1-1-2018

Sentiment inSitu - Automated Sentiment Analysis in the Native Language

Candice L. Lanius
University of Alabama in Huntsville

Sentiment in Situ - Automated Sentiment Analysis in the Native Language

Faculty Mentor – Dr. Candice L. Lanius, Lecturer in Communication Arts, College of Arts, Humanities, and Social Sciences.
Office phone: (256) 824-2122.
Email: candice.lanius@uah.edu
Mail: Morton Hall 342, Dept. of Communication Arts
No previous participation in the RCEU program.

Project Summary– Language is complicated. Most speakers in the process of learning a second language will agree, yet technologists are now attempting to teach machines to understand multiple languages as well. While a direct translation can be automated easily, translating the context and emotions behind large bodies of textual data is complicated. Everyday, 3.7 billion people around the world access the internet, producing the largest experimental archive on human behavior ever created. This mass of social data is beyond the reach of traditional research methods in both scope and magnitude. Despite early successes with building reliable and scalable research systems, there are still gaps in the qualitative data infrastructure necessary to make big social data projects valuable for the knowledge they produce.

The RCEU student will contribute to this project on improving sentiment analysis by creating and statistically validating a sentiment dictionary in their native language. Sentiment analysis is the process of automatically identifying the mood or attitude someone holds about another entity. An early example is Bollen, Mao, and Zeng (2011) who used sentiment analysis on Twitter data to predict positive or negative trends in the stock market. The current gold standard in sentiment analysis—the *Linguistic Inquiry Word Count* (LIWC)— has weaknesses for certain applications. The LIWC is based on psychological research and provides a reliable and valid assessment of the emotional valence of texts written in English. However, because the LIWC is based on average use, it is not adept at assessing unique cultural contexts such as online communities. When faced with unique cultural contexts, current researchers adapt the dictionary in an ad hoc fashion without testing and validating these additions. Currently, researchers use machine translation to convert their data into English before processing it for sentiment. This additional step muddies the data and destroys the nuances of human expression. This RCEU project is part of our larger program to address sentiment analysis problems by providing robust, reliable, and validated sentiment dictionaries that offer complexity and native language context.

Example: On September 8, 2017 an 8.2 magnitude earthquake struck Mexico. Many responded on Twitter's trending #PrayForMexico. One tweet posted by musician Alex Hoyer illustrates the problems with automated translation.

Google translate shows this as "TeamHoyer my prayers are with all of you. I send a hug to all my family and Mexican brothers from Argentina." A sentiment analysis tool designed for English texts could easily mistake this for a personal, romantic response. The software application that this RCEU project contributes to will use the native language and therefore understand that "ustedes" is the formal you. If "ustedes" appears next to "un abrazo" (hug), that indicates patriotic feelings rather than romantic attachment. These minor shifts in meaning add up in the large scale analysis performed using machine learning tools, so the student's work will be valuable for improving big social data technologies.

Student Prerequisites – The student must be a native speaker of one of the target languages: Spanish, Arabic, or Hindi. Beyond this, we seek students with an interest in researching social media or sentiment using big data, machine learning, or other technological means. Their major may be computer science, communication arts, psychology, etc.

Student Duties– The RCEU student will create a machine actionable dictionary for sentiment in their native language which will be paired with a software application already in development. In the final stages of the summer project, we will collaborate to run experiments to assess the efficacy and accuracy of their sentiment analysis dictionary compared to the old process of using machine translation to convert all text to English before running a sentiment analysis. The student will learn about research design and statistical analysis as they perform experiments involving large scale social media datasets related to cybersecurity issues. The student should have a strong ability to work independently and conduct primary and secondary research.

Deliverables include: 1) the sentiment dictionary, 2) a journal article on validation of the sentiment dictionary, and 3) a website housing the open source dictionary for free download.

Mentor Supervision and Interaction – The student will work in the eValuation and User Experience Lab (VUELab) alongside Dr. Lanius, Dr. Joy Robinson, Dr. Ryan Weber, and other RCEU students as they complete ongoing research projects. There will be a formal weekly meeting for progress reports and planning for the coming week and writing sessions once the initial dictionary has been completed.