Predicting the Outcome of Deliberative Democracy: A Research Proposal

Conor McKillop
School of Science and Engineering
University of Dundee
Dundee, United Kingdom
c.z.mckillop@dundee.ac.uk

Abstract

As liberal states across the world face a decline in political participation by citizens, deliberative democracy is a promising solution for the publics decreasing confidence and apathy towards the democratic process (Dahl et al., 2017). Deliberative dialogue is method of public interaction that is fundamental to the concept of deliberative democracy. The ability to identify and predict consensus in the dialogues could bring greater accessibility and transparency to the face-to-face participatory process. The paper sets out a research plan for the first steps at automatically identifying and predicting consensus in a corpus of German language debates on hydraulic fracking. It proposes the use of a unique combination of lexical, sentiment, durational and further derivative features of adjacency pairs to train traditional classification models. In addition to this, the use of deep learning techniques to improve the accuracy of the classification and prediction tasks is also discussed. Preliminary results at the classification of utterances are also presented, with an F1 between 0.61 and 0.64 demonstrating that the task of recognising agreement is demanding but possible.

1 Introduction

Liberal states across the world are facing a significant decline in political participation by citizens. The global voter turnout rate has dropped by more than 10% over the last 25 years (Groupe de la Banque mondiale, 2017), and this trend does not appear to be slowing down. The public have reported decreasing confidence and apathy towards the democratic process (Dahl et al., 2017). Deliberative Democracy represents a potential solution to these problems. Through the evaluation of different policy proposals using a process of truthful and rational discussion between citizens and authority, Deliberative democracy can enable consensual, well-justified, decision making. It can improve the political competence of citizens by; facilitating the exchange of arguments and sharing of ideas on proposals from authority (Estlund et al., 1989); reconfiguring democracy as a process of 'public reasoning' and connecting citizens with each other and with their governing institutions (Parkinson and Mansbridge, 2012; Dryzek, 2012).

Deliberative Dialogue is a structured, face-to-face method of public interaction. As a form of participatory process, it is fundamental to the concept of Deliberative Democracy (McCoy and Scully, 2002). There are many different forms of deliberative dialogue, including, but not limited to: citizens' assemblies, citizens' juries and planning cells. The European Commission's 'Future of Europe debates' (Directorate-General for Communication, 2017b) are an exemplar of hosting deliberative dialogue successfully at large scale.

The 'Future of Europe debates' are due to come to their natural conclusion after a two year long process that started with the release of the 'White paper on the future of Europe' in March of 2017 (ibid.). This white paper set out the main challanges and opportunities facing the 27 European Union (EU) member states for the next decade. To encourage citizens' participation, the Commission hosted a series of debates across cities and regions within Europe (Directorate-General for Communication, 2017a). At the debates, all members of the Commission engaged in dialogue with citizens and listened to their views and expectations concerning the future of Europe. The debates were well received, with 129 debates in more than 80 towns, attended by over 21,000 citizens (ibid.).

In the deliberative democratic process, one of the main aims is for informed agreement to be reached among all involved parties. However, in dialogues with larger citizenry, it is less likely that

consensus is reached between all participants (Peter, 2016). As can be seen with the 'Future of Europe debates', numbers in attendance can be high. Therefore, the ability to automatically identify, or even predict, consensus between participants in these dialogues can make the participatory process even more transparent and accessible. In the future, it could even provide authority with a tool for deciding when to move to an aggregative mechanism for deciding the outcome, such as majority voting.

2 Related Work

Previous work has reported some levels of success in the automatic classification of agreement and no agreement using machine learning techniques.

Galley et al. (2004) used a statistical approach, with Bayesian networks to model agreements and disagreements in conversational interaction. Simple Bayesian networks were trained with contextual features of adjacency pairs identified in an annotated corpus of meetings. With the recent advances in deep learning techniques, there is an opportunity to apply the techniques from this paper to multi-speaker debates

On the use of sentiment analysis to aid in the detection of agreement, as employed in this paper, a number of previous works have successfully applied the technique. For example, Thomas et al. (2006) used sentiment property for classifying support or opposition of proposed legislative speeches in transcripts from United States' Congress debates. Further work by Balasubramanyan et al. (2011) investigated classifying sentiment polarity of comments on a blog post, towards the topics in the blog.

Abbott et al. (2011) reported on automatically recognising disagreement between online posts. The paper presented the ARGUE corpus, containing thousands of quote and response pairs posted to an online debate forum. Abbott et al. proposed the use of simple classifiers to label a quote and response pair as in either agreement or disagreement. An improvement over baseline was achieved by the authors, though this was limited to informal, online political arguments.

The majority of research into the classification of agreement and disagreement has been heavily focused on postings in online forums and social networks. There has been very little work on the classification of agreement in face-to-face participatory process; the research area of this paper.

3 Data Set

The data set for the task is drawn from a total of 34 German language dialogues which all took place in an experimentally controlled environment. In each of the dialogues, there are four participants who were recorded discussing the topic of hydraulic fracking in Germany. The participants are tasked with coming to consensus around allowing or disallowing fracking within a time period of 60 minutes. Whole dialogues within the data set are annotated with either agreement or no agreement, by trusted annotators. These annotators also explicitly mark the utterance at which consensus occurs. All utterances are plain text, with a limited number of attributes, including the utterance identifier and speaker name.

This data set is composed of 20 dialogues where consensus is reached by the participants, 9 dialogues where no consensus took place and 5 dialogues where the session 'timed out' before any consensus was reached. By extracting single utterances from each dialogue, this is broken down into 1,376 utterances of agreement, 458 with no agreement and 240 with timeout. A manual investigation into the dialogues revealed that there was no clear difference in text between the dialogues of time out and no agreement.

For training and testing of the classifier, the data set was split into multiple subsets, with cross validation (Mosteller and Tukey, 1968) used to evaluate performance. The risk of overfitting by the classifier is minimised through the use of a 5-fold cross validation method.

4 Methodology

4.1 Tasks

There are two main goals of research which provide the body of work proposed in this paper. These two goals are:

- To identify where consensus has occurred between participants
- Prediction of whether it is likely that consensus between participants is going to occur

Of note is that these tasks are performed on a corpus of lower resource language.

4.2 Features

Work has already begun on the extraction of features from the data set in its current form without any further embellishment, such as the identification of *argumentation structure*, discussed in further detail in section 5of this paper.

Three distinct feature sets have been created from the data for use in machine learning techniques, these are termed:

- *Base Features* – Attributes connected to a single utterance
- *Derivative Features* – The change of Base Features across a pair of utterances
- *Second Derivative Features* – The change of Derivative Features between pairs of utterance pairs.

Base Features

A number of attributes from each singular utterance were extracted for input into the classifier responsible for identifying agreement and disagreement of utterances.

Lexical In order to capture basic lexical information, unigram and bigram features are extracted from each utterance. Text of an utterance is first processed before tokenisation and occurrence counting. In the text pre-processing: speaker names and punctuation are removed from the text, unicode characters normalised, German diacritics and ligatures translated[1], and finally words lemmatised.

Sentiment Prior work has shown that sentiment features can provide some value in the prediction of speakers' position on a topic, such as what the speaker supports or opposes (Pang and Lee, 2008). To access this information, an analysis of speaker sentiment within each utterance is undertaken. The SentimentWortschatz (SentiWS) (Remus et al., 2010) resource for German-language is used. The latest version[2] of the resource contains over 1,600 positive words and 1,800 negative words, or over 16,000 positive and 17,500 negative words when calculated to include inflections of every word. For each word in the resource, a polarity score, weighted between [-1; 1] is provided. It should be noted, that in cases where a word cannot be found in the resource, a 'neutral' score of 0 is used. For this work, a method was developed using SentiWS to give a score for each utterance in the corpus. By summing up the sentiment score for each word in the utterance, a total score for the utterance can be calculated. This total is then used as a feature for the classification model.

Durational Durational features for each utterance are also calculated. This includes, word count and character count, average word length and number of stop words.

Derivative Features

Adjacency Pairs Adjacency pairs, composed of two utterances from two speakers in succession are extracted from the dialogues and similarity measures are calculated for the features of each utterance in a pair.

Durational The change in Durational features between utterances in an adjacency pair.

Sentiment The change in Sentiment features between utterances in an adjacency pair to capture any possible shift in sentiment between speaker turns.

Similarity Measures To test the hypothesis that utterance pairs in agreement, are higher in similarity, this paper proposes using a similarity measure calculated between utterance pairs as a feature variable. An example of term based similarity, cosine similarity uses the cosine angle between the two vectors as a similarity measure. The spaCy[3] open-source software library for Natural Language Processing (NLP) will be used to calculate the similarity between the utterance text of two adjacency pairs.

Further Adjacency Pairs

Collection of Adjacency Pairs Similarity measures are calculated between a collection of two or more adjacency pairs.

4.3 Techniques

To classify an utterance as either agreement or no agreement, some work has already been undertaken using traditional machine learning models.

Traditional Classification Models

Support Vector Machine A *Support Vector Machine* (SVM) is a classifier that can be used to

[1] Translation as per the DIN 5007-2 standard.

[2] SentiWS v2.0 at the time of writing.

[3] Git repository for the library is hosted at: https://github.com/explosion/spaCy/.

perform identification of agreement on each utterance. SVMs are a versatile, supervised learning method that are well-suited to classification and regression tasks. The method produces non-linear boundaries using a linear boundary in a transformed version of the input feature space (Hastie et al., 2009). For the work in this paper, an SVM from the Scikit-learn open-source project (Pedregosa et al., 2011) was used. The input features to the classifier are from the aforementioned set, whilst the output is the binary label of agreement or no agreement.

Random Decision Forest The *Random Decision Forest* is a machine learning algorithm that is particularly suited for problems of both classification and regression. They operate by constructing and then average the results of a large collection of de-correlated decision trees (Hastie et al., 2009). The algorithm is particularly attractive for its high speed of classification and straight-forward training (Ho, 1995). A Random Decision Forest classifier from the Scikit-learn project (Pedregosa et al., 2011) is used for this work.

Naïve Bayes Another family of machine learning algorithms that remain popular and receives continuous levels of high usage, are *Naïve Bayes*. This is a method of classification that simplifies estimation by assuming that every attribute or feature contributes independently to the probability of a class (McCallum and Nigam, 1998). The family can often outperform more sophisticated alternatives (Hastie et al., 2009). However, when classifying text there is the potential for the model to adversely affect results if some adjustments are not made (Rennie et al., 2003).

Naïve Bayes Another family of machine learning algorithms that remain popular and receives continuous levels of high usage, are *Naïve Bayes*. This is a method of classification that simplifies estimation by assuming that every attribute or feature contributes independently to the probability of a class (McCallum and Nigam, 1998). The family can often outperform more sophisticated alternatives (Hastie et al., 2009). However, when classifying text there is the potential for the model to adversely affect results if some adjustments are not made (Rennie et al., 2003).

Deep Learning Models

For the second task discussed in this paper – predicting the point at which consensus between speakers is likely to occur – the use of a supervised, deep structured learning technique could possibly offer an advantage over the more traditional machine learning algorithms discussed previously.

RNN The *Recurrent Neural Network* (RNN) overcomes the shortcomings of traditional neural networks when dealing with sequential data, such as text. A class of artificial neural network, it uses connections between nodes to form a direct graph along a sequence (Graves, 2012). RNNs are limited to a short-term memory due to the 'vanishing gradient problem' (Bengio et al., 1994).

LSTM A class of RNN, *Long Short Term Memory* (LSTM) networks are capable of learning long-term dependencies. The repeating module of an LSTM has four neural network layers which interact to enable an RNN to remember inputs over a longer period of time (Graves, 2012). LSTMs reduce the problem of vanishing gradient (Chung et al., 2014). This will prove particularly important, due to the sequential nature of the adjacency pairs in the dialogues.

5 Proposed Work

Whilst work has been done using traditional machine learning algorithms to classify utterances, as per the first task described in section 4.1 of this paper, there remains work to be done in the use of deep learning models as a means for improved accuracy and performance in classification.

At present, the data set is mostly represented as plain text, with no further dimension to the utterances. One opportunity that could bring another dimension and realise unknown relationships in this data, is through the identification of argument structure within the discourse.

Argument structures are associated with, and constructed from, basic 'building blocks', and these components could also be identified. The blocks can come in the form of a *premise*, *conclusion* or *argumentation scheme*. There also exists a further opportunity for diversification of data through the analysis of relationships between argument pairs and their components. By modelling these structures, there arises the ability to gather a deeper understanding of what is being uttered by a

speaker (Lawrence et al., 2015). So, not only can the views expressed by a speaker be drawn from the argument structure, but it can also expose why these particular views are held.

Automatic identification or 'mining' of such argument structures would provide a significant time saving, allowing almost immediate use of the extracted model as features in a machine learning algorithm. However, despite the enormous growth in the field of Argument Mining, it is still difficult to identify argument structures with accuracy and reliability (Stede and Schneider, 2018). As a consequence of this, before the aforementioned advantages can be applied to this data set, it must be manually annotated by a human.

Manual annotation of the dialogues in this data set is not an insignificant cost, with regards to time and funding. As to guarantee the accuracy of the modelled arguments, annotation must follow predefined schemes, such as those set out by Reed and Budzynska (2011). The annotators carrying out the analysis must be trained to a sufficient level on the necessary schemes and also trusted. This work must be undertaken before the data can be put through the process responsible for identification and prediction of consensus. The manual annotation process of dialogues in the corpus is still ongoing.

Once the dialogues have been annotated, extraction of argumentative structure showing 'conflict' between two propositions should take place. The presence, count and exact arrangement of the propositions in conflict can then be used as an additional feature for training of the classifiers.

6 Preliminary Results

Classifier	Precision	Recall	F-measure
Naïve Bayes	0.63	0.66	0.61
SVM (Linear)	0.64	0.67	0.61
Random Forest	0.66	0.69	0.64

Table 1: Results of classification using traditional classifiers

Preliminary results related to the identification of agreement and no agreement in utterances can be seen in Table 1. This was a classification process using only the *Base Features* set and with traditional machine learning algorithms. These results suggest that the task as framed is feasible, though there is still significant opportunity for improvement.

7 Conclusion

The potential benefits resulting from the automatic identification and prediction of consensus between participants can be of significant advantage to government around the world. With only the preliminary results from classification of utterances into agreement and disagreement, it can be seen that the accuracy is nearing useable values. With the addition of advanced neural network models, such as LSTM, there is the possibility to increase the accuracy even further. The immediate goal after successfully classifying agreement and no agreement will be to predict where it is likely that agreement between participants is likely to occur.

Acknowledgements

The work reported in this paper has been supported, in part, by The Volkswagen Foundation (VolkswagenStiftung) under grant 92-182.

References

Rob Abbott, Marilyn Walker, Pranav Anand, Jean E. Fox Tree, Robeson Bowmani, and Joseph King. 2011. How can you say such things?!?: Recognizing disagreement in informal political argument. In *Proceedings of the Workshop on Languages in Social Media*, LSM '11, pages 2–11, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ramnath Balasubramanyan, William W. Cohen, Doug Pierce, and David P. Redlawsk. 2011. What pushes their buttons?: Predicting comment polarity from the content of political blog posts. In *Proceedings of the Workshop on Languages in Social Media*, LSM '11, pages 12–19, Stroudsburg, PA, USA. Association for Computational Linguistics.

Y. Bengio, P. Simard, and P. Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166.

Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, abs/1412.3555.

Viktor Dahl, Erik Amnå, Shakuntala Banaji, Monique Landberg, Jan Šerek, Norberto Ribeiro, Mai Beilmann, Vassilis Pavlopoulos, and Bruna Zani. 2017.

Apathy or alienation? political passivity among youths across eight european union countries. *European Journal of Developmental Psychology*, 15(3):284–301.

Directorate-General for Communication. 2017a. Citizens' dialogues on the future of europe. White Paper NA-01-17-787-EN-N, European Union, Brussels.

Directorate-General for Communication. 2017b. White paper on the future of europe. White Paper NA-02-17-345-EN-N, European Union, Brussels.

John S Dryzek. 2012. *Foundations and frontiers of deliberative governance*. Oxford University Press.

David M. Estlund, Jeremy Waldron, Bernard Grofman, and Scott L. Feld. 1989. Democratic theory and the public interest: Condorcet and rousseau revisited. *American Political Science Review*, 83(4):13171340.

Michel Galley, Kathleen McKeown, Julia Hirschberg, and Elizabeth Shriberg. 2004. 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*, page 669. Association for Computational Linguistics.

Alex Graves. 2012. *Supervised Sequence Labelling with Recurrent Neural Networks*, volume 385 of *Studies in Computational Intelligence*. Springer.

Groupe de la Banque mondiale. 2017. *World development report 2017: Governance and the law*. World Bank Group.

Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman. 2009. *The Elements of Statistical Learning*, second edition. Springer Series in Statistics. Springer-Verlag, New York.

Tin Kam Ho. 1995. Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 1, pages 278–282 vol.1.

J. Lawrence, M. Janier, and C. Reed. 2015. Working with open argument corpora. In *Proceedings of the 1st European Conference on Argumentation (ECA 2015)*, Lisbon. College Publications.

Andrew McCallum and Kamal Nigam. 1998. A comparison of event models for naive bayes text classification. In *Proceedings of the AAAI-98 Workshop on Learning for Text Categorization*, pages 41–48. AAAI Press.

Martha L. McCoy and Patrick L. Scully. 2002. Deliberative dialogue to expand civic engagement: What kind of talk does democracy need? *National Civic Review*, 91(2):117–135.

F. Mosteller and J Tukey. 1968. Data analysis, including statistics. In G. Lindzey and E. Aronson, editors, *Revised Handbook of Social Psychology*, volume 2, pages 80–203. Addison Wesley.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(12):1–135.

John Parkinson and Jane Mansbridge. 2012. *Deliberative systems: Deliberative democracy at the large scale*. Cambridge University Press.

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.

Fabienne Peter. 2016. *The Epistemology of Deliberative Democracy*, chapter 6. John Wiley & Sons, Ltd.

Chris Reed and Katarzyna Budzynska. 2011. How dialogues create arguments. In *Proceedings of the 7th Conference of the International Society for the Study of Argumentation (ISSA)*, pages 1633–1645. SicSat.

Robert Remus, Uwe Quasthoff, and Gerhard Heyer. 2010. Sentiws - a publicly available german-language resource for sentiment analysis. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Jason D. M. Rennie, Lawrence Shih, Jaime Teevan, and David R. Karger. 2003. Tackling the poor assumptions of naive bayes text classifiers. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, pages 616–623, Washington, D.C. AAAI Press.

Manfred Stede and Jodi Schneider. 2018. Argumentation mining. *Synthesis Lectures on Human Language Technologies*, 11(2):1–191.

Matt Thomas, Bo Pang, and Lillian Lee. 2006. Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. In *Proceedings of EMNLP*, pages 327–335.