangrove, inovasi, budidaya, mongabay, trivia, foto, kupulan berita, perusahaan, Aparatur Sipil Negara (ASN), nelayan, desa, Lembaga Swadaya Masyarakat (LSM)



Graphic 2.3: The structure of 31 hashtags in our dataset; yellow box is defined label in Bahasa Indonesia for this experiment and green box is special class because the article consists of many different articles

Detailed explanations of parts of hashtag classification dataset taxonomy

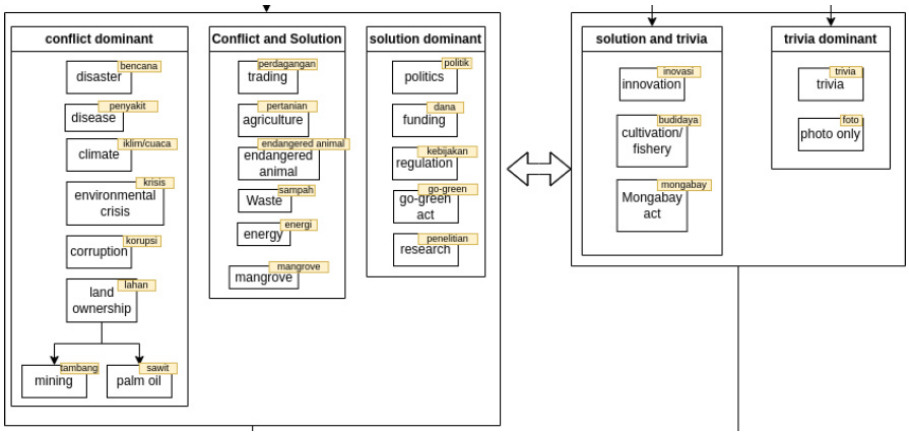
After analyzing the raw dataset, it became evident that the primary content of the articles could be categorized into two main themes: **conflict/solution** articles and **trivia** articles. Additionally, we encountered the outlier case, where some articles were compilations of various pieces on different topics. To account for this, we classified them under a specific category labeled **articles collection**



Graphic 2.4 'Main focus' part of hashtags taxonomy

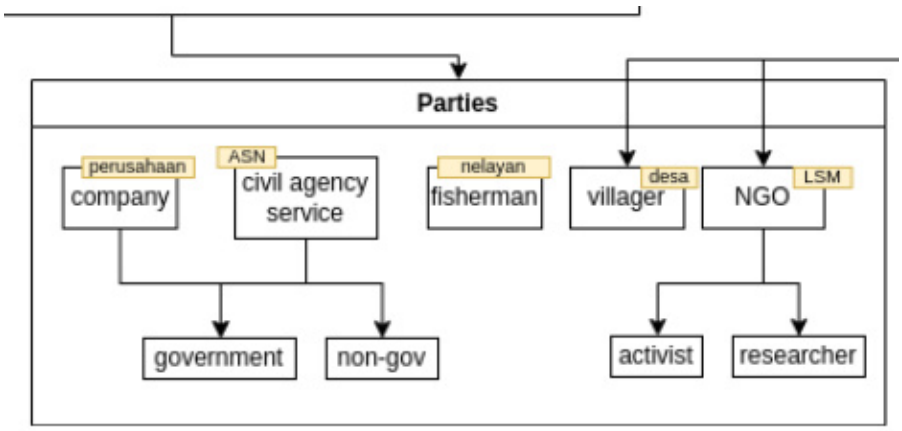
Within the non-outlier main topics (conflict, solution, and trivia), we further subdivided each topic into subtopics and generated hashtags based on frequent elements such as events, actions, and subjects found in the articles. For instance, within the **conflict/ solution** topic, we delineated subtopics as 'conflict-dominant,' 'conflict and solution,' and 'solution-dominant'.

Within the 'conflict-dominant' subtopic, articles predominantly covered issues related to **disasters, diseases, climate, environmental crises, corruption**, and **land ownership**. Notably, the land ownership issue included two prominent subjects: **mining** and **palm oil**. In total, we generated 31 hashtags, each represented as text in yellow boxes in the images below.



Graphic 2.5: Subtopic part of hashtags taxonomy

Additionally, we observed that conflict/solution articles featured several key participants, including **companies, civil service agencies, fishermen, villagers,** and **NGOs**. In contrast, trivia articles primarily featured villagers and NGOs. Taking into account the distribution of these participants' involvement in the articles, we assigned all the identified participants to corresponding hashtags.



Graphic 2.6: Subtopic part of hashtags taxonomy

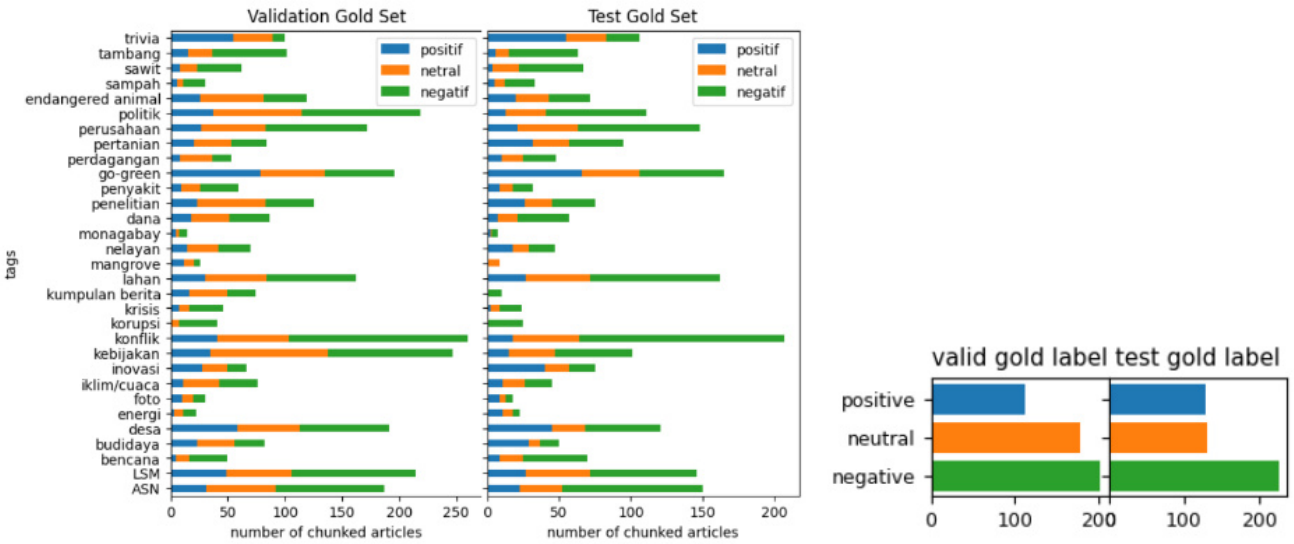
Once we had established the dataset taxonomy, we proceeded with the labeling process through Data Programming. This involved crafting labeling functions based on the taxonomy mentioned earlier. To reproduce data programming process in curating our dataset, you can access the labeling function in this Colab and the unlabeled dataset, then play around with our Data Programming extension.

After labeling dataset using data programming, we unearthed a notable feature in these articles: each chunked article can be associated with multiple hashtags. This reflects the diverse writing styles of the authors. For instance, a chunked article with a negative sentiment may be tagged with keywords like 'go-green,' 'conflicts,' 'corruption,' and 'NGOs', which indicate that chunked article discusses conflicts related to go-green action or regulation, highlights corruption issues, and mentions the involvement of NGOs.

Using the same exploration tools, we conducted an exploration of the dataset to create a sentiment classification dataset based on the labeled hashtag dataset. Furthermore, we employed Data Programming to apply labels to this dataset by utilizing a set of labeling functions for sentiment classification.

Constructed Dataset

These are the taxonomies of our weak-labeled dataset for hashtag classification and sentiment classification tasks.



Graphic 2.7: Distribution of curated sentiment and hashtags classification dataset

While each chunked article is categorized to a single sentiment class (positive, neutral, or negative), it's essential to recognize that each chunked article can encompass multiple hashtags. Consequently, each sentiment class (positive, neutral, or negative) is spread across nearly all hashtags, though in varying proportions. As previously mentioned, an article is divided into 2-3 chunks, and each chunk can possess a distinct sentiment class. This means that within a single article, multiple sentiment classes may be present.

Refer to previous graphic 2.7, an overarching trend in the sentiment classification reveals that structural tags related to **conflict** are predominantly associated with negative articles. On the other hand, a few tags from **trivia** and **solution** are more commonly observed in positive articles. Tags from **conflict and solution** are primarily included in neutral articles.

To be more clear, a fully constructed dataset can be accessed here.

Assess The Learnability of Dataset

To assess the learnability of our curated dataset, we conducted experiments involving the training and evaluation of this dataset using various standard NLP models, particularly BERT. We explored the dataset's learnability using three types of BERT pre-trained models: Indonesian bert base, multilingual bert base, and bert base. Our goal was to assess the effectiveness of our weak-labeled dataset across languages that are similar, included, or not included.

The results indicate that the Indonesian bert base model is the best-performing model for learning both dataset types (hashtag classification and sentiment classification), followed by the multilingual bert and bert base.

An important insight from these results is that the outcomes are significantly influenced by biases within the dataset. From graphic 2.7, it appears that most articles are written with a negative tone by the authors. Consequently, hashtag labels associated with negative sentiment (such as ‘civil service agency’, ‘disaster’, ‘energy’, ‘climate’, ‘conflict’, ‘crisis’, ‘land ownership’, ‘mangrove’, ‘trading’, ‘agriculture’, ‘company’, ‘politics’, ‘waste’, ‘palm oil’, ‘mining’) and negative label exhibit significantly better performance than other labels.

- **Hashtag Classification Result**
  In the context of hashtag classification, our model achieves a performance of 81.7% and 71.5% F1-score (micro average) on the validation and test sets, respectively. Notably, due to biases within the dataset, tags associated with the negative sentiment (such as 'civil service agency', 'disaster', 'energy', 'climate', 'conflict', 'crisis', 'land ownership', 'mangrove', 'trading', 'agriculture', 'company', 'politics', 'waste', 'palm oil', 'mining') attain an F1-score exceeding 70%. Conversely, tags primarily composed of non-negative sentiment classes do not perform as effectively.

PreModel	Aggr	R/A (Val)	F1-ma (Val)	F1-mi (Val)	R/A (Test)	F1-ma (Test)	F1-mi (Test)
indobert	CM	82.93	73.33	78.3	81.89	65.71	69.9
	MV	85.61	76.39	81.7	83.67	66.87	71.5
mbert	CM	81.16	69.86	75	80.01	62.18	66.8
	MV	85.14	75.49	80.7	82.85	64.95	70.2
bert	CM	70.07	47.07	55	69.56	44.45	49.2
	MV	73.34	53.25	61.23	73.16	51.96	55.4

Graphic 2.8: Validation and test results from tags classification experiment. The performance was gained from a model with the best validation score. **CM**: Using Covariance Matrix as label model; **MV**: Using Majority Voter as label model. R/A:ROC-AUC; F1-ma: F1-score macro average; F1-mi: F1-score micro average. **The red-bordered box** refers to the metrics mentioned in the preceding paragraph.

Tag	indobert		mbert		bert	
	CM	MV	CM	MV	CM	MV
ASN	84.5	88.1	83.3	85.6	56.3	54.5
LSM	69.3	72.6	71.9	74.4	41	47.5
bencana	82.4	86.1	83.8	82.7	36.5	42.7
budidaya	50.5	51.7	47.3	48.6	36.7	43.6
desa	42.9	43.6	42.1	43.5	42.1	44.6
energi	73.5	70.8	75.6	65.1	59.1	65
foto	55.8	55.8	50	59.1	64.9	65
iklim/cuaca	87.6	90.1	84.1	86	59.5	64.9
inovasi	48.1	52.3	21.7	41.5	27.1	25.3
kebijakan	46.9	48.5	44.9	49.1	42	48.4
konflik	77.3	80.2	73.4	79.5	62.9	63
korupsi	32.3	37.5	36.4	36.4	0	27.6
krisis	78	76.2	74.4	76.2	40.9	50
news-bulk	0	0	0	0	0	0
lahan	82.7	85.6	77.8	87.3	53.2	62.3
mangrove	94.7	94.7	88.9	85.7	94.7	90

Graphic 2.9a: F1-score per tag for tags classification task (part 1). The red-bordered box refers to the metrics mentioned in the preceding paragraph.

mangrove	94.7	94.7	88.9	85.7	94.7	90
nelayan	0	0	0	0	0	0
mongabay	88.4	89.7	89.7	89.7	51.5	73.4
dana	69.2	70.4	65.8	67.5	32.1	50.4
penelitian	72	67.7	68.7	69.8	58.8	57.4
penyakit	61.4	62.6	63.3	60.8	23.2	41
go-green	41.4	43.6	40.5	46.9	33.7	43.6
perdagangan	75.6	73.9	71.4	74.5	45.9	57.9
pertanian	85.2	84.6	79.3	81.3	49.5	62.9
perusahaan	78.3	83.3	78.2	83.4	58.4	68.8
politik	70.5	71.5	70	70.4	61.8	63.2
end-animal	68.2	65.6	43.5	49.3	38.4	52.3
sampah	76.9	79.5	67.5	78.5	20.7	41.4
sawit	89.8	93.2	88.4	94	75.2	85.7
tambang	89.3	92.8	89.7	93.5	55.7	72.1
trivia	64.3	60.4	56.3	53.2	56	46.2

Graphic 2.9b: F1-score per tag for tags classification task (part 2). The red-bordered box refers to the metrics mentioned in the preceding paragraph.

- **Sentiment Classification Result**
  In the context of the sentiment classification dataset, the F1-score (macro average) attained 56.37% on the validation set and 55.72% on the test set. The relatively lower performance is attributed to the model's challenge in learning the positive and neutral classes, primarily due to a bias towards negative sentiment. This bias impacts the performance of the negative (over 70%) and non-negative (around 40%) sentiment classes, ultimately affecting the overall averaged final score.

PreModel	Aggr	Acc (Val)	F1 (Val)	Acc (Test)	F1 (Test)
indobert	CM	60.77	56.37	59.79	55.72
	MV	58.74	53.97	58.97	54.12
mbert	CV	55.08	50.25	50.93	44.16
	MV	55.89	42.45	49.9	38.25
bert	CV	46.16	36.43	44.74	35.83
	MV	44.51	34.55	42.68	33.22

Graphic 2.10: Validation and test results from sentiment classification experiment. The performance was gained from a model with the best validation score. **CM**: Using Covariance Matrix as label model; **MV**: Using Majority Voter as label model. R/A:ROC-AUC; F1-ma: F1-score macro average; F1-mi: F1-score micro average. **The red-bordered box** refers to the metrics mentioned in the preceding paragraph.

Label	indobert		mbert		bert	
	CM	MV	CM	MV	CM	MV
positive	44.1	37.5	24.8	0	7	0
neutral	46.3	48.2	37.4	44.3	38.5	42.5
negative	76.7	76.7	70.3	70.4	61.9	57.1

Graphic 2.11: F1-score per label for sentiment classification task.

Summary

In summary, through this work, we have demonstrated the significant capability of our data programming, with other features (sort and column filter and search extension) assisting the data exploration process to curate **dataset with complex taxonomies**. We have unveiled the complexity of our dataset structure, highlighting the variety of hashtags that typically structure Indonesian conservation articles, shedding light on the distinct writing styles of authors, and their subjectivity towards conservation and environmental topics.

This work also contributes to the enrichment of Indonesian NLP tasks, which have traditionally been limited primarily to basic sentiment classification. It also sparks the idea of curating more diverse datasets in Indonesian and other under-represented languages, efficiently through our data programming. We have also observed the low performance in our experiments are attributed to biases within the dataset. For now, it could be an insight for us as the characteristics of the Indonesian editorial landscape, particularly in conservation and environmental topics, while also leaving us some challenges for future work in creating more equitable datasets. This, in turn, will enable NLP systems to contribute fairly in editorial area.

Based on our experience in utilizing Data programming in this work, besides automatically label the dataset, the labeling process through data programming can be done “transparently”. It means that, since labeling functions’ code adapt labeling rules, and the labeling functions are accessible for review by anyone involved in the project, any instances of mislabeled data can be tracked, corrected, and subsequently reprocessed through Data Programming. This process effectively minimizes the "black box" phenomenon often encountered in crowdsourced labeling, where mislabeled data becomes challenging to rectify due to the varying subjectivity of individual human annotators.

You can also read the findings in journal format in here: Utilizing Weak Supervision To Generate Indonesian Conservation Datasets

datasaur.ai