Scrutable Feature Sets for Stance Classification

Angrosh Mandya
Computing Science
University of Aberdeen, UK
angroshmandya@abdn.ac.uk

Advaith Siddharthan
Computing Science
University of Aberdeen, UK
advaith@abdn.ac.uk

Adam Wyner
Computing Science
University of Aberdeen, UK
azwyner@abdn.ac.uk

Abstract

This paper describes and evaluates a novel feature set for stance classification of argumentative texts; i.e. deciding whether a post by a user is for or against the issue being debated. We model the debate both as attitude bearing features, including a set of automatically acquired 'topic terms' associated with a Distributional Lexical Model (DLM) that captures the writer's attitude towards the topic term, and as dependency features that represent the points being made in the debate. The stance of the text towards the issue being debated is then learnt in a supervised framework as a function of these features. The main advantage of our feature set is that it is scrutable: The reasons for a classification can be explained to a human user in natural language. We also report that our method outperforms previous approaches to stance classification as well as a range of baselines based on sentiment analysis and topic-sentiment analysis.

1 Introduction

In recent years, stance classification for online debates has received increasing research interest (Somasundaran and Wiebe, 2010; Anand et al., 2011; Walker et al., 2012; Ranade et al., 2013; Sridhar et al., 2014). Given a post belonging to a two-sided debate on an issue (e.g. abortion rights; see Table 1), the task is classify the post as for or against the issue. The argumentative nature of such posts makes stance classification difficult; for example, one has to follow the reasoning quite closely to decide which of the posts in Table 1 argues for or against abortion.

In Table 1, the posts are monologic (independent of each other), but even with the availability of dialogic structure connecting posts, both humans and classifiers experience difficulties in stance classification (Anand et al., 2011), in part because posts that contain rebuttal arguments do not provide clear evidence that they are arguing for or against the main issue being debated. Stance classification is considered particularly challenging however when the posts are monologic since the lack of dialogic structure means all features for classification have to be extracted from the text itself. Indeed studies to classify such independent posts have previously found it difficult to even beat a unigram classifier baseline; for example, Somasundaran and Wiebe (2010) achieved only a 1.5% increase in accuracy from the use of more sophisticated features such as opinion and arguing expressions over a simple unigram model.

In this paper, we propose a new feature set for stance classification of independent posts that, unlike previous work, captures two key characteristics of such debates; namely, writers express their attitudes towards a range of topics associated with the issue being debated and also argue by making logical points. We model the debate using a combination of the following features.

- **topic-stance features** – a set of automatically extracted 'topic terms' (for abortion rights, these would include, for example, 'fetus', 'baby', 'woman' and 'life'), where each topic term is associated with a distributional lexical model (DLM) that captures the writer's stance towards that topic.
- **stance bearing terminology** – words related

Proceedings of the 3rd Workshop on Argument Mining, pages 60–69,
Berlin, Germany, August 7-12, 2016.

For abortion rights
If women (not men) are solely burdened by pregnancy, they must have a choice. Men are dominant in their ability to impregnate a woman, but carry no responsibilities afterward. If woman carry the entire burden of pregnancy, they must have a choice.

Against abortion rights
Life is an individual right, not a privilege, for unborn humans [...] The right to life does not depend, and must not be contingent, on the pleasure of anyone else, not even a parent or sovereign [...]

Table 1: Samples from posts arguing for and against abortion rights

by adjectival modifiers (amod) and the noun compound (nn) relations that carry stance bearing language.

- **logical point features** – features of the form subject-verb-object (SVO) extracted from the dependency parse that capture basic points being made.
- **unigrams and dependency features** – backoff features, useful for classifying short posts lacking other features.

The contributions of this paper are two fold. Using the features listed above, we learn the stance of the debate towards the issue in a supervised setting, demonstrating better classification performance than previous work. Second, we argue that our feature set lends itself to human scrutable stance classification, through features that are human readable.

The paper is organised as follows. In §2, we discuss related work on stance classification. In §3, we describe our methods to model online debates and in §4, we present and discuss the results achieved in this study. In §5, we present our conclusions.

2 Related work

Somasundaran and Wiebe (2010) developed a balanced corpus (with half the posts for and the other half against) of political and ideological debates and carried out experiments on stance classification pertaining to four debates on abortion rights, creation, gay rights and gun rights. They achieved an overall accuracy of 63.9% using a sentiment lexicon as well as an ngrams-based lexicon of arguing phrases derived from the manual annotations in the MPQA corpus (Wilson and Wiebe, 2005), barely outperforming a unigram baseline that achieved 62.5%. They also reported performance using the sentiment lexicon alone of only 55.0% and made the point that sentiment features alone were not useful for stance.

More recently, Hasan and Ng (2014) have focused on identifying reasons for supporting or opposing an issue under debate, using a corpus that provides information about post sequence, and with manually annotated reasons. The authors experiment with different features such as n-grams, dependency-based features, frame-semantic features, quotation features and positional features for stance classification of reasons. Nguyen and Litman (2015) proposed a feature reduction method based on the semi-supervised derivation of lexical signals of argumentative and domain content. Specifically, the method involved post-processing a topic-model to extract argument words (lexical signals of argumentative content) and domain words (terminologies in argument topics).

A larger number of studies have focused on the use of dialogic structure for stance classification. Anand et al. (2011) worked with debates that have rebuttal links between posts. With respect to stance classification, they achieved accuracies ranging from 54% to 69% using such contextual features. Walker et al. (2012) focused on capturing the dialogic structure between posts in terms of agreement relations between speakers. They showed that such a representation improves results as against the use of contextual features alone, achieving accuracies ranging from 57% to 64%. Several others have modelled dialogic structure in more sophisticated ways, reporting further improvements from such strategies (Ranade et al., 2013; Sridhar et al., 2014, for example).

For the related task of opinion mining, dependency parse based features have been shown to be useful. Joshi and Penstein-Rosé (2009) transformed dependency triples into 'composite backoff features' to show that they generalise better than regular dependency features. The composite backoff features replaces either head term or modifier term with its POS tag in a dependency relation to result in two

types of features for each relation. Greene and Resnik (2009) focused on 'syntactic packaging' of ideas to identify implicit sentiment. The authors proposed the concept of observable proxies for underlying semantics (OPUS) which involves identifying a set of relevant terms using relative frequency ratio. These terms are used to identify all relations with these terms in the dependency graph, which are further used to define the feature set. Paul and Girju (2010) presented a two-stage approach to summarise multiple contrasting viewpoints in opinionated text. In the first stage they used the topic-aspect model (TAM) for jointly modelling topics and viewpoints in the text. Amongst other features such as bag-of-words, negation and polarity, the TAM model also used the composite backoff features proposed by (Joshi and Penstein-Rosé, 2009).

In summary, many studies on stance classification have focused on the use of dialogic structure between posts (Anand et al., 2011; Walker et al., 2012; Ranade et al., 2013; Sridhar et al., 2014), but there has been less work on exploring feature sets for monologic posts, though a large body of such work exists for the related task of opinion mining. We are unaware of any attention paid to the scrutability of classifiers, though users might well be interested in why a post has been classified in a certain manner. To address these gaps, we consider again the task of stance classification from monologic posts, using the dataset created by Somasundaran and Wiebe (2010). We focus on modelling of the patterns within a post rather than connections between posts, and aim to design a competitive classifier whose decisions can be explained to a user.

3 Methods

As described earlier, the goals of this paper are two fold: (1) to develop a classifier for stance classification; and (2) employ the results of classification to create human readable explanations of the reasons for classification. Accordingly, we focus on the following features which lend themselves to human readable explanation, as discussed later: (a) topic-based distributional lexical models; (b) stance bearing relations; (c) points represented as subject-verb-object triplets.

3.1 Distributional Lexical Model of Topic

Dependency grammar allows us to identify syntactically related words in a sentence, by modelling the syntactic structure of a sentence using binary asymmetrical relations (De Marneffe and Manning, 2008). We use these relations to build a Distributional Lexical Model (DLM), excluding stop words such as determiners and conjunctions to obtain a set of content words connected to the topic term through syntax. The DLM is constructed in three steps:

Step 1. identify topic terms t_i in the sentence;
Step 2. for each t_i, identify all content words w_j in a dependency relation with t_i.
Step 3. for each w_j, identify all content words w_k in a dependency relation with w_j; i.e., identify words that are within two dependency relations of the topic term.

In order to derive the topic terms, we used MALLET (McCallum, 2002), which implements topic modelling using Latent Dirichlet Allocation (LDA) (Blei et al., 2003). Given a set of documents, MALLET produces a set of likely topics where each topic is a distribution over the vocabulary of the document set such that the higher probability words contribute more towards defining the topic. We configured MALLET to produce 10 set of likely topics for the collection of posts for a given political debate, and used the default setting of the top 19 words for each topic. As we required our topic words to be nouns, we filtered the 190 words by part of speech. After further removing repetitions of words in different topics, this resulted in 96, 105, 135, 105 and 110 distinct topic terms for the political debates on abortion rights, creation, gay rights, god and gun rights, respectively. Examples of such topic terms created for the domain of abortion rights are shown in Table 2.

For the sentence and dependency parse shown in Fig. 1 (with punctuation and word positions removed for simplicity), there are three topic terms: 'fetus', 'woman' and 'pregnancy', and the 3-steps above generate the following DLMs:

fetus:	'causes'; 'sickness'; 'discomfort'; 'pain'; 'woman'
woman:	'causes'; 'sickness'; 'discomfort'; 'pain'; 'pregnancy'; 'labor'
pregnancy:	'causes'; 'woman'; 'labor'

The fetus causes sickness discomfort and extreme pain to a woman during her pregnancy and labor.

det(fetus, the)
nsubj(causes, fetus)
dobj(causes, sickness)
dobj(causes, discomfort)
conj_and(sickness, discomfort)
dobj(causes, and)
conj_and(sickness, and)
amod(pain, extreme)
dobj(causes, pain)
conj_and(sickness, pain)
det(woman, a)
prep_to(causes, woman)
poss(pregnancy, her)
prep_during(woman, pregnancy)
prep_during(woman, labor)
conj_and(pregnancy, labor)

Figure 1: Dependency Parse (simplified to remove punctuation and word positions)

These features facilitate scrutability because we can explain a classification of a post as 'for abortion rights' with a sentence such as *"This post is classified as being in favour of abortion rights because it associates words such as 'causes', 'sickness', 'discomfort', 'pain' and 'woman' with the term 'fetus'."* Note that in practice only a few features will select for a particular stance, and this example (which uses all the word pairs) is just for illustration.

The process of deriving the model for the topic term *fetus* from a dependency tree is graphically shown in Fig. 2. As seen, the word *causes* (shown in thin dotted lines) is identified in Step 2 and the other words *discomfort*, *sickness*, *pain* and *woman* are obtained in Step 3 (shown in thick dotted lines). Non-content words are excluded from the model.

This method is aimed at identifying stance bearing words associated with topic terms in argumentative posts. The resulting graph for the post arguing for abortion in Table 4 is shown in Fig. 3, where the labelled arc indicates the sentence in which the relation appears, and the direction of the arrow indicates whether the topic term precedes or follows the related word. As seen, a topic word can be connected to different terms in the graph, e.g. *pain* and *causes* are connected to *fetus* in sentences 1 and 2.

Abortion Topic Terms
life; human; conception; embryo; choice; sex; vote; position; birth; rape; war; church; act; evil; fetus; person; body; womb; brain; baby; sperm; egg; cell; logic; people; argument; god; reason; law; woman; pregnancy; children; family; abortion; murder;

Table 2: Examples of topic terms produced by Mallet for the domain of abortion rights

3.2 Stance-bearing terminology

We also consider words connected by adjectival modifier (amod) and noun compound modifier (nn) relations from the dependency graph as features for the classifier. Given the political debate on abortion rights, phrases such as 'individual rights', 'personal choices', 'personal decision' and 'unwanted children' are used primarily in posts arguing for abortion rights. Similarly, phrases such as 'human life', 'unborn child', 'innocent child' and 'distinct DNA' provide good indicators that the posts is arguing against abortion rights. In the example in Fig. 1, the feature 'extreme-pain' is extracted in this manner. These features could be used in an explanation in a sentence such as *"This post is classified as being in favour of abortion rights because it contains subjective phrases such as 'extreme pain'."*

3.3 Modelling argumentative points

We also extract features aimed at modelling elementary points made in a debate. We do this in a limited manner by defining a point simply as a subject-verb-object triple from the dependency parse. More sophisticated definitions would not necessarily result in useful features for classification. For the sentence in Fig. 1, the following points are extracted to be used as features:

fetus-causes-sickness
fetus-causes-discomfort
~~fetus-causes-and~~
fetus-causes-pain

Non-content words are excluded from the analysis. This analysis could be used to construct explanations such as *"This post is classified as being in favour of abortion rights because it makes points such as 'fetus causes sickness', 'fetus causes discomfort' and 'fetus causes pain'."*

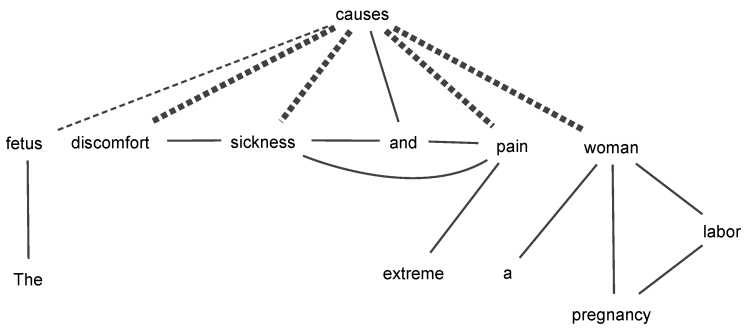
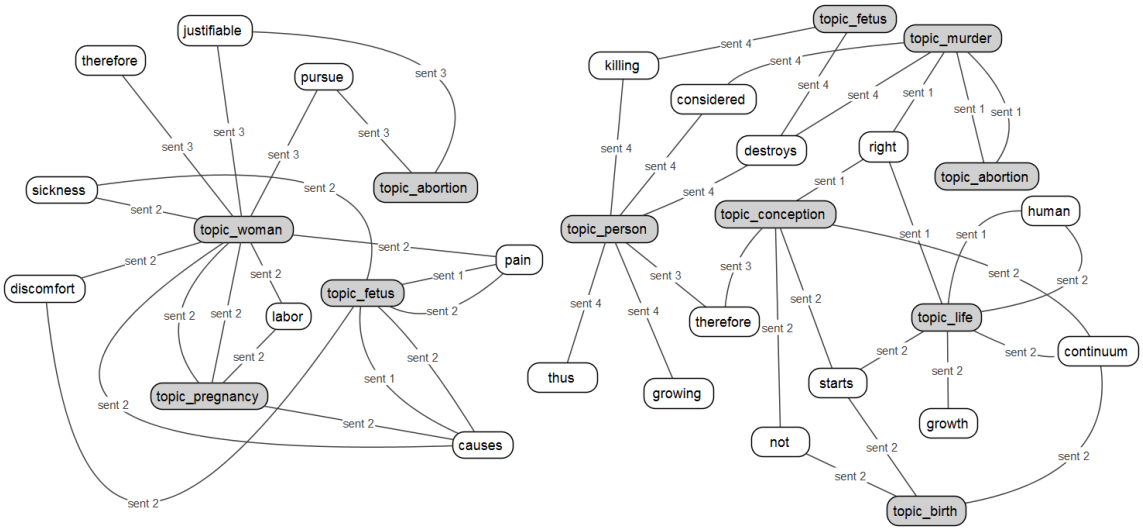


Figure 2: Deriving related words for 'fetus' from the dependency graph.



(a) Post arguing for abortion rights

(b) Post arguing against abortion rights

Figure 3: DLM models for the two posts in Table 4

3.4 Baselines

In addition to the features proposed above, we experimented with a variety of baselines for comparison.

3.4.1 Sentiment model

Our first baseline involved treating stance ('for' or 'against') as sentiment ('positive' or 'negative'). For this purpose, we used the Stanford sentiment tool [1] (Socher et al., 2013) to obtain sentence-level sentiment labels and provide these as features for stance classification of posts.

3.4.2 Topic-sentiment model

However, we do not expect a direct equivalence between sentiment and stance; for example, in Table 3, a negative sentiment is expressed in sentences arguing *for* abortion and a positive sentiment is expressed in sentences arguing *against* abortion. Our second baseline is to therefore model the stance of a post using features that indicate the sentiment of the writer towards key topics related to the issue being debated.

For example, let us consider the two sentences that argue for abortion in Table 3. Using topic modelling, we can identify topic terms such as 'fetus' and 'woman' in sentence 1. Further, using sentiment analysis the sentence can be identified to be negative. By tagging this sentiment to the topic terms contained in the sentence, we can associate a negative sentiment with topic terms 'fetus' and 'human'. Similarly for sentence 2, a negative sentiment can be associated with topic terms such as 'fetus', 'woman' and 'pregnancy'.

This model has the advantage over the sentiment analysis baseline that sentiment is asso-

[1] http://nlp.stanford.edu:8080/sentiment/

Sentences arguing for abortion rights
1. A fetus is no more a human than an acorn is a tree. 2. The fetus causes sickness, discomfort, and and extreme pain to a woman during her pregnancy and labor.

Sentences arguing against abortion rights
3. A fetus is uniquely capable of becoming a person; deserves rights, it is unquestionable that the fetus, at whatever stage of development, will inevitably develop the traits of a full-grown human person. 4. This is why extending a right to life is of utmost importance; the future of the unborn depends on it.

Table 3: Example sentences arguing for and against abortion rights

For abortion rights
The fetus causes physical pain; the woman has a right to self-defense. The fetus causes sickness, discomfort, and extreme pain to a woman during her pregnancy and labor. It is, therefore, justifiable for a woman to pursue an abortion in self-defense.

Against abortion rights
Human life and a right to life begin at conception; abortion is murder. Human life is continuum of growth that starts at conception, not at birth. The person, therefore, begins at conception. Killing the fetus, thus, destroys a growing person and can be considered murder.

Table 4: Example posts for and against abortion rights

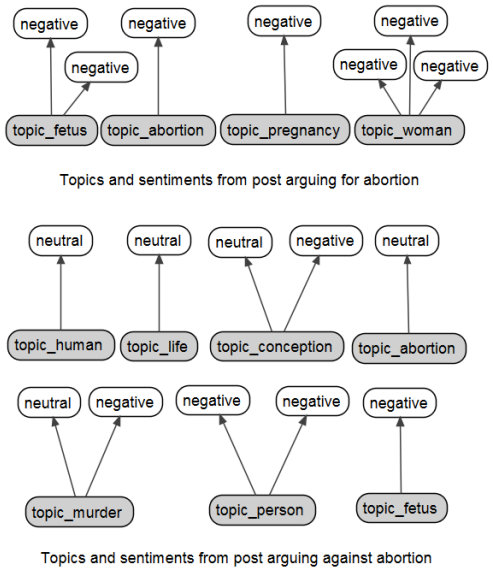


Figure 4: Topic-sentiment model for the two posts in Table 4

ciated with topic terms such as 'fetus', rather than the wider issue (abortion) being debated; here, a negative sentiment expressed towards a fetus is not a negative sentiment expressed towards abortion.

Applying the topic-sentiment model to the sentences in Table 3 arguing against abortion, we can associate a positive sentiment for topic terms such as 'fetus', 'person', 'stage', 'development' and 'human' in sentence 3, and for topic terms 'life' and 'unborn' in sentence 4.

We used Mallet as described in §3.1 to derive the topic terms. For an example of a topic-sentiment model, see Fig. 4, which shows the model obtained for the posts in Table 4.

3.4.3 Unigram model

We used a third baseline feature set containing all unigrams. The more realistic assumption here (compared to equating stance with sentiment) is that writers use different vocabularies to argue for or against an issue, and therefore a model can be learnt that predicts the likelihood of a class based solely on the words used in the post. As mentioned earlier, previous studies have struggled to outperform such a unigram model (Somasundaran and Wiebe, 2010).

3.4.4 Full dependency model

Our proposed feature set for stance classification using a distributional lexical model, stance bearing terminology and points was designed to be scrutable, but therefore made use of only a subset of word-pair features from the dependency graph. We also evaluated this against a baseline feature set which makes use of all word-pairs obtained from the dependency graph.

4 Evaluation

We used the dataset created by Somasundaran and Wiebe (2010) containing monologic posts about five issues: abortion, creation, gay rights, god and gun rights. Somasundaran and Wiebe (2010) reported results on a balanced subset of the corpus with equal numbers of posts for and against each issue. We adopted the same methodology as them to create a balanced subset and evaluated on our balanced dataset containing 4870 posts in total, with

Feature set	Abortion rights*	Creation	Gay rights	Existence of God	Gun rights	Average
Baselines						
B1	51.46	52.56	50.67	53.52	49.48	51.53
B2	57.72	56.64	60.89	61.23	61.77	59.65
B3	77.74	77.50	76.52	76.33	83.95	78.40
B4	87.10	86.13	86.94	85.12	87.71	86.60
Topic DLM						
D1	73.37	67.48	77.87	66.34	70.98	71.20
SVO, amod and nn						
S1	70.45	72.14	72.25	67.20	72.18	70.84
Combined Models						
C1 (D1+S1)	77.35	76.22	78.48	78.06	76.10	77.24
C2 (D1+S1+B3)	84.06	82.86	82.81	83.38	88.39	84.30
C3 (D1+S1+B3+B4)	89.40	87.99	90.18	88.05	93.51	89.26

*Development set

Table 5: Results of supervised learning experiments using Naive Bayes Multinomial model

1030, 856, 1478, 920 and 586 posts for domains of abortion rights, creation, gay rights, god and gun rights, respectively. We developed our ideas by manual examination of the abortion rights debate, leaving the other four debates unseen. We report results for both the development set and the four unseen test sets.

4.1 Classifier and Evaluation Metric

We conducted experiments using Multinomial Naive Bayes classifier implemented in the Weka toolkit (Hall et al., 2009). The Multinomial Naive Bayes model has been previously shown to perform better on text classification tasks with increasing vocabulary size, taking into account word frequencies (McCallum et al., 1998), and this was also our experience. For feature sets produced by each model described in the Methods section, we used the FilteredAttributeEval method available in Weka for feature selection, retaining all features with a score greater than zero. Feature counts were normalised by tf*idf. The performance of the classifier is reported using the accuracy metric, which is most appropriate for a balanced dataset.

4.2 Compared Models

Our discussion in §3 results in the following different models for stance classification. We present in the next section, the results of our experiments.

1. Baseline Models:
   - B1 Sentence level sentiment features.
   - B2 Topic-sentiment features.
   - B3 Unigram features.
   - B4 Dependency features composed of all word pairs connected by a dependency relation.
2. Distributional Lexical Models (DLM):
   - D1 Topic based features resulting from DLM discussed in §3.1.
3. SVO, amod and nn relations based model:
   - S1 The subject-verb-object (SVO) triplets, also broken up into SV and VO pairs, and the word pairs obtained from the amod and nn relations in the dependency parse.
4. Combined Models:
   - C1 D1+S1 - combining topic based features with SVO triplets and word-pairs from amod and nn relations.
   - C2 D1+S1+B3 - combining topic based features with SVO, word-pairs from amod and nn relations, and unigrams.
   - C3 D1+S1+B3+B4 - combining topic based features with SVO, word-pairs from amod and nn relations, unigrams and dependency features.

4.3 Results and analysis

Performance of various models: The 10-fold cross validation results for Multinomial Naive Bayes for different models are reported in Table 5.

As seen in Table 5, the baseline B4 using all relations from the dependency parse performs significantly better compared to other models that focus on selecting specific features for stance classification. The features that we introduce (D1 and S1) become competitive only when combined with one or more baseline models. C3, the best performing model, combines the unigram and dependency baselines with topic DLMs, SVO points, and stance bearing amod and nn relations, and outperforms previously published approaches

to stance classification described in §2 by a substantial margin.

With respect to scrutability, the features in C1, as described earlier in this paper, are easily explained in natural language. C2, the first competitive system, extends C1 with unigram features. These can be easily included in an explanation; for example, *"This post is classified as being in favour of abortion rights because it contains words such as 'extreme' and 'pain'."*. C3, which is the best performing classifier, also uses arbitrary dependency features that are harder to use in explanations. However, even when using C3, the classification decision for the vast majority of posts can be explained using features from C2. Table 6 explores the coverage of different features in the dataset, following feature selection.

Feature set	Coverage
Baselines	
B1	100.00%
B2	32.90%
B3	74.41%
B4	75.10%
Topic DLM	
D1	37.64%
SVO, amod and nn	
S1	40.56%
Combined Models	
C1 (D1+S1)	54.40%
C2 (D1+S1+B3)	80.45%
C3 (D1+S1+B3+B4)	86.58%

Table 6: Percentage of posts containing at least one feature for each feature set (following feature selection)

Poverty of sentiment-based models: While we expected our baseline model B1 that uses an off the shelf sentiment classifier to perform poorly on this task (see example in §3.4.2 for reasoning), we were slightly surprised by the poor performance of the topic-sentiment models (B2). Clearly there is more to stance classification than sentiment, and more effort into modelling the range of lexical associations with topic terms pays off for the distributional lexical models. The unigram model (B3) performed better than the topic-sentiment models (B2) and the off-the-shelf sentiment analysis tool (B1). This supports the results of Somasundaran and Wiebe (2010), who similarly found that sentiment features did not prove helpful, while unigram features were hard to beat. We additionally find that dependency features B4 provide an even stronger baseline.

Comparison with other systems: Our experiments are directly comparable to Somasundaran and Wiebe (2010) as we report results on the same dataset. Our best scoring system achieves an overall accuracy of 89.26%, in comparison to their overall accuracy of 63.63%, a statistically significant increase ($p < 0.0001$; z-test for difference in proportions). Further, our system performs better for each of the debate issues investigated.

While not directly comparable, our results also compare well to studies in dialogic stance classification. For example, Anand et al. (2011) achieved a maximum of 69% accuracy using contextual features based on LIWC, and Walker et al. (2012) obtained a highest of 64% using information related to agreement relations between speakers. Ranade et al. (2013) achieved 70.3% by focusing on capturing users' intent and Sentiwordnet scores. More recently, Hasan and Ng (2014) achieved an overall accuracy of 66.25% for four domains including abortion and gay rights, using features based on dependency parse, frame-semantics, quotations and position information. Their accuracy for abortion and gay rights was 66.3% and 65.7%, respectively. Our approach, unlike these, focuses on a finegrained modelling of the lexical context of important topic terms, and on dependency relations that relate to points and stance bearing phrases. Our results show that this is indeed beneficial.

4.4 Human readable explanations

While previous work in stance classification has primarily focused on the classifier, this is a topic where scrutability is of interest. A user might want to know why a post has been classified in a certain way, and a good response can build trust in the system. The features we have introduced in this paper lend themselves to the generation of explanations. Table 7 shows some example posts (selected to be short due to space constraints), the features (after feature selection) present in the posts, and the generated explanations. The points are generated from the SVO by including all premodifiers of the subject, verb and object in the sentence. The explanation sentence is

Post: Abortion is the woman's choice, not the father's The Father should be told that the woman is having an abortion but until he carries and gives birth to his own baby then it is not his choice to tell the woman that she has to keep and give a painful birth to this fetus.

Points derived using SVO information: ['Abortion is the woman choice'; 'it is not his choice']; *Unigrams:* ['fetus'; 'woman'; 'choice']; *No other features present*

Classified Stance: **For** Abortion rights

Explanation: This post has been classified as being in favour of abortion rights because it makes points such as 'abortion is the woman's choice' and 'it is not his choice', and uses vocabulary such as 'fetus', 'woman' and 'choice'.

Post: A dog is not a person. Therefore, it does not have rights. Positive feelings about dogs should have no bearing on the discussion. A fetus is not a person. Negative feelings about the metaphysically independent status of women should have no bearing on the discussion.

Points derived using SVO information: ['a fetus is not a person'; 'a dog is not a person']; *Unigrams:* ['fetus'; 'independent'; 'bearing']; *No other features present*

Classified Stance: **For** Abortion rights

Explanation: This post has been classified as being in favour of abortion rights because it makes points such as 'a fetus is not a person' and 'a dog is not a person', and uses vocabulary such as 'fetus', 'independent', and 'bearing'.

Post: God exists in the unborn as in the born

Unigrams: ['unborn']; *No other features present*

Classified Stance: **Against** Abortion rights

Explanation: This post has been classified as being against abortion rights because it uses vocabulary such as 'unborn'.

Post: Any abortions should not be aloud if you are stupid enough to get pregnant when you do not want a baby or selfish enough not to want to look after it when you find out it may have an illness then it is your own fault why should the life of an innocent unborn child be killed because of your mistake

amod features: ['unborn child'; 'innocent child']; *Unigrams:* ['baby'; 'unborn'; 'killed']; *No other features present*

Classified Stance: **Against** Abortion rights

Explanation: This post has been classified as being against abortion rights because it uses vocabulary such as 'baby', 'unborn' and 'killed' and subjective phrases such as 'unborn child' and 'innocent child'.

Table 7: Examples of explanations generated for stance classification

based on a very simple template that takes as input a list for each feature type, and populates slots based on which features are present.

5 Conclusions and future work

In this paper, we presented a new feature set for stance classification in online debates, designed to be scrutable by human users as well as capable of achieving high accuracy of classification. We showed that our proposed model significantly outperforms other approaches based on sentiment analysis and topic-sentiment analysis. We believe our models capture some of the subtleties of argumentation in text, by breaking down the stance towards the debated issue into expressed stances towards a variety of related topics, as well as modelling, albeit in a simple way, the notion of a point. However, this is just a first step; we do not yet model the sequence of points or topic-stance changes in the post, or dialogic structure connecting posts. Finally, stance classification is a staging post to more in-depth argumentation mining. Our ultimate goal is to model a richer argumentative framework including the support and rebuttal of claims, and the changing of opinion by users in online debates.

Acknowledgements

This work was supported by the Economic and Social Research Council [Grant number ES/M001628/1].

References

Pranav Anand, Marilyn Walker, Rob Abbott, Jean E Fox Tree, Robeson Bowmani, and Michael Minor. 2011. Cats rule and dogs drool!: Classifying stance in online debate. In *Proceedings of the 2nd workshop on computational approaches to subjectivity and sentiment analysis*, pages 1–9. Association for Computational Linguistics.

David M Blei, Andrew Y Ng, and Michael I Jordan.

2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Marie-Catherine De Marneffe and Christopher D Manning. 2008. The stanford typed dependencies representation. In *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pages 1–8. Association for Computational Linguistics.

Stephan Greene and Philip Resnik. 2009. More than words: Syntactic packaging and implicit sentiment. In *Proceedings of human language technologies: The 2009 annual conference of the north american chapter of the association for computational linguistics*, pages 503–511. Association for Computational Linguistics.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Kazi Saidul Hasan and Vincent Ng. 2014. Why are you taking this stance? identifying and classifying reasons in ideological debates. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 751–762, Doha, Qatar, October. Association for Computational Linguistics.

Mahesh Joshi and Carolyn Penstein-Rosé. 2009. Generalizing dependency features for opinion mining. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 313–316. Association for Computational Linguistics.

Andrew McCallum, Kamal Nigam, et al. 1998. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, pages 41–48. Citeseer.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. http://www.cs.umass.edu/ mccallum/mallet.

Huy V Nguyen and Diane J Litman. 2015. Extracting argument and domain words for identifying argument components in texts. *NAACL HLT 2015*, page 22.

Michael Paul and Roxana Girju. 2010. A two-dimensional topic-aspect model for discovering multi-faceted topics. *Urbana*, 51:61801.

Sarvesh Ranade, Rajeev Sangal, and Radhika Mamidi, 2013. *Proceedings of the SIGDIAL 2013 Conference*, chapter Stance Classification in Online Debates by Recognizing Users' Intentions, pages 61–69. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, volume 1631, page 1642. Citeseer.

Swapna Somasundaran and Janyce Wiebe. 2010. Recognizing stances in ideological on-line debates. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 116–124. Association for Computational Linguistics.

Dhanya Sridhar, Lise Getoor, and Marilyn Walker, 2014. *Proceedings of the Joint Workshop on Social Dynamics and Personal Attributes in Social Media*, chapter Collective Stance Classification of Posts in Online Debate Forums, pages 109–117. Association for Computational Linguistics.

Marilyn A Walker, Pranav Anand, Robert Abbott, and Ricky Grant. 2012. Stance classification using dialogic properties of persuasion. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 592–596. Association for Computational Linguistics.

Theresa Wilson and Janyce Wiebe. 2005. Annotating attributions and private states. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, pages 53–60. Association for Computational Linguistics.