Back up your Stance: Recognizing Arguments in Online Discussions

Filip Boltužić and **Jan Šnajder**
University of Zagreb
Faculty of Electrical Engineering and Computing
Text Analysis and Knowledge Engineering Lab
Unska 3, 10000 Zagreb, Croatia
{filip.boltuzic,jan.snajder}@fer.hr

Abstract

In online discussions, users often back up their stance with arguments. Their arguments are often vague, implicit, and poorly worded, yet they provide valuable insights into reasons underpinning users' opinions. In this paper, we make a first step towards *argument-based opinion mining* from online discussions and introduce a new task of *argument recognition*. We match user-created comments to a set of predefined topic-based arguments, which can be either attacked or supported in the comment. We present a manually-annotated corpus for argument recognition in online discussions. We describe a supervised model based on comment-argument similarity and entailment features. Depending on problem formulation, model performance ranges from 70.5% to 81.8% F1-score, and decreases only marginally when applied to an unseen topic.

1 Introduction

Whether about coffee preparation, music taste, or legal cases in courtrooms, arguing has always been the dominant way of rationalizing opinions. An argument consists of one or more premises leading to exactly one conclusion, while argumentation connects together several arguments (Van Eemeren et al., 2013). Across domains, argumentation differs in vocabulary, style, and purpose, ranging from legal (Walton, 2005) and scientific argumentation (Jiménez-Aleixandre and Erduran, 2007) to media (Walton, 2007) and social argumentation (Shum, 2008). When argumentation involves interactive argument exchange with elements of persuasion, we talk about debating. In the increasingly popular online debates, such as VBATES,[1] users can engage in debates over controversial topics, introduce new arguments or use existing ones.

Early computational approaches to argumentation have developed in two branches: logic-based approaches (Bos and Gabsdil, 2000; Lauriar et al., 2001) and argumentative zoning (Teufel and others, 2000). The latter aims to recognize argumentative sections of specific purpose in scientific papers, such as goals, related work, or conclusion. Moens et al. (2007) introduced argumentation mining as a research area involved with the automatic extraction of argumentation structure from free text, residing between NLP, argumentation theory, and information retrieval.

Prior work in argumentation mining has focused on official documents, such as legal cases (Palau and Moens, 2009), or moderated sources, such as debates (Cabrio and Villata, 2012). However, by far the largest source of opinions are online user discussions: comments on newspaper articles, social networks, blogs, and discussion forums – all argumentation arenas without strict rules. Despite the fact that the user-generated content is not moderated nor structured, one can often find an abundance of opinions, most of them backed up with arguments. By analyzing such arguments, we can gain valuable insight into the reasons underpinning users' opinions. Understanding the reasons has obvious benefits in social opinion mining, with applications ranging from brand analysis to political opinion mining.

Inspired by this idea, in this paper we take on the task of *argument-based opinion mining*. Instead of merely determining the general opinion or stance of users towards a given topic, in argument-based opinion mining we wish to determine the arguments on which the users base their stance. Unlike in argumentation mining, we are not ultimately interested in recovering the argumentation structure. Instead, we wish to recognize what arguments the user has used to back up her opinion.

[1] http://vbate.idebate.org/

Proceedings of the First Workshop on Argumentation Mining, pages 49–58,
Baltimore, Maryland USA, June 26, 2014.

As an example, consider a discussion on the topic *"Should gay marriage be legal?"* and the following comment:

> *Gay marriages must be legal in all 50 states. A marriage is covenant between 2 people regardless of their genders. Discrimination against gay marriage is unconstitutional and biased. Tolerance, education and social justice make our world a better place.*

This comment supports the argument *"It is discriminatory to refuse gay couples the right to marry"* and denies the argument *"Marriage should be between a man and a woman"*. The technical challenge here lies in the fact that, unlike in debates or other more formal argumentation sources, the arguments provided by the users, if any, are less formal, ambiguous, vague, implicit, or often simply poorly worded.

In this paper, we make a first step towards argument-based opinion mining from online discussions and introduce the task of *argument recognition*. We define this task as identifying what arguments, from a predefined set of arguments, have been used in users' comments, and how. We assume that a topic-dependent set of arguments has been prepared in advance. Each argument is described with a single phrase or a sentence. To back up her stance, the user will typically use one or more of the predefined arguments, but in their own wording and with varying degree of explicitness. The task of argument recognition amounts to matching these arguments to the predefined arguments, which can be either attacked or supported by the comment. Note that the user's comment may by itself be a single argument. However, we refer to it as *comment* to emphasize the fact that in general it may contain several arguments as well as non-argumentative text.

The contribution of our work is twofold. First, we present COMARG, a manually-annotated corpus for argument recognition from online discussions, which we make freely available. Secondly, we describe a supervised model for argument recognition based on comment-argument comparison. To address the fact that the arguments expressed in user comments are mostly vague and implicit, we use a series of semantic comment-argument comparison features based on semantic textual similarity (STS) and textual entailment (TE). To this end, we rely on state-of-the-art off-the-shelf STS and TE systems. We consider different feature subsets and argument recognition tasks of varying difficulty. Depending on task formulation, their performance ranges from 70.5% to 81.8% micro-averaged F1-score. Taking into account the difficulty of the task, we believe these results are promising. In particular, we show that TE features work best when also taking into account the stance of the argument, and that a classifier trained to recognize arguments in one topic can be applied to another one with a decrease in performance of less than 3% F1-score.

The rest of the paper is structured as follows. In the next section we review the related work. In Section 3 we describe the construction and annotation of the COMARG corpus. Section 4 describes the argument recognition model. In Section 5 we discuss the experimental results. Section 6 concludes the paper and outlines future work.

2 Related Work

Argument-based opinion mining is closely related to argumentation mining, stance classification, and opinion mining.

Palau and Moens (2009) approach argumentation mining in three steps: (1) argument identification (determining whether a sentence is argumentative), (2) argument proposition classification (categorize argumentative sentences as premises or conclusions), and (3) detection of argumentation structure or "argumentative parsing" (determining the relations between the arguments). The focus of their work is on legal text: the Araucaria corpus (Reed et al., 2008) and documents from the European Court of Human Rights.

More recently, Cabrio and Villata (2012) explored the use of textual entailment for building argumentation networks and determining the acceptability of arguments. Textual entailment (TE) is a generic NLP framework for recognizing inference relations between two natural language texts (Dagan et al., 2006). Cabrio and Villata base their approach on Dung's argumentation theory (Dung, 1995) and apply it to arguments from online debates. After linking the arguments with support/attack relations using TE, they are able to compute a set of acceptable arguments. Their system helps the participants to get an overview of a debate and the accepted arguments.

Our work differs from the above-described work in that we do not aim to extract the argumenta-

tion structure. Similarly to Cabrio and Villata (2012), we use TE as one of the features of our system to recognize the well-established arguments in user generated comments. However, aiming at argument-based opinion mining from noisy comments, we address a more general problem in which each comment may contain several arguments as well as non-argumentative text. Thus, in contrast to Cabrio and Villata (2012) who framed the problem as a binary yes/no entailment task, we tackle a more difficult five-class classification problem. We believe this is a more realistic task from the perspective of opinion mining.

A task similar to argument recognition is that of *stance classification*, which involves identifying a subjective disposition towards a particular topic (Lin et al., 2006; Malouf and Mullen, 2008; Somasundaran and Wiebe, 2010; Anand et al., 2011; Hasan and Ng, 2013). Anand et al. (2011) classified stance on a corpus of posts across a wide range of topics. They analyzed the usefulness of meta-post features, contextual features, dependency features, and word-based features for signaling disagreement. Their results range from 54% to 69% accuracy. Murakami and Raymond (2010) identify general user opinions in online debates. They distinguish between global positions (opinions on the topic) and local positions (opinions on previous remarks). By calculating user pairwise rates of agreement and disagreement, users are grouped into "support" and "oppose" sets.

In contrast to stance classification, argument recognition aims to uncover the reasons underlying an opinion. This relates to the well-established area of opinion mining. The main goal of opinion mining or sentiment analysis (Pang and Lee, 2008) is to analyze the opinions and emotions from (most often user-created) text. Opinions are often associated with user reviews (Kobayashi et al., 2007), unlike stances, which are more common for debates. Hasan and Ng (2013) characterize stance recognition as a more difficult task than opinion mining. Recently, however, there has been interesting work on combining argumentation mining and opinion mining (Chesñevar et al., 2013; Grosse et al., 2012; Hogenboom et al., 2010).

3 COMARG Corpus

For training and evaluating argument recognition models, we have compiled a corpus of user comments, manually annotated with arguments, to which we refer as COMARG. The COMARG corpus is freely available for research purposes.[2]

3.1 Data Description

As a source of data, we use two web sites: *Procon.org*[3] and *Idebate.org*.[4] The former is a discussion site covering ideological, social, political, and other topics. Users express their personal opinions on a selected topic, taking either the pro or con side. *Idebate.org* is a debating website containing online debates and an archive of past debates. Each archived topic contains a set of prominent arguments presented in the debate. Each argument is labeled as either for or against the topic. The arguments are moderated and edited to provide the best quality of information.

The two data sources are complementary to each other: *Procon.org* contains user comments, while *Idebate.org* contains the arguments. We manually identified near-identical topics covered by both web sites. From this set, we chose two topics: *"Under God in Pledge"* (UGIP) and *"Gay Marriage"* (GM). We chose these two topics because they have a larger-than-average number of comments (above 300) and are well-balanced between pro and con stances. For these two topics, we then took the corresponding comments and arguments from *Procon.org* and *Idebate.org*, respectively. As the users can post comments not relevant for the topic, we skim-read the comments and removed the spam. We end up with a set of 175 comments and 6 arguments for the UGIP topic, and 198 comments and 7 arguments for the GM topic. The comments are often verbose: the average number of words per comment is 116. This is in contrast to the less noisy dataset from Cabrio and Villata (2012), where the average comment length is 50 words.

Each comment has an associated stance (pro or con), depending on how it was classified in *Procon.org*. Similarly, each argument either attacks or supports the claim of the topic, depending on how it was classified in *Idebate.org*. To simplify the exposition, we will refer to them as "pro arguments" and "con arguments". Table 1 shows the arguments for UGIP and GM topics.

Users may attack or support both pro and con arguments. We will refer to the way how the argument is *used* (attacked or supported) as *argument*

[2] Freely available under the CC BY-SA-NC license from `http://takelab.fer.hr/data/comarg`

[3] `http://www.procon.org`

[4] `http://idebate.org`

"Under God in Pledge" (UGIP): *Should the words "under God" be in the U.S. Pledge of Allegiance?*		
(A1.1)	*Likely to be seen as a state sanctioned condemnation of religion*	Pro
(A1.2)	*The principles of democracy regulate that the wishes of American Christians, who are a majority are honored*	Pro
(A1.3)	*Under God is part of American tradition and history*	Pro
(A1.4)	*Implies ultimate power on the part of the state*	Con
(A1.5)	*Removing under god would promote religious tolerance*	Con
(A1.6)	*Separation of state and religion*	Con
"Gay Marriage" (GM): *Should gay marriage be legal?*		
(A2.1)	*It is discriminatory to refuse gay couples the right to marry*	Pro
(A2.2)	*Gay couples should be able to take advantage of the fiscal and legal benefits of marriage*	Pro
(A2.3)	*Marriage is about more than procreation, therefore gay couples should not be denied the right to marry due to their biology*	Pro
(A2.4)	*Gay couples can declare their union without resort to marriage*	Con
(A2.5)	*Gay marriage undermines the institution of marriage, leading to an increase in out of wedlock births and divorce rates*	Con
(A2.6)	*Major world religions are against gay marriages*	Con
(A2.7)	*Marriage should be between a man and a woman*	Con

Table 1: Predefined arguments for the two topics in the COMARG corpus

polarity. Typically, but not necessarily, users who take the pro stance do so by supporting one of the pro arguments, and perhaps attacking some of the con arguments, while for users who take the con stance it is the other way around.

3.2 Annotation

The next step was to annotate, for each comment, the arguments used in the comment as well as their polarity. For each topic we paired all comments with all possible arguments for that topic, resulting in 1,050 and 1,386 comment-argument pairs for the UGIP and GM topics, respectively. We then asked the annotators (not the authors) to annotate each pair. The alternative would have been to ask the annotators to assign arguments to comments, but we believe annotating pairs reduces the annotation efforts and improves annotation quality.[5]

[5]We initially attempted to crowdsource the annotation, but the task turned out to be too complex for the workers, resulting in unacceptably low interannotator agreement.

Label	Description: Comment...
A	...explicitly **attacks** the argument
a	...vaguely/implicitly **attacks** the argument
N	...makes **no use** of the argument
s	...vaguely/implicitly **supports** the argument
S	...explicitly **supports** the argument

Table 2: Labels for comment-argument pairs in the COMARG corpus

No, of course not. The original one was good enough. The insertion of Under God" between "Our nation" and "indivisible" is symbolic of how religion divides this country."
The Pledge of Allegiance reflects our morals and values. Therefore, it should reflect the ideas of all Americans not 80%. This country has no national religion, so why should we promote a god. Also, Thomas Jefferson, a founding father, was athiest.
I believe that since this country was founded under God why should we take that out of the pledge? Men and women have fought and gave their lives for this country, so that way we can have freedom and be able to have God in our lives. And since this country was founded under God and the Ten Commandments in mind, it needs to stay in. If it offends you well I am sorry but get out of this country!

Table 3: Example comments with low IAA from UGIP

Acknowledging the fact that user-provided arguments are often vague or implicit, we decided to annotate each comment-argument pair using a five-point scale. The labels are shown in Table 2. The labels encode the presence/absence of an argument in a comment, its polarity, as well as the degree of explicitness.

The annotation was carried out by three trained annotators, in two steps. In the first step, each annotator independently annotated the complete dataset of 2,436 comment-argument pairs. To improve the annotation quality, we singled out the problematic comment-argument pairs. We considered as problematic all comment-argument pairs for which (1) there is no agreement among the three annotators or (2) the ordinal distance between any of the labels assigned by the annotators is greater than one. Table 3 shows some examples of problematic comments. As for the arguments, the most problematic ones are A1.3 and A1.5 for the UGIP topic and arguments A2.1 and A2.7 for the GM topic (cf. Table 1).

In the second step, we asked the annotators to independently revise their decisions for the problematic comment-argument pairs. Each annotator re-annotated 515 pairs, of which for 86 the annotations were revised. In total, the annotation and

IAA	UGIP	GM	UGIP+GM
Fleiss' Kappa	0.46	0.51	0.49
Cohen's Kappa	0.46	0.51	0.49
Weighted Kappa	0.45	0.51	0.50
Pearson's r	0.68	0.74	0.71

Table 4: Interannotator agreement on the COMARG corpus

	Labels					
Topic	**A**	**a**	**N**	**s**	**S**	Total
UGIP	48	86	691	58	130	1,013
GM	89	73	849	98	176	1,285
UGIP+GM	137	159	1,540	156	306	2,298

Table 5: Distribution of labels in the COMARG corpus

subsequent revision took about 30 person-hours.

Table 4 shows the interannotator agreement (IAA). We compute Fleiss' multirater kappa, Cohen's kappa (averaged over three annotator pairs), Cohen's linearly weighted kappa (also averaged), and Pearson's r. The latter two reflect the fact that the five labels constitute an ordinal scale. According to standard interpretation (Landis and Koch, 1977), these values indicate moderate agreement, proving that argument recognition is a difficult task.

Finally, to obtain the the gold standard annotation, we took the majority label for each comment-argument pair, discarding the pairs for which there are ties. We ended up with a dataset of 2,249 comment-argument pairs. Table 6 shows examples of annotated comment-argument pairs.

3.3 Annotation Analysis

Table 5 shows the distribution of comment-argument pairs across labels. Expectedly, the majority (67.0%) of comment-argument pairs are cases in which the argument is not used (label **N**). Attacked arguments (labels **A** or **a**) make up 12.9%, while supported arguments (labels **S** or **s**) make up 20.1% of cases. Among the cases not labeled as **N**, arguments are used explicitly in 58.4% (labels **A** and **S**) and vague/implicit (labels **a** and **s**) in 41.5% of cases. There is a marked difference between the two topics in this regard: in UGIP, arguments are explicit in 55.3%, while in GM in 60.7% of cases. Note that this might be affected by the choice of the predefined arguments as well as how they are worded.

The average number of arguments per comment is 1.9 (1.8 for UGIP and 2.0 for GM). In GM, 62.8% of arguments used are pro arguments, while in UGIP pro arguments make up 52.2% of cases.

4 Argument Recognition Model

We cast the argument recognition task as a multi-class classification problem. Given a comment-argument pair as input, the classifier should predict the correct label from the set of five possible labels (cf. Table 2). The main idea is for the classifier to rely on comment-argument comparison features, which in principle makes the model less domain dependent than if we were to use features extracted directly from the comment or the arguments.

We use three kinds of features: textual entailment (TE) features, semantic text similarity (STS) features, and one "stance alignment" (SA) feature. The latter is a binary feature whose value is set to one if a pro comment is paired with a pro argument or if a con comment is paired with a con argument. This SA feature presupposes that comment stance is known a priori. The TE and STS features are described bellow.

4.1 Textual Entailment

Following the work of Cabrio and Villata (2012), we use textual entailment (TE) to determine whether the comment (the text) entails the argument phrase (the hypothesis). To this end we use the Excitement Open Platform (EOP), a rich suite of textual entailment tools designed for modular use (Padó et al., 2014). From EOP we used seven pre-trained *entailment decision algorithms* (EDAs). Some EDAs contain only syntactical features, whereas others rely on resources such as WordNet (Fellbaum, 1998) and VerbOcean (Chklovski and Pantel, 2004). Each EDA outputs a binary decision (*Entailment* or *NonEntailment*) along with the degree of confidence. We use the outputs (decisions and confidences) of all seven EDAs as the features of our classifier (14 features in total). We also experimented with using additional features (the disjunction of all classifier decisions, the maximum confidence value, and the mean confidence value), but using these did not improve the performance.

In principle, we expect the comment text (which is usually longer) to entail the argument phrase (which is usually shorter). This is also confirmed by the ratio of positive entailment decision across labels (averaged over seven EDAs), shown in

Id	Comment	Argument	Label
2.23.4	*All these arguments on my left are and have always been FALSE. Marriage is between a MAN and a WOMAN by divine definition. Sorry but, end of story.*	*It is discriminatory to refuse gay couples the right to marry.*	**s**
2.111.4	*Marriage isn't the joining of two people who have intentions of raising and nurturing children. It never has been. There have been many married couples whos have not had children. (...) If straight couples can attempt to work out a marriage, why can't homosexual couple have this same privilege? (...)*	*It is discriminatory to refuse gay couples the right to marry*	**s**
2.114.2	*(...) I truly believe that the powers behind the cause to re-define marriage stem from a stronger desire to attack a religious institution that does not support homosexuality, rather than a desire to achieve the same benefits as marriage for same sex couples. (...)"*	*Gay couples should be able to take advantage of the fiscal and legal benefits of marriage.*	**S**
2.101.2	*(...) One part of marriage is getting benefits from the other. Many married couples never have children but still get the benefits of marriage, should we take those benefits away because they don't have children? Another is the promise to be with each other for an eternity" etc. Marriage is also about being able to celebrate having each other. And last, marriage is about being there for each other. (...)"*	*Gay couples should be able to take advantage of the fiscal and legal benefits of marriage.*	**S**
2.157.2	*(...) There are no legal reasons why two homosexual people should not be allowed to marry, only religious ones (...)*	*Gay couples should be able to take advantage of the fiscal and legal benefits of marriage.*	**N**
1.45.2	*I am not bothered by under God but by the highfalutin christians that do not realize that phrase was never in the original pledge - it was not added until 1954. So stop being so pompous and do not offend my parents and grandparents who never used "under God" when they said the pledge. Let it stay, but know the history of the Cold War and fear of communism.*	*"Under God" is part of American tradition and history.*	**a**

Table 6: Example of comment-argument annotations from the COMARG corpus

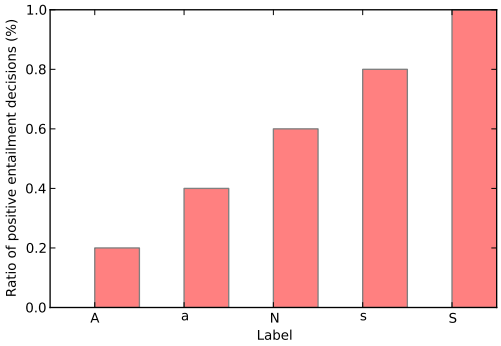

Figure 1: Ratio of positive entailment decisions across labels, scaled to a $[0, 1]$ interval

Fig. 1. Pro arguments have a higher ratio of positive entailment decisions than con arguments. Also, vaguely/implicitly supported arguments have a lower rate of entailment decisions than explicitly supported arguments.

4.2 Semantic Textual Similarity

Formally speaking, the argument should either be entailed or not entailed from the comment. The former case also includes a simple argument paraphrase. In the latter case, the argument may be contradicted or it may simply be a *non sequitur*. While we might expect these relations to be recognizable in texts from more formal genres, such as legal documents and parliamentary debates, it is questionable to what extent these relations can be detected in user-generated content, where the arguments are stated vaguely and implicitly.

To account for this, we use a series of argument-comment comparison features based on semantic textual similarity (STS). STS measures "the degree of semantic equivalence between two texts" (Agirre et al., 2012). It is a looser notion than TE and, unlike TE, it is a non-directional (symmetric) relation. We rely on the freely available TakeLab STS system by Šarić et al. (2012). Given a comment and an argument, the STS system outputs a continuous similarity score. We also compute the similarity between the argument and each sentence from the comment, which gives us a vector of similarities. The vector length equals the largest number of sentences in a comment, which in COMARG is 29. Additionally, we compute the maximum and the

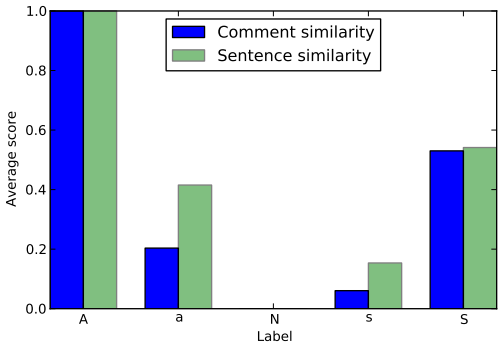

Figure 2: Average similarity score on sentence and comment level across labels, scaled to a $[0, 1]$ interval

mean of sentence-level similarities. In total, we use 31 STS features.

Fig. 2 shows the average comment- and sentence-level similarity scores across labels on COMARG, scaled to a $[0, 1]$ interval. Interestingly, attacked arguments on average receive a larger score than supported arguments.

5 Experimental Evaluation

5.1 Experimental Setup

We consider three formulations of the argument detection task. In the first setting (**A-a-N-s-S**), we consider the classification of a comment-argument into one of the five labels, i.e., we wish to determine whether an argument has been used, its polarity, as well as the degree of explicitness. In the second setting (**As-N-sS**), we conflate the two labels of equal polarity, thus we only consider whether an argument has been used and with what polarity. In the third setting (**A-N-S**), we only consider the comment-argument pairs in which arguments are either not used or used explicitly. This setting is not practically relevant, but we include it for purposes of comparison.

We compare to two baselines: (1) a majority class classifier (MCC), which assigns label **N** to every instance, and (2) a bag-of-words overlap classifier (BoWO), which uses the word overlap between the comment and the argument as the only feature.

For classification, we use the Support Vector Machine (SVM) algorithm with a Radial Basis Function kernel. In each setting, we train and evaluate the model using nested 5×3 cross-validation. The hyperparameters C and γ of the SVM are optimized using grid search. We rely on the well-known LibSVM implementation (Chang and Lin, 2011).

	A-a-N-s-S		**Aa-N-sS**		**A-N-S**	
Model	UGIP	GM	UGIP	GM	UGIP	GM
MCC baseline	68.2	69.4	68.2	69.4	79.5	76.6
BoWO baseline	68.2	69.4	67.8	69.5	79.6	76.9
TE	69.1	**81.1**	69.6	72.3	80.1	73.4
STS	67.8	68.7	67.3	69.9	79.2	75.8
SA	68.2	69.4	68.2	69.4	79.5	76.6
STS+SA	68.2	69.5	67.5	68.7	79.6	76.1
TE+SA	68.9	72.4	**71.0**	**73.7**	**81.8**	**80.3**
TE+STS+SA	**70.5**	72.5	68.9	73.4	81.4	79.7

Table 7: Argument recognition F1-score (separate models for UGIP and GM topics)

	UGIP → GM		GM → UGIP	
Model	**A-a-N-s-S**	**Aa-N-sS**	**A-a-N-s-S**	**Aa-N-sS**
STS+SA	69.4	69.4	68.2	68.2
TE+SA	**72.6**	**73.5**	**70.2**	**71.2**
STS+TE+SA	71.5	72.2	68.2	69.6

Table 8: Argument recognition F1-score on UGIP and GM topics (cross-topic setting)

5.2 Results

Table 7 shows the micro-averaged F1-score for the three problem formulations, for models trained separately on UGIP and GM topics. The two baselines perform similarly. The models that use only the STS or the SA features perform similar to the baseline. The TE model outperforms the baselines in all but one setting and on both topics: the difference ranges from 0.6 to 11.7 percentage points, depending on problem formulation, while the variation between the two topics is negligible. The STS model does not benefit from adding the SA feature, while the TE model does so in simpler settings (**Aa-N-sS** and **A-N-S**), where the average F1-scores increases by about 3 percentage points. This can be explained by referring to Fig. 1, which shows that even for the attacked arguments (labels **A** and **a**) entailment decisions are sometimes positive. In such cases, the stance alignment feature helps to distinguish between entailment (supported argument) and contradiction (attacked argument). Combining all three feature types gives the best results for the **A-a-N-s-S** setting and the UGIP topic.

The above evaluation was carried out in a within-topic setting. To test how the models perform when applied to comments and arguments from unseen topics, we trained each model on one topic and

	A-a-N-s-S				Aa-N-sS				A-N-S			
Model	P	R	F1	micro-F1	P	R	F1	micro-F1	P	R	F1	micro-F1
MCC baseline	13.8	20.0	16.3	68.9	23.0	33.3	27.2	68.9	26.0	33.3	29.2	77.9
TE+SA	**47.6**	26.6	27.9	71.1	**68.8**	**46.6**	**49.4**	**73.3**	**66.1**	**47.3**	**51.1**	**81.6**
STS+TE+SA	46.3	**27.2**	**28.6**	**71.6**	61.6	43.5	45.5	71.4	63.7	44.9	48.2	80.4

Table 9: Argument recognition F1-score for TE+SA and STS+TE+SA models on UGIP+GM topics

evaluated on the other. The results are shown in Table 8 (we show results only for the two problem formulations of practical interest). The difference in performance is small (0.7 on average). The best-performing model (TE+SA) does not suffer a decrease in performance. This suggests that the models are quite topic independent, but a more detailed study is required to verify this finding.

Finally, we trained and tested the TE+SA and STS+TE+SA models on the complete COMARG dataset. The results are shown in Table 9. We report macro-averaged precision, recall, and F1-score, as well as micro-averaged F1-score.[6] Generally, our models perform less well on smaller classes (**A**, **a**, **s**, and **S**), hence the macro-averaged F1-scores are much lower than the micro-averaged F1-scores. The recall is lower than the precision: the false negatives are mostly due to our models wrongly classifying comment-argument pairs as **N**. The STS+TE+SA model slightly outperforms the TE+SA model on the **A-a-N-s-S** problem, while on the other problem formulations the TE+SA model performs best.

5.3 Error Analysis

The vague/implicit arguments posed the greatest challenge for all models. A case in point is the comment-argument pair 2.23.4 from Table 6. Judging solely from the comment text, it is unclear what the user actually meant. Perhaps the user is attacking the argument, but there are certain additional assumptions that would need to be met for the argument to be entailed.

The second major problem is distinguishing between arguments that are mentioned and those that are not. Consider the comment-argument pairs 2.111.4 and 2.114.2 from Table 6. In the former case, classifier mistakenly predicts **S** instead of **s**. The decision is likely due to the low difference in argument-comment similarities for these two classes. In the latter example the classifier wrongly predicts that the argument is used in the comment.

The TE model in the majority of cases outperforms the STS model. Nonetheless, in case of the comment-argument pair 2.157.2 from Table 6, the STS-based model outperformed the entailment model. In this case, the word overlap between the argument and the comment in quite high, although they completely differ in meaning. Conversely, argument-comment 2.101.2 is a good example of when entailment was correctly recognized, whereas the STS model has failed.

6 Conclusion

In this paper we addressed the argument recognition task as a first step towards argument-based opinion mining from online discussions. We have presented the COMARG corpus, which consists of manually annotated comment-argument pairs. On this corpus we have trained a supervised model for three argument recognition tasks of varying difficulty. The model uses textual entailment and semantic textual similarity features. The experiments as well as the inter-annotator agreement show that argument recognition is a difficult task. Our best models outperform the baselines and perform in a 70.5% to 81.8% micro-averaged F1-score range, depending on problem formulation. The outputs of several entailment decision algorithms, combined with a stance alignment feature, proved to be the best features. Additional semantic textual similarity features seem to be useful in when we distinguish between vague/implicit and explicit arguments. The model performance is marginally affected when applied to an unseen topic.

This paper has only touched the surface of argument recognition. We plan to extend the COMARG corpus with more topics and additional annotation, such as argument segments. Besides experimenting with different models and feature sets, we intend to investigate how argument interactions can be exploited to improve argument recognition, as well as how argument recognition can be used for stance classification.

[6] We replace undefined values with zeros when computing the macro-averages.

References

Eneko Agirre, Mona Diab, Daniel Cer, and Aitor Gonzalez-Agirre. 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 385–393. Association for Computational Linguistics.

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.

Johan Bos and Malte Gabsdil. 2000. First-order inference and the interpretation of questions and answers. *Proceedings of Gotelog*, pages 43–50.

Elena Cabrio and Serena Villata. 2012. Combining textual entailment and argumentation theory for supporting online debates interactions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 208–212. Association for Computational Linguistics.

Chih-Chung Chang and Chih-Jen Lin. 2011. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27.

Carlos I Chesñevar, María Paula González, Kathrin Grosse, and Ana Gabriela Maguitman. 2013. A first approach to mining opinions as multisets through argumentation. In *Agreement Technologies*, pages 195–209. Springer.

Timothy Chklovski and Patrick Pantel. 2004. Verbocean: Mining the web for fine-grained semantic verb relations. In *EMNLP*, volume 2004, pages 33–40.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190. Springer.

Phan Minh Dung. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321–357.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Kathrin Grosse, Carlos Iván Chesñevar, and Ana Gabriela Maguitman. 2012. An argument-based approach to mining opinions from Twitter. In *AT*, pages 408–422.

Kazi Saidul Hasan and Vincent Ng. 2013. Extra-linguistic constraints on stance recognition in ideological debates. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 816–821.

Alexander Hogenboom, Frederik Hogenboom, Uzay Kaymak, Paul Wouters, and Franciska De Jong. 2010. Mining economic sentiment using argumentation structures. In *Advances in Conceptual Modeling – Applications and Challenges*, pages 200–209. Springer.

María Pilar Jiménez-Aleixandre and Sibel Erduran. 2007. Argumentation in science education: An overview. In *Argumentation in Science Education*, pages 3–27. Springer.

Nozomi Kobayashi, Kentaro Inui, and Yuji Matsumoto. 2007. Extracting aspect-evaluation and aspect-of relations in opinion mining. In *EMNLP-CoNLL*, pages 1065–1074.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Stanislao Lauriar, Johan Bos, Ewan Klein, Guido Bugmann, and Theocharis Kyriacou. 2001. Training personal robots using natural language instruction. *IEEE Intelligent Systems*, 16(5):38–45.

Wei-Hao Lin, Theresa Wilson, Janyce Wiebe, and Alexander Hauptmann. 2006. Which side are you on?: Identifying perspectives at the document and sentence levels. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, pages 109–116. Association for Computational Linguistics.

Robert Malouf and Tony Mullen. 2008. Taking sides: User classification for informal online political discourse. *Internet Research*, 18(2):177–190.

Marie-Francine Moens, Erik Boiy, Raquel Mochales Palau, and Chris Reed. 2007. Automatic detection of arguments in legal texts. In *Proceedings of the 11th International Conference on Artificial Intelligence and Law*, pages 225–230. ACM.

Akiko Murakami and Rudy Raymond. 2010. Support or oppose?: Classifying positions in online debates from reply activities and opinion expressions. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 869–875. Association for Computational Linguistics.

Sebastian Padó, Tae-Gil Noh, Asher Stern, Rui Wang, and Roberto Zanoli. 2014. Design and realization of a modular architecture for textual entailment. *Natural Language Engineering*, FirstView:1–34, 2.

Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation mining: The detection, classification and structure of arguments in text. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, pages 98–107. ACM.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135.

Chris Reed, Raquel Mochales Palau, Glenn Rowe, and Marie-Francine Moens. 2008. Language resources for studying argument. In *Proceedings of the 6th Conference on Language Resources and Evaluation (LREC 2008)*, pages 91–100.

Frane Šarić, Goran Glavaš, Mladen Karan, Jan Šnajder, and Bojana Dalbelo Bašić. 2012. Takelab: Systems for measuring semantic text similarity. In *Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 441–448, Montréal, Canada, 7-8 June. Association for Computational Linguistics.

Simon Buckingham Shum. 2008. Cohere: Towards web 2.0 argumentation. volume 8, pages 97–108.

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.

Simone Teufel et al. 2000. *Argumentative zoning: Information extraction from scientific text*. Ph.D. thesis, University of Edinburgh.

Frans H. Van Eemeren, Rob Grootendorst, Ralph H. Johnson, Christian Plantin, and Charles A. Willard. 2013. *Fundamentals of argumentation theory: A handbook of historical backgrounds and contemporary developments*. Routledge.

Douglas Walton. 2005. *Argumentation methods for artificial intelligence in law*. Springer.

Douglas Walton. 2007. *Media argumentation: dialectic, persuasion and rhetoric*. Cambridge University Press.