Entity-Based Sentiment Analysis on Tweets

Anderson Monken, Nicole Yoder
Georgetown University

Abstract

Recent work on sentiment detection of tweets and specific entities within tweets has utilized lexicons and a two-dimensional method, called SentiCircles. We recreate and add a new lexicon to the SentiCircles method and apply it to two datasets. The first is labeled at the entity-level and used to test our recreation and the second consists of tweets from a recent event: President Biden's address to the joint session of Congress on April 28, 2021. Our method has comparable accuracy to the original work, and we found that most of the politicians we analyzed were viewed neutrally.

1 Introduction

Since its start in 2006, Twitter has exponentially increased in popularity and become a key platform for microblogging and political discussions in particular. Twitter provides politicians a fast and free platform to communicate their messages, and citizens can respond to those messages and other public interactions involving politicians in real-time. An example of such political dialogue is the over 70,000 tweets written about President Biden's address to a joint session of Congress on April 28, 2021. We aim to analyze how various politicians were viewed during the address by detecting the sentiment (positive, negative, or neutral) of those entities using an adapted two-dimensional method from Saif et al. (2014, 2016) called SentiCircles.

We start by describing our evaluation and experimental datasets in Section 2. Then in Section 3, we give a background of sentiment detection of Twitter data and the SentiCircles method in particular. Section 4 outlines our preprocessing and recreated SentiCircles methodology. Our results are presented in Section 5 and then discussed in Section 6.

2 Datasets

We use the SentiCircles method on two datasets: an evaluation dataset and an experimental dataset. The evaluation dataset is the STS-Gold Entity data provided by Saif et al. (2013). The STS-gold dataset was acquired from GitHub where a researcher saved the original dataset since the author of the dataset did not keep an official version of it available.

STS-Gold is a subset of the Stanford Twitter Sentiment Corpus. It contains 405 tweets referring to 58 named entities, which were labeled by a group of Ph.D. students as positive, negative, neutral, mixed, or other. The most prevalent choice of the Ph.D. student labels is used to determine the positive, negative, and neutral labels. Three entities were removed from the dataset due to lack of data availability given our preprocessing method: Pride and Prejudice, lung cancer, and Trader Joe's. Table 1 displays the 55 remaining entities.

The experimental dataset consists of 70,412 tweets regarding President Biden's address to the joint session of Congress on April 28, 2021, and it will be referred to as the JS dataset. The tweets collected used at least one of the search phrases in Table 2 and were submitted between 12 pm (US Eastern) on April 28 and 12 pm on April 29. The tweet data contains information about the user, tweet time, and tweet text. The text can include references to topics using hashtags (#) or call-outs to specific users using the at-symbol (@). The entities chosen for the JS dataset are shown in Table 1.

3 Related Work

Understanding the sentiment of user-generated content is an essential part of viral conversations,

STS Gold Entities				JS Entities
Amy Adams	The Beatles	Brazil	Cancer	Joe Biden
Cavs	Coco Martin	Dallas	Dominique Wilkins	Joint Session
Emily	England	Facebook	Fever	Kamala Harris
Flu	Ginormica	Headache	iPhone	Mitch McConnell
iPod	Jasmine Tea	Jonas Brothers	Kardashian	Nancy Pelosi
Katy Perry	Kobe	Lakers	LeBron James	Ted Cruz
London	Macheist	Mary Jane	McDonalds	Tim Scott
Miley Cyrus	Obama	Omaha	Oprah	
Pancreatic Cancer	Pandora	Paulo	PSP	
Riyadh	Sam	Seattle	Sil	
Skin Cancer	Starbucks	Sydney Aquarium	Sydney	
Taylor Swift	Twitter	University	Usher	
Vegas	Verizon	Wii	Xbox	
YouTube	yyy			

Table 1: Entities analyzed in STS Gold and JS datasets.

support or opposition to a politician or country, or how groups feel about specific products, services, or companies. A vast body of research has developed using computational linguistics to answer the question of quantifying the positive and negative features of text. Traditional methods of conducting sentiment analysis involve simple lexicon-based counting methods. By identifying the number of positive and negative words in a text using a pre-existing dictionary, researchers like Taboada et. al (2011) have been able to label texts as positive or negative.

While lexicons can be a useful tool in determining sentiment, they have limitations when used with user-generated content, since the language used on Twitter, for instance, changes quickly and can be specific to the particular topics being discussed. Zhang et al. (2011) combined lexicon-based methods with learning-based methods to improve upon the weaknesses of both. Chi-square tests were used to add opinion indicators that were not in the lexicon (e.g. "looove") if the terms were associated with positive or negative terms in the lexicon. The lexicon and additional opinion indicators were then used to label the training data for the supervised classifiers. This avoids the intensive task of human annotation of training data.

Search Phrases (Case Insensitive)
#jointaddress
#jointsession
#presidentialaddress
#bidenaddress
sotu
"state of the union"
"joint address"

Table 2: Search phrases used to collect Joint Session tweets.

Da Silva et. al (2014) used lexicon features in addition to bag-of-words or feature hashing with ensembles of classifiers. They found that using an ensemble of classifiers (multinomial naive Bayes, SVM, random forest, and logistic regression) improved performance compared to the models used individually, and that BoW was less computationally efficient but more accurate than feature hashing. While supervised methods are effective at detecting sentiment, they often suffer from a lack of training data. We became interested in applying a supervised method of determining sentiment that does not require large amounts of training data to tweak parameters. SentiCircles met these criteria since the Joint Session dataset we collected is unlabeled and there is only limited evaluation data available at the entity-level.

3.1 SentiCircles

Saif et all. (2014, 2016) tweak the use of lexicon-based sentiment by modifying the polarity and strength of the sentiment based on a TF-IDF-like indicator and a sentiment lexicon to construct a polar coordinate setup for sentiment called SentiCircles. For a particular entity of interest, like Facebook or Taylor Swift, all the words in tweets with those entities are considered context words that affect the sentiment of the entity. By considering the entire corpus of tweets, the authors construct a SentiCircle by capturing all of the context words for a particular entity m. The radius, r_i, for a context word, c_i, is constructed using a term degree of correlation (TDOC) as shown in equation

1, where f(c_i, m) is a count of tweets that the context term with entity m, N represents the total number of terms and N_{ci} represents the total number of terms when c_i is in the tweet.

$$TDOC(m, c_i) = f(c_i, m) * APTARANORMAL \log\left(\frac{N}{N_{C_i}}\right)$$

Equation 1

The radius is normalized to between [0, 1] for all points to fit within a unit circle.

The angle, θ_i, represents the sentiment of context word c_i from a sentiment lexicon and could be based on part of speech tagging depending on the lexicon. Saif et. al choose SentiWordNet and two other sentiment lexicons for their analysis. The sentiment must be in the range [0, 1] because it is then multiplied by π so the strongest sentiment words have an angle close to π and objective words close to an angle of 0. The entire range of the angle falls in [0, 2π]. Context words with an angle close to zero fall into the neutral region, see Figure 1. These words do not contribute significantly to sentiment in either direction. Saif et. al choose to keep all words including stop words because a tweet without few strong sentiment words is less likely to be polarized. They also conduct negation processing.

After collecting all the points to construct a SentiCircle, the centroid of the context points is found to determine where the median context for entity m occurs. If the SentiMedian vector is within the neutral region, then the entity is given a neutral sentiment. Otherwise, the y-axis sign determines a positive or negative sentiment label. A hyperparameter, λ, is used to determine the y-axis bounds of the neutral region. The authors found success in their method with an F1 score of 80.03% when averaging success at identifying polarized (positive or negative) entities as well as identifying positive entities.

4 Methodology

We recreate the SentiCircles method in Python as described by Saif et al. (2014, 2016) and apply it after preprocessing the tweets to determine the overall sentiment for each entity. The additional lexicon VADER is applied to the SentiCircles framework to test performance.

4.1 Preprocessing

When the JS tweets were initially collected, retweets were not included and the STS Gold dataset had no retweets as well. Every emoji is replaced with text using the demoji Python library. For instance, '🦉' becomes 'owl'. Then URLs are removed, as they do not indicate sentiment. Similarly, random usernames like '@user1234' are removed, but usernames used by the entities in question (e.g. '@POTUS') are kept. The nltk Python library is used for both tokenization and part of speech tagging. Since the POS tags are inconsistent with the entity usernames, those had their tags changed to 'NNP'. All text is then set to lower case to match surface forms when the first letter of the word is capitalized.

Since the STS-Gold and JS datasets both had entities that could appear as either single tokens or bigrams, a token combiner procedure is set up to combine tokens like 'joe' and 'biden' into 'joe_biden'. After combining bigrams, various surface forms referring to an entity are unified. For example, Joe Biden has 11 different forms that are combined so that referring to 'President Biden', 'potus', or '#JoeBiden' all become 'joe_biden'.

4.2 SentiCircles*

Our implementation of SentiCircles closely follows Saif et. al (2014, 2016) as well as previously available public code methods. We use both the SentiWordNet lexicon as well as introduce the VADER lexicon (Valence Aware Dictionary and sEntiment Reasoner) as an alternative for determining the polarity of sentiment. Each entity has a SentiCircle calculated separately by determining the TDOC from the set of tweets and

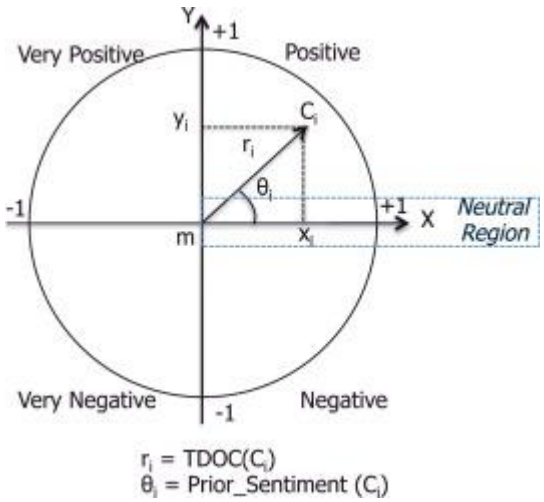

Figure 1: SentiCircles diagram from Saif et. al (2016)

an angle from a sentiment lexicon. It is not uncommon for a context word to be captured in the TDOC but not have an associated sentiment value in the lexicon. In these cases, the context word is dropped from the set of points going into the SentiCircle. The radii of the remaining points are normalized using a min-max scaler so the largest is 1 and the smallest is 0. These polar coordinates are converted to cartesian coordinates, and the SentiMedian vector is determined. The categorical sentiment (positive/negative/neutral) is applied to the SentiMedian vector with $\lambda = 0.05$ for the SentiWordNet lexicon, the same as Saif et. al (2014, 2016), and $\lambda = 0.0001$ for the VADER lexicon. The categorical sentiment values for each entity in the STS-Gold dataset are compared to the ground truth using three different binary classification tests:

1. Neutral (up), positive/negative (down)
2. Positive (up), neutral/negative (down)
3. Negative (up), neutral/positive (down)

SentiCircle results are also plotted on a unit circle to visualize the strongest context terms for specific entities.

5 Results

We find that the recreated SentiCircles method performs similar to the original in Saif et. al (2014, 2016). The results in Table 3 show the performance on the evaluation dataset for both VADER and SentiWordNet using the three binary classification schemes listed in section 4. We find that VADER outperforms SentiWordNet in accuracy and F1 score across all three tests. The average F1 score across the three tests for VADER is 68.73%. This is slightly below Saif et. al's F1 score of 80.03% but still outperforms the baseline from Saif et. al's paper by at least 20%. An example of the visual SentiCircle for Facebook is shown in Figure 2 for the SentiWordNet. The most prevalent negative context word for Facebook is "stupid" while the strongest positive ones are "good", "major", and "better". The overall SentiMedian for Facebook is positive since more points lie in the positive than the negative y-axis region.

The experimental results are shown in Table 4, with most entities being classified as neutral. The three non-neutral results are highlighted. Using the VADER lexicon, two Democratic entities (Pelosi and Harris) are labeled positive while McConnell, a Republican, is labeled negative. The SentiWordNet SentiCircle for Nancy Pelosi is shown in Figure 3, showing "best" as the strongest positive term. The SentiWordNet SentiCircle for Mitch McConnell is shown in Figure 4 with "poverty" and "not" as the strongest negative words.

6 Discussion

We find that SentiCircles using the VADER lexicon outperforms the SentiWordNet lexicon, though the visual results are much easier to understand using the SentiWordNet lexicon. This is because the VADER lexicon has a much more expansive set of terms, so there is an overwhelming number of context terms on the graph since the experimental dataset contains over 70,000 tweets.

Nancy Pelosi's terms are generally more positive than Mitch McConnell's, considering Figures 3 and 4. Some of the sentiment terms in Figure 4 appear to have a negative connotation. A common joke on social media is comparing Mitch McConnell to a turtle, which is a negative reflection on him. Turtle is one of the strongest context terms since it has such a large radius. Another positive term that is surprising for Mitch McConnell is "evil". One would have expected that to also be a negative term.

The lambda choice for each lexicon is based on Saif et. al's work by considering the density of entities that are considered neutral since most entities in the evaluation dataset were neutral. This hyperparameter choice is not performed under a train-test framework, which is a potential data leak in classification that could be improved by expanding the size and diversity of entities in an evaluation dataset to split into train-test datasets.

Future work on this topic will include more expansive use of part of speech tagging to influence the output of the sentiment lexicon. More domain-specific processing of context terms is important so that topic-related words such as "poverty" are not considered sentiment in the context of political discussions. More formal procedures to negate sentiment words will further improve performance as well. Other related work on ensemble models, and combining lexicon and machine learning can be combined with the SentiCircles method to further improve performance.

The Joint Session was overall relatively neutral, with some favor to Democrats. This makes sense since a Democratic president was giving the speech. Another manner to break down the

sentiment further would be to use the timestamp on the tweet to break out the corpus into several parts to separately evaluate the sentiment on politicians during the main speech as well as the rebuttal by the minority party.

References

Da Silva, N. F. F., Hruschka, E. R., & Hruschka, E. R. (2014). Tweet sentiment analysis with classifier ensembles. Decision Support Systems, 66, 170–179. https://doi.org/10.1016/j.dss.2014.07.003

Saif, H., Fernandez, M., He, Y., Alani, H.: Evaluation datasets for twitter sentiment analysis. In: Proceedings, 1st Workshop on Emotion and Sentiment in Social and Expressive Media (ESSEM) in conjunction with AI*IA Conference. Turin, Italy (2013) https://github.com/pollockj/world_mood/tree/master/sts_gold_v03

Saif, H., He, Y., Fernandez, M., & Alani, H. (2014). Semantic patterns for sentiment analysis of twitter. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (Vol. 8797, pp. 324–340). Springer Verlag. https://doi.org/10.1007/978-3-319-11915-1_21

Saif, H., He, Y., Fernandez, M., & Alani, H. (2016). Contextual semantics for sentiment analysis of Twitter. Information Processing and Management, 52(1), 5–19. https://doi.org/10.1016/j.ipm.2015.01.005

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-Based Methods for Sentiment Analysis. Computational Linguistics, 37(2), 267–307. https://doi.org/10.1162/COLI_a_00049

Zhang, L., Ghosh, R., Dekhil, M., Hsu, M., & Liu, B. (2011). Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis.Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

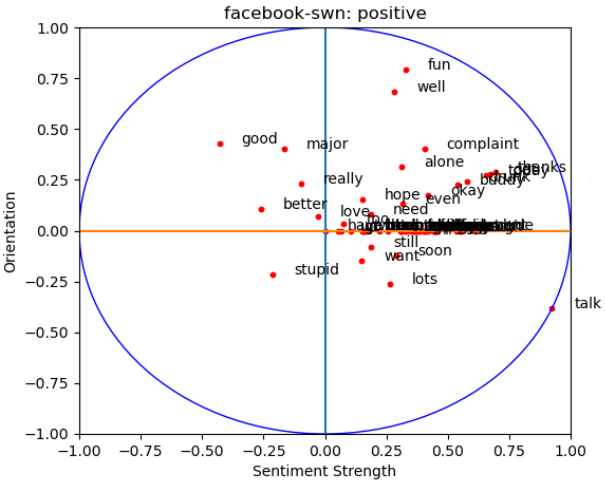

Figure 2: SentiCircles visualization of Facebook using SentiWordNet lexicon

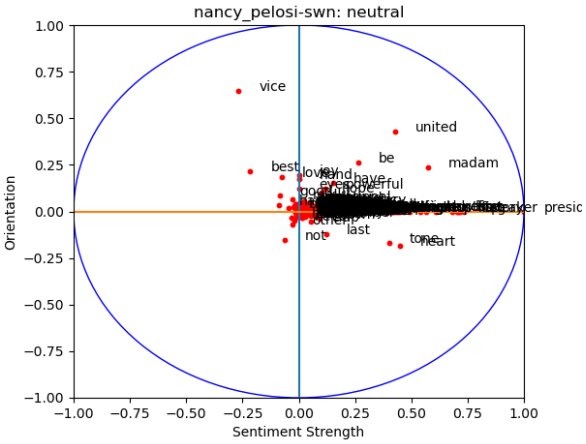

Figure 3: SentiCircles visualization for Nancy Pelosi using SentiWordNet

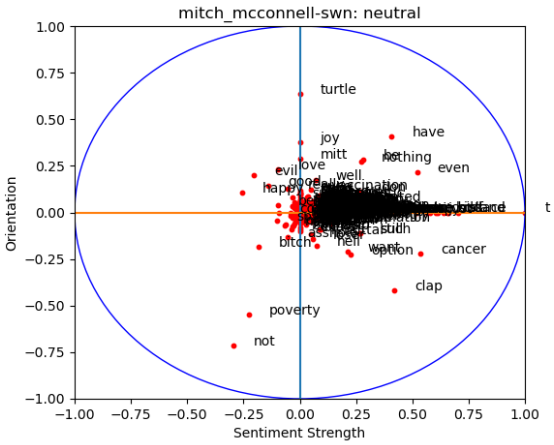

Figure 4: SentiCircles visualization for Mitch McConnell using SentiWordNet

Type of Test	Accuracy	Precision	Recall	F1 Score
SentiWordNet - subjectivity	0.618	0.900	0.486	0.632
VADER - subjectivity	0.618	0.660	0.892	**0.759**
SentiWordNet - positive vs. others	0.582	0.571	0.320	0.410
VADER - positive vs. others	0.564	0.512	0.840	**0.636**
SentiWordNet - negative vs. others	0.745	0.333	0.167	0.222
VADER - negative vs. others	0.873	0.778	0.583	**0.667**

Table 3: Results for evaluation dataset (STS Gold).

Entity	SentiWordNet Lexicon	VADER Lexicon
Joe Biden	Neutral	Neutral
Joint Session	Neutral	Neutral
Kamala Harris	Neutral	**Positive**
Mitch McConnell	Neutral	**Negative**
Nancy Pelosi	Neutral	**Positive**
Ted Cruz	Neutral	Neutral
Tim Scott	Neutral	Neutral

Table 4: Results for Joint Session dataset