MIT Technology Review Arabic Edition
http://techreviewpanarab.com/en/computing/analyzing-arabic-sentiment/

ARAB STORIES COMPUTING

Analyzing Arabic Sentiment

How deep learning could leapfrog Arabic sentiment analysis to the cutting edge.

By Benjamin Redd on October 26, 2015

It's happened to all of us. Our once-top-of-the-line smartphone starts to slow down, needs constant charging, or takes a fatal tumble. For many, that means it's time to scour Amazon, Reddit, and a dozen tech review sites to choose our next mobile companion. If we're lucky, after hours of work, we might find the right reviews to base our decision on. Not exactly an efficient—or scientific—approach.

But what if a computer—or network of computers—were able to find every review, every tweet, every webpage that mentions a specific smartphone model, and judge whether each mention was positive or negative, aggregating the results into an easy-to-digest summary of global opinion? Not only would it give us a complete picture, it would save hours of toil.

This is the burgeoning field of sentiment analysis, and it's big money—from corporations wanting to manage their brands to politicians looking for an edge in elections. At its heart, though, is the formidable task of analyzing human language in the real world and automatically figuring out whether it's positive, negative, or neutral. In English, where the most research has been conducted, even the best tools judge real tweets correctly just 65 to 71 percent of the time, according to a 2014 study. A lack of funding and attention has left research in Arabic in even worse straits.

But this doesn't necessarily mean Arabic sentiment analysis will lag behind forever, according to Wassim El Hajj, a professor of computer science at the American University of Beirut and member of the Opinion Mining in Arabic Project (OMA). "We're [almost] in the same boat," he contends—despite English

researchers' head start. Funded by a grant from the Qatar Foundation, OMA is looking not only to fill the gaps that impede current research, but also at technologies like deep learning that could leapfrog Arabic sentiment analysis to the cutting edge.

Arabic riddles

Arabic presents several language-specific difficulties for sentiment analysis. First, the dialects. Who tweets or comments in perfectly formed Modern Standard Arabic? Then there's the script—Arabic or Latin? If it's the latter, there's no universal standard for spelling. The OMA team is working on standard Arabic for the time being, but eventually wants to tackle both of these challenges.

But a more fundamental problem faces Arabic researchers: the language's rich morphology. Basically, Arabic words change form a lot—a three-letter root can parent dozens of (sometimes unrelated) words through vowel changes, prefixes, suffixes, and the like. So while English has different words for 'occupy', 'location', 'solve' and 'analysis', each of these concepts can be expressed by words derived from the h-l-l root in Arabic such as 'ihtilal', 'mahall', 'hall' and 'tahlil' respectively.

This has immediate consequences for one of the most popular sentiment analysis methods, the so-called 'bag of words' approach. The idea is that morphological changes don't matter a whole lot: 'great', 'greater', and 'greatest' are all close enough we don't need to know the shades of meaning expressed by their suffixes. Instead, we just need to be able to trace these words back to 'great' and know that this stem is positive. But without the affixes that relate words to one another, a sentence is reduced to a jumbled collection of words—a 'bag of words'. Combining the sentiments of the stems for each word in this bag, you can make a guess as to whether the overall sentence is positive or negative.

Clearly, this approach is too primitive for Arabic. But therein lies the opportunity: Arabic's complexity means researchers must look to the hottest area of study in natural language processing: deep learning. "We [already showed] that using deep learning is more promising than using the other traditional [methods]" in Arabic, says OMA researcher Ramy Baly, "Now we're investigating how to use [it]."

The basic idea is to create a program that can look at a raw dataset, learn the sentiment patterns from it, and then apply those patterns to classify new data. That's easier said than done. Baly is developing and testing different models for deep learning in Arabic sentiment analysis—there are a variety of approaches, and not all will be equally suited to the task.

But once an optimal model is identified, it ought to be able to handle Arabic's morphological richness relatively well. And not only that, because it learns itself, it may be able to be optimized to handle the other challenges of the language like dialectical differences. That is, if deep learning works—or works well enough—it could kill two birds with one stone.

When OMA's current funding ends in November 2016, Hajj says the team will have produced an open source tool allowing the public to visually explore opinions in Arabic around the world—something, he notes, that "doesn't exist yet in either [English or Arabic]." They've already released a sentiment mining app for Android and a fully downloadable sentiment lexicon for researchers, an upgraded version of which will be released in the coming months.

But work won't stop there, Hajj insists. The problems with natural language processing, especially in Arabic, are complex enough to keep researchers busy for years to come. Luckily, the benefits of such research should start rolling in much sooner. Maybe even in time for the next iPhone.