Agree to Disagree: Improving Disagreement Detection with Dual GRUs

Sushant Hiray*†, Venkatesh Duppada*‡
Seernet Technologies, LLC
Email: † sushant.hiray@seernet.io, ‡ venkatesh.duppada@seernet.io

***Abstract*—This paper presents models for detecting agreement/disagreement in online discussions. In this work we show that by using a Siamese inspired architecture to encode the discussions, we no longer need to rely on hand-crafted features to exploit the meta thread structure. We evaluate our model on existing online discussion corpora ABCD, IAC and AWTP. Experimental results on ABCD dataset show that by fusing lexical and word embedding features, our model achieves the state of the art performance of 0.804 average F1 score. We also show that the model trained on ABCD dataset performs competitively on relatively smaller annotated datasets (IAC and AWTP).**

1. Introduction

The rise of various discussion forums and social media websites has given people a lot of avenues to express their opinions. As multiple people join a particular discussion, participants often agree or disagree with views presented by others. Mining the agreement and disagreement (denoted (dis)agreement) signals helps detect presence of disputes, ideological stance of the participants [1] and unravel beliefs shaping the opinion in general. This can further be useful for detecting subgroups [2] [3], analyzing how well a new product is being received or analyzing the mood to predict the trends on stock markets [4].

In this work, we explore a Siamese [5] inspired deep neural network to detect the presence of (dis)agreement in online discussions between two posts, the quote and the response (Q-R pairs [6]). In this framework, the same neural network encoder is applied to two input sentences individually, so that both of the two sentences are encoded into sentence vectors in the same embedding space. Prior work in this problem primarily focused on using handcrafted features to exploit the meta thread structure. We show that by training on a sufficiently large dataset (ABCD) we can bypass the need for designing handcrafted features. Thus, the classifier can be used for (dis)agreement detection between any two posts, even when the underlying hierarchical relationship between the Q-R pairs isn't available. To the best of our knowledge, this is the first work to investigate detection of (dis)agreement using sentence based encoding.

* *These authors contributed equally to this work*

We detect (dis)agreement by performing a 3-way classification (agreement/disagreement/none) between the Q-R pairs on several existing annotated datasets. Due to the lack of a standard dataset, some prior work focused primarily on 2-way classification (agreement/disagreement).

In the following sections, we first discuss related work. Section 3 describes the datasets used for evaluation. Section 4 describes the various features used in the classifier. In section 5, we explain the model architecture. Section 6 details the experiments performed and their corresponding results. Section 7 performs error analysis on the results from the proposed system. Finally, we conclude in section 8 and suggest relevant future work.

2. Related Work

Previous work in this field focused a lot on spoken dialogues. [7], [8], [9] used spurt level agreement annotations from the ICSI corpus [10]. [11] presents detection of agreements in multi-party conversations using the AMI meeting corpus [12]. [13] presents a conditional random field based approach for detecting agreement/disagreement between speakers in English broadcast conversations

Recently, researchers have turned their attention towards (dis)agreement detection in online discussions. The prior work was geared towards performing 2-way classification of agreement/disagreement. [14] used various sentiment, emotional and durational features to detect local and global (dis)agreement in discussion forums. [15] performed (dis)agreement on annotated posts from the Internet Argument Corpus (IAC) [6]. They investigated various manual labelled features, which are however difficult to reproduce as they are not annotated in other datasets. To benchmark the results, we've also incorporated the IAC corpus in our experiments. Quite recently, [16] proposed a 3-way classification by exploiting meta-thread structures and accommodation between participants. They also proposed a naturally occurring dataset ABCD (Agreement by Create Debaters) which was about 25 times larger than prior existing corpus. We've trained our classifier on this larger dataset. [17] proposed (dis)agreement detection with an isotonic Conditional Random Fields (isotonic CRF) based sequential model. [18] proposed features motivated by theoretical predictions to perform (dis)agreement detection. However, they've used hand-crafted patterns as features and these features miss

978-1-5386-0680-3/17/$31.00

few real world scenarios reducing the performance of the classifier.

(Dis)agreement detection is related to other similar NLP tasks like stance detection and argument mining but is not exactly the same. Stance detection is the task of identifying whether the author of the text is in favor or against or neutral towards a target, while argument mining focuses on tasks like automatic extraction of arguments from free text, argument proposition classification and argumentative parsing [19] [20]. Recently there are studies on how people back up their stances when arguing where comments are classified as either attacking or supporting a set of pre-defined arguments [21]. These tasks (stance detection, argument mining) are not independent but have some common features because of which they are benefited by common building blocks like sentiment detection, textual entailment and sentence similarity [21] [22].

3. Data

In this work, we focus on 3-way classification (agreement/disagreement/none) between quote-response (Q-R) pairs for 3 prior existing in-domain datasets. These are described in the subsequent sub-sections.

3.1. Agreement by Create Debaters (ABCD)

The ABCD corpus [16] was curated from Create Debate website[1] where users can start a debate by asking a question. Although the website can support open ended as well as multiple sided debates, the corpus comprises only of the for-against debates. The corpus is annotated as follows: the side label corresponding to each post (*response*) determines whether the user agrees or disagrees with the previous post (*quote*). If the authors of both the posts are different, then they agree if the side labels are same or otherwise disagree. If the authors of both the posts is same, it is labeled as none as it implies that it is in continuation of the previous post. Also, the first post in a debate is usually setting up the premise of the debate, so it doesn't have a side attached to it. Hence all the Q-R pairs with the quote as the first post are labeled as none. Table 1 shows example Q-R pairs for each label type.

3.2. Agreement in Wikipedia Talk Pages (AWTP)

AWTP [23] is formatted in the same way as ABCD. Post-reply pairs are manually annotated with their (dis)agreement stance. Also additional mode information indicates the manner in which agreement or disagreement is expressed. The datasource for AWTP comprised primarily of Wikipedia Talk Pages and LiveJournal postings. Table 2 provides additional statistics for the corpus.

1. http://www.createdebate.com/

TABLE 1. EXAMPLES OF AGREEMENT/DISAGREEMENT/NONE IN QUOTE-RESPONSE PAIRS IN ABCD DATASET

Label	Post
None	*Quote*: Is Scientology a real religeon? Or is it a fake money making gimmick?
	Response: All religions are fake, there's an argument to be made the vast majority are money making gimmicks. Scientology is no more outlandish than any of the more widespread religions.
Agree	*Quote*: I am against suicide because you are basically not only harming yourself, but everyone else around you. Let's not mention it is a cowards way out. I also have a religious but I CAN explain that reason.
	Response: So true man people only harm the people they love by dying. I am not religious and religiously and non-religiously suicide is wrong.
Disagree	*Quote*: The majority of the information learned in school is irrelevant to real world skills. Besides, in a voluntary setting, most children would go to school via parents demands where school choice would be much more abundant.
	Response: Children learn math which is relevant, children learn history which is relvant, children learn the releveant languge to their country, children learn foreign languages which imporoves economic opportunites. My one friend grew up in Baghdad, Iraq, and they don't play when it comes to education. He started learning English in the 3rd grade I think through graduation which helped his economic ooportunities, and he is an artchitect so the math helped. Please excuse my typos. I have a learning disability.

TABLE 2. UNIQUE ANNOTATION COUNTS FOR AWTP DATASET

	Wikipedia			**LiveJournal**		
	Agree	Disagree	None	Agree	Disagree	None
Train	219	471	703	390	83	0
Dev	69	101	24	0	0	0
Test	62	107	79	0	0	0

3.3. Internet Argument Corpus (IAC)

The Internet Argument Corpus (IAC) [6] is a collection of corpora for research in political debate on internet forums. It consists of ~11,000 discussions, ~390,000 posts, and some ~73,000,000 words. It includes topic annotations, response characterizations, and stance. The 4forum posts were annotated using Mechanical Turk. The annotators were provided with a Q-R pair and they indicated the level of (dis)agreement on a scale of [-5, 5]. However, not all posts in a thread were annotated for (dis)agreement and roughly 6000 valid Q-R pairs were extracted. In accordance with prior work of this corpus [15], [16], [18], we converted the scalar values into corresponding (dis)agreement as follows: [-5, -1] is tagged as disagreement, [-1, 1] is tagged as none, [1, 5] is tagged as agreement. In case multiple annotators have tagged the same post, we combine them as follows. None annotations are ignored unless there are no other (dis)agreement tags. In all other cases, average annotation score is used as the final score of the post.

4. Feature Extraction

In this section we briefly mention the features experimented in this work.

4.1. Word Vectors

In the recent times distributed representations of words [24] (word2vec [25], GloVe [26]) has shown promise in many NLP tasks and is the driver for success of deep learning in NLP [27]. Word vectors encode semantics in low dimensional space and can be used efficiently for various NLP tasks [28] [29]. For this task of (dis)agreement classification, we use GloVe embeddings of 300 dimensions trained on Common Crawl with 840 billion tokens, 2.2 million vocabulary.

4.2. Lexicons

We used affect, sentiment, emotion, opinion lexicons for feature extraction because in many of the online discussions forums people tend to argue with emotion and opinion about a particular topic to convey their stance or belief. Making use of these affect lexicons will help us in classifying if a response (dis)agrees with quote or not. Prior work [14] using these lexicons have shown them to give good results. Among lexical features we used the following.

AFINN [30] word list are manually rated for valence with an integer between -5 (Negative Sentiment) and +5 (Positive Sentiment). Bing Liu [31] opinion lexicon extract opinion on customer reviews. +/-EffectWordNet [32] by MPQA group are sense level lexicons. The NRC Affect Intensity [33] lexicons provide real valued affect intensity. NRC Word-Emotion Association Lexicon [34] contains 8 sense level associations (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and 2 sentiment level associations (negative and positive). Expanded NRC Word-Emotion Association Lexicon [35] expands the NRC word-emotion association lexicon for twitter specific language. NRC Hashtag Emotion Lexicon [36] contains emotion word associations computed on emotion labeled twitter corpus via Hashtags. NRC Hashtag Sentiment Lexicon and Sentiment140 Lexicon [37] contains sentiment word associations computed on twitter corpus via Hashtags and Emoticons. SentiWordNet [38] assigns to each synset of WordNet three sentiment scores: positivity, negativity, objectivity. Negation lexicons collections are used to count the total occurrence of negative words. The Linguistic Inquiry Word Count (LIWC) [39] categorizes the words we use in everyday language to reveal our thoughts, feelings, personality, and motivations.

For each word in the sentence we calculate various metrics using these lexicons like number of negation words in a sentence, average negative and positive sentiment of the sentence etc. and use these as lexical feature vector to the system. The lexicon feature extractor was inspired from [40].

5. System Description

The task of (dis)agreement classification from Q-R pair comes under a much broader category of sentence pair modelling. A lot of NLP tasks like natural language inference, textual entailment, answer selection, paraphrase identification etc involve modelling a pair of sentences so that they perform well on a particular task or multitude of such tasks [41]. For this task we used an architecture which is inspired by Siamese network [5] and Stanford Natural Language Inference model [42] with identical networks encoding Q-R pair.

For each Q-R pair we extract two sets of features. First, GloVe word embeddings are fed to Gated Recurrent Units [43] to create a sentence embedding. Second, from each text a lexical feature vector is extracted as mentioned in subsection 4.2. Both these sentence embeddings from Q-R pairs are concatenated and then fed into fully connected layers to do 3 way classification. Figure 1 show the architecture of our model.

5.1. System Parameters

We have used relu non-linearity, dropout [44] for regularization and batch normalization [45] for accelerating training. The network is optimized with adam [46] optimizer with learning rate of 0.001. We have plotted sequence length of Q-R pairs in Figure 2 to better estimate the maximum sequence length for GRU layer. The average sequence length of quotes and responses is 46.28 and 64.935 respectively. The Impact of maximum sequence length of the text input to GRUs is computed and can be found in section 6. We used keras [47] deep learning framework to train our models.

6. Experiments

In this section, we evaluate our model on 3-way (dis)agreement classification. The general settings of the model have been defined in sub-section 5.1. In the upcoming sub-sections, we explore the variants of our model and compare our model with the state of the art models on benchmark datasets from section 3.

6.1. Features Used

We implemented three variants of the architecture: only lexical features are used, only GloVe embeddings are used and finally where both of them are used. As is evident and hypothesized, the model trained with lexical and word embeddings gave the best results. It is interesting to see that using just the lexicons, the model beat the previous best model which signifies the importance of lexical features. However, using only lexicons as features will suffer from all the disadvantages of any bag-of-words model. This is primarily because it cannot encode the temporal nature of language. This is where gated recurrent units or in general recurrent neural networks come into play. GRU's can

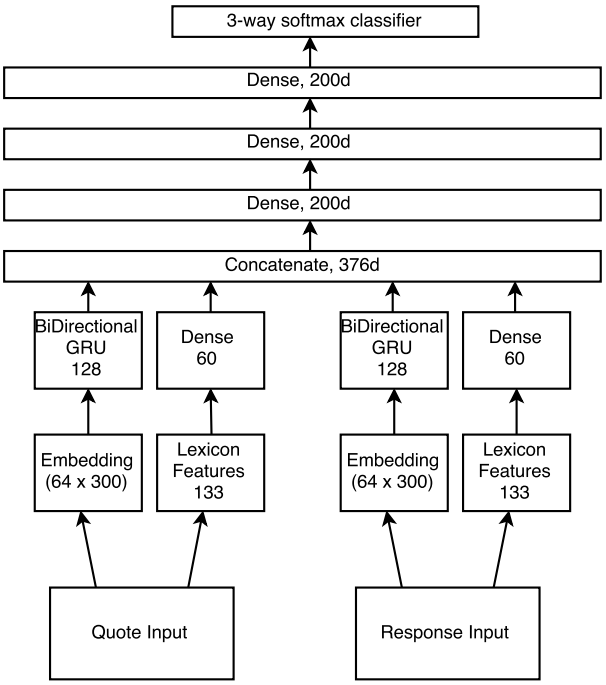

Figure 1. System Architecture

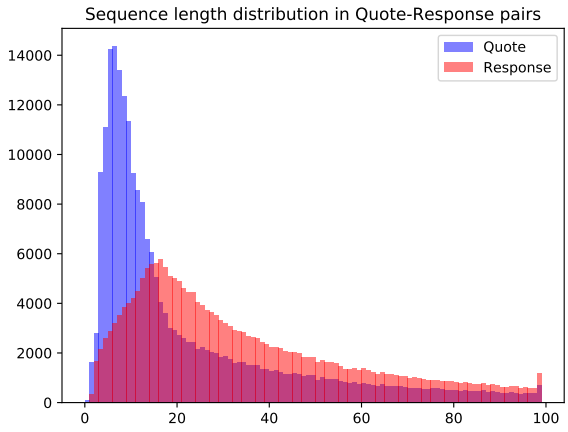

Figure 2. Distribution of sequence length in Q-R pairs. The graph shows the number of posts v/s sequence length.

successfully encode the temporal nature. Thus, by fusing both the GRU encoded embeddings and the lexicons, we achieved the state of the art results on the ABCD dataset by beating the previous best model's [16] average F1 score with margin of 4% percentage. Table 3 compares variants of the proposed system with the existing state of the art.

6.2. Maximum Input Sequence Length

We investigate the impact of varying the maximum input sequence length in Table 4. The results highlight some interesting insights. Since we are dealing with discussion posts, the average post size is larger than a few sentences. Hence, as we increase the maximum sequence length, the increase in performance is justified. However, once we increase the size after a certain threshold, a lot of the input sequences need to be padded with zeroes, thus causing a drop in performance.

TABLE 3. ANALYZING THE IMPACT OF CHANGING THE FEATURE VECTOR ON ABCD CORPUS. ENTRIES MARKED WITH **BOLD** ARE THE BEST PERFORMING VARIANTS AND THOSE MARKED WITH - ARE NOT AVAILABLE.

System	Precision	Recall	Weighted F1 Score
SOTA [16]	0.776	-	-
Lexicons	0.788	0.798	0.789
GRU	0.792	0.798	0.794
GRU + Lexicons	**0.812**	**0.815**	**0.804**

TABLE 4. INVESTIGATING THE IMPACT OF VARYING THE MAXIMUM SEQUENCE LENGTH ON THE OVERALL MODEL PERFORMANCE. ENTRIES MARKED WITH **BOLD** ARE THE BEST PERFORMING VARIANTS

Sequence Length	Precision	Recall	Weighted F1 Score
32	0.810	0.813	**0.806**
64	**0.812**	**0.815**	0.804
128	0.805	0.808	0.796

6.3. Transfer Learning

In the recent times, with the availability of huge amount of data a new paradigm in machine learning called Transfer Learning [48] has come into play. Transfer learning is the improvement of learning in a new task through the transfer of knowledge from a related task that has already been learned. Here we've applied transfer learning technique to smaller datasets and achieved competitive results. Table 5 and 6 enlist the results of various model architectures explored for testing the effectiveness of transferring learning from ABCD dataset to the smaller annotated datasets IAC and AWTP. The variants explored are as follows. *Direct*, where the model trained on ABCD model (referred as pre-trained model) is tested directly on the smaller dataset. *Tuning*: the model is seeded with the weights from the pre-trained model and trained with the smaller dataset. *Transfer*: The last 2 layers from the pre-trained model are stripped and replaced with new dense layers of size 100 and 50 and the model is trained on the smaller dataset. *Re-train last 2/3 layers*: All but last-2/3 layers of the pre-trained model are frozen and the remaining layers are trained on the smaller dataset.

7. Error Analysis

The ABCD dataset is scraped from online debate forum, Create Debate and is automatically labelled. This way of collecting Q-R pairs is not perfect and suffers from the following problems: people may be on the same/different side of debate but disagree/agree on some points (example 1 in Table 7) as the sides are for topic level not post

TABLE 5. TUNING MODEL TRAINED ON ABCD DATASET FOR IAC DATASET. ENTRIES IN **BOLD** ARE THE BEST PERFORMING VARIANTS AND THOSE MARKED WITH - ARE NOT AVAILABLE. THE MODEL VARIANTS ARE EXPLAINED IN SUB-SECTION 6.3

Model	Precision	Recall	Weighted F1 Score
SOTA [16]	-	-	**0.578**
Direct	0.473	0.364	0.285
Tuning	0.530	**0.572**	0.450
Transfer	0.508	0.559	0.465
Re-train last 2 layers	**0.531**	**0.572**	0.460
Re-train last 3 layers	0.523	0.570	0.428

TABLE 6. TUNING MODEL TRAINED ON ABCD DATASET FOR AWTP DATASET. ENTRIES IN **BOLD** ARE THE BEST PERFORMING VARIANTS AND THOSE MARKED WITH - ARE NOT AVAILABLE. THE MODEL VARIANTS ARE EXPLAINED IN SUB-SECTION 6.3

Model	Precision	Recall	Weighted F1 Score
SOTA [16]	-	-	0.389
Direct	0.434	0.445	0.389
Tuning	0.515	0.470	0.464
Transfer	0.525	0.437	0.412
Re-train last 2 layers	**0.534**	0.457	0.447
Re-train last 3 layers	0.477	**0.546**	**0.486**

or sentence level, off-topic social chitter-chatter on debate forum (example 2 in Table 7) etc. Our analysis indicated that in quite a few cases, the error in classification was in fact a case of incorrect label.

TABLE 7. HARD EXAMPLES OF ABCD DATASET

Label	Post
agree	*Quote*: JessHall01 my parents do not deserve it, they treat me like sh and that's unecessary. NO parents should even romotely lay their hands on a child. EVER.
	Response: kamranw I agree with you there. NO parent should hit their child. I completely feel your pain if that is the case. I would be curious to know how they treat you like shit though. Not sure how that is relevent to minors being sexual active either.
disagree	*Quote*: joecavalry I tell my kids to hurry up and eat their breakfast so that we can get to school on time. My little one is so slow that I tell her she eats breakslow ;)
	Response: Morgie7171 AWWW thats so cute..... hahahahahahahahahhaahahhha

8. Conclusion

We have trained a deep neural network with fusion of lexical and word vector based features to achieve state of the art results on 3-way (dis)agreement classification on the largest (dis)agreement classification dataset available till date (ABCD corpus). We've shown that by using this model, we no longer need to rely on hand-crafted features to exploit the meta-thread structure. We've also shown the benefit of transfer learning from pre-trained model on large corpora to achieve competitive results on small domain datasets. Till date, the research for (dis)agreement classification has not moved at the pace comparable to some other NLP tasks primarily because of unavailability of large standard dataset. Although ABCD dataset is large enough it suffers from many issues as it uses naturally occuring labels. We plan to enhance the ABCD dataset by performing semi-supervised tagging of labels by training models on hand-annotated datasets. With the availability of a large standard dataset for (dis)agreement classification, we can try various recent advanced state of the art architectures for modelling sentences pairs [41], [49] to further improve the performance.

References

[1] S. Somasundaran and J. Wiebe, "Recognizing stances in online debates," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*. Association for Computational Linguistics, 2009, pp. 226–234.

[2] A. Hassan, A. Abu-Jbara, and D. Radev, "Detecting subgroups in online discussions by modeling positive and negative relations among participants," in *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*. Association for Computational Linguistics, 2012, pp. 59–70.

[3] A. Abu-Jbara, M. Diab, P. Dasigi, and D. Radev, "Subgroup detection in ideological discussions," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, 2012, pp. 399–409.

[4] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of computational science*, vol. 2, no. 1, pp. 1–8, 2011.

[5] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a" siamese" time delay neural network," in *Advances in Neural Information Processing Systems*, 1994, pp. 737–744.

[6] M. A. Walker, J. E. F. Tree, P. Anand, R. Abbott, and J. King, "A corpus for research on deliberation and debate." in *LREC*, 2012, pp. 812–817.

[7] D. Hillard, M. Ostendorf, and E. Shriberg, "Detection of agreement vs. disagreement in meetings: Training with unlabeled data," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003–short papers-Volume 2*. Association for Computational Linguistics, 2003, pp. 34–36.

[8] M. Galley, K. McKeown, J. Hirschberg, and E. Shriberg, "Identifying agreement and disagreement in conversational speech: Use of bayesian networks to model pragmatic dependencies," in *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2004, p. 669.

[9] S. Hahn, R. Ladner, and M. Ostendorf, "Agreement/disagreement classification: Exploiting unlabeled data using contrast classifiers," in *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*. Association for Computational Linguistics, 2006, pp. 53–56.

[10] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke *et al.*, "The icsi meeting corpus," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, vol. 1. IEEE, 2003, pp. I–I.

[11] S. Germesin and T. Wilson, "Agreement detection in multiparty conversation," in *Proceedings of the 2009 international conference on Multimodal interfaces*. ACM, 2009, pp. 7–14.

[12] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos *et al.*, "The ami meeting corpus," in *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, vol. 88, 2005.

[13] W. Wang, S. Yaman, K. Precoda, C. Richey, and G. Raymond, "Detection of agreement and disagreement in broadcast conversations," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*. Association for Computational Linguistics, 2011, pp. 374–378.

[14] J. Yin, P. Thomas, N. Narang, and C. Paris, "Unifying local and global agreement and disagreement classification in online debates," in *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*. Association for Computational Linguistics, 2012, pp. 61–69.

[15] R. Abbott, M. Walker, P. Anand, J. E. Fox Tree, R. Bowmani, and J. King, "How can you say such things?!?: Recognizing disagreement in informal political argument," in *Proceedings of the Workshop on Languages in Social Media*. Association for Computational Linguistics, 2011, pp. 2–11.

[16] S. Rosenthal and K. McKeown, "I couldn't agree more: The role of conversational structure in agreement and disagreement detection in online discussions." in *SIGDIAL Conference*, 2015, pp. 168–177.

[17] L. Wang and C. Cardie, "Improving agreement and disagreement identification in online discussions with a socially-tuned sentiment lexicon," *arXiv preprint arXiv:1606.05706*, 2016.

[18] A. Misra and M. A. Walker, "Topic independent identification of agreement and disagreement in social media dialogue," in *Conference of the Special Interest Group on Discourse and Dialogue*, 2013, p. 920.

[19] M.-F. Moens, E. Boiy, R. M. Palau, and C. Reed, "Automatic detection of arguments in legal texts," in *Proceedings of the 11th international conference on Artificial intelligence and law*. ACM, 2007, pp. 225–230.

[20] R. M. Palau and M.-F. Moens, "Argumentation mining: the detection, classification and structure of arguments in text," in *Proceedings of the 12th international conference on artificial intelligence and law*. ACM, 2009, pp. 98–107.

[21] F. Boltuzic and J. Snajder, "Back up your stance: Recognizing arguments in online discussions." in *ArgMining@ ACL*, 2014, pp. 49–58.

[22] S. M. Mohammad, P. Sobhani, and S. Kiritchenko, "Stance and sentiment in tweets," *ACM Transactions on Internet Technology (TOIT)*, vol. 17, no. 3, p. 26, 2017.

[23] J. Andreas, S. Rosenthal, and K. McKeown, "Annotating agreement and disagreement in threaded discussion." in *LREC*, 2012, pp. 818–822.

[24] J. Turian, L. Ratinov, and Y. Bengio, "Word representations: a simple and general method for semi-supervised learning," in *Proceedings of the 48th annual meeting of the association for computational linguistics*. Association for Computational Linguistics, 2010, pp. 384–394.

[25] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[26] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation." in *EMNLP*, vol. 14, 2014, pp. 1532–1543.

[27] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3128–3137.

[28] O. Levy, Y. Goldberg, and I. Ramat-Gan, "Linguistic regularities in sparse and explicit word representations." in *CoNLL*, 2014, pp. 171–180.

[29] O. Levy, Y. Goldberg, and I. Dagan, "Improving distributional similarity with lessons learned from word embeddings," *Transactions of the Association for Computational Linguistics*, vol. 3, pp. 211–225, 2015.

[30] F. Å. Nielsen, "A new anew: Evaluation of a word list for sentiment analysis in microblogs," *arXiv preprint arXiv:1103.2903*, 2011.

[31] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2004, pp. 168–177.

[32] Y. Choi and J. Wiebe, "+/-effectwordnet: Sense-level lexicon acquisition for opinion inference." in *EMNLP*, 2014, pp. 1181–1191.

[33] S. M. Mohammad, "Word affect intensities," *arXiv preprint arXiv:1704.08798*, 2017.

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

[35] F. Bravo-Marquez, E. Frank, S. M. Mohammad, and B. Pfahringer, "Determining word–emotion associations from tweets by multi-label classification," in *WI'16*. IEEE Computer Society, 2016, pp. 536–539.

[36] S. M. Mohammad and S. Kiritchenko, "Using hashtags to capture fine emotion categories from tweets," *Computational Intelligence*, vol. 31, no. 2, pp. 301–326, 2015.

[37] S. M. Mohammad, S. Kiritchenko, and X. Zhu, "Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets," *arXiv preprint arXiv:1308.6242*, 2013.

[38] S. Baccianella, A. Esuli, and F. Sebastiani, "Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining." in *LREC*, vol. 10, 2010, pp. 2200–2204.

[39] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: Liwc and computerized text analysis methods," *Journal of language and social psychology*, vol. 29, no. 1, pp. 24–54, 2010.

[40] V. Duppada and S. Hiray, "Seernet at emoint-2017: Tweet emotion intensity estimator," in *Proceedings of the Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA)*, Copenhagen, Denmark, 2017.

[41] W. Yin, H. Schütze, B. Xiang, and B. Zhou, "Abcnn: Attention-based convolutional neural network for modeling sentence pairs," *arXiv preprint arXiv:1512.05193*, 2015.

[42] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," *arXiv preprint arXiv:1508.05326*, 2015.

[43] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[44] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting." *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[45] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, 2015, pp. 448–456.

[46] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[47] F. Chollet *et al.*, "Keras," https://github.com/fchollet/keras, 2015.

[48] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[49] Z. Wang, W. Hamza, and R. Florian, "Bilateral multi-perspective matching for natural language sentences," *arXiv preprint arXiv:1702.03814*, 2017.