Evaluation Datasets for Twitter Sentiment Analysis. A survey and a new dataset, the STS-Gold

Conference Paper · December 2013

CITATIONS
94

READS
1,659

4 authors, including:

Hassan Saif
The Open University (UK)
24 PUBLICATIONS **688** CITATIONS
SEE PROFILE

Harith Alani
The Open University (UK)
206 PUBLICATIONS **4,454** CITATIONS
SEE PROFILE

Evaluation Datasets for Twitter Sentiment Analysis

A survey and a new dataset, the STS-Gold

Hassan Saif[1], Miriam Fernandez[1], Yulan He[2] and Harith Alani[1]

[1] Knowledge Media Institute, The Open University, United Kingdom
{h.saif, m.fernandez, h.alani}@open.ac.uk
[2] School of Engineering and Applied Science, Aston University, UK
y.he@cantab.net

Abstract. Sentiment analysis over Twitter offers organisations and individuals a fast and effective way to monitor the publics' feelings towards them and their competitors. To assess the performance of sentiment analysis methods over Twitter a small set of evaluation datasets have been released in the last few years. In this paper we present an overview of eight publicly available and manually annotated evaluation datasets for Twitter sentiment analysis. Based on this review, we show that a common limitation of most of these datasets, when assessing sentiment analysis at target (entity) level, is the lack of distinctive sentiment annotations among the tweets and the entities contained in them. For example, the tweet "I love iPhone, but I hate iPad" can be annotated with a mixed sentiment label, but the entity iPhone within this tweet should be annotated with a positive sentiment label. Aiming to overcome this limitation, and to complement current evaluation datasets, we present STS-Gold, a new evaluation dataset where tweets and targets (entities) are annotated individually and therefore may present different sentiment labels. This paper also provides a comparative study of the various datasets along several dimensions including: total number of tweets, vocabulary size and sparsity. We also investigate the pair-wise correlation among these dimensions as well as their correlations to the sentiment classification performance on different datasets.

Keywords: Sentiment Analysis, Twitter, Datasets

1 Introduction

With the emergence of social media, the performance of sentiment analysis tools has become increasingly critical. In the current commercial competition, designers, developers, vendors and sales representatives of new information products need to carefully study whether and how do their products offer competitive advantages. Twitter, with over 500 million registered users and over 400 million messages per day,[3] has become a gold mine for organisations to monitor their reputation and

[3] http://www.alexa.com/topsites

brands by extracting and analysing the sentiment of the tweets posted by the public about them, their markets, and competitors.

Developing accurate sentiment analysis methods requires the creation of evaluation datasets that can be used to assess their performances. In the last few years several evaluation datasets for Twitter sentiment analysis have been made publicly available. The general evaluation dataset consists of a set of tweets, where each tweet is annotated with a sentiment label [1,8,16,22]. The most common sentiment labels are *positive*, *negative* and *neutral*, but some evaluation datasets consider additional sentiment labels such as *mixed*, *other* or *irrelevant* [1,23]. Instead of the final sentiment labels associated to the tweets, some datasets provide a numeric sentiment strength between -5 and 5 defining a range from negative to positive polarity [24,25]. In addition to sentiment labels associated to the tweets some evaluation datasets also provide sentiment labels associated to targets (entities) within the tweets. However, these datasets do not distinguish between the sentiment label of the tweet and the sentiment labels of the entities contained within it [23]. For example, the tweet "`iPhone 5 is awesome, but I can't upgrade :(`" presents a negative sentiment. However, the entity "`iPhone 5`" should receive a positive sentiment.

Aiming to overcome this limitation, we present STS-Gold, an evaluation dataset for Twitter sentiment analysis that targets sentiment annotation at both, tweet and entity levels. The annotation process allows a dissimilar polarity annotation between the tweet and the entities contained within it. To create this dataset a subset of tweets was selected from the Standford Twitter Sentiment Corpus [8] and entities were extracted from this subset of tweets by using a third-party entity extraction tool. Tweets and entities were manually annotated by three different human evaluators. The final evaluation dataset contains 2,206 tweets and 58 entities with associated sentiment labels. The purpose of this dataset is therefore to complement current state of the art datasets by providing entity sentiment labels, therefore supporting the evaluation of sentiment classification models at entity as well as tweet level.

Along with the description of the STS-Gold dataset, this paper summarises eight publicly available and manually annotated evaluation datasets for Twitter sentiment analysis. Our goal is to provide the reader with an overview of the existing evaluation datasets and their characteristics. To this aim, we provide a comparison of these datasets along different dimensions including: the total number of tweets, the vocabulary size and the degree of data sparsity. We also investigate the pair-wise correlation among these dimensions as well as their correlations to the sentiment classification performance on all datasets. Our study shows that the correlation between the sparsity and the classification performance is intrinsic, meaning that it might exists within the dataset itself, but not necessarily across the datasets. We also show that the correlations between sparsity, vocabulary size and number of tweets are all strong. However, the large number of tweets in a dataset is not always an indication for a large vocabulary size or a high sparsity degree.

The rest of the paper is structured as follows: Section2 presents an overview of the existing evaluation datasets for Twitter sentiment analysis. Section 3 describes STS-Gold, our proposed evaluation dataset. Section 4 presents a comparison study across the evaluation datasets. We conclude the paper in Section 5.

2 Twitter Sentiment Analysis Datasets

In this section we present 8 different datasets widely used in the Twitter sentiment analysis literature. We have focused our selection on those datasets that are: (i) publicly available to the research community, (ii) manually annotated, providing a reliable set of judgements over the tweets and, (iii) used to evaluate several sentiment sentiment analysis models. Tweets in these datasets have been annotated with different sentiment labels including: *Negative*, *Neutral*, *Positive*, *Mixed*, *Other* and *Irrelevant*. Table 1 displays the distribution of tweets in the eight selected datasets according to these sentiment labels.

Variations of the evaluation datasets are due to the particularities of the different sentiment analysis tasks. Sentiment analysis on Twitter spans multiple tasks, such as polarity detection (positive vs. negative), subjectivity detection (polar vs. neutral) or sentiment strength detection. These tasks can also be performed either at tweet level or at target (entity) level. In the following subsections, we provide an overview of the available evaluation datasets and the different sentiment tasks for which they are used.

Dataset	No. of Tweets	#Negative	#Neutral	#Positive	#Mixed	#Other	#Irrelevant
STS-Test	498	177	139	182	-	-	-
HCR	2,516	1,381	470	541	-	45	79
OMD	3,238	1,196	-	710	245	1,087	-
SS-Twitter	4,242	1,037	1,953	1,252	-	-	-
Sanders	5,513	654	2,503	570	-	-	1,786
GASP	12,771	5,235	6,268	1,050	-	218	-
WAB	13,340	2,580	3,707	2,915	-	420	3,718
SemEval	13,975	2,186	6,440	5,349	-	-	-

Table 1. Total number of tweets and the tweet sentiment distribution in all datasets

Stanford Twitter Sentiment Test Set (STS-Test)

The Stanford Twitter sentiment corpus (`http://help.sentiment140.com/`), introduced by Go et al. [8] consists of two different sets, training and test. The training set contains 1.6 million tweets automatically labelled as positive or negative based on emotions. For example, a tweet is labelled as positive if it contains :), :-), :), :D, or =) and is labelled as negative if it contains :(, :-(, or : (. Although automatic sentiment annotation of tweets using emoticons is fast, its accuracy is arguable because emoticons might not reflect the actual sentiment of tweets. In this study, we focus on those datasets that have been manually annotated. Therefore, although we acknowledge the relevance of the STS training dataset for building sentiment analysis models, we discard it from the rest of our study.

The test set (STS-Test), on the other hand, is manually annotated and contains 177 negative, 182 positive and 139 neutrals tweets. These tweets were

collected by searching Twitter API with specific queries including names of products, companies and people. Although the STS-Test dataset is relatively small, it has been widely used in the literature in different evaluation tasks. For example, Go et al. [8], Saif et al. [19,20], Speriosu et al. [23], and Bakliwal et al. [2] use it to evaluate their models for polarity classification (positive vs. negative). In addition to polarity classification, Marquez et al. [3] use this dataset for evaluating subjectivity classification (neutral vs. polar).

Health Care Reform (HCR)

The Health Care Reform (HCR) dataset was built by crawling tweets containing the hashtag "#hcr" (health care reform) in March 2010 [23]. A subset of this corpus was manually annotated by the authors with 5 labels (*positive, negative, neutral, irrelevant, unsure(other)*) and split into training (839 tweets), development (838 tweets) and test (839 tweets) sets. The authors also assigned sentiment labels to 8 different targets extracted from all the three sets (*Health Care Reform, Obama, Democrats, Republicans, Tea Party, Conservatives, Liberals, and Stupak*). However, both the tweet and the targets within it, were assigned the same sentiment label, as can be found in the published version of this dataset (`https://bitbucket.org/speriosu/updown`). In this paper, we consider all the three subsets (training, development and test) as one unique dataset for the analysis (see Section 4). The final datasets, as shown in Table 1, consists of 2,516 tweets including 1,381 negative, 470 neutral and 541 positive tweets.

The HCR dataset has been used to evaluate polarity classification [23,21] but can also be used to evaluate subjectivity classification since it identifies neutral tweets.

Obama-McCain Debate (OMD)

The Obama-McCain Debate (OMD) dataset was constructed from 3,238 tweets crawled during the first U.S. presidential TV debate in September 2008 [22]. Sentiment labels were acquired for these tweets using Amazon Mechanical Turk, where each tweet was rated by at least three annotators as either *positive, negative, mixed*, or *other*. The authors in [6] reported an inter-annotator agreement of 0.655, which shows a relatively good agreement between annotators. The dataset is provided at `https://bitbucket.org/speriosu/updown` along with the annotators' votes on each tweet. We considered those sentiment labels, which two-third of the voters agree on, as final labels of the tweets. This resulted in a set of 1,196 negative, 710 positive and 245 mixed tweets.

The OMD dataset is a popular dataset, which has been used to evaluate various supervised learning methods [10,23,21], as well as unsupervised methods [9] for polarity classification of tweets. Tweets' sentiments in this dataset were also used to characterize the Obama-McCain debate event in 2008 [6].

Sentiment Strength Twitter Dataset (SS-Tweet)

This dataset consists of 4,242 tweets manually labelled with their positive and negative sentiment strengths. i.e., a negative strength is a number between -1 (not negative) and -5 (extremely negative). Similarly, a positive strength is a

number between 1 (not positive) and 5 (extremely positive). The dataset was constructed by [24] to evaluate SentiStrenth (`http://sentistrength.wlv.ac.uk/`), a lexicon-based method for sentiment strength detection.

In this paper we propose re-annotating tweets in this dataset with sentiment labels (negative, positive, neutral) rather than sentiment strengths, which will allow using this dataset for subjectivity classification in addition to sentiment strength detection. To this end, we assign a single sentiment label to each tweet based on the following two rules inspired by the way SentiStrength works:[4] (i) a tweet is considered neutral if the absolute value of the tweet's negative to positive strength ratio is equals to 1, (ii) a tweet is positive if its positive sentiment strength is 1.5 times higher than the negative one, and negative otherwise. The final dataset, as shown in table 1, consists of 1,037 negative, 1,953 neutral and 1,252 positive tweets.

The original dataset is publicly available at `http://sentistrength.wlv.ac.uk/documentation/` along with other 5 datasets from different social media platforms including MySpace, Digg, BBC forum, Runners World forum, and YouTube.

Sanders Twitter Dataset

The Sanders dataset consists of 5,512 tweets on four different topics (Apple, Google, Microsoft, Twitter). Each tweet was manually labelled by one annotator as either *positive, negative, neutral*, or *irrelevant* with respect to the topic. The annotation process resulted in 654 negative, 2,503 neutral, 570 positive and 1,786 irrelevant tweets.

The dataset has been used in [3,12,5] for polarity and subjectivity classification of tweets.

The Sanders dataset is available at `http://www.sananalytics.com/lab`

The Dialogue Earth Twitter Corpus

The Dialogue Earth Twitter corpus consists of three subsets of tweets. The first two sets (WA, WB) contain 4,490 and 8,850 tweets respectively about the weather, while the third set (GASP) contains 12,770 tweets about gas prices. These datasets were constructed as a part of the Dialogue Earth Project[5] (`www.dialogueearth.org`) and were hand labelled by several annotators with five labels: *positive, negative, neutral, not related* and *can't tell (other)*. In this work we merge the two sets about the weather in one dataset (WAB) for our analysis study in Section 4. This results in 13,340 tweets with 2,580 negative, 3,707 neutral, and 2,915 positive tweets. The GASP dataset on the other hand consists of 5,235 negative, 6,268 neutral and 1,050 positive tweets.

The WAB and the GASP datasets have been used to evaluate several machine learning classifiers (e.g., Naive Bayes, SVM, KNN) for polarity classification of tweets [1].

[4] `http://sentistrength.wlv.ac.uk/documentation/SentiStrengthJavaManual.doc`

[5] Dialogue Earth, is former program of the Institute on the Environment at the University of Minnesota

SemEval-2013 Dataset (SemEval)

This dataset was constructed for the Twitter sentiment analysis task (Task 2) [16] in the Semantic Evaluation of Systems challenge (SemEval-2013).[6] The original SemEval dataset consists of 20K tweets split into training, development and test sets. All the tweets were manually annotated by 5 Amazon Mechanical Turk workers with negative, positive and neutral labels. The turkers were also asked to annotate expressions within the tweets as subjective or objective. Using a list of the dataset's tweet ids provided by [16], we managed to retrieve 13,975 tweets with 2,186 negative, 6,440 neutrals and 5,349 positives tweets.

Participants in the SemEval-2013 Task 2 used this dataset to evaluate their systems for expression-level subjectivity detection[15,4], as well as tweet-level subjectivity detection[14,18].

Summary: Based on the above reviews we can identify two main shortcomings of these datasets when using them to assess the performance of Twitter sentiment analysis models. The first shortcoming is the lack of specifications provided by some datasets (e.g., STS-Test, HCR, Sanders) about the annotation methodology used to assign sentiment labels to the tweets. For example [8] do not report the number of annotators. Similarly [23] do not report annotation agreement among annotators. The second shortcoming is that most of these datasets are focused on assessing the performance of sentiment analysis models working at tweet level but not at entity level (i.e., they provide human annotations for tweets but not for entities). In the few cases where the annotation process also targets entities as in the HCR dataset, these entities are assigned similar sentiment labels to the label of the tweet they belong to. Entity sentiment analysis is however a highly relevant task, since it is closely related to the problem of mining the reputation of individuals and brands in Twitter.

3 STS-Gold Dataset

In the following subsections we described our proposed dataset, STS-Gold. The goal of this dataset is to complement existing Twitter sentiment analysis evaluation datasets by providing a new dataset where tweets and entities are annotated independently, allowing for different sentiment labels between the tweet and the entities contained within it. The purpose is to support the performance assessment for entity-based sentiment analysis models, which is currently hardly addressed in the datasets that have been released to date (see Section 2).

3.1 Data Acquisition

To construct this dataset, we first extracted all named entities from a collection of 180K tweets randomly selected from the original Stanford Twitter corpus (see Section 2). To this end, we used AlchemyAPI,[7] an online service that allows for the extraction of entities from text along with their associated semantic concept class (e.g., Person, Company, City). After that, we identified the top most frequent semantic concepts and, selected under each of them, the top 2

[6] `http://www.cs.york.ac.uk/semeval-2013/task2/`

[7] `www.alchemyapi.com`

most frequent and 2 mid-frequent entities. For example, for the semantic concept *Person* we selected the top most frequent entities (Taylor Swift and Obama) as well as two mid frequent entities (Oprah and Lebron). This resulted in 28 different entities along with their 7 associated concepts as shown in Table 2.

Concept	**Top 2 Entities**	**Mid 2 Entities**
Person	Taylor Swift, Obama	Oprah, Lebron
Company	Facebook, Youtube	Starbucks, McDonalds
City	London, Vegas	Sydney, Seattle
Country	England, US	Brazil, Scotland
Organisation	Lakers, Cavs	Nasa, UN
Technology	iPhone, iPod	Xbox, PSP
HealthCondition	Headache, Flu	Cancer, Fever

Table 2. 28 Entities, with their semantic concepts, used to build STS-Gold.

The next step was to construct and prepare a collection of tweets for sentiment annotation, ensuring that each tweet in the collection contains one or more of the 28 entities listed in Table 2. To this aim, we randomly selected 100 tweets from the remaining part of the STS corpus for each of the 28 entities, i.e., a total of 2,800 tweets. We further added another 200 tweets without specific reference to any entities to add up a total of 3,000 tweets. Afterwards, we applied AlchemyAPI on the selected 3,000 tweets. Apart from the initial 28 entities the extraction tool returned 119 additional entities, providing a total of 147 entities for the 3,000 selected tweets.

3.2 Data Annotation

We asked three graduate students to manually label each of the 3,000 tweets with one of the five classes: (`Negative, Positive, Neutral, Mixed` and `Other`). The "`Mixed`" label was assigned to tweets containing mixed sentiment and "`Other`" to those that were difficult to decide on a proper label. The students were also asked to annotate each entity contained in a tweet with the same five sentiment classes. The students were provided with a booklet explaining both the tweet-level and the entity-level annotation tasks. The booklet also contains a list of key instructions as shown in this paper's appendix. It is worth noting that the annotation was done using Tweenator,[8] an online tool that we previously built to annotate tweet messages [20].

We measured the inter-annotation agreement using the Krippendorff's alpha metric [11], obtaining an agreement of $\alpha_t = 0.765$ for the tweet-level annotation task. For the entity-level annotation task, if we measured sentiment of entity for each individual tweet, we only obtained $\alpha_e = 0.416$ which is relatively low for the annotated data to be used. However, if we measured the aggregated sentiment for each entity, we got a very high inter-annotator agreement of $\alpha_e = 0.964$.

To construct the final STS-Gold dataset we selected those tweets and entities for which our three annotators agreed on the sentiment labels, discarding any

[8] `http://tweenator.com`

possible noisy data from the constructed dataset. As shown in Table 3 the STS-Gold dataset contains 13 negative, 27 positive and 18 neutral entities as well as 1,402 negative, 632 positive and 77 neutral tweets. The STS-Gold dataset contains independent sentiment labels for tweets and entities, supporting the evaluation of tweet-based as well as entity-based Twitter sentiment analysis models.

Class	Negative	Positive	Neutral	Mixed	Other
No. of Entities	13	27	18	-	-
No. of Tweets	1402	632	77	90	4

Table 3. Number of tweets and entities under each class

4 Comparative study of Twitter Sentiment Analysis Datasets

In this section, we present a comparison of the described datasets according to three different dimensions: the vocabulary size, the total number of tweets, and the data sparsity. We also study the pair-wise intrinsic correlation between these dimensions as well as their correlation with the sentiment classification performance (correlation are computed using the Pearson correlation coefficient). To this end, we perform a binary sentiment classification (positive vs. negative) on all the datasets using a Maximum Entropy classifier (MaxEnt). Note that no stemming or filtering was applied to the data since our aim by providing this comparison is not to build better classifiers. Instead, we aim at showing the particularities of each dataset and how these particularities may affect the performance of sentiment classifiers.

Vocabulary Size

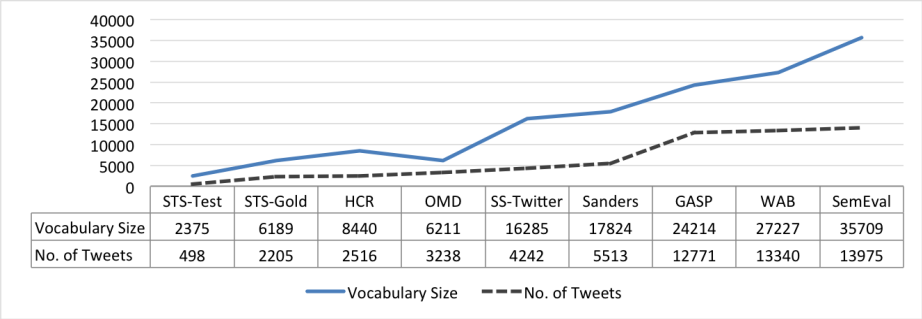
The vocabulary size of a dataset is commonly determined by the number of the unique word unigrams that the dataset contains. To extract the number of unigrams, we use the TweetNLP tokenizer [7], which is specifically built to work on tweets data.[9] Note that we considered all tokens found in the tweets including words, numbers, URLs, emoticons, and speical characters (e.g., question marks, intensifiers, hashtags, etc).

Figure 1 depicts the correlation between the the vocabulary size and the total number of tweets in the datasets. Although the correlation between the two quantities seems to be positively strong ($\rho = 0.95$), increasing the number of tweets does not always lead to increasing the vocabulary size. For example, the OMD dataset has higher number of tweets than the HCR dataset, yet the former has a smaller vocabulary size than the latter.

Data Sparsity

Dataset sparsity is an important factor that affects the overall performance of typical machine learning classifiers [17]. According to Saif et al. [20], tweets data

[9] The TweetNLP tokenizer can be downloaded from `http://www.ark.cs.cmu.edu/TweetNLP/`



	STS-Test	STS-Gold	HCR	OMD	SS-Twitter	Sanders	GASP	WAB	SemEval
Vocabulary Size	2375	6189	8440	6211	16285	17824	24214	27227	35709
No. of Tweets	498	2205	2516	3238	4242	5513	12771	13340	13975

Fig. 1. Total number of tweets and the vocabulary size of each dataset.

are sparser than other types of data (e.g., movie review data) due to a large number of infrequent words in tweets.

In this section, we aim to compare the presented datasets with respect to their sparsity. To calculate the sparsity degree of a given dataset we use the following formula from [13]:

$$S_d = 1 - \frac{\sum_i^n N_i}{n \times |V|} \tag{1}$$

Where N_i is the the number of distinct words in tweet i, n is the number of tweets in the dataset and $|V|$ the vocabulary size.

According to Figure 2, all datasets have a high sparsity degree, with SemEval being the sparsest. It is also worth noticing that there is a strong correlation between the sparsity degree and the total number of tweets in a dataset ($\rho = 0.71$) and an even stronger correlation between the sparsity degree and the vocabulary size of the dataset ($\rho = 0.77$).

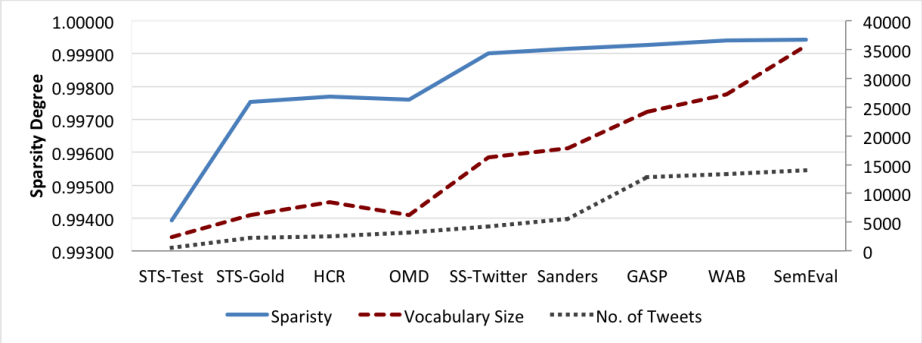


Fig. 2. Sparsity degree, vocabulary size and the total number of tweets across the datasets

Classification Performance

We perform a binary sentiment classification on all the datasets using a MaxEnt classifier from Mallet.[10] To this end, we selected for each dataset only the subset of positive and negative tweets.

Table 4 reports the classification results (using 10-fold cross validation) in accuracy and the average F-measure (F-average) on all datasets. The highest accuracy is achieved on the GASP dataset with 90.897%, while the highest average F-measure of 84.621% is obtained on the WAB dataset. It is also worth noticing that the per-class performance is highly affected by the distribution of positive and negative tweets in the dataset. For example, F-measure for detecting positive tweets (F-positive) is higher than F-measure for detecting negative tweets (F-negative) for positive datasets (i.e., datasets that have higher number of positive tweets than negative ones) such as STS-Test, SS-Twitter, WAB and SemEval. Similarly, F-negative score is higher than F-positive for negative datasets (i.e., datasets that have higher number of negative tweets than positive ones). However, the average accuracy for negative datasets is 84.53%, while it is 80.37% for positive tweets, suggesting that detecting positive tweets is more difficult than detecting negative tweets.

Dataset	STS-Test	STS-Gold	HCR	OMD	SS-Twitter	Sanders	GASP	WAB	SemEval
Accuracy	80.171	85.69	78.679	82.661	73.399	83.84	**90.897**	84.668	83.257
F-negative	79.405	89.999	85.698	86.617	69.179	84.964	94.617	83.745	68.668
F-positive	81.21	74.909	58.23	75.47	76.621	82.548	70.682	85.498	88.578
F-average	80.307	82.454	71.964	81.044	72.9	83.756	82.65	**84.621**	78.623

Table 4. Accuracy and the average harmonic mean (F measure) obtained from identifying positive and negative sentiment.

Makrehchi and Kamel [13] showed that the performance trend of text classifiers can be estimated using the sparsity degree of the dataset. In particular, they found that reducing the sparsity of a given dataset enhances the performance of a SVM classifier. Their observation is based on changing the sparsity degree of the same dataset by removing/keeping specific terms.

Figure 3 illustrates the correlation across all datasets between Accuracy and F-measure on the one hand, and the dataset sparsity on the other hand. As illustrated by this figure, there is almost no correlation ($\rho_{acc} = -0.06$, $\rho_{f1} = 0.23$) between the classification performance and the sparsity degree across the datasets. In other words, the sparsity-performance correlation is intrinsic, meaning that it might exists within the dataset itself, but not necessarily across the datasets. This is not surprising given that there are other dataset characteristics in addition to data sparsity, such as polarity class distribution, which may also affect the overall performance as we discussed earlier in this section.

[10] `http://mallet.cs.umass.edu/`

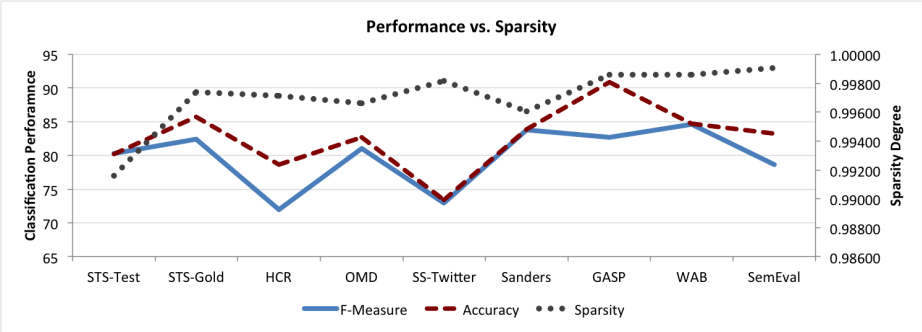


Fig. 3. F-Measure and the Sparsity degree of the datasets

5 Conclusions

In this paper, we provided an overview of eight publicly available and manually annotated evaluation datasets for Twitter sentiment analysis. Based on our review, we found that unlike the tweet level, very few annotation efforts were spent towards providing datasets for evaluating sentiment classifiers at the entity level. This motivated us to build a new evaluation dataset, STS-Gold, which allows for the evaluation of sentiment classification models at both the entity and the tweet levels. Our dataset, unlike most of the other datasets, distinguishes between the sentiment of a tweet and the sentiment of entities mentioned within it.

We also provided a comparative study across all the reported datasets in terms of different characteristics including the vocabulary size, the total number of tweets and the degree of sparsity. Finally, we studied the various pair-wise correlations among these characteristics as well as the correlation between the data sparsity degree and the sentiment classification performance across the datasets. Our study showed that the large number of tweets in a dataset is not always an indication for a large vocabulary size although the correlation between these two characteristics is relatively strong. We also showed that the sparsity-performance correlation is intrinsic, where it might exists within the dataset itself, but not necessarily across the datasets.

Acknowledgment

The work of the authors was supported by the EU-FP7 projects: ROBUST (grant no. 257859) and SENSE4US (grant no. 611242).

Appendix: Annotation Booklet

We need to manually annotate 3000 tweets with their sentiment label (Negative, Positive, Neutral, Mixed) using the online annotation tool "Tweenator.com". The task consists of two subtasks:

Task A. Tweet-Level Sentiment Annotation Given a tweet message, decide weather it has a positive, negative, neutral or mixed sentiment.

Task B. Entity-Level Sentiment Annotation Given a tweet message and a named entity, decided weather the entity received a negative, positive or neutral sentiment. The named entities to annotate are highlighted in yellow within the tweets.

Please note that:

- A Tweet could have a different sentiment from an entity within it. For example, the tweet "`iPhone 5 is very nice phone, but I can't upgrade :(`" has a negative sentiment. However, the entity "`iPhone 5`" receives a positive sentiment.
- "Mixed" label refers to a tweet that has mixed sentiment. For example, the "`Kobe is the best in the world not Lebron`" has a mixed sentiment.
- Some tweets might have emoticons such as :), :-), :(, or :-(. Please give less attention to the emoticons and focus more on the content of the tweets. Emoticons can be very misleading indicators sometimes.
- Try to be objective with your judgement and feel free to take a break whenever you feel tired or bored.

References

1. Asiaee T, A., Tepper, M., Banerjee, A., Sapiro, G.: If you are happy and you know it... tweet. In: Proceedings of the 21st ACM international conference on Information and knowledge management. pp. 1602–1606. ACM (2012)
2. Bakliwal, A., Arora, P., Madhappan, S., Kapre, N., Singh, M., Varma, V.: Mining sentiments from tweets. Proceedings of the WASSA 12 (2012)
3. Bravo-Marquez, F., Mendoza, M., Poblete, B.: Combining strengths, emotions and polarities for boosting twitter sentiment analysis. In: Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining. ACM (2013)
4. Chalothorn, T., Ellman, J.: Tjp: Using twitter to analyze the polarity of contexts. In: In Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013), Atlanta, Georgia, USA, June 2013. (2013)
5. Deitrick, W., Hu, W.: Mutually enhancing community detection and sentiment analysis on twitter networks. Journal of Data Analysis and Information Processing 1, 19–29 (2013)
6. Diakopoulos, N., Shamma, D.: Characterizing debate performance via aggregated twitter sentiment. In: Proceedings of the 28th international conference on Human factors in computing systems. ACM (2010)
7. Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., Smith, N.A.: Part-of-speech tagging for twitter: Annotation, features, and experiments. Tech. rep., DTIC Document (2010)
8. Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford (2009)
9. Hu, X., Tang, J., Gao, H., Liu, H.: Unsupervised sentiment analysis with emotional signals. In: Proceedings of the 22nd international conference on World Wide Web. pp. 607–618. International World Wide Web Conferences Steering Committee (2013)
10. Hu, X., Tang, L., Tang, J., Liu, H.: Exploiting social relations for sentiment analysis in microblogging. In: Proceedings of the sixth ACM international conference on Web search and data mining. pp. 537–546. ACM (2013)
11. Krippendorff, K.: Content analysis: an introduction to its methodology. (1980)

12. Liu, K.L., Li, W.J., Guo, M.: Emoticon smoothed language models for twitter sentiment analysis. In: AAAI (2012)
13. Makrehchi, M., Kamel, M.S.: Automatic extraction of domain-specific stopwords from labeled documents. In: Advances in information retrieval, pp. 222–233. Springer (2008)
14. Martınez-Cámara, E., Montejo-Ráez, A., Martın-Valdivia, M., Urena-López, L.: Sinai: Machine learning and emotion of the crowd for sentiment analysis in microblogs (2013)
15. Mohammad, S.M., Kiritchenko, S., Zhu, X.: Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In: In Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013), Atlanta, Georgia, USA, June 2013. (2013)
16. Nakov, P., Rosenthal, S., Kozareva, Z., Stoyanov, V., Ritter, A., Wilson, T.: Semeval-2013 task 2: Sentiment analysis in twitter. In: In Proceedings of the 7th International Workshop on Semantic Evaluation. Association for Computational Linguistics. (2013)
17. Phan, X.H., Nguyen, L.M., Horiguchi, S.: Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In: Proceedings of the 17th international conference on World Wide Web. pp. 91–100. ACM (2008)
18. Remus, R.: Asvuniofleipzig: Sentiment analysis in twitter using data-driven machine learning techniques (2013)
19. Saif, H., He, Y., Alani, H.: Semantic Smoothing for Twitter Sentiment Analysis. In: Proceeding of the 10th International Semantic Web Conference (ISWC) (2011)
20. Saif, H., He, Y., Alani, H.: Alleviating data sparsity for twitter sentiment analysis. In: Proceedings, 2nd Workshop on Making Sense of Microposts (#MSM2012) in conjunction with WWW 2012. Layon, France (2012)
21. Saif, H., He, Y., Alani, H.: Semantic sentiment analysis of twitter. In: Proceedings of the 11th international conference on The Semantic Web. Boston, MA (2012)
22. Shamma, D., Kennedy, L., Churchill, E.: Tweet the debates: understanding community annotation of uncollected sources. In: Proceedings of the first SIGMM workshop on Social media. pp. 3–10. ACM (2009)
23. Speriosu, M., Sudan, N., Upadhyay, S., Baldridge, J.: Twitter polarity classification with label propagation over lexical links and the follower graph. In: Proceedings of the EMNLP First workshop on Unsupervised Learning in NLP. Edinburgh, Scotland (2011)
24. Thelwall, M., Buckley, K., Paltoglou, G.: Sentiment strength detection for the social web. Journal of the American Society for Information Science and Technology 63(1), 163–173 (2012)
25. Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., Kappas, A.: Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology 61(12), 2544–2558 (2010)