Sentiment Analysis of Political Twitter Data – A Literature Review

**Introduction**

Since the advent of social media 15 years ago, we have witnessed an explosion in the amount of opinion data available digitally between traditional blogs, sites like Facebook, and microblogs like Twitter. There is a lot of noise to go along with all this data but the cost of determining which is which using humans is unimaginably high. For example, one estimate put acquiring all tweets for any given day around $64,000, which would waste almost $60k of that money because at most 10% of twitter data is relevant to any given event or object of study (Buntain, McGrath, & Behlendorf, 2017) Twitter sentiment analysis is still a young field but much progress has been made and researchers have used a variety of methods to attack the black hole that is the Twittersphere. Political scientists have used Twitter data to analyze everything from election results to candidate personalities. One of the unique qualities of social media data is that it can give group and individual level emotional responses to various events without extensive fieldwork or expensive surveys (McGrath, et.al., 2017). Leveraging such data enables policy makers to learn what issues are prevalent among their constituents by listening rather than asking. Yet, policy makers and citizens have not yet found useful ways of efficiently making use of this “mass conversation” (Osimo & Mureddu, 2012).

Current approaches used in the field range from dictionary-based methods such as LIWC, deep learning methods and Convolutional Neural Networks and everything in between, as well as brand new method currently in use at START called SentimentIt. Pioneered by Jacob Montgomery and David Carlson, SentimentIt uses expert coders and Mechanical Turkers to make pairwise, positivity (which of these tweets is more positive/negative) comparisons of each tweet in the relevant corpus to build a distribution that can be modeled by a machine learning algorithm. The goal of this paper will be to review these various methods, including their advantages and disadvantages, report on current developments at START INSPIRE, and finally relate the topic to the issue of Meaning in NLP.

**Dictionary-based Methods**

The first of the two main categories of Twitter Sentiment Analysis is the Dictionary-based method. One way these dictionaries are built is that each potential word is assigned a sentiment score. When you input a tweet into the program, each word receives a sentiment score and the average sentiment score across words determines the overall sentiment of the tweet. Alternatively, a popular method known as Linguistic Inquiry and Word Count or LIWC, utilizes several affective categories like “anger” or “greed” each of which have a pre-determined list of words. When you input a tweet to LIWC, you select which categories you want to measure and then LIWC counts how many words in the tweet belong to the categories of interest. The dimensions LIWC has created, and the words put in each category have been shown through various studies to capture important psychological properties such as status, social relationships, group processes, and emotionality (Tausczik & Pennebaker, 2010).

LIWC has been by far the most widely used Dictionary-based method in the field since its inception. Recently, the program was used to validate detailed psychological profiles of candidates from the 2008 US Presidential election using transcripts from their speeches and other sources of linguistics data (Kangas, 2014). But it has been used for than just whose language elicits more cognitive complexity or honesty or seeing who’s angry with what or at whom. Tumasjan et. al. (2011) sought to use LIWC in an analysis of Tweets surrounding the 2009 German Presidential Election to 4 topics:

1. Political deliberation through Twitter
2. Political preferences of the electorate
3. Party preference in individual accounts
4. Vote share prediction

They stated that political deliberation is common but most of it is carried out by several heavy users. They utilized 12 LIWC categories to build an ideological profile of voters based on their tweets as well as the candidates and political parties. The used a Euclidean distance measure to compare the LIWC data with existing information from the press and analysts. Voters’ twitter sentiment tended to overlap with their preferred party or candidate. The frequency and share of tweets among German political parties accurately predicted the outcome of the election, meaning the party that represented the largest proportion of Twitter traffic won.

Jurek et. al. (2015) constructed a lexicon-based sentiment analysis system with goal of analyzing Twitter content in real time. The sentiment scores of their vocabulary was based on SentiWordNet, which assigns a value to each word between -100 and 100. For words that have the potential to be both positive and negative, such as “lazy” in “lazy Sunday”, they calculated the probability that each word would appear in a sentence whose overall sentiment positive or negative. Their system also considered Negation, Intensifiers such as “quite” and “very”, in addition to utilizing a combining function that gives each sentence an overall sentiment score between -100 and 100 as opposed to just a positive or negative label. The combining function itself is represented as “the product of the average sentiment of the sentence and a coefficient that is based on the number of positive and negative words,” so that the program captures subtle differences in sentiment. This method achieved an accuracy of 77% compared to 69% for a baseline lexicon-based method that did not take negativity, intensity, or mixed sentiment into consideration. The twitter data used for testing was collected from the Stanford corpus.

***Advantages and Disadvantages***

Dictionary-based methods give the researcher and or analyst a lot of control over what the model learns about sentiment and therefore how it will classify the data. As we saw in the previous instances of this lexicon/dictionary-based method, dictionary builders can easily build functions in that look at important linguistic phenomenon like negativity and intensity. Dictionaries like LIWC are also very easy to implement. Any research can gather the necessary tweets in a CSV or text file and run them through the LIWC API. The main issue facing these methods is that they are costly and for the purposes of measuring sentiment in real-time may not be so practical. Language is a constantly evolving medium and the opinions that people express on Twitter surrounding elections and politics will vary by socio-political context. It is nearly impossible to build one dictionary that will provide adequate opinion analysis across elections in South Africa, Morocco, and the US.

**Deep Learning and Neural Networks**

Neural Networks (NN) are machine learning models that mimic the way neuroscientists believe the human brain functions. These networks require large amounts of input which then pass through interconnected nodes and arrive at an output layer. Deep Learning in the context of a Neural Network, simply refers to a type of NN that has multiple hidden layers between the input and output. These hidden layers learn features that aid the model in finding the correct output given the training and input data. I will first discuss different types of Neural Networks that have been designed for sentiment analysis as listed in Rojas‐Barahona (2016). These include Recurrent Neural Networks, Long Short-Term Memory models, Convolutional Neural Networks, and Recursive Neural Networks. Then, I will summarize previous studies applying these NNs to sentiment analysis tasks specifically using Twitter data.

A Recurrent Neural Network is one of the most basic NNs beyond a simple feed-forward network where the input layer connects to the output layer by activating all nodes in the hidden layer that are connected to it. An RNN adds a short-term memory component to the network structure allowing it to access information from the previous time step. In the case of analyzing sentences, it would consider the previous word. Long Short-Term Memory models are essentially RNNs that take more of the previous context into account when running natural language. The LSTM’s defining characteristic is its memory cell state that consists of 3 main gates: an input gate, an output gate, and forget gate. The functions inside of this memory cell examine the entirety of each sentence or piece of language that has been passed through the network at any given time point to determine if and what information should be remembered and stored for later usage or forgotten.

A Convolutional Neural Network, originally developed for use in computer vision, also resembles a Feed-Forward Network except each hidden layer is a set of features from a portion of the input(s). Each hidden layer that shares similar features share weights and a pooling layer determines the average or minima and maxima of the features in each layer to extract the most important and distinctive aspects of each feature.

A Recursive Neural Network, in the context of NLP, attempts to reconstruct the structure of the sentence by combining words into constituents like “my” and “dog” into the NP “my dog”. More specifically, Recursive Autoencoders (RAEs) create parent nodes for each possible binary combination of words in a sentence from left to right. The combinations with the lowest reconstruction error (distance between correct tree structure from training data and the realized tree structures) are determined to be correct representations. These methods have been made more syntactically relevant in papers such as Socher et. al. (2013), where Constituency trees guided the construction of the recursive tree structures.

At Sem Eval 2013, manually annotated and lexicon-based sentiment analysis systems faced off against a CNN trained using sentiment-specific word embeddings and other NN systems in a binary sentiment classification task of twitter data. These sentiment specific word embeddings consisted of sentiment annotated tweets, which were based on emoticons. Both the hand-annotated model (NRC) and the CNN - SSWE model classified about 84% of the tweets correctly (Nakov, et. al., 2013). Additionally, CNNs also perform the best or near the top on tasks involving determining the correct syntactic structure and sentiment of sentences (Sentiment Treebank) and determining the sentiment of movie reviews.

***Advantages and Disadvantages***

Neural Networks are, in my opinion, the best currently available approach to classification tasks, even sentiment analysis. They have an unbelievable amount of flexibility in terms of what they can take as input. Their structure can be modified depending on the needs or preferences of the individual project or assignment whether it be to build a syntactic structure of a tweet or learn the qualities of a picture that make it happy rather than sad. Perhaps most importantly, they are faster and cheaper than dictionary based methods because they require much less human input (sometimes none) and they can apply across socio-political domains on Twitter without requiring a huge new set of annotations (occasionally no annotations at all). However, to this day, exactly what Neural Network is learning inside of its hidden layers is not clear. This presents a problem for political scientists who want to be very specific about what is considered positive, negative, or neutral regarding tweets from different countries. Not only is sentiment relative across these countries but just because a tweet contains negative information does not mean the tweet itself is negative as is the case with news outlets reporting on bombings and other event reportings.

**Everything else**

The current section reviews two papers that do not fall into either of the previous categories. One paper discusses using CNNs to measure visual sentiment and the other paper compares the effectiveness of traditional opinion polls versus social media in measuring public opinion. Oliviera et. al. (2017) used sentiment analysis systems to measure the number of positive and negative tweets directed at candidates during the second round of the Brazilian Presidential Election in 2014. Researchers compared the twitter data and surveys of voting intentions to actual poll numbers and found that both correlated reasonably well with what happened in the election. Both the surveys and the twitter data were within 1 to 8 percent of each other. You et. al. (2015) uses a Progressively-trained Convolutional Neural Network (PCNN) to learn the visual sentiment of twitter images and the authors evaluated the PCNN against the sentiment ratings of Mechanical Turkers. The models the PCNN was trained against do not consider as many features and do not employ the unique training method the authors created for this paper. The Progressive CNN used in this paper was trained more on images that had stronger, more distinct labels to weed out noisy training data.

**Current efforts at START and Future Directions**

The goal of the START INSPIRE project is to measure political grievance in real-time to predict potential unrest. Social media is the medium by which we have decided to measure said grievance. Successful Twitter sentiment analysis methods have involved either dictionaries containing sentiment-annotated words, phrases, and or sentences, or Neural Networks. Dictionaries give you lots of control and customization of sentiment but Neural Networks can handle more data and are faster and cheaper. Both seem to suffer from issues of generalization. Any system you make is not going to apply across socio-political lines. The solution then seems to be a mixture of both. START has begun to employ an approach pioneered by Montgomery & Carlson (2016) called SentimentIt. This platform utilizes pairwise comparisons made by trained coders to build a distribution of tweets in a corpus. The tweets in this distribution could then be normalized into sentiment scores on a scale and then be used as training data for a Convolutional Neural Network, which has been shown to capture important syntactic features by outperforming LSTMs on SentiTreeBank in addition to incorporating sentiment word embeddings successfully in previous Twitter sentiment analysis tasks. This framework allows the researchers to train human coders on what information to consider depending on the country of interest to begin. It requires minimal human labor but maximizes the value of that labor. Finally, it takes advantage of the monetary and temporal efficiency of neural networks in learning sentiment classification while also utilizing a method (CNNs) that has performed well on tasks requiring solid knowledge of syntactic structure and complex semantic operations such as negativity and intensity.

**References**

Buntain, C., & McGrath, E., & Behlendorf, B. (2017). Information Retrieval from Microblog Data Resellers for Social Science Research. *In Proceedings of ACM Knowledge Discovery and Data Mining, Halifax, NS, CA, August 2017 (KDD’17).*

Jurek, A., Mulvenna, M. D., & Bi, Y. (2015). Improved lexicon-based sentiment analysis for social media analytics. *Security Informatics*, *4*(1), 9.

Kangas, S. E. (2014). What can software tell us about political candidates?: A critical analysis of a computerized method for political discourse. *Journal of Language and Politics*, *13*(1), 77-97.

Montgomery, J. M., & Carlson, D. (2016). Human computation scaling for measuring meaningful latent traits in political texts.

McGrath, E., & Donnay, K., & Dunford, E., & Backer, David., & Cunningham, D. (2017). Explaining electoral violence in Sub-Saharan Africa using grievance measures derived from social media. *Paper presented at ISA Conference, Baltimore, MD.*

Nakov, P., Rosenthal, S., Kiritchenko, S., Mohammad, S. M., Kozareva, Z., Ritter, A., ... & Zhu, X. (2016). Developing a successful SemEval task in sentiment analysis of Twitter and other social media texts. *Language Resources and Evaluation*, *50*(1), 35-65.

Oliveira, D. J. S., Bermejo, P. H. D. S., & dos Santos, P. A. (2017). Can social media reveal the preferences of voters? A comparison between sentiment analysis and traditional opinion polls. *Journal of Information Technology & Politics*, *14*(1), 34-45.

Osimo, D., & Mureddu, F. (2012). Research challenge on opinion mining and sentiment analysis. *Universite de Paris-Sud, Laboratoire LIMSI-CNRS, Bâtiment*, *508*.

Rojas‐Barahona, L. M. (2016). Deep learning for sentiment analysis. *Language and Linguistics Compass*, *10*(12), 701-719.

Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013, October). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)* (Vol. 1631, p. 1642).

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology*, *29*(1), 24-54.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2011). Election forecasts with Twitter: How 140 characters reflect the political landscape. *Social science computer review*, *29*(4), 402-418.

You, Q., Luo, J., Jin, H., & Yang, J. (2015). Robust image sentiment analysis using progressively trained and domain transferred deep networks. *arXiv preprint arXiv:1509.06041*.