

# Introduction to Data Analytics coursework

This is a report on three tasks: 1. Sentiment classification of a dataset containing news and social media (Maia et al., 2018), 2. Named entity recognition on a dataset from MEDLINE (Collier et al., 2004), and 3. Data visualisation on malnutrition data from the Global Health Observatory. The first two tasks were carried out using Jupiter notebook, and the final task was conducted using Tableau. For the first two tasks, the results indicate that while we are able to classify sentiment and perform named entity recognition, we could perhaps improve our models by making some additional changes to the data during data preparation, performing cross validation, and also capturing true negatives during performance testing. For task 3, the results indicate that while the users found the visualisation enabled them to carry out their tasks accurately, and in a timely manner, there is room for improvement when it comes to accessibility, and user experience, specifically relating to the geographic maps used in the visualisation.

## 1 Task 1: Sentiment classification

Here, sentiment analysis was performed on the FiQA Sentiment Analysis dataset.

### 1.1 Method and software implementation

A CountVectorizer was used to extract a bag of words that represents the vocabulary of the data. However, this alone is not enough to give us accurate sentiment analysis. That's because this single representation of the data doesn't take into account instances where two words, for example, '*does not*' represent a single meaning. The bag of words representation would count these as separate words, when in fact, for sentiment analysis, it would be better to understand them as having a single meaning. Therefore, the CountVectorizer has been set to also take into account bigrams for words as an additional feature.

Additionally, lemmatization was used to normalize the text. It reduces words to their root forms, which in return decreases the vocabulary size and ultimately should help to improve the performance of our classifier.

Finally, lexicon-based features for words in the data were also implemented. This works by counting how many words are in a positive lexicon or a negative lexicon. This was used because the training set might not have some words that the model could end up seeing in the validation/testing set. This should help to improve results, as the model will have been given some form of prior knowledge of the sort of data it might be dealing with in testing.

A logistic regression model was chosen as the model for sentiment analysis. This is because using a Naïve Bayes model, while simple, would have limitations that could impair its performance in a realistic setting. For example, it makes the assumption of conditional independence, which in reality is often not the case. The truth is that there are often dependencies between words in any given bag of words. For instance, if we have a phrase 'very bad', which is a bigram, then the unigram of 'bad' is very likely to occur. This suggests that there is a strong dependency between them, which Naïve Bayes ignores. A logistic regression model would enable us to take this into consideration.

The general process is as follows: we load the data and separate it into a training set, a validation set, and a final testing set. We use the training set to train the model. We use the validation set to see how well the model predicts sentiment labels based on the different combination of features that have been applied. This allows us to tune the design for our model to choose the optimal combination of features. We then use the testing set to see how well our optimized model generalizes to new data. The aim is to understand how our model performs using unigrams and bigrams as a feature, compared with how well it performs when using positive/negative lexicon occurrence as an additional feature, in

combination with unigrams and bigrams. The hypothesis was that the model should perform better after combining all these features.

## 1.2 Method evaluation and results

A confusion matrix was used to provide initial assessments of performance on the validation set and the testing set. The confusion matrix shows how many datapoints were correctly classified or incorrectly classified for each class label by our model. For example, the chart below shows that on the testing set, our model correctly predicted 115 docs as positive, and 34 as negative. Class label 0 represents negative, class label 1 is neutral and class label 2 is positive.

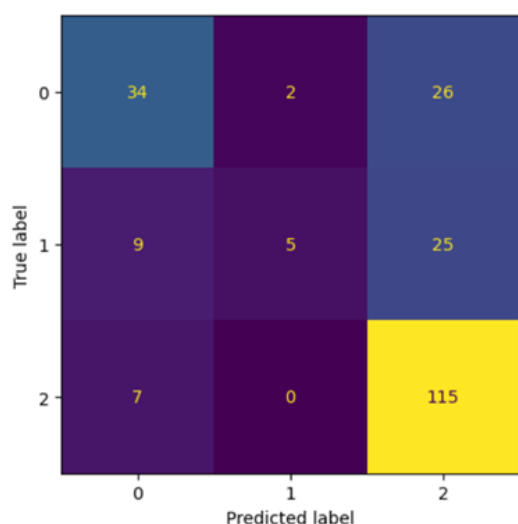


Figure 1 – Confusion matrix

Classifier-centric metrics available from the `classification_report` function in sci-kit learn were used. Because we are dealing with three different classes (negative, neutral, positive) it provides us with a handy breakdown of different relevant classification metrics for each class such as:

- **Accuracy** - Number of instances correctly predicted as a fraction of our total dataset.
- **Recall** - Fraction of positive instances in our dataset correctly predicted.
- **Precision** – How many instances were correctly predicted as a fraction of the total number of positive predictions that our model has made.
- **F1** - Considers both recall and precision for a balanced representation.

For example, below is the classification report for our model on the testing set:

	precision	recall	f1-score	support
0	0.68	0.55	0.61	62
1	0.71	0.13	0.22	39
2	0.69	0.94	0.80	122
accuracy			0.69	223
macro avg	0.70	0.54	0.54	223
weighted avg	0.69	0.69	0.64	223

Figure 2 - Classification report

The above metrics all have their own limitations. For example, accuracy is not useful if we happen to come across large class imbalances in our dataset. Additionally, if recall is too high, then precision will decrease, and vice versa.

The F1 score provides a balanced approach between precision and recall. But the F1 score isn't easy to interpret on its own, unlike accuracy which is simply a fraction or percentage. Additionally, it doesn't take into account true negatives (where the model has correctly avoided making a wrong

prediction). To illustrate, this means that two models could have the similar F1 scores, even though one has a much higher rate of true negatives.

The classification report provides us with a macro F1 average (shown in the figure above as 0.54), which is the mean of the F1 scores of each class. This gives equal weight to underrepresented classes. However, it must be noted that as this is a metric that uses aggregation, it could mean that we aren't really getting a clear picture of the nature of the errors for our model by just using this score.

### 1.3 Potential improvements

A good way to help us investigate why a model makes errors is to look at them, so that we can better understand its limitations in its current form. This enables us to improve our classifier or general approach in the future. Therefore, mislabelled documents for validation and final testing on the test set were printed out. This makes it possible to investigate the documents for anything that sticks out at us. For example, when looking at the errors found during validation using the positive/negative lexicon features, most documents that are mislabelled appear to have the dollar symbol (\$), and they were predicted as belonging to class 2 (positive), when in reality they are either negative or neutral. This suggests that the model sees the dollar symbol as a positive marker, which could tilt the documents to be classified towards positive.

This means that we could potentially improve our method by using regular expressions to replace phrases that include with the dollar symbol and other symbols with a neutral term so that they have less of an effect on the model's classification attempts.

### 1.4 Identifying common themes or topics associated with negative or positive sentiment

Topic modelling was used to identify general topics. Here, each document is considered as a single document from which potential topics and their related words are identified. The Latent Dirichlet Allocation (LDA) method was chosen to achieve this. The process is as follows; we pre-process the data by tokenizing and lemmatizing. We then create a dictionary based on the pre-processed data, which enables us to get a bag of words representing the data. We feed all of this to an instance of the LDA model and set its num\_topics parameter to 10 to identify a maximum of 10 topics for each single document.

The LDA model extracts potential topics and their distributions for the documents. For example, the figure below shows the topic distribution for a document, and we can see that there are three prominent topics (identified as topic 3, topic 4, and topic 6).

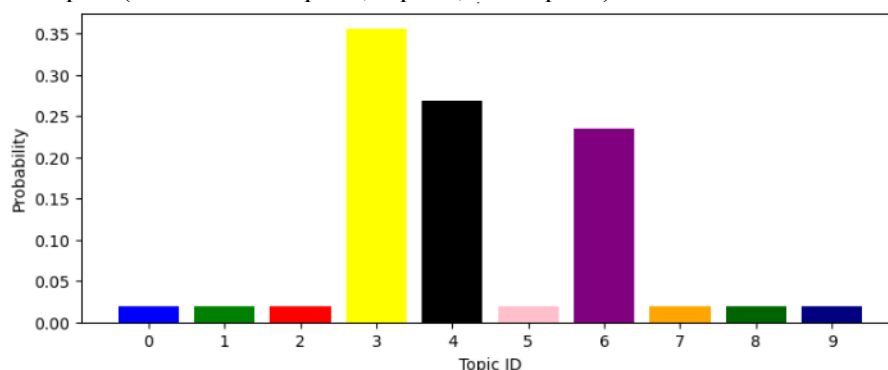


Figure 3 – Topic distribution for single document

Then, the validation set is used to enable us to identify how different topics are dispersed across different sentiment classes. This was done by first getting the topic distributions in sparse form, converted into dense vector representation, and then getting the means of those in a matrix representation for each label and the associated means for the different topics. This made it possible to generate a heatmap that shows how topics occur across the different sentiment classes. For instance, the figure

below shows that in the validation set, topic number 9 (it has the warmest colour) is common for documents across all classes despite their classification of negative, neutral, or positive.

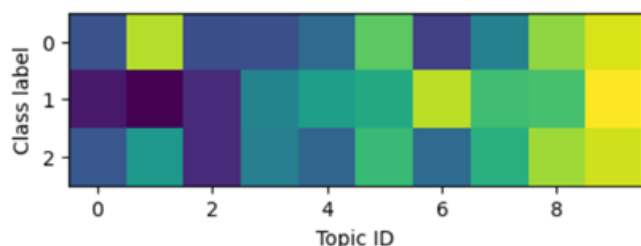


Figure 4 – Heatmap for topics on validation set

The LDA model also enables us to see which words are associated with topics so that we can understand what the topics are about. Using the model’s `print_topics` method, we are able to see that the top three words associated with this topic are ‘https’, ‘aapple’ and ‘tsla’. This could be because the two companies, Apple and Tesla, being two of the most discussed companies in the world, will have many headlines and tweets concerning them across all sentiments. The term ‘https’ here occurs because articles and tweets most likely contain links for advertisements or other articles for further discussion. Although, as the term ‘https’ on its own doesn’t really mean anything in the context of sentiment, it would be better if we used regular expressions to replace or remove instances of it and other URL related terms across the documents to prevent it from affecting our model. It should be noted that when run on the test set, the results showed the following:

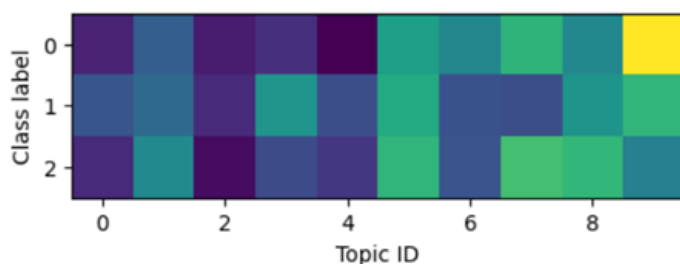


Figure 5 – Heatmap for topics on test set

While topic 9 still shows the most activity, across sentiment class labels, this isn’t as defined as in our validation set. This could be because the same number of instances don’t occur in our test set as opposed to the validation set. There are documents in our test set that aren’t in our validation set, which naturally could have slightly offset the emphasis on any single topic. Using cross validation could provide us with some better performance when closing the gap between our validation set results with our test results. Additionally, the LDA model relies on us to define a pre-set number of topics to identify, which in reality we can’t do accurately. One possible solution to this problem could be to use the Hierarchical Dirichlet Process (HDP) instead of LDP, which is able to learn the ideal number of topics to identify for a particular dataset. It provides a probability distribution for all possible topics numbers, which, while not perfectly accurate, would be better informed than starting with a randomly chosen fixed number as in the case of LDA.

## 2 Task 2: Named entity recognition

Here, named entity recognition was performed on the BioNLP 2004 dataset.

### 2.1 Chosen method

There are two options that were considered for this task. One being Hidden Markov Model (HMM), and the other one being Conditional Random Field (CRF). HMM is a generative model similar to Naïve Bayes and doesn’t directly learn to predict labels. This model has its own benefits (for instance, it’s faster), however, it is generally less accurate than CRF. Therefore, CRF was chosen for this task.

CRF learns to directly compute the probability of a sequence of labels for any set of tokens. It considers each sequence label to be dependent on the entire sequence of tokens in a given document. As such, the model takes into account all of the tokens for a document when predicting labels. It must be noted however, that while this model generally provides more accuracy, it comes with its own limitations. For example, it is more computationally expensive and complex, which means it costs more time and resources to train, compared with HMM.

## 2.2 How entity spans are encoded as tags for each token

An entity span is a smaller subsection of a larger phrase that identifies an entity such as a person, place, time/date, or organisation. For example, consider the sentence ‘Increased demand leads to higher stock prices for Gamma Health after turmoil.’ Here, ‘Gamma Health’ would be an entity span. With Named Entity Recognition (NER), we want to label each word in the sequence with an entity type or indicate that it is not an entity. For example, ‘Gamma Health’ could be labelled as an organization (ORG). Entities often span more than one single token, as is the case for ‘Gamma Health’. This means we need to be able to label where the entity starts (token= ‘Gamma’) and ends (token= ‘Health’). To do this, we would tag these tokens as follows:

- B-ORG for ‘Gamma’ token
- I-ORG for ‘Health’ token
- O- this tag would be set for the ‘after’ token that follows the entity in the sentence, to indicate the end of the ‘Gamma Health’ entity span. O (for Outside) means the word ‘after’ is not an entity.

For our named entity recognition task, we conduct this same process to identify the following named entity types on our dataset that consists of abstracts from MEDLINE. Namely:

1. B-DNA/I-DNA
2. B-protein/I-protein
3. B-cell\_type/I-cell\_type
4. B-cell\_line/I-cell\_line
5. B-RNA/I-RNA

## 2.3 Software implementation

Firstly, the dataset is loaded and then split into training, validation, and testing sets. The model, a CRFTagger instance, is then trained using the training set. The model is then given the validation set to predict tags for the validation set. The F1 score for each label is then calculated to see how well the model predicts against the validation set for every label. A macro averaged F1 score is also provided as a single F1 score that can be used to represent the general efficacy of the model. This same process is repeated with two other, customized, instances of the CRFTagger model:

1. A customized model that identifies the current, previous, and next words as additional features for the model to use. This helps us to identify named entities that are bigrams. This should provide improved performance over the version that doesn’t use these features.
2. A second customized model that adds parts of speech (POS) tags as an additional feature on top of the features from the previous model. Parts of speech tags are tags that represent the token’s syntactic role in a given sentence. This will help identify proper nouns, which is a good indicator for named entities. The expectation is that this will provide even better performance over the previous two models.

Of the three models, the model that performs best, having the best macro averaged F1 score, is then used to make predictions on the test set to see how well it generalises.

## 2.4 Method evaluation and results

The F1 score provides a balance between precision and recall. But the F1 score isn’t easy to interpret on its own, unlike plain accuracy. Also, we must recall that it doesn’t consider true negatives.

This means that our different models could have a similar F1 scores for the classes as shown below, even though one model could much higher rate of true negatives than the others.

We also calculate a macro F1 average for each model as in the previous task. However, this is still a metric that relies on aggregation. As such, we are not getting a clear picture of the errors in our classifications at this point. The F1 scores found for the models is as follows:

	DNA	protein	cell_type	cell_line	RNA	Macro Avg.
<b>Plain model</b>	0.649386	0.788591	0.682566	0.626761	0.700855	0.689632
<b>Prev-Next-WRD-Model</b>	0.692886	0.819590	0.751623	0.712238	0.694215	0.734110
<b>POS_model</b>	0.699780	0.821258	0.749798	0.713024	0.680672	0.732906

Figure 6 – Table showing F1 scores for models across entity types on validation set.

This indicates that while it was expected the POS\_model to perform best, the second model (Prev-Next-WRD\_model) appears to perform slightly better, even though it has less features. As a result, this model was chosen to run against the test set to see how well it generalises to unseen data. The results are as follows:

	DNA	protein	cell_type	cell_line	RNA	Macro Avg.
<b>Prev-Next-WRD-Model</b>	0.747716	0.724228	0.680238	0.581673	0.649351	0.676641

Figure 7 – Table showing F1 scores for the best performing model on the test set.

This shows that this model’s performance on the test set is close to that on the validation set, which indicates that our model might generalise well on unseen data.

However, as the F1 score doesn’t take into account true negatives, it may very well be that the POS\_model was able to identify more true negatives during validation, and despite a lower F1 score, could perform better in a real setting, and so could potentially be a better candidate to run against the test set. Especially when we consider that its macro averaged F1 score is not so different from that of the chosen model.

Additionally, we have only used a single validation set to run against. The POS\_model appearing to perform slightly worse, could simply be due to variance when training each model on the validation set (if we test the models multiple times, we could get different results). Using cross validation, we might be able to have a clearer understanding of which model performs better on average across various splits of the data set. Depending on the results, this could potentially lead us to choose the POS\_model instead to run on the test set in the future.

### 3 Task 3: Data visualisation

For this task, Tableau was used to create plots that should allow users to explore two datasets on child malnutrition.

#### 3.1 Task abstraction

First, we need to consider what we want the end-user to achieve, and how. We then map these to the three levels of **Actions** for visualisation (Munzner, 2014) that we want the user to be able to carry out, as follows:

- **Analyze:**
  - *Discovering:* Find new knowledge by, for instance, being able to verify if there is a link between wealth and malnutrition.
  - *Presenting:* To be able to succinctly tell a story that answers the questions.
  - *Enjoy:* Create enjoyable visualisation so the user spends more time exploring the data, which could lead to new insights or questions.
- **Search:**
  - *Lookup:* For example, by enabling the user to look up a specific country with regards to a feature of malnutrition.

- *Locate*: Enable users to locate a country on a world map showing malnutrition related information.
- *Browse*: For example, the user should be able to identify relevant countries based on a question. Such as countries where malnutrition has changed (up or down) over time.
- **Query**: Enabling the user to carry out search actions through by being able to:
  - *Identify* a specific target, such as how has malnutrition changed over time for India?
  - *Compare* multiple targets: How has underweight changed in the poorest quintiles compared with the richest?

In terms of **Targets**, we want the user to be able to view overall trends and outliers, such as whether malnutrition has increased or decreased over time for certain countries or globally. We also want the user to be able to find any correlations between attributes of interest. For example, whether there is a link between malnutrition prevalence and wealth quintile.

### 3.2 Visualisation techniques justification

The workbook displays data using the following visualisations:

- Stacked bar charts
- Line charts
- Geographic maps
- Stories and dashboards

These are formatted in ways to help ensure that the user's visual queries can be served effectively and rapidly. A subsection in this chapter is dedicated to each type of visualisation, listing some justifications for why it was used.

#### 3.2.1 Stacked bar chart

Average wasting prevalence in richest and poorest quintiles

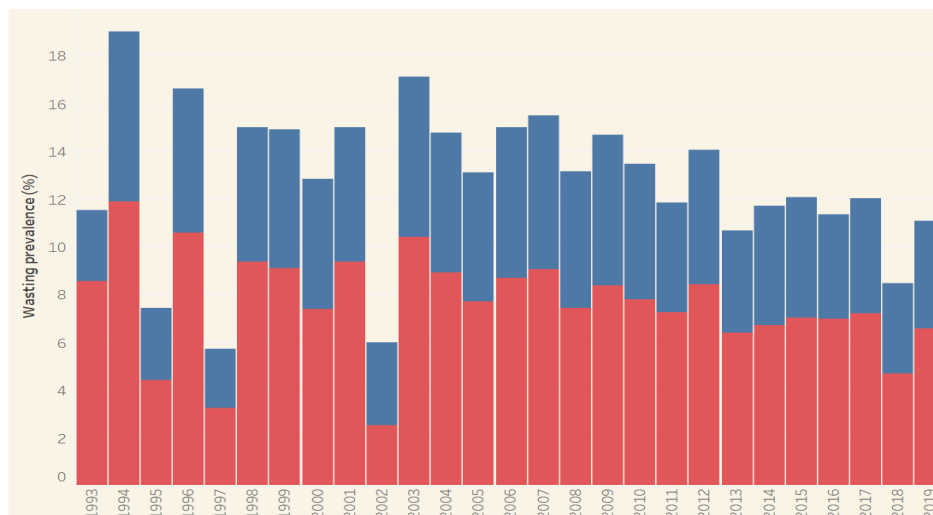


Figure 8 - Snapshot of a stacked bar chart in Tableau.

- **Height** (Area) magnitude channel enables us to indicate quantitative values such as percentage prevalence. This assists the user when making visual queries to determine relative sizes quickly and accurately.
- The **colour hue** identity channel enables us to indicate categorical labels, such as different quintiles. Colour is used to facilitate separability. For example, here red represents the poorest quintile, and green represents the richest. Stark distinct colours are used whenever possible to support contrast for visual distinctness.
- **Alignment** on a common scale is used in recognition of Weber's Law to enable the user to make more accurate comparative relative judgements about the length of the different bars.

- **Horizontal positioning** channel enables us to map to individual ordered years and show the comparative change year over year.

### 3.2.2 Line chart

Change in wasting prevalence worldwide across all quintiles

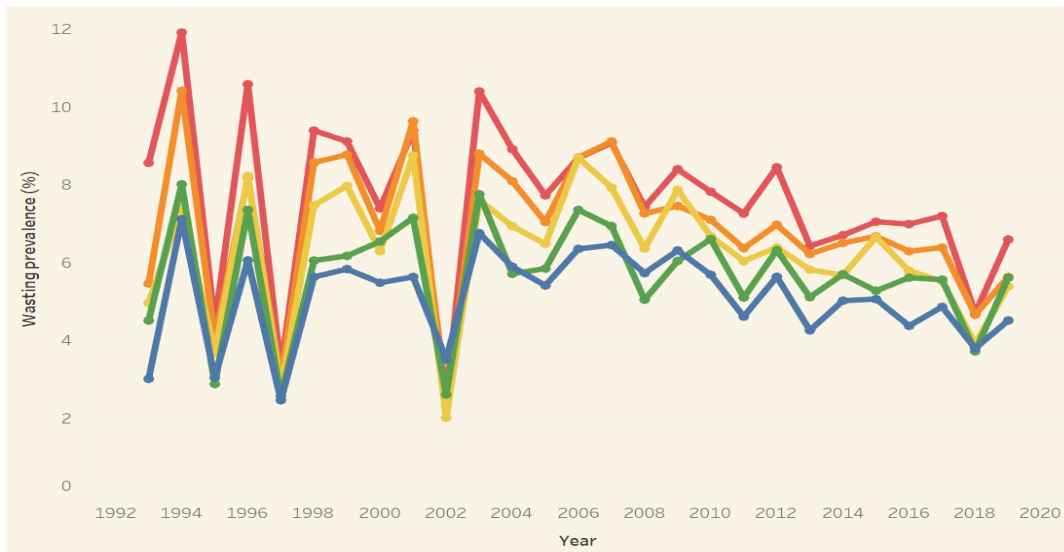


Figure 9 - Snapshot of a line chart.

- **Colour hue** channel is used as the identity channel to enable us to indicate categorical labels, such as quintiles. This gives us separability between labels.
- **Vertical positioning** is used as the magnitude channel to represent quantitative values such as percentages.
- **Horizontal positioning** channel enables us to map individual ordered years and show the comparative change year over year.

### 3.2.3 Geographic maps

Average wasting prevalence (%) for children under 0-1 and 2-5 years.

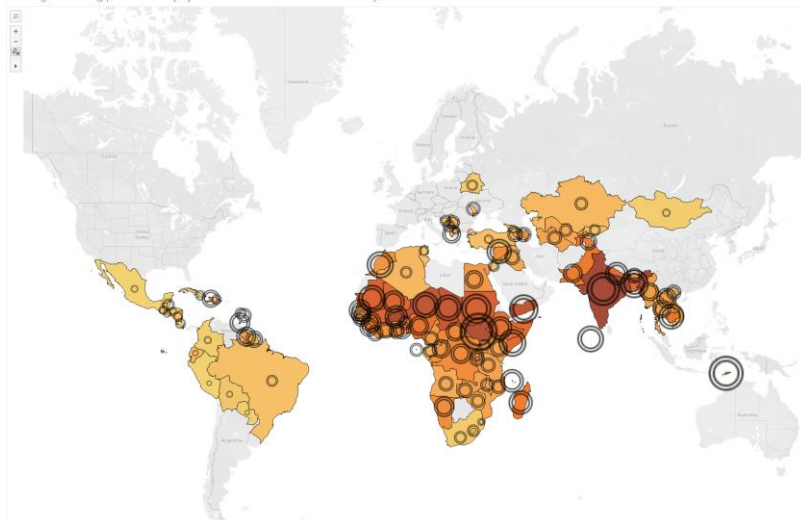


Figure 10 - Snapshot of a geographic map in tableau.

- **Colour saturation** is used as the magnitude channel to indicate quantitative values such as percentages. This facilitates discriminability between different countries.
- A second layer of **shapes** (circles in this case) is used to display the identity of an additional dimension, and shape **size** is used to indicate its quantitative magnitude. For example, here the circles relate to the stunting prevalence for a country for children aged 2-5 years, and size is



used to represent the value for each country. This enables us to achieve discriminability at a glance for this second layer. It also enables smaller countries with higher rates to be more visible, such as Timor-Leste (in the bottom right above Australia).

### 3.2.4 Stories and dashboards

To ease the cognitive load for the user in trying to answer the key questions, figures are also grouped together in dashboards for key features (such as a single dashboard for all data on under-weight percentage prevalence). Different dashboards are then combined using Stories, which are titled according to main the questions that the user needs to be able to answer. The user can start by looking at a story to help answer a question. For a more granular view, they can look at individual dashboards, and then move to individual charts for even more granularity. In some instances, paging is also used for some views to deal with the reality of limited display capacity (preventing too much information density) and human cognitive and perceptual capacity. Below is an example Story that consists of multiple dashboards (tabs) and multiple charts from each dashboard:

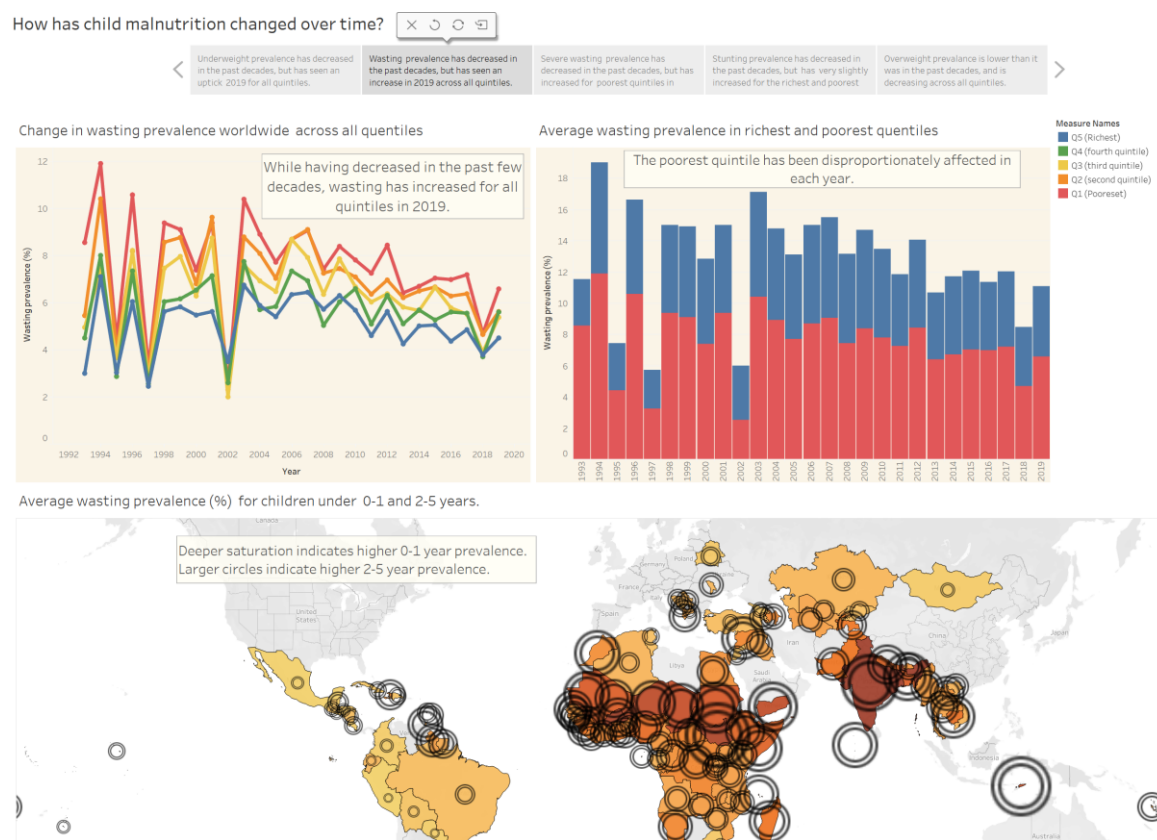


Figure 11 - Snapshot of a Story sheet in Tableau.

Consistent colours are used to identify quintiles across all Stories, for example, Q5 (richest) is always the same blue, and Q1 (poorest) is always the same red. This way users can become quickly familiarized with the colours across the figures in the stories, further reducing the cognitive burden for users during navigation. One thing to note is that colours can often be interpreted differently across cultures. In western cultures, red is seen as dangerous or at least has some general sense of warning, and so Q1, being the poorest quintile being assigned to red, might make some sense to us. However, in some cultures, such as the Chinese culture, red is a good and lucky colour associated with wealth, and so it would make more sense for Q5, the richest quintile, to have this colour in that case. This is why in general visualisation can be difficult to get right as it all often depends on the particular situation at hand.

### 3.3 Validation

To assess the quality of our visualisation, an online questionnaire was constructed (link in appendices 5.1). This questionnaire consists of questions that can be broken down into separate sections that are each concerned with a specific measurement. The first section consisting of three questions, is focused on evaluating the accuracy of task performance, i.e., whether users are able to get accurate answers from our visualisation. The three questions in this section ask for specific answers that requires the user to intuitively navigate the Stories in the workbook to find them. For example, ‘Has underweight prevalence increased or decreased for Nigeria?’ As questions like this one are more granular than the main, higher level, goal questions of the visualisation task, it means that if the users are able to accurately answer them, they are more likely to be able to answer the main goal questions that our workbook seeks to help address. The results (see appendices 5.2 for all results) for this section indicate that all users have managed to answer the questions correctly. This means they were able to accurately perform their tasks.

The second section of the questionnaire (question 4 and 5) is there to help us understand how much time the users needed to answer the first set of questions. We want to minimize the amount of time a user needs to find answers to questions in our workbook. A larger timeframe would indicate that our visualisation is perhaps too complicated, or simply not intuitive enough. The results indicate that the users on average were able to find answers to each question in around 1.33 minutes. This of course leaves for some room for improvement, but isn’t too much time considering the granularity of the questions that were asked, coupled with the fact that we asked the users to estimate how long it took them in minutes as opposed to seconds, 1 minute is the smallest estimate they could provide.

The third section is about the user experience of our users for the visualisation. We want to understand the users’ subjective view of the visualisation, along with whether the visualisation was accessible to them in general and where it might have failed in doing so. Therefore, this section includes questions like ‘How helpful did you find the colours?’, ‘How legible was the text to you?’, and also allows room for users who have visual impairments to communicate their experience as well. For this section the results definitely suggest there is good room for improvement. On a scale of 1 to 5, our users rated the visualisation an average of 3.33 when asked how accessible they found the visualisation. Since the users rated the text as legible and the colours as helpful, this result could be because they also generally didn’t find it easy to navigate the Stories due to their structure.

Indeed, when asked to rate how difficult they found the navigation, users rated it an average 2.67 out of 5. This suggest that while they didn’t find it too difficult, they also didn’t find it easy. And so, understanding why this is the case, could help us improve the user experience. In fact, when asked how we can make the visualisations more accessible, feedback from users ranged from changing the map type in the Stories, particularly the circles overlayed on the maps, to adding filters to the maps to select individual countries. As a result, making changes to our geographic maps in the visualisation in accordance with this feedback could make it easier for users to use, and help us to improve the overall usability and accessibility of the visualisation in the future.

## 4 Reference

- Jurafsky, Dan, and James H Martin. 2023. *Speech and Language Processing* Third edition, [https://web.stanford.edu/~jurafsky/slp3/ed3book\\_jan72023.pdf](https://web.stanford.edu/~jurafsky/slp3/ed3book_jan72023.pdf)
- Maia, Macedo & Handschuh, Siegfried & Freitas, Andre & Davis, Brian & McDermott, Ross & Zarrouk, Manel & Balahur, Alexandra. (2018). *WWW'18 Open Challenge: Financial Opinion Mining and Question Answering. WWW '18: Companion Proceedings of the The Web Conference 2018*. 1941- 1942. 10.1145/3184558.3192301.
- Munzner, T. (2014). *Visualization Analysis and Design* (1st ed.). A K Peters/CRC Press. <https://doi-org.bris.idm.oclc.org/10.1201/b17511>

- Nigel Collier, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Tateisi, and Jin-Dong Kim. 2004. *Introduction to the Bio-entity Recognition Task at JNLPBA*. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)*, pages 73–78, Geneva, Switzerland. COLING.

## 5 Appendices

### 5.1 Questionnaire

Link to survey: [https://forms.office.com/Pages/ResponsePage.aspx?id=MH\\_ksn3NTkql2rGM8aQVG6S0uZSy-b9IlnMYjzbpXIpURUtEQVo2M1FaRzg0QjRGTFhRTzZXS1JLMy4u](https://forms.office.com/Pages/ResponsePage.aspx?id=MH_ksn3NTkql2rGM8aQVG6S0uZSy-b9IlnMYjzbpXIpURUtEQVo2M1FaRzg0QjRGTFhRTzZXS1JLMy4u)

20/05/2023, 12:18

Survey for data visualisation on malnutrition in children



\* Required

\* This form will record your name, please fill your name.

Evaluating accuracy of task performance

1. How has underweight prevalence changed over time? \*

- ☐ It has increased
- ☐ It has decreased
- ☐ It has stayed the same

2. Comparing the richest vs poorest quintile, which one has suffered disproportionately from malnutrition across all years? \*

- ☐ Q1 (poorest)
- ☐ Q5 (richest)

3. Has underweight prevalence increased or decrease for Nigeria? \*

- ☐ Increased
- ☐ Decreased

## Time used to perform task

4. On average, how long (in minutes) did it take you to find the answer for each individual question above? \*

1	2	3	4	5
---	---	---	---	---

5. How difficult or easy was it for you to find the answers to the questions? \*

- ☐ Easy
- ☐ Neutral
- ☐ Difficult

## Evaluating user experience

6. How legible was the text to you? \*

- ☐ I could read the text easily
- ☐ Some small effort was needed to read the text
- ☐ The text was mostly not legible.
- ☐ I am unable to read any of the text (accessibility)

7. How helpful did you find the colours? E.g. The consistency of colours used for quintiles in graphs like bar charts and line charts. \*

- ☐ Helpful
- ☐ Neutral
- ☐ Unhelpful
- ☐ The colours are inaccessible to me.

8. On a scale of 1 to 5 (1 being easy and 5 being difficult), how difficult was it to navigate the Stories in the workbook? \*

1	2	3	4	5
---	---	---	---	---

9. On a scale of 1 to 5 (1 being inaccessible, 5 being very accessible), how would you rate the visualisations in terms of accessibility? \*

1	2	3	4	5
---	---	---	---	---

10. In as few words as possible what do you feel can done do to make the visualisations more accessible? \*

---

This content is neither created nor endorsed by Microsoft. The data you submit will be sent to the form owner.




## 5.2 Results

### Survey for data visualisation on malnutrition in children




3 Responses	08:05 Average time to complete	Active Status
----------------	-----------------------------------	------------------

[View results](#)

 [Open in Excel](#) ...

1. How has underweight prevalence changed over time?

[More Details](#)

 It has increased	0
 It has decreased	3
 It has stayed the same	0



2. Comparing the richest vs poorest quintile, which one has suffered disproportionately from malnutrition across all years?

[More Details](#)

 Q1 (poorest)	3
 Q5 (richest)	0



3. Has underweight prevalence increased or decrease for Nigeria?

[More Details](#)

 Increased	0
 Decreased	3



4. On average, how long (in minutes) did it take you to find the answer for each individual question above?

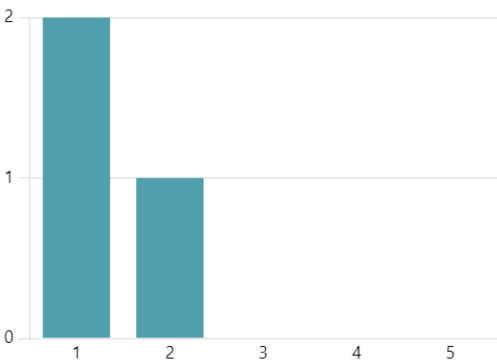
[More Details](#)



4. On average, how long (in minutes) did it take you to find the answer for each individual question above?

[More Details](#)

1.33  
Average Rating



5. How difficult or easy was it for you to find the answers to the questions?

[More Details](#)

Easy	1
Neutral	2
Difficult	0



6. How legible was the text to you?

[More Details](#)

I could read the text easily	3
Some small effort was needed t...	0
The text was mostly not legible.	0
I am unable to read any of the t...	0



7. How helpful did you find the colours? E.g. The consistency of colours used for quintiles in graphs like bar charts and line charts.

[More Details](#)

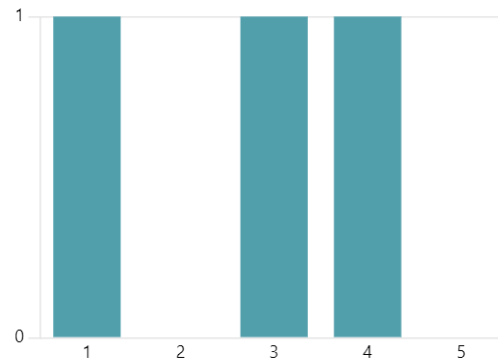
Helpful	3
Neutral	0
Unhelpful	0
The colours are inaccessible to ...	0



8. On a scale of 1 to 5 (1 being easy and 5 being difficult), how difficult was it to navigate the Stories in the workbook?

[More Details](#)

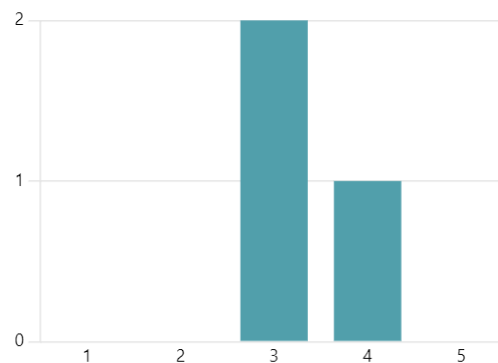
2.67  
Average Rating



9. On a scale of 1 to 5 (1 being inaccessible, 5 being very accessible), how would you rate the visualisations in terms of accessibility?

[More Details](#)

3.33  
Average Rating



10. In as few words as possible what do you feel can done do to make the visualisations more accessible?

[More Details](#)

3  
Responses

Latest Responses

"Add filters to be able to select specific countries on world map."  
"There seems to be a lot going on. In other to answer the questions I had to ..."  
"Perhaps change the type of map used at the bottom of story point 1 - How ..."