Stance Classification of Context-Dependent Claims

Roy Bar-Haim, Indrajit Bhattacharya,
Francesco Dinuzzo
Amrita Saha, and Noam Slonim

Stance Detection

Given a topic question and a claim extracted from some document, identify whether the claim supports or contrasts the topic.

Topic question => can be a question/ a prompt/ a statement

Claim => single sentence claim/ an article/ an essay

Stance Detection

Question: **Should More Gun Control Laws Be Enacted?**

Claim 1: **More gun control laws would reduce gun deaths. (Pro)**

Claim 2: **Gun control laws infringe upon the right to self defense and deny people a sense of safety. (Con)**

Claim 3: **The second amendment is not an unlimited right to own guns. (Pro)**

Stance Detection

Claim: **Should Abortion Be Legal?**

Article: The US Supreme Court has declared abortion to be a "fundamental right" guaranteed by the US Constitution. The landmark abortion case Roe v. Wade, decided on Jan. 22, 1973 in favor of abortion rights, remains the law of the land. The 7-2 decision stated that the Constitution gives "a guarantee of certain areas or zones of privacy," and that "This right of privacy... is broad enough to encompass a woman's decision whether or not to terminate her pregnancy.

Challenges

Challenges focused in this paper =>

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

More general challenges =>

- Understanding the relationship between target in topic and target in claim (abortion => killing of a human being, Barack Obama => B. Obama, Animals => dogs)
- Claim/ Article can contain subjective and objective texts
- Articles can contain positive/ negative sentiment towards the same target
- Categorizing an article as "related" to the topic

Stance Detection: An exceptional case

Topic: The sale of violent video games to minors should be banned.

Claim: Parents, not government bureaucrats, have the right to decide what is appropriate for their children.

Issue: no clear sentiment target in the claim that is either consistent or contrastive with the sale of violent video games to minors.

Author’s claim: Their 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 3 subtasks

Terminologies

Topic sentence => t   Claim => c

c includes a claim target x_c, defined as a phrase about which c makes a positive or a negative assertion.

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, for a topic t, the topic target x_t and topic sentiment s_t.

Terminologies

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 xc implies the opposite stance towards x_t .

The contrast relation between x_c and x_t, denoted R(xc, xt) ϵ {−1, 1} is 1 if x_c and x_t are consistent, and −1 if they are contrastive.

$$\text{Stance}(c, t) = s_c * R(x_c, x_t) * s_t$$

#	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.

Dataset Annotation

Labels are collected automatically.

One of the authors annotated 55 topics with topic target and topic sentiment.

5 annotators for doing the same in claims. First identify x_c, s_c, then $R(x_c, x_t)$

First, overlapping claim targets were clustered together. If no cluster contained the majority of the annotations, then the claim was labeled as incompatible with the model.

Dataset Annotation

- Majority cluster was found for 98.5% of the claims
- 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.
- Combining the labels for s_c, $R(x_c, x_t)$ and s_t correctly predicted the Pro/Con labels in the dataset for 99.6% of the compatible claims.
- 20% of the compatible claims have a contrastive relation with the topic target.

Open domain, Generic Target Identification

- A supervised learning problem, using an L2-regularized logistic regression classifier.
- Target candidates => noun phrases in the claim
- One training example from each such candidate phrase x and claim c in training set
- 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 (Jaccard similarity coefficient, threshold >= 0.6, determined empirically based on the training set)

Feature Set

Feature	Explanation
Syntactic & Positional	The dependency relation of x in c
Wikipedia	Whether the target is a Wiki title
Sentiment	The dependency relation connecting x to any sentiment phrase in the rest of c.
Relatedness	Between topic target & candidate target morphological similarity paths in WordNet and cosine similarity of word2vec embeddings

Claim Sentiment Classification

Lexicon-based sentiment analysis

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

Sentiment shifters application: a small lexicon of about 160 sentiment shifters. (Ex: not successful, prevented success, lack of success)The scope was defined as the k tokens following the shifter word

Claim Sentiment Classification

Sentiment weighting and score computation: sentiment term weight decays based on its distance from the claim target.

$$\text{Weight} = d \wedge -0.5$$

d is the distance in tokens between the sentiment term and the target.

p and n be the weighted sums of positive and negative sentiments detected in the claim, respectively.

$$\text{final sentiment score} = (p-n) / (p+n+1)$$

following Feldman et al. (2011).

Contrast Classification

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.

(God, atheism) => the anchor pair, defined as the pair of core phrases that establishes the semantic link between the targets.

Contrast Classification (Algorithm)

Input for the algorithm => 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.

Algorithm contd.

First, anchor candidates are extracted from x_c and x_t.

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

$$w(x) = tf(x) / df(x), \text{ where}$$

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 $w(u) * |r(u, v)| * w(v)$.

Algorithm contd.

$$\text{contrast score} = p(x_c, a_c) * r(a_c, a_t) * p(x_t, a_t)$$

where $p(u, v) \in [-1, +1]$ is the polarity towards v in u.

=> a small lexicon of stance flipping words (overlap with sentiment shifter words)

=>Contrast scores obtained for 5 relatedness measures are used as features in the contrast classifier (random forest classifier).

Generating anchor candidates

Besides single word tokens, considered phrases as anchor candidates

Candidates were generated from diverse sources,

- the output of the ESG syntactic parser (McCord, 1990; McCordet 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

Keep only dominant terms with respect to the topic, by applying a statistical significance test (Hyper-geometric test with Bonferroni correction).

Contrast Relations

Relatedness measures:

(i) morphological similarity (reliable for similarity only)

(ii) cosine similarity using word2vec embeddings (Mikolov et al., 2013) (reliable for similarity only)

(iii) reachability in Word- Net via synonym-antonym chains (Harabagiu et al., 2006) and

(iv) thesaurus-based synonym-antonym relations using polarity-inducing LSA (Yih et al., 2012)

=> works well at token level, not at phrase level

=> does not cover the training set well

Contrast Relations

A novel relatedness measure => consistent and contrastive cue-phrases (a list of 25 cue phrases)

Anchors are matched against 2 corpuses

- Query Logs => (450 million distinct queries) from the Blekko search engine (With over a million distinct queries containing the words vs, vs., or versus)
  - Ex: “God or atheism”, “political correctness vs freedom of speech”, “free trade vs protectionism” and “advertising and marketing”.
- Wikipedia headers => article titles, and section and subsection headers in
- Wikipedia (3 million in total).
  - Ex: “Military intervention vs diplomatic solution”.

Contrast Relations

Prob(Lex+ | u,v) = Freq(u, Lex+, v) / Freq(u,v)

Freq(u, v) => the number of documents (queries or headers), which contain u and v separated by at most 3 tokens

Score = Prob(Lex+ | u,v) if Prob(Lex+ | u,v) > Prob(Lex- | u,v) else -Prob(Lex- | u,v)

Experimental Setup

training set, comprising 25 topics (1,039 claims)

test set, comprising 30 topics (1,355 claims).

=> training set was used to train the target identification classifier and the contrast classifier as well as the baselines

=> A trade-off between presenting high-accuracy predictions to the user, and making

predictions for a large portion of the claims.

Coverage

- Threshold on the prediction confidence α
- #claims be the total number of claims.
- #predicted(α) as the number of corresponding predictions
- #correct(α) as the number of correct predictions

coverage(α) = #predicted(α) / #claims , and

accuracy() = #correct(α) / #predicted(α)

Baselines

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 ($R(x_c, x_t) = 1$), stance is then predicted as $s_c * s_t$

where st is the given topic target.

Baselines

Unigrams+Sentiment SVM: The unigram SVM with additional sentiment features.

A simplified sentiment analyzer:

- No target identification
- sentiment terms are weighted uniformly.
- Features: the sums of positive and negative sentiments (p and n), and the final sentiment score

Performance

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%

Performance

If the classifier outputs zero, predict the majority class in the train set with a constant, very low confidence.

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.

Claim target identification achieves accuracy of 0.752 for exact matching, and 0.813 for relaxed matching (using the Jaccard measure).

THANK YOU