Using Structured Representation and Data: A Hybrid Model for Negation and Sentiment in Customer Service Conversations

Amita Misra, Mansurul Bhuiyan, Jalal Mahmud, and Saurabh Tripathy
IBM-Research, Almaden
San Jose, CA, USA
amita.misra1|mansurul.bhuiyan|jumahmud|Saurabh.Tripathy2@ibm.com

Abstract

Twitter customer service interactions have recently emerged as an effective platform to respond and engage with customers. In this work, we explore the role of negation in customer service interactions, particularly applied to sentiment analysis. We define rules to identify true negation cues and scope more suited to conversational data than existing general review data. Using semantic knowledge and syntactic structure from constituency parse trees, we propose an algorithm for scope detection that performs comparable to state of the art BiLSTM. We further investigate the results of negation scope detection for the sentiment prediction task on customer service conversation data using both a traditional SVM and a Neural Network. We propose an antonym dictionary based method for negation applied to a CNN-LSTM combination model for sentiment analysis. Experimental results show that the antonym-based method outperforms the previous lexicon-based and neural network methods.

1 Introduction

Negation has been described as a polarity influencer (Wilson et al., 2009) and therefore it has to be taken into consideration while designing a sentiment prediction system, but how important it is in twitter customer service conversations? For example, both the customer service tweets in Table 1 have an explicit negation cue but the effect of cue words on the polarity differ. The first tweet has a negation cue [*don't*] that changes the positive polarity of the words in the scope [*think you do understand*]. Additionally, tweet 1 has a hashtag [*Misleading*] which could be a strong negative signal on its own. The second tweet has a cue word [*not*] but it does not negate the words in that sentence or change their polarity. The negation cue [*not*] in the second tweet is not a true negation cue, and hence it has no scope.

S.No	Tweet	Sentiment
1	@Username I *don't* think you do understand. Buyers and Sellers deserve to know facts,User actively prevents accurate feedback. #Misleading.	Negative
2	@Username Sorry to hear this. Have you had a chance to call/chat us? If *not*, we can look into options:	Positive

Table 1: Customer Service Conversation.

Negation can be expressed in different ways in natural language. It may be through the use of explicit negation cues such as *no, not and never* that have a morphologic indication of a negative meaning. This also includes a group of broad or semi negatives words (e.g. barely, hardly, and seldom) that have a negative meaning but are without any morphological negative. This has been also referred to as clausal or syntactic negation (Quirk et al., 1985; Givón, 1993). These cue words are often used to negate a statement or an assertion that expresses a judgment or an opinion. However in some contexts, these cue words function as exclamations, and not as true negation cues. These false cues do not change the sentiment polarity of the following expression, and hence do not have any associated scope. We define rules to identify true negation cues and their scopes more suited to conversational data than existing general review data.

The impact of negation has been studied in domains such as biomedical, literary texts, and online reviews (Szarvas et al., 2008; Morante et al., 2008; Councill et al., 2010; Reitan et al., 2015; Konstantinova et al., 2012); however, none of the previous corpora are conversational in nature. Scope definitions may depend on the domain. Reitan et al. (2015) showed that negation scope algo-

Proceedings of the 10th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 46–56
Minneapolis, June 6, 2019. ©2019 Association for Computational Linguistics

rithm trained on a twitter domain struggled when tested on a medical domain. Majority of the previous work in scope detection has been dominated by SVMs or Neural Networks, which require expensive annotated training data. Scope annotation is costly and time-intensive as all the scope conflicts have to be resolved by mutual discussion amongst expert annotators. Our main motivation is to create a system that does not require a huge amount of training data for scope detection, but has comparable performance to machine learning models that require annotated training data. The proposed method uses constituency parse trees and semantic knowledge to predict scope. The results in Table 7 show that the method is comparable to state of the art BiLSTM model from (Fancellu et al., 2016) on gold negation cues for scope prediction. Since our method does not need expensive training data, we could also use this method to predict on other negation data sets. However, our aim here was first to test if the predicted negation scope improves sentiment in conversations.

For a real time sentiment prediction system, we need both a cue prediction system to determine the true negation cues, and scope detection. As a first step, we use a data based approach to train an SVM to predict true negation cues. It's much faster and simpler to get annotated data for cue prediction, a binary task as compared to scope detection, which is a sequence labeling task. This is followed by a second step of constituency tree-based negation scope detection for predicted cues. The last step applies negation prediction coupled with antonym dictionary to improve the sentiment performance for a combination CNN-LSTM model.

The contributions of this paper are:

- Negation scope rules more suited to conversational data.
- A constituency-tree based approach for scope detection that uses both semantic and structural information, and does not require annotated data for scope.
- An antonym based negation applied to a combination CNN-LSTM model for sentiment prediction in conversations.

We begin with a discussion of related work in Section 2, followed by negation corpus in Section 3, and negation cue and scope detection experiments in Section 4. Next, we show the effect of introducing negation detection for the sentiment task in Section 5. We then compare and contrast the twitter conversational sentiment data to previous datasets in Section 6. Finally, Section 7 presents the conclusions and future directions.

2 Related Work

Initial studies on negation scope detection were performed in Biomedical domain including medical reports, biological abstracts, and papers (Szarvas et al., 2008). Morante and Daelemans (2009) used a 2 step approach: first, a decision tree to predict negation cues, followed by a CRF meta-learner to predict negation scope. The model used a combination of k-nearest neighbors, a support vector machine, and a CRF. The research in this field was further enhanced by a shared SemEval 2012 negation and scope resolution task (Morante and Blanco, 2012). The organizers released a cue and scope annotated corpus of Conan Doyle stories.

Read et al. (2012) described both, a rule-based and a data driven approach for scope resolution. Both the methods were driven by the hypothesis that syntactic units correspond to scope annotations. The rule-based approach used heuristic rules based on POS tags and constituent category labels, while machine learning used SVM based ranking of syntactic constituents. Limited rule based system obtained similar results to the data-driven system on a held-out set. This result was particularly note-worthy since getting sufficient scope annotation training data for every new domain is quite expensive, and requires trained annotators. A comparison of these results motivated us to further develop the rule-based system for the conversational domain using both semantic information and syntactic structure. (Councill et al., 2010; Lapponi et al., 2012; Reitan et al., 2015) used CRF-based sequence labeling using features from dependency tree. Packard et al. (2014) used hand-crafted heuristics to traverse Minimal Recursion Semantics (MRS). However, if a reliable representation for a sentence could not be created, their system used a fall back mechanism based on Read et al. (2012). Fancellu et al. (2016) showed that a neural network based model using a BiLSTM outperformed the previously developed classifiers on both scope token recognition and exact scope matching for in domain testing but not on a different domain. The authors noted that when tested on a different test set from Wikipedia, White (2012)'s model built on constituency-based features performed better.

A survey on the role of negation in sentiment analysis was done by (Wiegand et al., 2010) stating that negation expressions are ambiguous i.e. in some contexts do not function as a negation and, therefore, need to be disambiguated. Rules of composition were defined by Moilanen and Pulman (2007) on the syntactic representation of a sentence to account for negation and the modeling paradigm could be applied to determine the sub-sentential polarity of the sentiment expressed. (Councill et al., 2010) showed that a CRF based negation enhanced classifier improved the F-score of positive on-line reviews by 29.5% and 11.4% for negative. Much recent progress in the field has been in connection with the "The International Workshop on Semantic Evaluation" (SemEval) (Nakov et al., 2013). Since 2013 the workshop has included shared tasks on "Sentiment Analysis in Twitter". Most of the top performing systems submitted used just a simple punctuation model that assigns a negation cue scope over all the terms to the next punctuation (Tang et al., 2014; Miura et al., 2014; Mohammad et al., 2013). (Kiritchenko et al., 2014a) reported an improvement of up to 6.5 percentage points when handling negated context on the SemEval-2013 test set. Using the simple punctuation model for scope detection, an improvement of upto 6% was reported by (Reitan et al., 2015).

With the recent advances in deep learning and use of embeddings, the CNN and LSTM based models have shown to outperform traditional SVM and lexicon based methods for sentiment in twitter and review domain. Kim (2014) applied a CNN based model to numerous document classification tasks, and improved the sentiment state of the art using a CNN architecture with one layer of convolution trained using word vectors obtained from Mikolov et al. (2013) on 100 billion words of Google News. Yin and Schütze (2015) combined different word embeddings using multichannel CNN. Wang et al. (2016) showed that a combination CNN-LSTM outperformed CNN for sentiment task. Shin et al. (2017) integrated sentiment embeddings in a CNN to build simpler high-performing models with much smaller word embeddings. However, none of the previous work has explored negation coupled with antonyms to get a better sentence representation for sentiment prediction.

3 Conversational Negation Corpus

hardly	lack	lacking	lacks	neither
no	nobody	none	nothing	nowhere
cant	arent	dont	doesnt	didnt
havent	isnt	mightnt	mustnt	neednt
shouldnt	wasnt	werent	wouldnt	without
seldom	scarcely	wont	never	aint
barely	nor	not	hadnt	rather
hasnt	shant			

Table 2: Negation cue lexicon.

We selected conversations from the Twitter customer service pages of different companies and downloaded 89552 customer service tweets in total[1]. A lexicon of explicit cue words that may act as indicators of negation was primarily adopted from (Councill et al., 2010; Reitan et al., 2015). It was further extended to include semi negative words. The final set of cues used is shown in Table 2. We then extracted 23243 tweets containing explicit negation cues giving a frequency of 26%. In contrast, the equivalent numbers for BioScope corpus (Szarvas et al., 2008) and for Twitter corpus (Reitan et al., 2015) are 13.8 % and 13.5% respectively. Tottie (1991) presented a comprehensive taxonomy of English negations and stated that frequency of negation is 12.8% in written English. In another statistical study on negation, (Biber, 1999) reported that negation is much more frequent in conversation as compared to written discourse. Since we had a lot more negation cue occurrences, we divided the tweets into 5 different groups based on the number of negation cues present in each tweet. A random sample was selected from each group based on the number of instances in each group giving a dataset of 2000 tweets. A separate set of 100 tweets was used as a development set to help formulate the rules and study negation patterns. Every tweet was annotated by a pair of annotators. To test the robustness of guidelines, we measured inter-annotator agreements (IAA) for each pair of raters using the token level and full scope measures as used in previous work (Reitan et al., 2015). The token level is the percentage of tokens annotators agreed upon.

[1] Comapny names are anonymous for annotation

Since the average number of tokens in scope is far less than the number outside the scope, this is a skewed measure. For full scope, it is the percentage of scopes that have a complete and exact match amongst annotators (PCS). After an initial annotation phase of 1000 tweets, the average token level agreement was 0.95 and full scope was 0.78. All the scope conflicts were mutually resolved after discussion. Corpus statistics are shown in Table 3. The average number of tokens per tweet is 22.3, per sentence is 13.6 and average scope length is 2.9.

Total negation cues	2921
True negation cues	2674
False negation cues	247
Average scope length	2.9
Average sentence length	13.6
Average tweet length	22.3

Table 3: Cue and token distribution in the conversational negation corpus.

3.1 Annotation Guidelines

We define rules to identify true negation cues and their scopes more suited to twitter customer service conversational data than existing general review data, which has its own characteristics such as brevity and skewed distribution towards negative polarity (Sec. 6). The guidelines described here were adapted from Councill et al. (2010) but modified for customer service conversations. Nouns and adjectives are key indicators of sentiment (Hu and Liu, 2004; Pang and Lee, 2008) and hence we had a more restricted scope for noun and adjectives as compared to verbs and adverbs. In the following examples, the cue is underlined and the scope is marked in bold.

- Annotating the negation cue.
    - False Negation: Some negation cues can be used in multiple senses and hence the mere presence of an explicit cue in a sentence does not imply that it functions as a negator, (e.g., *He could not help me more*). Reitan et al. (2015) reported that in the twitter corpus the cue word *no* often occurs as an exclamation leading to erroneous predictions. Such cues should be marked as false negations.
    - Negation cues are not part of the scope.
- Annotating the Scope
    - Annotate the minimal span for scope.
    - Scope is continuous.
    - A noun or an adjective negated in a noun phrase: If only the noun or adjective is being negated then do not annotate the entire clause. Consider each term separately, (e.g., *There are no* **details** *on the return page)*.
    - A verb or an adverb phrase: By and large, the entire phrase is annotated, (e.g., *I do not* **want to update it anymore**).

We used a different scheme for annotating nouns and adverbs as compared to (Councill et al., 2010). Our nouns have a more restricted scope contrary to the previous work where typically the entire phrase is negated in a noun phrase.

4 Negation Cue and Scope Detection Experiments

We divided the dataset into train and test sets giving a training set of 2317 cues and test set of 604 cues to train both a cue detection and BiLSTM scope prediction.

4.1 Negation Cue Detection

The task of cue detection system is to determine if the potential cue word negates a concept in the sentence. It is based on the state-of-the-art cue classifier described by (Read et al., 2012; Velldal, 2011; Enger et al., 2017). A binary SVM classifier is used to disambiguate the cue for only the known cue words, considering the set of cue words as a closed class. Our baseline system uses the features and implementation as described in Enger et al. (2017). The features used are the word form, POS and lemma of the token, and lemmas for previous and next position. Adding simple features such as position of the cue word in the sentence, POS bigrams improves the F-score of false negation from a **0.61** baseline to **0.68** on a test set containing 47 false and 557 actual negation cues. See Table 4.

	F-Score		
	Baseline	**Proposed**	**Support**
False cues	0.61	0.68	47
Actual cues	0.97	0.98	557

Table 4: Cue classification on the test set.

4.2 Negation Scope Detection

Syntactic structure of the sentence has been often used to determine the scope of negation using supervised classifiers (Morante and Daelemans, 2009; Councill et al., 2010; Reitan et al., 2015; Read et al., 2012; Carrillo de Albornoz et al., 2012). Our work is inspired by the previous rule-based approach using constituency tree (Read et al., 2012; Carrillo de Albornoz et al., 2012; Velldal et al., 2012). We build on that work by adding rules based on semantic information, the position of the negation cue in the tree and the projection of its parent based on phrase structure. The syntax tree is obtained using Stanford CoreNLP (Manning et al., 2014). It is possible that the negation marker may be present in the main clause but semantically belong to the embedded clause. (Gotti et al., 2008) mention that semantic content of copula verbs is subsidiary to that of subject complement, (e.g., *A drunken worker does not become rich*, the negation marker "not" negates the subject complement "rich" rather than the copula verb "become"). Neg-raising is a linguistic phenomenon where certain predicates such as *think, believe* and *seem* occur in the main clause but may be interpreted to negate the complement clause (Fillmore, 1963; Horn, 1989). We move ahead in a linear order on either finding a copula verb or neg-raising predicates (NRPs). Table 5 contains the list of such verbs used. At this point, the algorithm branches based on POS tag of the token. We traverse the tree in an upward direction until we find a parent with the desired phrase tag as determined by the POS tag of the token. This method differs from the previous work that finds the first common ancestor enclosing the negation cue and the word immediately after it, and assumes all descendant leaf nodes to the right as its scope (Read et al., 2012). Our detailed algorithm for finding the scope is presented in Figure 1.

think	believe	seem	appear	feel
grow	look	prove	remain	smell
sound	become	might	are	am
been	has	were	was	is

Table 5: Neg-raising predicates (NRP) and copula verb.

1. Traverse the tokens in linear order and stop on finding any cue from the explicit cue lexicon.
2. Find the next first occurrence of noun, verb, adverb, adjective.
3. If the verb is an instance of copula verb or neg-raising, move to step2 else go to step4.
4. Branch depending upon POS tag of the token found in step2.
   (a) For nouns and adjectives:
   - Traverse the tree in upward direction level by level until you reach an ancestor with a tag of NP, VP, ADJP, SBAR* or S* for adjectives. For nouns, stop at NP, SBAR* or S*.
   - If a PP, VP, ADVP, SQ, SINV or SBAR* is a right child of the ancestor, then remove that child.
   - Get all the descendant leaves as scope.

   (b) For verbs and adverbs:
   - Traverse the tree in upward direction level by level until you reach an ancestor with a tag of VP, SBAR* or S*.
   - If there exists an SBAR*, SQ, or SINV tag as a right child of the ancestor then remove that child.
   - Get all the descendant leaves as scope.
5. Apply post-processing rules to align the scopes.

Figure 1: Negation scope detection .

Though we use SBAR* tags from syntax tree to determine the clause boundaries but it cannot detect all boundaries. We therefore also used explicit discourse connectives that signal a contrast relation, or a coordination to limit the scope. These connectives act as a boundary for an idea expressed in one clause. For example, *To be honest I am not angry but upset*, the scope of not as per the rules given in Figure 1, would be **angry but upset**. Once we find this scope, we use the discourse connective *'but'* to delimit the scope. The list of connectives used is given in Table 6. Morante and Blanco (2012) reported that for the SemEval shared task on negation scope detection, most of the systems were post processed to improve their performance. Read et al. (2012) formulated a set of slackening heuristics by removing certain constituents at the beginning or end of a scope. This improved the alignment of scopes from an initial 52.42% to 86.13%. Following a similar approach, the post-processing rules were designed and are given in Figure 2.

because	while	until	however	what
but	though	although	nothing	nowhere
whenever	&	and	nonetheless	whereas
whose	why	where	wherever	

Table 6: Prune-connective list

- If the scope contains a connective from the prune-connective list then delimit the scope before the connective.
- If the scope contains a punctuation then delimit the scope before the punctuation marker.
- Remove the negation cue from the scope.
- Remove the scope words before the cue word, if any.
- If no scope is found after using these rules then predict a default scope as all the tokens up to the first noun, adjective or verb.
- Include the tokens after the negation cue, upto the beginning of the predicted scope.

Figure 2: Post-processing heuristic rules.

4.3 Negation Scope Detection Evaluation

The algorithm is evaluated using two different measures; *token-level* and *scope-level*. Every token can be either in-scope or out of scope. We report the F-score for both in-scope and out-of-scope tokens. Since the output is a sequence, F-score metrics may be insufficient as it just considers individual tokens. We also report percentage of correct scopes (PCS). Results are given in Table 7. The out-of-scope token has a higher F-score as compared to in-scope. This is expected since scope tokens are restricted and less in number as compared to out-of-scope (See Table 3). The in-scope F-score is more important for the downstream task of sentiment as we apply negation on predicted in-scope tokens for sentiment. The results show that our proposed model is comparable to the BiLSTM model for sentences with gold cues that have an annotated scope, but our model does not require annotated data. For BiLSTM, we used the implementation provided by the authors [2]. We also implemented a punctuation model that marks as negated all terms from a negation cue to the next punctuation. Fancellu et al. (2017) mentioned punctuation alone as a strong predictor for negation scope detection task for a majority previous of negation corpora. Notably, it performs poorly on our data as our scope is more restricted. We next show that having a restricted scope is beneficial to the antonym based negation sentiment prediction.

	Punctuation	**BiLSTM**	**Proposed**
In-scope (F)	0.66	0.88	0.85
Out-scope (F)	0.87	0.97	0.97
PCS	0.52	0.72	0.72

Table 7: Negation classifier performance for scope detection with gold cues and scope.

5 Sentiment with Negation Detection Pipeline

Here we show the integration of predicted negation scope in sentiment prediction pipeline. We begin with an overview of the data preprocessing, features and modeling, followed by our experimental setup and results. Finally, a comparison of the prediction performance of different systems is presented.

5.1 Experimental Method

From our tweet collection, we discarded tweets containing images and Non-English characters and anonymized all user and company handles, giving a dataset of 21746 tweets. A sentiment annotation task was run on a data annotation platform. Each tweet was initially annotated by 5 annotators using a 4 point (0 to 3) Likert scale (Likert, 1932) indicating `Not-At-All`, `Slight`, `Moderate` and `Very` about their perception on the sentiment for a given tweet. We used a set of gold standard questions to filter out the bad annotators, computed the average score for each label, and assigned the maximum score. A tweet is assigned a sentiment label if the maximum score for that label is greater than 1 else it is discarded from the study, giving a labeled dataset of 17779 tweets. To compute the inter-annotator agreement, first we measured what percentage of the annotators out of 5 contributed to the final sentiment label and then took the average over all the tweets giving a 78.8% inter-annotator agreement.

5.1.1 Data Pre-processing

A cleaning module is incorporated to reduce sparsity when generating word-based features. We replaced all links/URL by a keyword URL, removed # from the hashtags, replaced all @*mentions*, and replaced emojis and emoticons with the word explanation. An entity recognition module is run to replace the identified entity using a keyword "ENT".

5.1.2 Features

TFIDF-based unigram features.
Existence of consecutive question and exclamation marks and capitalized words.
Emotion lexicon features: A count of the number of words in each of the 8 emotion classes from

[2] `https://github.com/ffancellu/NegNN`

the NRC emotion lexicon (Mohammad and Turney, 2010)
Sentiment lexicons used:
Bing Liu's Opinion Lexicon (Ding et al., 2008); *The MPQA Subjectivity Lexicon* (Wilson et al., 2005); *Sentiment140 Lexicon* (Kiritchenko et al., 2014b); *NRC Hashtag Sentiment Lexicon* (Kiritchenko et al., 2014b);
For a given tweet, we computed minimum, maximum, average and summation of positive and negative scores of the words in the tweet that lies within a negation scope, and the average sentiment score of the last word in the tweet.
Negation handling: Append "NOT_" to each word in the scope.

5.1.3 SVM Evaluation

Libsvm (Chang and Lin, 2011) is used to implement the SVM classifier. The tweet annotated dataset was divided into a train and test set (see Table 8 for the distribution). The training set was further split into a ratio of 85:15 for the validation set. The three parameters w1, w2 and C were tuned using the validation set. The variables w1 and w2 are the penalty associated to a class and C is the regularization. Table 8 shows the evaluation metric using Precision, Recall and F measure for each class in the test set.

We do not find a major difference for SVMs(w/o negation). This is in spite of using the standard features such as prefixing the tokens in scope with a keyword *NOT_* and changing the polarity of the sentiment-bearing words using sentiment lexicons as described in previous work. A possible reason is that customer service domain is more negative as compared to general review domain See Section 6 for detailed analysis.

5.2 Neural Network Evaluation

Baseline 1: Our first baseline is a single layer CNN as used in (Kim, 2014). The model consists of a 1D convolution layer of window size 2 and 64 different filters. The convolution layer takes as input the GloVe embeddings. Max pooling layer is used to reduce the output dimensionality but keep the most salient information.
Baseline 2: (Wang et al., 2016) presented a jointed CNN and LSTM architecture. The features generated from convolution and pooling operation can be viewed as local features similar to ngrams but cannot handle long term dpendencies. LSTM can handle CNN's limitation by preserving historical information for a long period of time. Using this as a motivation, we included a convolutional layer and max pooling layer before the input is fed into an LSTM. A bidirectional LSTM layer is stacked on the convolutions layer and the tweet representation is taken to the fully connected network.
Proposed Negation + Antonym CNN-LSTM : We modified the sentence representation learned by replacing a word in the negation scope with it's antonym. Using antonyms would reduce the Out-of-Vocabulary words as compared to prefixing a word with "NOT_" for learning word representations. Replacing all the words upto punctuation with antonyms could entirely change the sentence meaning and hence this required a more restricted and accurate scope detection. We get the predicted scopes from the scope detection model described in Section 4. The antonym list is obtained from AntNET (Rajana et al., 2017)

For the NN-based approaches, 20% data is used for validation and we save the model weights only if the validation accuracy improves. The outputs of the LSTM are fed through a sigmoid layer for binary classification. Regularization is performed by using a drop-out rate of 0.2 in the drop-out layer. The model is optimized using the (Kingma and Ba, 2014) optimizer. The deep network was implemented using the Keras package (Chollet et al., 2015). Hyper-parameter optimization for the neural network is performed using Hyperas, a python package, based on hyperopt (Bergstra et al., 2015). Results in Table 8 show that the antonym based learned representations are more useful for sentiment task as compared to prefixing with NOT_. The proposed CNN-LSTM-Our-neg-Ant improves upon the simple CNN-LSTM-w/o neg. baseline with F1 scores improving from 0.72 to 0.78 for positive sentiment and from 0.83 to 0.87 for negative sentiment. Hence Negation coupled with antonyms improves the sentiment prediction for a customer service domain.

6 Discussion

In this section, we aim to show the particularities of our dataset, suggesting the reasons why negation detection did not improve the performance of the lexicon-based SVM when previous work had seen huge performance gains, and intuitions on how the antonym based method gives improvement.

- Class Distribution

	Positive Sentiment		
Classifier	Precision	Recall	Fscore
SVM-w/o neg.	0.57	0.72	0.64
SVM-Punct. neg.	0.58	0.70	0.63
SVM-our-neg.	0.58	0.73	0.65
CNN	0.63	0.83	0.72
CNN-LSTM	0.71	0.72	0.72
CNN-LSTM-Our-neg-Ant	**0.78**	**0.77**	**0.78**
	Negative Sentiment		
	Precision	Recall	Fscore
SVM-w/o neg.	0.78	0.86	0.82
SVM-Punct. neg.	0.78	0.87	0.83
SVM-Our neg.	0.80	0.87	0.83
CNN	0.88	0.72	0.79
CNN-LSTM.	0.83	0.83	0.83
CNN-LSTM-our-neg-Ant	**0.87**	**0.87**	**0.87**
	Train		Test
Positive tweets	5121		1320
Negative tweets	9094		2244

Table 8: Sentiment classification evaluation, using different classifiers on the test set.

Our customer service dataset has a much larger number of negative tweets while the benchmark sentiment dataset used in most of the previous systems has positive class as the majority class (Kiritchenko et al., 2014b; Reitan et al., 2015; Nakov et al., 2013; Mohammad et al., 2013; Tang et al., 2014). Additionally, Reitan et al. (2015) reported that the classifier struggles with negative class prediction. A F-measure of 0.533 and 0.323 is reported by Reitan et al. (2015) and Councill et al. (2010) respectively, on negative class prediction. In contrast, our baseline classifier achieves a much higher F score of 0.82 on the negative class.

- Cue Word Distribution.
  The conversation negation corpus is annotated for both actual negation cues and sentiment. To see if there exists some correlation between the number of cues and sentiment, we calculated the percentage of positive and negative tweets with more than one cue. 19% of positive tweets contain more than one negation cue while for the negative class it is 48%. Though we need more evidence to support, but it is possible that the number of negation cues in these conversations is a strong indicator of negative class, hence the SVM based classifier had better prediction on negative class detection.

- Sarcasm and Irony
  Results in Table 8 show that the classifier struggles with positive class precision. A sentiment study on user-generated content by (Sarmento et al., 2009; Carvalho et al., 2009) has similar class distribution and results to ours. The sentences expressing negative opinions is almost the double of those expressing positive opinions and the precision of identifying negative opinions ($\approx$ 89%) is significantly higher than the precision of identifying positive opinions ($\approx$60%). They confirm the relevance of irony for sentiment analysis by an error analysis of their present classifier stating that a large proportion of misclassifications ($\approx$35%) derive from their system's inability to account for irony. We then performed some manual error analysis on the incorrect positive predictions for SVM and observed that some of the incorrect predictions were actually sarcastic, see Table 9. To get an insight on how our method improves these types of predictions, consider the example in Row2 in Table 9. It was predicted as positive by SVM, CNN and LSTM due to the positive word "Awesome". Our method detects negation cue word **"not"** with **"able"** in it's scope. The antonym dictionary then is used to replace **"able"** with **"incapable"**. Having a strong negative word corrects the prediction to negative. These results indicate that there is room for improvement for the positive class but negation handling may not be enough. A combination of negation and sarcasm may be a useful direction to explore in future for customer service conversations.

S.No	Tweet
1	Hi @username - Love u. I'd recommend not displaying the early bird button in the app if it's broken and not working
2	looks like I won't be able to vote because the train is running late. Awesome

Table 9: Negative sarcastic examples.

7 Conclusion and Future Work

This paper presented an approach to negation cue and scope detection in customer service interactions on Twitter and the impact of using this component for sentiment detection. We refined the annotation guidelines for scope representation, gathering a dataset of 2000 labeled tweets. Our rule based approach based on syntactic constituents does not require annotated scope data for training, but performs comparable to state of the art BiLSTM. To evaluate the effectiveness of negation modeling on sentiment detection, we performed experiments using both an SVM and CNN-LSTM Architecture. There was no significant improvement between the two lexicon based SVMs (with/without negation handling). The proposed antonym based negation for CNN-LSTM outperformed both a CNN and a combination CNN-LSTM that did not handle negation. The result and error analysis shows that customer service interactions have higher frequency of negation cues, are more skewed towards negative class, and are sometimes sarcastic. In future, we plan to study other language phenomenon such as sarcasm in combination with negation.

References

James Bergstra, Brent Komer, Chris Eliasmith, Dan Yamins, and David D Cox. 2015. Hyperopt: a python library for model selection and hyperparameter optimization. *Computational Science and Discovery*, 8(1):014008.

D. Biber. 1999. *Longman Grammar of Spoken and Written English*. Grammar Reference Series. Longman.

Jorge Carrillo de Albornoz, Laura Plaza, Alberto Díaz, and Miguel Ballesteros. 2012. UCM-I: A rule-based syntactic approach for resolving the scope of negation. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics, *SEM 2012, June 7-8, 2012, Montréal, Canada.*, pages 282–287. Association for Computational Linguistics.

Paula Carvalho, Luís Sarmento, Mário J. Silva, and Eugénio de Oliveira. 2009. Clues for detecting irony in user-generated contents: Oh...!! it's "so easy" ;-). In *Proceedings of the 1st International CIKM Workshop on Topic-sentiment Analysis for Mass Opinion*, TSA '09. ACM.

Chih-Chung Chang and Chih-Jen Lin. 2011. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3):27:1–27:27.

François Chollet et al. 2015. Keras. `https://keras.io`.

Isaac G. Councill, Ryan McDonald, and Leonid Velikovich. 2010. What's great and what's not: Learning to classify the scope of negation for improved sentiment analysis. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, NeSp-NLP '10, pages 51–59, Stroudsburg, PA, USA. Association for Computational Linguistics.

Xiaowen Ding, Bing Liu, and Philip S Yu. 2008. A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 international conference on web search and data mining*, pages 231–240. ACM.

Martine Enger, Erik Velldal, and Lilja Øvrelid. 2017. An open-source tool for negation detection: a maximum-margin approach. In *Proceedings of the EACL workshop on Computational Semantics Beyond Events and Roles (SemBEaR)*, pages 64–69, Valencia, Spain.

Federico Fancellu, Adam Lopez, Bonnie Webber, and Hangfeng He. 2017. *Detecting negation scope is easy, except when it isn't*, pages 58–63. Association for Computational Linguistics.

Federico Fancellu, Adam Lopez, and Bonnie L. Webber. 2016. Neural networks for negation scope detection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

C.J. Fillmore. 1963. The position of embedding transformations in a grammar. *Word*, 19.

T. Givón. 1993. *English grammar: a function-based introduction*. Number v. 2 in English Grammar: A Function-based Introduction. J. Benjamins Pub. Co.

M. Gotti, M. Dossena, and R. Dury. 2008. *English Historical Linguistics 2006: Selected papers from the fourteenth International Conference on English Historical Linguistics (ICEHL 14), Bergamo, Volume I: Syntax and Morphology*. English Historical Linguistics 2006. John Benjamins Publishing Company.

Laurence Horn. 1989. *A Natural History of Negation*. University of Chicago Press.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04. ACM.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1746–1751.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Svetlana Kiritchenko, Xiaodan Zhu, and Saif M. Mohammad. 2014a. Sentiment analysis of short informal texts. *J. Artif. Intell. Res.*, 50:723–762.

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

Natalia Konstantinova, Sheila CM De Sousa, Noa P Cruz Díaz, Manuel J Mana López, Maite Taboada, and Ruslan Mitkov. 2012. A review corpus annotated for negation, speculation and their scope. In *LREC*, pages 3190–3195.

Emanuele Lapponi, Erik Velldal, Lilja Ovrelid, and Jonathon Read. 2012. Uio2: Sequence-labeling negation using dependency features. In *Proceedings of 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 '12. Association for Computational Linguistics.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology*.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 3111–3119.

Yasuhide Miura, Shigeyuki Sakaki, Keigo Hattori, and Tomoko Ohkuma. 2014. Teamx: A sentiment analyzer with enhanced lexicon mapping and weighting scheme for unbalanced data. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 628–632. Association for Computational Linguistics.

Saif Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the 7th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2013, Atlanta, Georgia, USA, June 14-15, 2013*, pages 321–327. The Association for Computer Linguistics.

Saif M Mohammad and Peter D Turney. 2010. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 26–34. Association for Computational Linguistics.

Karo Moilanen and Stephen Pulman. 2007. Sentiment composition. In *Proceedings of Recent Advances in Natural Language Processing (RANLP 2007)*, pages 378–382.

Roser Morante and Eduardo Blanco. 2012. *sem 2012 shared task: Resolving the scope and focus of negation. In *Proceedings of 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 '12, pages 265–274, Stroudsburg, PA, USA. Association for Computational Linguistics.

Roser Morante and Walter Daelemans. 2009. A metalearning approach to processing the scope of negation. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, CoNLL '09, pages 21–29, Stroudsburg, PA, USA. Association for Computational Linguistics.

Roser Morante, Anthony M. L. Liekens, and Walter Daelemans. 2008. Learning the scope of negation in biomedical texts. In *2008 Conference on Empirical Methods in Natural Language Processing, EMNLP 2008, Proceedings of the Conference, 25-27 October 2008, Honolulu, Hawaii, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 715–724. ACL.

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)*, volume 2, pages 312–320.

Woodley Packard, Emily M. Bender, Jonathon Read, Stephan Oepen, and Rebecca Dridan. 2014. Simple negation scope resolution through deep parsing: A semantic solution to a semantic problem. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 69–78, Baltimore, USA.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Longman, London.

Sneha Rajana, Chris Callison-Burch, Marianna Apidianaki, and Vered Shwartz. 2017. Learning antonyms with paraphrases and a morphology-aware neural network. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 12–21. Association for Computational Linguistics.

Jonathon Read, Erik Velldal, Lilja Øvrelid, and Stephan Oepen. 2012. Uio1: Constituent-based discriminative ranking for negation resolution. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics, *SEM 2012, June 7-8, 2012, Montréal, Canada.*, pages 310–318. Association for Computational Linguistics.

Johan Reitan, Jorgen Faret, Bjorn Gamback, and Lars Bungum. 2015. Negation scope detection for twitter sentiment analysis. In *WASSA@EMNLP*.

Luís Sarmento, Paula Carvalho, Mário J. Silva, and Eugénio de Oliveira. 2009. Automatic creation of a reference corpus for political opinion mining in user-generated content. In *Proceedings of the 1st International CIKM Workshop on Topic-sentiment Analysis for Mass Opinion*, TSA '09, pages 29–36, New York, NY, USA. ACM.

Bonggun Shin, Timothy Lee, and Jinho D. Choi. 2017. Lexicon integrated CNN models with attention for sentiment analysis. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, WASSA@EMNLP 2017, Copenhagen, Denmark, September 8, 2017*, pages 149–158.

György Szarvas, Veronika Vincze, Richárd Farkas, and János Csirik. 2008. The bioscope corpus: Annotation for negation, uncertainty and their scope in biomedical texts. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, BioNLP '08, pages 38–45, Stroudsburg, PA, USA. Association for Computational Linguistics.

Duyu Tang, Furu Wei, Bing Qin, Ting Liu, and Ming Zhou. 2014. Coooolll: A deep learning system for twitter sentiment classification. In *Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval@COLING 2014, Dublin, Ireland, August 23-24, 2014.*, pages 208–212.

Gunnel Tottie. 1991. *Negation in English speech and writing : a study in variation*. San Diego : Academic Press. (Quantitative analyses of linguistic structure series).

Erik Velldal. 2011. Predicting speculation: A simple disambiguation approach to hedge detection in biomedical literature. *Journal of Biomedical Semantics*, 2(5). Supplement to the Fourth International Symposium on Semantic Mining in Biomedicine.

Erik Velldal, Lilja Øvrelid, Jonathon Read, and Stephan Oepen. 2012. Speculation and negation: Rules, rankers, and the role of syntax. *Computational Linguistics*, 38(2):369–410.

Xingyou Wang, Weijie Jiang, and Zhiyong Luo. 2016. Combination of convolutional and recurrent neural network for sentiment analysis of short texts. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 2428–2437.

James Paul White. 2012. Uwashington: Negation resolution using machine learning methods. 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 335–339. Association for Computational Linguistics.

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*, NeSp-NLP '10, pages 60–68, Stroudsburg, PA, USA. Association for Computational Linguistics.

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*, HLT '05, pages 347–354, Stroudsburg, PA, USA. Association for Computational Linguistics.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2009. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Comput. Linguist.*, 35(3):399–433.

Wenpeng Yin and Hinrich Schütze. 2015. Multichannel variable-size convolution for sentence classification. In *Proceedings of the 19th Conference on Computational Natural Language Learning, CoNLL 2015, Beijing, China, July 30-31, 2015*, pages 204–214.