Sentiment analysis is not solved! Assessing and probing sentiment classification

Jeremy Barnes, Lilja Øvrelid, Erik Velldal
University of Oslo
{jeremycb,liljao,erikve}@ifi.uio.no

Abstract

Neural methods for sentiment analysis have led to quantitative improvements over previous approaches, but these advances are not always accompanied with a thorough analysis of the qualitative differences. Therefore, it is not clear what outstanding conceptual challenges for sentiment analysis remain. In this work, we attempt to discover what challenges still prove a problem for sentiment classifiers for English and to provide a challenging dataset. We collect the subset of sentences that an (oracle) ensemble of state-of-the-art sentiment classifiers misclassify and then annotate them for 18 linguistic and paralinguistic phenomena, such as negation, sarcasm, modality, etc.[1] Finally, we provide a case study that demonstrates the usefulness of the dataset to probe the performance of a given sentiment classifier with respect to linguistic phenomena.

1 Introduction

Over the last 15 years, approaches to sentiment analysis which concentrated on creating and curating sentiment lexicons (Turney, 2002; Liu et al., 2005) or used n-grams for classification (Pang et al., 2002) have been replaced by models that are able to exploit compositionality (Socher et al., 2013; Irsoy and Cardie, 2014) or implicitly learn relations between tokens (Peters et al., 2018; Howard and Ruder, 2018; Devlin et al., 2018). These neural models push the state of the art to over 90% accuracy on binary sentence-level sentiment analysis.

Although these methods show a quantitative improvement over previous approaches, they are not often accompanied with a thorough analysis of the qualitative differences. This has led to the current situation, where we are aware of quantitative, but not qualitative differences between state-of-the-art sentiment classifiers. It also means that we are not aware of the outstanding conceptual challenges that we still face in sentiment analysis.

In this work, we attempt to discover what conceptual challenges still prove a problem for all state-of-the-art sentiment methods for English. To do so, we train and test three state-of-the-art machine learning classifiers (BERT, ELMo, and a BiLSTM) as well as a bag-of-words classifier on six sentence-level sentiment datasets available for English. We then collect the subset of sentences that all models misclassify and annotate them for 18 linguistic and paralinguistic phenomena, such as negation, sarcasm, modality or world knowledge. We present this new data as a challenging dataset for future research in sentiment analysis, which enables probing the problems that sentiment classifiers still face in more depth.

Specifically, the contributions of this work are:

- the creation of a challenging sentiment dataset from previously available data,
- the annotation of errors in this dataset for 18 linguistic and paralinguistic phenomena,
- a thorough analysis of the dataset,
- and finally presenting a practical use-case demonstrating how the dataset can be used to probe the particular types of errors made by a new model.

The rest of the paper is organized into related work (Section 2), a description of the experimental setup (Section 3), a brief description of the dataset (Section 4), an in-depth analysis (Section 5), a case-study that demonstrates the usefulness of the dataset (Section 6), and finally a conclusion (Section 7).

[1] The dataset is available at https://github.com/ltgoslo/assessing_and_probing_sentiment.

Proceedings of the Second BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, pages 12–23
Florence, Italy, August 1, 2019.

2 Related work

Neural networks are now ubiquitous in NLP tasks, often giving state-of-the-art results. However, they are known for being "black boxes" which are not easily interpretable. Recent interest in interpreting these methods has led to new lines of research which attempt to discover what linguistic phenomena neural networks are able to learn (Linzen et al., 2016; Gulordava et al., 2018; Conneau et al., 2018), how robust neural networks are to perturbations in input data (Ribeiro et al., 2018; Ebrahimi et al., 2018; Schluter and Varab, 2018), and what biases they propagate (Park et al., 2018; Zhao et al., 2018; Kiritchenko and Mohammad, 2018).

Specifically within the task of sentiment analysis, certain linguistic phenomena are known to be challenging. Negation is one of the aspects of language that most clearly affects expressions of sentiment and that has been studied widely within sentiment analysis (see Wiegand et al. (2010) for an early survey). The difficulties of resolving negation for sentiment analysis include determining negation scope (Hogenboom et al., 2011; Lapponi et al., 2012; Reitan et al., 2015), and semantic composition (Wilson et al., 2005; Choi and Cardie, 2008; Kiritchenko and Mohammad, 2016).

Verbal polarity shifters have also been studied. Schulder et al. (2018) annotate verbal shifters at the sense-level. They conclude that, although individual negation words are more frequent in the Amazon Product Review Data corpus, the overall frequency of negation words and shifters is likely similar. This suggests that there is a Zipfian tail of shifters which are not often handled within sentiment analysis.

Furthermore, the linguistic phenomenon of modality has also been shown to be problematic. Both Narayanan et al. (2009) and Liu et al. (2014) explore the effect of modality on sentiment classification and find that explicitly modeling certain modalities improves classification results. They advocate for a divide-and-conquer approach, which would address the various realizations of modality individually. Benamara et al. (2012) perform linguistic experiments using native speakers concerning the effects of both negation and modality on opinions, and similarly find that the type of negation and modality determines the final interpretation of polarity.

The sentiment models inspected in these analyses, however, were lexicon- and word- and n-gram-based models. It is not clear that neural networks have the same weaknesses, as they have been shown to deal with compositionality and long-distance dependencies to some degree (Socher et al., 2013; Linzen et al., 2016). Additionally, authors did not attempt to discover from the data what phenomena were present that could affect sentiment. In the current paper we aim to provide a systematic analysis of error types found across a range of datasets, domains and classifiers.

Label	MPQA	OP.	Sem.	SST	Ta.	Th.
++	–	379	–	1,852	–	–
+	193	879	3,499	3,111	923	2,727
0	527	–	4,478	2,242	1,419	1,779
−	413	399	1,310	3,140	1,320	1,828
−−	–	74	–	1,510	–	–
Total	1,133	1,731	9,287	11,855	3,662	6,334

Table 1: Statistics for the sentence-level annotations in each dataset.

3 Experimental setup

In these experiments, we test three state-of-the-art models for sentence-level sentiment classification. We choose to focus on sentence-level classification for three reasons: 1) sentence-level classification is a popular and useful task, 2) there is a large amount of high-quality annotated data available, and 3) annotation of linguistic phenomena is easier at sentence-level than document-level. It is also likely that most phenomena that occur at sentence-level, *e. g.*, negation, comparative sentiment, or modality, will transfer to other sentiment tasks.

3.1 Datasets

In order to discover a subset of sentences that all state-of-the-art models are unable to correctly predict, we collect six English-language datasets previously annotated for sentence-level sentiment from five domains (news wire, hotel reviews, movie reviews, twitter, and micro-blogs). Table 1 shows the statistics for each of the datasets.

MPQA The Multi-perspective Question Answer (MPQA) Opinion Corpus (Wilson et al., 2005) provides contextual polarity annotations for English news documents from world press. The annotations are private state frames, which include annotations for text anchor, source, target, and attitude type, among others. We extract sentiment labeled sentences by taking only those sentences that have

sentiment annotations. Additionally, we remove sentences that contain both positive and negative sentiment. This leaves a three-class (positive, neutral, negative) sentence-level dataset.

OpeNER The Open Polarity Enhanced Named Entity Recognition (OpeNER) sentiment datasets (Agerri et al., 2013) contain hotel reviews annotated for 4-class (strong positive, positive, negative, strong negative) sentiment classification. We take the English dataset, where self-attention networks give state-of-the-art results (Ambartsoumian and Popowich, 2018).

SemEval The SemEval 2013 tweet classification dataset (Nakov et al., 2013) contains tweets collected and annotated for three-class (positive, neutral, negative) sentiment. The state-of-the-art model is a Convolutional Network (Severyn and Moschitti, 2015).

Stanford Sentiment Treebank The Stanford Sentiment Treebank (Socher et al., 2013) contains 11,855 English sentences from movie reviews which have been annotated at each node of a constituency parse tree. Contextualized word representations combined with a bi-attentive sentiment network currently give state-of-the-art results (Peters et al., 2018).

Täckström dataset The Täckström dataset (Täckström and McDonald, 2011) contains product reviews which have been annotated at both document- and sentence-level for three-class sentiment, although the sentence-level annotations also have a "not relevant" label. We keep the sentence-level annotations, which gives 3,662 sentences annotated for three-class sentiment.

Thelwall dataset The Thelwall dataset derives from datasets provided with SentiStrength[2] (Thelwall et al., 2010). It contains microblogs annotated for both positive and negative sentiment on a scale from 1 to 5. We map these to single sentiment labels such that sentences which are clearly positive (pos $>=$ 3 and neg $<$ 3) are given the positive label, clearly negative sentences (pos $<$ 3 and neg $>=$ 3) the negative label, and clearly neutral sentences (3 $<$ pos $>$ 2 and 3 $<$ neg $>$ 2) the neutral. We discard all other sentences, which finally leaves 6,334 annotated sentences.

[2]The data are available at http://sentistrength.wlv.ac.uk/

3.2 Models

In order to gain an idea of what errors most models suffer from, we test three state-of-the-art models on the datasets. Additionally, we use a bag-of-words model as it is a strong baseline for text classification. For the SINGLE setup, we train all models on the training and development data for each dataset and test on the corresponding test set, therefore avoiding domain problems.

BERT The BERT model (Devlin et al., 2018) is a bidirectional transformer that is pretrained on two tasks: 1) a cloze-like language modeling task and 2) a binary next-sentence prediction task. It is pretrained on 330 million words from the BooksCorpus (Zhu et al., 2015) and English Wikipedia. We fine-tune the available pretrained model[3] on each sentiment dataset.

ELMo We use the bi-attentive classification network[4] from Peters et al. (2018). The network uses both word embeddings, as well as creating character-based embeddings from a character-level CNN-BiLSTM network. The word representations are first passed through a feedforward layer, and then through a sequence-to-sequence network with biattention. This new representation of the text is combined with the original representation and passed through another sequence-to-sequence network. Finally, a max, min, mean and self-attention pool representation is created from this last sequence. For classification, these features are sent to a maxout layer.

BiLSTM Bidirectional long short-term memory (BiLSTM) networks have shown to be strong baselines for sentiment tasks (Tai et al., 2015; Barnes et al., 2017). We implement a single-layered BiLSTM which takes pretrained skipgram embeddings as input, creates a sentence representation by concatenating the final hidden layer of both left and right LSTMs, and then passes this representation to a softmax layer for classification. Additionally, dropout serves as a regularizer.

Bag-of-Words classifier Finally, bag-of-words classifiers are strong baselines for sentiment and when combined with other features can still give

[3]https://github.com/google-research/bert

[4]https://s3-us-west-2.amazonaws.com/allennlp/models/sst-5-elmo-biattentive-classification-network-2018.09.04.tar.gz

state-of-the-art results for sentiment tasks (Mohammad et al., 2013). Therefore, we train a Linear SVM on a bag-of-words representation of the training sentences.

3.3 Model performance

Table 2 shows the accuracy of the models on the six tasks. Both methods that use pretrained language model classifiers (ELMo and BERT) are the best performing models, with an average of 11.8 difference between the language model classifiers and standard models (BOW and BILSTM). The error rates range between 8.3 on OpeNER and 20.5 on SST (see Table 3), indicating that there are differences in difficulty of datasets due to domain and annotation characteristics.

Additional experiments on a MERGED setup, where the labels from OpeNER and SST are mapped to the three-class setup, and a single model is trained on the concatenation of the training sets from all datasets, indicate that no clear performance gain is achieved. We therefore prefer to avoid the problem of domain differences and keep only the original results.

4 Challenging dataset

We create a challenging dataset by collecting the subset of test sentences that *all* of the sentiment systems predicted incorrectly (statistics are shown in Table 3). After removing sentences with incorrect gold labels, there are a total of 836 sentences in the dataset, with a similar number of positive, neutral, and negative labels and fewer strong labels. This is expected, as only two datasets have strong labels.

Furthermore, the main sources of examples are the SemEval task (249), Stanford Sentiment Treebank (452) and Thelwall datasets (215), while the Täckström dataset (129), MPQA (39) and OpeNER (29) contribute much less. This is a result of both dataset size and difficulty.

5 Dataset analysis

In order to give a clearer view of the data found in the dataset, we annotate these instances using 19 linguistic and paralinguistic labels. While most of these come from previous attempts to qualitatively analyze sentiment classifiers (Hu and Liu, 2004; Das and Chen, 2007; Pang and Lee, 2008; Socher et al., 2013; Barnes et al., 2018), others (incorrect label, no sentiment, morphology) emerged during the error annotation process. We further chose to manually annotate for the polarity of the sentence irrespective of the gold label in order to be able to locate possible annotation errors during our analysis. The annotation scheme and (manually constructed) examples of each label are shown in Table 6. Note that we did not limit the number of labels that the annotator could assign to each sentence and in principle they should assign all suitable labels during annotation.

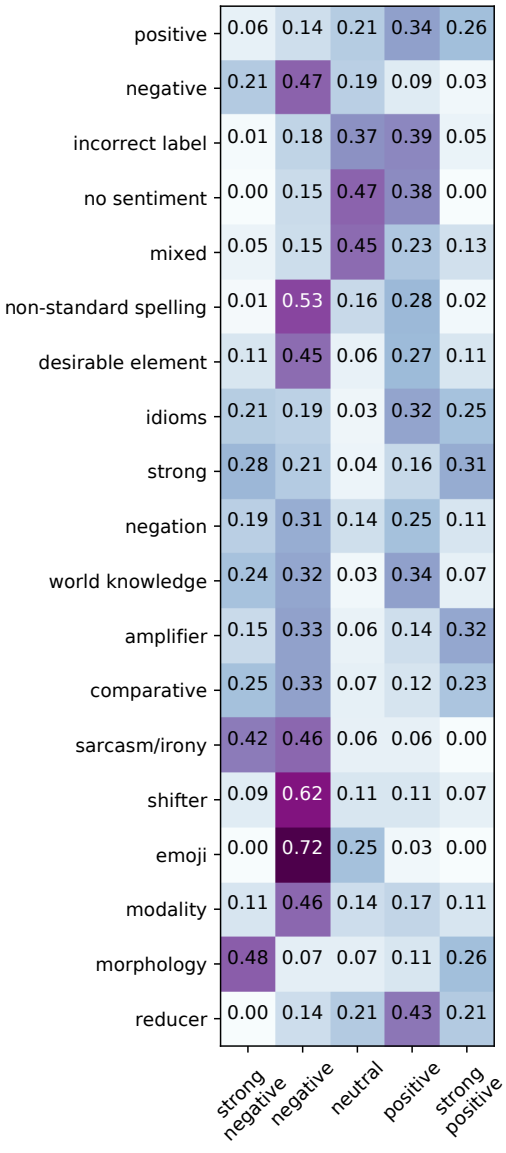

Figure 1: Distribution of labels across error categories.

An initial analysis of the errors shown in Table 5 and Figure 1 reveals that the most common errors come from the no-sentiment (214), mixed category

		MPQA	OpeNER	SemEval	SST	Täckström	Thelwall
Single	BOW	40.9	69.7	62.3	50.9	46.0	53.5
	BiLSTM	48.7	71.5	58.0	37.5	45.0	52.0
	ELMo	61.0	82.1	71.9	51.3	53.1	59.1
	BERT	**62.3**	**84.2**	**75.1**	**53.0**	**60.2**	**63.9**

Table 2: Accuracy of models on the sentiment datasets, where a different classifier is trained for each dataset.

Label	MPQA	OpeNER	SemEval	SST	Täckström	Thelwall	Total
++	–	8	–	87	–	–	95
+	16	9	59	49	46	9	188
0	1	–	45	75	31	48	200
–	16	2	47	51	18	116	250
– –	–	4	–	99	–	–	103
Total	33	23	151	361	95	173	**836**
% of original	14.5	6.6	6.4	16.3	12.9	13.6	11.7
avg. length	25.0	13.4	19.0	19.9	23.4	17.5	19.7

Table 3: Statistics of dataset, including the number of sentences from each dataset and for each label, the percentage of the original dataset kept in the dataset, and average length (in tokens) of sentences.

(185), non-standard spelling and hashtags (180), desirable elements (144), and the strong label (122).

The distribution of errors across labels (strong negative: 106, negative: 299, neutral: 303, positive: 296, strong positive: 109) compared to the gold distribution (strong negative: 294, negative: 1742, neutral: 2249, positive: 2402, strong positive: 475) shows that the strong negative is the most difficult and least common class, while positive is the easiest to classify. In the following we briefly discuss the error categories, also showing examples for each.

Mixed Polarity The largest set of errors, with 185 sentences labeled, are what we refer to as "mixed" polarity sentences. These are sentences where two differing polarities are expressed, either towards two separate entities, or towards the same entity. While the first can be solved by a more fine-grained approach (aspect-level or targeted sentiment), the second is more difficult and is often considered a category of its own (Shamma et al., 2009; Saif et al., 2013; Kenyon-Dean et al., 2018).

Strong Positive	It was **spot on**.
Positive	They're **on a roll**.
Neutral	It's a bit **hit-or-miss**.
Negative	I'm **pulling my hair out**.
Strong Negative	Madonna **can't act a lick**.

Table 4: Examples of idioms.

An analysis of the mixed category errors reveals that while most of the examples are in the "neutral" category (45%), the other 55% are annotated as having mostly positive or negative sentiment. This is a confusing situation for both annotators and sentiment classifiers, and a direct product of performing sentence-level classification rather than aspect-level. Nearly a third of the errors contain "but" clauses, which could be correctly classified by splitting them.

A more problematic situation is found among nearly 20% of the examples (34), where the annotator found the original label to be completely incorrect.[5]

Non-standard spelling Most errors in this category (180 total) are labeled either negative (49%) or positive (29%), with almost no strong positive or strong negative, which comes mainly from the fact that the noisier datasets do not contain the strong labels.

Around a third of the examples contain hashtags that clearly express the sentiment of the whole sentence, *e. g.*, "#imtiredof this SNOW and COLD weather!!!". This indicates the need to properly deal with hashtags in order to correctly classify sentiment.

Idioms Table 4 presents some examples of sentiment-bearing idioms that are taken from the challenge data set. In this category, errors (132

[5]We do not include examples where only the strength of the polarity was considered different, *i. e.*, positive vs. strong positive.

label	# examples
incorrect label	**277**
no sentiment	**214**
mixed	**185**
non-standard spelling	**180**
desirable element	**144**
idioms	132
strong	122
negation	97
world knowledge	81
amplifier	79
comparative	68
sarcasm/irony	58
shifter	50
emoji	46
modality	38
morphology	31
reducer	13

Table 5: Number of labels for each category in annotation study. **Bold** numbers indicate the five most frequent sources of errors. The total number of labels does not sum to the number of sentences in the dataset, as each sentence can have multiple labels.

sentences labeled) are spread relatively uniformly across labels. Learning these correctly from sentence-level annotations is unlikely, especially because they are seldom found repeatedly, even in a training corpus of decent size. Therefore, incorporating idiomatic information from external data sources may be necessary to improve the classification of sentences within this category.

Strong Labels This category (122 total) is particularly difficult for sentiment classifiers for several reasons. First, strong negative sentiment is often expressed in an understated or ironic manner. For example, "Better at putting you to sleep than a sound machine."

For strong positive examples in the dataset, there is often difficult vocabulary and morphologically creative uses of language, *e. g.*, "It is a kickass , dense sci-fi action thriller hybrid that delivers and then some.", while strong negative examples often contain sarcasm or non-standard spelling, *e. g.*, "All prints of this film should be sent to and buried on Pluto.".

Negation Negation, which accounts for 97 errors, directly affects the classification of polar sentence (Wiegand et al., 2010). Therefore, we look at the differences between correctly and incorrectly classified sentences containing negation, by analyzing 100 correctly and incorrectly classified sentences containing negation.

From our analysis, there is no specific negator that is more difficult to resolve regarding its effect on sentiment classification.

We also perform an analysis of negation scope under the assumption that when a negator occurs farther from its negated element, it is more difficult for the sentiment classifier to correctly resolve the negation. Let d be the distance between the negator n and the relevant sentiment element se, such that $d = |ind(se) - ind(n)|$ where the function ind calculates the index of a token in a sentence. We find that the incorrectly classified examples have an average d of 2.7, while the correctly classified examples had 2.5. This seems to rule out a problem of negation scope as the underlying difference.

High-level or clausal negation occurs when the negator negates a full clause, rather than an adjective or noun phrase, *e. g.*, "I don't think it is a particularly interesting film". In the dataset this phenomenon is found more prevalently in the incorrectly classified examples (8%) versus the correctly classified examples (3%), but does not occur often in absolute terms.

The main source of difference regarding correctly classifying examples involving negation seems to be irrelevant negation. Irrelevant negation refers to cases where a sentence contains a negation but where the sentiment-bearing expression is not within the scope of negation. In our data, there is a strong difference in the distribution of irrelevant negation in correctly and incorrectly classified examples (80% vs. 25%, respectively), suggesting that sentiment classifiers learn to ignore most occurrences of negation.

World Knowledge Examples from the dataset where world knowledge is necessary to correctly classify a sentence (81 sentences) include comparisons with entities commonly associated with positive or negative polarity, *e. g.*, "Elicits more groans from the audience than Jar Jar Binks, Scrappy Doo and Scooby Dumb, all wrapped up into one.", analogies, *e. g.*, "Adam Sandler is to Gary Cooper what a gnat is to a racehorse.", or rating scales, *e. g.*, "10/10 overall".

This category is also highly correlated with sarcasm and irony. In fact, irony is often defined as "violating expectations" (Hao and Veale, 2010),

positive	"It was good."
negative	"It was bad."
negation	"It was not good."
strong	"It was incredible."
amplifier	"It was really good."
reducer	"It was kind of bad."
desirable element	"It had a pool."
comparative	"It was better than the first hotel."
shifter	"They denied him the scholarship"
modality	"I would have loved the room if it been bigger."
world knowledge	"It was 2 minutes from the beach." vs. "It was 2 hours from the beach."
morphology	"It was un-fricking-believable."
non-standard spelling	"It was awesoooome."
idioms	"It's not my cup of tea."
sarcasm/irony	"I love it when people yell at me first thing in the morning."
emoji	":)"
no sentiment	"The president will hold a talk tomorrow."
mixed	"The plot was nice, but a little slow."
incorrect label	Any clearly incorrect label.

Table 6: Categories and examples for error annotation guidelines.

which presupposes that we possess a world knowledge containing expectations of a situation.

Amplified Amplifiers occur mainly in negative and strong positive examples, such as "It's an awfully derivative story." Most of the amplified sentences found in the dataset (71/79) contain amplifiers other than "very", such as "super", "incredibly", or "so".

Comparative Comparative sentiment, with 68 errors, is known to be difficult (Hu and Liu, 2004; Liu, 2012), as it is necessary to determine which entity is on which side of the inequality. Sentences like "Will probably stay in the shadow of its two older, more accessible Qatsi siblings" are difficult for sentiment classifiers that do not model this phenomenon explicitly.

Sarcasm/Irony Sarcasm and irony (58 errors), which are often treated separately from sentiment analysis (Filatova, 2012; Barbieri et al., 2014), are present mainly in negative and strong negative examples in the dataset. Correctly capturing sarcasm and irony is necessary to classify some negative and strong negative examples, *e. g.*, "If Melville is creatively a great whale, this film is canned tuna."

Shifters Shifters (50 errors), such as "abandon", "lessen", or "reject" are less common within the dataset, but normally move positive polarity words towards a more negative sentiment. The most common shifter is the word "miss", used as in "We miss the quirky amazement that used to come along for an integral part of the ride."

Emoji While the models handle most occurrences of emojis well, they falter more on the negative examples (46 errors). More than half of the examples in the dataset present positive emoji with a negative gold label, such as "Pricess Leia is going to be gutted! :-)."

Modality None of the state-of-the-art sentiment systems deals explicitly with modality (38 total errors). While in many of the examples modality does not express a different sentiment than the same sentence without modality, in the dataset there are examples that do, *e. g.*, "Still, I thought it could have been more."

Morphology While not the most prominent label (31 errors), the examples in the dataset that contain morphological features that effect sentiment are normally strong positive or strong negative. This most often contains creative use of English morphology, *e. g.*, "It was fan-freakin-tastic!" or "It's hyper-cliched".

Reducers Reducers (13 errors), such as "kind of", "less", or "all that" cooccur with both positive and negative polar words within the dataset, and

label	Sent.	Phrases	Rel. Imp.
overall	23.0	31.1	10.5
positive	19.0	**26.9**	9.8%
negative	23.1	**35.0**	15.5%
mixed	21.2	**26.5**	6.7%
no-sentiment	37.6	**42.6**	8.1%
non-strd spelling	40.3	**43.5**	3.8%
desirable	25.7	**28.7**	4.0%
idioms	13.7	**23.1**	11.0%
strong	15.5	**23.7**	9.7%
negation	23.9	**38.6**	19.3%
world know.	14.9	**21.6**	19.6%
amplified	13.9	**31.9**	20.9%
comparative	11.7	**13.3**	1.8%
irony	**20.8**	18.8	-2.5%
shifters	**33.3**	24.4	-11.8%
emoji	33.3	**50.0**	25.0%
modality	20	**22.9**	3.6%
morphology	**18.5**	**18.5**	0%
reduced	7.7	**23.1**	16.7%

Table 7: Per category accuracy and relative improvement (last column) of BERT model trained on SST sentences (8,544) and SST phrases (155,019).

tend to lead to positive or neutral sentiment, *e. g.*, "It was a lot less hassle."

6 Case study: Training with phrase-level annotations

As a case study for the usage of the dataset presented here, we evaluate a model that has access to more compositional information. Besides having sentence-level annotations, the SST dataset also contains annotations for each phrase in a constituency tree, which gives a considerable amount more training data, specifically 155,019 annotated phrases vs. 8,544 annotated sentences. It has been claimed that this data allows models to learn more compositionality (Socher et al., 2013). Therefore, we fine-tune the best performing model (BERT) on this data and test on our dataset. The BERT model trained on phrases achieves 55.1 accuracy on the SST dataset, versus 53.0 for the model trained only on sentence-level annotations.

Table 7 shows that the model trained on the SST phrases performs overall much better than the model trained on SST sentences[6] on the dataset. Using the error annotations in the challenge data set, we find that results improve greatly on the sentences which contain the labels negation, world knowledge, amplified, emoji, and reduced, while performing worse on irony, shifters and equally on morphology. This analysis seems to indicate that phrase-level annotations help primarily with learning compositional sentiment (negation, amplified, reduced), while other phenomena, such as irony or morphology do not receive improvements. This confirms that training on the phrase-level annotations improves a sentiment model's ability to classify compositional sentiment, while also demonstrating the usefulness of our dataset for introspection.

7 Conclusion and future work

In this paper, we tested three state-of-the-art sentiment classifiers and a baseline bag-of-words classifier on six English sentence-level sentiment datasets. We gathered the sentences that all methods misclassified in order to create a dataset. Additionally, we performed a fine-grained annotation of error types in order to provide insight into the kinds of problems sentiment classifiers have. We will release both the code and the annotated data with the hope that future research will utilize this resource to probe sentiment classifiers for qualitative differences, rather than rely only on quantitative scores, which often obscure the plentiful challenges that still exist.

Many of the phenomena found in the dataset, *e. g.*, negation or modality, have been discussed in depth in (Liu, 2012). However, the dataset that resulted from this work demonstrates that modern neural methods still fail on many examples of these phenomena. Additionally, our dataset enables a quick analysis of qualitative differences between models, probing their performance with respect to the linguistic and paralinguistic categories of errors.

Additionally, many of the findings from this paper are likely to vary to a degree for other languages, due to typological differences, as well as differences in available training data. The annotation method proposed in this paper, however,

[6] It is important to realize that the SST-sentence model has 0 accuracy on the subset of the dataset taken from the SST dataset, but not on the sentences taken from the other datasets.

should enable the creation of similar analyses and datasets in other languages.

We expect that this approach to creating a dataset is also easily transferable to other tasks which are affected by linguistic or paralinguistic phenomena, such as hate speech detection or sarcasm detection. It would be more useful to have some knowledge of the phenomena that could affect the task beforehand, but a careful error analysis can also lead to insights which can be translated into annotation labels.

Regarding ways of moving forward, there are already many sources of data for the linguistic phenomena we have analyzed in this work, ranging from datasets annotated for negation (Morante and Blanco, 2012; Liu et al., 2018), irony (Van Hee et al., 2018), emoji (Barbieri et al., 2018), as well as datasets for idioms (Muzny and Zettlemoyer, 2013) and their relationship with sentiment (Jochim et al., 2018). We believe that discovering ways to explicitly incorporate this available information into state-of-the-art sentiment models may provide a way to improve current approaches. Multi-task learning (Caruana, 1993) and transfer learning (Peters et al., 2018; Devlin et al., 2018; Howard and Ruder, 2018) have shown promise in this respect, but have not been exploited for improving sentiment classification with regards to these specific phenomena.

Acknowledgements

This work has been carried out as part of the SANT project, funded by the Research Council of Norway (grant number 270908).

References

Rodrigo Agerri, Montse Cuadros, Sean Gaines, and German Rigau. 2013. OpeNER: Open polarity enhanced named entity recognition. *Sociedad Española para el Procesamiento del Lenguaje Natural*, 51(Septiembre):215–218.

Artaches Ambartsoumian and Fred Popowich. 2018. Self-attention: A better building block for sentiment analysis neural network classifiers. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 130–139.

Francesco Barbieri, Jose Camacho-Collados, Francesco Ronzano, Luis Espinosa Anke, Miguel Ballesteros, Valerio Basile, Viviana Patti, and Horacio Saggion. 2018. SemEval 2018 task 2: Multilingual emoji prediction. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 24–33, New Orleans, Louisiana.

Francesco Barbieri, Horacio Saggion, and Francesco Ronzano. 2014. Modelling sarcasm in twitter, a novel approach. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 50–58.

Jeremy Barnes, Toni Badia, and Patrik Lambert. 2018. MultiBooked: A Corpus of Basque and Catalan Hotel Reviews Annotated for Aspect-level Sentiment Classification. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan.

Jeremy Barnes, Roman Klinger, and Sabine Schulte im Walde. 2017. Assessing state-of-the-art sentiment models on state-of-the-art sentiment datasets. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 2–12, Copenhagen, Denmark.

Farah Benamara, Baptiste Chardon, Yannick Mathieu, Vladimir Popescu, and Nicholas Asher. 2012. How do negation and modality impact on opinions? In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, pages 10–18.

Richard Caruana. 1993. Multitask learning: A knowledge-based source of inductive bias. In *Proceedings of the Tenth International Conference on Machine Learning*, pages 41–48. Morgan Kaufmann.

Yejin Choi and Claire Cardie. 2008. Learning with compositional semantics as structural inference for subsentential sentiment analysis. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 793–801.

Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single \$&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136.

Sanjiv R. Das and Mike Y. Chen. 2007. Yahoo! for amazon: Sentiment extraction from small talk on the web. *Management Science*, 53(9):1375–1388.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. Hotflip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 31–36.

Elena Filatova. 2012. Irony and sarcasm: Corpus generation and analysis using crowdsourcing. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205.

Yanfen Hao and Tony Veale. 2010. An ironic fist in a velvet glove: Creative mis-representation in the construction of ironic similes. *Minds and Machines*, 20(4):635–650.

A. Hogenboom, P. Van Iterson, B. Heerschop, F. Frasincar, and U. Kaymak. 2011. Determining negation scope and strength in sentiment analysis. pages 2589–2594.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339.

Minqing Hu and Bing Liu. 2004. Mining opinion features in customer reviews. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2004)*, pages 168–177.

Ozan Irsoy and Claire Cardie. 2014. Deep recursive neural networks for compositionality in language. In *Advances in Neural Information Processing Systems*, volume 3, pages 2096–2104.

Charles Jochim, Francesca Bonin, Roy Bar-Haim, and Noam Slonim. 2018. SLIDE - a sentiment lexicon of common idioms. In *Proceedings of the 11th Language Resources and Evaluation Conference*, Miyazaki, Japan.

Kian Kenyon-Dean, Eisha Ahmed, Scott Fujimoto, Jeremy Georges-Filteau, Christopher Glasz, Barleen Kaur, Auguste Lalande, Shruti Bhanderi, Robert Belfer, Nirmal Kanagasabai, Roman Sarrazingendron, Rohit Verma, and Derek Ruths. 2018. Sentiment analysis: It's complicated! In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1886–1895.

Svetlana Kiritchenko and Saif Mohammad. 2016. The effect of negators, modals, and degree adverbs on sentiment composition. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 43–52.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53.

Emanuele Lapponi, Jonathon Read, and Lilja Øvrelid. 2012. Representing and resolving negation for sentiment analysis. In *Proceedings of the 2012 IEEE 12th International Conference on Data Mining Workshops*, ICDMW '12, pages 687–692, Washington, DC, USA.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167.

Bing Liu, Minqing Hu, and Junsheng Cheng. 2005. Opinion Observer: Analyzing and comparing opinions on the web. In *Proceedings of the 14th international World Wide Web conference (WWW-2005)*, Chiba Japan.

Qianchu Liu, Federico Fancellu, and Bonnie Webber. 2018. NegPar: A parallel corpus annotated for negation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan.

Yang Liu, Xiaohui Yu, Bing Liu, and Zhongshuai Chen. 2014. Sentence-level sentiment analysis in the presence of modalities. In *Computational Linguistics and Intelligent Text Processing*, pages 1–16, Berlin, Heidelberg. Springer Berlin Heidelberg.

Saif M. Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013)*.

Roser Morante and Eduardo Blanco. 2012. *SEM 2012 shared task: Resolving the scope and focus of negation. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 265–274, Montréal, Canada.

Grace Muzny and Luke Zettlemoyer. 2013. Automatic idiom identification in Wiktionary. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1417–1421, Seattle, Washington, USA.

Preslav Nakov, Sara Rosenthal, Zornitsa Kozareva, Veselin Stoyanov, Alan Ritter, and Theresa Wilson. 2013. Semeval-2013 task 2: Sentiment analysis in twitter. In *Second Joint Conference on Lexical and*

*Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*.

Ramanathan Narayanan, Bing Liu, and Alok Choudhary. 2009. Sentiment analysis of conditional sentences. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 180–189.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 79–86.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237.

Johan Reitan, Jørgen Faret, Björn Gambäck, and Lars Bungum. 2015. Negation scope detection for twitter sentiment analysis. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 99–108.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically equivalent adversarial rules for debugging nlp models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 856–865.

Hassan Saif, Miriam Fernández, Yulan He, and Harith Alani. 2013. Evaluation datasets for twitter sentiment analysis: A survey and a new dataset, the sts-gold. In *1st Interantional Workshop on Emotion and Sentiment in Social and Expressive Media: Approaches and Perspectives from AI (ESSEM 2013)*, Turin, Italy.

Natalie Schluter and Daniel Varab. 2018. When data permutations are pathological: the case of neural natural language inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4935–4939.

Marc Schulder, Michael Wiegand, Josef Ruppenhofer, and Stephanie Køser. 2018. Introducing a Lexicon of Verbal Polarity Shifters for English. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan.

Aliaksei Severyn and Alessandro Moschitti. 2015. Unitn: Training deep convolutional neural network for twitter sentiment classification. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 464–469.

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*, WSM '09, pages 3–10, New York, NY, USA.

Richard Socher, Alex Perelygin, Jy Wu, Jason Chuang, Chris Manning, Andrew Ng, and Chris Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642.

Oscar Täckström and Ryan McDonald. 2011. Semi-supervised latent variable models for sentence-level sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 569–574.

Kai Sheng Tai, Richard Socher, and Christopher D Manning. 2015. Improved semantic representations From tree-structured long short-term memory networks. *Association for Computational Linguistics 2015 Conference*, pages 1556–1566.

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

Peter Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*.

Cynthia Van Hee, Els Lefever, and Veronique Hoste. 2018. SemEval-2018 task 3: Irony detection in English tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 39–50, New Orleans, Louisiana.

Michael Wiegand, Alexandra Balahur, Benjamin Roth, Dietrich Klakow, and Andrés Montoyo. 2010. A survey on the role of negation in sentiment analysis. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, pages 60–68.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *The IEEE International Conference on Computer Vision (ICCV)*.