# Three-Class Sentiment Classification of Instagram Text: An INFO 539 NLP Project

**Kendra Lyons**
University of Arizona, Tucson, AZ, USA.
kendramlyons@email.arizona.edu

## Abstract

Emotional appeals and opinions are often expressed in the captions of Instagram and other social media posts made by activists, grassroots organizations and non-profit businesses. Examining these posts with sentiment analysis can be useful for understanding how positive, neutral or negative they are. Pre-trained GloVe embeddings are used for measuring semantic similarity and dissimilarity of words and texts. This project used a supervised machine learning algorithm to classify text from Instagram posts based on cosine similarity between the centroid vectors of the text and selected GloVe vectors of positive, neutral and negative terms. A logistic regression classifier was used to predict the sentiment of Instagram posts as positive, neutral or negative. This approach yielded a small performance improvement compared to a baseline bag-of-words (BoW) logistic regression model.

## 1 Introduction

Sentiment analysis of text from social media has been performed using supervised, unsupervised and mixed machine learning approaches. It is useful for many reasons, including as a means of gauging how people are reacting to organizations, services or products. Performing sentiment analysis on social media content can help researchers understand not only how platform users feel about goods and services, but also how they feel about issues of public interest.

The spectrum that ranges from grassroots activists to non-profits to multinational NGOs is well-represented on Instagram. Their posts often make emotional appeals or share strong opinions and feelings about issues of public concern. Digital activism is one of many online spheres that people participate in that can impact how they feel about the world and themselves. Instagram is one very popular venue for digital activism and a good source for sentiment-rich text.

| **Raw Post Text** | **Sentiment Label** |
|---|---|
| Happy#NationalLibraryWeek! <U+0001F4DA> <U+2063> <U+2063> Now more than ever, we celebrate and thank the contributions of our nation's libraries and librarians, as well as promote library use and support. <U+2063> . . . | Positive |
| I love/hate what I study. Some days can get really emotional. Especially when I read and study politics and most of them lack humanity and empathy. Anyways, trying to finish this degree to make someone's life easier! Lessss go <U+0001F4C7> . . . | Neutral |
| When travelling on your own path, make sure you look around and reach out to help others. You never know how the tables may turn later on. . . . | Neutral |
| "What's happening in Yemen is the worst humanitarian crisis on the planet. 16 million people—over half of the country's population—are going hungry."-@repmarkpocan . . . | Negative |

Table 1: Examples of *positive, neutral,* and *negative* Instagram texts. The texts labeled as neutral express contrasting sentiments, making them more challenging to categorize. In this case, the positive and negative examples were easier to categorize.

There is a lot of sentiment analysis research on data from Twitter, but that is not true for Instagram. The research that does exist often focuses on images as well as text. However, the text caption of an Instagram post, at least in the "Instactivism" sphere, can be the most interesting part. Often, they are rich with sentiment, opinion, emotion and expression, making them well-suited for sentiment analysis. This may be of interest to social and political scientists, journalists and other researchers, especially those who seek to compare opinions on topics across platforms with different demographics and "vibes" (i.e. audiences/co-users on Instagram have been said to be more positive in general compared to those on Twitter).

For this project, we collected and manually labeled a dataset of the text and hashtags from 496 unique Instagram posts. Each text was classified as *positive, neutral* or *negative*. Several examples have been provided in Table 1. The positive and negative examples were not difficult to label, but the neutral examples both express ambiguous sentiment. Two options for dealing with ambiguity were categorizing them as "neutral" or adding one or more new classes; in this case they were labeled as "neutral" when both sentiments were expressed strongly. Indeed, uncertainty around which labels to apply to complex cases was an considerable source of error.

We trained a logistic regression model using this new dataset and pre-trained GloVe embeddings and obtained an F1 performance score of 63% and 54% on the development and test sets, two-point improvements over the baseline unigram model scores. Although a non-parametric bootstrap resampling test did not confirm that these results were statistically significant, they were obtained with only 298 labeled samples for training, 100 for development and 99 for testing. Notably, these small improvements were made without direct use of neural networks or other computationally-intensive machine learning techniques, and not in combination with any other features.

## 2 Related Work

There is a great deal of research about sentiment classification of texts from social media. Past research has investigated methods for the detection of sentiment conveyed through emoji or emoticons. In industry, there is interest in creating live-feed sentiment analysis tools that would allow organizations to monitor public opinion around a given topic based on a live feed of social media posts.

One such project was undertaken by Arunachalam and Sarkar at IBM (2013), who suggested that governments use sentiment analysis for tracking and (ideally) responding to public opinion. Some might cringe at the possible privacy implications of this. Others might observe that existing forms of political engagement with public opinion data (including monitoring opinions on social media) do not have much effect on what policy is actually implemented. However, governments aren't the only entities with an interest in understanding public opinion. Grassroots and non-profit initiatives working to raise awareness and promote social change could also use this approach to inform their public outreach strategies.

Arunachalam and Sarkar's approach used a modified TF-IDF approach involving topic-modelling and high-performance computing. They incorporated data from multiple social media platforms including Twitter, Facebook, Flickr and YouTube. This project, on the other hand, utilized the modest computational resources of a laptop and data from only one platform. In both cases, though, texts were selected based on their relevance to a topic of interest.

Sobhani, Mohammad and Kiritchenko (2016) addressed another problem relevant to this task, working to automatically detect not only sentiment in Twitter posts, but stance, as well. Stance is related to sentiment, but more specific, describing whether someone is in favor of or against a specific target topic. Sobhani et al. applied both sentiment and stance annotations to a large set of Tweets, noting that sentiment scores were useful but not sufficient for determining stance toward an issue. They used a linear-kernel SVM classifier with word- and character-level n-grams as well as "sentiment features from lexicons and word-embedding features" (Sobhani et al., 2016). In this project we stuck with simple unigram-based logistic regression, and did not annotate for stance, but this dataset would be a good candidate for such labelling in the future. For both projects, "query hashtags" were used to acquire a collection of texts related to a set of topics.

Another approach was taken by Chen et al. (2018), who incorporated emoji and emoticon characteristics for binary classification of Twitter posts. They used deep learning with both convolutional neural networks and long short-term memory for classification. Their research utilized pre-trained GloVe embeddings, as did this project, along with emoji embeddings trained by neural networks.Their method improved recall by about 4 percentage points. Notably, their text-preprocessing incorporated hashtag segmentation, a strategy which would likely improve the performance of the model used in this study. Their study also converted all text to lowercase and separated suffixes from contractions prior to model training, which this study did not.

## 3 Approach

This problem is framed as a three-class classification problem. The three target classes are *positive, neutral* and negative. Sentiment analysis is sometimes simplified with a binary class approach, but it ignores ambiguity in text such as when multiple contrasting sentiments are expressed at once (for reference, see the neutral examples in Table 1), which is often the case in texts found on social media. Sometimes it is easiest to classify such ambiguous texts as neutral, even though they actually convey strong feelings.

There were also challenges in deciding how to pre-process text before tokenization. For an example of how a text was cleaned, see Appendix A.

### 3.1 Data collection

The data used were scraped from Instagram in two batches, the first in April of 2020 and the second in April of 2021. They were collected for this study and labeled by hand (by one inexperienced and tired grad student) as positive, neutral or negative. Emojis were not rendered or taken into account during labeling. The data set consists of the text (including hashtags) of 496 unique Instagram posts that contain at least one of the following hashtags: *#grassroots, #socialmovement(s) #mutualaid, #protest(s), #filibuster* or *#communityorganization(s)*.

The data set is imbalanced in favor of positive posts, which represent 59% of the data. Neutral and negative labels represent only 22% and 19% of the data, respectively. The labeled data set was split using random, stratified sampling. The training set contains 60% and the development and test sets each contain 20% of the data. The classes in each subset are representative of the overall distribution of positive, neutral and negative texts.

### 3.2 Experiments

We evaluate the performance of two models. Performance is measured by F1 score, which is reported along with precision and accuracy in Table 3.

### 3.3 Models

**Baseline Unigram Model:** The only features included in the baseline model were unigram (BoW) features. The classifier used was logistic regression, tuned to account for the imbalanced class proportions. Text was pre-processed for both models the same way. See Table 6 for an example of a text before and after cleaning.

| Positive | Neutral | Negative |
|---|---|---|
| yes | maybe | no |
| positive | neutral | negative |
| good | okay | bad |
| agree | alright | disagree |
| like | middle | dislike |

Table 2: A selection of pre-trained GloVe vectors used for cosine similarity comparison with centroid vectors of texts. For a complete list, see Table 7 in Appendix A

**GloVe Regression Model:** The experimental model built upon the baseline model to make use of pre-trained GloVe embeddings (Pennington et al., 2014) trained on two billion tweets (27B tokens, 1.2M vocab, uncased) with 50 dimensions. The centroid vector of each tokenized, uncased text was calculated based on the tokens it contained that were also present in the GloVe collection. Then, the cosine similarity was calculated between each text's centroid vector and a series of selected positive, neutral and negative GloVe vectors. These cosine similarity scores were inputted as features (along with unigrams). For examples of the words corresponding to these selected vectors, see Table 2.

### 3.4 Hyperparameter Tuning

The scikit-learn logistic regression classifier class weight parameter was set to "balanced". It was shown to improve performance for the classes that were under-represented in the data, while still maintaining very good performance for over-represented positive examples. When not used, the classifier performed worse on neutral and negative examples, but slightly better on positive examples.

Other hyperparameters were tested but not implemented. Bigrams were included with unigrams but did not improve performance during development. Max_df and min_df were adjusted but the resulting models mostly performed worse than the default. Stop words were not removed, although stop word removal using sklearn's English stop words was tested. It was found to decrease performance for under-represented neutral and negative cases, and correctly identifying as many of those as possible was a priority.

| Model | Precision | Recall | F1 |
|---|---|---|---|
| Unigram Baseline (development) | 0.61 | 0.62 | 0.61 |
| Unigram Baseline (test) | 0.55 | 0.51 | 0.52 |
| GloVe Experimental (development) | 0.64 | 0.63 | 0.63 |
| GloVe Experimental (test) | 0.56 | 0.53 | 0.54 |

Table 3: Task performance across the baseline and experimental models on development and test sets.

### 3.5 Results

Baseline and experimental performance is given in Table 3. Overall the experimental results show a slight improvement over the baseline in F1, precision and recall, although it was not shown to be statistically significant by a non-parametric bootstrap re-sampling test. P-values are reported in Table 4. Results from all tests, including those run while tuning hyperparameters, are available in the data folder on GitHub.

| Model | P-Value |
|---|---|
| GloVe Experimental (development) | 0.093 |
| GloVe Experimental (test) | 0.3996 |

Table 4: P-values were calculated using a non-parametric bootstrap resampling test.

### 3.6 Error Analysis

Error categories and proportions are shown in Table 5 and described below. Figure 1 shows the distribution of each error category across the validation and test sets. Figure 2 shows the relative proportions of misclassification types by error category.

| Prop. | Error Class |
|---|---|
| 21.2% | Short texts or rare tokens |
| 18.2% | Gold label uncertainty |
| 16.9% | Advertising something |
| 15.2% | Advocating for something |
| 15.2% | Making a suggestion |
| 13,6% | Conveying urgency or need |

Table 5: Error classes and percentages of all 66 errors from the validation and test sets.

**Short or Rare (21.2%):** Texts that were either very short or mostly comprised of unique hashtags or other tokens unlikely to be found in other training texts or the pre-trained GloVe embeddings made up 21.2% of validation and test errors. For such texts, negative and positive texts were both commonly misclassified as neutral (*negative-neutral* and *positive-neutral*).

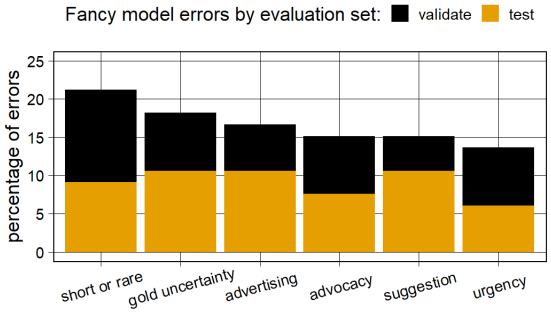


Figure 1: All error categories were represented in both development and test sets, but with some imbalance in the proportions.

**Gold Uncertainty (18.2%):** Texts that were difficult to select a label for due to the presence of multiple contrasting sentiments or other ambiguity made up 18.2% of errors. The complexity led to indecision on the part of the annotator about whether the manually applied or predicted label was correct. These were mainly *neutral-positive, negative-positive*, and *negative-neutral* misclassifications.

**Advertising (16.7%):** Texts whose primary purpose was to advertise events, products or services made up 16.7% of overall errors. Advertising texts were exclusively *neutral-positive, positive-neutral* and *positive-negative* misclassifications.

**Advocacy (15.2%):** Texts that advocated **for** or **against** an idea, cause, institution or person made up 15.2% of errors. They were mostly *neutral-positive* and *positive-neutral* misclassifications.

**Suggestion (15.2%):** Texts that made a suggestion to try or do something made up 15.2% of errors, as well. This is the category with the most even distribution of misclassification types. All types are present in proportions that are even relative to those in other categories, but the largest category is *positive-negative* misclassification.

**Urgency (13.6%):** Texts that expressed urgency, often including a request for aid or action of some kind, made up 13.6% of all errors. These texts represented the bulk of the *negative-positive* classifications.

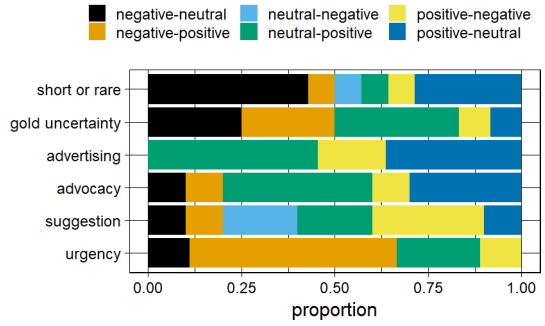


Figure 2: Proportion of prediction error types (true-predicted) by error category. Different kinds of misclassified texts had different compositions of prediction error types.

## 4 Conclusion

Future research could build on this model to improve performance. This approach using cosine similarity and pre-trained GloVe embeddings is likely to be useful in combination with other features not included in this analysis; even other features already present in the data such and post length and hashtag count may prove to be helpful. The data set could also be expanded to include more negative and neutral examples, or padded by replicating existing ones. A more balanced data set would also improve performance.

During this project, some important lessons were learned. The complexity of the texts collected and the resulting uncertainty on the part of the person doing the labelling presented challenges throughout. A future strategy to overcome this would be having multiple people label each text, then selecting the most common label as the "gold" standard.

## 5 Project Site

Data, scripts and figures are available at: https://github.com/kendramlyons/instagram-sentiment-analysis.

## References

Ravi Arunachalam and Sandipan Sarkar. 2013. The new eye of government: Citizen sentiment analysis in social media. In *Proceedings of the IJCNLP 2013 Workshop on Natural Language Processing for Social Media (SocialNLP)*, pages 23–28, Nagoya, Japan. Asian Federation of Natural Language Processing.

Chia-Ping Chen, Tzu-Hsuan Tseng, and Tzu-Hsuan Yang. 2018. Sentiment analysis on social network: Using emoticon characteristics for twitter polarity classification. In *International Journal of Computational Linguistics & {C} Chinese Language Processing, Volume 23, Number 1, June 2018*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Parinaz Sobhani, Saif Mohammad, and Svetlana Kiritchenko. 2016. Detecting stance in tweets and analyzing its interaction with sentiment. In *Proceedings of the fifth joint conference on lexical and computational semantics*, pages 159–169.

## A Appendix

| **Raw and pre-processed text** |
|---|
| There's a difference between "going with the flow" and "strategizing for the flow". Going with the flow has it's benefits, but strategizing is what helps you create lasting impact through effective programs and actions. Choose your approach wisely. #strategygame #strategy #nonprofitorganization #nonprofitorganizations #ngo #communityservice #givingbacktothecommunity #communityorganization #communityorganizations #changemaker |
| There's a difference between "going with the flow" and "strategizing for the flow" Going with the flow has it's benefits but strategizing is what helps you create lasting impact through effective programs and actions Choose your approach wisely strategygame strategy nonprofitorganization nonprofitorganizations ngo communityservice givingbacktothecommunity communityorganization communityorganizations changemaker |

Table 6: Example of an Instagram post text before and after cleaning. Quotation marks were missed accidentally but apostrophes and exclamation points were kept intentionally.

| **Sentiment** | **Selected GloVe embeddings** |
|---|---|
| Positive | "yes", "positive", "good", "agree", "like", "happy" |
| Neutral | "maybe", "neutral", "okay", "alright", "middle", "might", "unsure", "moderate", "inform" |
| Negative | "no", "negative", "bad", "disagree", "dislike", "not", "hate", "despise", "angry", "mad", "pathetic", "ugh", "stupid" |

Table 7: More vectors were selected for neutral and negative terms than positive terms. The addition of specific "positive" terms such as "thank" actually decreased performance during development.