**Project Short Description**

**SuFooBar**

Michael Arnold and Yi Li

**Problem Description**

Twitter has been a great social networking tool and has shaped our society's political landscape. Sometimes, it can be difficult to understand the majority's feelings concerning certain candidates. Our program will determine a person's support for a political candidate, or whether they have neutral feelings, based on the content of their tweets.

We will be focusing on Donald Trump and Hillary Clinton, and using tweets that are tagged with debate hashtags. Using tweets from the first two debates to train our algorithm, we will then use the last debate as the test cases. Because this election has seen a surge in tweets concerning the candidates, we will only use a subset of the total tweets for each debate.

**Proposed Solution and Implementation Details**

**(1). Baseline system**

For the randomly sampled tweets (~300) corpus, we will use "Donald Trump" and "Hillary Clinton" as the keywords to perform simple search. And the total counts for each candidate's name within each tweet will be used as the metrics for sentiment determination. If the counts are the same, the tweet is neutral.

Example: In a specific tweet, if the total count for "Donald Trump" and its immediate variants (e.g. DonaldTrump) is 5 and "Hillary Clinton" and its immediate variations appear 3 times, we will label this tweet as "pro-Republican".

**(2).Improvement strategy using lexical features**

Approach 1: It is conceivable that tweets supportive of a specific candidate will likely use distinct hashtags/word types, such as #VoteHillary2016, #USAforTrump2016 and #votetrump. Thus, we will generate a list of such word types, and perform a RE (regular expression)-based search algorithm to locate such instances.

Example: We will perform search using keywords such as #VoteHillary2016 and #MakeAmericaGreatAgain.

Approach 2: The tweets supporting a specific candidate are likely to include negative information of his/her competitors. We can use NLP tokenization (using the NLTK package) to find the specified keywords or negative/positive verbs used in the tweet. This will help us determine each tweet’s general political inclination.

Example: Tweets involving tokens such as Benghazi are more likely to be anti-Hillary Clinton, while tokens such as harassment are more likely to be anti-Donald Trump.

**(3). Improvement strategy using syntactic features**

Approach 1: In supportive tweets of a specific candidate, it is likely that his/her name will be closely associated with positive/negative information, such as "support Hillary Clinton". Thus, we could generate the syntactic parsing structures for all sentences in a specific tweet, and extract the minimal constituents containing the candidate's name. This will be done using the NLTK library for Python. We will then analyze the other tokens within these constituents to determine its inclination.

Example: After syntactic parsing analysis, we locate a specific VP (verb phrase) "vote Donald Trump", and this could be determined as pro-Republican. Similarly, another NP (noun phrase) "Hillary Clinton scandal" could be used to determine it as anti-Hillary Clinton.

Approach 2: In a specific sentence, it is likely to associate the candidate's name with positive or negative adjectives. Thus, we could perform POS taggings for all the sentences and for the ones including the candidates' names and adjectives (JJs), we could use the positive/negative adjectives as the metrics to determine its sentiment.

Example: If a sentence contains "greedy...Hillary Clinton", it is likely this tweet is anti-Hillary Clinton. Similarly, a sentence such as "shameless...Trump" most likely is anti-Donald Trump.

**(4). Improvement strategy using semantic features**

Approach 1: In a specific sentence, if the Beneficiary refers to certain negatively perceived actors/groups, such as Wall Street and top 1%, this tweet is more likely to be against the Agent of this sentence. Therefore, we could perform thematic relation analysis and determine the agents and beneficiary for the sentiment analysis.

Example: A specific sentence contains "Donald Trump (Agent) proposed the tax reform plan...for the top 1% (Beneficiary)", and this thematic relation could be used to determine that this tweet is anti-Donald Trump.

Approach 2: It is notable that each candidate has his/her own core constituents, whose concerns are the focus of the candidate's messengers. Thus, the tweets supporting a specific candidate are more likely to include Agents of such specific constituents.

Example: Hillary Clinton's campaign noticeably caters for groups such as women, immigrants or LGBT community. In contrast, Donald Trump often mentions groups such as veterans, law enforcement agents, or (negatively) mexicans. Thus, in a specific sentence, if the Agent falls into the first group, this tweet is more likely to be pro-Hillary Clinton. Similarly, if the Agent refers to groups such as illegal immigrants or veterans, it's more likely to be pro-Donald Trump.

**Programming tools:**

(1). NLTK: NLTK uses many other tools, wrapped up into one python package. We will use the NLTK package to do parsing, tagging, tokenization, and semantic analysis.

**Architectural diagram:**

