Using context to identify the language of face-saving

Nona Naderi
Department of Computer Science
University of Toronto
Toronto, Ontario, Canada
nona@cs.toronto.edu

Graeme Hirst
Department of Computer Science
University of Toronto
Toronto, Ontario, Canada
gh@cs.toronto.edu

Abstract

We created a corpus of utterances that attempt to save face from parliamentary debates and use it to automatically analyze the language of reputation defence. Our proposed model that incorporates information regarding threats to reputation can predict reputation defence language with high confidence. Further experiments and evaluations on different datasets show that the model is able to generalize to new utterances and can predict the language of reputation defence in a new dataset.

1 Introduction

Goffman (1967) defines *face*, or *reputation*, as "the positive social value a person effectively claims for himself by the line others assume he has taken during a particular contact. Face is an image of self delineated in terms of approved social attributes". Criticisms and persuasive attacks pose threats to reputation or face and they are common in all social interactions. Allegations are often made against organizations (e.g., companies and governments) and individuals (e.g., medical practitioners and politicians), and various argumentation tactics and persuasive strategies are used in response to these allegations to attempt to defend the respondent's reputation and thereby save face. Previous studies on reputation defence mostly use manual content analysis, such as the studies by Benoit and Henson (2009) and Zhang and Benoit (2009) on political cases, and by Penman (1990) and Tracy (2011) on courtroom cases. While these studies reveal much about reputation defence strategies in various social settings, they do not analyze in detail the actual language used in the defence of reputation.

Here, we examine political speeches and investigate whether we can detect the language of reputation defence. We created a corpus of reputation defence,[1] in which the annotations are based on the structure of parliamentary debate. This corpus is based on the oral question period of a Westminster-style parliamentary system, specifically that of Canada, where the government of the day is held accountable for its actions and tries to defend its reputation.[2] Using this naturally annotated data lets us avoid the subjectivity of manual analysis, any interpretation by the annotators, and any annotation inconsistencies. We investigate whether we can predict the language of reputation defence and whether the context in which the reputation defence occurs can help in identifying this language. We first perform experiments on a sampled dataset from Canadian parliamentary proceedings of 1994–2014. We then explore the performance of our approaches on two different governments. We show that the context of reputation defence is effective in its recognition.

2 Related work

Reputation defense is more broadly related to Aristotelian ethos (Aristotle, 2007) or one's credibility that is reflected through the use of language. Previous studies on face-saving and reputation management focused on identifying various persuasive strategies and their effectiveness (Benoit, 1995; Coombs and Holladay, 2008; Burns and Bruner, 2000; Sheldon and Sallot, 2008). In the NLP field, Naderi and Hirst (2017) performed a manual annotation analysis on reputation defence strategies in Parliament and proposed a computational model to identify strategies of denial, excuse, justification, and concession. Naderi and Hirst (2018) further proposed two approaches to

[1] The data is freely available at http://www.cs.toronto.edu/~nona/data/data.html

[2] https://www.ourcommons.ca/About/Compendium/Questions/c_d_principlesguidelinesoralquestions-e.htm

Proceedings of the 5th Workshop on Argument Mining, pages 111–120
Brussels, Belgium, November 1, 2018. ©2018 Association for Computational Linguistics

automatically annotate unlabeled speeches with defence strategies. Another related NLP study focuses on extracting ethos from the United Kingdom's parliamentary debates; Duthie and Budzynska (2018) used a set of features, such as sentiments and part-of-speech tags, to extract negative and positive references. Here, instead, we are interested in studying whether we can classify a speech as reputation defence or not, and whether the context can improve this classification.

3 Reputation defence

The main purpose of the oral question period in a Westminster-style parliamentary system is to hold the government accountable for its actions and to highlight the inadequacies of the government.[3] Members of the opposition and government backbenchers both may ask questions, and government ministers must respond. The questions asked by the opposition members are confrontational, intended to criticize or embarass the government, and are considered reputation threats; the answers to these questions by government ministers try to defend the government's choices and the ministers' reputations. Therefore, these questions and answers are a rich dataset for characterizing the language of reputation attack and the language of reputation defence. Government backbenchers can also pose questions. However, these questions are most often friendly and promotional questions, and the answers given to these questions try to promote the government's plans. Thus these questions and their answers are ordinary reputation-building or reputation-enhancing pairs. They thus act as negative examples.

This dichotomy between the two types of questions in Parliament is supported by qualitative studies such as those of Peŕez de Ayala (2001), Ilie (2006), and Bates et al. (2012). Peŕez de Ayala (2001) describes Question Time in the U.K. House of Commons as a ”face-threatening genre” and examines politeness strategies used in the face-threatening language of a set of questions. Bates et al.'s (2012) analysis shows that government backbenchers ask either questions that allow the minister to talk about the government's policies and positions, or questions that are straightforward to answer. While concerns with reputation are of particular importance not only for politicians but are salient in all social encounters, gathering a dataset of reputation threats and defences from encounters other than parliamentary settings is challenging. Hence, we use the available parliamentary proceedings for characterizing these languages.

The following question posed by the opposition in the Canadian Parliament and the Minister's reply to it is an example of a reputation threat and the defence made in response. In the example, the [Deputy] Prime Minister is confronted by an opposition member with a persuasive attack, and he tries to defend and justify the actions of the government:[4]

Example 3.1 ***Q.*** *Mr. Speaker, the former finance minister continues to amaze the crowds with his dance of the veils, with the ethics counsellor standing just off stage catching whatever is shed. The first layer was the blind trust that no one could see through. Next came blind management. Now we are down to the last and flimsiest layer, the supervisory agreement. Could the Prime Minister explain why the former finance minister was allowed the opportunity for hands on management by the ethics counsellor while all other ministers adhered to the stricter blind trust or blind management agreements?*

A. *Mr. Speaker, the arrangements that were in place were those that were appropriate to the circumstances and, in fact, reflect the views of the Parker commission that reviewed these matters in the past. The former minister complied entirely with the requirements before him.*

The next example shows a non-threatening question and answer pair, where the question is posed by a government backbencher.[5]

Example 3.2 ***Q.*** *Mr. Speaker, my question is for the Minister of the Environment. Recently we have been reading more and more articles in the media concerning high levels of sulphur in fuels, air pollution and health problems that result from these high levels. On this issue could the minister tell the*

[3] The Westminster system originated in the United Kingdom and is used in Commonwealth nations, such as Canada, Australia, India, and New Zealand. The tradition of question time for government accountability is practiced under different names in these countries; in the United Kingdom, it is known as *oral questions*, in Canada as *oral question period*, in Australia and New Zealand as *question time*, and in India as *question hour*.

[4] 2003-02-20, John Reynolds (Q) and John Manley, Deputy Prime Minister, representing the Prime Minister (A).

[5] 2001-06-04, Shawn Murphy (Q) and David Anderson (A).

House what actions are being taken to deal with the issue of high sulphur levels in fuels in Canada?

A. *Mr. Speaker, the announcement I made earlier this year covers gasoline, diesel and fuel oils outside road fuels. It will reduce the amount of sulphur in gasoline from its average now of 360 parts per million to 30 parts per million. In on road diesel, the figure will go from 500 parts per million to 15. The dates for this are the end of 2004 for gasoline and June 1, 2006, for diesel.*

4 Data

We extracted our Canadian data from the Lipad[6] dataset of the Canadian parliamentary proceedings (Hansard) from 1994 to 2014. This data consists of the proceedings of the 35th to 41st Canadian parliaments. We focused on only the first question and answer pair of each topic of discussion during the oral question period of parliament sessions in order to minimize dependency on the broader topical context. We created a balanced corpus by randomly sampling the same number of questions posed by the opposition members (reputation threats) as those asked by the government backbenchers (friendly non-threats). This resulted in 9,048 pairs of questions and answers on more than 1,600 issues over the 20-year period.

To further analyze reputation defence strategies used by different governments, we extracted the question and answer pairs from parliaments with different governing parties. The Liberal Party was the government in the 36th, 37th, and 38th Parliaments, and the Conservative Party was the government in the 39th, 40th, and 41st Parliaments. This allows us to examine the language of reputation defence used by different political ideologies. Furthermore, by training and testing models on parliaments with different governing parties, we can ensure that the models are not affected by the ideology of the speaker and the topic of day or interest of the accuser. Table 1 shows the statistics of these datasets, which, unlike the 1994–2014 dataset, are not balanced.

Party	Parliaments	Opposition	Government
Liberal	36, 37, 38	11,090	1,736
Conservative	39, 40, 41	11,504	2,004

Table 1: Corpus statistics; *Party* shows the governing party; *Opposition* shows the number of questions asked by the opposition members and their respective answers, *Government* shows the number of questions asked by the government backbenchers and their respective answers.

5 Reputation threat analysis

A principled analysis of the language of face-threats or accusations themselves falls outside the scope of this work, but here we characterize the differences between the questions asked by opposition members (reputation threats) and questions asked by government backbenchers (friendly non-threats). We randomly sampled 3,400 questions asked by the oppositions and 3,400 questions asked by the government backbenchers. We performed our analysis using Linguistic Inquiry and Word Count (LIWC) (Tausczik and Pennebaker, 2010), which is widely used in social science studies. Table 2 presents the ratio of averages between reputation threats and non-threat questions for a set of LIWC features, including *anger*, *negative* and *positive* emotions, *achievement*, and *cognitive processes*. Ratios greater than 1.0 indicate features that are more prominent in reputation threats and ratios less than 1.0 indicate features that are more prominent in non-threats. The results show that, unsurprisingly, anger and negative emotions used more in reputation threats than non-threats, whereas positive emotions are used more in non-threats. These features are motivated by theories, such as Brown and Levinson (1987) and Partington (2003) that recognize varying degrees of politeness in threatening or saving the addressee's face. Achievements are used more in non-threats and cognitive processes are used more in reputation threats. This is consistent with theories (Mulholland, 2003) that recognize mentioning the consequences of the fault as one mode of accusation.

6 Approach

Convolutional Neural Networks (CNN) have been shown to be effective for classification tasks (Kim, 2014). Here, we used a CNN model to represent the question and answer pairs for binary classifications of face-saving language. We first represented each word in the question and the answer with its associated pre-trained embedding. We then applied a convolution operation to each possible window of x words from the question and the answer to produce a feature map, similar to the ap-

[6] LInked PArliamentary Data, `https://www.lipad.ca`

Feature	Ratio	Text
Anger	1.15	**Opp:** Prime Minister has the **annoying** habit of blindly exonerating . . .
Negative emotion	1.35	**Opp:** We all know there is a **nasty** trade dispute going on between . . .
Positive emotion	0.69	**Gov:** . . . presenting new and exciting **opportunities** . . .
Achievement	0.82	**Gov:** . . . foundation has **successfully** concluded agreements with . . .
Cognitive processes	1.20	**Opp:** . . . Minister of the Environment **ought to** read the U.S. . . .

Table 2: Ratios of linguistic features in opposition questions to government backbenchers' questions. Text shows an example for each feature. **Opp** shows an opposition question and **Gov** shows a government backbencher's question.

proach of Kim (2014). We then applied a sliding Max Pooling and concatenated the representation of the question and the answer. We used 20 and 10 filters for the five-fold cross-validation and cross-parliament experiments, respectively. We used filter windows of 3 and 4, a dropout of 0.8, and mini-batch sizes of 32 and 50 for five-fold cross-validation and cross-parliament experiments, respectively.

Recurrent neural networks have been used effectively in NLP for sequence modeling. Here, we further used two long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) networks[7] with 128 units to represent questions and answers, separately. The LSTM layers were then passed to a dropout layer (Hinton et al., 2012) with a rate of 0.6. We then merged the two representations. For all our Neural Network models, we initialized our word representations using the publicly available GloVe pre-trained word embeddings (Pennington et al., 2014)[8] (300-dimensional vectors trained on Common Crawl data), and restricted the vocabulary to the 5,000 most-frequent words. The models were trained with binary cross-entropy with the Adam optimizer (Kingma and Ba, 2014) for 10 and 5 epochs for five-fold cross-validation and cross-parliament experiments, respectively. We also tried encoding the questions and answers using a layer of Gated Recurrent Units (GRU) (Cho et al., 2014) with shared parameters, but this model performed worse than the other models, and for brevity we do not report the results here.

We further trained an SVM classifier (using the scikit-learn package (Pedregosa et al., 2011)) with all possible combinations of words extracted from the cross-product of questions and answers to capture the interaction between reputation threat and reputation defence. The features are tuples of word pairs from question and answer pairs. We removed word pairs that occurred fewer than 80 times in the datasets. Our use of this set of features is inspired by the effectiveness of word pairs in classifying discourse relations (Biran and McKeown, 2013; Pitler et al., 2009) regardless of their sparsity issue.

7 Evaluation and results

We approach the recognition of the face-saving language as a binary supervised classification task. Our baselines are majority class (which is always answers given to the opposition questions), an SVM model trained with answer unigram vectors (weighted using *tf–idf*, represented with the notation '-Answers' in the result tables), and one layer of GRU to model answer sequences. Since reputation defence is expressed in response to the reputation threat, we further considered the question as the context of the reputation defence and trained an SVM model with question and answer unigrams (weighted using *tf-idf*, represented by the notation '-Questions&Answers' in the result tables). For comparison, we further include the results of an SVM model trained on only unigrams from questions ('-Questions'). We also use one layer of GRU to model the concatenation of question and answer pairs as one sequence. The SVM model trained on word pairs is represented with the notation '-Questions×Answers' in the result tables.

In the cross-parliament setting, we used the 36^{th}, 37^{th}, and 38^{th} parliaments with Liberal governments and the 39^{th}, 40^{th}, and 41^{st} parliaments with Conservative governments. We first performed a five-fold cross-validation on the Liberal and Conservative governments individually (three parliaments each), and then performed a

[7] Using https://keras.io/

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

Model	**Accuracy**	**F$_1$**	**Precision**	**Recall**
(1) Canada 1994–2014; Opposition: 4,524; Government: 4,524				
Majority	50.00			
Unigrams-Answers	76.57	76.57	76.59	76.57
Unigrams-Question&Answers	88.00	88.00	88.01	88.00
Unigrams-Questions	90.10	90.10	90.11	90.10
1 GRU(128)-Answers	81.60	82.64	77.27	89.99
1 GRU(128)-Questions&Answers	**94.39**	**94.23**	**93.94**	**94.91**
CNN(128)-Questions&Answers	91.40	91.16	90.54	92.41
2 LSTMs(128)-Questions&Answers	92.26	91.92	93.34	91.04
Word-pairs-Questions×Answers	91.46	91.46	91.47	91.46
(2) Parliaments 36, 37, 38; Opposition: 11,090; Government: 1,736				
Majority	86.47			
Unigrams-Answers	88.57	88.26	88.10	88.57
Unigrams-Questions&Answers	92.77	92.59	92.50	92.77
Unigrams-Questions	93.59	93.43	93.43	93.59
1 GRU(128)-Answers	90.89	94.91	91.53	98.70
1 GRU(128)-Questions&Answers	**95.72**	**97.52**	96.52	98.66
CNN(128)-Questions&Answers	94.50	96.87	95.12	**98.81**
2 LSTMs(128)-Questions&Answers	94.11	96.52	**97.23**	95.99
Word-pairs-Questions×Answers	95.06	94.95	94.98	95.06
(3) Parliaments 39, 40, 41; Opposition: 11,504; Government: 2,004				
Majority	85.16			
Unigrams-Answers	87.27	86.95	86.82	87.27
Unigrams-Questions&Answers	95.87	95.75	95.78	95.87
Unigrams-Questions	97.45	97.41	97.42	97.45
1 GRU(128)-Answers	91.05	94.93	91.63	98.63
1 GRU(128)-Questions&Answers	**98.33**	**99.02**	98.77	**99.30**
CNN(128)-Questions&Answers	97.10	98.31	97.50	99.20
2 LSTMs(128)-Questions&Answers	97.11	98.27	**98.98**	97.63
Word-pairs-Questions×Answers	97.48	97.43	97.45	97.48

Table 3: The performance of different models for binary classification of reputation defence language using five-fold cross-validation on (1) a balanced set from 1994–2014; (2) three Liberal governments; (3) three Conservative governments.

cross-parliament classification. For all datasets and models, we randomly used 10% of the training data as the development set. We evaluated the performance of reputation defence classification using the metrics *Accuracy, Precision, Recall,* and F_1. Table 3 shows the results of five-fold cross-validation on a balanced set from all parliaments in the period 1994–2014, on just the Liberal governments, and on just the Conservative governments. Both CNN and LSTM models improve the classification compared to the baselines. In general, we can see that all the models that rely only on the answer or reputation defence perform poorer than the models that rely also on the questions. The best model achieves an accuracy and F_1 measure of above 98% on the parliaments with Conservative governments. The highest accuracy and F_1 measure on the Liberal dataset is above 95% and 97%, respectively.

Table 4 shows the results of the cross-parliament classification. We trained the models on all Liberal parliaments, and tested them on all Conservative governments, and then vice versa. The SVM model trained using question-and-answer unigrams is a strong baseline. Both the CNN and LSTM models improved F_1 measure compared to the baseline models. On the cross-parliament classification setting, again the models trained on both questions and answers perform better. The overall performance of the neural net models across parliaments is poorer than the classification performance within parliaments. This can be explained by the differences in framing strategies used in the language of defence by the two parties, which each defend their actions and choices from their own point of view. The SVM model trained on the words extracted from the cross-product of questions and answers (word-pairs) achieves the best accuracy, reaching an accuracy and F_1 measure above 92% across parliaments. These results show that reputation defence language can be detected with high accuracy regardless of differences in ideologies and framing strategies.

An error analysis shows that most errors occurred in the classification of answers to non-threat questions. One reason for this is that while the government ministers do not defend themselves in the answers in response to the government backbenchers, they do try to enhance their image. Consider the following example[9]:

Example 7.1 ***Q.*** *Mr. Speaker, my question is for the Minister of the Environment. Over the weekend, the leader of the Bloc Québécois had the temerity to claim that the 2005 budget did not serve the interests of the people in Quebec. I know full well that the environment is very important to the people in my riding. Could the minister tell the House how the environmental initiatives contained in the budget will benefit Quebec?*

A. *Mr. Speaker, Quebeckers are impatiently awaiting the greenest budget since Confederation. Very successful contacts have been established with the Government of Quebec for the use of the partnership fund. Projects are sprouting up all over for the climate fund, for new investments, for national parks and for investment in renewable and wind energy. Mayors are waiting for green investments for cities and municipalities through the new deal, the green municipal fund, the EnerGuide program for cities and so on. Quebec must not be blocked, but greened even more.*

We further examined the cases where a reputation defence was erroneously assigned a non-defence label. These cases require real-world knowledge to determine that they are indeed reputation defence. Here is an example[10]:

Example 7.2 ***Q.*** *Mr. Speaker, this country was built upon common interests by and for the people here. We cannot allow the House of Commons to introduce a bill which, in reality, provides a recipe for destroying this country. Does the government realize that this draft bill is an avowal of failure by this government as far as the future of the federation is concerned?*

A. *No, Mr. Speaker. This bill is a follow-up to the Supreme Court judgment referring back to the political stakeholders the responsibility to establish the conditions of clarity under which they would agree to negotiate the secession of a province from Canada, and it seems to me that one of those stakeholders is the Canadian House of Commons.*

The models that rely on only the answer have particular difficulty in distinguishing these cases.

[9] 2005-05-31, David Smith (Q) and Stéphane Dion (A).

[10] 1999-12-13, André Bachand (Q) and Stéphane Dion (A).

Model	**Accuracy**	**F$_1$**	**Precision**	**Recall**
Train 36, 37, 38 (Opp: 11,090; Gov: 1,736) and test 39, 40, 41 (Opp: 11,504; Gov: 2,004)				
Majority	85.16			
Unigram-Answers	82.22	82.63	83.10	82.22
Unigrams-Questions&Answers	89.60	89.23	89.02	89.60
Unigrams-Questions	91.56	91.07	91.04	91.56
GRU(128)-Answers	84.02	91.21	85.23	98.25
GRU(128)-Questions&Answers	83.48	90.83	85.65	96.84
CNN(128)-Questions&Answers	85.86	92.32	86.53	**99.10**
2 LSTMs(128)-Questions&Answers	85.27	91.88	86.10	98.66
Word-pairs-Questions×Answers	**93.59**	**93.36**	**93.33**	93.59
Train 39, 40, 41 (Opp: 11,504; Gov: 2,004) and test 36, 37, 38 (Opp: 11,090; Gov: 1,736)				
Majority	86.47			
Unigram-Answers	86.95	85.44	84.87	86.95
Unigrams-Questions&Answers	90.34	89.10	89.40	90.34
Unigrams-Questions	91.14	90.52	90.42	91.14
GRU(128)-Answers	86.29	92.58	86.49	**99.71**
GRU(128)-Questions&Answers	85.58	92.14	86.49	98.75
CNN(128)-Questions&Answers	86.75	92.73	87.67	98.55
2 LSTMs(128)-Questions&Answers	86.72	**92.78**	87.10	99.45
Word-pairs-Questions×Answers	**92.95**	92.31	**92.62**	92.95

Table 4: The performance of different models for binary classification of reputation defence in the cross-parliament setting. **Opp** shows the number of opposition members' questions and their respective answers and **Gov** shows the number of government backbenchers' questions and their respective answers.

Model	**Accuracy**	**F$_1$**	**Precision**	**Recall**
Train 36, 37, 38 and test 39, 40, 41 (balanced, 3400 instances train and 3400 test)				
Majority	50.00			
Unigrams-Answers	67.94	67.92	67.99	67.94
+NRC Emotion (anger+pos+neg)	69.77	69.70	69.94	69.77
+Bigrams	73.41	73.33	73.73	73.41
+Vagueness cue words	73.85	73.75	74.22	73.85
Word-pairs-Questions×Answers	83.97	83.95	84.14	83.97
Train 39, 40, 41 and test 36, 37, 38 (balanced, 3400 instances train and 3400 test)				
Majority	50.00			
Unigrams-Answers	71.24	70.68	72.99	71.24
+NRC Emotion (anger+pos+neg)	71.71	71.14	73.57	71.71
+Bigram	73.71	72.91	76.88	73.71
+Vagueness cue words	73.88	73.91	76.98	73.88
Word-pairs-Questions×Answers	83.77	83.67	84.82	83.77

Table 5: The performance of different models for binary classification of reputation defence in the cross-parliament setting with the balanced data (1700 instances of each class).

8 Analyzing the language of defence

To help discover more about the underlying structure of the data, we conducted an exploratory feature analysis. We created two balanced datasets from the two governments, where each dataset consists of 3,400 question and answer pairs (1,700 questions asked by opposition members and 1,700 questions asked by government backbenchers). The question and answer pairs were selected randomly. In this setting, we focused only on the text of the answers or reputation defence.

We consider emotions, such as positive, negative, and anger. For extracting these features, we used the NRC Word-Emotion Association Lexicon (NRC Emotion lexicon)[11]. This lexicon provides manually assigned association scores for basic emotions including *anger, fear, joy, sadness, disgust, anticipation, trust, surprise*, and sentiments (*positive* and *negative*) (Mohammad and Turney, 2013). It consists of 14,182 unigrams that are manually annotated through crowdsourcing. We compute the total association scores of the lexicon words in the answer for each class of emotions and sentiments.

We further examined the NRC VAD Lexicon[12] for our analysis. This lexicon provides valence (positiveness–negativeness / pleasure / displeasure), arousal (active–passive), and dominance (dominant–submissive) scores for 20K English words (Mohammad, 2018). These dimensions have been used for analysis of human interaction (Burgoon and Hale, 1984). We use the total score of each dimension in the answer as a feature. We also consider vagueness cue words (Bhatia et al., 2016; Lebanoff and Liu, 2018). This set of features (40 cue words) is represented by the frequency of the vagueness cues in the answer. The use of these features is motivated by theories such as that of Fraser (2012) that suggest that hedge words can be used to avoid face-threatening acts. We also use bigrams as additional features. We performed the classification using SVM. The results of the binary classification of face-saving language on the balanced data of the cross-parliament setting is presented in Table 5.

		Predicted	
		Non-defence	Defence
Actual	Non-defence	1,360	340
	Defence	549	1,151

Table 6: Confusion matrix for the best performing model that relies only on features extracted from answers, including unigrams and bigrams, NRC emotions (anger+pos+neg), and vagueness cues. Trained on 36,37,38 (3,400 instances) and tested on 39,40,41 (3,400 instances).

		Predicted	
		Non-defence	Defence
Actual	Non-defence	1,368	332
	Defence	213	1,487

Table 7: Confusion matrix for the model trained on word pairs. Trained on 36,37,38 (3,400 instances) and tested on 39,40,41 (3,400 instances).

The only emotion that contributed to the classification was anger. The positive impact of anger on the classification performance is in line with theories such as those of Mulholland (2003) and Benoit (1995) that find that attacking the accuser is a type of face-saving strategy. Both positive and negative sentiments also improved the performance of the classification, as did vagueness cues and bigrams. However, using valence, arousal, and dominance hurt the performance.

The confusion matrices for the best model trained on the features extracted from the answers (unigrams and bigrams + NRC Emotions including negative and positive sentiments and anger + vagueness cues) and the model trained on word pairs are presented in Tables 6 and 7, respectively. Both models are trained on 3,400 instances from the 36th, 37th, and 38th parliaments and tested on 3,400 instances from the 39th, 40th, and 41st parliaments.

9 Conclusion

Face-saving language is employed in everyday human interaction. In this study, we introduced the task of automatically recognizing the language of face-saving. We created a corpus of reputation-defence language on various issues from parliamentary proceedings that is freely available. We further presented two neural network approaches to classify this language. We showed that the context of reputation defence is important for this classification task. Our results supported our annotation decision based on the adversarial struc-

[11] http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

[12] http://saifmohammad.com/WebPages/nrc-vad.html

ture of the parliament and showed that our corpus is appropriate for analyzing the language of reputation defence. A practical application of our model will be to analyze human behavior and to examine the effectiveness of reputation defence in various social settings.

Acknowledgements

This research is financially supported by an Ontario Graduate Scholarship, the Natural Sciences and Engineering Research Council of Canada, and the University of Toronto. We thank the anonymous reviewers for their thoughtful comments and suggestions.

References

Aristotle. 2007. *On Rhetoric: A Theory of Civic Discourse* (translated by G.A. Kennedy). Oxford University Press.

Soledad Peŕez de Ayala. 2001. FTAs and Erskine May: Conflicting needs? Politeness in question time. *Journal of Pragmatics*, 33(2):143–169.

Stephen R. Bates, Peter Kerr, Christopher Byrne, and Liam Stanley. 2012. Questions to the Prime Minister: A comparative study of PMQs from Thatcher to Cameron. *Parliamentary Affairs*, 67(2):253–280.

William L. Benoit. 1995. *Accounts, Excuses, and Apologies: A Theory of Image Restoration Strategies*. State University of New York Press, Albany.

William L. Benoit and Jayne R. Henson. 2009. President Bush's image repair discourse on Hurricane Katrina. *Public Relations Review*, 35(1):40–46.

Jaspreet Bhatia, Travis D. Breaux, Joel R. Reidenberg, and Thomas B. Norton. 2016. A theory of vagueness and privacy risk perception. In *Requirements Engineering Conference (RE), 2016 IEEE 24th International*, pages 26–35. IEEE.

Or Biran and Kathleen McKeown. 2013. Aggregated word pair features for implicit discourse relation disambiguation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 69–73, Sofia, Bulgaria. Association for Computational Linguistics.

Penelope Brown and Stephen C. Levinson. 1987. *Politeness: Some Universals in Language Usage*, volume 4. Cambridge University Press.

Judee K. Burgoon and Jerold L. Hale. 1984. The fundamental topoi of relational communication. *Communication Monographs*, 51(3):193–214.

Judith P. Burns and Michael S. Bruner. 2000. Revisiting the theory of image restoration strategies. *Communication Quarterly*, 48(1):27–39.

KyungHyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259,*.

W. Timothy Coombs and Sherry J. Holladay. 2008. Comparing apology to equivalent crisis response strategies: Clarifying apology's role and value in crisis communication. *Public Relations Review*, 34(3):252–257.

Rory Duthie and Katarzyna Budzynska. 2018. A Deep Modular RNN Approach for Ethos Mining. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4041–4047. International Joint Conferences on Artificial Intelligence Organization.

Bruce Fraser. 2012. Pragmatic competence: The case of hedging. In Gunther Kaltenböck, Wiltrud Mihatsch, and Stefan Schneider, editors, *New Approaches to Hedging*, pages 15–34. Brill.

Erving Goffman. 1967. *Interaction Ritual: Essays on face-to-face interaction.* Aldine.

Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R. Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580.*

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Cornelia Ilie. 2006. Parliamentary discourses. In Keith Brown, Anne H. Anderson, Laurie Bauer, Margie Berns, Graeme Hirst, and Jim Miller, editors, *Encyclopedia of Language and Linguistics*, second edition, pages 188–196. Elsevier, Oxford.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980.*

Logan Lebanoff and Fei Liu. 2018. Automatic detection of vague words and sentences in privacy policies. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium. Association for Computational Linguistics.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184. Association for Computational Linguistics.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Joan Mulholland. 2003. *A Handbook of Persuasive Tactics: A Practical Language Guide*. Routledge.

Nona Naderi and Graeme Hirst. 2017. Recognizing reputation defence strategies in critical political exchanges. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 527–535, Varna, Bulgaria.

Nona Naderi and Graeme Hirst. 2018. Automatically labeled data generation for classification of reputation defence strategies. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Alan Partington. 2003. *The linguistics of political argument: The spin-doctor and the wolf-pack at the White House*. Routledge.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Robyn Penman. 1990. Facework & politeness: Multiple goals in courtroom discourse. *Journal of Language and Social Psychology*, 9(1-2):15–38.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conferene on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Emily Pitler, Annie Louis, and Ani Nenkova. 2009. Automatic sense prediction for implicit discourse relations in text. 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 2-Volume 2*, pages 683–691. Association for Computational Linguistics.

Catherine A. Sheldon and Lynne M. Sallot. 2008. Image repair in politics: Testing effects of communication strategy and performance history in a faux pas. *Journal of Public Relations Research*, 21(1):25–50.

Yla R. Tausczik and James W. Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.

Karen Tracy. 2011. A facework system of minimal politeness: Oral argument in appellate court. *Journal of Politeness Research. Language, Behaviour, Culture*, 7(1):123–145.

Ernest Zhang and William L. Benoit. 2009. Former Minister Zhang's discourse on SARS: Government's image restoration or destruction? *Public Relations Review*, 35(3):240–246.