

Ph.D. Assessment

Context

On a daily basis, we receive hundreds of thousands of text inputs from our users, some are more critical than others. Our approach so far has been to let someone handle this manually. The manual approach is not scalable as overall the performance of the individual in charge is that of 4K input per week which is not enough for our daily quote.

Why do we need research?

The input text that we are receiving is written in the Algerian Dialect. This dialect suffers from a lack of research and documentation which makes it difficult to rely on existing models to handle this. So given the amount of data that we have. It is our desire to make a breakthrough in the field of sentiment analysis with the Algerian Dialect by making contributions in the following areas

- Algerian Dialect Documentation and benchmark creation
- Algerian text transliteration
- Algerian text sentiment analysis and chat bot

The Task

Your objective for this assignment is not to work on the Algerian dialect immediately but rather to work on a language with similar attributes, which is Hindi. Hindi is also multilingual with a lot of words borrowed from English which makes it somehow easier to analyze than the Algerian dialect. Another aspect of the Indian language or at least the corpus that you would be working with is that no transliteration is needed, so all the text is already written using the Latin alphabet.

Your work would be based on the attached paper and you will be evaluated on

- Paper review and understanding carried in a small report or slides.
- Paper Implementation
 - Clear processing Pipeline

- Both models
- Comprehensive results with comments

Extra points will be awarded for

- Best practices
 - Coding
 - Comments
 - Project packaging
- Additional processing
- Data Visualization