**Characteristics/Features of a Tweet**

* Up to 280 characters (including Unicode)
* RT to indicate Retweet.
* #s to tag messages with index keywords or topics (no spaces or punctuation included)
* @s to mention another account/user.
* Can include links, which all begin with standard http:// notation.
* Some tweets are surrounded in quotations.

**Patterns**

No distinguishable patterns (word structure). @ and # appear in any format as well. Only noticeable thing is the addition of links at the end, besides some few exceptions.

**Filtering**

Since the point of the model is to be able to recognize and approximate what language the tweet is made in, some filtering is required. Anything followed by a @ and # should be filtered out. @ is used to mention another user. However, this serves as no indication on what kind of language the tweet is written is. Even if the username was a person’s name, assuming a language based on a person’s name is incorrect profiling as a person can be multilingual.

Similarly, # is simply used to included a searchable keyword or topic. It is possible for multiple languages to reference a common event. Going through some tweets, I can easily some tweets that are written in Spanish but include the following hashtags: #running, #me, #friends, #carnivals; all of which are English words.

Numbers and website links should also be avoided as they serve no indications of what language is currently being spoken.