Sentiment Analysis: It's Complicated!

[1*]Kian Kenyon-Dean, [1*]Eisha Ahmed, [1†]Scott Fujimoto, [1†]Jeremy Georges-Filteau, [1†]Christopher Glasz, [1†]Barleen Kaur, [1†]Auguste Lalande, [1#]Shruti Bhanderi, [1#]Robert Belfer, [1#]Nirmal Kanagasabai, [1#]Roman Sarrazingendron, [1#]Rohit Verma, and [2]Derek Ruths

[1,2]McGill University, Department of Computer Science

[1] {first.last}@mail.mcgill.ca

[2] derek.ruths@mcgill.ca

Abstract

Sentiment analysis is used as a proxy to measure human emotion, where the objective is to categorize text according to some predefined notion of sentiment. Sentiment analysis datasets are typically constructed with gold-standard sentiment labels, assigned based on the results of manual annotations. When working with such annotations, it is common for dataset constructors to discard "noisy" or "controversial" data where there is significant disagreement on the proper label. In datasets constructed for the purpose of Twitter sentiment analysis (TSA), these controversial examples can compose over 30% of the originally annotated data. We argue that the removal of such data is a problematic trend because, when performing real-time sentiment classification of short-text, an automated system cannot know *a priori* which samples would fall into this category of disputed sentiment. We therefore propose the notion of a "complicated" class of sentiment to categorize such text, and argue that its inclusion in the short-text sentiment analysis framework will improve the quality of automated sentiment analysis systems as they are implemented in real-world settings. We motivate this argument by building and analyzing a new publicly available TSA dataset of over 7,000 tweets annotated with 5x coverage, named MTSA. Our analysis of classifier performance over our dataset offers insights into sentiment analysis dataset and model design, how current techniques would perform in the real world, and how researchers should handle difficult data.

1 Introduction

The goal of sentiment analysis is to determine the attitude or emotional state held by the author of a piece of text. Automatic sentiment classification that can quickly garner user sentiment is useful for applications ranging from product marketing to measuring public opinion. The volume and availability of short-text user content makes automated sentiment analysis systems highly attractive for companies and organizations, despite potential complications arising from their short length and specialized use of language. The popularity of Twitter as a social media platform on which people can readily express their thoughts, feelings, and opinions, coupled with the openness of the platform, provides a large amount of publicly accessible data ripe for analysis, being a well established domain for sentiment analysis as reflecting real-world attitudes (Pak and Paroubek, 2010; Bollen et al., 2011). In this paper, we look into Twitter sentiment analysis (TSA) as a suitable, core instance of general short-text sentiment analysis (Thelwall et al., 2010, 2012; Kiritchenko et al., 2014; Dos Santos and Gatti, 2014), and encourage the methods and practices presented to be applied across other domains.

Building a TSA model that can automatically

Tweet text	+	-	0
Members came in today for lunch to learn more about competitive events.	0	0	5
15 year old with an iPhone X, like DAMN girl, Whatcha gonna do with that much power in your hands? Facebook? Snapchat? That's it?	0	2	3
i am really missing the food my family makes rn	2	2	1

Table 1: Example tweets from our dataset over varying levels of annotator labellings; **+**, **-**, **0** stand for POSITIVE, NEGATIVE, OBJECTIVE.

[*]These authors contributed equally to this work.
[†]These authors contributed equally to this work.
[#]These authors contributed equally to this work.

determine the sentiment of a tweet has received significant attention over the past several years. However, since most state-of-the-art TSA models use machine learning to tune their parameters, their performance – and relevance to a real-world implementation setting – is highly dependent on the dataset on which they are trained.

TSA dataset construction has, unfortunately, received less attention than TSA model design. Many commonly used TSA datasets make assumptions that do not hold in a real-world implementation setting. For example, it is a common practice for studies to discard tweets on which there is high annotator disagreement. While some argue that this is done to remove noise resulting from poor annotator quality, this argument does not hold when considering that these datasets present high rates of unanimous annotator agreement[1]. This suggests that the problem is not poor annotators, but, rather, difficult data that does not fall into the established categories of sentiment.

Consider the sample tweets in Table 1 drawn from our dataset, one with unanimous agreement on an OBJECTIVE label, one with 60% agreement, and one with complete disagreement. We observe that, as the amount of disagreement across annotations increases, so too does the clarity of what the tweet's gold standard label really should be. Though the issues we raise may seem obvious, the absence of their proper treatment in the existing literature suggests the need to systematically consider their implications in sentiment analysis.

In this paper, we propose the inclusion of a COMPLICATED class of sentiment to indicate that the text does not fall into the established categories of sentiment. We offer insights into the differences between tweets that receive different levels of inter-annotator-agreement, providing empirical evidence that tweets with differing levels of agreement are qualitatively different from each other.

Our claims are supported by empirical analysis of a new TSA dataset, the McGill Twitter Sentiment Analysis dataset (MTSA), which we release publicly with this work[2]. The dataset contains 7,026 tweets across five different topic-domains, annotated with 5x coverage. We release this dataset with the raw annotation results, and hope that researchers and organizations will be able to analyze our dataset and build models that can be applied in real-world sentiment analysis settings.

2 Current Problems in TSA

The field of Twitter Sentiment Analysis (TSA) has seen a considerable productive work over the past several years, and several large reviews and surveys have been written to highlight the trends and progress of the field, its datasets, and the methods used for building automatic TSA systems (Saif et al., 2013; Medhat et al., 2014; Martínez-Cámara et al., 2014; Giachanou and Crestani, 2016).

There are a variety of methods for constructing TSA datasets along a variety of domains, ranging from very specific (e.g., OMD (Shamma et al., 2009)) to general (e.g., SemEval 2013-2014 (Nakov et al., 2016)). While there is the popular Stanford Twitter corpus, constructed with noisy labellings (Go et al., 2009), the more common method of constructing TSA datasets relies on manual annotation (usually crowd-sourced) of tweet sentiment to establish gold-standard labellings according to a pre-defined set of possible label categories (often POSITIVE, NEGATIVE, and NEUTRAL) (Shamma et al., 2009; Speriosu et al., 2011; Thelwall et al., 2012; Saif et al., 2013; Nakov et al., 2016; Rosenthal et al., 2017).

One of the earliest manually annotated TSA datasets, the Obama-McCain Debate (OMD) (Shamma et al., 2009) was released with the specific annotator votes for each tweet, rather than a final specific label assignment. Nonetheless, most work on this dataset filters out tweets with less than two-thirds agreement (Speriosu et al., 2011; Saif et al., 2013) (Table 2). Unfortunately, many later dataset releases have not followed the example of the OMD; the designers of such datasets have opted instead to release only the resultant labelling according to a motivated (but constraining) label-assignment schema, often removing tweets with high inter-annotator disagreement from the final dataset release (Saif et al., 2013; Nakov et al., 2016; Rosenthal et al., 2017).

The assumptions and implications resulting from such design choices should be carefully considered by researchers before deciding on how to construct or analyze sentiment analysis datasets. Indeed, a current limitation in the field is the lack of attention paid to label-assignment schemes, which ultimately determine the gold-standard labellings of samples. We argue that researchers

[1] Annotator disagreement information has proven useful in other areas of sentiment analysis (Wilson et al., 2005).

[2] Download at https://github.com/networkdynamics/mcgill-tsa

Name	# Annotated	Discarded	Coverage	Labels	Ref.
OMD	3,238	1,087 (33.6%)[3]	3x	**+**, **-**, MIXED, OTHER	(Shamma et al., 2009)
STS-Gold	3,000	794 (26.5%)	3x	**+**, **-**, **0**, MIXED, OTHER	(Saif et al., 2013)
SemEval 2013-14 (B)	17,048	Unknown	5x	**+**, **-**, **0**	(Nakov et al., 2016)
SemEval 2017	12,379	None (0%)	5x	**s+**, **w+**, **w-**, **s-**, **0**	(Rosenthal et al., 2017)

Table 2: Summary of some of the major TSA datasets used in recent work. Symbols **+**, **-**, **0** stand for POSITIVE, NEGATIVE, and NEUTRAL, respectively; prefixes **s**, **w** stand for STRONGLY and WEAKLY.

should consider whether or not the choices made during dataset construction adequately reflect a situation in which automatic sentiment analysis systems would be used in real-world settings.

2.1 General Trends in TSA Datasets

In the SemEval 2017 Task 4 (Rosenthal et al., 2017), a thorough 5x coverage annotation scheme is used (each tweet is annotated by at least five people). Annotations were made on a five-point scale, with categories STRONGLYNEGATIVE, WEAKLYNEGATIVE, NEUTRAL, WEAKLYPOSITIVE, and STRONGLYPOSITIVE. If at least three out of five of the annotators gave the same labelling, that was accepted as the final annotation. Otherwise, the authors used an averaging scheme (mapping the labels to integers $-2, -1, 0, 1, 2$) to determine the final label, taking the average of the labellings and rounding according to a specific criterion. This is highly problematic. For example, if a controversial tweet receives two STRONGLYNEGATIVE, two STRONGLYPOSITIVE, and one NEUTRAL labelling, it will have a resultant label of NEUTRAL. Yet, the tweet would certainly not be "neutral", it would be qualitatively different from a tweet with unanimous agreement on a NEUTRAL labelling. In Section 5, we provide empirical results supporting this claim, discovering that high-disagreement data is qualitatively different from high-agreement data.

Nakov et al. (2016) provide a thorough exploration into the specific design decisions and considerations made during the construction of the 2013-2014 SemEval shared task for short-text sentiment analysis. In Subtask B, annotators determined the overall polarity of a piece of text, according to a ternary labelling scheme between POSITIVE, NEGATIVE, or NEUTRAL. The final label of the sentence was "determined based on the majority of the labels" according to 5x coverage. The designers thus discarded sentences where there was no majority annotator agreement, since such sentences "are likely to be controversial cases" (p. 40); they do not report how much data was discarded.

Saif et al. (2013) constructed a new dataset, the STS-Gold, by taking into account several limitations of the TSA datasets they reviewed. In their study, 3,000 tweets were labelled with 3x coverage. Any tweet without unanimous agreement on the label was discarded; this decision was justified by the argument that they did not want "noisy data" in their dataset. Thus, they discarded 794 tweets, or 26.5% of their originally annotated data. While we argue that this is a problematic design decision, we note that discarding data in this way successfully isolated unanimous-agreement from majority-agreement data, thus avoiding conflating tweets with different levels of agreement, unlike in the 2013-14 and 2017 SemEval tasks.

The annotation scheme for the STS-Gold resolves one of the problems in the SemEval 2017 Task, as it provides an option for labelling a MIXED category, capturing tweets bearing multiple conflicting sentiments. It also provides the OTHER category for tweets where it is "difficult to decide on a proper label". Interestingly, the dichotomy between the high frequency of high-disagreement tweets (794 total) compared to the low frequency of tweets unanimously labelled as OTHER (4 total) is consistent with our findings on the COMPLICATED label (Section 3.3).

[3]Note that the entire OMD dataset was released with annotator votes, but most studies remove that proportion tweets where there was not at least two-thirds agreement on label.

The challenges and possible approaches to manual sentiment annotation have been previously discussed by Mohammad (2016), who offers important insights into how questions and problem descriptions should be posed to annotators.

2.2 Summary of TSA Problems

Based on analysis of the design choices of the three datasets described above, and on the thorough overview of other datasets found in (Saif et al., 2013), we conclude that there are two primary limitations in the standard TSA datasets.

First, the lack of distinction between data with majority- vs. unanimous-agreement on the annotated label (Nakov et al., 2016; Rosenthal et al., 2017). In the analysis (Sections 3.3 and 6) of our TSA dataset, we observe a clear qualitative difference between majority-agreement and unanimous-agreement data, suggesting that these sets of data should not necessarily be treated in the same way.

Second, the systematic removal of controversial (or, high-disagreement) data (Saif et al., 2013; Nakov et al., 2016). We argue that this tendency is problematic because any automatic sentiment analysis system to be implemented in a real-world setting cannot know *a priori* which tweets will be "noisy" or "controversial". An automatic sentiment analysis system trained on such a dataset will inevitably mislabel such tweets as they appear in a real-world implementation setting.

We therefore suggest that the following paradigm become the norm in the field: in releasing sentiment analysis datasets, *researchers should provide the specific annotations obtained for each sample* (as was done by Shamma et al. (2009)), in addition to the resultant labelling based on the label-assignment scheme they decide upon. Additionally, *data with high levels of annotator disagreement should not be discarded*, rather, it should be included in dataset releases.

3 Building the MTSA Dataset

The absence of a TSA dataset containing raw annotations and sufficient coverage to identify sources of annotator disagreement necessitated the creation of a new annotated dataset. Here, we provide an overview of the development of a new McGill TSA (MTSA) dataset composed of 7,026 tweets annotated with 5x coverage.

Topic	**Count**	**% of Total**
Sports	1752	24.9
Food	1729	24.6
Media	1697	24.2
Commercial Tech.	1353	19.3
General	495	7.0
Total	**7026**	100

Table 3: Distribution of annotated tweets by topic.

3.1 Data Collection

Tweets were collected from Twitter's streaming API, filtered for English tweets that contained at least one English token, that were posted by users in North American time-zones. Each tweet had to contain at least one keyword from a topic cloud relating to *Food* (example keywords: "weight", "breakfast", "protein"), *Media* ("cinema", "gameofthrones", "reggae"), *Commercial Technology* ("microsoft", "laptop", "iphone"), or *Sports* ("spurs", "hockey", "habs"). Using this topic cloud and a diverse set of keywords per topic (average of 38 hand-selected keywords per topic), we collected tweets with the intent to represent the general sentiment surrounding a specific topic, while reducing the bias that would result by relying on a single topic or keyword. A further subset of tweets (categorized as *General*) was collected from the stream, without any keyword filters, in order to further broaden the representative scope of our dataset.

We additionally filtered out tweets containing external links or images, arguing that analysis of these multimodal tweets is a separate problem, belonging to the domain of Multimodal Sentiment Analysis (Poria et al., 2016; Soleymani et al., 2017). After the entire filtering process[4] was complete, we obtained 7,026 tweets across the different topics, which would be annotated with 5x coverage. The distribution of these tweets is seen below in Table 3.

3.2 Data Annotation

Data annotation was crowd-sourced using the CrowdFlower platform[5]. All qualified CrowdFlower contributors had the opportunity to complete the task, which was presented as: carefully

[4]See supplemental material for full enumeration of the specific filters used and the keywords used for each topic.

[5]Website: https://www.crowdflower.com

read the tweet, determine whether or not it expresses sentiment (e.g., OBJECTIVE or not), if it does, categorize the sentiment as being either POSITIVE, NEGATIVE, or COMPLICATED. In the instructions, COMPLICATED was presented as the preferable option when the sentiment expressed in the tweet was ambiguous, mixed or could be interpreted as both positive and/or negative. After a one-line description of the meaning of each category, the contributor was presented with examples of tweets belonging in each category before starting the task.

In order to be considered qualified to complete the task, the contributor had to correctly answer at least 8 of 10 test questions, which we manually selected and labelled. When a user failed a test question, they were presented with the correct answer and a corresponding justification to ensure that they understood the task.

We experimented with the inclusion of test questions from the COMPLICATED category during screening, and found that this was a major source of protest among high-quality annotators. Indeed, it may be paradoxical to expect annotators to agree on tweets that cause significant disagreement. Furthermore, due to the heterogeneous nature of this class, such test questions would risk biasing the annotators' notion of the category. As such, we limited our test questions to OBJECTIVE, POSITIVE, and NEGATIVE tweets.

Users who successfully passed the initial test questions annotated a maximum of 400 tweets. Of those tweets, 10% were additional hidden test questions to continuously assess the quality of the annotators; an accuracy of at least 80% on these test questions was the threshold for including their annotations in the dataset. In the end, a total of 35,926 tasks were completed by 181 trusted contributors, resulting in 7,026 annotated tweets.

3.3 Dataset Analysis

The annotated tweets are categorized by four agreement levels: *Unanimous* (5 out of 5 agreed on the label), *Consensus* (exactly 4 out of 5 agreed), *Majority* (exactly 3 out of 5 agreed), or *Disputed* (maximum 2 out of 5 agreed). The distribution of agreement rates was consistent across topics (see supplemental material), thus the entire dataset is merged for the remainder of the analysis.

Annotator agreement distribution. Tweets with at least *Consensus* agreement compose 64% of the dataset (4505 tweets), and tweets with at least *Majority* agreement compose 92% of the dataset (6473 tweets; see Table 4). The decision to discard tweets with significant annotator disagreement, as previously done in TSA research, would result in the loss of 8% to 34% of the annotated tweets in our dataset, depending on whether to filter to a minimum *Majority* or *Consensus* agreement, respectively. Interestingly, these numbers are consistent with the proportion of discarded tweets in previous literature (Table 2).

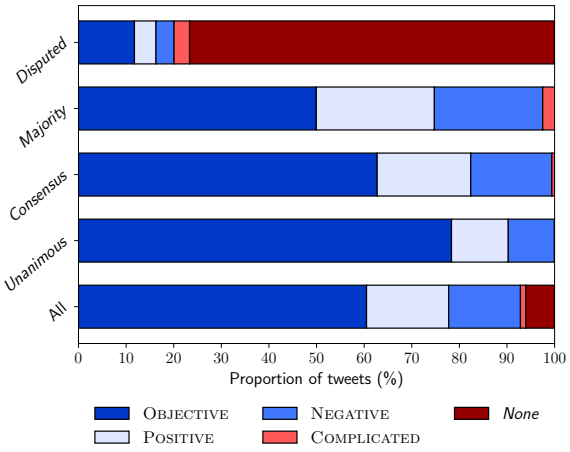

Figure 1: Most frequent annotation by annotator agreement rates. Of the 553 *Disputed* tweets, only 129 had a single most frequent annotation.

Sentiment and annotator agreement. Tweets that caused more disagreement among the human annotators were found to be more sentiment-laden (majority label of POSITIVE, NEGATIVE, or COMPLICATED; Figure 1). Objective tweets composed 78% (1892 tweets), 63% (1311), and 50% (983) of the *Unanimous*, *Consensus*, and *Majority* subsets of annotated tweets, respectively.

COMPLICATED label usage. Use of the COMPLICATED label by annotators was infrequent, and of those tweets with high inter-annotator agreement, almost exclusively limited to tweets that expressed clear, mixed sentiment. For example, the single tweet that received a unanimous COMPLICATED annotation had clear mixed sentiment: "the iPhone 6s is so big and hard to use but I still like it". There were a total of 13 tweets with at least *Consensus* agreement for the COMPLICATED label (see supplemental material). These specific tweets largely corresponded to the MIXED label used in previous TSA datasets (Shamma et al.,

Agreement	Count	% of Total
Unanimous	2415	34.4
Consensus	2090	29.7
Majority	1968	28.0
Disputed	553	7.9
Total	7026	100

Table 4: Annotator agreement rates. *Unanimous* stands for 100% annotator agreement, *Consensus* 80%, *Majority* 60%, and *Disputed* <60%.

Label	Count	% of Total
OBJECTIVE	4186	59.6
POSITIVE	1187	16.9
NEGATIVE	1038	14.8
COMPLICATED	62	0.9
Disputed	553	7.9
Total	7026	100

Table 5: Distribution of tweets across classes, where the label given is the result of majority vote.

2009; Saif et al., 2013). Other types of ambiguous tweets that did not clearly fall clearly within OBJECTIVE, POSITIVE, and NEGATIVE categories were not consistently identified as COMPLICATED by annotators. Rather, those tweets were a source of significant disagreement.

4 Classifying Tweet Sentiment

Here, we present the construction of shallow classifier and the experiments performed to study the phenomenon of annotator disagreement. Our objective was not to build a state-of-the-art classifier with optimal accuracy rates, rather, we sought to understand how the inclusion or exclusion of tweet subsets based on annotator disagreement impacts classification accuracy.

4.1 Preprocessing and Feature Extraction

To use machine learning methods with textual data, it is necessary to represent the data in a vector space such that each sample has the same dimensionality, despite varying sequence lengths. We concatenated three different standard feature extraction methods to build vector representations of tweets: N-Grams (unigrams and bigrams), mean word embedding (GLoVE embeddings built from twitter data (Pennington et al., 2014)[6]), and SentiWordNet scores (Baccianella et al., 2010).[7]

4.2 Experimental Design

As described in Section 2, most recent work in TSA has agglomerated tweets together based on the majority labelling. For example, a tweet annotated with a *Majority* agreement labelling (e.g., 3 OBJECTIVE and 2 NEGATIVE) would be given the label OBJECTIVE, just as one with *Unanimous* agreement on an OBJECTIVE labelling. In our experiments with our collected dataset (Section 3) we seek to determine whether or not there is a qualitative difference between high- versus low-agreement data.

Experiment I. In the first experiment setting, we agglomerate tweets according to the traditional practice for assigning labels based on annotations (Section 2); e.g., we remove tweets with at least a *Majority* voted label as COMPLICATED, and remove the *Disputed* tweets (that is, we remove 8.75% of our annotated data for these experiments), creating a 3-class classification problem. We experiment over four different sets of our data in this scenario: the full dataset (minus the COMPLICATED 8.75%); tweets with exactly *Majority* agreement; tweets with exactly *Consensus* agreement; and tweets with exactly *Unanimous* agreement on the label (see Figure 1 for the label distributions over each of these subsets). Additionally, when making predictions on a specific subset, we present results from training solely on the subset versus training on all of the data in this setting.

Experiment II. In the second experiment setting, we sought to determine the impact of including controversial samples, making a 4-class classification problem. Samples that were labelled with at least *Majority* agreement on a COMPLICATED label, and samples with *Disputed* agreement, were all assigned the label COMPLICATED. We thus used the entirety of our dataset for this experiment, where the COMPLICATED class accounted for 8.75% (615) of the samples, with the rest of the samples being given the majority-vote labelling.

Methods. For both experiments, we use a logistic regression classifier with balanced training set class weights, using the feature set described in

[6] https://nlp.stanford.edu/projects/glove/

[7] See supplemental material for full elaboration of the preprocessing decisions and features extracted.

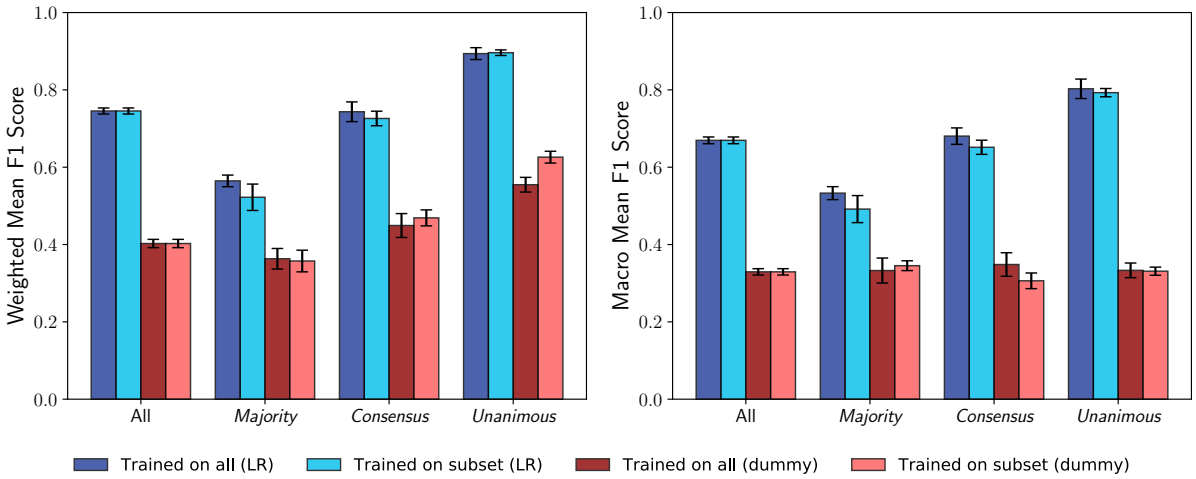

Figure 2: *Experiment I.* Weighted- and macro-F1-scores, obtained by testing logistic regression (LR) and a stratified random guesser (dummy) on the different agreement-level subsets, as described in Section 4.2. Black bars indicate plus/minus one standard deviation from the mean, as computed from the accuracies obtained across each of the 5 cross-validated folds.

Section 4.1. Preliminary experiments with feature ablation, whether or not to balance the train set classes, and different models (SVM with linear or RBF kernel, Random Forests, Naive Bayes, and K-Nearest-Neighbors), proved that this model variant was the best. We compare to a stratified random guesser, which predicts according to the distribution of classes in the training set (e.g., if 50% of the training set has samples labelled as OBJECTIVE, it will guess OBJECTIVE 50% of the time). To account for possible variance in the results, we use 5-fold cross validation over the full dataset, where the accuracy reported is the average over the specific scores obtained on each fold.

4.3 Evaluation

We evaluate with weighted- and macro-F1-scores to assess classifier performance. F1-score is a common way to measure classifier performance in sentiment analysis as it computes the harmonic mean between precision and recall. In multi-class classification, we obtain a one-versus-all F-score, F_c, for each class c in our set of possible classes, $\mathcal{C}$. Weighted F-score weights each F-score by its support in the test set; if there are n_c samples in the test set belonging to class c, then the weighted F-score is expressed by $\mathbf{F}_{weighted}$ in Equation 1.

$$\mathbf{F}_{weighted} = \frac{1}{(\sum n_c)} \sum_{c \in \mathcal{C}} n_c F_c \qquad (1)$$

Naturally, the weighted F-score is influenced by the frequency of samples in a class; so, in our case, it is biased toward the OBJECTIVE class due to its large frequency compared to the other classes (Table 5; Figure 1). Thus, we also report the macro F-score, which averages the F-scores for each class without considering their support, expressed by $\mathbf{F}_{macro}$ in Equation 2. This score evaluates model performance isolated from the class distribution, allowing us to determine if a change in accuracy is the result of simply a change in distribution of classes or a change in model generalization ability.

$$\mathbf{F}_{macro} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} F_c \qquad (2)$$

5 Results

In Figure 2, we present the results for Experiment I (Section 4.2). We note that the presented accuracy is higher when evaluated with weighted F-score versus macro F-score. Since both weighted- and macro-F1-score increase as we move along to higher agreement subsets, we conclude that the accuracy improvement is not solely due to a change in distribution of classes. Rather, there must be a qualitative difference between high- vs. low-agreement tweets, otherwise the accuracy would have been the same across agreement levels.

In Figure 3, we present the normalized confusion matrix obtained from Experiment II. We observe that the model poorly classifies COMPLICATED tweets. Although the model uses balanced class weights for training, it predicts OB-

Test subset	Trained on all			Trained on subset		
	Precision	*Recall*	*F1-Score*	*Precision*	*Recall*	*F1-Score*
All	0.681	0.660	0.669	–	–	–
Majority	0.552	0.524	**0.533**	0.502	0.488	0.491
Consensus	0.689	0.674	**0.680**	0.680	0.633	0.652
Unanimous	0.789	0.821	**0.803**	0.843	0.761	0.793

Table 6: *Experiment I. Macro*-F1 score results for precision, recall, and F1-score, as shown visually in Figure 2. These are the mean scores across the 5 cross-validated folds. Bold numbers indicate the improvement of training on all data versus just the subset being tested upon.

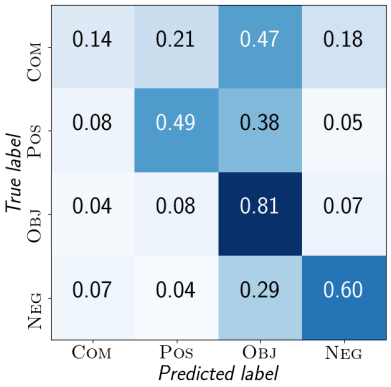

Figure 3: *Experiment II.* Normalized confusion matrix from using logistic regression.

Label	*Precision*	*Recall*	*F1*
COMPLICATED	0.199	0.138	0.163
POSITIVE	0.599	0.605	0.602
OBJECTIVE	0.766	0.806	0.786
NEGATIVE	0.510	0.487	0.498
Total – *weighted*	0.650	0.667	0.658
Total – *macro*	0.518	0.509	0.512

Table 7: *Experiment II.* Results across evaluation metrics, as shown visually in Figure 3. Results are computed by agglomerating all predictions made across each of the 5 cross-validated folds.

JECTIVE the majority of the time, where each other class is most frequently mistaken as OBJECTIVE. The final weighted-F1-score and macro-F1-scores, were, respectively: 65.8% and 51.2% with logistic regression, and 41.1% and 24.7% with the stratified random guesser. This large difference between weighted and macro is largely due to the poor classifier performance on the COMPLICATED class.

6 Discussion

The interpretation of expressed sentiment is an inherently subjective exercise, the gold-standard of which is the sentiment perceived by other humans. Thus, it is crucial to better understand sentiment annotation itself to inform future classifier design.

Annotator disagreement is not human error. Our results show that annotator disagreements cannot simply be attributed to human error. There is a clear decrease in classifier performance when testing on subsets of tweets with lower annotator agreement (Figure 2), suggesting that tweets across these subsets are qualitatively different from each other. From a probabilistic perspective, this means that samples that obtain high annotator agreement are generated by a different real-world function than those that obtain low annotator agreement. This perspective is further justified by the fact that classifier performance is roughly the same when training on the full dataset versus when training just on the specific agreement level subsets. Future work should explore how to handle this data, and we recommend reporting results on the different subsets by agreement-level.

On defaulting to the majority label. When each tweet is assigned a gold-standard label according to the majority annotation, we demonstrated that there are qualitative differences between tweets with *Majority*, *Consensus*, and *Unanimous* agreement. As exemplified by the sample tweets in Table 1, the differences between the two tweets with a majority OBJECTIVE annotation is reflected in the inter-annotator disagreement. We have shown that the subtleties in sentiment expression are masked by simply taking the majority label, and future work would involve factoring in these varying levels of agreement on labels during the model design process.

Standards for sentiment analysis datasets. To advance the field of short-text sentiment analysis, it is necessary to change common practices in dataset design and development. First, future datasets should be released with the raw annotator label assignments without discarding any annotated data. This would allow other researchers to experiment with different means of handling annotation-disagreement during the model design process. Secondly, we argue that sufficient resolution of short-text sentiment annotations requires at least 5x coverage. Our dataset, MTSA, of 7,026 tweets was constructed with 5x annotation coverage, a resolution at which we can just begin to distinguish these subsets of tweets. Higher coverage may be needed still to identify and understand these annotator disagreements. In contrast, the differences between these two subsets would be masked using the 3x coverage commonly found in other datasets.

Identifying ambiguous data. Results from Experiment II, and analysis of our COMPLICATED tweets, reveal that detecting high-disagreement tweets is a difficult task for both classifiers and humans. The poor performance of human annotators on identifying ambiguous tweets in our study, and the fact that high disagreement affected up to one third of the samples across TSA datasets, suggests that "complicatedness" is a real phenomenon. The optimal way to handle and identify this data requires further research. It is, however, an essential problem to solve, as real-world implementations of automated sentiment analysis systems will inevitably be confronted with such data. Such a system may be able to leverage the raw annotations during training, which is why we release the MTSA dataset with the raw annotation results included, and suggest all others do this as well.

7 Conclusion

In this paper, we highlight the need to better engage with how humans actually annotate data in short-text sentiment analysis dataset construction by constructing the new McGill Twitter Sentiment Analysis (MTSA) dataset. Future work involves leveraging raw human annotations to improve sentiment analysis classifiers, and finding ways to better detect and understand the "complicated" property in these samples that cause high annotator disagreement. Additionally, we encourage researchers to use MTSA in the development of other methods for short text sentiment analysis, including unsupervised, lexicon-based, and rule-based methods.

Acknowledgements

This work was the product of a class project pursued collectively by students in the COMP 767 graduate seminar in Social Media Analytics at McGill University, taught by Derek Ruths. This work was funded by the Discovery Grant Accelerator Supplement 2017-05165.

References

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*. volume 10, pages 2200–2204.

Johan Bollen, Huina Mao, and Alberto Pepe. 2011. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *ICWSM* 11:450–453.

Cícero Nogueira Dos Santos and Maira Gatti. 2014. Deep convolutional neural networks for sentiment analysis of short texts. In *COLING*. pages 69–78.

Anastasia Giachanou and Fabio Crestani. 2016. Like it or not: A survey of twitter sentiment analysis methods. *ACM Computing Surveys (CSUR)* 49(2):28.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford* 1(2009):12.

Svetlana Kiritchenko, Xiaodan Zhu, and Saif M Mohammad. 2014. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research* 50:723–762.

Eugenio Martínez-Cámara, M Teresa Martín-Valdivia, L Alfonso Urena-López, and A Rturo Montejo-Ráez. 2014. Sentiment analysis in twitter. *Natural Language Engineering* 20(1):1–28.

Walaa Medhat, Ahmed Hassan, and Hoda Korashy. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal* 5(4):1093–1113.

Saif Mohammad. 2016. A practical guide to sentiment annotation: Challenges and solutions. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. pages 174–179.

Preslav Nakov, Sara Rosenthal, Svetlana Kiritchenko, Saif M Mohammad, Zornitsa Kozareva, Alan Ritter, Veselin Stoyanov, and Xiaodan Zhu. 2016. Developing a successful semeval task in sentiment analysis of twitter and other social media texts. *Language Resources and Evaluation* 50(1):35–65.

Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*. volume 10.

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.

Soujanya Poria, Erik Cambria, Newton Howard, Guang-Bin Huang, and Amir Hussain. 2016. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing* 174:50–59.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. pages 502–518.

Hassan Saif, Miriam Fernandez, Yulan He, and Harith Alani. 2013. Evaluation datasets for Twitter sentiment analysis: a survey and a new dataset, the STS-Gold. In *CEUR Workshop Proceedings*. volume 1096, pages 9–21.

David A Shamma, Lyndon Kennedy, and Elizabeth F Churchill. 2009. Tweet the debates: understanding community annotation of uncollected sources. In *Proceedings of the first SIGMM workshop on Social media*. ACM, pages 3–10.

Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. 2017. A survey of multimodal sentiment analysis. *Image and Vision Computing* 65:3–14.

Michael Speriosu, Nikita Sudan, Sid Upadhyay, and Jason Baldridge. 2011. Twitter polarity classification with label propagation over lexical links and the follower graph. In *Proceedings of the First workshop on Unsupervised Learning in NLP*. Association for Computational Linguistics, pages 53–63.

Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. 2012. Sentiment strength detection for the social web. *Journal of the Association for Information Science and Technology* 63(1):163–173.

Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. 2010. Sentiment strength detection in short informal text. *Journal of the Association for Information Science and Technology* 61(12):2544–2558.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*. Association for Computational Linguistics, pages 347–354.