Retrieving Topical Sentiments from Online Document Collections

Matthew Hurst and Kamal Nigam

Intelliseek Applied Research Center, Pittsburgh, USA

ABSTRACT

Retrieving documents by subject matter is the general goal of information retrieval and other content access systems. There are other aspects of textual content, however, which form equally valid selection critieria. One such aspect is that of sentiment or polarity - indicating the users opinion or emotional relationship with some topic. Recent work in this area has treated polarity effectively as a discrete aspect of text. In this paper we present a lightweight but robust approach to combining topic and polarity thus enabling content access systems to select content based on a certain opinion *about* a certain topic.

1. INTRODUCTION

As the new marketing manager for BrightScreen, a supplier of LCD screens for personal digital assistants you would like to understand what positive and negative impressions the public holds about your product. Your predecessor left you 300,000 customer service emails sent to BrightScreen last year that address not only screens for PDAs, but the entire BrightScreen product line. Instead of trying to manually sift through these emails to understand the public sentiment, can document retrieval and text analysis techniques help you quickly determine what aspects of your product line are viewed favorably or unfavorably?

One way to address BrightScreen's business need would be a text mining toolkit that automatically identifies just those email fragments that are topical to LCD screens and also express positive or negative sentiment. These fragments will contain the most salient representation of the consumers' likes and dislikes specifically with regard to the product at hand. The goal of the research presented in this paper is to reliably extract polar sentences about a specific topic from a corpus of data containing both relevant and irrelevant text.

Recent advances in the fields of text mining, information extraction and information retrieval have been motivated by a similar goal: to exploit the hidden value locked in huge volumes of unstructured data. Much of this work has focused on categorizing documents into a predefined topic hierarchy, finding named entities (entity extraction), clustering similar documents, and inferring relationships between extracted entities and meta data.

An emerging field of research with much perceived benefit, particularly to certain corporate functions such as brand management and marketing, is that of sentiment or polarity detection. For example, sentences such as `I hate its resolution` or `The BrightScreen LCD is excellent` indicate authorial opinions about the BrightScreen LCD. Sentences such as `The BrightScreen LCD has a resolution of 320x200` indicates factual objectivity. To effectively evaluate the public's impression of a product, focusing on the small minority of sentences containing subjective language can create a much more efficient evaluation process.

Recently, several researchers have addressed techniques for analyzing a document and discovering the presence or location of sentiment or polarity within them. Weibe[1] discovers subjective language by doing a fine-grained NLP-based textual analysis. Pang et al[2] use a machine learning classification-based approach to determine if a movie review as a whole is generally positive or negative about the movie.

This prior art makes significant advances into this novel area. However, they do not consider the relationship between polar language and topicallity. In taking a whole-document approach, Pang[2] sidesteps any issues of topicality by assuming that each document addresses a single topic (a movie), and that the preponderance of the expressed sentiment is about the topic. In the domain of movie reviews this may be a good assumption (though

Further author information: (Send correspondence to Matthew Hurst.)
Matthew Hurst: E-mail: mhurst@intelliseek.com

it is not tested), but this assumption does not generalize to less constrained domains.* Weibe's[1] approach does a good job of capturing the local context of a single expression, but with such a small context, the subject of the polar expression is typically captured by just the several base noun words, which are often too vague to identify the topic in question.

This paper strikes out a middle ground between these two approaches and argues for a fusion of polarity and topicality. One approach to performing this task is to do a full NLP-style analysis of each sentence and understand at a deep level the semantics of the sentence, how it relates to the topic, and whether the sentiment of the expression is positive or negative. In the absense of comprehensive NLP techniques, we instead approximate the topicality judgment with a statistical machine learning classifier and the polarity judgment with shallow NLP techniques. The system we describe asserts that any sentence that is both polar and topical is polar about the topic in question. However, when these modules are run separately there are no guarantees that a sentence that is judged to be both topical and polar is expressing anything polar about the topic. For example the sentence `It has a BrightScreen LCD screen and awesome battery life` does not say anything positive about the screen. This paper demonstrates that the underlying combination assumption made by this system is sound, resulting in high-precision identification of these sentences.

In summary, in an industrial application setting, the value of polarity detection is very much increased when married with an ability to determine the topic of a document or part of a document. In this paper, we outline methods for recognizing polar statements and for determining the topic of a document segment.

2. POLARITY

Texts can be broadly categorized as subjective or objective.[1] Those that are subjective often carry some indication of the author's opinion, or evaluation of a topic as well as some indication of the author's emotional state with respect to that topic. For example, the statement `this is an excellent car` indicates the author's evaluation of the car in question; `I hate it!` reflects the author's emotional state with respect to the topic. An additional type of statement informative in this context is that which indicates a desirable or undesirable condition. These statements may be deemed objective. For example, `it is broken` may well be objective but is still describing an undesirable state.

An idealized view of polarity detection would be able to accept any document, or subsection thereof, and provide an indication of the polarity: the segment is either positive, negative, mixed or neutral.

However, this is only half of the story. Firstly, the classification of polar segments has a dependancy on many other aspects of the text. For example, the adjective `huge` is negative in the context `there was a huge stain on my trousers` and positive in the context `this washing machine can deal with huge loads`. There is no gurantee that the information required to resolve such ambiguities will be present in the observable segment of the document.

Secondly, knowing that a piece of text is positive is only as useful as our ability to determine the topic of the segment. If a brand manager is told that this set of documents is positive and this set is negative, they cannot directly use this information without knowing, for example, which are positive about their company and which are positive about the competition.

The polar phrase extraction system was implemented with the following steps.

In the set up phase, a lexicon is developed which is tuned to the domain being explored. For example, if we are looking at digital cameras, phrases like 'blurry' may be negative and 'crisp' may be positive. Care is taken not to add ambiguous terms where possible as we rely on assumptions about the distribution of the phrases that we can detect with high precision and its relationship to the distribution of all polar phrases. Note that our lexicon contains possibly 'incorrect' terms which reflect modern language usage as found in online messages. For example, there is an increasing lack of distinction between certain classes of adverbs and adjectives and so many adjectives are replicated as adverbs.

*It is noted that the data used in that paper contained a number of reiviews about more than one movie. In addition, the domain of movie reviews is one of the more challanging for sentiment detection as the topic matter is often of an emotional character: there are bad characters that make a movie enjoyable.

At run time, the input is tokenized. The tokenized input is then segmented into discrete chunks. The chunking phase consists of the following steps. Part of speech tagging is carried out using a statistical tagger trained on Penn Treebank data.† Semantic tagging adds polar orientation information to each token (positive or negative) where appropriate using the prepared polarity lexicon. Simple linear POS tag patterns are then applied to form the chunks. The chunk types that are derived are basic groups (noun, adjective, adverb and verb) as well as determiner groups and an 'other' type.

The chunked input is then further processed to form higher-order groupings of a limited set of syntactic patterns. These patterns are designed to cover expressions which associate polarity with some topic, and those expressions which toggle the logical orientation of polar phrases (`I have never liked it.`). This last step conflates simple syntactic rules with semantic rules for propogating the polarity information according to any logical toggles that may occur.

If the text `This car is really great` were to be processed, firstly the tokenization step would result in the sequence {this, car, is, really, great}. Part of speech tagging would provide {this_DT, car_NN, is_VB, really_RR, great_JJ}. Assuming the appropriate polarity lexicon, additional information would be added thus: {this_DT, car_NN, is_VB, really_RR, great_JJ;+} where '+' indicate a positive lexical item. Note that features are encoded in a simplified frame structure which is a tree. The standard operations of unification (merging), test for unifiability and subsumption are available on these structures.

The chunking phase would bracket the token sequence as follows: {(this_DT)_DET, (car_NN)_BNP, (is_VB)_BVP, (really_RR, great_JJ)_BADJP}. Note that the basic chunk categories are {DET, BNP, BADVP, BADJP, BVP, OTHER}.

The interpretation phase then carries out two tasks: the elevation of semantic information from lower constituents to higher, applying negation logic where appropriate, and assembling larger constituents from smaller. Rules are applied in a certain order. In this example, a rule combining DET and BNP chunks would work first over the sequence, followed by a rule that forms verb phrases from BNP BVP BADJP sequences whenever polar information is found in a BADJP.

Note that there is a restriction of the applicability of rules related to the presence of polar features in the frames of at least one constituent (be it a BNP, BADJP, BADVP or BVP).

The simple syntactic patterns are: Predicative modification (`it is good`), Attributive modification (`a good car`), Equality (`it is a good car`), Polar clause (`it broke my car`).

Negation of the following types are captured by the system: Verbal attachment (`it is not good`, `it isn't good`), Adverbal negatives (`I never really liked it`, `it is never any good`), Determiners (`it is no good`).

3. POLARITY EVALUATION

We wish to evaluate three aspects of our approach: the performance of the topic classifier on sentences, the performance of the polarity recognition system and the assumption that polar sentences that are on topic contain polar language about that topic.

Using our message harvesting and text mining toolkit, we acquired 20,000 messages from online resources (usenet, online message boards, etc.). Our message harvesting system harvests messages in a particular domain (a vertical industry, such as 'automotive', or a specific set of products). Messages are then automatically tagged according to some set of topics of interest to the analyst.

Our experiment proceeded as follows. We selected those messages which were tagged as being on topic for a particular topic in the domain being studied (982 messages). These messages were then segmented into sentences (using a naive sentence boundary detection algorithm) resulting in (16, 616 sentences). The sentences were then tagged individually by the topic classifier and the polarity recognition system.

We then selected at random 250 sentences for each of the evaluation tasks (topic, polarity, topic & polarity) and hand labeled them as follows.

†We note that taggers trained on clean data, when applied to the noisey data found in our domain, are less accurate than with their native data.

polarity : positive, negative (in a multi-label environment this results in four possible combinations).

topic : topical, off-topic (a binary labelling).

topic and polarity : positive-correlated, negative-correlated, positive-uncorrelated, negative-uncorrelated, topical, off-topic. The positive-correlated label indicates that the sentences contained a positive polar segment that referred to the topic, positive-uncorrelated indicates that there was some positive polarity but that it was not associated with the topic in question.

As our system is designed to detect relative degrees of opinion we are far more interested in precision than recall. A far greater issue than recall is the potential bias that our set of classifiers might impose on the data. This aspect is not measured here due to the labour intensive nature of the task.

The results for the polarity task from this hand labelling are shown in Table 1. Sentences judged to have positive polarity were detected with a precision of 82%. Negative sentences were judged to be detected with a precision of 80%.

		predicted	
		pos	neg
truth	pos	139	
	notPos	30	
	neg		70
	notNeg		17

Table 1. Precision of polarity for hand labeled sentences. Positive: 82%; negative: 80%

4. IDENTIFYING TOPICAL SENTENCES WITH A DOCUMENT CLASSIFIER

In the previous section we approached the task of assessing the sentiment of a sentence through a shallow NLP approach. In this section, we take a different approach for determining the topicality of a sentence. We treat the topicality judgment as a text classification problem and solve it with machine learning techniques.

In the standard text classification approach, representative training examples are provided along with human judgements of topicality. From these, a learning algorithm forms a generalization hypothesis that can be used to determine topicality of previously unseen examples. Typically, the types of text that form the training examples are the same type as those seen during the evaluation and appplication phases for the classifier. That is, the classifier assumes the example distribution remains constant before and after training.

In our application of finding topical polar sentences within a text mining system, that would mean a classifier that was trained over a distribution of hand-labeled sentences. Unfortunately, such an idealized setup is not available as part of our system; generally speaking, there are not enough configuration resources to train such a classifier. Instead, in our text mining system for a specific marketing domain, a machine learning text classifier is trained to assess topicality on *whole messages* and thus expects to predict whether or not a whole message is relevant to the given topic. In this section we explore how to use such a text classifier trained on whole messages to accurately predict sentence-level topicality.

4.1. Classifying Topical Messages

The provided classifier is trained with machine learning techniques from a collection of documents that have been hand-labeled with the binary relation of topicality. The underlying classifier is a variant of the Winnow classifier,[3–5] an online learning algorithm that finds a linear separator between the class of documents that are topical and the class of documents that are irrelevant. Documents are modeled with the standard bag-of-words representation that simply counts how many times each word occurs in a document. Winnow learns a linear classifier of the form

$$h(x) = \sum_{w \in V} f_w c_w(x) \tag{1}$$

where $c_w(x)$ is 1 if word w occurs in document x and 0 otherwise. f_w is the weight for feature w. If $h(x) > V$ then the classifier predicts topical, and otherwise predicts irrelevant. The basic Winnow algorithm proceeds as:

1. Intialize all f_w to 1.
2. For each labeled document x in the training set:

2a. calculate $h(x)$.

2b. If the document is topical, but Winnow predicts irrelevant, update each weight f_w where $c_w(x)$ is 1 by:

$$f_w* = 2 \tag{2}$$

2c. If the document is irrelevant, but Winnow predicts topical, update each weight f_w where $c_w(x)$ is 1 by:

$$f_w/ = 2 \tag{3}$$

In a setting with many irrelevant features, no label noise and a linear separation of the classes, Winnow is theoretically guaranteed to quickly converge to a correct hypothesis. Empirically, we have found Winnow to be a very effective document classification algorithm, rivaling the performance of Support Vector Machines[6] and k-Nearest Neighbor,[7] two other state-of-the-art text classification algorithms. We use Winnow because it is more computationally efficient than SVMs and easier to apply than kNN.

4.2. Classifying Topical Sentences

We use a straightforward and ad-hoc technique of adapting a given document classifier into a high precision/low recall sentence classifier. If a document is judged by the classifier to be irrelevant, we predict that all sentences in that document are also irrelevant. If a document is judged to be topical, then we further examine each sentence in that document. Given each sentence and our text classifier, we simply form a bag-of-words representation of the sentence as if an entire document consisted of that single sentence. We then run the classifier on the derived pseudo-document. If the classifier predicts topical, then we label the sentence as topical and proceed with the sentiment analysis for that sentence. If the classifier predicts irrelevant, we skip the sentiment analysis and proceed on to the next sentence.

A machine learning classifier expects the training document distribution and the testing distribution to be similar, and any theoretical guarantees of performance are abandoned when this type of adaptation is performed. However, we have empirically observed quite good performance from this technique. We find that sentence-level classification tends to maintain high precision but have significantly lower recall than the performance of the classifier over whole documents. For the class of linear separator document classifiers, this result is expected when the frequency of topical training documents is relatively small (significantly less than 50%). Since a sentence is substantially shorter than the average document, there will be many fewer features in a sentence bag-of-words than in a document bag-of-words. In the extreme case, an document with no words will always be classified as irrelevant, because the default always-on feature will predict irrelevant, since the topic is relatively rare. With just a very few features on for a sentence, the words in the sentence need to be very topical in order for the classifier to predict positive. Thus, many sentences that are truly topical will not be classified as such, because the strength of their word weights will not be enough to overcome the default feature's weight. This leads directly to a loss in recall. On the other hand, the sentences that are predicted positive tend to have a large frequency of topical words, making the prediction of positive sentences still have the high precision that the classifier had on the document level.

4.3. Results and Discussion

We use the same experimental setup as described in the previous section. We trained a Winnow classifier by hand-labeling 731 messages, 246 which were topical. On our test collection, 982 messages were predicted to be topical by the classifier. Precision was measured at 85.4% (117/137 on a randomly selected test set) on the message level. The 982 messages contained 16,616 sentences, 1262 of which were judged to be topical by the classifier. These sentences came from 685 different documents, indicating that that 70% of documents judged to be topical also had at least one sentence predicted to be topical. A random sample of 224 of the 1262 topical sentences were hand labeled. Precision on this set was estimated at 79% (176/224). These results show that applying a message-level classifier in a straightforward fashion on the sentence level still maintains about the same precision that was seen on the document level. However, this approach clearly results in a loss of recall, as a significant number of messages predicted to be topical did not have any sentences predicted as topical.

Instead of adapting a document classifier to label individual sentences, it would be more appropriate to have a sentence-level classifier for the task. However, the difficulty arises with the training data; in practice, it is not feasible to collect labeled training data over single sentences. It is much more efficient and practical to ask a human for labeling judgments about an entire document. Traditional machine learning techniques cannot train a sentence-level classifier from labels for whole documents, because there is no knowledge about which sentences in a topical document are actually topical, and which are superfluous to the topic. However, we believe that the multiple instance classification setting[8] would be applicable to this task. In this setting labels are provided for a collection of instances. The provided label is positive if at least one instance in the collection is positive, and is negative if every instance in the collection is negative. This maps well to the sentence-document task. Each instance will be a sentence, and each document will be a collection of sentences that occurs in the document. The underlying assumption here is that if a document is positive for some topic, there must be at least one sentence that, taken by itself, is also positive for that topic. In general, this need not always be the case, but will be true for the preponderance of documents. In our ongoing work, we are exploring the use of such techniques in this domain.

5. POLARITY AND TOPIC

The goal of the work presented in this paper is to reliably extract polar sentiments about a topic. Our system asserts that a sentence judged to be polar and also judged to be topical is indeed expressing polarity about the topic. This relationship is asserted without any NLP-style evidence for a connection between the topic and the sentiment, other than their apparent locality in the same sentence. This section tests the assumption that at the locality of a sentence, a message that is both topical and polar actually expresses polar sentiment about the topic.

		predicted	
		topic& positive	topic & negative
truth	topic & positive	137	
	other	52	
truth	topic & negative		37
	other		15

Table 2. Topic/Polarity combinations: 72% precision (72% for positive, 71% for negative)

Using the polarity and topic modules described and tested in the previous two sections, the system identify sentences that are judged to be topical and have either positive or negative sentiment. These sentences are predicted by the system to be saying either positive or negative things about the topic in question. Out of the 1262 sentences predicted to be topical, 316 sentences were predicted to have positive polarity and 81 were predicted to have negative polarity. The precision for the intersection - testing the assumption that a topical sentence with polar content is polar about that topic - is show in Table 2. The results show the overall precision was 72%. Since the precision of the polarity module was 82% and the topic module 79%, an overall precision of 72% demonstrates that the locality assumption holds in most instances.

Below are five randomly selected sentences predicted to be negative topical and five randomly selected sentences predicted to be positive topical. These show typical examples of the sentences discovered by our system.

- Compared to the PRODUCT's screen this thing is very very poor.
- In multimedia I think the winner is not that clear when you consider that PRODUCT-A has a higher resolution screen than PRODUCT-B and built in camera.
- I never had a problem with the PRODUCT-A, but did encounter the "Dust/Glass Under The Screen Problem" associated with PRODUCT-B.
- broken PRODUCT screen
- It is very difficult to take a picture of a screen.
- The B&W display is great in the sun.
- The screen is at 70 setting (255 max) which is for me the lowest comfortable setting.
- At that time, superior screen.
- Although I really don't care for a cover, I like what COMPANY-A has done with the rotating screen, or even better yet, the concept from COMPANY-B with the horizontally rotating screen and large foldable keyboard.
- The screen is the same (both COMPANY-A & COMPANY-B decided to follow COMPANY-C), but multimedia is better and more stable on the PRODUCT.

6. CONCLUSIONS

Determining the sentiment of an author by text analysis requires the ability to determine the polarity of the text as well as the topic. We believe that topic detection is generally best solved by a trainable classification algorithm and that polarity detection is generally best solved by a grammatical model. The approach described in this paper takes independant topic and polarity systems and combines them, under the assumption that a topical sentence with polarity contains polar content on that topic. We tested this assumption and determined it to be viable for the domain of online messages. This sytem provides the ability to retrieve messages (in fact, parts of messages) that indicate the authors sentiment to some particular topic, a valuable capability.

The detection if polarity is a semantic or meta-semantic interpritive problem. A complete linguistic solution to the problem would need to deal with word sense issues and some form of discourse modeling (which would ultimately require a reference resolution component) in order to determine the topic of the polar expressions. Our approach restricts these open problems by constraining the data set, and specializing the detection of polarity. These steps by no means address directly these complex linguistic issues, but taken in conjunction (and with some additional aspects pertaining to the type of expressions found in the domain of online messages) the problem is constrained enough to produce perfectly reliable results.

There are many assumptions made in this approach.

- The assumption that sentences on a topic which are polar (as detected by our system) is tested here and deemed to be reliable for our analytical needs.
- The assumption that sentences that are determined to be topical follow the distribution of messages that are topical is not evaluated in this paper though an initial investigation suggests that it is supportable.
- The assumption that the set of polar expressions modeled by our grammatical approach to polarity detection follows the distribution of all polar expressions is not evaluted here.

- The assumption that the distribution of polar expressions on a topic that are scoped sententially is directly related to the entire set of polar topical statements - the assumption that allows us to project our results onto the entire data set - has not been evaluated in this paper, though an initial investigation into this area suggests that it is a valid assumption.

In summary, further evaluation is required to ensure that the inferences that we wish to make about the results produced by our system are reliable.

In the future, we intend to evaluate the entire set of assumptions outlined above as well as make improvements to our topic classification and polarity detection systems.

Topic classification may be improved by:

- lightweight anaphora resolution to improve topic recall on the sentence level. Throwing in candidate resolution phrases from previous sentences if anaphora are detected in the given sentence.
- multiple instance training.

and polarity detection may be improved by:

- automated construction of lexica per domain.
- a larger set of phrase patterns.
- improved modelling of negation and other polarity toggling grammatical devices (such as `it solved my problem`).

ACKNOWLEDGMENTS

William Cohen, Natalie Glance, Matthew Siegler, Robert Stockton, Takashi Tomokiyo, Jennifer Washburn.

REFERENCES

1. J. Wiebe, T. Wilson, and M. Bell, "Identifying collocations for recognizing opinions," in *Proceedings of ACL/EACL '01 Workshop on Collocation*, (Toulouse, France), July 2001.
2. B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? sentiment classification using machine learning techniques," in *Proceedings of EMNLP 2002*, 2002.
3. N. Littlestone, "Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm," *Machine Learning* **2**, pp. 285–318, 1988.
4. A. Blum, "Empirical support for winnow and weighted-majority based algorithms: results on a calendar scheduling domain," *Machine Learning* **26**, pp. 5–23, 1997.
5. I. Dagan, Y. Karov, and D. Roth, "Mistake-driven learning in text categorization," in *EMNLP '97, 2nd Conference on Empirical Methods in Natural Language Processing*, 1997.
6. T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *Machine Learning: ECML-98, Tenth European Conference on Machine Learning*, pp. 137–142, 1998.
7. Y. Yang, "An evaluation of statistical approaches to text categorization," *Information Retrieval* **1**(1/2), pp. 67–88, 1999.
8. T. G. Dietterich, R. H. Lathrop, and T. Lozano-Perez, "Solving the multiple-instance problem with axis-parallel rectangles," *Artificial Intelligence* **89**(1-2), pp. 31–71, 1997.