Identifying Appropriate Support for Propositions in Online User Comments

Joonsuk Park
Department of Computer Science
Cornell University
Ithaca, NY, USA
jpark@cs.cornell.edu

Claire Cardie
Department of Computer Science
Cornell University
Ithaca, NY, USA
cardie@cs.cornell.edu

Abstract

The ability to analyze the adequacy of supporting information is necessary for determining the strength of an argument.[1] This is especially the case for online user comments, which often consist of arguments lacking proper substantiation and reasoning. Thus, we develop a framework for automatically classifying each proposition as UNVERIFIABLE, VERIFIABLE NON-EXPERIENTIAL, or VERIFIABLE EXPERIENTIAL[2], where the appropriate type of support is *reason*, *evidence*, and *optional evidence*, respectively[3]. Once the existing support for propositions are identified, this classification can provide an estimate of how adequately the arguments have been supported. We build a gold-standard dataset of 9,476 sentences and clauses from 1,047 comments submitted to an eRulemaking platform and find that Support Vector Machine (SVM) classifiers trained with n-grams and additional features capturing the verifiability and experientiality exhibit statistically significant improvement over the unigram baseline, achieving a macro-averaged F_1 of 68.99%.

1 Introduction

Argumentation mining is a relatively new field focusing on identifying and extracting argumentative structures in documents. An *argument* is typically defined as a conclusion with supporting premises, which can be conclusions of other arguments themselves (Toulmin, 1958; Toulmin et al., 1979; Pollock, 1987). To date, much of the argumentation mining research has been conducted on domains like news articles, parliamentary records and legal documents, where the documents contain well-formed explicit arguments, i.e. propositions with supporting reasons and evidence present in the text (Moens et al., 2007; Palau and Moens, 2009; Wyner et al., 2010; Feng and Hirst, 2011; Ashley and Walker, 2013).

Unlike documents written by professionals, online user comments often contain arguments with inappropriate or missing justification. One way to deal with such implicit arguments is to simply disregard them and focus on extracting arguments containing proper support (Villalba and Saint-Dizier, 2012; Cabrio and Villata, 2012). However, recognizing such propositions as part of an argument,[4] and determining the appropriate types of support can be useful for assessing the adequacy of the supporting information, and in turn, the strength of the whole argument. Consider the following examples:

> *How much does a small carton of milk cost?*$_1$ *More children should drink milk*$_2$*, because children who drink milk everyday are taller than those who don't*$_3$*. Children would want to drink milk, anyway*$_4$*.*

Firstly, **Sentence 1** does not need any support, nor is it part of an argument. Next, **Proposition 2** is an *unverifiable* proposition because it cannot be proved with objective evidence, due to the value judgement. Instead, it can be supported by a reason explaining why it may be true. If the reason, **Proposition 3**, were not true, the whole ar-

[1] In this work, even unsupported propositions are consider part of an argument. Not disregarding such implicit arguments allows us to discuss the types of support that can further be provided to strengthen the argument, as a form of assessment.

[2] Verifiable Experiential propositions are verifiable propositions about personal state or experience. See Table 1 for examples.

[3] We are assuming that there is no background knowledge that eliminates the need of support.

[4] Not all sentences in user comments are part of an argument, e.g. questions and greetings. We address this in Section 4.1

Proceedings of the First Workshop on Argumentation Mining, pages 29–38,
Baltimore, Maryland USA, June 26, 2014.

gument would fall apart, giving little weight to **Proposition 2**. Thus, an objective evidence supporting **Proposition 3**, which is a *verifiable* proposition, could be provided to strengthen the argument. Lastly, as **Proposition 4** is *unverifiable*, we cannot expect an objective evidence that proves it, but a reason as its support. Note that providing a reason why **Proposition 3** might be true is not as effective as substantiating it with a proof, but is still better than having no support. This shows that not only the presence, but also the type of supporting information affects the strength of the argument.

Examining each proposition in this way, i.e. with respect to its verifiability, provides a means to determine the desirable types of support, if any, and enables the analysis of the arguments in terms of the adequacy of their support. Thus, we propose the task of classifying each proposition (the elementary unit of argumentation in this work) in an argument as UNVERIFIABLE, VERIFIABLE PUBLIC, or VERIFIABLE PRIVATE, where the appropriate type of support is *reason*, *evidence*, and *optional evidence*, respectively. To perform the experiments, we annotate 9,476 sentences and clauses from 1,047 comments extracted from an eRulemaking platform.

In the remainder of the paper, we describe the annotation scheme and a newly created dataset (Section 2), propose a supervised learning approach to the task (Section 3), evaluate the approach (Section 4), and survey related work (Section 5). We find that Support Vector Machines (SVM) classifiers trained with n-grams and other features to capture the verifiability and experientiality exhibit statistically significant improvement over the unigram baseline, achieving a macro-averaged F_1 score of 68.99%.

2 Data

We have collected and manually annotated sentences and (independent) clauses from user comments extracted from an eRulemaking website, *Regulation Room*[5]. Rulemaking is the process by which U.S. government agencies make new regulations and enact public policy; its digital counterpart — *eRulemaking* — moves the process to online platforms (see, e.g. (Park et al., 2012)). By providing platforms in which the public can discuss regulations that interest them, government agencies hope to enlist the expertise and experience of participants to create better regulations. In many rulemaking scenarios, agencies are, in fact, required to obtain feedback from the public on the proposed regulation as well as to address all substantive questions, criticisms or suggestions that are raised (Lubbers, 2006). In this way, public comments can produce changes in the final rule (Hochschild and Danielson, 1998) that, in turn, can affect millions of lives. It is crucial, therefore, for rule makers to be able to identify credible comments from those submitted.

Regulation Room is an experimental website operated by Cornell eRulemaking Initiative (CeRI)[6] to promote public participation in the rulemaking process, help users write more informative comments and build collective knowledge via active discussions guided by human moderators. *Regulation Room* hosts actual regulations from government agencies, such as the U.S. Department of Transportation.

For our research, we collected and manually annotated 9,476 propositions from 1,047 user comments from two recent rules: Airline Passenger Rights (serving peanuts on the plane, tarmac delay contingency plan, oversales of tickets, baggage fees and other airline traveller rights) and Home Mortgage Consumer Protection (loss mitigation, accounting error resolution, etc.).

2.1 Annotation Scheme

To start, we collected 1,147 comments and randomly selected 100 of them to devise an annotation scheme for identifying appropriate types of support for propositions and to train annotators. Initially, we allowed the annotators to define the span for a propositions, leading to various complications and a low inter-annotator reliability. Thus, we introduced an additional step in which comments were manually sliced into propositions (or non-propositional sentences) before being given to the annotators. A proposition or sentence found this way was split further if it consisted of two or more independent clauses. The sliced comments were then coded by two annotators into the following four disjoint classes (See Figure 1 for an overview):

Verifiable Proposition [Experiential(VERIF_{EXP}) and Non-experiential(VERIF_{NON})]. A proposition is verifiable if it contains an objective asser-

[5]http://www.regulationroom.org

[6]http://www.lawschool.cornell.edu/ceri/

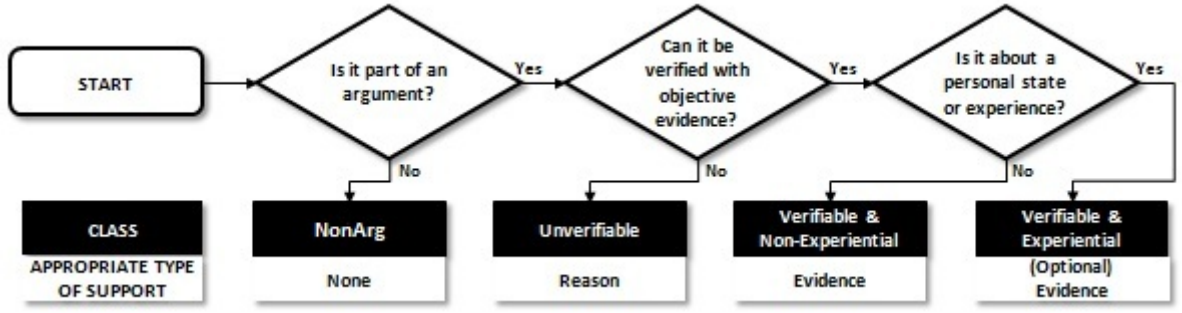

Figure 1: Flow chart for annotation (*It* refers to the sentence (or clause) being annotated)

	#	proposition
VERIF$_{EXP}$	1	I've been a physician for 20 years.
	2	My son has hypolycemia.
	3	They flew me to NY in February.
	4	The flight attendant yelled at the passengers.
VERIF$_{NON}$	5	*They can have inhalation reactions.*
	6	since *they serve them to the whole plane.*
	7	*Peanuts do not kill people.*
	8	Clearly, *peanuts do not kill people.*
	9	I believe *peanuts do not kill people.*
	10	*The governor said* that he enjoyed it.
	11	*food allergies are rare*
	12	*food allergies are seen in less than 20% of the population*
UNVERIF	13	Again, keep it simple.
	14	Banning peanuts will reduce deaths.
	15	I enjoy having peanuts on the plane.
	16	others are of uncertain significance
	17	banning peanuts is a slippery slope
NONARG	18	Who is in charge of this?
	19	I have two comments
	20	http://www.someurl.com
	21	Thanks for allowing me to comment.
	22	- Mike

Table 1: Example Sentences.

* Italics is used to illustrate *core clause* (Section 3.2).

tion, where *objective* means "expressing or dealing with facts or conditions as perceived without distortion by personal feelings, prejudices, or interpretations."[7] Such assertions have truth values that can be proved or disproved with objective evidence[8]:

Consider the examples from Table 1. propositions 1 through 7 are clearly verifiable because they only contain objective assertions. propositions 8 and 9 show that adding subjective expressions such as "Clearly" (e.g. sentence 8) or "I believe that" (e.g. sentence 9) to an objectively verifiable proposition (e.g. sentence 7) does not affect the verifiability of the proposition. Sentence 10 is considered verifiable because whether or not the governor *said* "he enjoyed the peanuts" can be verified with objective evidence, even though whether he really does or not cannot be verified.

For the purpose of identifying an appropriate type of support, we employ a rather lenient notion of objectivity: an assertion is objectively verifiable if the *domain of comparison* is free of interpretation. For instance, sentence 11 is regarded as objectively verifiable, because it is clear, i.e. it is not open for interpretation, that *percentage of the population* is the metric under comparison even though the *threshold* is purely subjective[9]. The rationale is that this type of proposition can be sufficiently substantiated with objective evidence (e.g. published statistics showing the percentage of people suffering from food allergies). Another way to think about it is that sentence 11 is a loose way of saying a (more obviously) verifiable sentence 12, where the commenter neglected to mention the threshold. This is fundamentally different from propositions 13 through 16 for which objective evidence cannot exist[10].

A verifiable proposition can further be distinguished as experiential or not, depending on whether the proposition is about the writer's personal state or experience (VERIF$_{EXP}$) or something non-experiential (VERIF$_{NON}$). This difference determines whether objective evidence is mandatory or optional with respect to the credibility of the comment. Evidence is optional when the evidence contains private information or is practically impossible to be provided: While propositions 1 through 3 can be proved with pictures of official documents, for instance, the commenters may not want to provide them for privacy reasons. Also, the website interface may not al-

[7] http://www.merriam-webster.com/

[8] The correctness of the assertion or the availability of the objective evidence does not matter.

[9] One may think anything less frequent than the average is rare and another may have more stricter notion.

[10] Objective evidence may exist for propositions that provide *reasons* for propositions 13 through 16.

Regulation	VERIF$_{NON}$	VERIF$_{EXP}$	UNVERIF	Subtotal	NONARG	Total	# of Comments
APR	1106	851	4413	6370	522	6892	820
HMCP	251	416	1733	2400	186	2586	227
Total	1357	1267	6146	8770	708	9476	1047

Table 2: Class Distribution Over Sentences and Clauses

low pictures to be uploaded in comment section, which is the case with most websites. sentence 4 is practically impossible to prove unless the commenter happened to have recorded the conversation, and the website interface allows multimedia files to be uploaded. This is different from propositions 5 through 12, which should be (if valid, that is) based on non-experiential knowledge the commenter acquired through objective evidence available to the public.

In certain domains, VERIF$_{EXP}$ propositions—sometimes referred to as *anectotal evidence*—provide the novel knowledge that readers are seeking. In eRulemaking, for instance, agencies accept a wide variety of comments from the public, including accounts of personal experience with the problems or conditions the new regulation proposes to address. If these accounts are relevant and plausible, the agencies may use them, even if they include no independent substantiation. Taking it to an extreme, even if the "experience" is fake, the "experience" and opinions based on them are valuable to the agencies as long as the "experience" is realistic.

Unverifiable Proposition (UNVERIF). A proposition is unverifiable if it cannot be proved with objective evidence. UNVERIF propositions are typically opinions, suggestions, judgements, or assertions about what will happen in the future. (See propositions 13 through 17.) Assertions about the future are typically unverifiable, because there is no direct evidence that something will happen. A very prominent exception is a prediction based on a policy of organizations, i.e. "The store will be open this Sunday." where the policy serves as a direct evidence.

Non-Argumentative (NONARG). A sentence or clause is in this category if it is not a proposition, i.e. it cannot be verified with objective evidence and no supporting reason is required for the purpose of improving the comment quality. Examples include question, greeting, citation, and URL. (See sentences 18 through 21.)

2.2 Annotation Results

The resulting distribution of classes is shown in Table 2. Note that even though we employed a rather lenient definition of objective propositions, the distribution is highly skewed towards UNVERIF propositions. This is expected because the comments are written by people who want to express their opinions about a regulation. Also, NONARG sentences comprise about 7% of the data, suggesting that most comment propositions need to be supported with a reason or evidence for maximal credibility.

The inter-coder reliability checked on 30% of the data is moderate, yielding an *Unweighted Cohen's* κ of 0.73. Most of the disagreement occurred in propositions like "Airlines have to provide compensation for both fees and lost bags" in which it is not clear from the context whether it is an opinion (UNVERIF) or a law (VERIF$_{NON}$). Also, opinions that may be verifiable (e.g. "The problems with passenger experience are not dependant on aircraft size!") seem to cause disagreement among annotators.

3 Proposition Type Classification

3.1 Learning Algorithm

To classify each proposition in an argument as VERIF$_{NON}$, VERIF$_{EXP}$, or UNVERIF, we train multiclass Support Vector Machines (SVM) as formulated by Crammer and Singer (2002), and later extended by Keerthi et al.(2008). We use the LibLinear (Fan et al., 2008) implementation. We experimented with other multiclass SVM approaches such as 1-vs-all and 1-vs-1 (all-vs-all), but the differences were statistically insignificant, consistent with Hsu and Lin's (2002) empirical comparison of these methods. Thus, we only report the performance of the Crammer and Singer version of Multiclass SVM.

3.2 Features

The features are binary-valued, and the feature vector for each data point is normalized to have the unit length: "Presence" features are binary features indicating whether the given feature is present in the proposition or not; "Count" features

are numeric counts of the occurrence of each feature is converted to a set of three binary features each denoting 0, 1 and 2 or more occurrences. We first tried a *binning* method with each digit as its own interval, resulting in binary features of the form $featCnt_n$, but the three-interval approach proved to be better empirically, and is consistent with the approach by Riloff and Shoen (1995).

The features can be grouped into three categories by purpose: Verifiability-specific (VER), Experientiality-specific (EXP) and Basic Features, each designed to capture the given proposition's verifiability, experientiality, and both, respectively. Now we discuss the features in more detail.

3.2.1 Basic Features

N-gram Presence A set of binary features denote whether a given unigram or bigram occurs in the proposition. The intuition is that by examining the occurrence of words or phrases in VERIF$_{NON}$, VERIF$_{EXP}$, and UNVERIF propositions, the classes that have close ties to certain words and phrases can be identified. For instance, when a proposition contains the word *happy*, the proposition tends to be UNVERIF. From this observation, we can speculate that *happy* is highly associated with UNVERIF, and *went*, VERIF$_{EXP}$. n-gram presence, rather than the raw or normalized frequency is chosen for its superior performance (O'Keefe and Koprinska, 2009).

Core Clause Tag (CCT) To correctly classify propositions with main or subordinate clauses that do not affect the verifiability of the proposition (e.g. propositions 8 through 10 in Table 1, respectively), it is necessary to distinguish features that appear in the main clause from those that appear in the subordinate clause. Thus, we employ an auxiliary feature that adds clausal information to other features by tagging them as either *core* or *accessory* clause.

Let's consider propositions 7, 9 and 10 in Table 1: In all three examples, the *core clause* is italicized. In single clause cases like proposition 7, the entire proposition is the core clause. However, for proposition 9, the core clause is the subordinate clause introduced by the main clause, i.e. "I believe" should be ignored, since the verifiability of "peanuts do not kill people" is not dependent on it. It is the opposite for proposition 10: the main clause "The governor said" is the core clause, and the rest need not be considered. The reason is that "said" is a speech event, and it is possible to objectively verify whether or not the governor verbally expressed his appreciation of peanuts.

To realize this intuition, we use syntactic parse trees generated by the Stanford Parser (De Marneffe et al., 2006). In particular, Penn Treebank 2 Tags contain a clause-level tag *SBAR* denoting a "clause introduced by a subordinating conjunction" (Marcus et al., 1993). The "that" clause in proposition 10 spans a subtree rooted by *SBAR*, whose left-most child has a lexical value "that." Similarly, the subordinate (non-italicized) clause in proposition 9 falls in a subtree rooted by *SBAR*, whose only child is *S*. Once the main clause of a given proposition is identified, all features set off by the clause are tagged as "core" and the rest are tagged as "accessory." If a speech event is present, the tags are flipped.

3.2.2 Verifiability-specific Features (*VER*)

Parts-of-Speech (POS) Count Rayson et al. (2001) have shown that the POS distribution is distinct in imaginative vs. informative writing. We expect this feature to distinguish UNVERIF propositions from the rest.

Sentiment Clue Count Wilson et al. (2005) provides a subjectivity clue lexicon, which is a list of words with sentiment strength tags, either strong or weak, along with additional information, such as the sentiment polarity, *Part-of-Speech Count* (POS), etc. We suspect that propositions containing more sentiment words is more likely to be UNVERIF.

Speech Event Count We use the 50 most frequent *Objective-speech-event* text anchors crawled from the *MPQA 2.0* corpus (Wilson and Wiebe, 2005) as a speech event lexicon. The speech event text anchors refer to words like "stated" and "wrote" that introduce written or spoken propositions attributed to a source. propositions containing speech events such as proposition 10 in Table 1 are generally VERIF$_{NON}$ or VERIF$_{EXP}$, since whether the attributed source has indeed made the proposition he allegedly made is objectively verifiable regardless of the subjectivity of the proposition itself.

Imperative Expression Count Imperatives, i.e. commands, are generally UNVERIF (e.g. "Do the homework now!" that is, we expect there to be no objective evidence proving that the homework should be done right away.), unless the sentence is a law or general procedure (e.g. "The library should allow you to check out books." where the

context makes it clear that the writer is claiming that the library lends out books.) This feature denotes whether the proposition begins with a verb or contains the following: *must, should, need to, have to, ought to*.

Emotion Expression Count These features target specific tokens "!", and "..." as well as fully capitalized word tokens to capture the emotion in text. The rationale is that expression of emotion is likely to be more prevalent in UNVERIF propositions.

3.2.3 Experientiality-specific Features (*EXP*)

Tense Count propositions written in past tense are rarely VERIF$_{NON}$, because even in the case that the statment is verifiable, they are likely to be the commenter's past experience, i.e. VERIF$_{EXP}$. Future tense are typically UNVERIF because propositions about what will happen in the future are often unverifiable with objective evidence, with exception being propositions like predictions based on policy of organizations, i.e. "Fedex will deliver on Sunday." To take advantage of these observations, three binary features capture each of three tenses: *past, present,* and *future*.

Person Count First person narratives can suggest that the proposition is UNVERIF or VERIF$_{EXP}$, except for rare cases like "We, the passengers,..." in which the first person pronoun refers to a large body of individuals. This intuition is captured by having binary features for: *1st, 2nd* and *3rd person*.

4 Experiments

4.1 Methodology

A Note on Argument Detection A natural first step in argumentation mining is to determine which portions of the given document comprise an argument. It can also be framed as a binary classification task in which each proposition in the document needs to be classified as either argumentative or not. Some authors choose to skip this step (Feng and Hirst, 2011), while others make use of various classifiers to achieve high level of accuracy, as Palau and Moens achieved over 70% accuracy on Araucaria and ECHR corpus (Reed and Moens, 2008; Palau and Moens, 2009).

As we have discussed in Section 1, our setup is a bit unique in that we also consider implicit arguments, where propositions are not supported with explicit reason or evidence, as argumentative. As a result, only about 7%($\frac{\text{NONARG}}{\text{TOTAL}}$ in Table 2) of our entire dataset is marked as non-argumentative, most of which consists of questions and greetings. By simply searching for specific unigrams, such as "?" and "thank", we achieve over 99% F_1 score in determining which propositions are part of an argument.

The remaining experiments were done without non-argumentative propositions, i.e. NONARG in Table 2.

Experimental Setup We first randomly selected 292 comments as held-out test set, resulting in the distribution shown in Table 4. Then, VERIF$_{NON}$ and VERIF$_{EXP}$ in the training set were oversampled so that the classes are equally distributed. During training, five fold cross-validation was done on the training set to tune the *C* parameter to *32*. Because the micro-averaged F_1 score can be easily boosted on datasets with highly skewed class distribution, we optimize for the macro-averaged F_1 score.

Preprocessing was kept at a minimal level: capital letters were lowercased after counting fully capitalized words, and numbers were converted to a *NUM* token.

	VERIF$_{NON}$	VERIF$_{EXP}$	UNVERIF	Total
Train	987	900	4459	6346
Test	370	367	1687	2424
Total	1357	1267	6146	8770

Table 4: # of propositions in Train and Test Set

4.2 Results & Analysis

Table 3 shows a summary of the classification results. The best overall performance is achieved by combining all features (*UNI+BI+VER+EXP*), yielding 68.99% macro-averaged F_1, where the gain over the baseline is statistically significant according to the bootstrap method with 10,000 samples (Efron and Tibshirani, 1994; Berg-Kirkpatrick et al., 2012).

Core Clause Tag (CCT) We do not report the performance of employing feature sets with *Core Clause Tag (CCT)* in Table 3, because the effect of *CCT* on each of the six sets of features is statistically insignificant. This is surprising at first, given the strong motivation for distinguishing the core clause from auxiliary clause, as addressed in the previous section: Subordinate clauses like "I believe" should not cause the entire proposition to be classified as UNVERIF, and clauses like "He said" should serve as a queue for VERIF$_{NON}$ or VERIF$_{EXP}$, even if an unverifiable clause follows

Feature Set	UNVERIF vs All			VERIF$_{NON}$ vs All			VERIF$_{EXP}$ vs All			Average F_1	
	Pre.	Rec.	F_1	Pre.	Rec.	F_1	Pre.	Rec.	F_1	Macro	Micro
UNI(base)	85.24	79.43	82.23	42.57	51.89	46.77	61.10	66.76	63.80	64.27	73.31
UNI+BI	82.14	89.69*	85.75*	51.67*	37.57	43.51	73.48*	62.67	67.65*	65.63	77.64*
VER	88.52*	52.10	65.60	28.41	61.35*	38.84	42.41	73.02*	53.65	52.70	56.68
EXP	82.42	4.45	8.44	20.92	76.49*	32.85	31.02	82.83*	45.14	28.81	27.31
VER+EXP	89.40*	49.50	63.72	29.25	71.62*	41.54	50.00	79.56*	61.41	55.55	57.43
UNI+BI+ VER+EXP	86.86*	83.05*	84.91*	49.88*	55.14	52.37*	66.67*	73.02*	69.70*	**68.99***	77.27*

Table 3: Three class classification results in % (Crammer & Singer's Multiclass SVMs)

Precision, recall, and F_1 scores are computed with respect to each one-vs-all classification problem for evaluation purposes, though a single machine is built for the multi-class classification problem, instead of 3 one-vs-all classifiers. The star (*) indicates that the given result is statistically significantly better than the unigram baseline.

Fts		*UNI*	*UNI$_{CCT}$*
UNVERIF	+	should, whatever, responsibility	should$_C$, should$_A$, understand$_C$
	-	previous, solve, florida, exposed, reacted, reply, kinds	exposed$_C$, solve$_C$, NUM$_C$, florida$_C$, reacted$_C$, pool$_C$, owed$_C$
VERIF$_{NON}$	+	impacted, NUM, solve, cars, pull, kinds, congress	impacted$_C$, solve$_C$, cars$_C$, NUM$_C$, pool$_C$, writing$_C$, death$_C$, link$_C$
	-	should, seems, comments	should$_C$, comments$_C$
VERIF$_{EXP}$	+	owed, consumed, saw, expert, interesting, him, reacted, refinance	owed$_C$, consumed$_C$, expert$_C$, reacted$_C$, happened$_C$, interesting$_C$
	-	impacted, wo	impacted$_C$, wo$_C$, concern$_C$, died$_C$

Table 5: Most Informative Features for *UNI* and *UNI$_{CCT}$*

10 Unigrams with the largest weight (magnitude) with respect to each class (+ : positive weight / - : negative weight).

it. Our conjecture turned out to be wrong, mainly because such distinction can be made for only a small subset of the data: For instance, over 83% of the unigrams are tagged as *core* in the *UNI* feature set. Thus, most of the important features for feature sets with *CCT* end up being features with *core* tag, and the important features for feature sets with and without *CCT* are practically the same, as shown in Table 5, resulting in statistically insignificant performance differences.

Informative Features The most informative features reported in Table 6 exhibit interesting differences among the three classes: Sentiment bearing words, i.e. "should" and strong sentiment clues, are good indicators of UNVERIF, whereas person and tense information is crucial for VERIF$_{EXP}$. As expected, the strong indicators of UNVERIF and VERIF$_{EXP}$, namely "should" and PER$_{1st}$ are negatively associated with VERIF$_{NON}$. It is intriguing to see that the heavily weighted features of VERIF$_{NON}$ are non-verb content words, unlike those of the other classes. One explanation for this is that VERIF$_{NON}$ are rarely indicated by specific cues; instead, a good sign of VERIF$_{NON}$ is the absences of cues for the other classes, which are often function words and verbs. What is remaining, then, are non-verb content words. Also, certain content words seem to be more likely to bring about factual discussions. For instance, technical terms like"allergen" and "airborne," appear in verifiable non-experiential propositions as "The FDA requires labeling for the following 8 allergens."

Feature Set		*UNI+BI+VER+EXP*
UNVERIF	+	should,StrSentClue$_{>2}$, VB$_{>2}$
	-	StrSentClue$_0$, VBD$_{>2}$, air, since, no_one, allergic, not_an
VERIF$_{NON}$	+	die, death, reaction, person, allergen, airborne, no_one, allergies
	-	PER$_{1st}$, should
VERIF$_{EXP}$	+	VBD$_{>2}$, PER$_{1st}$, i_have, his, he, him, time_!
	-	VBZ$_{>2}$, PER$_{2nd}$

Table 6: Most Informative Features for *UNI+BI+VER+EXP*

10 Features with the largest weight (magnitude) with respect to each class (+ : positive weight / - : negative weight).

Non-n-gram Features Table 3 clearly shows that the three non-n-gram features, *VER*, *EXP*, and *VER+EXP*, do not perform as well as the n-gram features. But still, the performance is impressive, given the drastic difference in the dimensionality of the features: Even the combined feature set, *VER+EXP*, consists of only about 100 features, when there are over 8,000 unigrams and close to 70,000 bigrams. In other words, the non-n-gram features are effectively capturing characteristics of each class. This is very promising, since this shows that a better understanding of the types of proposition can potentially lead to a more concise set of features with equal, or even better, performance.

Also notice that *VER* outperforms *EXP* for the most part, even with respect to VERIF$_{NON}$ *vs All* and VERIF$_{EXP}$ *vs All*, except for recall. This is in-

triguing, because *VER* are mostly from subjectivity detection domain, intended to capture the subjectivity of words in the propositions leveraging on pre-built lexia. Simply considering subjectivity of words should provide no means of distinguishing VERIF$_{NON}$ from VERIF$_{EXP}$. One of the reasons for *VER*'s superior performance over *EXP* is that *EXP* by itself is inadequate for the classification task: *EXP* consists of only 6 (or 12 with CCT) features denoting the person and tense information. Another reason is that *VER*, in a limited fashion, does encode experientiality: For instance, past tense propositions can be identified with the existence of *VBD*(verb, past tense) and *VBN*(verb, past participle).

5 Related Work

Argumentation Mining The primary goal of argumentation mining has been to identify and extract argumentative structures present in documents, which are often written by professionals (Moens et al., 2007; Wyner et al., 2010; Feng and Hirst, 2011; Ashley and Walker, 2013). In certain cases, the specific document structure allows additional means of identify arguments (Mochales and Moens, 2008). Even the work on online text data, which are less rigid in structure and often contain insufficiently supported propositions, focus on the extraction of arguments (Villalba and Saint-Dizier, 2012; Cabrio and Villata, 2012). We, however, are interested in the assessment of the argumentative structure, potentially providing recommendations to readers and feedback to the writers. Thus it is crucial that we also process unsubstantiated propositions, which we consider as implicit arguments. Our approach should be valuable for processing documents like online user comment where arguments may not have adequate support and an automatic means of analysis can be useful.

Subjectivity Detection Work to distinguish subjective from objective propositions (e.g.(Wiebe and Riloff, 2005)), often a subtask for sentiment analysis (Pang and Lee, 2008), is relevant to our work since we are concerned with the objective verifiability of propositions. In particular, previous work attempts to detect certain types of subjective proposition: Conrad et al. (2012) identify *arguing subjectivity* propositions and tag them with argument labels in order to cluster argument paraphrases. Others incorporate this task as a component for solving related problems, such as answering opinion-based questions and determining the writer's political stance (Somasundaran et al., 2007; Somasundaran and Wiebe, 2010). Similarly, Rosenthal and McKeown (2012) identify *opinionated* propositions expressing beliefs, leveraging from previous work in sentiment analysis and belief tagging. While the class of *subjective* propositions in subjectivity detection strictly contains UNVERIF propositions, it also partially overlaps with the VERIF$_{EXP}$ and VERIF$_{NON}$ classes of our work: We want to identify verifiable assertions within propositions, rather than determine the subjectivity of the proposition as a whole (e.g. proposition 8 in Table 1 is classified as a VERIF$_{NON}$, though "Clearly" is subjective.). We also distinguish two types of verifiable propositions, which is necessary for the purpose of identifying appropriate types of support.

6 Conclusions and Future Work

We have proposed a novel task of automatically classifying each proposition as UNVERIFIABLE, VERIFIABLE NONEXPERIENTIAL, or VERIFIABLE EXPERIENTIAL, where the appropriate type of support is *reason*, *evidence*, and *optional evidence*, respectively. This classification, once the existing support relations among propositions are identified, can provide an estimate of how well the arguments are supported. We find that Support Vector Machines (SVM) classifiers trained with n-grams and other features to capture the verifiability and experientiality exhibit statistically significant improvement over the unigram baseline, achieving a macro-averaged F_1 score of 68.99%. In the process, we have built a gold-standard dataset of 9,476 propositions from 1,047 comments submitted to an eRulemaking platform.

One immediate avenue for future work is to incorporate the identification of relations among the propositions in an argument to the system to analyze the adequacy of the supporting information in the argument. This, in turn, can be used to recommend comments to readers and provide feedback to writers so that they can construct better arguments.

Acknowledgments

This work was supported in part by NSF grants IIS-1111176 and IIS–1314778. We thank our annotators, Pamela Ijeoma Amaechi and Simon Boehme, as well as the Cornell NLP Group and the reviewers for helpful comments.

References

Kevin D. Ashley and Vern R. Walker. 2013. From information retrieval (ir) to argument retrieval (ar) for legal cases: Report on a baseline study. In *JURIX*, pages 29–38.

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An empirical investigation of statistical significance in nlp. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 995–1005, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Alexander Conrad, Janyce Wiebe, Hwa, and Rebecca. 2012. Recognizing arguing subjectivity and argument tags. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, ExProM '12, pages 80–88, Stroudsburg, PA, USA. Association for Computational Linguistics.

Koby Crammer and Yoram Singer. 2002. On the algorithmic implementation of multiclass kernel-based vector machines. *J. Mach. Learn. Res.*, 2:265–292, March.

Marie-Catherine De Marneffe, Bill Maccartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *In Proc. Intl Conf. on Language Resources and Evaluation (LREC*, pages 449–454.

B. Efron and R.J. Tibshirani. 1994. *An Introduction to the Bootstrap*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. *J. Mach. Learn. Res.*, 9:1871–1874, June.

Vanessa Wei Feng and Graeme Hirst. 2011. Classifying arguments by scheme. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 987–996, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jennifer L. Hochschild and Michael Danielson, 1998. *Can We Desegregate Public Schools and Subsidized Housing? Lessons from the Sorry History of Yonkers, New York*, chapter 2, pages 23–44. University Press of Kansas, Lawrence KS, edited by clarence stone edition.

Chih-Wei Hsu and Chih-Jen Lin. 2002. A comparison of methods for multiclass support vector machines. *Trans. Neur. Netw.*, 13(2):415–425, March.

S. Sathiya Keerthi, S. Sundararajan, Kai-Wei Chang, Cho-Jui Hsieh, and Chih-Jen Lin. 2008. A sequential dual method for large scale multi-class linear svms. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 408–416, New York, NY, USA. ACM.

Jeffrey S. Lubbers. 2006. *A Guide to Federal Agency Rulemaking*. American Bar Association Chicago, 4th ed. edition.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank. *COMPUTATIONAL LINGUISTICS*, 19(2):313–330.

Raquel Mochales and Marie-Francine Moens. 2008. Study on the structure of argumentation in case law. In *Proceedings of the 2008 Conference on Legal Knowledge and Information Systems: JURIX 2008: The Twenty-First Annual Conference*, pages 11–20, Amsterdam, The Netherlands, The Netherlands. IOS Press.

Marie-Francine Moens, Erik Boiy, Raquel Mochales Palau, and Chris Reed. 2007. Automatic detection of arguments in legal texts. In *Proceedings of the 11th International Conference on Artificial Intelligence and Law*, ICAIL '07, pages 225–230, New York, NY, USA. ACM.

Tim O'Keefe and Irena Koprinska. 2009. Feature selection and weighting methods in sentiment analysis. In *Proceedings of the 14th Australasian Document Computing Symposium*.

Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation mining: The detection, classification and structure of arguments in text. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, ICAIL '09, pages 98–107, New York, NY, USA. ACM.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135, January.

Joonsuk Park, Sally Klingel, Claire Cardie, Mary Newhart, Cynthia Farina, and Joan-Josep Vallbé. 2012. Facilitative moderation for online participation in erulemaking. In *Proceedings of the 13th Annual International Conference on Digital Government Research*, dg.o '12, pages 173–182, New York, NY, USA. ACM.

John L. Pollock. 1987. Defeasible reasoning. *Cognitive Science*, 11:481–518.

Paul Rayson, Andrew Wilson, and Geoffrey Leech. 2001. Grammatical word class variation within the british national corpus sampler. Language and Computers.

Raquel Mochales Palau Rowe Glenn Reed, Chris and Marie-Francine Moens. 2008. Language resources for studying argument. In *Proceedings of the 6th conference on language resources and evaluation - LREC 2008*, pages 91–100. ELRA.

Ellen Riloff and Jay Shoen. 1995. Automatically acquiring conceptual patterns without an annotated corpus. In *In Proceedings of the Third Workshop on Very Large Corpora*, pages 148–161.

Sara Rosenthal and Kathleen McKeown. 2012. Detecting opinionated claims in online discussions. In *ICSC*, pages 30–37.

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*, CAAGET '10, pages 116–124, Stroudsburg, PA, USA. Association for Computational Linguistics.

Swapna Somasundaran, Josef Ruppenhofer, and Janyce Wiebe. 2007. Detecting arguing and sentiment in meetings. In *Proceedings of the SIGdial Workshop on Discourse and Dialogue*.

Stephen E. Toulmin, Richard Rieke, and Allan Janik. 1979. *An Introduction to Reasoning*. Macmillan Publishing Company.

S.E. Toulmin. 1958. *The Uses of Argument*. Cambridge University Press.

Maria Paz Garcia Villalba and Patrick Saint-Dizier. 2012. Some facets of argument mining for opinion analysis. In *COMMA*, pages 23–34.

Janyce Wiebe and Ellen Riloff. 2005. Creating subjective and objective sentence classifiers from unannotated texts. In *In CICLing2005*, pages 486–497.

Theresa Wilson and Janyce Wiebe. 2005. Annotating attributions and private states. In *Proceedings of ACL Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*.

Theresa Wilson. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *In Proceedings of HLT-EMNLP*, pages 347–354.

Adam Wyner, Raquel Mochales-Palau, Marie-Francine Moens, and David Milward. 2010. Semantic processing of legal texts. chapter Approaches to Text Mining Arguments from Legal Cases, pages 60–79. Springer-Verlag, Berlin, Heidelberg.