# Assessing Social License to Operate from the Public Discourse on Social Media

**Chang Xu**
University of Melbourne
xu.c3@unimelb.edu.au

**Cécile Paris**
CSIRO Data61
cecile.paris@data61.csiro.au

**Ross Sparks**
CSIRO Data61
ross.sparks@data61.csiro.au

**Surya Nepal**
CSIRO Data61
surya.nepal@data61.csiro.au

**Keith VanderLinden**
Calvin University
kvlinden@calvin.edu

## Abstract

Organisations are monitoring their Social License to Operate (SLO) with increasing regularity. SLO, the level of support organisations gain from the public, is typically assessed through surveys or focus groups, which require expensive manual efforts and yield quickly-outdated results. In this paper, we present SIRTA (**S**ocial **I**nsight via **R**eal-Time **T**ext **A**nalytics), a novel real-time text analytics system for assessing and monitoring organisations' SLO levels by analysing the public discourse from social posts. To assess SLO levels, our insight is to extract and transform peoples' *stances* towards an organisation into SLO levels. SIRTA achieves this by performing a chain of three text classification tasks, where it identifies task-relevant social posts, discovers key SLO risks discussed in the posts, and infers stances specific to the SLO risks. We leverage recent language understanding techniques (e.g., BERT) for building our classifiers. To monitor SLO levels over time, SIRTA employs quality control mechanisms to reliably identify SLO trends and variations of multiple organisations in a market. These are derived from the smoothed time series of their SLO levels based on exponentially-weighted moving average (EWMA) calculation. Our experimental results show that SIRTA is highly effective in distilling stances from social posts for SLO level assessment, and that the continuous monitoring of SLO levels afforded by SIRTA enables the early detection of critical SLO changes.

## 1 Introduction

Social License to Operate (SLO) represents the ongoing acceptance (or lack thereof) of an organisation's standard business practices or operating procedures by the general public (or the society at large) (Moffat and Zhang, 2014; Gunningham et al., 2004; Moffat et al., 2016). It captures the opinion of the public towards a business. Low SLO levels can increase business risks significantly, and, in the worst case scenarios, prevent the operation of an organisation. To obtain a high SLO level, organisations typically need to build trust with the community and then work to maintain that trust. Traditionally, the SLO of an organisation is evaluated using surveys and focus groups (Moffat and Zhang, 2014), during which a diversity of opinions is collected and the results then quantified. These effective techniques provide in-depth analysis. They are, however, manual practices and thus expensive to do on a frequent basis (Moffat and Zhang, 2014). In addition, the samples of a survey are often limited, and, as the time intervals between consecutive surveys are usually long, an organisation might not detect critical changes in its SLO levels in a timely fashion, leading to exposure to potential risks. The public discussions continuously taking place on social media, where people are not shy about expressing their opinions about a number of topics, including companies and specific projects, provide an opportunity to monitor SLO in real-time, on a continuous basis and at scale. This is what we aim to do in this work.

We first determined the possible facets of SLO for our domain, specifically *economic* (e.g., the public is in favor of a project because it will create jobs), *environmental* (e.g., the public believes the company has a good/bad environmental record) and *social* (e.g., the public believes the company addresses - or not - its social responsibilities). We then built SIRTA (**S**ocial **I**nsight via **R**eal-Time **T**ext **A**nalytics), a





*Proceedings of the 28th International Conference on Computational Linguistics: Industry Track*, pages 146–159
Barcelona, Spain (Online), December 12, 2020

novel automated system that combines advanced text analytics with real-time monitoring techniques to assess and monitor the SLO levels of a collection of organisations (in the same industry) over time. By taking the "pulse" of the public towards an organisation in real time, through the lens of social media, this tool complements the in-depth analysis done through surveys and focus groups, providing an early indication of trends, and potentially informing the design of in-depth surveys.

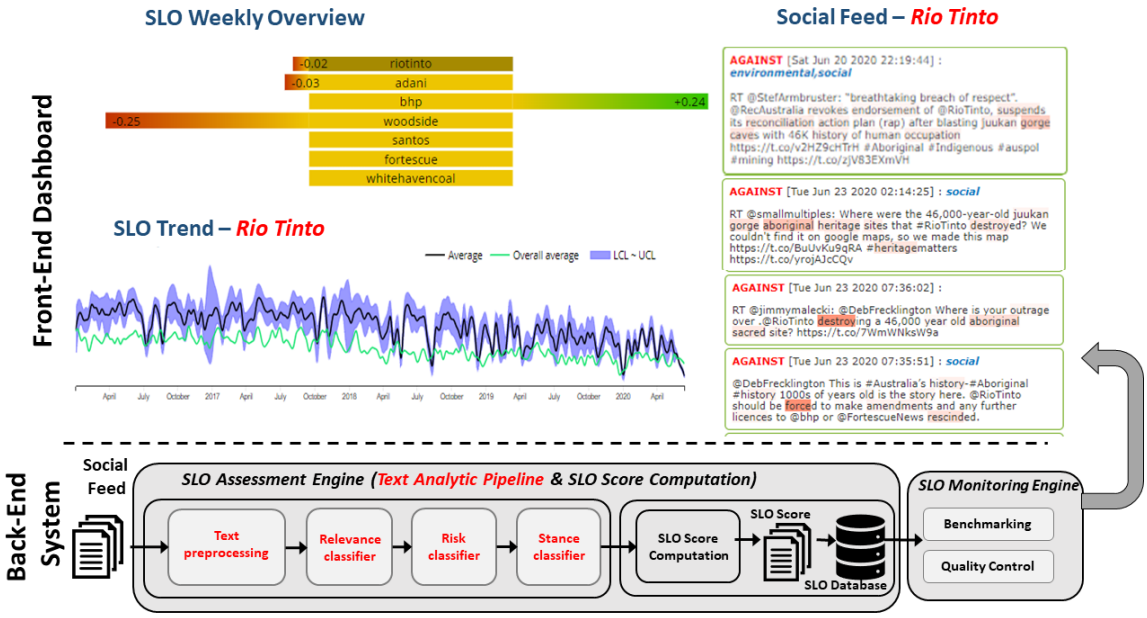



Figure 1: The dashboard of SIRTA for SLO assessment and monitoring plus the SIRTA architecture.

Figure 1 shows the dashboard of SIRTA for monitoring several major mining companies in the country. Its main functionality is demonstrated in three panels: 1) the *SLO Weekly Overview*, a list of real-time (weekly) numerical scores representing the SLO levels for the organisations under consideration, 2) the *SLO Trend*, which plots the long-term trend of the SLO level of a selected organisation (here, Rio Tinto), compared with the general trend of the market (all the organisations together), and 3) the *Social Feed*, with the most recent social media posts (e.g., tweets) about the selected organisation, with both the stances and SLO risk categories (i.e., *environment*, *social*, and *economic*) identified.

To carry out the SLO assessment, SIRTA extracts opinion information from posts published on social media (Twitter), along the different SLO facets, and then transforms that information into SLO scores. In contrast to many opinion mining systems that rely primarily on sentiment analysis, e.g., (Pang et al., 2008), we focus on *stance detection* (Mohammad et al., 2016a), which is more suitable for our task because it indicates whether someone is *for*, *neutral* or *against* a specific company, not just whether the surface sentiment of their posts is positive or negative. The novel aspect of SIRTA's SLO assessment engine includes a specialised text classification pipeline (see the Text Analytic Pipeline on the bottom of Figure 1), where three chained text classification tasks are performed for opinion extraction: 1) *relevance classification*, for finding posts contributing to the SLO assessment, 2) *risk classification*, for identifying the different facet(s), or *SLO risk(s)*, being discussed in the posts, and 3) *risk-aware stance classification*, for detecting stances in the posts that are specific to each SLO risk. The outcome from text analytic pipeline is fed into the SLO score computation component for converting the stances into numerical SLO scores. To train and evaluate the classifiers for each task above, we created both a silver standard and a gold standard dataset and employed state-of-the-art language understanding models such as BERT (Devlin et al., 2019). To monitor the derived SLO scores, a monitoring engine keeps track of the time series of the SLO scores of multiple organisations operating in a market. Specifically, it leverages Control Charts (Kan, 2002), a powerful tool for statistical process control. The monitoring engine discovers if an organisation is experiencing significant changes in its SLO score by contrasting its time series with a benchmark of the market.

We conduct several quantitative experiments to evaluate the performance of our classifiers, and thus the effectiveness of our text analysis pipeline. We then present a case study, which suggests that SIRTA can identify periods of unusual changes early. This confirms our original hypothesis that we could harness social media to monitor SLO in real-time and at scale, in a relatively inexpensive manner, reserving the more expensive, traditional methods for circumstances where a more detailed assessment is required.

## 2 SIRTA: Real-Time Text Analytics for SLO Assessment and Monitoring

### 2.1 System Overview

The bottom part of Figure 1 illustrates the architecture of SIRTA, which consists of two processing modules: the *SLO assessment engine* and the *SLO monitoring engine*. The *assessment engine* takes social feeds as inputs and generates SLO scores by first extracting opinions from social feeds via text analytics. Specifically, it performs the three text classification tasks mentioned earlier to extract, in real-time, opinions from a stream of posts, and then calculates the SLO scores.

To enable appropriate monitoring (the detection of a significant change), the opinions are aggregated regularly in different time frames (e.g., weekly). These are computed and stored in a dedicated database.

The *monitoring engine* keeps track of the time series of different organisations' scores as well as that of the overall market over time. It first computes a benchmark SLO time series representing the context (market) where those organisations operate. This allows one to see when an organisation's score departs significantly from the benchmark. To identify such a departure, the monitoring engine applies quality control techniques (Kan, 2002) to compute control limits for bounding an organisation's time series and a departure occurs when the benchmark falls out of the bound.

### 2.2 SLO Assessment Engine

The assessment engine transforms social posts into an organisation's SLO score with two modules: a text analytic pipeline and the SLO score computation. In the text analytic pipeline, three sequential tasks are performed: *relevance classification*, *risk classification*, and *risk-aware stance classification*.

#### 2.2.1 Relevance Classification: Finding Task-Relevant Posts

SIRTA uses the Twitter API to collect all tweets containing the names of the organisations under consideration. It is, of course, inevitable that posts irrelevant to our task are also collected by our system. We thus need to discard these irrelevant posts and keep only the posts that can contribute to the SLO assessment. This is done through the relevance classification task, and a binary *relevance classifier* $C_r$ was trained for this task. The classifier $C_r$ reads a post $\mathbf{x}_i$ and assigns it a relevance label $\hat{\mathbf{y}}_r$. To train $C_r$, we minimised the negative log-likelihood of the ground truth label: $\mathcal{L}_r = -\sum_{i=1}^{N_r} \mathbf{y}_r \log \hat{\mathbf{y}}_r$, where $\mathbf{y}_r$ is $\mathbf{x}_i$'s true relevance label, and $N_r$ the training data size. To facilitate discussion, we use $R_o$ to denote the set of all relevant posts discussing organisation $o$ detected by $C_r$.

#### 2.2.2 Risk Classification: Discovering Key SLO Risks Mentioned in a Post

As mentioned earlier, SLO can have many facets, or, put differently, SLO poses risks along various dimensions. In turn, an individual post can relate to different SLO risks. Consider, for example, the following two posts about mining companies. The first one is solely expressing an opinion about the company's handling of environmental concerns, thus contributing (negatively) to the SLO risk of environment for this company. In contrast, the second negatively mentions a company's actions with respect to both environmental and social concerns, thus contributing (again negatively) to the SLO risks of both environment and social for this company.

- *They don't even know which aquifer is the source of the Doongmabulla Springs, but Adani* [**the company name**] *is belittling crucial environmental studies as "paperwork".* (SLO risk factor: **environmental**)
- *We are at BHP* [**the company name**] *HQ protesting against their toxic Olympic Dam uranium mine that fuels war and breaches land rights #uprootthesystem #nonukes #KeepItInTheGround* (SLO risk factors: **environmental** and **social**)

Being able to identify the specific SLO risks discussed in the public discourse allows an organisation to better identify and manage them. We took a two-step approach to detecting the risk factors mentioned

in a post. First, we identified the SLO risks for our domain (mining), based on knowledge from domain experts and a literature survey. They are *economic*, *social* and *environment*. We note that, while these risks are fairly general, SLO risks might be different in different domains. Let $K$ be the set of these risks: $K = \{economic, environmental, social\}$.

We then trained a *multi-label risk classifier* $C_k$ to find all potential risk factors mentioned in a post $\mathbf{x}_i \in R_o$. The training involved minimising the *one-vs-all* loss, a commonly-used objective for multi-label classification (Tsoumakas and Katakis, 2007). Formally, we calculated a binary cross entropy between the logits and ground-truth labels of the same training example,

$$\mathcal{L}_k = \mathbf{y}_i^k \cdot \log C_k(\mathbf{x}_i) + (1 - \mathbf{y}_i^k) \cdot \log(1 - C_k(\mathbf{x}_i)) \tag{1}$$

where $C_k(\mathbf{x}_i)$ produces the logits. $\mathbf{y}_i^k$ is the corresponding risk label of $\mathbf{x}_i$, which is multi-hot encoded; $\mathbf{y}_{ij}^k = 1$ as long as $\mathbf{x}_i$ belongs to the $j$th risk, otherwise $\mathbf{y}_{ij}^k = 0$.

#### 2.2.3 Stance Classification: Revealing Risk-Specific Stances

The first two classifiers identified the overall relevance of a post to our task, and to which specific SLO risk the post is relevant. In the final text analytics task, we extract the opinion of a post. Sentiment analysis (Liu, 2012) is a common opinion mining technique, but recent studies have shown that one's sentiment may not always reflect one's attitude towards a target (Sobhani et al., 2016; Mohammad et al., 2017). For example, consider the following post about a mining company, "This is huge! Our momentum is unstoppable. Every day we're closer to stopping Adani and saving our Reef". Although the sentiment here is positive ("This is huge!"), the author's attitude towards the mining company (Adani) is negative. Therefore, instead of extracting the sentiment of the post, we propose to extract the *stance* of the author implied in their posts (Mohammad et al., 2016a; Augenstein et al., 2016; Sun et al., 2018), which could be *for*, *against*, or *neutral* towards a target (an organisation). The stance could be extracted in two ways. We could employ a general stance classifier (to obtain the stance of any relevant post) or a *risk-aware stance classifier*, that is a classifier specifically designed to detect the stance in posts discussing a specific risk factor. We posit that the language used to express stances vary with different risk factors (e.g., "create jobs" for *economic* vs. "destroy the reef" for *environmental*), and thus a risk-aware stance classifier would be more effective. Our experiments show that training a stance classifier for each risk factor indeed allows us to capture the stance more accurately ($\sim$4% boost) than training a generic stance classifier to work across the classes (see Section 3.3 below).

To train these classifiers, we minimised the negative log-likelihood of the ground truth label: $\mathcal{L}_s^j = -\sum_{i=1}^{N} \mathbf{y}_s^j \log C_s^j(\mathbf{x}_i)$, where $\mathbf{y}_s^j$ is the true stance label of the post $\mathbf{x}_i$ for the $j$th risk factor.

#### 2.2.4 SLO Score Computation: Transforming Stances into SLO Scores

The final step in the assessment engine is to quantify the stances derived from the text analytic pipeline to obtain an SLO score, based on the degree of the opinion expressed in each post for an organisation $o$ using the set of relevant posts $R_o$. With all the stance classifiers $\{C_s^j\}_{j=1}^{|K|}$ developed (one for each risk factor), given a post, $\mathbf{x}$, its overall SLO score is derived by averaging over the stances across all risk factors: $s = \frac{1}{|K|} \sum_{j=1}^{|K|} C_s^j(\mathbf{x})$[1]. To produce the final SLO score for an organisation $o$, we aggregate the SLO scores of all relevant posts $R_o$ via averaging: $s_o = \frac{1}{|R_o|} \sum_{\mathbf{x}_i \in R_o} s_i$.

### 2.3 SLO Monitoring Engine

The changing nature of an organisation's operational context can impact its SLO score. Changes could be due, for example, to a change of a company's CEO, changes in the general trend of the overall market, or a major event. Assessing such changes thus requires that we keep track of not only the time series of a company's SLO scores but also the time series of scores of other companies operating in that market sector. This allows us to see when a company's score departs significantly from the average score across similar organisations, which can be seen as a benchmark of the context/market. Such information can drive strategic action at critical points in time. SIRTA's SLO monitoring engine is

[1]The stance of an absent risk factor will not be included in the summation.

designed to track a comparable set of organisations across time. To achieve this, it first obtains the market benchmark by averaging over all organisations' SLO time series. This ensures that the larger or more topical organisations (i.e., the ones that are discussed more often) do not dominate in the comparison. All organisations are thus comparable as they have faced the same market conditions over the same period. Then, the engine seeks to monitor the departure of each organisation's time series from the benchmark over a period of time (e.g., one week), which is computed as follows.

Let $s^t_{o,i}$ be the $i$th SLO score of organisation $o$ in period $t$, and $n^t_o$ the number of its SLO scores in $t$, the average SLO score of $o$ in $t$ is then given by $\bar{s}^t_o = \sum_{i=1}^{n^t_o} s^t_{o,i}/n^t_o$ and standard deviation $\sigma^t_o = \sum_{i=1}^{n^t_o}(s^t_{o,i} - \bar{s}^t_o)^2/n^t_o$. For organisations with sufficient observations in $t$, the Shewhart chart (Kan, 2002) with upper control limit (UCL) and lower control limit (UCL) is given by

$$UCL_o = \bar{s}^t_o + 3\sigma^t_o/\sqrt{n^t_o} \quad \text{and} \quad LCL_o = \bar{s}^t_o - 3\sigma^t_o/\sqrt{n^t_o} \tag{2}$$

Then a departure of the organisation $o$ from the benchmark in $t$ occurs if the benchmark is below $LCL_o$ or above $UCL_o$. For organisations with zero observations in $t$, we use the exponentially-weighted moving average (EWMA) for the monitoring. Specifically, in period $t$, we compute the moving average as $a^t_o = 0.05\bar{s}^t_o + 0.95a^{t-1}_o$ if $\bar{s}^t_o$ exists, otherwise $a^t_o = a^{t-1}_o$. Similarly the moving standard deviation is defined as $v^t_o = 0.05\sigma^t_o + 0.95v^{t-1}_o$ if $\sigma^t_o$ exists, otherwise $v^t_o = v^{t-1}_o$. Then the control charts for this case is given by $UCL_o = a^t_o + 3v^t_o/\sqrt{39}$ and $LCL_o = a^t_o - 3v^t_o/\sqrt{39}$.

## 3 Training and Evaluating the SLO Assessment Engine

We now present how we developed, trained and evaluated the classifiers for the SLO assessment engine. We first created training and test data sets for each task in the text classification pipeline. We used these data sets to train a number of modules, experimenting with several state-of-the-art techniques. Finally, we evaluated the classifiers, in order to choose the best ones to incorporate into SIRTA.

### 3.1 Data sets

**Data sets for stance classification**. We collected tweets about different mining organisations posted in the country from January 1, 2016 up to 23 October, 2019. We obtained *silver* standard labels for these tweets with rules that automatically determine the stance labels based on specific meta signals such as hashtags and Twitter account names. While the full set of rules is presented in Appendix, some examples are: 1) *favour* - a tweet by a mining company-owned account, e.g., *adaniaustralia*; 2) *against* - a tweet contains disapproving hashtags, e.g., *#stopadani*; and 3) *neutral* - a tweet from known mining-related news sources, e.g., *MiningNewsNet*. We do not rely on the content of a tweet to determine its label. To test the accuracy of the auto-coding, we randomly sampled 24 tweets from the resulting training set and asked three coders to manually code them as stance for, against or neutral. The coding had a Fleiss Kappa score of 0.71, in the "substantial agreement" range, and the majority code from this manual coding matched the auto-coding in all cases. We took this as evidence that the auto-coder based on the simple rules listed above provided largely accurate codings. We note that such silver training set would inevitably contain noise (e.g., a news source account may occasionally post a positive news report about a mining company), but we suspect that this would not harm the performance much (as shown later in our experiments) due to the large scale of the training set. To prepare the test set, we manually created a *gold* standard dataset by asking three human coders to annotate 274 tweets[2] using specific annotating guidelines (see Appendix). The statistics summary of the training/test sets for this task are shown in Table 1.

| | **#train** | **#test** |
|---|---|---|
| Favour | 24,255 | 111 |
| Against | 19,311 | 103 |
| Neutral | 19,317 | 60 |
| Total | 62,883 | 274 |

Table 1: Statistics summary of the datasets for SLO stance classification.

**Data sets for risk classification**. The training set for this task shares the same tweets as in the above task, except that each tweet is now associated with one or more SLO risk labels. As already mentioned, our risk labels were: *social*, *economic*, and *environmental*. We again obtained *silver* risk labels by using

[2]A Fleiss Kappa score of 0.88.

rules on matching the tweet contents with specific keywords (e.g., "community" for *social*, "environment" and "greatbarrierreef" for *environment*, and 'jobs' for *economic*). For the test set, we asked three human coders to annotate 300 tweets. Table 2 shows the statistics summary[3] of the training/test sets.

| | **Social** | **Economic** | **Environmental** | **Other** | **One-Label** | **Two-Label** | **Three-Label** | **Total** |
|---|---|---|---|---|---|---|---|---|
| Train | 12,960 | 13,540 | 7,405 | 35,878 | 56,605 | 5,654 | 624 | 62,883 |
| Test | 140 | 72 | 61 | 101 | 207 | 84 | 9 | 300 |

Table 2: Statistics summary of the datasets for SLO category classification.

**Data set for relevance classification**. Finally, we built the training/test sets for the relevance classification task. For the training set, we considered all the tweets used in the stance classification task as relevant, as they were collected with rules for ensuring they were mining-related and informative to stance determination. Then, to get the irrelevant tweets and a balanced data set, we randomly sampled the Twitter stream[4] to obtain the same number of tweets (62,883). The resulting training set contains 125,764 tweets in total (50% *relevant* and 50% *irrelevant*). For the test set, as we lacked a gold standard set, 5-fold cross-validation was used instead.

### 3.2 Classifiers and Training Details

We pre-processed the data in the above data sets as follows. For each tweet, tokenisation was done via the CMU Tweet Tagger (Owoputi et al., 2013), and character elongations were shrunk (e.g., "yeees" $\rightarrow$ "yes"). We removed all hashtags and mentions[5]. We also replaced all URLs, and year, time, cash with place holders (e.g., "slo_url"). All text were down-cased. Stop words were retained, because of the stance-indicative information they can contain (e.g., "not").

We followed the best practice of training text classification models by implementing four classifiers with state-of-the-art neural network models as the baselines: 1) **fastText** (Joulin et al., 2017): an efficient classification model trained on word vectors created with subword information; 2) **BiLSTM** (Augenstein et al., 2016): a bidirectional LSTM trained on word vectors pretrained with GloVe word embeddings (Pennington et al., 2014) (glove.twitter.27B, 200d); 3) **CNN** (Kim, 2014): a convolutional neural network for sentence classification; 4) **BERT** (Devlin et al., 2019): a general-purpose pre-training contextual model for sentence encoding and classification.

The following configurations were used for training the classifiers: 1) **fastText**: learning rate of 0.1 was used, and the training did not stop until 10 epochs had passed; 2) **BiLSTM**: the hidden sizes of both LSTM and the followed dense layer were set to 256. A step learning rate scheduler was used, where the learning rate was set to 0.5 initially and then decayed by 10% after each epoch. A dropout layer was placed after the dense layer with a dropout rate of 0.3; 3) **CNN**: four 1D convolutional layers of 256 filters were chained as the sentence encoder, with the sequential filter sizes as 2, 3, 4, and 5. The same learning rate scheduler and dropout layer as those in BiLSTM were used; 4) **BERT**: the $\text{BERT}_{\text{BASE}}$ (uncased) was used. The learning rate was set to $10^{-5}$. The maximum number of wordpieces was set to 128. The batch size for each training step was 16 for BERT (due to GPU memory limits) and 128 for others. Early stopping was applied with a patience of 3.

### 3.3 Experimental Results

We now report the results of all the text classifiers in SIRTA's SLO assessment engine.

**Relevance Classification** All the classifiers achieved reasonable results on this task, as shown in Table 3, suggesting the distributions of the relevant and irrelevant tweets are easily separable. Among the classifiers, fastText and BiLSTM obtained the highest scores in different training set settings, while BERT, as a cutting-edge modelling tool for text, surprisingly failed to show its potential on this task.

[3]More details of the rules/guidelines for the silver/gold label acquisition are in Appendix.

[4]https://developer.twitter.com/en/docs/labs/sampled-stream/api-reference/get-tweets-stream-sample-v1

[5]The removal of all hashtags/mentions allows us to train models that generalise and are not specific to the seen hashtags/mentions in the training data. The removal of the hashtags is also because some of them were already used for obtaining the silver standard labels of the training data.

| %Train | fastText | BiLSTM | CNN | BERT |
|---|---|---|---|---|
| 25% | 95.9±0.9 | 94.5±1.0 | 93.5±1.5 | 89.7±1.0 |
| 50% | 96.0±0.9 | 94.9±2.4 | 93.9±1.5 | 91.3±1.4 |
| 75% | 96.1±0.9 | 96.0±1.1 | 93.9±1.5 | 92.9±2.1 |
| 100% | 96.2±0.9 | 96.2±1.1 | 94.3±1.4 | 93.2±1.7 |

Table 3: Accuracy on relevance classification.

This could be caused by the small batch size (16) used for BERT training in order to avoid the out-of-GPU-memory issue; a small batch size usually makes SGD updates less effective on each batch. Another reason behind this could be that the BERT model is over-complex for such an easy task (the classes are easily separable), potentially leading to overfitting.

**Risk Classification** Table 4 shows the classification results. This task is more challenging than the previous one, as evidenced by the generally lower accuracy attained by all classifiers. BERT performed the best across all risk factors, exhibiting its superiority on this more demanding modelling task. However, considering its complexity, the improvements gained by BERT were not proportionally outstanding (2%∼2.9%). fastText also performed well, better than both BiLSTM and CNN, demonstrating that it is also a cost-effective choice for this task.

The performance on the *Environmental* risk factor was better than all other factors, which may suggest that it is easier to recognise a post discussing the environmental than one discussing social or economic issues. The performance on the *Social* risk was the worst. We found that *Social* samples dominate the multi-label samples in the test set (88.2%); such multi-label posts are harder to classify. As a result, the classifiers make more mistakes on the *Social* samples.

| Risk Factor | fastText | BiLSTM | CNN | BERT |
|---|---|---|---|---|
| Social | 57.7±0.3 | 57.3±0.2 | 55.2±0.4 | 60.7±2.0 |
| Economic | 64.3±0.2 | 63.9±0.6 | 64.3±0.3 | 64.5±0.3 |
| Environmental | 68.3±0.4 | 67.3±0.5 | 68.6±0.3 | 70.9±0.3 |
| Other | 71.1±0.5 | 68.9±1.4 | 69.7±0.8 | 73.2±0.7 |
| Average | 65.4±0.4 | 64.4±0.7 | 64.4±0.5 | 67.3±0.8 |

Table 4: Accuracy on risk classification.

**Stance Classification** To validate the hypothesis that a risk-aware classifier is more accurate than a generic stance classifier, we compared two experiments: 1) we trained four individual risk-specific stance classifiers using data from the corresponding risk ($\mathcal{R}$), and 2) we trained a single generic stance classifier using data on all risks (not differentiating the risk labels). For both experiments, we split the test set into subsets on different risks. For each risk-specific classifier, we tested it on the respective risk subset. The generic stance classifier was tested on all risk subsets. The results are shown in Table 5, where we observe the risk-aware stance classifiers provided performance gains across all risk-classifier

| SLO Risk | fastText | BiLSTM | CNN | BERT |
|---|---|---|---|---|
| Social | 70.8±1.3 | 76.1±1.3 | 75.9±0.9 | 74.7±2.0 |
| Social ($\mathcal{R}$) | 71.3±2.6 (**0.5**) | 76.9±3.7 (**0.8**) | 77.7±2.6 (**1.8**) | 76.2±2.2 (**1.5**$^{***}$) |
| Economic | 59.6±2.6 | 63.1±2.5 | 62.2±2.4 | 66.3±3.8 |
| Economic ($\mathcal{R}$) | 62.2±2.5 (**2.6**) | 68.4±2.0 (**5.3**$^{***}$) | 66.5±2.7 (**4.3**$^{**}$) | 71.2±1.5 (**4.9**$^{**}$) |
| Environmental | 61.5±2.2 | 67.0±4.0 | 67.5±4.5 | 69.4±2.7 |
| Environmental ($\mathcal{R}$) | 68.0±2.9 (**6.5**$^{***}$) | 68.4±1.2 (**1.4**) | 68.0±4.0 (**0.5**) | 72.5±1.6 (**3.2**) |
| Other | 49.1±2.7 | 53.0±3.4 | 55.9±1.7 | 58.1±1.0 |
| Other ($\mathcal{R}$) | 56.7±1.4 (**7.6**$^{***}$) | 57.5±2.7 (**4.5**$^{*}$) | 56.3±1.3 (**0.4**) | 61.1±4.9 (**3.0**) |
| Overall | 58.5±1.5 | 64.8±2.8 | 65.4±2.4 | 67.2±1.9 |
| Overall ($\mathcal{R}$) | 64.3±2.0 (**5.8**$^{**}$) | 67.8±2.4 (**3.0**) | 67.1±2.7 (**1.7**) | 71.7±4.0 (**4.5**) |

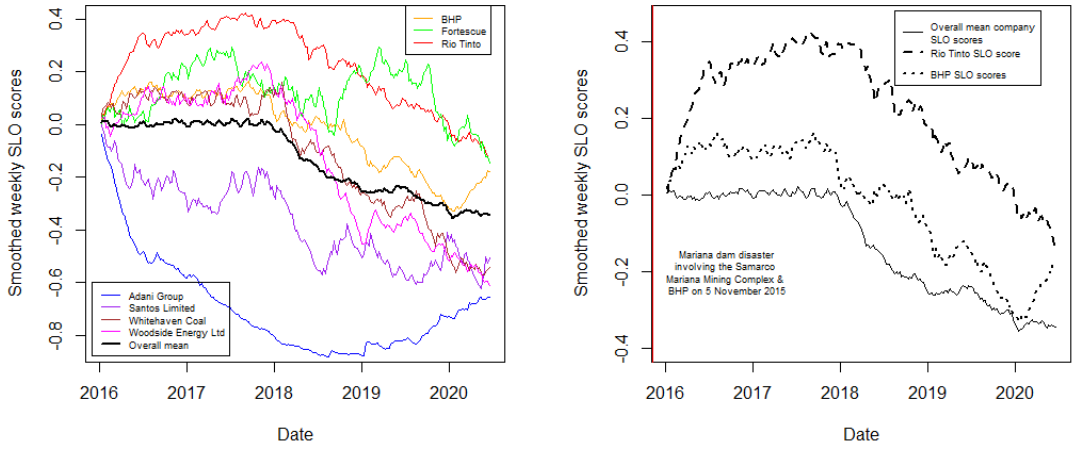
(Two-tailed t-test: $^{***}$ $p < 0.01$; $^{**}$ $p < 0.05$; $^{*}$ $p < 0.1$)

Table 5: Accuracy on stance classification. Performance gains of the risk-aware methods over the corresponding non-risk ones are shown in the parentheses.

combinations, although the gains are not necessarily statistically significant in all cases. This validates our hypothesis that the language used to express stances generally varies when people discuss different SLO risk factors, and training specialised stance classifiers for different risks could better capture the underlying risk-specific language variations.

### 3.4 Implementation and Deployment of SIRTA

SIRTA was built by using Apache Kafka and ELK stack (Elasticsearch, Logstash, and Kibana) for constructing the real-time text classification pipeline in SIRTA's SLO assessment engine. It continuously obtains streaming tweets, which are then fed into the analytics pipeline. For the classifier configuration, based on our evaluation in the previous subsection (§3.3), we deployed fastText for the relevance classification task and BERT for both the risk and stance classification tasks. The monitoring dashboard was implemented with NodeJS and D3. SIRTA was deployed on a web server with a dockerised form.




(a) SLO trends and variations of the monitored mining companies

(b) A case study of detection of SLO score changes in relation to BHP and Rio Tinto

Figure 2: The monitoring of SLO scores of seven major mining companies in the country.

### 3.5 Monitoring in Practice

We have been using SIRTA to monitor a number of major mining companies in the country. Figure 2a shows the chart of the trends and variations of their (EWMA[6]) SLO scores over a four-year span, from 2016 to 2020, on a weekly basis (averaging over a one-week window). Among these companies, Adani has had consistently the lowest score over time. This is aligned with our observations on Twitter of the numerous campaigns against the company.[7] SIRTA captured these negative opinions and trend over time. Rio Tinto, at the other extreme, had maintained a generally higher SLO profile until very recently, when it destroyed the Juukan Gorge, an ancient Aboriginal sacred site in Australia.[8] BHP also started with a high SLO, on par with Rio Tinto, but we then see a big departure from Rio Tinto in early 2016. This is likely due to the Mariana dam disaster[9] in South America late 2015.[10] Overall, the mean time series of SLO scores (black) was essentially steady until 2018 (although there is a decrease in early 2016, also probably due to the dam disaster), and then decreased significantly, indicating the general public in the country has become more negative about the mining sector overall. This is likely due to increased

[6]Exponentially-Weighted Moving Average, as a measure to account for data scarcity cases.

[7]These campaigns often have the textual prefix "StopAdani" in their Twitter account names.

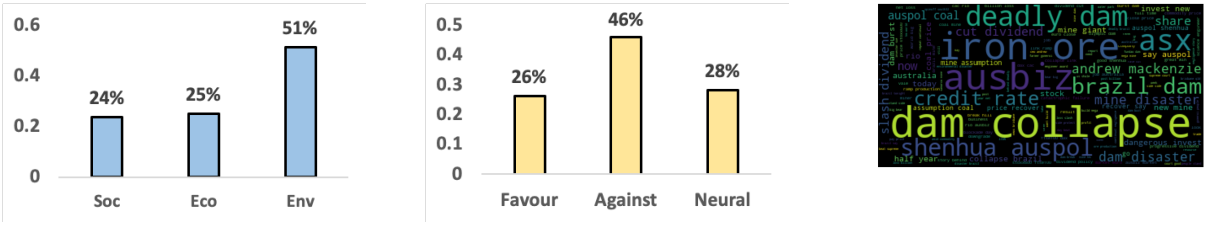
[8]See news articles: www.business-humanrights.org/en/australia-rio-tinto-mining-blast-destroys-ancient-aboriginal-sacred-site and www.ft.com/content/6db79b46-8e46-4e89-8688-97064effbc61 – accessed June 24th, 2020.

[9]https://en.wikipedia.org/wiki/Mariana_dam_disaster

[10]Unfortunately, we lack the data before 1st Jan, 2016.

concerns about the environment and the public awareness of a major project by Adani, with the 'stop adani' movement becoming very active in early 2018.

An advantage of SIRTA is its ability to detect SLO changes, thus allowing for prompt mitigating actions being taken. To demonstrate this, we look at two companies, BHP and Rio Tinto. Both experienced significant changes in SLO scores during the monitoring period - See Figure 2b. We again observe the drop in the sector's SLO in 2018. With respect to BHP, there is a sharp departure from Rio Tinto in early 2016, most likely from the Mariana dam disaster, as mentioned above. Another point of particular interest is the sharp decline in Rio Tinto's SLO score in 2020, probably due to the destruction of the sacred site. After that, Rio Tinto and BHP appear converging in their scores, while Rio Tinto was performing much higher earlier. We notice a recent rise in BHP's SLO, potentially because of an announcement to postpone the destruction of other ancient caves until they had a chance to discuss with the community. We note that BHP, Rio Tinto, and Fortescue are essentially Iron Ore companies and have moved out of fossil fuels, while the other companies are related to fossil fuels. The figure shows that Iron Ore companies generally do better than mining companies (when no major event like the destruction of ancient sites or a dam disaster occurs), reinforcing the hypothesis that the downward trend in the mean SLO score is due to climate change concerns about using fossil fuels.




(a) SLO risk distribution

(b) Stance distribution

(c) Word cloud of tweet contents

Figure 3: Further analysis on the tweets posted between Jan and March 2016 about BHP.

To verify our hypothesis that the departure from Rio Tinto in early 2016 is due to the dam collapse, we did a further text analysis on all the tweets about BHP in our database between Jan and March of 2016. The results are shown in Figure 3. We found that most of those tweets were discussing environmental issues (51%, as opposed to about 25% for both economic and social issues in Figure 3a) and holding an against stance towards the company (46% in Figure 3b). We also drew a word cloud of the contents of those tweets (Figure 3c), which shows the frequent use of words such as "deadly", "dam", and "collapse'. All the findings above clearly suggest the occurrence of the negative change of BHP's SLO scores during that period was related (at least partly) to the dam collapsing event.

### 3.6 Cross-Validation with Survey-based Approaches

We contacted Voconiq[11], a company focused on measuring social license to operate with mining companies in Australia and overseas. Voconiq has pioneered the development of social science tools to provide insights for their clients on their social license to operate. Voconiq employs survey data, unstructured qualitative data, workshops and interviews with community members and company employees living in mining communities. Voconiq collects data about a similar set of companies as SIRTA. We shared our results with the CEO and Co-Founder, Dr Moffat. He told us he believed our work "added an important piece of research and technical development to the field of SLO research and practice". Dr Moffat also made some observations about the insights gained from SIRTA, compared to the patterns Voconiq observed. The first observation was the results in response to the Juukan Gorge incident, evident in the Twitter data in Quarter 2, 2020 (Figure 2a). The patterns observed through SIRTA in terms of company specific patterns of community responding were similar to those observed in his own work utilising data collected from monthly community surveys. Publicly available data regarding community sentiment toward Rio Tinto in Pilbara communities showed a drop in community sentiment corresponding with a similar drop in the SLO scores in SIRTA.

[11] https://voconiq.com/

Dr Moffat's second observation pointed to the fact that SIRTA and Voconiq "listen to different voices", reinforcing our hypothesis that looking at SLO from a social media perspective can provide complementary information to other methods. While large events like the San Marco dam collapse or the Juukan Gorge incident are of a magnitude that affect community sentiment within local mining communities and at a larger societal scale in similar ways, typically the sentiment of community members at these two scales are different. Local communities are more supportive of mining companies typically (often because of the jobs they provide), and they have more realistic understanding of both the benefits and impacts of mining operations. In contrast, data collected at a societal level (e.g., from social media) often reflects a different set of issues and agendas. These differences are evident in divergences Dr Moffat observed when looking at the insights from the Twitter data through SIRTA. He emphasised, however, that this was not a problem but rather a strength of our work. Data from social media provides a unique, and often leading, indicator of community sentiment, allowing companies and other stakeholders to combine these perspectives with those of local community residents for a more three-dimensional (and accurate) understanding of SLO at multiple scales.

## 4 Related Work

Traditionally, organisations determine SLO using surveys and focus groups (Moffat and Zhang, 2014), which are effective but also expensive to run. Our work seeks to complement these in-depth qualitative methods, harnessing social media to provide a real-time view of social license for organisations and/or specific projects, with early detection of changes - especially downward changes which might need to be addressed through immediate action. It can also inform the design of the focus groups and surveys, by providing information as to current concerns of the public.

Social media monitoring systems have been built to cover social phenomena (Wan and Paris, 2015; Larsen et al., 2015; Joshi et al., 2019), but none has focused on social license. Yet, this is potentially a very important application of social media analytics, as, increasingly, companies need to ensure they have such license, either as an organisation as a whole, or for specific projects they intend to carry out. In addition, an important aspect of our work is to couple the text analytics with a statistical monitoring engine to ensure insights from the text are appropriately put into an overall historical and sector contexts.

Stance detection in social media has gained much attention in recent years. The SemEval-2016 Task 6 challenge (Mohammad et al., 2016b) focused on stance classification of tweets discussing controversial political positions (e.g., abortion and climate change) and opposing political candidates (e.g., Clinton and Trump) (Mohammad et al., 2016b). Following this work, we focus on stance and code tweet instances as stance-for, against and neutral with respect to target companies. Note that, in our application domain, the zero-sum context inherent in the political domain used for SemEval-2016 does not apply; rejection of one company doesn't necessarily imply support of other companies. Indeed, in our specific case, tweet authors with environmentalist inclinations tend to reject all mining companies.

## 5 Conclusion

In this paper, we present SIRTA, a novel real-time text analytic system coupled with sophisticated monitoring techniques to help organisations manage their social license to operate (SLO) over time. Our experimental results and a case study show its effectiveness and applicability. The work could be furthered in a number of directions. First, our multi-label risk classifier currently does not consider the potential correlations among the risk factors mentioned in a post. It might be helpful to examine whether these correlations exist, and, if they do, refine the model. Second, our current strategy for aggregating the SLO scores of individual posts treats each post equally. We are considering employing a weighting function instead. Third, over time, the underlying distributions of the social media posts may shift, and our text classification models may need to be updated requiring a retraining policy. Finally, we plan to extend SIRTA to other organisations, sector or technology and extend our text analytics to support languages other than English (such as Korean and Japanese).

## References


Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance detection with bidirectional conditional encoding. *arXiv preprint arXiv:1606.05464*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Neil Gunningham, Robert A. Kagan, and Dorothy Thornton. 2004. Social License and Environmental Protection: Why Businesses Go beyond Compliance. *Journal of the American Bar Foundation*, 39(2):307–341.

Aditya Joshi, Ross Sparks, James McHugh, Sarvnaz Karimi, Cecile Paris, and Raina MacIntyre. 2019. Harnessing tweets for early detection of an acute disease event. *Epidemiology*, 31:90 – 97.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431. Association for Computational Linguistics, April.

Stephen H Kan. 2002. *Metrics and models in software quality engineering*. Addison-Wesley Longman Publishing Co., Inc.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, October. Association for Computational Linguistics.

M. Larsen, T. T. Boonstra, P. Batterham, B. B. O'Dea, C.Paris, and H. Christensen. 2015. We feel: Mapping emotions on Twitter. *IEEE Journal of Biomedical and Health Informatics (JBHI)*, 9:1246 – 1252.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Kieren Moffat and Airong Zhang. 2014. The paths to social licence to operate: An integrative model explaining community acceptance of mining. *Resources Policy*, 39(1):61–70.

Kieren Moffat, Justine Lacey, Airong Zhang, and Sina Leipold. 2016. The social licence to operate: a critical review. *Forestry*, 89:477–488.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016a. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41.

Saif M. Mohammad, Swetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016b. SemEval-2016 Task 6: Detecting stance in tweets. In *Proceedings of the International Workshop on Semantic Evaluation*, pages 31–41. ACM, jun. `https://www.aclweb.org/anthology/S/S16/S16-1003.pdf`.

Saif M Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2017. Stance and sentiment in tweets. *ACM Transactions on Internet Technology (TOIT)*, 17(3):26.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of NAACL-HLT 2013*, pages 380–390. ACL.

Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Parinaz Sobhani, Saif Mohammad, and Svetlana Kiritchenko. 2016. Detecting stance in tweets and analyzing its interaction with sentiment. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 159–169.

Qingying Sun, Zhongqing Wang, Qiaoming Zhu, and Guodong Zhou. 2018. Stance detection with hierarchical attention network. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2399–2409.

Grigorios Tsoumakas and Ioannis Katakis. 2007. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13.

Stephen Wan and Cécile Paris. 2015. Understanding public emotional reactions on Twitter. In *Proceedings of the Ninth International AAAI Conference on Web and Social Media*, pages 715–716. AAAI.

# A Appendix

## A.1 Datasets

### A.1.1 Query Terms for SLO Relevance Annotation

| Type | Query Terms |
|---|---|
| Hashtag | #StopAdani, #GoAdani, #StopBHP, #StopRioTinto, #StopFortescue, #StopSantos, #nonewcoal, #NoNewCoalMines |
| Username | StopAdani, StopAdaniNoon, StopAdaniMelbs, StopAdaniK, StopadaniB, stopadanieltham, StopAdaniTSV, stopadanisydney, StopadaniGC, StopAdaniCairns, StopadaniW, StopAdaniGTown, stopadaninoosa, stopadanibowen, adani_stop, StopBHP, AdaniOnline, bhp, RioTinto, SantosLtd, FortescueNews, kennecottutah, NSWMC, CMEWA, QRCouncil, WoodsideEnergy, MiningNewsNet, ozmining, miningcomau, Mining_EnergySA, AUMiningMonthly, MineralsCouncil, Austmine, MiningWeeklyAUS, AuMiningReview |
| Text | stopadani, adani, goadani, santosltd, santos, bhpbilliton, bhp, riotinto, rio tinto, woodside, woodsideenergy, woodside petroleum, woodside energy, fortescuenews, fortescue, metals, fortescue, whitehaven, whitehavencoal, iluka, ilukaresources, iluka resources, oilsearchltd, oil search, cuestacoal, cuesta coal, cuesta, cqc, newmont, newmont mining |

### A.1.2 SLO Risk Annotation

**Query Terms for Automatic Training Set Annotation:**

| Category | Instances | Query Terms |
|---|---|---|
| Social | 12960 | culture, support, live, land, public, approve, traditional, humanity, licence, human, labor, moral, national, land, donate, local, farm, vote, trust, life, multinational, regional, party, generation |
| Economic | 13540 | fund, financial, business, work, economic, import, money, job, employ, invest, spend, pay, market, cost, productivity, deposit, donation, asset |
| Environmental | 7405 | environment, destroy, approve, reef, insanity, enviro, climate, danger, greatbarrierreef, climatechange, reefnotcoal, flood, renewable, river, groundwater, poison, agriculture, save, protect, threaten |

**Guidelines for Manual Test Set Annotation:**
(1) Select *Economic* if the message is about the economic value of the company or its production, about shareholders, or about any employment/staff related issues (hiring, firing, Health and Safety). Examples:

- share price movements, or an indication that the company is doing well/bad.
- jobs (new hires(+ve) and cuts(-ve)). (e.g., "[company]" is recruiting/laying off")
- big economic wins.
- Positive or negative movement in terms of commodity price (e.g., iron price per ton).
- Positive/negative economic forecasts for the industry/company.
- Mentions of shareholders.
- Health and safety matters or working conditions.

(2) Select Social and cultural if the focus of the message is about how the company interacts with the community and how its activity affects the community. Sample topics include: health and education, community support services, social engagement with government, the cultural value of a site (e.g., sacred sites) used by the organisation. Any protest activity or activity trying rally for a cause is taken as "Social and Cultura" (this does not include shareholders revolt, which would be "Economic and Employment"). Examples:

- Government calls for a company to be investigated.
- Wining and dining government officials.
- Any interactions between the government and the mining company.

- Company partnerships with schools, hospitals, communities, that relates to funded programs or infrastructure.
- Any danger to valued sacred sites or cultural artefacts, e.g., art.
- Any dangers to the cultural way of life for local inhabitants.
- Any community issue that surfaces in relation to mine operations or management.
- Protests.

(3) Select *Environmental* if the message is related to natural environment, including fauna, flora, water, air, climate, etc. Examples:

- Discussions about the impact on the natural environment.
- Comments on environmental impact of the company, its decisions and actions.

### A.1.3 SLO Stance Annotation

**Query Terms:**

| Stance | Instances | Query Terms |
|---|---|---|
| Favour | 24255 | GoAdani, AdaniOnline, bhp, RioTinto, SantosLtd, FortescueNews, kennecottutah, NSWMC, CMEWA, QRCouncil, WoodsideEnergy |
| Against | 19311 | #StopAdani, #StopBHP, #StopRioTinto, #StopFortescue, #StopSantos, #nonewcoal, #NoNewCoalMines, StopAdani, StopAdaniNoon, StopAdaniMelbs, StopAdaniK, StopadaniB, stopadanieltham, StopAdaniTSV, stopadanisydney, StopadaniGC, StopAdaniCairns, StopadaniW, StopAdaniGTown, stopadaninoosa, stopadanibowen, adani_stop, StopBHP |
| Neutral | 19317 | MiningNewsNet, ozmining, miningcomau, MiningEnergySA, AUMiningMonthly, MineralsCouncil, Austmine, MiningWeeklyAUS, AuMiningReview |

**Guidelines for Manual Test Set Annotation:**
*for*: The coder infers from the tweet and its context that the author supports the target either because:

- the tweet explicitly supports the target.
- the tweet supports something/someone else aligned with or supporting the target or rejects something/someone else not aligned with or supporting the target.
- the tweet can be seen, in context, to support the target, either because:
  - the tweet author's profile lists positions consistent with support of the target.
  - the tweet discourse context places the tweet in support the target either by echoing support for the target in other tweets or by opposing rejection for the target in other tweets.

*against*: The coder infers from the tweet and its context that the author rejects the target.
*neutral*: The coder infers from the tweet or its context that the author neither supports nor rejects the target because:

- the tweet states no position consistent with support or rejection of the target.
- the tweet re-posts information only, with no clear hint as to the author's stance.
- the tweet context gives no hints as to the tweet author's stance.