**Challenge Overview:**

*Language is a powerful means of communication for all members of a society. A shared, common language is one of the most important features of a community, be it geographical or online communities. The continuous use of languages with shared roots, can provide a historical narrative of different communities of people. The study of language is a multidisciplinary field, with varied mediums, such as oral, written or visual. The aim of today’s challenge is to observe the differences & commonalities & use these features to build a language recognition model.*

**The challenge this year will be separated into two problems.**

1. **The Language Identification Task – 60%**
2. **The Sentiment Analysis Task – 40%**

**Brief rules for successful entries:**

**Both tasks should be attempted, as the final score to determine a winner will be aggregated from combined scores of both tasks.**

**The use of external datasets is permitted.**

**The use of external APIs is NOT permitted.**

**With these conditions in place, it is necessary for all teams to submit a copy of their commented code, any external datasets (if used) and/or a concise, detailed description of their methodology.**

**Please note; submissions will close at 5pm, with winners announced at 7.30pm – feel free to enjoy the food, beverages & entertainment provided during the breaks!**

**The Specifics:**

**Task 1: Language Analysis:**

The first part of the challenge is to build a model which auto classifies a piece of text with the correct language it was written in. With a file given of 7 different languages, the data will be one sentence per id. The file is encoded with UTF-8 encoding.

The 7 languages are as follows:

* ENGLISH,
* FRENCH,
* GERMAN,
* ITALIAN,
* PORTUGUESE,
* SPANISH,
* JAPANESE

The data is provided in the following file: Languages.csv with the following columns:

1. ID - a unique id per row
2. Text - the sentence
3. Language - a label determining the language of text.

**Task 2: Language Analysis:**

The second task will be to do sentiment classification of one language for positive & negative sentiment. The data is in the file called: Sentiment.csv, with the following columns:

1. ID - a unique id per row
2. Text - the sentence
3. Sentiment - a label determining the sentiment of text.

**Approaching the challenge:**

**Initial Steps – Getting to know your data!**

To begin, some preparation should be done with the data, so the following are some steps which may be useful:

1. Open & Read the data with a tool of your choice (e.g. Python, R) – scan the fields to understand the structure.
2. **Beware!!** Both files contain Text data, which is **encoded in UTF-8.** To preserve the original encoding, UTF-8 should be considered, any other encoding may reduce the data integrity & therefore reduce model accuracy. (Note: ASCII encoding is not sufficient, as it only covers the English language’s character set!)
3. Check all fields for missing data or corrupted data.
4. Consider the variations of class labels – decide on one encoding per class & stick with it!
5. Begin with exploratory analysis, to review the different distributions of languages or sentiment.
6. Have you considered splitting your data for training & testing your model?
7. Will Upper or Lower case lettering & punctuation affect your model?

**Scoring:**

**Task 1: Language Analysis:**

Task 1 will be scored as a weighted sum: of the number correctly classified entities per language, as ratio of the total number of entities in a language.

**Task 2: Language Analysis:**

And Task 2 will be similarly scored, as a weighted sum of the correct count per sentiment grouping:

**Final Score:**

With the final score, given as follows:

**Files Submissions:**

**Please ensure all submissions contain the following header:**

**id, label**

**Task 1: Language Analysis:**

For task 1, a file called **language\_interim.csv or language\_final.csv** (Comma Separated/delimiter) should be in the form of two columns:

1. ID – the unique id of the text
2. Label – the classification language label from the model.\*

\*The results label should be keyed as follows:

|  |  |
| --- | --- |
| LANGUAGE | KEY\_ID |
| ENGLISH | **1** |
| FRENCH | **2** |
| GERMAN | **3** |
| ITALIAN | **4** |
| PORTUGUESE | **5** |
| SPANISH | **6** |
| JAPANESE | **7** |

A sample of the expected results format is available on the Google Drive shared with the email address under which your team provided on registration for the Datathon. The path for the language identification is as follows: **/home/teams/\*team\_name\*/results/language/sample\_results.csv**

**Task 2: Sentiment Analysis:**

For task 2, a file called **sentiment\_interim.csv or sentiment\_final.csv** (Comma Separated/delimiter) should be submitted with two columns in the following form:

1. ID – the unique id of the text
2. Label – the classification sentiment label from the model.\*\*

\*\*The results label should be keyed as follows:

|  |  |
| --- | --- |
| Sentiment | KEY\_ID |
| NEGATIVE | **0** |
| POSITIVE | **1** |

A sample of the expected results format is available on the Google Drive shared with the email address under which your team provided on registration for the Datathon. The path for the language identification is as follows: **/home/teams/\*team\_name\*/results/sentiment/sample\_results.csv**

**Important:**

To assess how your model(s) is(are) progressing, an **interim file (called: language\_interim.csv & sentiment\_interim.csv)** will be available for both tasks. An automatic scorer will run all day to collect any files submitted for the interim results.

At the end of the day, the final evaluation file **(called: language\_final.csv & sentiment\_final.csv)** can be submitted for evaluation, this will only be scored once & will determine the final result of the challenge.

This final result files will be used to identify the most accurate model in order to avoid the potential of the winning model having simply over-fit to the interim data. Due to this the final standings may be different to what was visible on the leaderboard before the end of the competition. However, the leaderboard will still give a good indication of overall model accuracy.

**Uploading & downloading files:**

As per the Google Drive link shared with your registered team email earlier in the week, please find all necessary datasets under **/home/datasets**, all relevant documentation under **/home/docs** and please submit your results for task1 under /home/teams/\*team\_name\*/results/language and task2 under /home/teams/\*team\_name\*/results/sentiment

**Data: /home/datasets**

**Documentation: /home/docs**

**Please submit your results for task 1, language identification: /home/teams/\*team\_name\*/results/language**

**Please submit your results for task 1, sentiment recognition: /home/teams/\*team\_name\*/results/sentiment**

**Finally, please submit any commented code, pseudocode or methodology under: /home/teams/\*team\_name\*/results/src**

**Disclaimer:**

**The views and opinions expressed in the dataset provided are those of the authors and do not necessarily reflect the official policy or position of AIB or its employees. All content provided is publicly available & opinions, views, commentary are not intended to malign any religion, ethnic group, club, organization, company, individual or anyone or anything.**

**Best of luck, enjoy hacking & if you have any questions, please ask the data scientists available!!!**

Datathon Schedule:

**8am -** Registration opens

**8-10am** - Breakfast served

**9.00am** - Opening announcement & housekeeping

**12.30 - 2pm** - Lunch served

**3pm** - Cupcakes, tea and coffee refresh

**5pm** - Datathon Submissions

**6pm** - Finger food & beverages served - opportunity to network.

**6.30pm**– Hackathon finalists teams pitch on stage

**7.30pm**- Awards presentation for Datathon & Hackathon.

**8pm** - Close

**Please note; submissions will close at 5pm, with winners announced at 7.30pm – feel free to enjoy the food, beverages & entertainment provided!**