PolariCQ: Polarity Classification of Political Quotations

Rawia Awadallah
Max-Planck Institute for Informatics
Saarbruecken, Germany
rawadall@mpi-inf.mpg.de

Maya Ramanath
Indian Institute of Technology, Delhi
New Delhi, India
ramanath@cse.iitd.ac.in

Gerhard Weikum
Max-Planck Institute for Informatics
Saarbruecken, Germany
weikum@mpi-inf.mpg.de

ABSTRACT

We consider the problem of automatically classifying quotations about political debates into both topic and polarity. These quotations typically appear in news media and online forums. Our approach maps quotations onto one or more topics in a category system of political debates, containing more than a thousand fine-grained topics. To overcome the difficulty that pro/con classification faces due to the brevity of quotations and sparseness of features, we have devised a model of quotation expansion that harnesses antonyms from thesauri like WordNet. We developed a suite of statistical language models, judiciously customized to our settings, and use these to define similarity measures for unsupervised or supervised classifications. Experiments show the effectiveness of our method.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Linguistic Processing*

General Terms

Algorithms, Experimentation

Keywords

Web Information Extraction, Political Opinion Mining

1. INTRODUCTION

Political controversial topics such as “Greece bailout”, or “Arab Spring revolts” are discussed in great depth in political debates, newspapers, and other forms of social media. In contrast to standard approaches for sentiment analysis on products (cameras, movies, etc.) [10, 7], the sophisticated nature of political opinions calls for more advanced linguistic techniques. A typical task considers a number of controversial topics such as “immigration” or “abortion” and a set of stakeholders like politicians, and aims to classify the stakeholders’ opinions into pro or con categories for the respective topics [2]. Such analyses are typically based on aggregating many statements and work well for coarse-grained topics. However, political analysts are often interested in individual and brief statements, as quoted in news media, and their *pro/con polarity* with regard to *fine-grained debates* such as “deporting illegal immigrants” or “immigration amnesty”. *Quotations* are a prominent form of highlighting opinions in newspapers, and all kinds of social media.

The problem addressed in this paper is to automatically classify quotations onto fine-grained topics of controversial nature, and to assign a pro/con polarity for each quotation-topic pair. Note that a quotation, given as a short text and the quoted person or party, can refer to several topics. Table 1 shows an example with quotations on immigration. We are interested in identifying the pro/con stances in the inputs, not just on the broad issue of immigration but on fine-grained topics (e.g. ***con*** Mexican border fence, ***con*** deporting illegal immigrants), for quotation 1). The problem is much more demanding than traditional forms of sentiment mining for various reasons: **(1)** *fine granularity:* the number of debated topics that the classifier must consider is potentially very high, in the thousands rather than the usual tens; **(2)** *brevity of statements:* quotations are usually very short texts (often only a single sentence), so that whatever linguistic features are used tend to be sparse; **(3)** *topic-dependent polarity:* the same quotation can have different polarities for different fine-grained topics.

Our approach maps quotations onto one or more topics in a category system of political debates. We use the fine-grained categories (called *debates*) of `debatepedia.org` as our target. For each pair of quotation and relevant debate, our classifier targets, the debates in debatepedia, come with articles and user discussions, and computes a pro/con polarity. For example, for the input texts in Table 1, the output of our method is the topic/polarity pairs shown at the bottom of each quotation. We define language models (LMs), with judiciously chosen features (including bigrams) for each debate, and then use a scoring function based on the query likelihood as a similarity measure that is fed into different kinds of unsupervised or supervised classifiers (kNN, SVM, LDA). The high sparsity in the debates themselves is addressed by smoothing the debates’ LMs via thematically related debates. To overcome the difficulty caused by the brevity of quotations and sparseness of features, we have devised a method of *quotation expansion* that harnesses thesauri like WordNet. We use synonyms and antonyms (i.e., words for opposite senses, e.g., “censorship” or “regulation” as antonyms of words like “neutrality” or “freedom”) to conceptually expand the text of a quotation. This approach leads to a novel form of enriched feature space: a quotation is then represented by an expanded entailment/contradiction language model.

In Section 2, different features models are introduced. The quotation expansion is described in Section 3. Details about the pro-

CIKM’12, October 29–November 2, 2012, Maui, HI, USA.
Copyright 2012 ACM 978-1-4503-1156-4/12/10 ...$10.00.

Table 1: Quotations about immigration.

Quotation (1)*:** “The alternatives to the Specter bill are senseless. The enforcement-only approach – building a 700-mile wall and engaging in a campaign of mass deportation and harassment to rip 12 million people from the national fabric would destroy families and weaken the economy.” ***con Mexican border fence , ***con*** deporting illegal immigrants
Quotation (2)*:** “I think that the wall could help with the economy” ***pro Mexican border fence
Quotation (3)*:** “The fence should be finished, but that mass deportations are not the answer. Until we build that border, we should neither have storm troopers come in and throw people out of the country nor should we provide amnesty” ***pro Mexican border fence, ***con*** deporting illegal immigrants

posed LM based models for topic and pro/con classification are given in Section 4. Section 5 presents our experimental evaluation. Section 6 positions our contributions with regard to related work. We conclude with Section 7.

2. FEATURES MODEL

Topic and Sentiment Terms. A **topic term** is a term which describes a topic, while a **sentiment term** is one which describes an opinion. We assume nouns to be topic terms, while verbs, adjectives and adverbs are sentiment terms. For example, quotation (1) in Table 1 has the topic terms “wall”, “campaign”, “economy”, “deportation”, etc., while its sentiment terms are “destroy”, “weaken”, “senseless”, etc.

Unary and Binary Features. We define unary features and binary features as follows.

DEFINITION 2.1. *A* **unary feature** *is denoted as* $\langle u \rangle$ *where* u *is either a topic term or a sentiment term.*

For example, quotation (1) in Table 1 has the unary features: ⟨wall⟩, ⟨economy⟩, ⟨deportation⟩, ⟨destroy⟩, ⟨weaken⟩, etc.

DEFINITION 2.2. *For a given quotation* $\mathcal{Q}$*, let* $\mathcal{Q}^T$ *and* $\mathcal{Q}^S$ *denote the set of its topic terms and sentiment terms respectively. A* **binary feature***, denoted* $\langle t, s \rangle$*, consists of* $t \in \mathcal{Q}^T$ *and* $s \in \mathcal{Q}^S$*, such that,* t *and* s *are connected by a dependency relation. The dependency relation is determined by parsing the sentence in* $\mathcal{Q}$ *in which they co-occur using a dependency parser* `nlp.stanford.edu`*.*

For example, quotation (1) has as binary features: ⟨wall, weaken⟩, ⟨deportation, destroy⟩, ⟨economy,weaken⟩, etc.

3. QUOTATION EXPANSION

Our approach of pro and con classification is built on the intuition that opinions which are in agreement with each other have expressions which are in agreement to each other, while opinions which disagree have expressions which are in disagreement. For example in Table 1 quotation (1) which has the expression “…and weaken the economy.” is in *disagreement* with the expression in quotation (2) “…could help with the economy”, while the expression “…mass deportation is not the answer” in quotation (3) is in *agreement* with the expression “engaging in a campaign of mass deportation…would destroy families” in quotation (1).

In order to capture the notion of *agreement* and *disagreement* for a given quotation, we focus specifically on the binary features of the quotation. That is, the topic and sentiment term pair $\langle t, s \rangle$ which are in a dependency relationship with each other. The key idea that we propose is to expand the topic term and the sentiment term with both their synonyms as well as their antonyms (see Table 2). For this expansion, we use the WordNet thesaurus, which gives synonyms and antonyms for many concepts. In order to map a word observed in a quotation onto its proper word sense, that is, the WordNet concept denoted by the potentially ambiguous word, we use the most-common-sense heuristics which has been used effectively in many applications [11].

Table 2: Agreement/disagreement with original features

Quotation: “The *{wall}* could *[help]* with the *{economy}*”	
Topic terms	wall, economy
Topic term synonyms	{fence, border}, {saving}
Topic term antonyms	{ }, {spending, expend}
Sentiment terms	help
Sentiment term synonyms	{help, assist, support}
Sentiment term antonyms	{destroy, weaken, not help}
Original feature	(economy, help)
Features in agreement	(economy, assist), (economy, support)
Features in disagreement	(economy, destroy), (spending, support)

Let t^+, t^- denote a synonym and antonym of a topic term respectively. Analogously, s^+ and s^- denote a synonym and antonym of a sentiment term. The possible expansions of a binary feature $\langle t, s \rangle$ are the pairs $\langle t^+, s^+ \rangle$, $\langle t^-, s^- \rangle$, $\langle t^-, s^+ \rangle$, $\langle t^+, s^- \rangle$. The first two are in agreement with the original feature $\langle t, s \rangle$, while the last two are in disagreement. As an example, consider the quotation in Table 2 and the binary feature ⟨economy, help⟩. A synonym for the topic term “economy” is “saving” while its antonyms could include “spending” and “expend”. Similarly, for the sentiment term “help”, synonyms include “support” and “assist”. while it’s antonyms are “destroy”, “weaken”. Therefore, the expanded binary features include ⟨economy, weaken⟩ ($\langle t^+, s^- \rangle$), which is in *disagreement* with the original feature, as well as ⟨saving, assist⟩ ($\langle t^+, s^+ \rangle$), which is in *agreement* with the original feature. Table 2 shows more examples of these binary features.

DEFINITION 3.1. *For a topic term* t_i*, the set of* ***topic term synonyms*** *is denoted as* $\mathcal{T}_i$*, and the set of* ***topic term antonyms*** *is denoted as* $\bar{\mathcal{T}}_i$*. Analogously,* $\mathcal{S}_i$ *and* $\bar{\mathcal{S}}_i$ *denote the sentiment term synonyms and antonyms of a* ***sentiment term*** s_i*, respectively.*

DEFINITION 3.2. *For a given binary feature* $\langle t_i, s_i \rangle$ *present in the quotation, we define the set of* **agreement features** *as* $AF = \{\langle t'_i, s'_i \rangle | t'_i \in \mathcal{T}_i, s'_i \in \mathcal{S}_i\} \cup \{\langle t'_i, s'_i \rangle | t'_i \in \bar{\mathcal{T}}_i, s'_i \in \bar{\mathcal{S}}_i\}$

DEFINITION 3.3. *For a given binary feature* $\langle t_i, s_i \rangle$ *present in the quotation, we define the set of* **disagreement features** *as* $DF = \{\langle t'_i, s'_i \rangle | t'_i \in \mathcal{T}_i, s'_i \in \bar{\mathcal{S}}_i\} \cup \{\langle t'_i, s'_i \rangle | t'_i \in \bar{\mathcal{T}}_i, s'_i \in \mathcal{S}_i\}$

4. TOPIC & SENTIMENT CLASSIFICATION

We first use topic features to map a quotation onto one or more debates. Then, for each of the identified debates, we use the joint topic-sentiment unary and binary features, optionally with expansion, for inferring the pro/con polarity of the quotation.

4.1 Mapping Quotation to Topics

We estimate an LM for each debate with unary topic features as terms and then compute the query likelihood. Let $P_{\mathcal{D}}$ denote ***the language model of a debate*** $\mathcal{D}$. Then,

$$P_{\mathcal{D}}(w) = (1 - \lambda)P(w|\mathcal{D}) + \lambda P(w|\mathcal{C}_{\mathcal{D}})$$

where $\mathcal{D}$ is a debate on a fine-grained topic, w is a topic term, and $\mathcal{C}_{\mathcal{D}}$ is the set of debates belonging to the same category as $\mathcal{D}$ in Debatepedia. Let $score(\mathcal{D}) = P(\mathcal{Q}|\mathcal{D})$ denote the probability that $\mathcal{D}$ generates quotation $\mathcal{Q}$. Then,

$$P(\mathcal{Q}|\mathcal{D}) = \prod_{w_i \in \mathcal{Q}^T} P(w_i|\mathcal{D})$$

where $\mathcal{Q}^T$ is the set of topic terms in $\mathcal{Q}$. This is the ***the quotation likelihood*** of $\mathcal{Q}$ with respect to $\mathcal{D}$. The set of topics for $\mathcal{Q}$ is now $\mathcal{Q}^{\mathcal{D}} = \{\mathcal{D}|score(\mathcal{D}) > \sigma\}$, where σ is a threshold (in our experiments, $\sigma = 0.01$, with an average of 3 topics per quotation).

4.2 Pro/Con Classification

Once we have a set of topics $\mathcal{Q}^{\mathcal{D}}$ for the given quotation $\mathcal{Q}$, our task is to classify the polarity of $\mathcal{Q}$ on each $\mathcal{D} \in \mathcal{Q}^{\mathcal{D}}$. For every debate in Debatepedia, there is a set of pro documents and a set of con document. For a debate $\mathcal{D} \in \mathcal{Q}^{\mathcal{D}}$, we define $\mathcal{D}^+$ as the concatenation of *all* pro documents, and $\mathcal{D}^-$ as the concatenation of all con documents for that debate. Given a pro and a con document for $\mathcal{D}$, we compute the quotation likelihoods: $P(\mathcal{Q}|\mathcal{D}^+)$ and $P(\mathcal{Q}|\mathcal{D}^-)$. If $(P(\mathcal{Q}|\mathcal{D}^+) > P(\mathcal{Q}|\mathcal{D}^-))$, the quotation is classified as pro, otherwise, we classify it as con. In effect, this is a kNN classifier (k nearest neighbors) with $k = 1$ in our 2-class settings. We estimate the quotation likelihood with respect to $\mathcal{D}^+$ as follows.

$$P(\mathcal{Q}|\mathcal{D}^+) = \prod_{w_i \in \mathcal{Q}} P(w_i|\mathcal{D}^+)$$

Analogously, we also estimate the quotation likelihood with respect to $\mathcal{D}^-$. The query is represented as a set of features where each feature is denoted as w_i. While we can use the terms in the quotation and debates as is, this is unlikely to give us good results (as we show in our experiments) because of the sparsity of terms in the quotation. To overcome this problem, we make use of our features model for the estimation of the language models of the debates.

LM over n-grams. This model denoted **LM-NG**, uses n-grams as features to represent the queries and the documents. For a given debate $\mathcal{D}$, we estimate the language model of $\mathcal{D}^+$ over all possible n-grams, where $n \leq 3$ as follows.

$$P_{\mathcal{D}^+}(w) = (1-\lambda)P(w|\mathcal{D}^+) + \lambda P(w|\mathcal{C}_D)$$

where w is an n-gram and $\mathcal{C}_D$ is the background corpus consisting of all debates in the same branch of Debatepedia. We estimate $P(w|\mathcal{D}^+)$ as $\frac{\#(w;\mathcal{D}^+)}{\Sigma_i \#(w_i;\mathcal{D}^+)}$. $P(w|C_D)$ is estimated in a similar manner. The LM for $\mathcal{D}^-$is estimated analogously. Finally, the quotation likelihoods given $\mathcal{D}^+$ or $\mathcal{D}^-$ are computed.

LM over unary features. This model is denoted as **LM-UNA**. We estimate the LM of a document as a mixture model of two LMs, one considers topic terms and the other considers sentiment terms. The topic terms and sentiment terms together form the unary features. The language model of $\mathcal{D}^+$ is estimated as:

$$P_{\mathcal{D}^+}(w) = \alpha P_{\mathcal{D}^+_T}(w) + (1-\alpha)P_{\mathcal{D}^+_S}(w)$$

where w is a unary feature, $P_{\mathcal{D}^+_T}(w)$ is the probability of w in the topic LM of $\mathcal{D}^+$, and $P_{\mathcal{D}^+_S}(w)$ is the probability of w in the sentiment LM of $\mathcal{D}^+$, and α is a parameter which determines the importance of each. The topic LM and sentiment LM are estimated as before, with the universe of terms consisting of topic terms and sentiment terms, respectively. Analogously, we estimate $\mathcal{D}^-$ LM.

LMs with binary and unary features. Recall that binary features are $\langle t, s\rangle$ pairs, where t is a topic term and s is a sentiment term and the two are in a parse dependency relationship. We now estimate the LM of $\mathcal{D}^+$ and $\mathcal{D}^-$ over both unary and binary features.

$$P_{\mathcal{D}^+}(w) = \beta P_{\mathcal{D}^+_U} + (1-\beta)P_{\mathcal{D}^+_B}$$

where w is a unary or binary feature, $P_{\mathcal{D}^+_U}$ is the unary LM of $\mathcal{D}^+$, $P_{\mathcal{D}^+_B}$ is the binary LM, and both LMs are estimated as explained in the previous section. β is a weighting factor. The LM for $\mathcal{D}^-$ is estimated in an analogous manner. In this LM denoted as **LM-BIN-I**, we assume that the features are independent analogous to assuming independence among unigrams in the standard LM techniques.

In order to increase the accuracy of the LM, we propose the modeling of limited dependence among features denoted as **LM-BIN-D**. We consider a universe of terms consisting of pairs of features $\langle f_i, f_j\rangle$ where f_i and f_j could be the unary or the binary features. For example the quotation in Table 2 has as features pairs $\langle fence, \langle economy, help\rangle\rangle$, $\langle\langle fence, help\rangle, \langle economy, support\rangle\rangle$, etc. This is similar to modeling bi-grams in the standard LM setting, but with a crucial difference. While bi-grams are naturally two consecutive unigrams, we cannot insist that our feature pair are consecutive. Instead, we make the default assumption that the feature pair occur in the same sentence. That is, the frequency of a feature pair is the number of sentences in which they co-occur. With this in mind, we estimate our new LM as the interpolation of two LMs,

$$P_{\mathcal{D}^+}(w) = \beta P_{\mathcal{D}^+_U} + (1-\beta)P_{\mathcal{D}^+_{pair}}$$

where w is now either a unary feature, or a feature pair. We can now compute the likelihood of generating the quotation from either the pro document or the con document, in a straightforward way. However, since our binary features are confined to the scope of the same sentence, we can alternatively compute the likelihood of generating a *sentence* from the two polarities' documents. As a quotation typically consists of few sentences, we can subsequently aggregate over these sentence likelihoods. This is,

DEFINITION 4.1. *The **quotation score given the pro document***

$$Score(\mathcal{Q}|\mathcal{D}^+) = MAX_{sen_i \in \mathcal{Q}}(P(sen_i|\mathcal{D}^+))$$

sen_i is a sentence in the quotation, MAX denotes the maximum over the likelihoods of the quotation's sentences, and $P(sen_i|\mathcal{D}^+)$ is computed as,

$$P(sen_i|\mathcal{D}^+) = \prod_{w \in sen_i} P_{\mathcal{D}^+}(w)$$

Analogously, we define the ***quotation score given the con document*** $score(\mathcal{Q}|\mathcal{D}^-)$.

Entailment and contradiction model. As our final variant, we make use of the *agreement* and *disagreement* features of Section 3. The intuition here is that, not only should a pro document (respectively, con) agree with the *agreement* expressions, but the con document (respectively, pro) should agree with the *disagreement* expressions.

DEFINITION 4.2. *Given the agreement features $\mathcal{Q}^+$ and disagreement features $\mathcal{Q}^-$ of a quotation $\mathcal{Q}$, in the **entailment and contradiction model (EC)**, the probability of generating $\mathcal{Q}$ given the pro or the con document:*

$$P(\mathcal{Q}|\mathcal{D}^+) = (1-\lambda)P(\mathcal{Q}^+|\mathcal{D}^+) + \lambda P(\mathcal{Q}^-|\mathcal{D}^-)$$
$$P(\mathcal{Q}|\mathcal{D}^-) = (1-\lambda)P(\mathcal{Q}^+|\mathcal{D}^-) + \lambda P(\mathcal{Q}^-|\mathcal{D}^+)$$

*λ is a weight parameter to determine the importance of the agreement features of the quotation versus the disagreement features. If $\lambda = 0$ which means we consider only the agreement features, we denote the model in this case as **E**. We estimate the above probabilities (e.g. $P(\mathcal{Q}|\mathcal{D}^+)$) using the models described above (e.g. LM-UNA, and LM-BIN).*

5. EVALUATION

We evaluated the effectiveness of the proposed features models and the quotation expansion model in combination with several classifiers: LM-based kNN (k nearest neighbors, here with the special case $k = 1$), LDA, and SVM. We report the classification of quotations into pro/con polarities. The metrics of interest are precision and recall, both micro-averaged over all quotations in a test set and macro-averaged over the classes of pro/con. We created our own training and test datasets as explained next.

5.1 Setup

Quotations datasets. Debatepedia (`debatepedia.org`) is focused on political controversies. It consists of ca. 1,700 topics called debates such as “Deporting illegal immigrants”. Each debate has two types of short documents (quotations) debating it, pro documents and con documents. The quotation belongs to one or more debates in Debatepedia. We extracted 142,253 quotations from Debatepedia, and created our experimental sets.

Test dataset. We compiled, by random sampling, a held-out set of 250 pro and 250 con quotations from various different topics as a *test set*. These 500 quotations, each belonging to one or more topics, covered a total of 73 different fine-grained topics from Debatepedia. Since the topics as well as the polarities are given in advance, the ground truth for the classifiers is known.

Development dataset. Hyper-parameter tuning is performed on separate *development set* of 200 pro and con quotations sampled from Debatepedia for 73 topics that occur in the test dataset.

Training dataset. For training supervised classifiers, we sampled 4,400 quotations from Debatepedia for 73 topics that occur in the test dataset.

Topic Documents variants. We tried variations of preparing pro and con documents.

DNone with original features only. Only the unary and binary features extracte from the debate documents are used as features to represent the debates.

DExp with expansion. In addition to the unary and binary features, their synonyms and antonyms are added to the features set. Therefore, the final set of features for a debate includes its unary and binary features and their expansions with synonyms and antonyms.

Methods under comparison. We compared our family of LM-based methods against each other and against two baselines.

LDA (Latent Dirichlet Allocation). A state-of-the-art latent-topic clustering method (implemented using `mallet.cs.umass.edu`). For pro/con classification, LDA is configured with two latent dimensions; it is run separately for each Debatepedia topic.

SVM (Support Vector Machine). A supervised discriminative classifier (implemented using `svmlight.joachims.org`). For each Debatepedia topic, we train a binary classifier with a linear kernel. Equipped with various feature models (n-grams, unary with expansions and binary with expansions), both LDA and SVM were trained with the quotations in the training dataset described in Section 5, and tuned with the separate development set.

LM-based classification. We studied the LM-based methods described in Section 4 on different test sets, using different feature models, and tuned with the separate development set: (1) the n-grams model: **LM-NG**, (2) the entailment model given the pro LM and the con LM, in combination with the unary model only denoted as **LM-E-UNA**, or the binary and the unary models assuming either binary features independence (BIN-I), denoted as **LM-E-BIN-I**, or binary features dependence (BIN-D), denoted as **LM-E-BIN-D**, and finally (3) the entailment and contradiction model given the pro LM and the con LM, in combination with the unary model only **LM-EC-UNA**, or the binary and the unary models assuming either binary features independence, denoted as **LM-EC-BIN-I**, or binary features dependence, denoted as **LM-EC-BIN-D**.

Table 3: Micro-averaged precision/recall for LM-based pro/con classification on Debatepedia test set

	DNone	DExp
LM-NG	0.68/0.73	0.68/0.75
LM-E-UNA	0.68/0.66	0.69/0.64
LM-E-BIN-I	0.71/0.68	0.70/0.65
LM-E-BIN-D	0.69/0.75	0.72/0.76
LM-EC-UNA	0.65/0.67	0.66/0.69
LM-EC-BIN-I	0.70/0.76	0.72/0.75
LM-EC-BIN-D	0.73/0.76	**0.74/0.78**

Table 4: Micro- and macro-averaged precision/recall for LM-based pro/con classifications on the ProCon test set

	micro P	micro R	macro P	macro R
LM-NG	0.67	0.68	0.70	0.69
LM-EC-UNA	0.64	0.67	0.65	0.69
LM-EC-BIN-D	**0.72**	**0.70**	**0.71**	**0.71**

5.2 Results

LM-based Classification. We used two datasets in the experiments with family of LM-based methods on classifying quotations into pro/con polarities for each topic: (1) the **Debatepedia test dataset**, and (2) **ProCon test dataset**: `www.procon.org` is a political website which provides pro and con quotations by politicians on specific topics. Each quotation is tagged by both the topic and the stance (pro or con). We collected 400 quotations on various topics as our second test set.

Results of Debatepedia dataset. Results of the LM-based approaches with the different topic documents variants are shown in Table 3. These are the results found at $\alpha = 0.64$ for the LMs over the unary features (**LM-UNA**), $\beta = 0.18$ for the LMs over the unary and the binary features (**LM-BIN**), and $\lambda = 0.24$ for the entailment and contradiction model (**EC**). The hyper-parameter values are automatically determined from the development set. At test level $\alpha = 0.05$ using the paired two sample t-test, we found that the results of the different techniques on **DNone** and **DExp** are not statistically different. This means that the expansion of the debates did not significantly improve the results. On the other hand, the difference in the results between the **E** models which use only the agreement features of the quotations and the **EC** models which use both the agreement and the disagreement features is statistically significant. So quotation expansion using both the synonyms and the antonyms improves the results. In addition, the difference in the results of the **BIN-I** model, and the **BIN-D** model were statistically significant, too, which means that considering the dependency of the binary features in each sentence improves the results.

Results of ProCon dataset. We evaluated a set of 235 quotations from the set of quotations from the ProCon test set assigned to the topics in Debatepedia test set. We considered the classification models (**LM-EC-BIN-D**, **LM-EC-UNA**, and **LM-NG**). For this experiment, we used the hyper-parameter values of the LM determined from the development set (e.g. $\alpha = 0.64$, $\beta = 0.18$, and $\lambda = 0.24$). Table 4 shows the results of the three different models. These results are statistically significant at test level $\alpha = 0.05$ using the paired two sample t-test.

LDA & SVM. We conducted experiments on the Debatepedia test set in order to evaluate the effectiveness of the proposed feature models and quotation expansion model on the pro/con classification task with two different classifiers SVM and LDA, in comparison to our LM-based methods. Table 5 shows the micro- and macro-averaged precision and the micro- and macro-averaged recall of LDA and SVM compared to the LM-based approaches in

Table 5: LM pro/con classifiers micro- and macro-averaged precision (P) and recall (R) compared to SVM and LDA

	BIN-D with expansion			
	micro P	micro R	macro P	macro R
LDA	0.72	0.76	0.70	0.74
SVM	0.71	0.72	0.71	0.78
LM	**0.74**	**0.78**	**0.70**	**0.80**
	UNA with expansion			
	micro P	micro R	macro P	macro R
LDA	0.63	0.70	0.67	0.77
SVM	0.60	0.68	0.65	0.72
LM	0.66	0.69	0.71	0.76
	NG			
	micro P	micro R	macro P	macro R
LDA	0.67	0.77	0.69	0.81
SVM	0.63	0.74	0.67	0.81
LM	0.68	0.75	0.73	0.80

combination with different feature models (e.g. **LM-EC-BIN-D, LM-EC-UNA, and LM-NG**). The results are statistically significant at test level $\alpha = 0.05$ using the paired two sample t-test. We notice that the binary features model with expansion improved the results of both the SVM and the LDA classifiers. Moreover, our LM-based methods outperformed SVM and LDA by a significant margin for both precision and recall.

5.3 Discussion

For the pro/con assignment, our best method **LM-EC-BIN-D** achieves almost 74% precision. It uses the richest features, the dependent pairs of binary features and the entailment-contradiction expansions. While one may have hoped for even higher precision, this is actually a decent result given the sophisticated nature and stylistic subtleties of political quotations. The gains over the simpler alternatives are statistically significant. This shows that the novel elements in our features model and quotation expansion are indeed decisive for achieving good precision on this difficult classification task. The experimental results also show that our features model are decisive for achieving good pro/con classification precision using classification methods such as SVM and LDA. The overall winner, however, in this comparison is the LM-based method with rich features and quotation expansion.

6. RELATED WORK

Many previous works on sentiment analysis rely on training classifiers with annotated training data [10, 7]). In our work, we follow an alternative approach of using language models (LMs) to classify opinions, thus reducing the dependence on annotated training data. Other studies of sentiment analysis focus on detecting text polarity given that the classified documents are part of social media like online debates, blogs and twitters [4, 1]. These studies rely on the sentiment orientations of the features of the text (positive and negative), which are rich in these types of media. Some other studies further consider the linkage among documents to detect polarities [4]. However positive/negative features and hyperlinks are very sparse in news media. Prior works addressed the problem of detecting general perspectives (e.g. ideologies or political parties) of given texts, e.g., Republicans versus Democrates, or Palestinian versus Israeli [12, 9]. These works use statistical methods and train classifiers on a set of perspective-annotated documents in order to learn a set of discriminative n-grams of each perspective. This requires manually annotated documents, which is not practical if we move to a finer level of granularity.

Coarse-grained classification is taken further by the work of [6], which aims to annotate political speeches and parliament debates, but does not deal with fine-grained topics. In [8] semantic taxonomies are used to identify aspects of topics, and analyzes opinions on these aspects rather than topics as a whole. This applies to opinion mining on politicians (with aspects such as Vietnam war, Watergate affair, etc.), but it does not address the polarity issue of these opinions. In [5] opinions at the collection level are examined with each collection on a topic coming from a different perspective. A latent topic model is devised to discover the common topics across all the perspectives. For each topic, the opinions from each perspective are summarized. A related task is addressed in [3]. They focus on predicting the sentiment polarity of comments on blog postings. Their approach models mixed-community responses, to identify the topics and the responses they evoke in different sub-communities. In contrast to all of the above, our work finds the underlying fine grained topics and the opinion on each topic at the quotation-level.

7. CONCLUSIONS

We addressed the problem of automatically classifying quotations about political debates which appear in news media and online forums, by politicians or other opinion makers, into fine-grained controversial topics and a pro/con polarity for each topic. We proposed a *topic/polarity classification* approach that maps quotations onto one or more topics in a category system of political debates on fine-grained topics. Our method builds on the estimation of statistical language models on a variety of advanced features designed to overcome the brevity of quotations. We showed the effectiveness of our techniques through systematic experiments on more than 1000 quotations on a variety of topics. Our best method achieved a precision of about 74%.

8. REFERENCES

[1] L. Adamic and N. Glance. The political blogosphere and the 2004 us election: divided they blog. *Workshop on Link Discovery*, 2005.

[2] R. Awadallah, M. Ramanath, and G. Weikum. Harmony and dissonance: organizing the people's voices on political controversies. In *WSDM*, 2012.

[3] R. Balasubramanyan, W. Cohen, D. Pierce, and D. Redlawsk. Modeling polarizing topics: When do different political communities respond differently to the same news? In *ICWSM*, 2012.

[4] C. Burfoot, S. Bird, and T. Baldwin. Collective classification of congressional floor-debate transcripts. In *ACL*, 2011.

[5] Y. Fang, L. Si, N. Somasundaram, and Z. Yu. Mining contrastive opinions on political texts using cross-perspective topic model. In *WSDM*, 2012.

[6] R. Kaptein, M. Marx, and J. Kamps. Who said what to whom?: capturing the structure of debates. In *SIGIR*, 2009.

[7] B. Liu. Sentiment analysis and subjectivity. In *Handbook of Natural Language Processing, Second Edition*. 2010.

[8] Y. Lu, H. Duan, H. Wang, and C. Zhai. Exploiting structured ontology to organize scattered online opinions. In *COLING*, 2010.

[9] D. Nguyen, E. Mayfield, and C. Rosé. An analysis of perspectives in interactive settings. In *SOMA*, 2010.

[10] B. Pang and L. Lee. Opinion mining and sentiment analysis. *FnT in IR*, 2(1-2), 2008.

[11] S. Ponzetto and R. Navigli. Knowledge-rich word sense disambiguation rivaling supervised systems. In *ACL*, 2010.

[12] X. Zhou, P. Resnick, and Q. Mei. Classifying the political leaning of news articles and users from user votes. In *ICWSM*, 2011.