Alexandra Gamez

CSCE 489

Project Proposal

**Multilingual Emoji Prediction**

Twitter is a place where people share ideas, give their opinions on different topics, discuss popular current events, and find out about the what’s happening in the world. It is a place rich with information about what the world is talking about. This is a place where one could find great insight, analyze behaviors, and patterns. Knowing what people care about and being able to predict that is valuable. Corporations pay big money for this type of information and it powers things like Netflix’s suggestion engine.

This year SemEval has suggested a problem of great importance. Given a tweet in a specific language, what emoji is it most likely to use? The task is simple. Predict the emoji of a tweet. The data set contain tweets in English and in Spanish which use one of the 20 most popular emojis in each language. The problem is split in two parts. First, figure out the language. Second, given the language and the tweet predict the emoji.

This is a classification problem. It is also a prediction problem. Scope of the project deals with something silly like emoji prediction but the general approach to solving this problem could easily be applied to something non-trivial. It is, after all, a classification problem and this could easily be applied to something important such as a sign language translator which is something that could help a lot of people.

The problem will be approached in the following manner. There are many predictive algorithms out there. Hidden Markovs, Naïve Bayes, etc. It’s hard to pick which one to choose. The problem seems simple enough that the goal of this project is to not only solve the emoji prediction problem but also to implement several of these algorithms and compare their performance.

In the end, I will have used my knowledge of algorithms learned in the class and have increased my understanding by applying my knowledge on fun and interesting project.