# Stance Classification of Context-Dependent Claims

**Roy Bar-Haim**[1], **Indrajit Bhattacharya**[2], **Francesco Dinuzzo**[3*]
**Amrita Saha**[2], and **Noam Slonim**[1]

[1]IBM Research - Haifa, Mount Carmel, Haifa, 31905, Israel
[2]IBM Research - Bangalore, India
[3]IBM Research - Ireland, Damastown Industrial Estate, Dublin 15, Ireland
{roybar,noams}@il.ibm.com, {indrajitb,amrsaha4}@in.ibm.com

## Abstract

Recent work has addressed the problem of detecting relevant claims for a given controversial topic. We introduce the complementary task of *claim stance classification*, along with the first benchmark dataset for this task. We decompose this problem into: (a) open-domain target identification for topic and claim (b) sentiment classification for each target, and (c) open-domain contrast detection between the topic and the claim targets. Manual annotation of the dataset confirms the applicability and validity of our model. We describe an implementation of our model, focusing on a novel algorithm for contrast detection. Our approach achieves promising results, and is shown to outperform several baselines, which represent the common practice of applying a single, monolithic classifier for stance classification.

## 1 Introduction

The need for making persuasive arguments arises in many domains, including politics, law, marketing, and financial and business advising. On-demand generation of pro and con arguments for a given controversial topic would therefore be of great practical value. Natural use cases include *debating support*, where the user is presented with persuasive arguments for a topic of interest, and *decision support*, where the pros and cons of a given proposal are presented to the user.

A notable research effort in this area is the IBM Debater® project whose goal is "to develop technologies that can assist humans to debate and reason"[1]. As part of this research, Levy et al. (2014) have developed *context-dependent claim detection*. Given a controversial *topic*, such as

(1) *The sale of violent video games to minors should be banned* ,

their system extracts, from corpora such as Wikipedia, *Context-Dependent Claims (CDCs)*, defined as "general, concise statements that directly support or contest the given Topic". A claim forms the basis of an argument, being the assertion that the argument aims to establish, and therefore claim detection may be viewed as a first step in automated argument construction. Recent research on claim detection (Levy et al., 2014; Lippi and Torroni, 2015) was facilitated by the IBM argumentative structure dataset (Aharoni et al., 2014), which contains manually collected claims for a variety of topics, as well as supporting evidence.

In this work we introduce the related task of *Claim Stance Classification*: given a topic, and a set of claims extracted for it, determine for each claim whether it supports or contests the topic. Sorting extracted claims into *Pro* and *Con* would clearly improve the usability of both debating and decision support systems. We introduce the first benchmark for this task, by adding Pro/Con annotations to the claims in the IBM dataset.

Based on the analysis of this dataset, we propose a semantic model for predicting claim stance. We observed that both the debate topic and a supporting/contesting claim often contain a *target* phrase, about which they make a positive or a negative statement. The pro/con relation can then be determined by the sentiments of the topic and the claim towards their targets, as well as the semantic relation between these targets. For example, suppose that a topic expresses support for *freedom of*

*Present affiliation - Amazon.

[1]http://researcher.ibm.com/researcher/view_group.php?id=5443



*Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 251–261,
Valencia, Spain, April 3-7, 2017. 

*speech*. A *Pro* claim may support it by arguing in favor of *free discussion*, or alternatively by criticizing *censorship*. We say that *freedom of speech* and *free discussion* are *consistent* targets, while *freedom of speech* and *censorship* are *contrastive*. Accordingly, we suggest that claim stance classification can be reduced to simpler, more tractable sub-problems:

1. Identify the targets of the given topic and claim.
2. Identify the polarity (*sentiment*) towards each of the targets.
3. Determine whether the targets are consistent or contrastive.

While our model seems intuitive, it was not clear a priori how well it captures the semantics of claims in practice. Some types of claims do not fit into this decomposition. Consider the following *Con* claim for the topic given in (1):

(2) *Parents, not government bureaucrats, have the right to decide what is appropriate for their children.*

In this example, there is no clear sentiment target in the claim that is either consistent or contrastive with *the sale of violent video games to minors*. Nevertheless, extensive data annotation confirmed that our model is applicable to about 95% of the claims in the dataset, and for these claims, Pro/Con relations can be accurately predicted by solving the above sub-problems. Furthermore, our analysis reveals that *contrastive targets* are quite common, and thus must be accounted for. Our model highlights intriguing subproblems such as *open-domain target identification* and *open-domain contrast detection* between a given pair of phrases, which have received relatively little attention in previous stance classification work. We hope that the annotated data collected in this work will facilitate further research on these important subtasks.

We developed a classifier for each of the above subtasks. Most notably, we present a novel method for the challenging task of contrast detection. Empirical evaluation confirms that our modular approach outperforms several strong baselines that employ a single, monolithic classier.

## 2 Related Work

Previous work on *stance classification* focused on analyzing debating forums (Somasundaran and Wiebe, 2009; Somasundaran and Wiebe, 2010; Walker et al., 2012b; Hasan and Ng, 2013; Walker et al., 2012a; Sridhar et al., 2014), congressional floor debates (Thomas et al., 2006; Yessenalina et al., 2010; Burfoot et al., 2011), public comments on proposed regulations (Kwon et al., 2007), and student essays (Faulkner, 2014). Most of these works relied on both generic features such as sentiment, and topic-specific features learned from labeled data for a closed set of topics. Simple classifiers with unigram or ngram features are known to be hard to beat for these tasks (Somasundaran and Wiebe, 2010; Hasan and Ng, 2013; Mohammad et al., 2016).

In addition to content-based features, previous work also made use of various types of contextual information, such as agreement/disagreement between posts or speeches, author identity, conversation structure in debating forums, and discourse structure. Collective classification has been shown to improve performance (Thomas et al., 2006; Yessenalina et al., 2010; Burfoot et al., 2011; Hasan and Ng, 2013; Walker et al., 2012a; Sridhar et al., 2014).

The setting of ad-hoc claim retrieval, which we address in this work, is different in several respects. First, topics are not known in advance. They may be arbitrarily complex, and belong to any domain. Second, much of the contextual information that was exploited in previous work is not available in this setting. In addition, claims are short sentences, while previous work typically addressed text spanning one or more paragraphs. Moreover, since we may want to present to the user only claims for which we are confident about stance, reliable confidence ranking of our predictions is important. We explore this aspect in our evaluation.

Consequently, our approach relies on generic sentiment analysis, rather than on topic or domain-specific features. We focus on precise semantic analysis of the debate topic and the claim, including target identification, and contrast detection between the claim and the topic targets. While sentiment analysis is a well-studied task, open-domain target identification and open-domain contrast detection between two given phrases have received little attention in previous work.

*Consistent/contrastive* targets were previously discussed by Somasundaran et al. (2009) [2], who

[2]Termed *same/alternative* in their paper.

used them in conjunction with discourse relations to improve the prediction of opinion polarity. However, these targets and relations were not automatically identified, but rather taken from a labeled dataset. Somasundaran and Wiebe (2009) considered debates comparing two products, such as *Windows* and *Mac*. In comparison, topics in our setting are not limited to product names, and the scope of contrast we address is far more general.

Cabrio and Villata (2013) employ textual entailment to detect *support/attack* relations between arguments. However, as illustrated in Table 1, claims typically refer to the pros and cons of the topic target, but do not entail or contradict the topic.

A recent related task is the SemEval 2016 tweets stance classification (Mohammad et al., 2016). In particular, in its weakly supervised subtask (Task B), no labeled training data was provided for the single assessed topic (*Donald Trump*). Beyond the obvious differences in language and content between claims and tweets, the setting of this task is rather different from ours: the topic was known in advance to the participants, and an unlabeled corpus of related tweets was provided. Top performing systems took advantage of this setting, and developed offline rules for automatically labeling the domain corpus. In our setting, the topic is not known in advance, and obtaining a large collection of claims for a given topic does not seem feasible.

## 3 The Claim Polarity Dataset

The IBM argumentative structure dataset published by Aharoni et al. (2014) contains claims and evidence for 33 controversial topics. In this work we used an updated version of this dataset, which includes 55 topics. Topics were selected at random from the debate motions database at the International Debate Education Association (IDEA) website[3]. Motions are worded as "This house ...", in the tradition of British Parliamentary debates. Claims and evidence were manually collected from hundreds of Wikipedia articles. The dataset contains 2,394 claims.

By definition, all claims in the dataset either support or contest the topic, and Aharoni et al. give a few examples for *Pro* and *Con* claims in their paper. However, the dataset itself does not include stance annotations. We enhanced the dataset with polarity annotations as follows. The polarity of each claim with respect to the motion (*Pro/Con*) was assessed by five annotators, and the final label was determined by the majority annotation.[4] Table 1 shows examples of motions, claims and their pro/con labeling.

[3] http://idebate.org/

## 4 Semantic Model for Claim Stance Classification

In this section we propose a model for predicting the stance of a claim $c$ towards a topic sentence $t$.

We assume that $c$ includes a *claim target* $x_c$, defined as a phrase about which $c$ makes a positive or a negative assertion. Specifically, it is defined as the most explicit and direct sentiment target in the claim. The *claim sentiment* $s_c \in \{-1, 1\}$ is the sentiment of the claim towards its target, where 1 denotes positive sentiment and $-1$ denotes negative sentiment. Similarly, we define for a topic $t$ the *topic target* $x_t$ and *topic sentiment* $s_t$.

We say that the claim target $x_c$ is *consistent* with the topic target $x_t$ if the stance towards $x_c$ implies the *same* stance towards $x_t$ . Similarly, $x_c$ and $x_t$ are *contrastive* if the stance towards $x_c$ implies the *opposite* stance towards $x_t$ . The *contrast relation* between $x_c$ and $x_t$, denoted $\mathcal{R}(x_c, x_t) \in \{-1, 1\}$ is 1 if $x_c$ and $x_t$ are consistent, and $-1$ if they are contrastive. Using the above definitions, we define the stance relation between $c$ and $t$ as

$$Stance(c, t) = s_c \times \mathcal{R}(x_c, x_t) \times s_t \qquad (1)$$

where $Stance(c, t) \in \{-1, 1\}$, 1 indicates *Pro* and $-1$ indicates *Con*. Rows 1-8 in Table 1 show examples for $x_c$, $s_c$, $x_t$, $s_t$ and $\mathcal{R}(x_c, x_t)$. It is easy to verify that the model correctly predicts the claim polarity for these examples. For instance, row 3 has $x_c$="Unity", $x_t$="Multiculturalism", $s_c = 1$, $\mathcal{R}(x_c, x_t) = -1$, $s_t = 1$, and the resulting stance is $1 \times (-1) \times 1 = -1$ (*Con*).

**Continuous model:** The above model produces binary output (+1/-1). In practice, it would be desirable to obtain confidence ranking of the model predictions, which would allow presenting to the user only the top $k$ predictions, or predictions whose confidence is above some threshold. We therefore implemented a continuous variant of the model, where $s_c$, $s_t$, $\mathcal{R}(x_c, x_t)$ and the resulting stance score are all real-valued numbers in [-1,1].

[4] Note that while we considered the original motion phrasing for Pro/Con labeling, the original dataset only contains motion themes as the topics, e.g. *boxing* for "This house would ban boxing".

| # | Debate Topic (Motion) | | Claim | |
|---|---|---|---|---|
| 1 | This house believes that **advertising** is harmful. $\ominus$ | $\Leftrightarrow$ | **Marketing** promotes consumerism and waste. $\ominus$ | Pro |
| 2 | This house would ban **boxing**. $\ominus$ | $\Leftrightarrow$ | **Boxing** remains the 8th most deadly sport. $\ominus$ | Pro |
| 3 | This house would embrace **multiculturalism**. $\oplus$ | $\not\Leftrightarrow$ | **Unity** is seen as an essential feature of the nation and the nation-state. $\oplus$ | Con |
| 4 | This house supports **the one-child policy of the republic of China**. $\oplus$ | $\not\Leftrightarrow$ | **Children with many siblings** receive fewer resources. $\ominus$ | Pro |
| 5 | This house would **build hydroelectric dams**. $\oplus$ | $\Leftrightarrow$ | As an alternative energy source, **a hydroelectric power source** is cheaper than both nuclear and wind power. $\oplus$ | Pro |
| 6 | This house believes that it is sometimes right for the government to restrict **freedom of speech**. $\ominus$ | $\Leftrightarrow$ | **Human rights** can be limited or even pushed aside during times of national emergency. $\ominus$ | Pro |
| 7 | This house would abolish **the monarchy**. $\ominus$ | $\Leftrightarrow$ | **Hereditary succession** is outdated. $\ominus$ | Pro |
| 8 | This house would **unleash the free market** $\oplus$ | $\not\Leftrightarrow$ | Virtually all developed countries today successfully promoted their national industries through **protectionism**. $\oplus$ | Con |
| 9 | This house supports **the one-child policy of the republic of China**. $\oplus$ | | If, for any reason, the single child is unable to care for their older adult relatives, the oldest generations would face a lack of resources and necessities. | Con |

Table 1: Sample topic and claim annotations. Targets are marked in bold. $\oplus$/$\ominus$ denote positive/negative sentiment towards the target, and $\Leftrightarrow$/$\not\Leftrightarrow$ denote consistent/contrastive targets.

For each real-valued prediction, the class is given by its sign, and the confidence is given by its absolute value.

## 5 Model Assessment via Manual Data Annotation

We assessed the validity and applicability of the proposed model through manual annotation of the IBM dataset.[5] The labeled data was also used to train and assess sub-components in the model implementation. This section describes the annotation process and the analysis of the annotation results.

**Annotation Process:** Each of the 55 topics was annotated by one of the authors for its target $x_t$ and sentiment $s_t$. $x_t$ was used as an input for the claim annotation task. Each claim was labeled independently by five annotators who were given the definitions for claim target $x_c$, claim sentiment $s_c$ and the contrast relation $\mathcal{R}(x_c, x_t)$ (cf. Section 4). The annotators were first asked to identify $x_c$ and $s_c$. If successful, they proceeded to determine $\mathcal{R}(x_c, x_t)$.

The final claim labels were derived from the five individual annotations as follows. First, overlapping claim targets were clustered together. If no cluster contained the majority of the annotations ($\geq$3), then the claim was labeled as incompatible with our model. If a majority cluster was found, we discarded annotations where the target was not in this cluster, and selected $x_c$, $s_c$ and $\mathcal{R}(x_c, x_t)$ based on the majority of the remaining annotations. We required absolute majority agreement ($\geq$3) for $s_c$ and $\mathcal{R}(x_c, x_t)$, otherwise the claim was labeled as incompatible with our model.

Rows 1-8 in Table 1 show some examples of annotated claims in our dataset. Row 9 is an example of a claim that was found incompatible with our model.

**Data Annotation Results:** Majority cluster was found for 98.5% of the claims, and for 92.5% of the claims, the majority of the annotators agreed on the exact boundaries of the target. 94.4% of the claims were found to be compatible with our model. Furthermore, combining the labels for $s_c$, $\mathcal{R}(x_c, x_t)$ and $s_t$ as in Equation (1) correctly predicted the Pro/Con labels in the dataset (which were collected independently and were not presented to the annotators) for 99.6% of the compatible claims. Given that the pro/con labels are approximately balanced (55.3% are Pro, 44.7% are Con), this result provides a clear and strong evidence for the applicability and validity of the proposed model. This near-perfect correspondence also indicates the high quality of both *Pro*/*Con* labels and the model-based annotations.

Similar to pro/con labels, claim sentiment is approximately balanced between positive and negative (55% negative vs. 45% positive). Interestingly, 20% of the compatible claims have a con-

[5]The IBM Debating Technologies group in IBM Research has already released several data resources, found here: `https://www.research.ibm.com/haifa/dept/vst/mlta_data.shtml`. We aim to release the resource presented in this paper as well, as soon as we obtain the required licenses.

trastive relation with the topic target. Since contrastive targets flip polarity, stance classification would fail in these cases, unless these cases are correctly identified and accounted for. This highlights the importance of contrast classification for claim pro/con analysis. We discuss contrast detection in Section 7.

## 6 Target Extraction and Targeted Sentiment Analysis

Next, we describe an implementation of the stance classification model. This section provides a concise description of target identification and targeted sentiment analysis. The next section presents in more detail our novel contrast detection algorithm. We assume that for the user, directly specifying the topic target $x_t$ and the topic sentiment $s_t$ (e.g., *<boxing, Con>*) is as easy as phrasing the topic as a short sentence (*"This house would ban boxing"*), in terms of supervision effort. Therefore, we focus on finding $x_c$ and $s_c$, the claim target and sentiment, and assume that $x_t$ and $s_t$ are given.

### 6.1 Claim Target Identification

Previous work on *targeted/aspect-based* sentiment analysis focused on detecting in user reviews sentiment towards products and their components (Popescu and Etzioni, 2005; Hu and Liu, 2004b), or considered only named entities as targets (Mitchell et al., 2013). Here we address a more general problem of open domain, generic target identification. Table 1 illustrates the diversity and complexity of claim targets.

We set up the problem of claim target identification as a supervised learning problem, using an $L2$-regularized logistic regression classifier. Target candidates are the noun phrases in the claim, obtained from its syntactic parse[6]. We create one training example from each such candidate phrase $x$ and claim $c$ in our training set. The feature set is summarized in Table 2. Candidate phrases that exactly match the true target or overlap significantly with it are considered positive training examples, while the other candidates are considered negative examples. We measured overlap using the Jaccard similarity coefficient, defined as the ratio between the number of tokens in the intersection and the union of the two phrases, and considered an overlap of 0.6 or higher as significant overlap[7]. The candidate with the highest classifier confidence is predicted to be the target.

| |
|---|
| **Syntactic and Positional:** The dependency relation of $x$ in $c$; whether $x$ is a direct child of the root in the dependency parse tree for $c$; the minimum distance of $x$ from the start or the end of the chunk containing it. |
| **Wikipedia:** whether $x$ is a Wikipedia title, (e.g. *human rights*) |
| **Sentiment:** The dependency relation connecting $x$ to any sentiment phrase in the rest of $c$. The (Hu and Liu, 2004a) sentiment lexicon was used. For example, *Hereditary succession* is the sentiment target of *outdated*, indicated by the subject-predicate relation connecting them (Table 1, row 7). |
| **Topic relatedness:** Semantic similarity between $x$ and the topic target , e.g. *Marketing* and *advertising* (Table 1, row 1). We consider morphological similarity, paths in WordNet (Miller, 1995; Fellbaum, 1998), and cosine similarity of word2vec embeddings (Mikolov et al., 2013). |

Table 2: Features extracted for a target candidate $x$ in a claim $c$. Examples are taken from Table 1.

### 6.2 Claim Sentiment Classification

This component determines the sentiment of the claim towards its target. Given our open-domain setting, and the relatively small amount of training data available, we followed the common practice of lexicon-based sentiment analysis (Liu, 2012, pp. 50–53)[8]. Our method is similar to the one described by Ding et al. (2008), and comprises the following steps:

**Sentiment matching:** Positive and negative terms from the sentiment lexicon of Hu and Liu (2004a) are matched in the claim.

**Sentiment shifters application:** Sentiment shifters (Polanyi and Zaenen, 2004) reverse the polarity of sentiment words, and may belong to various parts of speech, e.g. "*not* successful$^+$", "*prevented* success$^+$", and "*lack* of success$^+$". We manually composed a small lexicon of about 160 sentiment shifters. The scope was defined as the $k$ tokens following the shifter word.[9]

**Sentiment weighting and score computation:** Following Ding et al., sentiment term weight decays based on its distance from the claim target. We used a weight of $d^{-0.5}$, where $d$ is the distance in tokens between the sentiment term and the target. Let $p$ and $n$ be the weighted sums of positive

[6] We used the ESG parser (McCord, 1990; McCord et al., 2012).

[7] Determined empirically based on the training set.

[8] Our sentiment analyzer was found to outperform the Stanford sentiment analyzer (Socher et al., 2013) on claims.

[9] We experimentally set $k = 8$ based on the training data.

and negative sentiments detected in the claim, respectively. The final sentiment score is then given by $\frac{p-n}{p+n+1}$, following Feldman et al. (2011).

## 7 Contrast Classification

The most challenging subtask in our model implementation is determining the contrast relation between the topic target $x_t$, and the claim target $x_c$. Previous work has focused on word-level contrast and synonym-antonym distinction (Mohammad et al., 2013; Yih et al., 2012; Scheible et al., 2013). The algorithm presented in this section addresses complex phrases, as well as consistent/contrastive semantic relations that go beyond synonyms/antonyms.

### 7.1 Algorithm

Consider the targets *atheism* and *denying the existence of God*. The relation between these targets is determined based on the contrastive relation between *God* and *atheism*, which is flipped by the negative polarity towards *God*, resulting in a consistent relation between the targets. We call the pair (*God*, *atheism*) the *anchor pair*, defined as the pair of core phrases that establishes the semantic link between the targets.

The following algorithm generalizes this notion, analogously to our claim-level model. The input for the algorithm includes $x_c$, $x_t$ and a relatedness measure $r(u, v) \in [-1, +1]$ over pairs of phrases $u$ and $v$. Positive/negative values of $r$ indicate a consistent/contrastive relation, respectively, and the absolute value indicates confidence.

First, anchor candidates are extracted from $x_c$ and $x_t$, as detailed in the next subsection. The anchor pair is selected based on the association strength of each anchor with the debate topic domain, as well as the strength of the semantic relation between the anchors. Term association with the domain is given by a TF-IDF measure $w(x) = tf(x)/df(x)$, where $tf(x)$ is the frequency of $x$ in articles that were identified as relevant to the topic in the labeled dataset, and $df(x)$ is its overall frequency in Wikipedia. We choose in $(x_c, x_t)$ the anchor pair $(a_c, a_t)$ that maximizes $w(u) \times |r(u, v)| \times w(v)$.

The contrast score is then predicted as $p(x_c, a_c) \times r(a_c, a_t) \times p(x_t, a_t)$, where $p(u, v) \in [-1, +1]$ is the polarity towards $v$ in $u$. Negative polarity is determined by the presence of words such as *limit, ban, restrict, deny* etc. We manually developed a small lexicon of stance flipping words, which largely overlaps with our sentiment shifters lexicon. We employ several relatedness measures, described in the next subsection, and the contrast scores obtained for these measures are used as features in the contrast classifier, implemented as a random forest classifier.

The above approach can be extended to find the top-K anchor pairs for complex targets. We use $K = 3$ in our experiments. When considering additional anchor pairs beyond the top-ranked pair $(a_c, a_t)$, we multiply the above contrast score by $sgn(r(b_c, b_t))$ for each such additional pair $(b_c, b_t)$. Thus, these pairs may affect the sign of the contrast score but not its magnitude. Anchor pair assignment is computed using the Hungarian Method (Kuhn, 1955).

### 7.2 Contrast Relations

We initially implemented the following known relatedness measures: (i) morphological similarity, (ii) cosine similarity using word2vec embeddings (Mikolov et al., 2013), (iii) reachability in WordNet via synonym-antonym chains (Harabagiu et al., 2006) and (iv) thesaurus-based synonym-antonym relations using polarity-inducing LSA (Yih et al., 2012). Note that the measures (i) and (ii) above take values only in $[0, 1]$, and thus are indicative of similarity but not of contrast. All these measures suffer from two limitations: (a) They only operate at the token level, while our anchors are often phrases (b) Their coverage on our data is insufficient, in particular for contrastive anchors.

We developed a novel relatedness measure that addresses these limitations, and is used in conjunction with the other measures. Our method is based on co-occurrence of the anchor pair with consistent and contrastive cue-phrases. For example, *"vs"*, *"or"* and *"against"* are contrastive cue phrases, while *"and"*, *"like"* and *"same as"* are consistent cue phrases. We compiled a list of 25 cue phrases.

The anchors are matched in a corpus we composed from the union of two complementary sources, which were found particularly effective for this task:

**Query logs:** We obtained 2.2 billion queries (450 million distinct queries) from the Blekko® search engine. With over a million distinct queries containing the words *vs*, *vs.*, or *versus*, it is an abundant resource for detecting contrast. Some

examples are: *"God **or** atheism"*, *"political correctness **vs** freedom of speech"*, *"free trade **vs** protectionism"* and *"advertising **and** marketing"*.

**Wikipedia headers:** We considered article titles, and section and subsection headers in Wikipedia (3 million in total). For example, *"Military intervention **vs** diplomatic solution"*.

Compared to full sentences, both queries and headers are short, concise texts, and therefore are less likely to suffer from contextual errors (in which the context alters the meaning of the matched pattern).

The score returned by our method is calculated as follows. Let $Lex^+$ and $Lex^-$ be the lexicons of consistent and contrastive cue phrases, respectively. Let $Freq(u, v)$ be the number of documents (queries or headers), which contain $u$ and $v$ separated by at most 3 tokens, and $Freq(u, Lex^+, v)$ is the size of the subset of these documents, which also contain a consistent cue phrase between $u$ and $v$. We then define the probability $P(Lex^+|u, v)$ as $\frac{Freq(u,Lex^+,v)}{Freq(u,v)}$. $P(Lex^-|u, v)$ is defined analogously for the contrastive lexicon. The returned score is $P(Lex^+|u, v)$ if $P(Lex^+|u, v) > P(Lex^-|u, v)$, and $-P(Lex^-|u, v)$ otherwise. We also experimented with other scoring methods, based on pointwise mutual information between the concurrences of the the pair ($u$,$v$) and the lexicon cue phrase, as well as statistical significance tests for their co-occurrence. However, the above method was found to perform best on our data.

**Generating anchor candidates:** Candidate anchors for measures (i)-(iv) are all single tokens. For our method, we additionally considered phrases as anchors. Candidates were generated from diverse sources, including the output of the ESG syntactic parser (McCord, 1990; McCord et al., 2012), the TagMe Wikifier (Ferragina and Scaiella, 2010), named entities recognized with the Stanford NER (Finkel et al., 2005) and multiword expressions in WordNet. Candidates subsumed by larger candidates were discarded. Following Levy et al. (2015), we kept only dominant terms with respect to the topic, by applying a statistical significance test (Hyper-geometric test with Bonferroni correction).

Overall, our method detects many consistent and contrastive pairs missed by previous methods.

### 7.3 Classification Output

The contrast classifier outputs a score in the $[0, 1]$ interval indicating the likelihood of $x_t$ and $x_c$ being consistent. We found that while it still cannot predict reliably contrastive targets, this consistency confidence score performs well on ranking the targets according to their likelihood of being consistent. We therefore use this score to re-rank our predictions, so that claims that are likely to have consistent targets would rank higher.

## 8 Evaluation

### 8.1 Experimental Setup

We evaluated the overall performance of the system, as well as the performance of individual components. The dataset was randomly split into a training set, comprising 25 topics (1,039 claims), and a test set, comprising 30 topics (1,355 claims). The training set was used to train the target identification classifier and the contrast classifier in our system, as well as the baselines described below.

We explore the trade-off between presenting high-accuracy predictions to the user, and making predictions for a large portion of the claims. This tradeoff is controlled by setting a threshold on the prediction confidence, and discarding predictions below that threshold. Let $\#claims$ be the total number of claims. Given some threshold $\alpha$, we define $\#predicted(\alpha)$ as the number of corresponding predictions, and $\#correct(\alpha)$ as the number of correct predictions. We then define: $coverage(\alpha) = \frac{\#predicted(\alpha)}{\#claims}$, and $accuracy(\alpha) = \frac{\#correct(\alpha)}{\#predicted(\alpha)}$.

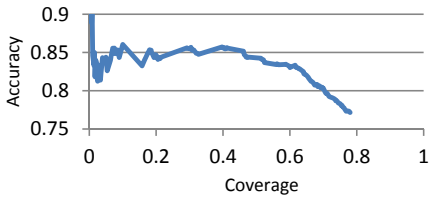
We consider the macro averaged $accuracy(\alpha)$ and $coverage(\alpha)$ over the test topics. Our evaluation focuses on the following question: suppose that we require a minimum coverage level, what is the highest accuracy we can obtain? The result is determined by an exhaustive search over threshold values. This assessment was performed for varying coverage levels.

The following configurations were evaluated. The first two configurations represent known strong baselines in stance classification (cf. Section 2).

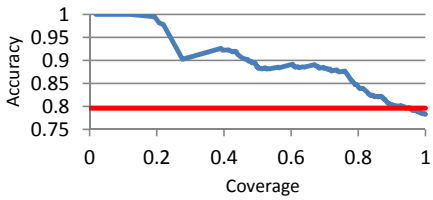
**Unigrams SVM:** SVM with unigram features. The SVM classifier gets the claim as an input, and aims to predict the claim sentiment $s_c$. Assuming consistent targets ($\mathcal{R}(x_c, x_t) = 1$), stance is then predicted as $s_c \times s_t$, where $s_t$ is the given topic

| Configuration | Accuracy@Coverage 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1.0 |
|---|---|---|---|---|---|---|---|---|---|---|
| Baselines | | | | | | | | | | |
| Unigrams SVM | 0.688 | 0.688 | 0.659 | 0.612 | 0.587 | 0.563 | 0.560 | 0.554 | 0.554 | 0.547 |
| Unigrams+Sentiment SVM | 0.717 | 0.717 | 0.717 | 0.709 | 0.693 | 0.691 | 0.687 | 0.668 | 0.655 | 0.632 |
| Our System | | | | | | | | | | |
| Sentiment Score | 0.752 | 0.720 | 0.720 | 0.720 | 0.720 | 0.720 | 0.636 | 0.636 | 0.636 | 0.636 |
| +Targeted Sentiment | 0.770 | 0.770 | 0.770 | 0.749 | 0.734 | 0.734 | **0.706** | 0.632 | 0.632 | 0.632 |
| +Contrast Detection | **0.849** | **0.847** | **0.836** | **0.793** | **0.767** | **0.740** | 0.704 | 0.632 | 0.632 | 0.632 |
| Our System+Unigrams SVM | 0.784 | 0.758 | 0.749 | 0.743 | 0.730 | 0.711 | 0.682 | **0.671** | **0.658** | **0.645** |

Table 3: Stance classification results. Majority baseline accuracy: 51.9%




(a) Sentiment (majority baseline: 56.2%)




(b) Contrast (majority baseline: 79.6%)

Figure 1: Performance of Sub-Components

sentiment.

**Unigrams+Sentiment SVM:** The unigram SVM with additional sentiment features. We employed here a simplified version of the sentiment analyzer (cf. Section 6.2), in which target identification is not performed, and sentiment terms are weighted uniformly. The following three features were used: the sums of positive and negative sentiments ($p$ and $n$), and the final sentiment score.

The next three configurations are incremental implementations of our system. For each configuration, only the difference from the previous configuration is specified.

**Sentiment Score:** Predicts $s_c$ as the sentiment score of the simplified sentiment analyzer. Stance is predicted as $s_c \times s_t$, similar to the SVM baselines.

**+Targeted Sentiment:** Employs the targeted sentiment analyzer described in Section 6.2.

**+Contrast Detection:** Full implementation of our model. Stance score is further multiplied by the output of the contrast classifier, $\mathcal{R}(x_c, x_t)$, predicted for the extracted claim target $x_c$ and the topic target $x_t$. As discussed in the previous section, this aims to rank higher claims with consistent targets.

Lastly, we tested a combination of our system with the unigrams SVM baseline.

**Our System+Unigrams SVM:** Adding the targeted sentiment score as a feature to the unigrams SVM. The SVM output is multiplied by the contrast classifier score.

For each configuration, if the classifier outputs zero[10], we predict the majority class in the train set with a constant, very low confidence.

### 8.2 Results, Analysis and Discussion

The results are shown in Table 3. Comparing the two baselines highlights the importance of sentiment in our open-domain setting, in which no topic-specific training data is available.

Using only the simple sentiment score outperforms the baselines for coverage rates $\leq 0.6$. For higher coverage rates the performance drops from 72% to 63.6%. This happens since the sentiment analyzer makes predictions for 69.4% of the claims, and the remaining claims are given the majority class with a fixed low confidence, as described above. For coverage rates $\geq 0.7$, these claims are added together (since they all match the same threshold), and thus accuracy is actually computed over the whole test set.

Targeted sentiment analysis improves over the non-weighted *Sentiment Score* baseline. It makes predictions for 77.4% of the claims[11], and similar to the previous configuration, accuracy drops accordingly from 70.6% to 63.2% for higher coverage rates ($\geq 0.8$).

Re-ranking based on target consistency confi-

[10]This can happen, for example, if the sentiment analyzer does not match any sentiment term in the claim.

[11]Coverage is improved since sentiment weighting breaks ties between positive and negative sentiments, which result in zero predictions of the simple analyzer.

dence substantially improves accuracy for lower coverage rates ($\leq 0.6$). For instance, the classifier achieves accuracy of 79.3% over 40% the claims, and 83.6% for 30% of the claims.

Finally, combining our system with the unigrams SVM allows the classifier to make predictions for claims that are not covered by the targeted sentiment analyzer, and consequently this configuration achieves the best accuracy for high coverage rates ($\geq 0.8$). It outperforms the SVM baselines for both low and high coverage rates.

Overall, the results confirm that our modular approach outperforms the common practice of monolithic classifiers for stance classification, in particular for making high-accuracy stance predictions for a large portion of the claims. Each component was shown to contribute to the overall performance.

We also assessed the performance for each subtask on the test set. Claim target identification achieves accuracy of 0.752 for exact matching, and 0.813 for relaxed matching (using the Jaccard measure, as in Section 6.1). Figure 1 shows accuracy vs. coverage curves for targeted claim sentiment analysis and contrast detection. Both components achieve higher accuracy for lower coverage rates, illustrating the effectiveness of their confidence score. As mentioned above, the sentiment analyzer makes a prediction for nearly 80% of the claims, and is shown to perform well. The contrast classifier, while not outperforming the majority baseline over the whole dataset, achieves accuracy that is much higher than the baseline for lower coverage rates.

## 9 Conclusion

This work is the first to address claim stance classification with respect to a given topic. We proposed a model that breaks down this complex task into simpler, well defined subtasks. Extensive data annotation and analysis has confirmed the applicability and accuracy of this reduction. The annotated dataset, which we plan to share with the community, is another contribution of this work.

The work also presented a concrete implementation of our model, using the collected labeled data to train each component, and demonstrated its effectiveness empirically. We plan to improve each of these components in future work.

## Acknowledgments

We would like to thank Yonatan Bilu, Ido Dagan, and Charles Jochim for their helpful feedback on this work.

## References

Ehud Aharoni, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund, and Noam Slonim. 2014. A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. In *Proceedings of the First Workshop on Argumentation Mining*, pages 64–68, Baltimore, Maryland, June. Association for Computational Linguistics.

Clinton Burfoot, Steven Bird, and Timothy Baldwin. 2011. Collective classification of congressional floor-debate transcripts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1506–1515, Portland, Oregon, USA, June. Association for Computational Linguistics.

Elena Cabrio and Serena Villata. 2013. A natural language bipolar argumentation approach to support users in online debate interactions. *Argument & Computation*, 4(3):209–230.

Xiaowen Ding, Bing Liu, and Philip S. Yu. 2008. A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, pages 231–240, New York, NY, USA. ACM.

Adam Faulkner. 2014. Automated classification of stance in student essays: An approach using stance target information and the Wikipedia link-based measure. In *Proceedings of the Twenty-Seventh International Florida Artificial Intelligence Research Society Conference, FLAIRS*.

Ronen Feldman, Benjamin Rosenfeld, Roy Bar-Haim, and Moshe Fresko. 2011. The Stock Sonar - sentiment analysis of stocks based on a hybrid approach. In *Innovative Applications of Artificial Intelligence (IAAI-11)*.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge, MA.

Paolo Ferragina and Ugo Scaiella. 2010. Tagme: On-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 1625–1628, New York, NY, USA. ACM.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs

sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 363–370, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Sanda Harabagiu, Andrew Hickl, and Finley Lacatusu. 2006. Negation, contrast and contradiction in text processing. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1*, AAAI'06, pages 755–762. AAAI Press.

Kazi Saidul Hasan and Vincent Ng. 2013. Stance classification of ideological debates: Data, models, features, and constraints. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1348–1356, Nagoya, Japan, October. Asian Federation of Natural Language Processing.

Minqing Hu and Bing Liu. 2004a. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 168–177, New York, NY, USA. ACM.

Minqing Hu and Bing Liu. 2004b. Mining opinion features in customer reviews. In *Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI-2004)*.

H. Kuhn. 1955. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(3).

Namhee Kwon, Liang Zhou, Eduard Hovy, and Stuart W. Shulman. 2007. Identifying and classifying subjective claims. In *Proceedings of the 8th Annual International Conference on Digital Government Research: Bridging Disciplines & Domains*, dg.o '07, pages 76–81. Digital Government Society of North America.

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context dependent claim detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1489–1500, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

Ran Levy, Liat Ein-Dor, Shay Hummel, Ruty Rinott, and Noam Slonim. 2015. Tr9856: A multi-word term relatedness benchmark. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 419–424, Beijing, China, July. Association for Computational Linguistics.

Marco Lippi and Paolo Torroni. 2015. Context-independent claim detection for argument mining. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, pages 185–191.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Morgan & Claypool Publishers*.

Michael C McCord, J William Murdock, and Branimir K Boguraev. 2012. Deep parsing in watson. *IBM Journal of Research and Development*, 56(3/4):3.1–3.15.

Michael C. McCord. 1990. Slot grammar: A system for simpler construction of practical natural language grammars. In R. Studer, editor, *Natural Language and Logic: Proc. of the International Scientific Symposium, Hamburg, FRG*, pages 118–145. Springer, Berlin, Heidelberg.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, NIPS'13, pages 3111–3119, USA. Curran Associates Inc.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, November.

Margaret Mitchell, Jacqui Aguilar, Theresa Wilson, and Benjamin Van Durme. 2013. Open domain targeted sentiment. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1643–1654, Seattle, Washington, USA, October. Association for Computational Linguistics.

Saif M. Mohammad, Bonnie J. Dorr, Graeme Hirst, and Peter D. Turney. 2013. Computing lexical contrast. *Computational Linguistics*, 39(3):555–590.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California, June. Association for Computational Linguistics.

Livia Polanyi and Annie Zaenen. 2004. Contextual valence shifters. In *Working Notes — Exploring Attitude and Affect in Text: Theories and Applications (AAAI Spring Symposium Series)*.

Ana-Maria Popescu and Oren Etzioni. 2005. Extracting product features and opinions from reviews. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 339–346, Vancouver, British Columbia, Canada, October. Association for Computational Linguistics.

Silke Scheible, Sabine Schulte im Walde, and Sylvia Springorum. 2013. Uncovering distributional differences between synonyms and antonyms in a word

space model. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 489–497, Nagoya, Japan, October. Asian Federation of Natural Language Processing.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, October. Association for Computational Linguistics.

Swapna Somasundaran and Janyce Wiebe. 2009. Recognizing stances in online debates. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 226–234, Suntec, Singapore, August. Association for Computational Linguistics.

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, Los Angeles, CA, June. Association for Computational Linguistics.

Swapna Somasundaran, Galileo Namata, Janyce Wiebe, and Lise Getoor. 2009. Supervised and unsupervised methods in employing discourse relations for improving opinion polarity classification. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 170–179, Singapore, August. Association for Computational Linguistics.

Dhanya Sridhar, Lise Getoor, and Marilyn Walker. 2014. Collective stance classification of posts in online debate forums. In *Proceedings of the Joint Workshop on Social Dynamics and Personal Attributes in Social Media*, pages 109–117, Baltimore, Maryland, June. Association for Computational Linguistics.

Matt Thomas, Bo Pang, and Lillian Lee. 2006. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 327–335, Sydney, Australia, July. Association for Computational Linguistics.

Marilyn Walker, Pranav Anand, Rob Abbott, and Ricky Grant. 2012a. 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, Montréal, Canada, June. Association for Computational Linguistics.

Marilyn A. Walker, Pranav Anand, Rob Abbott, Jean E. Fox Tree, Craig Martell, and Joseph King. 2012b. That is your evidence?: Classifying stance in online political debate. *Decis. Support Syst.*, 53(4):719–729, November.

Ainur Yessenalina, Yisong Yue, and Claire Cardie. 2010. Multi-level structured models for document-level sentiment classification. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1046–1056, Cambridge, MA, October. Association for Computational Linguistics.

Wen-tau Yih, Geoffrey Zweig, and John Platt. 2012. Polarity inducing latent semantic analysis. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1212–1222, Jeju Island, Korea, July. Association for Computational Linguistics.