Machine Leaning Techniques for Sentiment Classification

Ishdeep Singla[1] Ramneet[2]
[1]Assistant Professor [2]Lecturer
[1]Chandigarh Group of Colleges, Mohali, Punjab, India [2]Longowal Polytechnic College, Mohali, Punjab, India

Abstract— We consider the problem of classifying documents not by topic, but by overall sentiment, e.g., determining whether a review is positive or negative. Using movie reviews as data, we find that standard machine learning techniques definitively outperform human-produced baselines. How- ever, the three machine learning methods we employed (Naive Bayes, maximum entropy classification, and support vector ma- chines) do not perform as well on sentiment classification as on traditional topic-based categorization. We conclude by examining factors that make the sentiment classification problem more challenging.
Key words: Sentiment Classification

I. INTRODUCTION

Today, very large amounts of information are avail- able in on-line documents. As part of the effort to better organize this information for users, researchers have been actively investigating the problem of automatic text categorization.

The bulk of such work has focused on topical categorization, attempting to sort documents according to their subject matter (e.g., sports vs. politics). However, recent years have seen rapid growth in on-line discussion groups and review sites (e.g., the New York Times' Books web page) where a crucial characteristic of the posted articles is their sentiment, or overall opinion towards the subject matter for example, whether a product review is positive or negative. Labeling these articles with their sentiment would provide succinct summaries to readers; indeed, these labels are part of the appeal and value-add of such sites as www.rottentomatoes.com, which both labels movie reviews that do not consume. Sentiment classification would also be helpful in business intelligence applications (e.g. MindfulEye's Lexant system1) and recommender systems (e.g., Terveen et al. (1997), Tatemura (2000)), where user input and feedback could be quickly summarized; in- deed, in general, free-form survey responses given in natural language format could be processed using sentiment categorization. Moreover, there are also potential applications to message filtering; for example, one might be able to use sentiment information to recognize and discard "flames"(Spertus, 1997).

In this paper, we examine the effectiveness of applying machine learning techniques to the sentiment classification problem. A challenging aspect of this problem that seems to distinguish it from traditional topic-based classification is that while topics are of- ten identifiable by keywords alone, sentiment can be expressed in a more subtle manner. For example, the sentence "How could anyone sit through this movie?" contains no single word that is obviously negative. (See Section 7 for more examples). Thus, sentiment seems to require more understanding than the usual topic-based classification. So, apart from presenting our results obtained via machine learning techniques, we also analyze the problem to gain a better under- standing of how difficult it is.

II. PREVIOUS WORK

This section briefly surveys previous work on non- topic-based text categorization.

One area of research concentrates on classifying documents according to their source or source style, with statistically-detected stylistic variation (Biber,1988) serving as an important cue. Examples include author, publisher (e.g., the New York Times vs. The Daily News),native-language background, and "brow" (e.g., high-brow vs. "popular", or low-brow) (Mosteller and Wallace, 1984; Argamon-Engelson ettain explicit rating indicators and normalizes the different rating schemes that individual reviewersal., 1998; Tomokiyo and Jones, 2001; Kessler et al.,1997).

Another, more related area of research is that of determining the genre of texts; subjective genres, such as "editorial", are often one of the possible categories (Karlgren and Cutting, 1994; Kessler et al., 1997; Finn et al., 2002). Other work explicitly attempts to find features indicating that subjective language is being used (Hatzivassiloglou and Wiebe, 2000; Wiebe et al., 2001). But, while techniques for genre categorization and subjectivity detection can help us recognize documents that express an opinion, they do not address our specific classification task of determining what that opinion actually is. Most previous research on sentiment-based classification has been at least partially knowledge-based. Some of this work focuses on classifying the semantic orientation of individual words or phrases, using linguistic heuristics or a pre-selected set of seed words (Hatzivassiloglou and McKeown, 1997; Turney and Littman, 2002). Past work on sentiment-based categorization of entire documents has often involved either the use of models inspired by cognitive linguistics (Hearst, 1992; Sack, 1994) or the manual or semi-manual construction of discriminant-word lexicons (Huettner and Subasic, 2000; Das and Chen, 2001; Tong,2001). Interestingly, our baseline experiments, described in Section 4, show that humans may not always have the best intuition for choosing discriminating words.

Turney's (2002) work on classification of reviews is perhaps the closest to ours.2 He applied a specific unsupervised learning technique based on the mutual information between document phrases and the words "excellent" and "poor", where the mutual information is computed using statistics gathered by a search engine. In contrast, we utilize several completely prior-knowledge-free supervised ma- chine learning methods, with the goal of understanding the inherent difficulty of the task.

All rights reserved by www.ijsrd.com

III. THE MOVIE-REVIEW DOMAIN

For our experiments, we chose to work with movie reviews. This domain is experimentally convenient because there are large on-line collections of such re- views, and because reviewers often summarize their overall sentiment with a machine-extractable rating indicator, such as a number of stars; hence, we did not need to hand-label the data for supervised learning or evaluation purposes. We also note that Turney (2002) found movie reviews to be the most difficult of several domains for sentiment classification, reporting an accuracy of 65.83% on a 120- document set (random-choice performance: 50%). But we stress that the machine learning methods and features we use are not specific to movie reviews, and should be easily applicable to other domains as long as sufficient training data exists. Our data source was the Internet Movie Database (IMDb) archive of the rec.arts.movies.reviews newsgroup.3 We selected only reviews where the author rating was expressed either with stars or some numerical value (other conventions varied too widely to allow for automatic processing). Ratings were automatically extracted and converted into one of three categories: positive, negative, or neutral. For the work described in this paper, we concentrated only on discriminating between positive and negative sentiment. To avoid domination of the corpus by a small number of prolific reviewers, we imposed a limit of fewer than 20 reviews per author per sentiment category, yielding a corpus of 752 negative and 1301 positive reviews, with a total of 144 reviewers represented. This dataset will be available on-line at http://www.cs.cornell.edu/people/pabo/-movie-review-data/ (the URL contains hyphens only around the word "review").

IV. UNDERSTANDING PROBLEM THOROUGHLY

Intuitions seem to differ as to the difficulty of the sentiment detection problem. An expert on using machine learning for text categorization predicted relatively low performance for automatic methods. On the other hand, it seems that distinguishing positive from negative reviews is relatively easy for humans, especially in comparison to the standard text categorization problem, where topics can be closely related. One might also suspect that there are certain words people tend to use to express strong sentiments, so that it might suffice to simply produce a list of such words by introspection and rely on them alone to classify the texts.

To test this latter hypothesis, we asked two graduate students in computer science to (independently) choose good indicator words for positive and negative sentiments in movie reviews. Their selections, shown in Figure 1, seem intuitively plausible. We then converted their responses into simple decision procedures that essentially count the number of the proposed positive and negative words in a given document. We applied these procedures to uniformly- distributed data, so that the random-choice baseline result would be 50%. As shown in Figure 1, the 2 Indeed, although our choice of title was completely independent of his, our selections were eerily similar.

	Proposed word lists	Accuracy	Ties
Human 1	positive: dazzling, brilliant, phenomenal, excellent, fantastic	68%	65%
Human 2	positive: gripping, mesmerizing, riveting, spectacular, cool, awesome, thrilling, badass, excellent, moving, exciting	44%	29%

Fig. 1: baseline results for human wordlists. Data: 700 positive and 700 negative reviews.

	Proposed word lists	Accuracy	Ties
Human 3 + stats	positive: love, wonderful, best, great, superb, still, beautiful	79%	6%

Fig. 2: Results for baseline using introspection and simple statistics of the data (include test data)

accuracy — percentage of documents classified correctly — for the human-based classifiers were 58% and 64%, respectively.4 Note that the tie rates — percentage of documents where the two sentiments were rated equally likely — are quite high5 (we chose a tie breaking policy that maximized the accuracy of the baselines).

While the tie rates suggest that the brevity of the human-produced lists is a factor in the relatively poor performance results, it is not the case that size alone necessarily limits accuracy. Based on a very preliminary examination of frequency counts in the entire corpus (including test data) plus introspection, we created a list of seven positive and seven negative words (including punctuation), shown in Figure 2. As that figure indicates, using these words raised the accuracy to 69%. Also, although this third list is of comparable length to the other two, it has a much lower tie rate of 16%. We further observe that some of the items in this third list, such as "?" or "still", would probably not have been proposed as possible candidates merely through introspection, although upon reflection one sees their merit (the question mark tends to occur in sentences like "What was the director thinking?"; "still" appears in sentences like "Still, though, it was worth seeing").

We conclude from these preliminary experiments that it is worthwhile to explore corpus-based techniques, rather than relying on prior intuitions, to select good indicator features and to perform sentiment classification in general. These experiments also pro- vide us with baselines for experimental comparison; in particular, the third baseline of 69% might actually be considered somewhat difficult to beat, since not claim that our list was the optimal set of four- teen words).

V. MACHINE LEARNING METHODS

Our aim in this work was to examine whether it suffices to treat sentiment classification simply as a special case of

All rights reserved by *www.ijsrd.com*

topic-based categorization (with the two "topics" being positive sentiment and negative sentiment), or whether special sentiment-categorization methods need to be developed. We experimented with three standard algorithms: Naive Bayes classification, maximum entropy classification, and sup- port vector machines. The philosophies behind these three algorithms are quite different, but each has been shown to be effective in previous text categorization studies.

To implement these machine learning algorithms on our document data, we used the following standard bag-of-features framework. Let {f1, . . . , fm} be a predefined set of m features that can appear in a document; examples include the word "still" or the bigram "really stinks". Let ni(d) be the number of times fi occurs in document d. Then, each document d is represented by the document vector d~ := (n1(d), n2(d), . . . , nm(d))

VI. Discussions and Conclusion

The results produced via machine learning techniques are quite good in comparison to the human- generated baselines discussed in Section 4. In terms of relative performance, Naive Bayes tends to do the worst and SVMs tend to do the best, although the differences aren't very large.

On the other hand, we were not able to achieve accuracies on the sentiment classification problem com- parable to those reported for standard topic-based categorization, despite the several different types of features we tried. Unigram presence information turned out to be the most effective; in fact, none of the alternative features we employed provided consistently better performance once unigram presence was incorporated. Interestingly, though, the superiority of presence information in comparison to frequency information in our setting contradicts previous observations made in topic-classification work (McCallum and Nigam, 1998).

What accounts for these two differences — difficulty and types of information proving useful — between topic and sentiment classification, and how might we improve the latter? To answer these questions, we examined the data further. (All examples below are drawn from the full 2053-document corpus.)

As it turns out, a common phenomenon in the documents was a kind of "thwarted expectations" narrative, where the author sets up a deliberate contrast to earlier discussion: for example, "This film should be brilliant. It sounds like a great plot, the actors are first grade, and the supporting cast is good as well, and Stallone is attempting to deliver a good performance. However, it can't hold up" or "I hate the Spice Girls. [3 things the author hates about them]... Why I saw this movie is a really, really, really long story, but I did, and one would think I'd despise every minute of it. But Okay, I'm really ashamed of it, but I enjoyed it. I mean, I admit it's a really awful movie ...the ninth floor of hell...The plot is such a mess that it's terrible. But I loved it." 15

In these examples, a human would easily detect the true sentiment of the review, but bag-of-features classifiers would presumably find these instances difficult, since there are many words indicative of the opposite sentiment to that of the entire review. Fundamentally, it seems that some form of discourse analysis is necessary (using more sophisticated tech)

This phenomenon is related to another common theme, that of "a good actor trapped in a bad movie": "AN AMERICAN WEREWOLF IN PARIS is a failed attempt... Julie Delpy is far too good for this movie. Her imbues Serafine with spirit, spunk, and humanity. This isn't necessarily a good thing, since it prevents us from relaxing and enjoying AN AMERICAN WEREWOLF IN PARIS as a completely mindless, campy entertainment experience. Delpy's injection of class into an otherwise classless production raises the specter of what this film could have been with a better script and a better cast ... She was radiant, charismatic, and effective ...."

References

[1] Shlomo Argamon-Engelson, Moshe Koppel, and Galit Avneri. 1998. Style-based text categoriza- tion: What newspaper am I reading? In Proc. of the AAAI Workshop on Text Categorization, pages 1–4.
[2] Adam L. Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra. 1996. A maximum entropy approach to natural language processing. Computational Linguistics, 22(1):39–71.
[3] Douglas Biber. 1988. Variation across Speech and Writing. Cambridge University Press.
[4] Stanley Chen and Ronald Rosenfeld. 2000. A survey of smoothing techniques for ME models. IEEE Trans. Speech and Audio Processing, 8(1):37–50.
[5] Sanjiv Das and Mike Chen. 2001. Yahoo! for Amazon: Extracting market sentiment from stock message boards. In Proc. of the 8th Asia Pacific Finance Association Annual Conference (APFA-2001).
[6] Stephen Della Pietra, Vincent Della Pietra, and John Lafferty. 1997. Inducing features of random fields. IEEE Transactions on Pattern Analysis and Ma- chine Intelligence, 19(4):380–393.
[7] Pedro Domingos and Michael J. Pazzani. 1997. On the optimality of the simple Bayesian classifier un- der zero-one loss. Machine Learning, 29(2-3):103–130.
[8] Aidan Finn, Nicholas Kushmerick, and Barry Smyth. 2002. Genre classification and domain transfer for information filtering. In Proc. of the Eu-ropean Colloquium on Information Retrieval Re-search, pages 353–362, Glasgow.
[9] Vasileios Hatzivassiloglou and Kathleen McKeown.1997. Predicting the semantic orientation of adjectives. In Proc. of the 35th ACL/8th EACL, pages 174–181.
[10] Vasileios Hatzivassiloglou and Janyce Wiebe. 2000. Effects of adjective orientation and gradability on sentence subjectivity. In Proc. of COLING.
[11] Marti Hearst. 1992. Direction-based text interpretation as an information access refinement. In Paul Jacobs, editor, Text-Based Intelligent Sys- tems. Lawrence Erlbaum Associates.
[12] Alison Huettner and Pero Subasic. 2000. Fuzzy typing for document management. In ACL

All rights reserved by *www.ijsrd.com*

2000 Companion Volume: Tutorial Abstracts and Demonstration Notes, pages 26–27.
[13] Thorsten Joachims. 1998. Text categorization with support vector machines: Learning with many rel- evant features. In Proc. of the European Confer- ence on Machine Learning (ECML), pages 137–142.
[14] Thorsten Joachims. 1999. Making large-scale SVM learning practical. In Bernhard Sch¨olkopf and Alexander Smola, editors, Advances in Kernel Methods - Support Vector Learning, pages 44–56. MIT Press.
[15] Jussi Karlgren and Douglass Cutting. 1994. Recognizing text genres with simple metrics using discriminant analysis. In Proc. of COLING.
[16] Brett Kessler, Geoffrey Nunberg, and Hinrich Schu¨tze. 1997. Automatic detection of text genre. In Proc. of the 35th ACL/8th EACL, pages 32–38.
[17] David D. Lewis. 1998. Naive (Bayes) at forty: The independence assumption in information retrieval. In Proc. of the European Conference on Machine Learning (ECML), pages 4–15. Invited talk.
[18] Andrew McCallum and Kamal Nigam. 1998. A comparison of event models for Naive Bayes text classification. In Proc. of the AAAI-98 Workshop on Learning for Text Categorization, pages 41–48.
[19] Frederick Mosteller and David L. Wallace. 1984. Applied Bayesian and Classical Inference: The Case of the Federalist Papers. Springer-Verlag.
[20] Kamal Nigam, John Lafferty, and Andrew McCallum. 1999. Using maximum entropy for text classification. In Proc. of the IJCAI-99 Workshop on Machine Learning for Information Filtering, pages 61–67.
[21] Ted Pedersen. 2001. A decision tree of bigrams is an accurate predictor of word sense. In Proc. of the Second NAACL, pages 79–86.
[22] Warren Sack. 1994. On the computation of point of view. In Proc. of the Twelfth AAAI, page 1488. Student abstract.
[23] Ellen Spertus. 1997. Smokey: Automatic recognition of hostile messages. In Proc. of Innova- tive Applications of Artificial Intelligence (IAAI), pages 1058–1065.
[24] Junichi Tatemura. 2000. Virtual reviewers for collaborative exploration of movie reviews. In Proc. of the 5th International Conference on Intelligent User Interfaces, pages 272–275.
[25] Loren Terveen, Will Hill, Brian Amento, David McDonald, and Josh Creter. 1997. PHOAKS: A sys- tem for sharing recommendations. Communica- tions of the ACM, 40(3):59–62. Laura Mayfield Tomokiyo and Rosie Jones. 2001. You're not from round here, are you? Naive Bayes detection of non-native utterance text. In Proc. of the Second NAACL, pages 239–246.
[26] Richard M. Tong. 2001. An operational system for detecting and tracking opinions in on-line discus- sion. Workshop note, SIGIR 2001 Workshop on Operational Text Classification.
[27] Peter D. Turney and Michael L. Littman. 2002. Unsupervised learning of semantic orientation from a hundred-billion-word corpus. Technical Report EGB-1094, National Research Council Canada.
[28] Peter Turney. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In Proc. of the ACL.
[29] Janyce M. Wiebe, Theresa Wilson, and Matthew Bell. 2001. Identifying collocations for recognizing opinions. In Proc. of the ACL/EACL Workshop on Collocation.
[30] Yorick Wilks and Mark Stevenson. 1998. The grammar of sense: Using part-of-speech tags as a first step in semantic disambiguation. Journal of Nat- ural Language Engineering, 4(2):135–144.

All rights reserved by *www.ijsrd.com*