On Assessing the Sentiment of *General* Tweets

Sifei Han[1] and Ramakanth Kavuluru[1,2]⋆

[1] Department of Computer Science
[2] Division of Biomedical Informatics, Department of Biostatistics
University of Kentucky, Lexington, KY
{eric.s.han,ramakanth.kavuluru}@uky.edu

Abstract. With the explosion of publicly accessible social data, sentiment analysis has emerged as an important task with applications in e-commerce, politics, and social sciences. Hence, so far, researchers have largely focused on sentiment analysis of texts involving entities such as products, persons, institutions, and events. However, a significant amount of chatter on microblogging websites may not be directed at a particular entity. On Twitter, users share information on their general state of mind, details about how their day went, their plans for the next day, or just conversational chatter with other users. In this paper, we look into the problem of assessing the sentiment of publicly available *general* stream of tweets. Assessing the sentiment of such tweets helps us assess the overall sentiment being expressed in a geographic location or by a set of users (scoped through some means), which has applications in social sciences, psychology, and health sciences. The only prior effort [1] that addresses this problem assumes equal proportion of positive, negative, and neutral tweets, but a casual observation shows that such a scenario is not realistic. So in our work, we first determine the proportion (with appropriate confidence intervals) of positive/negative/neutral tweets from a set of 1000 randomly curated tweets. Next, adhering to this proportion, we use a combination of an existing dataset [1] with our dataset and conduct experiments to achieve new state-of-the-art results using a large set of features. Our results also demonstrate that methods that work best for tweets containing popular named entities may not work well for general tweets. We also conduct qualitative error analysis and identify future research directions to further improve performance.

1 Introduction

Sentiment analysis (or opinion mining) has gained significant attention from the computer science research community over the last decade due to the rapid growth in e-commerce and the practice of consumers writing online reviews for products and services they have used. Movies, restaurants, hotels, and recently even hospitals and physicians are being reviewed online. Manually aggregating all information available in a large number of textual reviews is impractical. However, discovering different aspects of the product/service that the review

⋆ corresponding author

is discussing and the corresponding evaluative nature of the review for each of them is computationally challenging given the idiosyncratic and informal nature of customer reviews. Sentiment analysis has also been essential in gleaning information from customer surveys that companies routinely conduct. Due to our direct involvement in an ongoing project, we also observe that companies consult researchers to conduct sentiment analysis of emails of their employees to assess personnel morale and to improve organizational behavior and decision making. Recently, in the field of healthcare, researchers have focused on identifying emotions in suicide notes [25] and predicting county level heart disease mortality using Twitter language usage [6].

Due to the short informal nature of messages called *tweets* and the asymmetric network structure, since its introduction in 2006, Twitter has grown into one of the top 10 visited websites in the world with 100 million daily active users who generate over 500 million tweets per day [28]. Instead of going to a product website, Twitter users (henceforth *tweeters*) discuss their opinions and express their sentiments on different topics to their followers and all other users (if they wish). Given a recent study [14] reveals that over 95% of Twitter profiles are public (the default setting), Twitter has become an interesting platform to track sentiment on different topics and to assess the general mood of tweeters at specific locations and times. The informal text also poses challenges through (sometimes intentionally) misspelled words, neologisms, and other short forms that do not occur in dictionaries. Emoticons, abbreviations, user mentions, and hashtags also add to the complexity of analyzing tweet sentiment. We request the readers to refer to a recent survey [15] for details on general approaches to sentiment analysis on Twitter.

Most current efforts that analyze tweet sentiment directly or indirectly focus on tweets that contain popular topics or entities and tend to use datasets that are skewed to contain fewer neutral tweets. The ongoing series of Semantic Evaluation (SemEval) tasks added a Twitter sentiment analysis track in 2013 [19, 26] in which the dataset selection was done based on the presence of a popular named entity in the tweet *and* the presence of at least one word with positive or negative sentiment score > 0.3 in SentiWordNet 3.0 [2], a lexical resource that contains positive, negative, and objectivity scores for synsets in WordNet. Although this is justified for tasks that involve analyzing sentiment of tweets that discuss a popular topic or entity, several tweets from the Twitter firehose may not discuss a popular topic. Tweeters might be chatting with others, sharing information on how their day went or their plans for the next few days, or just tweeting about how they feel at the time. Consider the following tweets from our dataset

- `I feel so accomplished i had 1 Liter of water ! 1 more to go`
- `Good thing i have no work today`
- `Well headed to dorm to nap then open gym to practice`
- `Started a new diet where I only eat fast food`
- `Math exam tomorrow is going to kick my butt`
- `I swear this cold is gonna get the best of me`

We noticed that most tweets in our dataset (randomly selected using Twitter streaming API) have the general nature as those in these six examples. Although some of them contain sentiment words, many do not and most tweets are not on any popular topic. This seems to be in line with the original intention of Twitter creators: until November 2009 Twitter had "What are you doing?" as the prompt displayed to the users when they log in; since then this has been changed to "What's happening?" Although this sample tweet list is slightly biased to show sentiment expressing tweets, we notice that many tweets are neutral or objective in nature.

However, to gauge the aggregate sentiment from a geographic location or sentiment expressed by a user group, it is essential to be able to determine the sentiment of all tweets (without any other topic based or sentiment word based selection bias). Such scenarios arise naturally in social sciences, psychology, and health sciences especially in the domain of mental health and substance abuse. For example, researchers might want to analyze the sentiment of tweet streams from users who identify themselves as smokers or vapers (e-cigarette users) or users from a particular area that reports low health rankings (`http://www.countyhealthrankings.org/`). Further selection of users can be done based on predicted age group, gender, race, or ethnicity [13, 20, 24]. The results can also be extended to interview or counseling narratives of mental health patients by aggregating sentiment expressed in each sentence. However, except for the lone effort by Agarwal et al. [1], we are not aware of attempts to determine sentiment of a set of *general* tweets[3] collected using the Twitter streaming API. Even in their effort, Agarwal et al. assume equal proportion for the three sentiment classes, which our manual analysis shows is not realistic. So in our current effort

1. We manually estimate the proportion of [positive : negative : neutral] tweets to be 29% (26–32%) : 18% (16–21%) : 53% (50–56%) from a sample of 1000 randomly selected tweets selected from a set of 20 million tweets collected through Twitter streaming API in 2013. We also estimate that only 10% (7–13%) of the tweets have named entities in them. The 95% confidence intervals of the proportions calculated using Wilson score [31] are shown in parentheses.
2. Adhering to this estimated class proportion, we combine the dataset used by Agarwal et al. [1] with our dataset to create a larger dataset and conduct experiments with a broad set of features to identify a combination of features that offers the best performance (macro average of positive and negative sentiment F-scores). We also show that our best model improves over Agarwal et al.'s results on their original dataset with equal class proportions. Furthermore, we also show that a system comparable to the top performer [17] in SemEval tasks may not suffice for our general tweets, warranting identification of high performing feature subsets.

[3] Note that some of these tweets may contain named entities or popular topics but we are not prescreening those that contain such themes

3. We analyze the confusion matrix and manually identify causes for certain types of errors and future research directions to improve sentiment analysis of general tweets.

2 Background and Related Work

Sentiment analysis has emerged as an important sub-discipline within natural language processing research in computer science. Given it is very difficult for human users to exhaustively read and understand large numbers of potentially subjective narratives, automated methods have gained prominence over the past decade pursued first as document level classification tasks [21, 27] and subsequently as sentence level [9], and recently as phrase level [19, 32] tasks. Unsupervised approaches (e.g., [27]) that take advantage of sentiment lexicons, supervised approaches (e.g., [22]) that employ statistical learning, and semi-supervised approaches (e.g., [33]) that automatically generate training data have evolved as different alternatives that are currently being used in a hybrid fashion to obtain state-of-the-art results. Purely lexicon based approaches suffer from low recall and statistical learning approaches that rely only on tweet content and labeled data often offer low precision, especially with smaller training datasets; this has been mitigated to some extent with the advent of Internet crowd sourcing opportunities such as Amazon Mechanical Turk for generating large training datasets. Although manually building high coverage sentiment lexicons is impractical, automated approaches to induce them have resulted in significant performance gains [11]. This has proven especially useful for Twitter data given its 140 character limit and the extremely informal nature of communication, due to which popular hand-built lexicons were found insufficient.

One of the first notable attempts in sentiment analysis for tweets was by Go et al. [8] who used supervised learning and emoticon based distance supervision to acquire training data. Researchers have also focused on target dependent sentiment classification [4, 10] where the sentiment is associated with a target concept. From 2013, a shared task [19] on Twitter sentiment analysis has been added to the annual SemEval workshop. Researchers at NRC-Canada entered the best performer [17] in the 2013 SemEval task. They designed a sophisticated hybrid sentiment analysis system that incorporates both hand-built and automatically constructed sentiment lexicons as features, besides using the conventional ngram features, in a supervised learning framework. Recently, they improved upon their results [11] by generating separate lexicons for affirmative and negated contexts. Although these efforts significantly advance the state-of-the-art in tweet sentiment analysis, they all use datasets that have been curated to contain popular topics during the collection period. Named entities or event names such as `iPhone, Gaddafi, AT&T, Kindle,` and `Japan Earthquake` are used and in the case of SemEval tasks, additionally, presence of sentiment expressing words is also required, thus inherently skewing the dataset to subjective tweets [19]. It is not clear whether methods that produce the best results on these datasets work best for general tweets as indicated in Section 1, where we discuss applications of sentiment analysis of general tweets.

Agarwal et al. [1] curated a set of random tweets collected using Twitter streaming API and conducted supervised learning experiments with a broad set of features also incorporating some sentiment lexicons. As they point out, their effort is the first in looking at such tweets without any pre-screening constraints and to our knowledge is the only such attempt. They introduce a new tree kernel representation of tweets and show that this representation performs on par with traditional approaches that involve content based features and sentiment features including emoticons, parts of speech, lexicon based prior polarity scores, and presence of hashtags, user mentions, and URLs. However, they assume that the positive, negative, and neutral classes are equal in proportion, which our analysis shows is not realistic. Hence, in our current effort we first estimate the proportion of the three classes, build a representative dataset, and conduct experiments to identify feature combinations that achieve best performance.

3 Datasets and Performance Measures

We sampled tweets using Twitter streaming API periodically in 2013 and collected over 20 million English tweets from which we curated two randomly selected tweet datasets one with 10,000 tweets and another with 1000 tweets. We used the larger dataset to conduct automated analysis of different characteristics of general tweets and the smaller dataset to manually determine the distribution of positive, negative, and neutral class distributions.

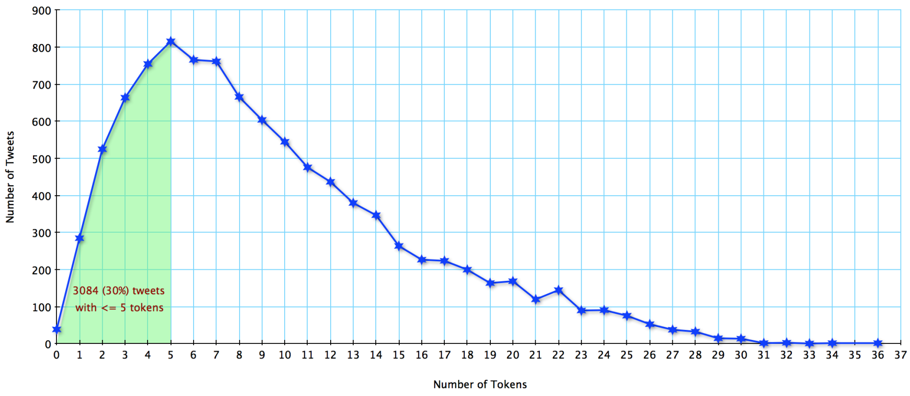

Fig. 1. Distributions of tweets with different numbers of tokens from a dataset of 10,000 randomly selected tweets

In Figure 1 we plot the number of tokens in a tweet on the x-axis and the corresponding number of tweets in the larger dataset of 10,000 tweets. We consider each user mention, emoticon, URL, and hashtag as an individual token. As we can see from the figure, 30.8% (29.9–31.7%) of the tweets have fewer than 6 tokens. Based on this dataset, we also observe that 12.5% (11.9–13.2%) of general tweets have URLs, 36.1% (35.2–37.1%) contain user mentions, 13.1% (12.5–13.8%) use hashtags, and 5.3% (4.9–5.8%) have emoticons in them, where the ranges in parentheses represent the 95% confidence intervals computed based on Wilson score [31].

We have two annotators independently perform a three way classification of smaller 1000 tweet dataset instances into positive, negative, or neutral classes. The two annotators were given general instructions on classifying the tweets and were later asked to discuss and resolve disagreements through discussion. We had *moderate* agreement ($\kappa = 0.54$) based on the general rule of thumb [12] on observer agreement for categorical data. Based on the consolidated judgments we estimate the class proportion ratio [positive : negative : neutral] to be 29% (26–32%) : 18% (16–21%) : 53% (50–56%) with 95% confidence intervals shown in parentheses. Although the proportions are different, the proportion of positive tweets is larger than that of negative tweets even in the SemEval datasets [19]. Adhering to our estimated proportion, we combine our dataset with the dataset used in Agarwal et al. [1] to build a consolidated larger dataset of 3523 tweets with 1844 neutral, 1011 positive, and 668 negative tweets. Since neutral tweets are the majority, we first merge neutral tweets from both datasets, and randomly select positive and negative tweets according to our estimated proportion. Since the dataset in [1] has non-English tweets, we use Natural Language Toolkit (NLTK [3]) `'words'` corpus' English word subset to first automatically filter tweets when at least 40% of words in the tweet are English words; these filtered tweets are subsequently manually filtered to obtain only English tweets. We undertook this pre-screening process, given misspellings and other neologisms may not be in the NLTK dataset. Given tweets in Agarwal et al. dataset are annotated by a single person, we also annotated that dataset and found *substantial* agreement ($\kappa = 0.8$). Disagreements were resolved by an arbiter. Given there are three classes, when the arbiter disagreed with both annotations ($< 1\%$ cases), the corresponding tweets were discarded.

Besides classifier accuracy (proportion of all tweets correctly classified into the corresponding classes), we also assess the macro average of F-score of the positive and negative classes, which we term as F-Sent for simplicity for the rest of this paper. Given F_+ and F_- are F-scores for the positive and negative classes respectively, then F-Sent $= (F_+ + F_-)/2$. This measure takes into account the FPs and FNs caused (including those due to neutral tweets) in classifying positive and negative sentiment categories but does not directly incorporate credit for correctly classifying neutral tweets. It is well known and has been used as the main measure in the SemEval [19] Twitter sentiment analysis tasks.

4 Supervised Classification Framework

We follow the hybrid approach of employing sentiment lexicons as features in the supervised framework while also using the conventional content based feature (e.g., n-grams) and Twitter specific features (e.g., emoticons and hashtags).

Our main classifier is the well known linear support vector machine from the LIBLINEAR [7] library made available in the scikit-learn [23] machine learning framework. We use automatic class weighting supported through the classifier and use the default one-vs-rest approach for three way classification of sentiment categories. Free text is pre-processed in general to minimize the noise in the feature space for text classification. For our tweet dataset we fully replicate the approach used by Agarwal et al. [1, Section 4] by replacing tweet targets (user mentions using the @ symbol) and URLs with specialized tokens since specific user mentions and URLs tokens often do not constitute meaningful features. Emoticons are replaced by their polarity based on the emoticon polarity dictionary built by Agarwal et al. [1]. We replace negation words (e.g., not, no, never, n't, cannot) with a single "NOT" token and expand popular slang acronyms (e.g., rofl, lol) to full forms. We incorporate a large set of features used by Agarwal et al. [1] and Kiritechnko et al. [11] and introduce new lexico-syntactic features that combine sentiment expressing words with their parts of speech and dependency edges involving them.

Lexical features: We use word unigrams and bigrams (henceforth called just ngrams) as the base features with feature weighting based on Naives Bayes (NB) scores [29] computed using the training data. We refer to this way of using NB scores as input to an SVM classifier as NBSVM as introduced by Wang and Manning [29]. The numbers of tokens with all capitalized letters, hashtags, elongated words[4], contiguous sequences of question marks, exclamation marks, or a combination of both are also included in the feature list.

Syntactic features: We incorporate numbers of each part-of-speech (POS) tag type as a feature. Since sentiment lexicons often record word polarity scores without specifying the word POS (except the MPQA subjectivity lexicon), it is difficult to compensate for the polarity scores that might be incorrectly considered as features. For example, consider the sentences "I like watching movies" and "It smells like popcorn". Sentiment lexicons might have a high positive score for the word 'like' but it does not apply for the second sentence. So we include a new lexico-syntactic binary feature <w>-POS(w) where 'w' is a sentiment expressing word found in sentiment lexicons and POS(w) is its part-of-speech as observed in an input tweet. For such words, based on the dependency parse [5] of a tweet, we also introduce another lexico-syntactic binary feature <w>-<g/d>-<dtype>, where 'dtype' is the type of a dependency relation involving 'w' and 'g/d' is determined based on whether the relation has 'w' as a governor (g) or dependent (d). For example running Stanford parser on "it smells like popcorn" generates dependencies `prep(smells, like)` and `pobj(like, popcorn)` which give the features `like-d-prep` and `like-g-pobj`. This is to capture the effects of syntactic relations involving sentiment words on the overall tweet sentiment.

[4] We do not consider elongated versions for ngrams and shorten such tokens as in [1]. However, we look at the original tweet text to determine the number of such words.

Sentiment lexicon based features: A sentiment lexicon typically has a list of words with the corresponding polarity expressed simply as a binary positive or negative categorization. More recent lexicons, especially those curated automatically, assign numerical scores that indicate the polarity of the word where a positive (negative) value typically indicates a positive (negative) polarity and the magnitude of the value corresponds to strength of the sentiment. For our experiments we use the hand-built Bing Liu lexicon [9], MPQA subjectivity lexicon [32], and the NRC-Canada Emotion Lexicon [18]. For these lexicons, since numerical scores are not explicitly provided, we choose appropriate integer scores based on the polarity and any corresponding strength/intensity information available following the approach by Kiritchenko et al. [11]. We also use an automatically created sentiment lexicon, the Hashtag Sentiment Lexicon (HSLex), constructed by researchers at NRC-Canada using a dataset of tweets with a few hashtagged emotion words. We use their latest version [11] of this lexicon where they generate different scores for affirmative and negated contexts. Given these different lexicons the actual features are as follows:

1. For each lexicon used, the total score of all sentiment expressing ngrams that occur in the tweet, the total score of only positive (negative) ngrams, the maximum score among positive (negative) ngrams, score of the last token in the tweet, and all these scores computed separately for unigrams and bigrams. The scores for ngrams in a negated context (as identified in [22]) within a tweet are obtained from the negated context lexicon for the automatically created HSLex lexicon [11].
2. For each lexicon used, the total numbers of sentiment expressing ngrams, negation words, positive ngrams, negative ngrams, and all these counts computed separately for unigrams and bigrams. We also include the numbers of positive (negative) emoticons and also their presence (so binary) as the last token of the tweet.
3. For each lexicon used, similar to how scores and counts for ngrams are computed (items 1 and 2 in this list), we also incorporate as features, sentiment (aggregated) scores and counts for different parts of speech that occur in a tweet. This is based on the link between the unigrams in the tweet and the associated POS tags and the presence of such unigrams in the lexicons.

5 Experiments, Results, and Discussion

We split our dataset into 80% training and 20% test sets using stratified sampling with class proportions maintained according to the distribution in the full dataset. We ran 5-fold cross validation hundred times (using distinct shuffles) on the training dataset to identify the best feature combination among all features described in Section 4. The best combination chosen was the one that had the maximum average F-Sent score over those 100 iterations. Given the large number of features, for computational tractability of considering all possible combinations, we divided all features into ten distinct groups: 1. four groups from the lexicon based features (corresponding to the list at the end of Section 4 with separate score and count groups from the third item); 2. three groups from syntactic features (POS tag counts and the two new lexico-syntactic features as singleton groups); and 3. three groups from lexical features (ngrams, NBSVM weighting, and all Twitter specific features such as all-caps words and elongated words as one group). The best feature combination based on our experiments is the union of all features excluding the following features: POS tag counts, emoticon counts, lexical features such as elongated or all-caps words, and the lexico-syntactic feature that joins a sentiment word with its POS tag. Using this best feature combination, we used cross validation and grid search to identify the best regularization parameter and tolerance value for stopping criteria for the SVM classifier. We finally trained using the best feature combination and parameter settings and ran our model on the test set to obtain accuracy of 70.70% and F-Sent of 62.87%. However, since this is based on a single 80:20 split of our dataset, we repeated our experiments over hundred distinct 80:20 splits and obtain results shown in Table 1 which shows a mean F-Sent of 62.28 with a 95% confidence interval of 61.82–62.75%.

Table 1. Average accuracy (Acc.) and F-Sent over 100 distinct 80%-20% train-test splits when using all features, the best feature combination with feature group ablated performances

Features	Train Stats		Test Stats		
	Acc.	F-Sent	Acc.	F-Sent	95% CI F-Sent
Best Combination	71.56%	64.48%	70.47%	62.28%	61.82–62.75%
– Dependency Feature	71.50%	64.44%	70.45%	62.17%	61.75–62.60%
– NBSVM	69.64%	61.76%	70.13%	62.49%	62.02–62.95%
– Ngram+NBSVM	67.01%	58.60%	67.33%	58.99%	58.57–59.40%
– Lexicon Features	65.29%	53.23%	65.56%	53.43%	52.90–53.96%
All features	68.52%	60.11%	69.05%	60.87%	60.42–61.32%
NRC-Lite	68.52%	58.09%	68.57%	59.39%	58.97–59.82%

In Table 1, rows 2–5 indicate the performance if we remove feature classes from the best combination. The biggest drop in performance is obtained when lexicon features removed resulting in a 9% drop in F-Sent and 6% drop in accuracy. As noted by Kiritchenko et al. [11], the drop due to ngrams ablation is significantly less (row 4) compared to removing lexicon features. Dropping the NBSVM weighting causes 2.72% loss in F-Sent in training but shows a negligible increase in performance over the best combination test average. Although test accuracy drops when ablating NBSVM weighting, it is also negligible in contrast with the corresponding drop of nearly 2% in training. The drop in performance due to the dependency based feature is also not significant. These results for NBSVM weighting and dependency features could be due to the sparsity of token frequencies and (word, dependency type) pair frequencies, respectively, and need further investigation with a larger dataset. The penultimate row of Table 1 shows that identifying the best combination results in 1.41% improvement in test F-Sent score and over 3% improvement for the training F-Sent score. We also experimented with a system comparable to that used by NRC-Canada researchers [11] which we call NRC-Lite since we removed certain features, specifically, word cluster scores, ngrams that are longer than 2 tokens (given the sparsity), non-contiguous ngrams, and lexicon features from the Sentiment140 lexicon [11]. From the final row in Table 1, we notice that our best performer test F-Sent is 2.91% higher than that for NRC-Lite and the gains are over 6% for training F-Sent. However, many of our features are from [11] and identifying the best feature combination might have been the key in obtaining higher performance.

Although most current approaches directly model sentiment analysis into a direct three way classification problem, there is a more intuitive two stage approach where a binary classifier first distinguishes tweets that carry sentiment from objective/neutral tweets. For such tweets identified through the first stage model, a second binary classifier identifies whether the tweet expresses positive or negative sentiment. We curated separate subsets of our dataset for the corresponding classifiers in these two stages and experimentally identified the best feature combination for both types of classifiers in exactly the same way as we did for the direct three way classification. Finally, over hundred different 80-20% train-test splits of our data, we obtained a mean test F-sent of 60.25% (59.79–60.71%) and mean accuracy of 69.93% (69.61–70.25%) with 95% confidence intervals show in parentheses. Compared with our best results (first row of Table 1), we notice a drop of 2% in F-Sent and 1% in accuracy. This has been our experience with the two-stage approach even in other text classification domains that have hierarchical class structures. Given neutral tweets cause the most errors (more later in Section 6), we believe that the first stage classifier propagates errors to the second stage to an extent that limits the overall performance of the approach.

Table 2. Average performance measures with our best combination based on 5-fold cross-validation using 100 distinct shuffles of the original Agarwal et al. [1] dataset with equal class proportions

Measures	Agarwal et al.	Our Best Combination	
		Mean	95% CI
Accuracy	60.50%	62.85%	62.78–62.92%
F-Sent	60.23%	64.68%	64.61–64.76%
F_+	59.41%	64.07%	63.98–64.16%
F_-	61.04%	65.30%	65.21–65.39%
$F_{neutral}$	60.15%	59.74%	59.65–59.82%

We conducted additional experiments to see how our best feature combination performs on the original general tweet dataset by Agarwal et al. [1] with equal proportions for the three classes (1709 tweets per class). We followed their approach of five-fold cross validation and obtained results as shown in Table 2, which indicates an improvement of over 4% in F-Sent and 2% in accuracy. Since our features are geared towards identifying tweets with polarity, we notice a significant increase in F_+ and F_- and a negligible drop in $F_{neutral}$.

6 Qualitative Error Analysis

In Table 3, we display the test and training error confusion matrices where the rows represent ground truth and columns are predicted classes. A glance at them shows that in both scenarios most errors involve neutral tweets. To be precise, 86% of test errors and 89% of training errors are caused due to neutral tweets. This is the main motivation for our effort and this observation strongly backs our belief that datasets with a realistic distribution of the three classes should be used without pre-screening bias.

Table 3. Confusion Matrices for Training and Test Datasets

T\P	+	-	N
+	142	13	47
-	16	71	46
N	48	36	284

(a) Test Matrix

T\P	+	-	N
+	542	38	230
-	45	304	186
N	166	132	1177

(b) Training Matrix

Given this situation, we manually analyzed a few misclassified tweets. Since we did not impose a constraint on the length of the tweets, we found several examples of short tweets that have been misclassified. Consider the negative tweet, `I’m not fine`, misclassified as a positive tweet. The main clue is the negation word followed by the word ‘fine’. However, in the automatically created HSLex [11], we find a score of 0.832 for `fine_NEG`. While this might not be the main reason, there is not much additional information to rely on for the learner for such short tweets. We believe a specific customization of the features and classification framework for short tweets might be essential. Although our best combination did not include elongated words and other tweet specific lexical features, based on our manual analysis, we believe these features might play a crucial role for shorter tweets. Based on our observation in Figure 1, since 30% of tweets are short tweets with fewer than 6 tokens, we believe this to be an interesting research direction. However, our initial experiments on building two separate classifiers for short ($\leq$ 5 tokens) and long ($>$ 5 tokens) tweets did not result in overall performance improvements, potentially due to the very small size of the training dataset for short tweets. However, we noticed that the percentage of neutral tweets increases by 10% in short tweets compared to the full dataset. This further confirms that a more involved customization for short tweets is essential.

Consider this positive tweet misclassified as a neutral tweet: `@jenna_bandi ohhhh my lorddddd your a lifesaver`. The bigram "ohhh my" has a positive score of 1.64 but "your a" has a negative score of -1.05 in HSLex. Due to a missing space between 'life' and 'saver', we missed important evidence given both words have positive scores. Splitting up potential bigrams into constituent unigrams (in addition to retaining the original token) might provide more evidence toward the correct sentiment of the tweet. Tweets that start out positive (negative) but end up conveying a more negative (positive) sentiment latter might need special handling. Consider this negative tweet misclassified as neutral: `@vixxybabyy hopefully a GED !!! But even that might not happen for this one`. Researchers have had success [30] by simply splitting the tweet in the middle (in terms of word count) and treating tokens in both halves as having a separate feature type. We will employ this approach in our future work.

We end with an example of a neutral tweet misclassified as a positive tweet: `This ice is supppppa cold, but then again it is ice`. This tweet showcases the complexity of identifying neutral tweets. Even if the slang word is correctly identified as 'super', the overall sentiment might still be positive given 'super' and 'super cold' both have positive sentiment scores in HSLex.

7 Conclusion

Most current efforts in sentiment analysis of Twitter data are focused on datasets that are biased toward tweets that contain popular named entities and sentiment expressing words. While there is merit to this focus, it is also important to consider general tweets most of which (90% according to our estimate) do not contain named entities and are essentially conversational chatter about tweeters' daily activities and personal situations or their general mood. Sentiment analysis of such tweets can help study the sentiment expressed via Twitter by different groups of tweeters based on demographics (age, race/ethnicity, gender, location) and additional attributes (e.g., smokers, vapers) across time and to correlate this information [16] with additional locational information (county health rankings, urban/rural indices). To our knowledge, there is only one earlier effort that looks at general tweets by Agarwal et al. [1], although the authors assume equal proportion of positive, negative, and neutral tweets.

In this paper, we first estimated the proportion of the three classes using manual annotation of a random sample of 1000 tweets selected from over 20 million tweets collected in 2013. Our analysis showed that class proportions are skewed and that more than half of the tweets are neutral justifying additional efforts for general tweet sentiment analysis. Based on the estimated proportion, we constructed a new dataset and conducted experiments using well known features, including those derived from sentiment lexicons. We also introduced additional lexico-syntactic features based on part-of-speech tags and dependency parses for sentiment expressing words. Unlike prior efforts, we identified best feature combinations based on repeated cross validation experiments on different shuffles of the training data. We demonstrated that our best feature combination provides statistically significant performance improvements over using all features as indicated by non-overlapping 95% confidence intervals (last column of the first and penultimate rows in Table 1), which justifies our approach of identifying feature subsets instead of simply using a large set of features. Our feature ablation experiments demonstrated that the lexicon based features contribute the most to the performance of our models, corroborating the findings of other researchers [11]. Additionally, we also showed that models based on our best feature combination outperform prior approaches on the original equal proportioned dataset of general tweets by Agarwal et al. [1]. At the time of this writing, our current effort is the first to study the distribution of the sentiment classes in general tweets and the associated sentiment analysis of such tweets based on a dataset constructed according to the estimated class proportions.

We conducted qualitative error analysis of our results and identified important future research directions. Besides improvements identified in Section 6, we believe that a larger dataset might be more suitable for further research in assessing the sentiment of general tweets from the Twitter stream, especially in building separate classifiers for short and long tweets. Given the presence of a large number of neutral tweets, it might also be more desirable to employ more than two annotators through an online crowd sourcing approach.

Acknowledgments

Many thanks to anonymous reviewers for their detailed comments that greatly helped improve the quality of this paper. We are also grateful to Apoorv Agarwal for providing the dataset used in his prior efforts. The project described in this paper was supported by the National Center for Advancing Translational Sciences (UL1TR000117). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

1. A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Languages in Social Media*, pages 30–38. Association for Computational Linguistics, 2011.
2. S. Baccianella, A. Esuli, and F. Sebastiani. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 2200–2204, 2010.
3. S. Bird, E. Klein, and E. Loper. *Natural Language Processing with Python.* O'Reilly Media, 2009.
4. L. Chen, W. Wang, M. Nagarajan, S. Wang, and A. P. Sheth. Extracting diverse sentiment expressions with target-dependent polarity from twitter. In *Proceedings of the Sixth International Conference on Weblogs and Social Media, ICWSM*, pages 50–57, 2012.

5. M. de Marneffe, B. MacCartney, and C. Manning. Generating Typed Dependency Parses from Phrase Structure Parses. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, pages 449–454, 2006.
6. J. C. Eichstaedt, H. A. Schwartz, M. L. Kern, G. Park, D. R. Labarthe, R. M. Merchant, S. Jha, M. Agrawal, L. A. Dziurzynski, M. Sap, et al. Psychological language on twitter predicts county-level heart disease mortality. *Psychological science*, 26(2):159–169, 2015.
7. R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.
8. A. Go, R. Bhayani, and L. Huang. Twitter sentiment classification using distant supervision. Technical report, Dept. of Computer Science, Stanford Univ., 2009.
9. M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM, 2004.
10. L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao. Target-dependent twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 151–160. Association for Computational Linguistics, 2011.
11. S. Kiritchenko, X. Zhu, and S. M. Mohammad. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, pages 723–762, 2014.
12. J. Landis and G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.
13. W. Liu and D. Ruths. What's in a name? using first names as features for gender inference in twitter. In *Proceedings of the AAAI Spring Symposium: Analyzing Microtext*, pages 10–16, 2013.
14. Y. Liu, C. Kliman-Silver, and A. Mislove. The tweets they are a-changin': Evolution of twitter users and behavior. In *Proceedings of the Eighth AAAI International Conference on Weblogs and Social Media (ICWSM)*, 2014.
15. E. Martínez-Cámara, M. T. Martín-Valdivia, L. A. Ureña-López, and A. R. Montejo-Ráez. Sentiment analysis in twitter. *Natural Language Engineering*, 20(01):1–28, 2014.
16. L. Mitchell, M. R. Frank, K. D. Harris, P. S. Dodds, and C. M. Danforth. The geography of happiness: Connecting twitter sentiment and expression, demographics, and objective characteristics of place. *PloS one*, 8(5):e64417, 2013.
17. S. M. Mohammad, S. Kiritchenko, and X. Zhu. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the Annual SemEval Workshop*, pages 321–327, 2013.
18. S. M. Mohammad and P. D. Turney. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 26–34. Association for Computational Linguistics, 2010.
19. P. Nakov, Z. Kozareva, A. Ritter, S. Rosenthal, V. Stoyanov, and T. Wilson. Semeval-2013 task 2: Sentiment analysis in twitter. *Proc. SemEval*, 2013.
20. D. Nguyen, R. Gravel, D. Trieschnigg, and T. Meder. "how old do you think i am?" a study of language and age in twitter. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 439–448, 2013.
21. B. Pang and L. Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual*

meeting on Association for Computational Linguistics, page 271. Association for Computational Linguistics, 2004.

22. B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics, 2002.
23. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
24. M. Pennacchiotti and A.-M. Popescu. A machine learning approach to twitter user classification. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 281–288, 2011.
25. J. P. Pestian, P. Matykiewicz, M. Linn-Gust, B. South, O. Uzuner, J. Wiebe, K. B. Cohen, J. Hurdle, and C. Brew. Sentiment analysis of suicide notes: A shared task. *Biomedical informatics insights*, 5(Suppl 1):3, 2012.
26. S. Rosenthal, P. Nakov, A. Ritter, and V. Stoyanov. Semeval-2014 task 9: Sentiment analysis in twitter. *Proc. SemEval*, 2014.
27. P. D. Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 417–424. Association for Computational Linguistics, 2002.
28. Twitter, Inc. Registration with United States securities and exchanges commission. `http://www.sec.gov/Archives/edgar/data/1418091/000119312513390321/d564001ds1.htm`, 2013.
29. S. Wang and C. D. Manning. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 90–94. Association for Computational Linguistics, 2012.
30. W. Wang, L. Chen, K. Thirunarayan, and A. P. Sheth. Harnessing Twitter "big data" for automatic emotion identification. In *2012 International Conference on Social Computing (SocialCom)*, pages 587–592. IEEE, 2012.
31. E. B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.
32. T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics, 2005.
33. N. Yu and S. Kubler. Semi-supervised learning for opinion detection. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, volume 3, pages 249–252. IEEE, 2010.