Rosette Delivers Accurate and Fast Language Detection for Social Media Monitoring

Hootsuite is the most widely used social media management platform, trusted by more than 18 million people and employees at 80% of the Fortune 1000. Clients of Hootsuite are empowered to fully leverage social media, and meet customers where they congregate. Its platform builds strong internal cultures, uncovers emotionally rich consumer insights, and unifies the customer experience. A key piece of social data analysis is segmentation by location, gender, and language.

The Challenge

Segmenting social media messages by language is important for brands that only do business in particular regions. If, say, a brand doesn't do business in Romania, it won't be interested in tweets in Romanian.

Some global brands differentiate themselves by region. For example, international burger chains have different offerings for different regions. In India — which has a large population of vegetarians — the menu might include a spicy paneer and a salsa bean burger that are not offered elsewhere. There often is a dedicated team working on social media posts in each language, and their other local market knowledge is wasted if a message is misdirected.

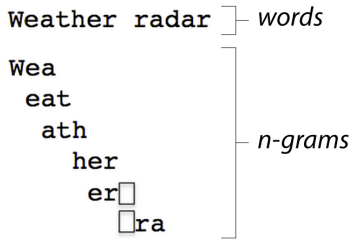
Hootsuite Insights and Analytics is the workhorse that enables the big data analysis of social media. Hootsuite Development Manager Mihai Caraman recalls that before Rosette Language Identifier was

integrated, they had a home-grown language detector that had to be refreshed every three to four years, as new words entered the vocabulary of each language.

"It was a challenge to keep pace with evolving expressions, so if we trained with a certain set of data, three years later there would be new expressions [that required an update]," Caraman said. "That was a considerable effort. So we decided to use a solution from people who are dedicated to it."

The Solution

Caraman and his team chose Rosette Language Identifier because it's fast and uses a statistical profile for each of the 55 languages it supports. Large samples of each language are broken into n-grams (overlapping sequences of n characters), and sorted by frequency.



In the case where n=3, the n-gram technique breaks up characters into overlapping trigrams.

The n-grams follow language-specific characteristics that do not change as the language evolves. For example, English words frequently use the ending "-ight," whereas that combination of letters would be very rare in French. In fact, new expressions in each language do not vary from the characteristics captured by the statistical profile. As a result, the profiles do not need updating.

Recent improvements to Rosette Language Identifier include a short string algorithm for greater accuracy on text of one to three words up to a sentence. This feature is useful for tweets, search queries, and photo captions. To achieve this greater accuracy on short texts, the algorithm looks beyond n-grams to script and word characteristics.

The Impact

With Rosette reliably tagging the language of social media posts, Hootsuite clients can also identify how many social media representatives they need in each language. This is vital because they must respond to their customers via Twitter, Facebook, LinkedIn, or whatever social media platform is used.

"In customer support you need to specialize in different languages, and hire people who speak that specific language," Caraman explained. "So when a brand comes to us, we can identify if their customer support is well staffed by identifying the number of people reaching out to them in different languages."

Most importantly, social media posts are being sorted by language to be directed to the right person for a response.

"Because language identification is done so well by Rosette, we can focus on other things. It's one thing we take for granted." Caraman said. "Compared to others, the precision of the detection is good, so we rarely revisit that decision [to use Rosette]."