Sentiment-Stance-Specificity (SSS) Dataset: Identifying Support-based Entailment among Opinions.

Pavithra Rajendran, Danushka Bollegala, Simon Parsons
University of Liverpool, University of Liverpool, Kings College London
Pavithra.Rajendran@liverpool.ac.uk, danushka.bollegala@liverpool.ac.uk, simon.parsons@kcl.ac.uk

Abstract

Computational argumentation aims to model arguments as a set of premises that either support each other or collectively support a conclusion. We prepare three datasets of text-hypothesis pairs with support-based entailment based on opinions present in hotel reviews using a distant supervision approach. Support-based entailment is defined as the existence of a specific opinion (premise) that supports as well as entails a more general opinion and where these together support a generalised conclusion. A set of rules is proposed based on three different components — *sentiment*, *stance* and *specificity* to automatically predict support-based entailment. Two annotators manually annotated the relations among text-hypothesis pairs with an inter-rater agreement of 0.80. We compare the performance of the rules which gave an overall accuracy of 0.83. Further, we compare the performance of textual entailment under various conditions. The overall accuracy was 89.54%, 90.00% and 96.19% for our three datasets.

Keywords: argument mining, stance classification, structured argumentation

1. Introduction

Argument mining (Abbas and Sawamura, 2008; Palau and Moens, 2009) deals with the extraction of argument components and structures from natural language texts. It has drawn attention from both the argumentation and NLP communities with the introduction of ArgMining workshops [1]. In computational argumentation, an argument can be defined as a collection of premises together (linked argument) or individually (convergent argument) which are related to a conclusion (Palau and Moens, 2009). Each premise provides a *support* in the form of logical reasoning for, or evidence in support of, the conclusion to which it is connected.

It has been suggested that, in natural language texts, this support relation can be interpreted as meaning either (a) one premise is inferred from another premise (Janier et al., 2014) or (b) one premise provides evidence that supports another premise (Park and Cardie, 2014). In either case, it is natural to interpret the relationship as a form of entailment.

In this paper, we consider a subtype of entailment, which we term *support-based entailment*, where a support relation exists between the text and the hypothesis. Despite the unstructured nature of natural language texts, they provide meta-linguistic attributes such as stance, sentiment, and specificity that can be exploited for detecting support-based entailment.

We create a dataset of text-hypothesis pairs from opinions collected from a set of hotel reviews where the text provides support to the corresponding hypothesis. As an example, we consider online reviews that are comprised largely of opinionated texts that talk about various aspects of a product or service. Consider the examples shown in Fig. 1 where we have two different reviews with overall star ratings of 1 and 2. There we see a collection of *opinions*, that is sentence-level statements that talk about one or more aspects of a product or service. These are the basic units that we deal with in our work. Human annotation of argument structures and relation among them is a complicated task which is domain-dependent and hence manually annotating huge data is costly and difficult (Matthias and Stein, 2016). To combat this, we use a distant supervision approach by manually creating a set of rules based on meta-linguistic attributes such as *stance*, *sentiment* and *specificity*. These rules automatically label a set of sentences, which is then used to train a classifier for predicting support-based entailment.

Now, we give a few examples to explain how the three meta-linguistic attributes are useful. First, let us consider two opinions with the same sentiment as follows:

"not good enough for a Hotel charging these prices"

"the problem with the hotel is the staff"

Both these opinions have the same negative sentiment, but there exists no support or entailment relation between them. Suppose, we consider two opinions with the same stance (here, we refer stance as the standpoint taken towards a particular topic) as follows:

"the staff were helpful and polite"

"the staff was great"

Both these opinions have the same sentiment and stance towards the aspect *staff*. The only difference is that, in the first opinion the stance does not contain the stance expressed linguistically, whereas it is expressed in the second. Rajendran et al. (2017) use a supervised approach to classify opinions as implicit/explicit based on how the stance is expressed linguistically. This classification can help to relate the opinions as: the first opinion (text) supports as well as entails the second opinion (hypothesis).

Suppose, we consider two opinions as follows:

"the staff was great"

[1] https://argmining2017.wordpress.com/

"overall, great service!"

While these two opinions have the same positive sentiment and both are explicit opinions, the first opinion has a stance towards the aspect *staff* and the second opinion has a stance towards the aspect *service*. In such cases, sentiment and stance alone would be insufficient. If we were given a knowledge base that can relate *staff* with *service*, then, it can be useful to relate these opinions as: first opinion (text) supports as well as entails the second opinion (hypothesis). The remaining sections are given below.

- Section 2. gives an overview of the related works.
- Section 3. gives a description about the support-based entailment relation and the three meta-linguistic attributes – *sentiment*, *stance* and *specificity* that are useful to predict the same.
- Section 4. gives a description about the different support-based entailment rules (SER) that are proposed to predict the support-based entailment relation.
- Section 5. describes the SSS dataset that we create using opinions extracted a set of hotel reviews and SER.
- Section 6. describes the experiments carried out on the SSS dataset using existing textual entailment methods and the results are reported.
- Section 7. presents the conclusion of our work.

2. Related Work

A detailed study of previous work in argument mining has been presented by Lippi and Torroni (Lippi and Torroni, 2015; Lippi and Torroni, 2016). Few papers have dealt with the problem of mining arguments from online reviews (Wyner et al., 2012; Villalba and Saint-Dizier, 2012). Using computational argumentation techniques to deal with real-world problems such as opinion mining (Dragoni et al., 2016), sentiment analysis (Rajendran et al., 2016) and detecting deceptive reviews (Cocarascu and Toni, 2016) has been tackled so far. Boltuzic et al. (2014) combine stance, textual entailment and semantic similarity to identify relations between arguments and comments presented in a debate. We propose a way of detecting support-based entailment such that a specific opinion *supports* as well as *entails* a corresponding generalised opinion. Cabrio and Villata (2012) consider entailment to be a form of support relation that occurs between arguments present in debates. We differ from this work, since we define a support relation based on structured argumentation using three different components – sentiment, stance and specificity that can also predict entailment. Previously, Grosse et al. (2012) have explored constructing opinion analysis trees that aggregate opinions present in a Twitter dataset based on the specificity property. Our work is not to aggregate opinions but to construct argument structures that are able to persuade an audience towards a particular conclusion.

Stance classification (Mohammad et al., 2016; Augenstein et al., 2016; Anand et al., 2011) relates to classifying whether a given statement is for or against a known target, which is explicitly stated or not. Sobhani et al. (2016) investigate the relation between stance and sentiment on a set of Twitter data where the target need not be present explicitly. Ebrahimi et al. (2016) propose a model that integrates stance, sentiment and target features jointly as a three way interaction for classifying stance in a set of tweets. We use sentiment as a way of identifying stance present in opinions where the target is explicitly present. Also, we are interested in how the stance is expressed and use this as a feature to identify support-based entailment relation.

Textual entailment deals with identifying whether a hypothesis can be inferred from a given text, which is directional and differs from semantic similarity measures. Yokote et al. (2012) propose a model that transforms similarity measures into a non-linear transformation for predicting textual entailment. Zanzotto et al. (2005) investigate on identifying patterns based on subject verb relation to identifying entailment. In their paper, they argue that the logical entailment present between the text and hypothesis is not captured properly. In contrast, we are interested in a subtype of entailment that can predict the *support* relation based on argumentation theory.

3. Support-based Entailment

The three components of the proposed method are explained below. Based on these, we manually identify a set of support-based entailment rules (SER) for predicting the support-based entailment between a text (T) and a hypothesis (H).

Opinion and Premise: We take an *opinion* to be a sentence-level statement, which might be either positive or negative in sentiment, and talks about an aspect or several aspects of a product/service. For example, *service*, *location* are aspects of hotels in the hotel domain.

We consider a *premise* as a simple atomic unit that talks about one particular aspect. Hence, any opinion that talks about several aspects can be considered as a collection of several premises that may or may not be related.

Sentiment: The positive/negative sentiment of an opinion is taken into consideration. We ignore objective opinions as it cannot be used to match the global sentiment (overall star rating). As a first step we only consider TH pairs as opinions with the same sentiment.

Stance: Previously (Rajendran et al., 2017), we explained how to classify the stance expressed by an opinion as implicit/explicit. In both, the stance (for/against) is expressed by the reviewer, but explicit opinions have the stance explicitly expressed using (1) direct approval/disapproval or (2) words/phrases by the reviewer that have a stronger intensity of expression with respect to the topic in discussion. General cues such as *recommend*, *great*, *worst* indicate direct expressions and are useful in identifying explicit opinions. Specific cues that are related to domain-based targets can help in identifying implicit opinions. For

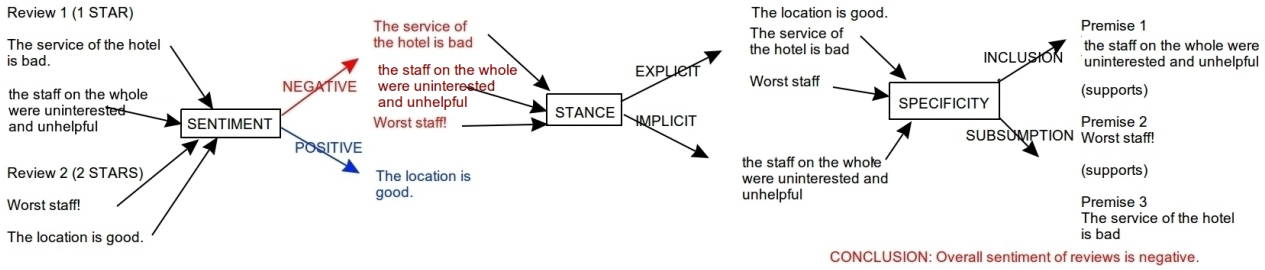

Figure 1: Opinions from two reviews are extracted and distinguished based on their local sentiment, stance, and specificity. All opinions that do not match the overall sentiment of the reviews are discarded. The rest of the opinions are then classified as explicit or implicit and using subsumption and inclusion relation, these opinions are combined such that one supports another.

example, *lightweight laptop* has a positive stance towards the target *laptop* whereas *the storyline of the book is lightweight* has a negative stance towards the target *book*. Also, opinions can express justification such as reasons that express stance implicitly. An example is provided in Fig. 1.

Specificity: A knowledge base (KB) is created based on the domain and the aspects present where one aspect is a sub-class of the other. Given such a KB, we describe three domain-based ontology relations between two premises that make use of the implicit/explicit nature of the opinions in which the aspects are present.

Suppose an aspect is present in a given opinion, we consider the opinion to contain a premise about that particular aspect. We thus represent each such premise as $\mathcal{P}(attr, op, stance)$ where $attr$ is the *aspect* present in an opinion *Op* which is classified as implicit/explicit and represented as *Stance*. We define the three relations below.

Def. 1 (Subsumption, $\sqsubseteq_{sub}$). Two premises present within an opinion, $\mathcal{P}(attr1, op1, exp) \sqsubseteq_{intrasub}$ (intra-subsumption) $\mathcal{P}(attr2, op1, exp)$ if *attr1* is a sub-class of *attr2*.
Two premises present in two different opinions, $\mathcal{P}(attr1, op1, exp) \sqsubseteq_{intersub}$ (inter-subsumption) $\mathcal{P}(attr2, op2, exp)$ if *attr1* is a sub-class of *attr2*.

Def. 2 (Inclusion, $\sqsubseteq_{inc}$). Two premises, one present in an implicit opinion and the other present present in an explicit opinion satisfies $\mathcal{P}(attr1, op1, imp) \sqsubseteq_{inc}$ (is-inclusive of) $\mathcal{P}(attr2, op2, imp)$ such that *attr1* and *attr2* are the same.

Def. 3 (Equivalence, $\equiv$). $\mathcal{P}(attr1, op1, exp) \equiv$ (equivalent) $\mathcal{P}(attr2, op2, exp)$ if *attr1* and *attr2* are same. $\mathcal{P}(attr1, op1, imp) \equiv$ (equivalent) $\mathcal{P}(attr2, op2, imp)$ if *attr1* and *attr2* are same.

4. Support-based Entailment Rules (SER)

Our definition of a premise states that an opinion with *n* aspects contains *n* premises. For example,

*"and the **service** from the **staff** was extremely poor"*

contains two premises, one about the *service* and the other about the *staff*.
We are not interested in decomposing the opinion into different premises based on the linguistic structure but instead focus on identifying text-hypothesis (TH) pairs. Our motivation behind creating the dataset is to identify TH pairs that can help in forming argument structures from these premises using implicit and explicit opinions. A simple structure would be of the form $(implicit_1, explicit_1, explicit_2)$ with different relations as follows:

- Inclusion relation between a premise present in $implicit_1$ and a premise in $explicit_1$. Both premises are about the same aspect.
- Intra-subsumption relation between two different premises present within $explicit_1$. The same can be said for $explicit_2$.
- Inter-Subsumption/Equivalence relation between a premise in $explicit_1$ and a premise in $explicit_2$.

All these relations require two premises. For every opinion (text or hypothesis), our rules are designed to consider atmost two premises at a time and whether those two premises are related or not. For example,

Op 1: *the **hotel** was exceptionally clean, the* service *was very friendly at all times and nothing seemed to be too much and the* location *is quiet and peaceful...*

Op 2: *this is very nice **hotel** that exceeded our expectations*

Op1 contains three premises $\mathcal{P}(hotel, Op1, imp)$, $\mathcal{P}(service, Op1, imp)$ and $\mathcal{P}(location, Op2, imp)$. Op2 contains one premise $\mathcal{P}(hotel, Op2, exp)$.
In the above example, we can consider atmost two premises at a time, which means we have the following premise pairs:-

- $(\mathcal{P}(hotel, Op1, imp), \mathcal{P}(service, Op1, imp))$
- $(\mathcal{P}(hotel, Op1, imp), \mathcal{P}(location, Op2, imp))$
- $(\mathcal{P}(service, Op1, imp), \mathcal{P}(location, Op2, imp))$

Rule	# Aspects (Text)	#Aspects (Hypothesis)	Text	Hypothesis	Relation
Rule 1	>1	>1	*a* $\sqsubseteq_{intrasub}$ *b*	*c* $\sqsubseteq_{intrasub}$ *d*	*b* $\sqsubseteq_{intersub}$ *d* or *b* $\equiv$ *d* and *a* $\sqsubseteq_{intersub}$ *c* or *a* $\equiv$ *c*
Rule 2	>1	1	*a* $\sqsubseteq_{intrasub}$ *b*	*c*	*b* $\sqsubseteq_{intersub}$ *c* or *b* $\equiv$ *c*
Rule 3	>1	1	*a,b* and not related	*c*	*a* $\sqsubseteq_{intersub}$ *c* and *b* $\sqsubseteq_{intersub}$ *c*
Rule 4	>1	1	*a,b* and not related	*c*	*a* $\equiv$ *c* or *b* $\equiv$ *c*
Rule 5	1	1	*a*	*c*	*a* $\sqsubseteq_{intersub}$ *c*
Rule 6	1	1	*a*	*c*	*a* $\equiv$ *c*
Rule 1	1	1	*a*	*c*	*a* $\sqsubseteq_{inc}$ *b*
Rule 2	1	>1	*a*	*b* $\sqsubseteq_{intrasub}$ *c*	*a* $\sqsubseteq_{inc}$ *b*
Rule 3	>1	>1	*a,b* and not related	*c* $\sqsubseteq_{intrasub}$ *d*	*a* $\sqsubseteq_{inc}$ *c* and *b* $\sqsubseteq_{inc}$ *d*
Rule 4	>1	1	*a,b* and not related	*c*	*a* $\sqsubseteq_{inc}$ *c* or *b* $\sqsubseteq_{inc}$ *c*
Rule 5	1	>1	*a*	*b,c* and not related	*a* $\sqsubseteq_{inc}$ *b* or *a* $\sqsubseteq_{inc}$ *c*
Rule 6	>1	>1	*a,b* and not related	*c,d* and not related	*a* $\sqsubseteq_{inc}$ *c* or *b* $\sqsubseteq_{inc}$ *d*

Table 1: Each proposed rule for subsumption (top) and inclusion (bottom) relation is presented. The number of aspects (premises) that must be present in text and hypothesis is given. Conditions that must hold true in text, hypothesis and between them is also given. Here, we consider *a*,*b*,*c* and *d* to represent the aspects (premises) present.

Rule	Text	Hypothesis	Relation
Rule 1	and the **service** from the **staff** was extremely poor (*staff*$_{text}$ $\sqsubseteq_{intrasub}$ *service*$_{text}$)	it is the worst **service** i have seen in a five star **hotel** (*service*$_{hyp}$ $\sqsubseteq_{intrasub}$ *hotel*$_{hyp}$)	*service*$_{text}$ $\sqsubseteq_{intersub}$ *hotel*$_{hyp}$, *staff*$_{text}$ $\sqsubseteq_{intersub}$ *service*$_{hyp}$, *service*$_{text}$ $\equiv$ *service*$_{hyp}$
Rule 2	**location** of the **hotel** is really well placed - you're in the middle of everything (*location*$_{text}$ $\sqsubseteq_{intrasub}$ *hotel*$_{text}$)	overall a very good **hotel** (*hotel*$_{hyp}$)	*hotel*$_{text}$ $\equiv$ *hotel*$_{hyp}$
Rule 3	weak **service** for very high **prices** (*service*$_{text}$, *prices*$_{text}$)	i would not plan to stay at this **hotel** again (*hotel*$_{hyp}$)	*service*$_{text}$ $\sqsubseteq_{intersub}$ *hotel*$_{hyp}$, *prices*$_{text}$ $\sqsubseteq_{intersub}$ *hotel*$_{hyp}$
Rule 4	weak **service** for very high **prices** (*service*$_{text}$, *prices*$_{text}$)	however this is probably the worst **service** we have ever experienced (*service*$_{hyp}$)	*service*$_{text}$ $\equiv$ *service*$_{hyp}$
Rule 5	great **location** (*location*$_{text}$)	i absolutely loved this **hotel** (*hotel*$_{hyp}$)	*location*$_{text}$ $\sqsubseteq_{intersub}$ *hotel*$_{hyp}$
Rule 6	i absolutely loved this **hotel** (*hotel*$_{text}$)	overall a very good **hotel** (*hotel*$_{hyp}$)	*hotel*$_{text}$ $\sqsubseteq_{intersub}$ *hotel*$_{hyp}$
Rule 1	**hotel** infrastructure is in need of serious upgrading (*hotel*$_{text}$)	so believe me when i say do not stay at this **hotel** (*hotel*$_{hyp}$)	*hotel*$_{text}$ $\sqsubseteq_{inc}$ *hotel*$_{hyp}$
Rule 2	the **staff** that we encountered were very friendly and helpful (*staff*$_{text}$)	and the **service** from the valet and front desk **staff** is very good (*staff*$_{hyp}$ $\sqsubseteq_{intrasub}$ *service*$_{hyp}$)	*staff*$_{text}$ $\sqsubseteq_{inc}$ *staff*$_{hyp}$
Rule 4	to their credit the management was more responsive and very apologetic for the condition of my **room** and the rude treatment by their **staff** (*room*$_{text}$, *staff*$_{text}$)	dissapointed from the **room** (*room*$_{hyp}$)	*room*$_{text}$ $\sqsubseteq_{inc}$ *room*$_{hyp}$
Rule 5	the **staff** was not friendly nor helpful (*staff*$_{text}$)	overall its a dark dated **hotel** let down badly by the unhelpful and rude **staff** (*hotel*$_{text}$, *staff*$_{hyp}$)	*staff*$_{text}$ $\sqsubseteq_{inc}$ *staff*$_{hyp}$

Table 2: Examples for different rules satisfying subsumption (top) and inclusion (bottom) relations.

- ($\mathcal{P}(hotel, Op2, exp)$,$\mathcal{P}(hotel, Op1, imp)$)
- ($\mathcal{P}(hotel, Op2, exp)$,$\mathcal{P}(service, Op1, imp)$)
- ($\mathcal{P}(hotel, Op2, exp)$,$\mathcal{P}(location, Op2, imp)$)

Out of these, ($\mathcal{P}(hotel, Op2, exp)$, $\mathcal{P}(hotel, Op1, imp)$) is related by the inter-subsumption relation. Further, if an opinion contains more than one premise, then rules based on a single premise cannot be considered. In the above example, Op2 can be considered for rules based on a single premise whereas Op1 cannot be considered.

Let us consider another case where a text that contains 3 premises *a,b* and *c* with *a* and *b* related. For a given hypothesis, one rule will be satisfied based on the related premises *a* and *b* while some other rule might be satisfied based on two premises that are not related (eg. *a* and *c*). We predict the support-based entailment in a TH pair if at least one of the rules is satisfied. This is to ensure that there are no

Data	Rev	Exp	Imp	Sub	Inc
FA	369	264	720	Rule 1: 14	Rule 1: 271
				Rule 2: 138	Rule 2: 25
				Rule 3: 27	Rule 3: 6
				Rule 4: 218	Rule 4: 619
				Rule 5: 193	Rule 5: 147
				Rule 6: 218	Rule 6: 344
SA	707	1001	4359	Rule 1: 92	Rule 1: 1790
				Rule 2: 566	Rule 2: 137
				Rule 3: 82	Rule 3: 55
				Rule 4: 344	Rule 4: 3418
				Rule 5: 842	Rule 5: 933
				Rule 6: 1834	Rule 6: 1799
UA	3271	564	5933	Rule 1: 34	Rule 1: 3708
				Rule 2: 467	Rule 2: 148
				Rule 3: 55	Rule 3: 33
				Rule 4: 119	Rule 4: 4726
				Rule 5: 428	Rule 5: 2189
				Rule 6: 1354	Rule 6: 3053

Table 3: In each dataset: total number of reviews (Rev) present, total number of explicit opinions (Exp) and implicit opinions (Imp) found and total number of TH pairs satisfying each rule in SER based on subsumption (Sub) and inclusive (Inc) relation is present.

duplicate pairs created.

If a text/hypothesis can contain a single premise or atmost two premises, then nine different combinations are possible based on whether inter-subsumption is present in the text/hypothesis or not. This holds for both subsumption-based and inclusion-based rules. Based on our definition of *support-based entailment*, a specific premise supports a more generalised premise. Thus, we ignore rules based on subsumption relation that look into hypothesis containing non-related premises. This means we have only six different combinations to deal with. Moreover, implicit opinions (text) cannot have any inter-subsumption relation and hence three of those combinations are ruled out. Thus, we have a total of six different rules based on each of inclusion and subsumption. These rules are present in Table 1.

Given two explicit opinions of same sentiment, we apply the rules based on the subsumption relation. Fistly, we check for intra-subsumption related premises within each text and hypothesis and apply the corresponding rules. If not, rules based on unrelated and single premises are applied. Given an implicit opinion and an explicit opinion of same sentiment, we apply the rules based on the inclusion relation. Single premises within the text and hypothesis are checked first and the corresponding rules are applied. Otherwise, hypotheses with related premises are considered and the rule is applied. Then, text and hypothesis with unrelated premises are considered and the rules are applied accordingly.

5. Sentiment-Stance-Specificity[2] (SSS) Dataset

We use an existing hotel reviews corpus, ArguAna (Wachsmuth et al., 2014b) to create our datasets. The data for each hotel contains a balanced set of reviews based on the overall star rating for that hotel. Each review contains manually annotated local sentiment of the statements (pos, neg or obj), aspects present and the overall star rating.

First, we create a knowledge base using a list of aspects extracted from the ArguAna corpus. For example, *(Location $\sqsubseteq_{sub}$ Hotel), (Service $\sqsubseteq_{sub}$ Hotel), (Cleanliness $\sqsubseteq_{sub}$ Hotel), (Staff $\sqsubseteq_{sub}$ Service), (Restaurant service $\sqsubseteq_{sub}$ Service) etc.*

We used the manually annotated dataset of 1288 implicit/explicit opinions created in (Rajendran et al., 2017) which was annotated by two annotators with an inter-rater agreement of Cohen's Kappa = 0.70. Finally, three different dataset were created for our experiment using the proposed rules (few examples in Table 2):

1. **Fully annotated (FA)** This contains a balanced set of 369 reviews from 15 different hotels present in the ArguAna corpus. As explained previously, the local sentiment of statements and aspects present in them are manually annotated. Further, using the definitions from (Rajendran et al., 2017), the extracted opinions are manually annotated as explicit or implicit. There are 264 explicit opinions and 720 implicit opinions present. The SER rules predicted 2220 TH pairs with support-based entailment.

2. **Semi-annotated (SA)** This contains a balanced set of 707 reviews from 33 different hotels present in the ArguAna corpus. Here, the extracted opinions are automatically classfied as explicit or implicit using an SVM-based classifier (Rajendran et al., 2017) with the following features:

   - Surface based features - Unigrams, bigrams and adjective-noun pairs (count of adjective-noun pairs present).
   - Average embedding based feature - For each word, we use the Glove-based (Pennington et al., 2014) word embedding and average these embeddings for an opinion.

   We train a linear SVM classifier using the Scikit-learn[3] package for an undersampled dataset containing 494 explicit opinions and 894 implicit opinions respectively. We use this undersampled data as our training data. We performed a cross-validation on the unbalanced data containing 494 explicit opinions and 1367 implicit opinions to obtain the cost parameter value C of the SVM as 1.0. The cross-validation accuracy of the training data using the above mentioned features is 80% for explicit opinions and 87% for implicit opinions respectively.

   There are 1001 explicit opinions and 4359 implicit opinions present. The SER rules predicted 11892 TH pairs with support-based entailment.

[2] goo.gl/cfBHc7

[3] scikit-learn.org

Experiment	FA	SA	UA
SER	89.54	90.00	96.19
Non-SER	76.18	72.69	88.01
Subsumption based SER	81.63	75.82	92.11
Subsumption based Non-SER	73.91	67.93	86.21
Inclusion based SER	95.83	96.49	97.68
Inclusion based NON-SER	76.87	73.84	88.31
Implicit-Explicit Entailment	75.94	71.03	87.89
Subsumption			
-Rule 1	100.0	83.69	100.0
-Rule 2	86.95	92.40	96.14
-Rule 3	44.44	52.43	80.0
-Rule 4	89.44	93.89	99.15
-Rule 5	62.69	46.67	83.64
-Rule 6	86.69	81.35	92.17
Inclusion			
-Rule 1	92.61	93.74	94.76
-Rule 2	96.0	95.62	96.62
-Rule 3	100.0	94.59	100.0
-Rule 4	97.25	98.50	98.47
-Rule 5	89.79	92.60	95.56
-Rule 6	95.63	97.72	98.59
Random sentiment (SER)	45.62	45.31	47.98
Random sentiment (Non-SER)	38.64	36.37	44.02

Table 4: An experiment was run on each dataset by (a) SER — TH pairs satisfying either of the six subsumption or six inclusion rules (b) Non-SER — TH pairs that do not satisfy any of the 12 rules. (c) Subsumption and Inclusion — TH pairs satisfying each individual rule and (d) Random sentiment — assigning sentiment of opinions present in TH pairs of SER and Non-SER randomly. Accuracy is reported.

3. **Unannotated (UA)** Reviews from 30 different hotels that are unannotated and not present in the ArguAna corpus are used. Here, the reviews are unbalanced. For each statement, local sentiment is automatically classfied as positive, negative or objective using the SVM-based classfier used in the ArguAna (Wachsmuth et al., 2014a) tool. All statements predicted as positive or negative were considered as opinions. We extract a list of aspects manually annotated in the ArguAna corpus and use this to identify aspects present in opinions. The opinions are automatically classfied as implicit or explicit as mentioned for the previous dataset.

   There are 564 explicit opinions and 5933 implicit opinions present. The SER rules predicted 16314 TH pairs with support-based entailment.

6. Experiments and Results

6.1. Performance of SER

In each of the above datasets, we predicted support-based entailment relation using the SER and present the total number of predicted cases in Table. 3. We extracted 160 TH pairs based on the SER as well as those that do not satisfy them. The proportion of TH pairs based on the SER is higher than those that do not satisfy them. We do this to understand whether the pairs extracted using SER rules are accepted by human annotators as well. Two annotators were asked to manually annotate whether the pairs satisfy support-based entailment or not. No information about the rules were provided. The inter-rater agreement was calculated using Cohen's Kappa as 0.80. To test the performance of the SER, we took the intersection of the two annotations as the ground truth data and the accuracy of the SER prediction was 0.83. We also considered the union of the two annotations as the ground truth data which gave the accuracy of the SER prediction as 0.93.

6.2. Performance of Textual Entailment

We use the Excitement Open Platform (EOP) (Magnini et al., 2014) to automatically predict textual entailment in support-based entailment relation and investigate using three different training sets – standard RTE-3 (Giampiccolo et al., 2007), SICK (Marelli et al., 2014) and EXCITEMENT (Kotlerman et al., 2015). The EOP tool takes a text and a hypothesis as input and predicts whether text (T) entails the hypothesis (H) or not. We use the TH pairs that are predicted as support-based entailment using the 12 different SER (Table. 1). Four different entailment decision algorithms (EDA) present in the EOP were used to test the support-based entailment present in the *Fully Annotated* dataset – MaxEntClassificationEDA, AdArteEDA, EditDistanceEDA and PSOEDA. Among these the MaxEntClassificationEDA which is based on the maximum entropy classifier gave the best performance with the RTE-3 dataset and overall accuracy of 89.54 % on the *FA* dataset and hence we use this classifier and the training data for other experiments.

We evaluate the performance of automatically predicting entailment by conducting the following experiments on the three different datasets. The accuracy of correct prediction in each of these experiments is listed in Table. 4 and we describe the experiments below.

Subsumption based SER Based on subsumption rules, two explicit opinions are paired with each other.

Subsumption based Non-SER Two explicit opinions are paired with each other if they do not match any of the subsumption rules.

Inclusion based SER Based on inclusion rules, an implicit opinion is paired with an explicit opinion.

Inclusion based Non-SER An implicit opinion is paired with an explicit opinion, if it does not match any of the inclusion rules.

SER We use pairs extracted in both **Subsumption based SER** and **Inclusion based SER**.

Non-SER We use pairs extracted in both **Subsumption based Non-SER** and **Inclusion based Non-SER**.

Subsumption Text-hypothesis pairs are extracted according to each individual subsumption rule.

Inclusion Text-hypothesis pairs are extracted according to each individual inclusive rule.

Implicit-Explicit Entailment Here, we predict entailment by pairing an explicit opinion with an implicit opinion as text and hypothesis without any rules. The only condition is that both must be of the same sentiment. This is to understand how textual entailment is able to differentiate between explicit and implicit opinions.

Random sentiment For each opinion in each pair present in SER and Non-SER, we randomly assign a local sentiment and predict *support based entailment* relation based on this misinformation.

From Table. 6. we can observe that the overall accuracy of SER outperforms that of Non-SER, which shows that our method is effective for predicting support-based entailment across all datasets. The individual cases, case 3 and 5 in the subsumption category do not perform better than the remaining cases. One reason could be that these two cases are strictly based upon the subsumption relation whereas the rest of them are a combination of both the subsumption and the equivalence relation. Given that these two cases are strictly based on the subsumption relation, it is evident that textual entailment does not depend on the domain ontology and does not consider specificity as a property for prediction.

There is not much difference among the cases present in inclusion, mainly because we differentiate between identical aspects based on the implicit/explicit opinion classification. It is best to compare the accuracy of inclusion-based SER with implicit-explicit entailment to analyse how the implicit/explicit classification affects textual entailment. The performance of inclusion-based SER is better and means that implicit/explicit opinion classification helps in better prediction.

We also experimented by randomly assigning incorrect sentiment (random sentiment baseline) and as expected the accuracy was lowered in comparison with SER.

It has to be noted that the inconsistency in the textual entailment results (Table. 6.) may be higher for the unannotated dataset, even though the results are higher. This is due to the following reasons: (1) the sentiment of the opinions as well as implicit/explicit classification are predicted automatically and (2) only a limited number of aspects are identified.

7. Conclusion

We present three datasets of TH pairs based on a subtype of entailment, which we term as support-based entailment that predicts the support relation between a specific premise and a generalised premise using sentiment, stance and specificity. A distant supervision approach is carried out by using a set of proposed rules based on three components — *sentiment*, *stance* and *specificity*. The performance of these rules against manually annotated 160 TH pairs is measured in terms of accuracy as 0.83. Experiments on the three datasets for the textual entailment task shows that the rules are able to predict the entailment relation but existing textual entailment method is not able to capture support-based entailment. We believe that our datasets will be useful to expedite research in argument mining.

8. Future Work

As part of future work, manually evaluating the unannotated/semi-annotated datasets would be a costly task. Instead, using semi-supervised approaches for automatically classifying implicit/explicit opinions can help in reducing the noise in labels. These datasets can also be useful for learning deep-learning models for predicting *support-based entailment* relation. We will need to evaluate whether such deep-learning models are able to capture the relation without any information such as *sentiment*, *stance* and *target* given explicitly. As of now, we consider only aspects that are explicitly present in an opinion. Given that a lot of existing work (Wang et al., 2011; Hai et al., 2011) in NLP have dealt with identifying explicit and implicit aspects present in online reviews, our work can benefit from this. Another direction for future work is to use the dataset to create argument structures similar to OVA+ structures (Janier et al., 2014).

9. Bibliographical References

Abbas, S. and Sawamura, H. (2008). A first step towards argument mining and its use in arguing agents and its. In *KES*, pages 149–157.

Anand, P., Walker, M., Abbott, R., Tree, J. E. F., Bowmani, R., and Minor, M. (2011). Cats rule and dogs drool!: Classifying stance in online debate. In *WASSA*, pages 1–9.

Augenstein, I., Rocktaschel, T., Vlachos, A., and Bontcheva, K. (2016). Stance detection with bidirectional conditional encoding. In *EMNLP*, pages 876–885.

Boltuzic, F. and Snajder, J. (2014). Back up your stance: Recognizing arguments in online discussions. In *ArgMining@ ACL*, pages 49–58.

Cabrio, E. and Villata, S. (2012). Combining textual entailment and argumentation theory for supporting online debates interactions. In *ACL*, pages 208–212.

Cocarascu, O. and Toni, F. (2016). Detecting deceptive reviews using argumentation. In *PrAISe*, pages 1–8.

Dragoni, M., Pereira, C. D. C., Tettamanzi, A. G., and Villata, S. (2016). Smack: An argumentation framework for opinion mining. In *IJCAI*, pages 4242–4243.

Ebrahimi, J., Dou, D., and Lowd, D. (2016). A joint sentiment-target-stance model for stance classification in tweets. In *COLING*, pages 2656–2665.

Giampiccolo, D., Magnini, B., Dagan, I., and Dolan, B. (2007). The third pascal recognizing textual entailment challenge. In *ACL-PASCAL*, pages 1–9.

Grosse, K., Chesñevar, C. I., and Maguitman, A. G. (2012). An argument-based approach to mining opinions from twitter. In *AT*, pages 408–422.

Hai, Z., Chang, K., and Kim, J.-j. (2011). Implicit feature identification via co-occurrence association rule mining. *CiCLing*, pages 393–404.

Janier, M., Lawrence, J., and Reed, C. (2014). Ova+: an argument analysis interface. In *COMMA*, pages 463–464.

Kotlerman, L., Dagan, I., Magnini, B., and Bentivogli, L. (2015). Textual entailment graphs. *Natural Language Engineering*, 21:699–724.

Lippi, M. and Torroni, P. (2015). Argument mining: A machine learning perspective. In *TAFA*, pages 163–176.

Lippi, M. and Torroni, P. (2016). Argumentation mining: State of the art and emerging trends. *TOIT*, 16(2):10.

Magnini, B., Zanoli, R., Dagan, I., Eichler, K., Neumann, G., Noh, T.-G., Pado, S., Stern, A., and Levy, O. (2014). The excitement open platform for textual inferences. In *ACL*, pages 43–48.

Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., and Zamparelli, R. (2014). A sick cure for the evaluation of compositional distributional semantic models. In *LREC*, pages 216–223.

Matthias, K. A.-K. H. W. and Stein, H. J. K. B. (2016). Cross-domain mining of argumentative text through distant supervision. In *NAACL-HLT*, pages 1395–1404.

Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X.-D., and Cherry, C. (2016). A dataset for detecting stance in tweets. In *LREC*, pages 3945–3952.

Palau, R. M. and Moens, M.-F. (2009). Argumentation mining: the detection, classification and structure of arguments in text. In *ICAIL*, pages 98–107.

Park, J. and Cardie, C. (2014). Identifying appropriate support for propositions in online user comments. In *ACL*, pages 29–38.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543.

Rajendran, P., Bollegala, D., and Parsons, S. (2016). Assessing weight of opinion by aggregating coalitions of arguments. In *COMMA*, pages 431–438.

Rajendran, P., Bollegala, D., and Parsons, S. (2017). Identifying argument based relation properties in opinions. In *PACLING*, page to appear.

Sobhani, P., Mohammad, S., and Kiritchenko, S. (2016). Detecting stance in tweets and analyzing its interaction with sentiment. In *SEM*, pages 159–169.

Villalba, M. P. G. and Saint-Dizier, P. (2012). A framework to extract arguments in opinion texts. *IJCINI*, 6:62–87.

Wachsmuth, H., Trenkmann, M., Stein, B., and Engels, G. (2014a). Modeling review argumentation for robust sentiment analysis. In *COLING*, pages 553–564.

Wachsmuth, H., Trenkmann, M., Stein, B., Engels, G., and Palakarska, T. (2014b). A review corpus for argumentation analysis. In *CICLing*, pages 115–127.

Wang, H., Lu, Y., and Zhai, C. (2011). Latent aspect rating analysis without aspect keyword supervision. In *SIGKDD*, pages 618–626.

Wyner, A., Schneider, J., Atkinson, K., and Bench-Capon, T. J. (2012). Semi-automated argumentative analysis of online product reviews. In *COMMA*, pages 43–50.

Yokote, K., Bollegala, D., and Ishizuka, M. (2012). Similarity is not entailment - jointly learning similarity transformations for textual entailment. In *AAAI*, pages 1720–1726.

Zanzotto, F. M., Pazienza, M. T., and Pennacchiotti, M. (2005). Discovering entailment relations using textual entailment patterns. In *ACL*, pages 37–42.