Language-Model-based Pro/Con Classification of Political Text

Rawia Awadallah, Maya Ramanath, and Gerhard Weikum
Max-Planck Institute for Informatics
Saarbrücken, Germany
rawadall@mpi-inf.mpg.de, ramanath@mpi-inf.mpg.de, weikum@mpi-sb.mpg.de

ABSTRACT

Given a controversial political topic, our aim is to classify documents debating the topic into pro or con. Our approach extracts topic related terms, pro/con related terms, and pairs of topic related and pro/con related terms and uses them as the basis for constructing a pro query and a con query. Following standard LM techniques, a document is classified as pro or con depending on which of the query likelihoods is higher for the document. Our experiments show that our approach is promising.

Categories and Subject Descriptors

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

General Terms

Algorithms, Experimentation

Keywords

Language Models, Sentiment Analysis, Text Classification

1. INTRODUCTION

The popularity of online forums such as film and book review sites, online political fora, personal blogs, the comments section on newspaper articles, etc., allow people to post their views and opinions on a wide range of topics. The proliferation of such opinion oriented content has led to renewed interest in sentiment analysis and opinion mining techniques to facilitate the *automatic* analysis and classification of opinions. Automated analysis of opinions has a wide range of applications, including, advertising, political policy formulation, business intelligence applications, etc.

In this paper, we focus on the problem of classifying political opinions (expressed for example, in online debate forums) on controversial questions such as, "Should felons be given voting rights?" and "Should the death penalty remain a legal option in America?" into "pro" opinions and "con" opinions. Figure 1 shows an example of the kind of documents we would like to classify.

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.

Topic: "Should the *death penalty* **remain** a **legal** option?"
***Pro*:**
"The death *penalty* for heinous crimes in which the circumstances warrant capital *punishment* should ***not*** be **altered**. The **laws** that expand the number of federal crimes *punishable* by *death*, including terrorism and narcotics trafficking by drug kingpins should be **retained** ."
***Con*:**
"Since I was a law student at Harvard, I have been ***against*** the **continuation** of *death penalty*. It does not deter. It is severely discriminatory against minorities, especially since they're given ***no*** competent **legal** counsel defense in many cases."

Figure 1: A discussion topic with pro and con documents.

Many previous works on sentiment analysis address the problem of *query independent* opinion mining. For example, classifying a movie review into positive or negative. However, our setting is *query dependent* and the classification of a document into pro or con can change depending on the topic (see [1]). For example, in Figure 1, if the wording of the topic were changed to "Should the death penalty become illegal in America?", then the first document is actually a con document while the second document is pro. Moreover, most previous methods rely on training classifiers with annotated training data (see [2] for an overview and [3] for an example of classifying political text), which often has to be annotated manually. In this work, we follow an alternative approach of using language models (LMs) to classify opinions, thus reducing the dependence on annotated training data.

1.1 Our Approach

Our approach follows the query likelihood method [4], where a query is regarded as a sample of the document and its likelihood gives the measure of relevance of the document to the query. Clearly, in our setting, given a discussion topic and a document which debates it, the document will always be highly relevant to the topic. However, our goal is not to quantify relevance, but to classify the document based on its opinion. If a document contains expressions which are in agreement to similar expressions in the topic statement, then it is likely that the document is a pro document. Otherwise, it is likely to be a con document. For example, in the pro document in Figure 1, the expression "capital punishment should not be altered" is in agreement with the discussion topic expression "death penalty remain a legal option", while, in contrast, the con document contains the expression, "against the continuation of death penalty". Finding such expressions and using them for pro/con classification is one of the key challenges that we address in this work.

In brief, our technique is based on the following steps. First, we map both the discussion topic and documents to a set of "interesting patterns". Second, we make use of the interesting patterns from

Table 1: Synonyms and Antonyms of terms

Term	Synonyms	Antonyms
penalty	*punishment*, *retribution*	*award*, *pardon*
legal	**sound**, **lawful**	**illegal**, **unlawful**, not **legal**
remain	**stay**, **continue**	**change**, **alter**, not **remain**

the discussion topic to construct two queries: a "pro" query and a "con" query. Third, for a given test document, an LM is estimated based on the interesting patterns in the document and in the background corpus. And finally, both the pro query and the con query likelihoods are computed and the document is classified as pro or con based on the likelihood values. In the rest of this paper, we formalize our technique and present results of our experiments.

2. PRO/CON CLASSIFICATION MODEL

Interesting patterns. Given a controversial discussion topic and the corpus of documents debating it, we first identify the "discussion vocabulary", consisting of two kinds of terms: topical terms describing the topic, and pro/con terms describing opinions on the topic. In the discussion topic, nouns are assumed to be topic-related while verbs, adjectives, and adverbs are assumed to be pro/con related. For both types of terms, we look up WordNet for synonyms and antonyms. Table 1 shows examples of synonyms and antonyms for both topical as well as pro/con terms of the discussion topic in Figure 1. Note that the term itself can be negated and is indicated with a "not" and negations of terms are detected during parsing. In Figure 1, topic's topical terms and their synonyms and antonyms in the pro/con documents are *italicized*, while topic's pro/con terms and their synonyms and antonyms are in **bold**. Negations are ***italicized and bold***.

Let T_s denote a topic term or its synonym and T_a denote an antonym of a topic term. Similarly, PC_s and PC_a denote a pro/con term or its synonym and an antonym of a pro/con term. We construct two types of interesting patterns: (1) binary patterns *(lexical pairs)*: $\langle \mathsf{T}_s, \mathsf{PC}_s \rangle$, $\langle \mathsf{T}_a, \mathsf{PC}_a \rangle$, $\langle \mathsf{T}_a, \mathsf{PC}_s \rangle$, $\langle \mathsf{T}_s, \mathsf{PC}_a \rangle$ (e.g. $\langle \mathsf{penalty}, \mathsf{remain} \rangle$, $\langle \mathsf{penalty}, \mathsf{against\ continuation} \rangle$, etc.), and (2) unary patterns: T_s, T_a, PC_s, PC_a. Let B and U denote the set of binary and unary patterns respectively.

Constructing queries. Similar to general ontology based query expansion, the interesting patterns described above are used to construct two queries: a pro query Q^+ and a con query Q^-,

$$Q^+ = \{\langle \mathsf{T}_s, \mathsf{PC}_s \rangle\} \cup \{\langle \mathsf{T}_a, \mathsf{PC}_a \rangle\} \cup \{\mathsf{T}_s\} \cup \{\mathsf{PC}_s\}$$
$$Q^- = \{\langle \mathsf{T}_s, \mathsf{PC}_a \rangle\} \cup \{\langle \mathsf{T}_a, \mathsf{PC}_s \rangle\} \cup \{\mathsf{T}_a\} \cup \{\mathsf{PC}_a\}$$

Estimating the LM of a document. An LM M_D of a test document D is estimated as an interpolation of a binary pattern and a unary pattern LMs over all interesting patterns, thus benefiting from both LMs and overcoming the sparseness problem [4]:

$$P_{M_D}(pat_i|D) = (1-\alpha)P_B(pat_i|D) + \alpha P_U(pat_i|D)$$

where $P_B(pat_i|D)$: LM of D over binary patterns, $P_U(pat_i|D)$: LM of D over unary patterns, pat_i: a pattern and α: a weight parameter. A Unary pattern LM $P_U(pat_i|D)$ is estimated as:

$$P_U(pat_i|D) = (1-\lambda)P(pat_i|D) + \lambda P(pat_i|C)$$

where $pat_i \in U$, C: a background corpus and λ: a smoothing parameter. $P(pat_i|D)$ and $P(pat_i|C)$ are estimated as:

$$P(pat_i|D) = \frac{c(pat_i; D)}{\sum_{pat_j \in D} c(pat_j; D)}$$

$$P(pat_i|C) = \frac{c(pat_i; C)}{\sum_{pat_j \in C} c(pat_j; C)}$$

where $c(pat_i; D)$ and $c(pat_i; C)$ denote the frequency of pat_i in document D and corpus C, respectively. The binary LM $P_B(pat_i|D)$ is estimated in an analogous manner.

Classifying the document. Given M_D, the estimated LM of test document D, we estimate the query likelihoods of both the pro and con queries, assuming independence between patterns. That is,

$$P(Q^+|M_D) = \prod_{pat_i \in Q^+} P(pat_i|D)$$

$$P(Q^-|M_D) = \prod_{pat_i \in Q^-} P(pat_i|D)$$

The document is classified as pro if $P(Q^+|M_D) > P(Q^-|M_D)$ and con otherwise.

3. EXPERIMENTAL EVALUATION

Table 2: Precision and Recall of all Techniques

	OWN	LM-term	SVM
D1 (prec.,recall)	**0.66,0.68**	0.64,0.66	0.64,0.65
D2 (prec.,recall)	**0.67,0.67**	0.63,0.62	0.65,0.63

We used two datasets to evaluate our method: one from `http://www.procon.org` (dataset D1) and another from `http://www.opposingviews.com`(dataset D2). Both websites contain controversial political questions. Each question has a clearly marked (pro or con) set of documents debating it and thus serving as the ground truth for evaluation. We chose around 350 questions and their corresponding documents from each dataset.

We evaluated our method against two methods: a trained SVM classifier with our patterns as features, and an LM-based method which considers only unary patterns (denoted LM-term). The results in Table 2 show the differences (which are statistically significant) in both precision and recall between our method and the other two methods on both datasets.

4. CONCLUSIONS AND FUTURE WORK

We proposed an LM-based method for classifying political texts into pro or con, based on a controversial discussion topic. We evaluated our proposal and showed that it is promising. In this work, we considered topical and pro/con unigrams. A natural extension is to extend this to n-grams. Our datasets were known to contain formal political opinions expressed in non-emotive language and so, we were able to ignore many other tricky issues, such as, for example, dealing with noisy informal text, and identifying opinion-bearing sentences relevant to the topic. In the future, we plan to extend our techniques to work on other datasets (e.g. newspaper articles, blogs).

5. REFERENCES

[1] K. Eguchi and V. Lavrenko. Sentiment retrieval using generative models. In *EMNLP*, 2006.

[2] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Found. and Trends in IR*, 2(1-2):1–135, 2008.

[3] M. Thomas, B. Pang, and L. Lee. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *EMNLP*, 2006.

[4] C. Zhai. Statistical language models for information retrieval. *Found. and Trends in IR*, 2(3):137–215, 2008.