Chapter 20

Towards a Robust Metric of Polarity

Kamal Nigam and Matthew Hurst
Intelliseek Applied Research Center
5001 Baum Blvd, Suite 644
Pittsburgh, PA 15213, USA
{knigam, mhurst}@intelliseek.com

Abstract

This chapter describes an automated system for detecting polar expressions about a specified topic. The two elementary components of this approach are a shallow NLP polar language extraction system and a machine learning based topic classifier. These components are composed together by making a simple but accurate collocation assumption: if a topical sentence contains polar language, the polarity is associated with the topic. We evaluate our system, components and assumption on a corpus of online consumer messages.

Based on these components, we discuss how to measure the overall sentiment about a particular topic as expressed in online messages authored by many different people. We propose to use the fundamentals of Bayesian statistics to form an aggregate authorial opinion metric. This metric would propagate uncertainties introduced by the polarity and topic modules to facilitate statistically valid comparisons of opinion across multiple topics.

Keywords: natural language processing, text classification, sentiment analysis, text mining, metrics.

1. Introduction

In the field of market research, one largely untapped data source is unsolicited first-person commentary freely available on the internet through blogs, Usenet, and web sites with discussion boards. Traditional methods of market research include surveys and focus groups. With these methods it is relatively easy to collect a limited amount of data in a structured form amenable to statistical analysis. In contrast, the characteristics of unsolicited first-person commentary include (1) a huge volume of mostly irrelevant content that (2) is created by a non-random sample of consumers, and (3) is available as unstructured text instead of checkboxes or rankings on a survey form.

With the proper tools, these seeming disadvantages become advantages because each increases the richness of the data available. The huge volume of total data means that typically there is also a large amount of topical relevant data. Typically, the authors of this commentary are key targets for marketers—they are disproportionately influential, spreading their opinions in large public forums. Finally, the unstructured nature of the data allows a level of detail and unfiltered feedback that is not available by forcing everyone to have an opinion on a survey form.

The goal of our research is to create text analysis techniques that facilitate real-world market research over first-person commentary from the internet. An emerging field of research related to this is that of automatically identifying sentiment or polarity in unstructured text. For example, sentences such as *I hate the BrightScreen LCD's resolution* and *My BrightScreen LCD had many dead pixels* indicate negative authorial opinion and objective but negatively oriented description respectively.

In a previous paper (Hurst and Nigam, 2004) we demonstrated an algorithm for identifying subjective or polar sentences about a particular topic of interest, such as a product offering or corporate brand. The goal of that work was to identify sentences that could be efficiently scanned by a marketing analyst to identify salient quotes to use in support of positive or negative marketing conclusions. To this end, the work focused on achieving high-precision results without concern for the recall of the algorithm. Given the volume of text under consideration by an analyst, high recall was not necessary.

Our previous work enabled discovery of anecdotal evidence in support of a marketing finding, but it did not provide any technique for assessing the overall average opinion of the authorial public. In this paper, we take the first steps toward automated techniques that assess at an aggregate level the orientation of a corpus of unsolicited first-person commentary regarding a particular topic. That is, we seek text analysis techniques that result in a well-founded metric score that represents public opinion about a topic such as a product offering or corporate brand. If such an automated technique exists, it can be used to efficiently evaluate brands in the marketplace. For example, it could score different makes of automobiles based on unsolicited customer satisfaction feedback in blogs, Usenet, and message board discussions.

The general approach that we take is:

- Segment the corpus into individual expressions (sentences, in our case).
- Use a general-purpose polar language module and a topic classifier to identify individual polar expressions about the topic of interest.
- Aggregate these individual expressions into a single score, taking into account the known and measured performance characteristics of the polarity and topic modules as well as other properties of the corpus.

This paper describes and evaluates our techniques for the first two steps of this process and presents our thoughts and some initial empirical exploration detailing how we plan to proceed on the third step.

2. Related Work

Agrawal et al. (2003) describe an approach to opinion mining that relies on the link structure implied by citations in newsgroup postings. A subset of topical message is derived using a simple

keyword filter and the graph described by the link structure is partitioned into 'for' and 'against' sub-graphs. An explicit assumption is made (and tested) that citations represent 'antagonistic' standpoints. An implicit assumption is made that there is a single topic per posting and a poster is either 'for' or 'against' that topic. Our own work suggests that the distribution of topical segments is not so trivially modeled. However, work is needed to clarify the nature of 'topics', their granularity (in terms of textual expression - do some topics require long tracts of text?) and their taxonomy.

Pang et al. (2002) describe a set of initial experiments using supervised text classification methods. The domain is movie reviews. An assumption is made that each review is about an individual movie (one that doesn't hold on inspecting the data). They evaluate a number of algorithms using a bag-of-words representation. Interestingly, the labeling of the data comes from user supplied star ratings common in the review genre. As these stars are part of the discourse, and consequently the context of the text, it is not clear what dependencies hold between the textual content of the documents and the stars. If I provide all my polar information by the star mechanism, I am free to use any language I choose to discuss the various aspects of the movie that I have strong feelings about. Dave et al. (2003) describe a similar approach applied to the domain of product reviews. Both of these papers report an exploration of the space of supervised learning algorithms and feature sets that improve performance. Interestingly, neither of them found any real benefit from linguistically motivated features including stemming and a simple transformation of tokens following a negating word.

The domain of movie reviews is certainly a popular one for work in the area of automated opinion mining. GoogleMovies provides an online classification mechanism for movie reviews. Again, as with Pang et al. (2002), there are issues in GoogleMovies to do with topic. Many of the 'reviews' encountered on the site are actually plot synopses.

It is notable that the literature to date refers to systems that make assumptions about the topicality of the texts being classified. Movie reviews are assumed to be restricted to one movie and about only that movie, work on consumer goods reviews makes similar assumptions and the network based methods described by Agrawal et al. (2003) use a simple method for selecting messages that contain content on a topic but which has no control for multi-topic messages or a notion of a 'main' topic.

The work described in this paper, and earlier work reported by Hurst and Nigam (2004), aims to explore the intersection of polar and topical language with the ultimate aim of deriving reliable models of attitudes toward predefined topics.

This work might be compared to Nasukawa and Yi (2003) which adopts a similar approach, but in which topicality is derived from the recognition of fixed terms in noun phrases as derived by shallow parsing. There are two aspects to comparing to the approach described in Nasukawa and Yi (2003), which relies wholly on shallow parsing methods, and that described here, which is a hybrid of shallow parsing and machine learning. Using shallow parsing for topic discovery limits the topics to those which are discovered by the shallow parser as noun chunks, and which can be mapped (i.e. interpreted) to appropriate semantic objects. The topics are limited to those that are captured in a certain grammatical relationship with the polar expression as determined by the grammatical patterns and the semantic lexicon. The advantage of this approach is that the associations have more precision as they are constrained grammatically. The machine learning approach admits a broader class of topic (no constraints on the topic being described by a single

noun phrase) and a more robust interpretation (when we view the classification as the discovery of a semantic object). Our hybrid approach does not rely on grammatical constraints for association, other than the sentential proximity assumption. Consequently, what we lose on precision we gain in recall.

Recent work (Engstrom, 2004) has looked at the association problem from a trained classifier point of view. The results reported there emphasize the problem of topicality when adopting a machine learning approach and gives support to the intuition that there is a strong linguistic aspect to the solution.

3. Classes of Polar Expression

Defining the language of polarity is a challenging task. However, when creating labeled data for training and evaluation, a definition is vital to making judgments.

We focus on two general aspects of expression. The first we term *opinion*. Statements used by the author to communicate opinion reflect a personal state of the author (Wiebe et al., 2001). The second we term *evaluative factual*. Statements with this aspect are objective but describe what is generally held to be a desirable or undesirable state. In other words, the first class reflects the users' personal evaluation and the second reflects the assumption of a social evaluation.[1]

For example*, I love this car* reflects the authors personal state and is an opinion. However, *The screen is broken* is clearly an objective expression describing what is generally accepted to be an undesirable state of affairs and is thus an evaluative factual. The notion of generally accepted evaluation is an interesting one as it is to some degree context and time dependent. At some point, the phrase *it has a color screen* will be positive. However, at some point later in time when all devices have a color screen, this will not necessarily be a polar phrase.

Opinion may be communicated indirectly via the use of emotive language—an indirect form of polar statement. For instance *The screen is frickin' broken again!* contains both emotive language as well as an objective reflection of the state of the world. This example shows that any single statement can easily mix both opinion and evaluative factive aspects of expression.

It is tempting to refer to intuition when describing opinionated or subjective language. However, the subtlety of expression requires that some lines be drawn even if they only serve to help us tackle a simpler problem. The literature in this novel field is often lacking in definition.[2] We have identified four features of the language of opinion that will be useful taxonomically for creating labeled data as well as constructing model driven analysis of polar language.

The first dimension that we call out is that of explicit versus implicit language. Explicit expressions of opinion include:

---

[1]The term *sentiment* is often used in this field. As we are including both opinion (subjective) and factual (objective) expressions, we defer to the term *polarity* indicating the common feature of *orientation*.

[2] Wiebe *et al*. (2001), however, provide a useful definition of subjective language in terms of textual representations of *private states*: they are represented in text either directly or indirectly. The class of direct statements allows us to build up lexical items corresponding to states, for example concern, disgust, etc. Indirect textual representations of private states appear as *expressive subjective elements* (Banfield, 1982). Private states are by definition subjective. However, descriptions of these states may be objective, thus Wiebe's work on sources and nested sources to chain the mentioning and author attribution of descriptions.

- Direct statements: *I like it*.
- Subjective evaluative language: *It is good.*

Implicit expression, on the other hand, involves sarcasm, irony, idiom and other deeper cultural referents:

- *It's really jumped the shark*. (cultural referent)
- *It's great if you like dead batteries*. (irony/sarcasm)
- *I give this two thumbs up*.

Of course, one might argue that lexicalized idioms, sarcasm, etc. are not distinct classes of expression but exist on a continuum with the explicit expression.

The next dimension is that of the matter about which an opinion is being expressed. This may be an established 'real world' concept, or it may be a hypothetical 'possible worlds' concept. For example *I love my new car* is an expression regarding an established object, where as *I am searching for the best possible deal* describes something that may or may not exist.

A third aspect that concerns us is that of modality, conditionals and temporal aspects of the language used. Statements such as *I might enjoy it* have a clear evaluative element (*enjoy it*) but do not express a definite opinion. *If it were larger*... describes a condition that perhaps must be met before the author admits an opinion. Such expression may also involve language indicating time: *The version coming in May is going to rock*!.

Finally, there is the matter of attribution. *I think you will like it* suggests that the author has a model of my likes and dislikes. It might mean that the author likes it and assumes I have the same taste, and so on. Quite possibly nobody likes it!

The above is an attempt to describe the space of expressions and their relationship to the author's communicative intent. For purposes of this paper, we define polar language to include both explicit and implicit expressions, 'real world' concepts and not hypothetical concepts, reject modal, conditional and temporal language and accept expressions regardless of attribution. This definition is used in driving the algorithmic approach to polarity recognition, and is consistent with the labeling criteria in the evaluation section.

4. Recognizing Polar Language

Our system begins by identifying polar language in individual sentences. To that end, a 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. Each item in the lexicon is a pairing of a word and its part-of-speech. Note that our lexicon contains possibly 'incorrect' terms that 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.

At run time, the input is tokenized. The tokenized input is then segmented into discrete chunks. The chunking phase consists of the following steps. The input is tagged with part of speech information. 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 that associate polarity with some topic, and those expressions that toggle the logical orientation of polar phrases (*I have never liked it*.). This last step conflates simple syntactic rules with semantic rules for propagating 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 '+' indicates a positive lexical item. Note that features are encoded in a simplified frame structure that 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 used to combine semantic features are: Predicative modification (*it is good*), Attributive modification (*a good car*), Equality (*it is a good car*), and Polar clause (*it broke my car*). Negations of the following types are captured by the system: Verbal attachment (*it is not good*, *it isn't good*), Adverbial negatives (*I never really liked it*, *it is never any good*), Determiners (*it is no good*), and Superordinate scope (*I don't think they made their best offer*).

5. Topic Detection in Online Messages

In the previous section we approached the task of assessing the polarity 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. For some types of topics, a well-written hand-built rule can suffice to identify a topic. (For example, imagine writing a rule to match the topic *Toyota Corolla*.) However, it's often the case that a topic is more accurately recognized from a complex language expression that is not easily captured by a rule. Thus, we often approach topic classification with machine learning techniques.

In the standard text classification approach, representative training examples are provided along with human judgments 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 application phases for the classifier. That is, the classifier assumes the example distribution remains constant before and after training.

Given our application of identifying topical polar sentences, the requisite distribution for training would be a distribution of hand-labeled sentences. However, hand-labeling individual sentences for building a classifier can be extremely expensive. For example, in our test domain fewer than 3% of all sentences were found to be topical. On the other hand, labeling entire messages provides much more labeled data with lower cost. Therefore, in our text mining system, a machine learning text classifier is trained to assess topicality on *whole messages*. We then use this classifier to accurately predict topicality at the sentence level, even though sentence distribution is quite different than whole message distribution.

5.1 Classifying Text with Winnow

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 (Littlestone, 1988; Blum, 1997; Dagan et al., 1997), 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 discards the ordering of words and notices only whether or not a word occurs in a document. Winnow learns a linear classifier of the form

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

where $c_w(x) = 1$ if word w occurs in document x and $c_w(x) = 0$ 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:

- Initialize all f_w to 1.
- For each labeled document x in the training set:
  - Calculate $h(x)$.
  - If the document is topical, but Winnow predicts irrelevant, update each weight f_w where $c_w(x) = 1$ by $f_w = f_w \times 2$
  - If the document is irrelevant, but Winnow predicts topical, update each weight f_w where $c_w(x) = 1$ by $f_w = f_w \div 2$

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 (Joachims, 1997) and k-Nearest Neighbor (Yang, 1999), two other state-of-the-art text classification algorithms. We use Winnow because it is more

computationally efficient than SVMs at training time and more computationally efficient than kNN at application time.

5.2 Using a Whole-document Classifier on 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, otherwise we label the sentence as irrelevant.

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 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, a 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.

6. The Intersection of Topic and Polarity

In the previous two sections we described fairly general-purpose tools for identifying polar expressions and topical expressions within sentences. However, each of these modules does so without any knowledge of the other. If a sentence is assessed as having both a polar expression and a topical expression, the independence of these judgments does not obviously lead us to conclude that the polar expression was with reference to the topic in question.

However, our system does assert 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. It is an empirical question whether or not this is a reasonable assumption to make. Our empirical results presented later demonstrate that this assumption generally holds with high accuracy in our domain of online messages.

The system we have described to this point is a shallow NLP-based system that assesses polar orientation of sentences and a machine learning-based text classifier for assessing topicality of

individual sentences. Sentences that are predicted as both topical and polar are then identified by the text analysis system as being polar about the topic. The next section evaluates the performance of the individual modules as well as the overall identification of topical polar sentences. The following section discusses how these results and algorithms might be combined to create a metric for aggregating these identified sentences into an overall score.

7. Empirical Analysis

In this section we describe a corpus for evaluating topical polarity and present experimental results showing that we can automatically identify topical sentences with positive or negative orientation.

7.1 Experimental Testbed

Using the Intelliseek message harvesting and text mining toolkit, we acquired about 34,000 messages from online resources (blogs, Usenet, and online message boards). Our message harvesting system collects messages in a particular domain (a vertical industry, such as 'automotive', or a specific set of products). From these messages, a trained topic classifier was built and a polarity lexicon was customized.

We hand-labeled a separate random sample of 822 messages for topicality, 88 (11%) which were topical. We hand-labeled all 1298 sentences in the 88 topical messages for topicality, polarity (positive and/or negative), and the correspondence between them. For the 7649 sentences in messages that were not topical, every sentence was automatically labeled as topically irrelevant, and thus containing no topical polarity either. Out of 8947 total sentences, just 147 (1.6%) have polar expression about the specified topic.

We evaluate the polarity module in isolation using the 1298 sentences with the complete polarity labels. We use the full dataset for evaluating topic and its combination with polarity. To evaluate the difficulty of the task for humans, we had a second labeler repeat the hand labeling on the 1298 sentences. Human agreement numbers are presented along with the algorithmic performance numbers in the next section.

7.2 Experimental Results

Below are several randomly selected examples of sentences predicted to be positive and negative about the topic of interest in our domain. This gives some idea for both the success of the algorithm, the types of errors it makes, and the sorts of marketing insights that can be gathered by quickly scanning topical polar sentences.

Sentences predicted as topical positive:

- The B&W display is great in the sun.
- 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.
- At that time, superior screen.
- 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.
- The screen is at 70 setting (255 max) which is for me the lowest comfortable setting.

	Algorithmic		Human Agreement	
	Precision	**Recall**	**Precision**	**Recall**
Positive	77%	43%	82%	78%
Negative	84%	16%	78%	74%

Table 1. Performance of the polarity analysis module compared to human agreement measured over messages relevant to a specific topic.

Sentences predicted as topical negative:

- Compared to the PRODUCT's screen this thing is very very poor.
- 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.
- 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.

Table 1 shows the results of the polarity module in isolation. Note that the precision of identifying both positive and negative topical language is very similar to the precision given by human agreement. This is indicative both of the difficulty of the task given the vagaries of language and the success of the algorithm at identifying these expressions. The automated recall, though, is significantly lower than human performance. One of the main reasons for this is the grammatical distinction between explicit and implicit polar language (c.f. the definitions section). Our approach to polar language detection is grammatical, suitable for many explicit expressions. However, the grammatical approach is less appropriate for the indirect language of implicit polarity that is generally more semantic in nature and may be better modeled by sets of cue phrases.

Also note that the recall of negative polarity is quite a bit lower than the recall of positive polarity. This confirms our anecdotal observation that language used for negative commentary is much more varied than that used for positive commentary. This observation in part drives the Bayesian approach to metric generation outlined in the next section.

Table 2 shows the performance of the trained topic classifier when measured over whole messages as well as individual sentences. Note that the precisions of applying the topic classifier on the message-level and on the sentence-level are very similar, while the sentence-level classifier has lower recall. This result is expected as described in an earlier section. In future work, we are looking towards various anaphora resolution techniques to improve the recall of the topic classifier on the sentence level.

	Algorithmic		Human Agreement	
Topicality	**Precision**	**Recall**	**Precision**	**Recall**
Message	71%	88%	---	---
Sentence	71%	77%	88%	70%

Table 2. Performance of the topic classifier compared to human agreement when run on both whole messages and individual sentences.

	Algorithmic		Human Agreement	
	Precision	**Recall**	**Precision**	**Recall**
Positive Topical	65%	43%	76%	62%
Negative Topical	65%	23%	80%	62%

Table 3. The performance of algorithmically identifying polar sentences about a specific topic compared to human performance.

We used the ground truth data to test our basic assumption for correlating topic and polar language. Our system assumes that any expression of polarity in a topical sentence is expressing polarity about the topic itself. We examined topical sentences that also contained positive polarity. 91% (90/99) of the time, the polarity was about the topic. In topical sentences that contained negative polarity, 80% (60/75) of the negative expressions concerned the topic. These statistics validate our basic assumption as a light-weight mechanism for correlating topic and polarity and represent an upper bound on our precision measurements for the recognition of topical polar sentences.

Table 3 shows the end results of identifying polar topical sentences given a polarity extraction system and a topic classifier. Unsurprisingly, the precision for both positive and negative topical extraction is lower than for positive and negative extraction in isolation. The correlation assumption between topic and polarity does not always hold, and the topic classifier adds in additional error. However, it is encouraging to notice that the drop in precision is less than would be suggested if all three sources of error were independent. This suggests that a certain amount of salient locality exists, where sentences that are topical are easier to identify polarity in, and vice-versa.

8. Metrics for Topic and Polarity

In this section we discuss how to use the polarity and topic modules to compile an aggregate score for a topic based on expressions contained in the data. We envision an aggregate topical orientation metric to be a function of:

- The total number of topical expressions
- The underlying frequency of topical expressions
- The underlying frequency of positive topical expressions
- The underlying frequency of negative topical expressions

For example, one very simplistic metric might be just the ratio of positive to negative expressions about a topic. The actual functional form of the metric may be driven more by marketplace requirements, but certain properties are very desirable. Ideally, such a metric would be able to propagate any uncertainty in the estimates of the various true frequencies. That is, the metric should not only support a single estimation of orientation, but also include some confidence in its measure. This will allow us to compare two or more competing topics of interest and say with a quantifiable probability that one topic is more favorably received than another.

Given the functional form of a polarity metric, one naive way of calculating the metric would be to plug in the values of the empirically measured frequencies of topic and polarity. If we believed that every topic classifier had the same accuracy, and that our polarity module performed equally in all domains, this might be a reasonable first pass. However, we believe that the performance of these modules will vary from domain-to-domain and topic-to-topic. Thus it is necessary to have some idea of the accuracy of each of our components in order to estimate the true frequencies from the empirical ones.

We plan to treat the estimation of the true underlying frequencies of topic and polarity as an exercise in Bayesian statistics. That is, we posit a probabilistic model for how the data are generated and use the data to estimate the parameters of the model. The model we propose has a set of parameters that are fixed for each domain and topic:

- With probability p_{topic} any expression will be written about specified topic.
- With probability $p_{pos|topic}$ any topical expression will be positive about the topic
- With probability $p_{neg|topic}$ any topical expression will be negative about the topic.

In practice, we observe expressions by seeing the output of our topic and polarity modules. These are not perfect observers, and they cloud the data by the following process:

- With probability $p_{topic,falsePos}$ we observe a true irrelevant expression as a topical one.
- With probability $p_{topic,falseNeg}$ we miss observing a true topical expression.
- With probability $p_{pos,falsePos}$ we observe a positive expression when there is none.
- With probability $p_{pos,falseNeg}$ we miss observing a positive expression.
- With probability $p_{neg,falsePos}$ we observe a negative expression when there is none.
- With probability $p_{neg,falseNeg}$ we miss observing a negative expression.

Using this explicit generative process of the data, we can use our observed data with standard statistical techniques to estimate the true underlying parameters of interest, p_{topic}, $p_{pos|topic}$, and $p_{neg|topic}$. These parameters are exactly the inputs needed by our hypothesized metric. Because we are working in the world of Bayesian statistics, we also get variances on our estimates that can be propagated through our metric. One nice property of Bayesian statistics is that the more data available for the estimation process, the smaller the measured variances become.

One requirement for this estimation process is the reliance on prior probability distributions for each of the model's parameters. We expect that uninformative priors will not serve well in this role. The whole point of the explicit modeling process is to get beyond the empirical estimates to a more robust estimate of the true underlying model parameters—the frequencies of polar topical expressions. To this end we plan to build empirical priors for each of our model parameters. We will do this by hand-labeling sets of data for a variety of topics and build empirical priors based on the distribution of the measured precision, recall and frequencies of each of our modules. These

informative priors will give us a more solid underpinning to our estimation process, resulting in more statistically valid metrics.

We have implemented a simplified version of the metric estimation described above. Our initial implementation assumes that the performance of polarity and topic is equivalent across topics, and thus has only a single set of estimation parameters. The metric is a 1-10 normalization of the ratios of the MAP estimates of the frequencies of positive and negative polarity for the topic. A 1.0 score indicates a very negative rating, a 10.0 indicates a very positive rating, and a 5.0 indicates a balanced rating of positive and negative. Table 4 shows the results of measuring polarity for location topics in a data set of messages about Caribbean destinations. Note that the

Location	Buzz %	Polarity
St. Lucia	2.3	10.0
Barbados	3.9	10.0
Aruba	6.9	8.0
Antigua	2.1	7.2
Grand Bahama	0.5	6.9
St. Bart's	0.7	6.4
Curacao	1.3	6.2
Jamaica	12.6	5.8
Grand Cayman	4.1	5.2
Belize	2.2	4.3
Cuba	6.3	4.1

Table 4. Example output of topical polarity scores. Buzz % shows the frequency of the topic, where the Polarity score is a 1-10 normalization of aggregate sentiment.

polarity score and the frequency of the topic are not correlated. By drilling down on these scores by reading the supporting positive and negative statements, an analyst can quickly determine that:

- Barbados and Aruba score well due to a good general opinion of dining out, snorkeling and beach activities.
- Cuba has a low score due to poor snorkeling and beach activities
- Grand Bahama's medium score comes from above average opinion of snorkeling, moderate opinion of dining out and a slightly lower opinion of beach activities.

9. Conclusions and Future Work

This paper has described continued work in the detection of topic and polarity. We have outlined a proposal for a metric relating the two that may be used as an aggregate measure of authorial sentiment on a particular topic drawn from online messages. We have described the components of a system working toward an implementation of this metric, and presented an evaluation of their performance with respect to a hand-labeled data set. In addition, we have tested the assumption that topical sentences that contain polar language are polar on that topic. We believe that our investigation supports this assumption.

There are a number of necessary steps required to complete the system and to improve the performance of its elements:

- Improve polarity recognition. Improvements can be made in both precision and recall. These issues may be addressed both by improvements and extensions to the underlying grammatical system and by the application of novel methods perhaps seeded by the results of this algorithm.
- Improve recall for topic. A number of assumptions regarding the collocation of topical sentences within paragraphs can be tested to improve the selection of topical sentences. For example, all the sentences in a paragraph starting with a topical sentence may be assumed to also be topical.
- Implement and test the full version of the metric described above.

10. References

Agrawal, R., Rajagopalan, S., Srikant, R., and Xu, Y. (2003) Mining newsgroups using networks arising from social behavior. In *Proceedings of the 12th World Wide Web Conference.*

Banfield, A. (1982) *Unspeakable Sentences.* Boston: Routledge and Kegan Paul.

Blum, A. (1997) Empirical support for Winnow and weighted-majority based algorithms: Results on a calendar scheduling domain. *Machine Learning* 26:5-23.

Dagan, I., Karov, Y, and Roth, D. (1997) Mistake-driven learning in text categorization. In *EMNLP '97, 2nd Conference on Empirical Methods in Natural Language Processing.*

Dave, K., Lawrence, S., and Pennock, D. M. (2003) Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th World Wide Web Conference.*

Engstrom, C. (2004) *Topic Dependence in Sentiment Classification.* Master's thesis, Cambridge University.

GoogleMovies. http://24.60.188.10:8080/demos/googlemovies/googlemovies.cgi.

Hurst, M., and Nigam, K. (2004) Retrieving topical sentiment from online document collections. In *Proceedings of the 11th Conference on Document Recognition and Retrieval.*

Joachims, T. (1998) Text categorization with support vector machines: Learning with many relevant features. In *Machine Learning: ECML-98 Tenth European Conference on Machine Learning,* 137-142.

Littlestone, N. (1998). Learning quickly when irrelevant features abound: A new linear-threshold algorithm. *Machine Learning* 2:285-318.

Nasukawa, T., and Yi, J. (2003) Sentiment analysis: Capturing favorability using natural language processing. In *Proceedings of K-CAP '03.*

Pang, B., Lee, L., and Vaithyanathan, S. (2002) Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of EMNLP 2002.*

Wiebe, J., Wilson, T., and Bell, M. (2001) Identifying collocations for recognizing opinions. In *Proceedings of ACL/EACL '01 Workshop on Collocation.*

Yang, Y. (1999) An evaluation of statistical approaches to text categorization. *Information Retrieval* 1(1/2): 67-88.