

ZACHARY HAROIAN

MOST @ MENTIONS

NATURAL LANGUAGE PROCESSING

Natural Language Processing, usually shortened as NLP, is a branch of artificial intelligence that deals with the interaction between computers and humans using the natural language. The ultimate objective of NLP is to read, decipher, understand, and make sense of the human languages. One of the more relevant applications of NLP is to analyze large amounts of text to identify whether the text is positive or negative. This allows for a massive amount of data and opinions to be extracted without manual parsing.

Because tweets are inherently short and lack formal grammatical structure, full of acronyms, slang, emojis, and names, it makes it hard to formalize a structure to correctly parse them. The method implemented in this research was the Bag-of-words model. This model disregards sentence structure in favor of word frequency, relying on word selection to determine what the intended meaning is. This method is typically used for smaller sets of data, where context may be incomplete, such as movie reviews or email filtering. While common words like “the”, “a”, “to” are usually the highest frequency, they are not the best representation of the text and certain stopwords are chosen to reduce computation time and eliminate words that do not directly contribute to the meaning of the sentence.

Sentiment Analysis is contextual mining of text which identifies and extracts subjective information in material, and helps to understand what the general sentiment is for a given topic. This categorizes text based on a value given to each word, positive or negative in polarity, or 0 to 1 in subjectivity. This helps identify what the person is communicating, without including the content that they were discussing. This is used with Bag of Words to classify tweets about certain candidates.

CANDIDATE MENTIONS

There were multiple problems we ran into on the way. The biggest one is that users will not use hashtags consistently. Some users add hashtags to the end of their tweet as a way to contribute to a larger conversation, but not directly tied to a sentence, such as “I love going down to Walden Pond!” #fishing.” Other users replace words within the sentences with hashtags, such as “I love #fishing at Walden Pond!” Others still would respond to prompts or questions posed by a hashtag as part of a game, such as “ThingsMoreTrustedThanHillary” where the user’s response is only a sentence fragment without the hashtag, such as “#ThingsMoreTrustedThanHillary the water in Flint.”

This poses a problem, as the hashtags cannot and should not be removed, as they are usually an integral part of the message being delivered. But, a machine cannot read them easily. There is also a lack of consistency within how people use them. Some people use CamelCase, which replaces spaces with capitalizing each word to make it more readable. This is the simplest to deal with – as long as a lowercase letter is followed by an upper case letter, it can be split apart into words. Some people use all lower/uppercase, which is much harder to interpret. If we were performing this research again, it would probably be worth investing some time into a greedy algorithm to match words in the hashtag to split them up. There is a danger in doing this, of course, as hashtags don't always contain real/full words.

Overall, this selection of natural language processing did not produce very insightful data. I believe the reason for this is that the Twitter data is especially hard for the NLP to correctly analyze because of conflicting clauses. Many of the tweets will compliment one candidate, and then put down the other, resulting in a net zero polarity, and both candidate are mentioned. Without a more advanced method of NLP, more accurate information is not possible. The other problem is that the tweets contain such a small amount of text that one positive or negative word could disproportionately outweigh other words, or the tone could be sarcastic and actually mean the opposite of the dictionary definition.

WORDS USED WITH EACH CANDIDATE

The figure displays eight scatter plots arranged in a 4x2 grid, comparing subjectivity (y-axis) against polarity (x-axis) for tweets from Hillary Clinton (left column) and Donald Trump (right column). The plots are categorized by account type: HashTagGamer, NewsFeed, RightTroll, LeftTroll, and NonEnglish. Each plot shows a distribution of points colored by account category. The legend indicates: HashTagGamer (blue), NewsFeed (orange), RightTroll (green), LeftTroll (red), and NonEnglish (purple). The y-axis (subjectivity) ranges from 0.0 to 1.0, and the x-axis (polarity) ranges from -1.0 to 1.0. The plots show varying degrees of clustering and separation between the categories, with some categories like LeftTroll showing higher subjectivity and others like HashTagGamer showing more varied polarity.

[illegible]

Debate 1
Before: 9/26/16 1:00 a.m. - 9/27/16 00:59 a.m. UTC
During: 9/27/16 1:00 a.m. - 2:29 a.m. UTC
After: 9/27/16 2:30 a.m. - 9/28/16 2:30 a.m. UTC

Debate 2
Before: 10/9/16 1:00 a.m. - 10/10/16 00:59 a.m. UTC
During: 10/10/16 1:00 a.m. - 2:29 a.m. UTC
After: 10/10/16 2:30 a.m. - 10/11/16 2:30 a.m. UTC

Debate 3
Before: 10/18/16 1:00 a.m. - 10/19/16 00:59 a.m. UTC
During: 10/19/16 1:00 a.m. - 2:29 a.m. UTC
After: 10/19/16 2:30 a.m. - 10/20/19 2:30 a.m. UTC

