TakeLab

Laboratorij za analizu teksta i inženjerstvo znanja
Text Analysis and Knowledge Engineering Lab
Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva
Unska 3, 10000 Zagreb, Hrvatska

UNIVERSITY OF ZAGREB
FACULTY OF ELECTRICAL ENGINEERING AND COMPUTING

BACHELOR'S THESIS no. 5323

Claim and Stance Classification in Online Discussions Using Machine Learning

Toni Kukurin

Zagreb, June 2017

UNIVERSITY OF ZAGREB
FACULTY OF ELECTRICAL ENGINEERING AND COMPUTING
BACHELOR THESIS COMMITTEE

Zagreb, 3 March 2017

BACHELOR THESIS ASSIGNMENT No. 5323

Student: **Toni Kukurin (0036485224)**
Study: Computing
Module: Computer Science

Title: **Claim and Stance Classification in Online Discussions Using Machine Learning**

Description:

Online user comments are a valuable source of information for the analysis of opinions on events and their protagonists, political decisions, political subjects, ideological issues, contested topics, etc. Stance classification is a new research area at the intersection of sentiment analysis and argumentation mining, concerned with the automatic detection of stances expressed in text. The task is especially challenging in the context of user comments on the Internet, due to the informality and brevity of texts.

The topic of this thesis is the application of machine learning to claim and stanceclassification in online user comments in English. Do a literature study on short text classification, with an emphasis on stance classification. Devise and implement a model for claim classification based on their type within informal argumentation (fact, value, policy) and for claim stance classification with respect to a given topic (for, against, neutral). For model development and testing, use the dataset from (Hasan and Ng, 2014), and carry out an additional annotation round at the level of individual claims. Carry out a detailed evaluation of the model, comparison against a baseline, a statistical analysis of the results, and an error analysis. All references must be cited, and all source code, documentation, executables, and datasets must be provided with the thesis.

Issue date: 10 March 2017
Submission date: 9 June 2017

Mentor:

Associate Professor Jan Šnajder, PhD

Committee Chair:

Full Professor Siniša Srbljić, PhD

Committee Secretary:

Assistant Professor Tomislav Hrkać, PhD

Contents

1. Introduction

Online forums and various other types of question-answering websites often provide valuable information regarding a wide array of different topics. Controversial questions such as ones related to the existence of God and gay marriage particularly entice users to share their opinions and exchange beliefs by engaging in argumented discussions.

Such exchanges usually present two opposing stances, evolving naturally with new information as time goes by and a larger number of users start participating. We are interested in parsing this information in order to retrieve relevant arguments from the opposing sides and hopefully gain a better understanding of the topics being debated. Automatic retrieval and classification of such data is usually referred to as *argumentation* or *opinion* mining, and poses an interesting problem in a subfield of machine learning called Natural Language Processing (NLP). More formally, Palau and Moens [15] describe argumentation mining as a field of research which *aims to detect the arguments presented in a text document, the relations between them and the internal structure of each individual argument.*

By virtue of the ubiquity of online discussion platforms,[1] argumentation mining with specific focus on web debates has become an increasingly interesting topic in NLP research over the past few years. Classifying arguments from online sources proves to be a particularly challenging problem since the text often contains spelling errors, informal and domain-specific phrasing, as well as abysmal grammatical constructs.

In this thesis we will be focusing on the question of argument classification with emphasis on short online texts written in English; we intend to devise and implement a Machine Learning model for claim classification based on argument type within informal argumentation (namely, whether the argument can be classifed as fact, value or policy) and for claim stance classification with respect to a

[1] https://www.quora.com/, http://www.opposingviews.com, http://wiki.idebate.org, http://www.createdebate.com, to name a few

given topic.

The following chapter will provide a short overview of related work in the field. We will follow with a summary and exploratory evaluation of our dataset, focusing on corpora statistics and describing how the data is labeled. We continue with model and feature description, and ultimately provide evaluation results and our conclusions in the final two sections.

2. Related Work

Argumentation is a multidisciplinary research field, and so we can find relevant papers for this topic not only in computer science, but also in various other domains such as psychology and linguistics. While the former focuses mainly on automatic information retrieval and classification, the latter two disciplines approach argumentation from a more exploratory standpoint and often provide guidelines by which argumentation mining experiments are set-up. Our focus will be directed towards related work coming from the field of computer science.

Lippi and Toroni [12] recently surveyed the state of the art in the field, citing the problem at hand to be commonly phrased as a three-level pipeline consisting of (1) sentence and supporting claim detection, (2) argument link prediction and (3) inference of between-argument connections. While in principle a single experiment could be set up to implement all three steps of the sequence, current argumentation mining approaches are still very much in their early stages and researchers usually direct their focus only to a few of its select parts.

In related work we commonly identify four argumentation mining subtasks regarding different steps of the pipeline, and shortly describe them in the following sections. Namely, these are claim detection, discourse relation detection, stance prediction and argument recognition.

2.1. Claim Detection

While its definition is fairly controversial, an argument is in linguistics research commonly defined to minimally consist of a single premise (or evidence) and a conclusion (or claim) [15]. Claim detection tasks usually conduct their experiments with this definition in mind, and try to separate argumentative from non-argumentative text segments by searching for a model of this premise-to-conclusion pattern.

Levy et al. [10] were the first to propose the task of so-called *Context-*

Dependent Claim Detection (CDCD), where they define a topic as *a short phrase that frames the discussion*, and a *Context Dependent Claim* (CDC) as a *general, concise statement that directly supports or contests the given Topic*. This idea drives their research focusing on lexical and semantic features to detect document-wide argument boundaries given specific topic words. They label sentences from Wikipedia articles, which have since become a common resource for these types of tasks, noting a small percentage (around 30 among 300,000) of positive examples. Their proposed approach to the problem achieves a mean precision score of 12%.

Lippi and Torroni [11] follow their work by defining *Context-Independent Claim Detection* (CICD), and research the feasibility of extracting argumentative segments given a generic set of corpora. They propose an SVM-based method relying on tree kernel features to measure similarity between sentences, achieving comparable results over the same corpus (precision level of 10.5%).

2.2. Discourse Relation Detection

Discourse relation detection is tasked with discovering existing relations within a given corpus of text. The most accepted discourse theory is Rhetorical Structure Theory (RST), which defines twenty-three rhetorical relations [15]; distinctions are additionally made between the relations depending on whether they are bound by explicit discourse connectives (e.g., "since", "however"). Pitler et al. [18] assert that explicit relations are easy to identify, achieving 93% classification accuracy for the most general senses using a model which exclusively considers the discourse connective being used.

For implicit relation detection, early works by Marcu and Echibabi [13] exploit arguments with explicit connectives to train unsupervised machine learning models. They assume that implicit relations and explicitly signaled relations mostly only differ with regard to their inclusion of a discourse connective, and thus synthetically construct training samples for implicit arguments by simply deleting discourse connectives (for instance, *I like ice-cream because it's tasty* would be reduced to *I like ice-cream it's tasty*). They report extremely promising results, but Blair-Goldensohn et al. [2] subsequently refute their reports. It is discovered that real-world performance of models from [13] is not on-par with results achieved using synthetic data.

In contrast to Marcu and Echibabi's unsupervised attempt, Pitler et al. [17] try to discover implicit discourse relations using a supervised approach and eval-

uate different models on the Penn Discourse Treebank (PDTB). They test a wide array of features, including polarity tags, Levin verb classes, length of verb phrases, modality, context, and various lexical features. By analyzing features from both explicit and implicit texts, they confirm findings in [2] and conclude that annotated explicit relations are not as helpful feature sources as annotated implicit relations. Although varying levels of success are reported depending on relation type (up to 16% absolute improvement for some relations), the paper nevertheless provides valuable insight into the reasoning behind their feature choices and critical overview of previous work.

2.3. Stance Prediction

Given a topic and a text corpora, stance prediction is concerned with determining a particular text's stance with respect to the given topic. This task is usually approached as a supervised learning problem.

One of the earlier works from Somasundaran and Wiebe [19] shows that a classifier based exclusively on unigrams already makes a strong baseline for claim classification on ideological debate posts, and manages to only marginally outperform it using word lexicons and ngram features. Reported overall accuracies are 63.93% for their model against a 62.50% unigram baseline.

Anand et al [1] confirm these findings in subsequent experiments, but note that unigram scores they achieve appear significantly worse than Somasundaran and Wiebe's (yielding 54.00% accuracy). The discrepancies were attributed to a difference in their respective datasets (Somasundran and Wiebe's dataset is described as having a larger amount of rebuttal posts), indicating cross-corpora classification as a possibly interesting venture for future research.

The work presented in previous papers is in concordance with Pang and Lee's notes in their textbook on opinion mining [16], confirming the robustness of word independence assumptions for stance classification.

2.4. Argument Recognition

Boltužić and Šnajder [3] define the argument recognition problem to comprise of *matching user-created comments to a set of predefined topic-based arguments, which can be either attacked or supported in the comment.* They classify comment-to-argument pairs on a discrete five-level scale depending on whether a particular

comment attacks or defends a given argument, and its stance. They use the ComArg[1] corpus for this task, and report a 70.5% to 81.8% micro-averaged F1 score across different topics.

In a following related study [4], they investigate whether providing context in claim-matching tasks by including human-created implicit premises can bridge the gap between user claims and main claims in text, also reporting statistically significant improvements.

Similar to a combination of [3] and [4], Faulkner [6] researches document-level argument classification but tries to automatically bridge the comment-to-topic gap. They use a corpus composed of student essays paired up with seven different essay prompts, labeling each pair on a three-level scale (*for, against, neither*) and focusing on two main features for their model. First, using the MPQA lexicon of stance words and the Stanford Dependency Parser, they extract *POS-generalized subtrees* which indicate structure found in different types of arguments posed in the essays. Their second feature is Witten and Milne's Wikipedia Link-based Measure (WLM) score [20], which is calculated for each essay-prompt pair. They observe results in support of this approach, particularly noting the strength of WLM scores in bridging the gap between user comments and main claims. It is concluded that topic features can help capture subtle differences from similarly structured texts, as in their examples: "*Prisoners should be reintegrated into society*," and "*Prisoners should be punished*." (opposing topic words being *reintegrated* and *punished*). Using the aforementioned approach, Faulkner reports F1 scores around 82.0%, only slightly better than what was achieved in [3].

[1] http://takelab.fer.hr/data/comarg/

3. Data

We will use the same dataset as Hasan and Ng [7], which is made up of data obtained from an online debate on gay marriage. The corpus contains a total of 99 user-created posts which our annotator has manually processed, first into argumentative segments and then into their corresponding paraphrases. We hope to capture more complete contextual information this way and reduce noisiness in user arguments. An example of an extracted paraphrased segment follows (boldfaced part of the original text has been paraphrased):

Original text *Men can be attracted to Men.. BUT... Gay Marriage shouldn't be accepted.* ***Think about it, What will happen if Gays are accepted? There will be an increasing amount of Gays.*** *That will create population decrease everywhere, and that's going to lead to an end in the human population.*

Sentence paraphrase *Accepting gay people will lead to an increasing amount of gay people.*

This was done for each argumentative segment within a post, for a total of 920 paraphrased segments including some duplicates – removing the duplicates reduces the set to 876 uniques. The following two sections will provide a summary of the annotated data, first focusing on segment paraphrases and then shortly describing the original texts.

3.1. Segment paraphrases

Each paraphrased segment is annotated with type and stance information by the same annotator who extracted and interpreted the segments. We have also made an additional annotation pass through the (already paraphrased) data independently in order to gain a better understanding of it, as well as to obtain measures

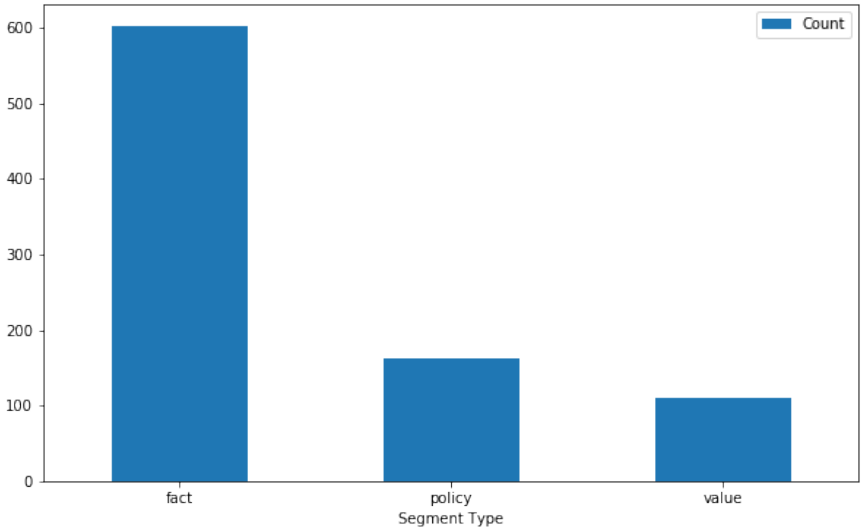

Figure 3.1: Histogram of segment type counts.

for inter-annotator agreement. Albeit agreement scores fall somewhat short of substantial (examined later in this chapter), we treat the original annotator's labels as the gold standard, and summary reports presented from here onward refer to information obtained on these labels.

3.1.1. Type classification

Arguments are in persuasive speech classified into three categories depending on whether they try to answer questions of fact, value, or policy. Their distributions in our dataset are visible in Figure 3.1, and examples for each of the segment types can be found in Table 3.1.

Segment paraphrase	Segment type
A country shouldn't deny marriage rights just because of the way of living.	policy
A child's character is not determined solely with having gay parents.	fact
Adopting a child is selfish if you are gay.	value

Table 3.1: Example segment type classification from the dataset. Classification was done by the thesis author and an independent annotator, giving a kappa score of 0.57.

As their name suggests, arguments of fact refer to factual information; they do not necessarily hold true and are not necessarily verifiable, but are considered

true from the viewholder's standpoint. These seem to be most common (in the particular examined dataset, at least), numbering at 602 (68.72% within the entire dataset).

Second most common stance type is policy, with 163 instances (18.6%). Policy questions refer to something which, according to the viewholder, *should* be enforced or, in other words, they specify a policy which is to be followed.

Finally, 111 (12.67%) value stance types are the ones which usually concern moral values and assert something which the viewholder deems right or wrong, worthy or unworthy. In the particular case of gay marriage, these are often related to either family values or Christianity (commonly manifested as references to God and parenting).

From Table 3.2 we observe that policy and value labels have fairly equal distributions of stances, in contrast to factual arguments which seem to be more often employed when presenting neutral information ($p < 0.00022$ under the chi-square test for independence when comparing frequencies of 5-point stance labels; policy and value arguments don't exhibit statistically significant differences). Intuitively, this might make sense considering that factual stance types cover a more broader range of arguments which are used for stating something in support of a main claim. For example, in the sentence "***Too many kids in the foster care program are being ignored in some states*** *because they wont let a gay couple take them in.*" the first segment (boldfaced) gives a neutral statement in support of the (pro-gay adoption) second segment.

	-2	-1	0	1	2
fact	29.57	11.96	14.62	8.80	35.05
policy	23.31	5.52	7.98	9.20	53.99
value	28.83	9.01	8.11	4.50	49.55

Table 3.2: Distribution of stances within different type labels (table values represent percentages).

As previously mentioned, we have made an additional pass through the sentence paraphrases and labeled them to determine the level of inter-annotator agreement. Cohen's Kappa score between the two annotations was 0.56, rendering the inter-annotator agreement somewhat weak by usual threshold levels. We consider the task of segment type classification to be somewhat subjective, so the

weak agreement score does not necessarily come as a big surprise. Some of the examples where we find disagreement are shown in Table 3.3

Segment paraphrase	Segment type
Being gay is a betrayal of Jesus.	value / fact
Children should have both a mother and a father.	policy / value
By natural means, infertility is wrong.	value / fact
Family values should be respected.	policy / value

Table 3.3: Some of the more controversial paraphrases with regard to type classification

3.1.2. Stance classification

Claim stances were annotated on a discrete scale from negative to positive two (yielding five classes, namely -2, -1, 0, 1 and 2), inferred from sentence paraphrases. Table 3.4 presents argument type percentages within our five stance classes; we can observe that per-stance distributions for the most part mimic ones observed across the entire corpus.

	fact	policy	value
-2	71.77	15.33	12.90
-1	79.12	9.89	10.99
0	80.0	11.82	8.18
1	72.60	20.55	6.85
2	59.60	24.86	15.54

Table 3.4: Distribution of type labels within different stances (table values represent percentages).

While sentence paraphrases contain more contextual information than original user sentences, their classification by stance nevertheless proves non-trivial in practice as we often find arguments relating to, e.g., child abuse and procreation as supporting claims to our main topic. In these cases the paraphrases were treated by subjectively trying to reason whether they are framed in a way which would result in a positive or negative stance towards gay marriage. As an example:

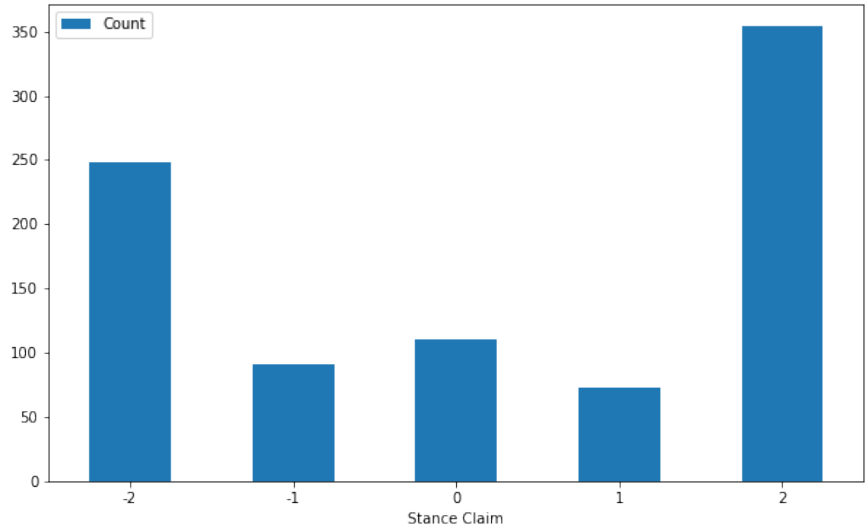

Figure 3.2: Stance claim count distribution.

Likely positive with regards to gay marriage *It is legal for a divorced parent to raise a child alone.*

Likely negative with regards to gay marriage *Having children is the only way to advance our culture and race.*

The first argument would probably reason that gay parents should be able to adopt children and, by extension, offer a mostly positive view on gay partnerships; the second, on the other hand, most likely tries to appeal to a procreation argument and thus invalidate gay relationships. Taking this into account, the first paraphrase would be classified as moderately positive (+1) while the second would be classified as moderately negative (-1). Table 3.5 contains some other paraphrases with their respective stance labels.

Segment paraphrase	Stance
Allowing gay marriage will bring us closer to true equality.	2
Being gay is a natural phenomenon.	1
A child needs proper upbringing to be a good adult.	0
By natural means, infertility is wrong.	-1
According to the Bible, gay marriage is unnatural.	-2

Table 3.5: Example stance claim classification from the dataset.

As in the previous section, we have made an additional pass through the

data independently from the original annotator and labeled segment paraphrases according to their stance. Since the graded classification is moderately subjective, for calculating inter-annotator agreement we consider two labels to be a match when their signs are equal (i.e., we separate labels into three groups: -2 and -1; 0; 1 and 2); such evaluation provided a weak kappa score of 0.52, and it remains to be seen whether this is an anomaly for the particular dataset (caused by e.g. many controversial statements) or if perhaps systematic errors in labeling caused the discrepancies.

3.2. Original text

As previously mentioned, the dataset provided contained a total of 99 posts. They were annotated on a binary scale with regards to their stance towards gay marriage (either for or against) by the post author, without any additional processing, as described in [7]. This data was fairly balanced, with 49 posts being against (labeled as -1) and 50 posts being for (labeled as +1) gay marriage.

Considering the fact that posts weren't preprocessed, we note that the data is somewhat noisy and can contain misspelled words, as well as poor grammatical constructs. Ubiquitous online slang such as usage of emoticons and all-uppercase typing, however, isn't that common. We present two representative instances of user posts:

Positive post *Marriage should be allowed regardless of the participants gender. The only things I can think of to disallow marriage are age and consent. Marriage is about the people getting married, strangers not approving in a marriage should not even be a factor to be considered. Also, a lot of people like to include religion in this argument. With so many different beliefs, I think it should be left out, but as a Christian, isn't it taught that God loves everyone? and that Jesus says not to judge? So, why shouldn't God approve, and who is one person to decide another person's orientation?*

Negative post *Well IF science 'says' that your fate is decided at birth then I'm going to say it how it is. Bullshit. Science only points to an increase likelyhood of being gay. Never 100% definately going to be a cocksucker. A lot of the general 'science' is flawed and sometimes doesn't even give cause. Sexuality isn't predetermined at birth. Anybody who says so is just trying to*

give gays a free pass. We should start doing that with criminals. Ohh wait they are.....

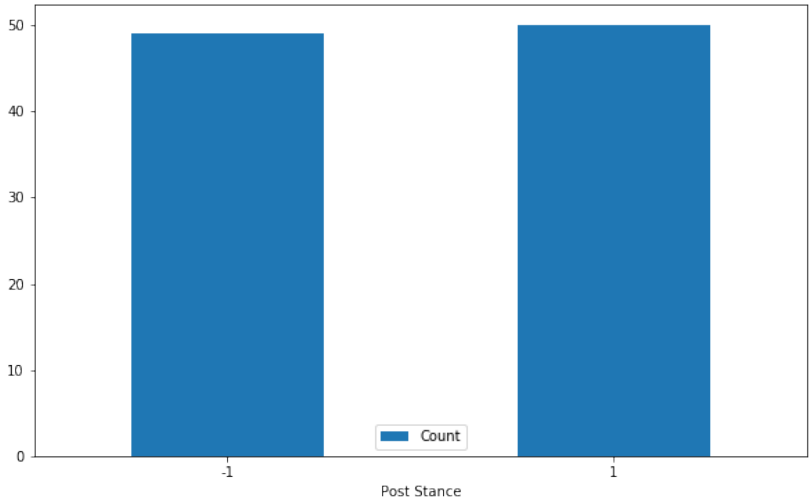

Figure 3.3: Post stance count distribution.

It might be interesting to note here that classifying a post simply by considering its individual segment stance labels (as marked by our annotator) might be a worthwhile pursuit. We've investigated this by classifying a post as positive if the sum of its individual segment labels is greater than zero and negative if the sum is less than zero. We treated cases where the sum equals zero as incorrect, since the classifier wouldn't be able to non-arbitrarily determine stance in these cases. This procedure provides correct post-level labels in 88 out of 99 total posts.

As one might expect, incorrectly classified posts are the ones indicating opposing opinions, for example the following user text where overall stance can be deemed positive for gay marriage in the legal sense, but against it in a more traditional sense:

"*In my opinion, marriage is between a man and a woman. If we are talking about marriage in the sense of a union recognized by the church, then absolutely not. A marriage in the church is for a man and a woman ONLY. HOWEVER! I do think that if two men, or two women want to be recognized legally as a couple, then they ought to have the right to a civil union. They would be legally recognized as married, but not recognized by the church. If they wanted to have a "ceremony" of sorts, that's fine, it just won't be recognized by the church. Legal marriage: yes. Holy marriage: no.*"

This is, however, still a fairly small sample size, and some topics more controversial than gay marriage might yield different success rates by relying on this classification process. Further research and a more diversified topic range might provide better insight into this matter.

4. Model

To our knowledge, type classification within informal argumentation hasn't yet been explored and this thesis would be the first to do so. Since a common approach in the stance prediction domain is to utilize a bag-of-words baseline model, we will be doing so for both stance and type classification. This is well aligned with previous work in argumentation mining, such as [19] and [16].

The following sections provide an overview of features we used in our experiments.

4.1. Preprocessing

We considered different standard preprocessing steps, and mention them in the evaluation section where they provide significant difference applied to a given model. For all models, we performed lowercasing and punctuation removal; other preprocessing steps considered on a case-by-case basis were stopword filtering, stemming and lemmatization.

The Natural Language Toolkit (NLTK)[1] software package for Python was tasked with these, using its implementations of Porter stemmer, WordNet lemmatizer and an English stopword collection counting 153 words.

4.2. Common features

Features described in this section are common to both stance and type classification. In the evaluation chapter, however, we only report the best-performing ones for each of the two classification tasks.

Segment-level metadata Inclusion of basic metadata such as average word counts, sentence length, punctuation counts and average word length in

[1] http://www.nltk.org/

some experiments proves to be fruitful, especially with regards to online comments which many times tend to express emotion via for instance repeated punctuation marks (e.g. '...', ':)', multiple sequential exclamation points, and similar). We tried including these as a "cheap" way of potentially observing some improvements.

Ngrams In a similar fashion to unigram counts, 1 to 3-grams were tested in hopes of capturing some common phrase dependencies among the different classes.

Skipgrams Unlike ngrams, skipgrams allow for gaps between word tuples in text. The reasoning behind trying out this approach is the same as for ngrams.

Word2Vec Using a model trained on Google News' corpus [14],[2] we simply sum each individual word's vectors elementwise and append this sum as a feature to our model. The word-embedded values are represented here as a vector of 300 real numbers.

Stanford Universal Dependencies We computed the Universal Dependencies and phrase structure tree representations of both paraphrased and their corresponding original text segments using the well-known Stanford Parser [5].[3] A few different features were constructed from data transformed this way:

Dependency type counter Hypothesizing that some stance and type information might be found from the dependencies, a basic 'bag-of-words' similar to the one applied in our baseline was tested on Universal Dependency types.

Composite Back-off Features Following [9], we also computed Composite Back-off Features (`h_bo` variant from the paper), transforming tagged items so that each head word gets replaced by its POS tag and modifier remains as-is. When classifying posts, word dependencies obtained from original sentences were appended as features for each individual user text.

[2]https://code.google.com/archive/p/word2vec/

[3]https://nlp.stanford.edu/software/lex-parser.shtml

4.3. Post-level specific features

In general, when applying the transformations mentioned above, the same process has been used for posts and segments alike. Two feature sets were however tried out exclusively on original user texts for stance classification, namely oracle and predicted values from the post's segment-level constituents:

Oracle labels from segment-level classification We were interested in exploring whether, as our preliminary dataset analysis suggests, post-level results can be improved by including segment-level labels as features. We've used two approaches from these values:

Average label value The first approach is similar to the simplistic idea our data analysis in Section 3.2 uses; we extract all segment stances and use their average as the single model feature.

Label sequence This model is constructed by appending all (type, stance) pairs to form a feature vector, padding and performing imputation on values where necessary so as not to accidentally introduce additional bias to the data.

Predicted labels for segment-level classification This was done in a similar vein to the above, only difference being that we re-run the experiment using our two best classifier results in place of oracle labels.

5. Evaluation

All models were evaluated using the SKLearn[1] software package and its default implementation of the Support Vector Machine [8] classifier. Namely, the classifier employs an RBF kernel and a one-v-one multiclass classification strategy.

Standard macro-averaged F1-score was used as the evaluation metric for all segment-level models. Post-level evaluation uses binary-averaged F1, considering the fact there are only two possible post-level stances. We performed 5-fold cross validation with oversampled training data on sentence paraphrases, original sentences, as well as original user posts (for the stance classification task). Test sets were not balanced, but rather constructed to represent the actual distribution of our dataset. The following sections report our findings.

5.1. Type classification

Table 5.1 shows our results for type classification. We have three baselines, namely a majority classifier (*Majority*), a binary bag-of-words classifier (*Baseline*) and a limited-dictionary binary bag-of-words classifier (*Baseline-L*).

The computed baselines perform very well for type classification on both fronts, although we observe that segment paraphrases give far better results than original segment text. This is most likely related to the reduced noisiness of the data, as well as additional contextual information ingrained within segment paraphrases.

It should also be mentioned that even small differences in user texts and paraphrases tend to have an effect on classifier output. To illustrate this point, similar segment-paraphrase pairs such as "*and feelings can't be controlled.*" (segment) and "*Feelings can't be controlled.*" (segment paraphrase) get different labels assigned. This, as well as performance degradation upon excluding stopwords, indicates fairly strong influence of non-topic words on classifier output.

[1] http://scikit-learn.org/

	Segment paraphrase	Segment text
Majority	0.27	0.27
Baseline	0.58	0.50
Baseline-L	**0.66**	0.50
W2V	0.63	**0.58**
W2V-NoStop	0.48	0.50
W2V-Should	0.63	0.55
CBF	0.42	0.39
UD-C	0.45	0.40

Table 5.1: Evaluation results for classification by type (mean values of 5-fold cross-validated F1 scores). Majority classifier (labeling all instances as 'fact') is included as an additional reference point.

We observe that policy and factual arguments in all tests tend to have a relatively high precision. Arguments questioning value are on the other hand characterized by poorer precision scores, mostly at the expense of recall for factual arguments. Table 5.2 displays a representative confusion matrix for the task.

A classifier limited to only the 70 most common words from the train set (*Baseline-L*) shows significant improvements over baseline methods on segment paraphrases, however it is questionable whether this is just a peculiarity of our dataset or if any relevant conclusions can be made from the fact. For now, we make an educated guess in assuming that the less-frequent words are among our topic words, which are often discussed using different argument types within a single user post (e.g. the commenter might provide a fact to establish credibility, followed by a policy he believes should be followed – segment paraphrases will in those cases often contain overlapping words).

Hoping to further examine the importance of topic-level words for our stance classification task, we tested a word2vec model with and without stopwords (*W2V* and *W2V-NoStop*, respectively) and observed significant result degradation for the second case. Some cue lexemes such as '*should*' were noted to be particularly common for distinguishing segment types, especially within segment paraphrases. This is reinforced by observing that *W2V-Should*, which removes all stopwords except for *should*, gives results almost identical to the "raw" word2vec.

Other than word-based models, we also tested methods relying on counting Composite Back-off Features (*CBF*) and Universal Dependencies (*UD-C*), in-

	Fact	Policy	Value
Fact	79	7	34
Policy	12	16	14
Value	8	2	12

Table 5.2: Representative confusion matrix for a baseline segment paraphrase classification (rows represent actual values while columns are our classifier predictions)

	Fact	Policy	Value
Fact	111	0	9
Policy	23	9	0
Value	21	0	1

Table 5.3: Representative confusion matrix for a segment paraphrase classification using Composite Back-off Features (rows represent actual values while columns are our classifier predictions)

tending to confirm whether some common structure exists in different argument types. Using these specific methods, we haven't achieved satisfactory results, which is best illustrated by analyzing confusion matrix 5.3. We observe behavior close to a majority classifier, accounting for a very weak recall on policy and value argument types.

5.2. Stance classification

Due to the fact that we have a fairly high number of inter-annotator disagreements regarding stance, we carry out evaluation of segment paraphrases and segment text only on the examples with matching stance annotation signs. We further limit both the train and test data to only strong and neutral stances (i.e., labels -2, 0 and 2) reducing our dataset to 519 samples of segment paraphrases and original segments. Evaluation is otherwise performed as explained at the beginning of this chapter. We will first focus on segment-level stance classification in this section (summarized in Table 5.4), and then transition into reporting evaluation results for post-level stance classification (summarized in Table 5.5).

A simple bag-of-words approach offers a fair improvement over the majority classifier in segment-level evaluation. Segment texts have a worse F1 score than

	Segment paraphrase	Segment text
Majority	0.22	0.22
Baseline	0.46	0.41
Ngrams	0.51	0.39
W2V	**0.57**	**0.46**
CBF	0.45	0.37
UD-C	0.41	0.38

Table 5.4: Sentence-level results for classification by stance (mean values of 5-fold cross-validated F1 scores). Majority classifier (labeling all instances as positive, i.e. +2) is included as an additional reference point.

paraphrases, and we hypothesize this is again likely related to the lack of contextual information. As an example, consider the segment text "*Just read A Child's Machiavelli, everything rings true.*" and its corresponding paraphrase "*According to A Child's Machiavelli, kids are extremely brutal.*" It is clear that the latter provides additional context (*kids are extremely brutal*) necessary to obtain any kind of interesting information for our use-case from a bag-of-words standpoint. To take an even more drastic example from the dataset into account, it is virtually impossible even for humans to classify original segment text of "*Yes!!*" without additional context while its corresponding paraphrase "*Gay marriage should be allowed.*" makes for a relatively easy classification target.

Using Composite Back-off Features (*CBF*) gives results similar to the baseline, though it should be noted that we hypothesize they could be a more robust cross-domain solution. The same thing is observed when counting Universal Dependencies (*UD-C*) for segment classification. We unfortunately cannot confirm our suspicions without obtaining additional data, and therefore leave it for a future experiment. Using stemming with 1-3-grams (*Ngrams*) appears to be effective in capturing some additional common context from segment paraphrases, and improves the F-score by a few points. We observed marginally worse performance utilizing skipgrams (for gap sizes and number of tokens in range 1-3) and decided against including them in the evaluation table.

Finally, the best results on stance classification overall were observed with word embeddings, where the word2vec model outperformed all others by a relatively large margin. Unlike in our tests regarding type, removing stopwords here does not change classifier output by a lot which goes in concordance with our

intuition that topical words provide much more useful information for this task.

Post-level results for models described above, along with four additional post-specific models, are presented in Table 5.5. The additional model types were tested in order to explore the robustness of document classification from its argumentative segment constituents, and include evaluations on our oracle ("gold-standard") labels, as well as on labels predicted using our best segment-level classifiers for both paraphrases and original segments. The averaging model (*Average*) with oracle features gives best results among all others, reaching a mean 0.91 F1 binary-averaged score, and is also strong when averaging predicted stances. Interestingly enough, the variants using predicted features show results identical to a post-level word2vec model, they themselves having been constructed by classifying a segment-level word2vec model.

As suggested by our preliminary dataset analysis, it also does not come as a surprise that the second-best model turns out to be the one constructed by appending oracle stance label sequences (*ST*), for which we achieve a 0.90 mean F1 binary-averaged score. Although the sequence of type labels alone (*FVP*) does not perform comparably (as observed from F1 scores of 0.37, 0.26 and 0.34 for paraphrases, original segments and oracle label values, respectively), it is interesting to note that evaluating predicted stance and type labels jointly (*ST+FVP*) shows some improvements over stance-only models, weakly hinting at a possible positive interaction of biases introduced by interactions of the two given models.

Following the previous observation, we hypothesize that conjoining segment-based features (*FVP+ST*) together with our previous best post-level model (*W2V*) might also display positive interaction with regards to post-level stance classification. While in our tests a combined model of the two (*W2V+FVP+ST*) does not manage to achieve significant improvements over a word-embeddings-only approach, it does perform marginally better and possibly suggests that obtaining more accurate segment-level predictions would give fruitful ground for future research in the domain of post-level stance classification.

	Post-level score		
Majority	0.67		
Baseline	0.55		
Ngrams	0.55		
W2V	**0.73**		
CBF	0.46		
UD-C	0.44		
	Paraphrases	*Original*	*Oracle*
Average	0.73	0.71	**0.91**
ST	0.70	0.68	0.90
FVP	0.37	0.26	0.34
ST+FVP	0.71	0.69	0.88
W2V+FVP+ST	0.75	0.75	0.77

Table 5.5: Post-level results for classification by stance (5-fold means of binary-averaged F1). *Paraphrases* and *Original* represent results obtained by running our best segment-level classifiers on paraphrased and original segments, respectively. Majority classifier (labeling all instances as positive, i.e. +1) is included as an additional reference point.

6. Conclusion

In this thesis, we have studied the application of machine learning to claim and stance classification in online user comments in English. Two main subtasks were approached separately, namely the classification of comments by stance and the classification of comments' argumentative constituents by both stance and type within informal argumentation (fact, value or policy).

We first devised models targeting text segments, and found that reducing noisiness of the data by classifying on paraphrased rather than original content is effective in improving results both for stance, as well as type classification tasks. We have also confirmed assertions from previous work in the domain regarding the robustness of a word independence assumption for classifying short, argumentative text segments.

The second task involving post-level classification was approached twofold; we first tackled it by simply employing previously devised segment-level models on the entire post content, managing to achieve results which slightly improve upon a majority baseline. Our second approach is a bit more interesting insofar that we combined segment-level information in order to obtain post-level stance predictions. While statistically significant improvements were not observed except with oracle data, we do regard the results promising and note that exploring post-level classification by accumulating segment-level results might be an interesting venture for future research.

We further propose that it would likely be interesting to investigate how the approaches studied in this thesis generalize to different domains and whether the statistics obtained in our dataset can also be observed in other corpora. Considering our fairly weak kappa-scores, running the experiments on data re-labeled by multiple annotators could provide valuable insight into how well humans actually agree when it comes to classification of short argumentative sentence paraphrases, and whether different results can be expected on data with higher levels of inter-annotator agreement.

References

[1] Pranav Anand, Marilyn Walker, Rob Abbott, Jean E Fox Tree, Robeson Bowmani, and Michael Minor. Cats rule and dogs drool!: Classifying stance in online debate. In *Proceedings of the 2nd workshop on computational approaches to subjectivity and sentiment analysis*, pages 1–9. Association for Computational Linguistics, 2011.

[2] Sasha Blair-Goldensohn, Kathleen McKeown, and Owen Rambow. Building and refining rhetorical-semantic relation models. In *HLT-NAACL*, pages 428–435, 2007.

[3] Filip Boltužić and Jan Šnajder. Back up your stance: Recognizing arguments in online discussions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 49–58. Citeseer, 2014.

[4] Filip Boltužić and Jan Šnajder. Fill the gap! analyzing implicit premises between claims from online debates. *ACL 2016*, page 124, 2016.

[5] Danqi Chen and Christopher D Manning. A fast and accurate dependency parser using neural networks. In *EMNLP*, pages 740–750, 2014.

[6] Adam Faulkner. Automated classification of stance in student essays: An approach using stance target information and the wikipedia link-based measure. *Science*, 376(12):86, 2014.

[7] Kazi Saidul Hasan and Vincent Ng. Why are you taking this stance? identifying and classifying reasons in ideological debates. In *EMNLP*, pages 751–762, 2014.

[8] Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. Support vector machines. *IEEE Intelligent Systems and their Applications*, 13(4):18–28, 1998.

[9] Mahesh Joshi and Carolyn Penstein-Rosé. Generalizing dependency features for opinion mining. In *Proceedings of the ACL-IJCNLP 2009 conference short papers*, pages 313–316. Association for Computational Linguistics, 2009.

[10] Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. Context dependent claim detection. 2014.

[11] Marco Lippi and Paolo Torroni. Context-independent claim detection for argument mining. In *IJCAI*, volume 15, pages 185–191, 2015.

[12] Marco Lippi and Paolo Torroni. Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology (TOIT)*, 16(2):10, 2016.

[13] Daniel Marcu and Abdessamad Echihabi. An unsupervised approach to recognizing discourse relations. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 368–375. Association for Computational Linguistics, 2002.

[14] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[15] Raquel Mochales Palau and Marie-Francine Moens. Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the 12th international conference on artificial intelligence and law*, pages 98–107. ACM, 2009.

[16] Bo Pang, Lillian Lee, et al. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135, 2008.

[17] Emily Pitler, Annie Louis, and Ani Nenkova. Automatic sense prediction for implicit discourse relations in text. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 683–691. Association for Computational Linguistics, 2009.

[18] Emily Pitler and Ani Nenkova. Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 13–16. Association for Computational Linguistics, 2009.

[19] Swapna Somasundaran and Janyce Wiebe. Recognizing stances in ideological on-line debates. In *Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, page 116, 2010.

[20] Ian Witten and David Milne. An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In *Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy, AAAI Press, Chicago, USA*, pages 25–30, 2008.

Klasifikacija tvrdnji i stavova na internetskim raspravama pomoću strojnog učenja

Sažetak

Forumi i razne ostale web-stranice s mogućnošću korisničke interakcije često pružaju vrijedne informacije o širokom spektru zanimljivih tema te nam je iz tog razloga zanimljiva automatska analiza takvih izvora podataka. Polazeći iz te ideje, u ovom radu implementiramo sustav strojnog učenja za klasifikaciju stava i podjelu argumenata prema tipu za kratke *online* tekstove pisane engleskim jezikom. Koristimo pristup nadziranog učenja nad ručno označenim podacima, parafrazama argumentativnih segmenata korisničkih tekstova, kao i neprerađenim korisničkim tekstovima. Vrednujemo i analiziramo implementirane sustave, izlažemo naše zaključke o prikupljenim rezultatima te pružamo smjernice za daljnji rad unutar dane domene.

Ključne riječi: prirodna obrada jezika, strojno učenje, analiza argumentacije, klasifikacija stavova, klasifikacija tvrdnji

Claim and Stance Classification in Online Discussions Using Machine Learning

Abstract

Online forums and various other types of question-answering websites often provide valuable information regarding a wide array of different topics, thereby making automatic analysis of the information from these sources an interesting research task. Building on the previous statement, in this thesis we devise and implement a Machine Learning model for classifying stance and type in short online user comments written in English. This is done by employing a supervised learning approach on manually-extracted argumentative text segments and their paraphrases, as well as original user texts. We evaluate and analyze our models, provide conclusions on obtained results, and offer suggestions for future exploratory work to be performed in the given domain.

Keywords: natural language processing, machine learning, argumentation mining, stance detection, claim classification