Vocabulary Choice as an Indicator of Perspective

Beata Beigman Klebanov, Eyal Beigman, Daniel Diermeier
Northwestern University and Washington University in St. Louis
beata,d-diermeier@northwestern.edu, beigman@wustl.edu

Abstract

We establish the following characteristics of the task of perspective classification: (a) using term frequencies in a document does not improve classification achieved with absence/presence features; (b) for datasets allowing the relevant comparisons, a small number of top features is found to be as effective as the full feature set and indispensable for the best achieved performance, testifying to the existence of perspective-specific keywords. We relate our findings to research on word frequency distributions and to discourse analytic studies of perspective.

1 Introduction

We address the task of perspective classification. Apart from the spatial sense not considered here, perspective can refer to an agent's role (doctor vs patient in a dialogue), or understood as "a particular way of thinking about something, especially one that is influenced by one's beliefs or experiences," stressing the manifestation of one's broader perspective in some specific issue, or "the state of one's ideas, the facts known to one, etc., in having a meaningful interrelationship," stressing the meaningful connectedness of one's stances and pronouncements on possibly different issues.[1]

Accordingly, one can talk about, say, opinion on a particular proposed legislation on abortion within pro-choice or pro-life perspectives; in this case, perspective essentially boils down to opinion in a particular debate. Holding the issue constant but relaxing the requirement of a debate on a specific document, we can consider writings from pro- and con- perspective, in, for example, the death penalty controversy over a course of a period of time. Relaxing the issue specificity somewhat, one can talk about perspectives of people on two sides of a conflict; this is not opposition or support for any particular proposal, but ideas about a highly related cluster of issues, such as Israeli and Palestinian perspectives on the conflict in all its manifestations. Zooming out even further, one can talk about perspectives due to certain life contingencies, such as being born and raised in a particular culture, region, religion, or political tradition, such perspectives manifesting themselves in certain patterns of discourse on a wide variety of issues, for example, views on political issues in the Middle East from Arab vs Western observers.

In this article, we consider perspective at all the four levels of abstraction. We apply the same types of models to all, in order to discover any common properties of perspective classification. We contrast it with text categorization and with opinion classification by employing models routinely used for such tasks. Specifically, we consider models that use term frequencies as features (usually found to be superior for text categorization) and models that use term absence/presence (usually found to be superior for opinion classification). We motivate our hypothesis that presence/absence features would be as good as or better than frequencies, and test it experimentally. Secondly, we investigate the question of feature redundancy often observed in text categorization.

2 Vocabulary Selection

A line of inquiry going back at least to Zipf strives to characterize word frequency distributions in texts and corpora; see Baayen (2001) for a survey. One of the findings in this literature is that a multinomial (called "urn model" by Baayen) is not a good model for word frequency distributions. Among the many proposed remedies (Baayen, 2001; Jansche, 2003; Baroni and Evert, 2007; Bhat and Sproat, 2009), we would like to draw attention to the following insight articulated

[1] Google English Dictionary, Dictionary.com

Proceedings of the ACL 2010 Conference Short Papers, pages 253–257,
Uppsala, Sweden, 11-16 July 2010.

most clearly in Jansche (2003). Estimation is improved if texts are construed as being generated by two processes, one choosing which words would appear at all in the text, and then, for words that have been chosen to appear, how many times they would in fact appear. Jansche (2003) describes a two-stage generation process: (1) Toss a z-biased coin; if it comes up heads, generate 0; if it comes up tails, (2) generate according to $F(\theta)$, where $F(\theta)$ is a negative binomial distribution and z is a parameter controlling the extent of zero-inflation.

The postulation of two separate processes is effective for predicting word frequencies, but is there any meaning to the two processes? The first process of deciding on the vocabulary, or word types, for the text – what is its function? Jansche (2003) suggests that the zero-inflation component takes care of the multitude of vocabulary words that are not "on topic" for the given text, including taboo words, technical jargon, proper names. This implies that words that are chosen to appear are all "on topic". Indeed, text segmentation studies show that tracing recurrence of words in a text permits topical segmentation (Hearst, 1997; Hoey, 1991). Yet, if a person compares abortion to *infanticide* – are we content with describing this word as being merely "on topic," that is, having a certain probability of occurrence once the topic of abortion comes up? In fact, it is only likely to occur if the speaker holds a pro-life perspective, while a pro-choicer would avoid this term.

We therefore hypothesize that the choice of vocabulary is not only a matter of topic but also of perspective, while word *recurrence* has mainly to do with the topical composition of the text. Therefore, tracing word frequencies is not going to be effective for perspective classification beyond noting the mere presence/absence of words, differently from the findings in text categorization, where frequency-based features usually do better than boolean features for sufficiently large vocabulary sizes (McCallum and Nigam, 1998).

3 Data

Partial Birth Abortion (PBA) debates: We use transcripts of the debates on Partial Birth Abortion Ban Act on the floors of the US House and Senate in 104-108 Congresses (1995-2003). Similar legislation was proposed multiple times, passed the legislatures, and, after having initially been vetoed by President Clinton, was signed into law by President Bush in 2003. We use data from 278 legislators, with 669 speeches in all. We take only one speech per speaker per year; since many serve multiple years, each speaker is represented with 1 to 5 speeches. We perform 10-fold cross-validation splitting by speakers, so that all speeches by the same speaker are assigned to the same fold and testing is always inter-speaker.

When deriving the label for perspective, it is important to differentiate between a particular legislation and a pro-choice / pro-life perspective. A pro-choice person might still support the bill: "I am pro-choice, but believe late-term abortions are wrong. Abortion is a very personal decision and a woman's right to choose whether to terminate a pregnancy subject to the restrictions of Roe v. Wade must be protected. In my judgment, however, the use of this particular procedure cannot be justified." (Rep. Shays, R-CT, 2003). To avoid inconsistency between vote and perspective, we use data from pro-choice and pro-life non-governmental organizations, NARAL and NRLC, that track legislators' votes on abortion-related bills, showing the percentage of times a legislator supported the side the organization deems consistent with its perspective. We removed 22 legislators with a mixed record, that is, those who gave 20-60% support to one of the positions.[2]

Death Penalty (DP) blogs: We use University of Maryland Death Penalty Corpus (Greene and Resnik, 2009) of 1085 texts from a number of pro- and anti-death penalty websites. We report 4-fold cross-validation (DP-4) using the folds in Greene and Resnik (2009), where training and testing data come from different websites for each of the sides, as well as 10-fold cross-validation performance on the entire corpus, irrespective of the site.[3]

Bitter Lemons (BL): We use the GUEST part of the BitterLemons corpus (Lin et al., 2006), containing 296 articles published in 2001-2005 on http://www.bitterlemons.org by more than 200 different Israeli and Palestinian writers on issues related to the conflict.

Bitter Lemons International (BL-I): We collected 150 documents each by a different per-

[2]Ratings are from: http://www.OnTheIssues.org/. We further excluded data from Rep. James Moran, D-VA, as he changed his vote over the years. For legislators rated by neither NRLC nor NARAL, we assumed the vote aligns with the perspective.

[3]The 10-fold setting yields almost perfect performance likely due to site-specific features beyond perspective per se, hence we do not use this setting in subsequent experiments.

son from either Arab or Western perspectives on Middle Eastern affairs in 2003-2009 from http://www.bitterlemons-international.org/. The writers and interviewees on this site are usually former diplomats or government officials, academics, journalists, media and political analysts.[4] The specific issues cover a broad spectrum, including public life, politics, wars and conflicts, education, trade relations in and between countries like Lebanon, Jordan, Iraq, Egypt, Yemen, Morocco, Saudi Arabia, as well as their relations with the US and members of the European Union.

3.1 Pre-processing

We are interested in perspective manifestations using common English vocabulary. To avoid the possibility that artifacts such as names of senators or states drive the classification, we use as features words that contain only lowercase letters, possibly hyphenated. No stemming is performed, and no stopwords are excluded.[5]

Table 1: Summary of corpora

Data	#Docs	#Features	# CV folds
PBA	669	9.8 K	10
BL	296	10 K	10
BL-I	150	9 K	10
DP	1085	25 K	4

4 Models

For generative models, we use two versions of Naive Bayes models termed *multi-variate Bernoulli* (here, NB-BOOL) and *multinomial* (here, NB-COUNT), respectively, in McCallum and Nigam (1998) study of event models for text categorization. The first records presence/absence of a word in a text, while the second records the number of occurrences. McCallum and Nigam (1998) found NB-COUNT to do better than NB-BOOL for sufficiently large vocabulary sizes for text categorization by topic. For discriminative models, we use linear SVM, with presence-absence, normalized frequency, and tfidf feature weighting. Both types of models are commonly used for text classification tasks. For example, Lin et al. (2006) use NB-COUNT and SVM-NORMF for perspective classification; Pang et al. (2002) consider most and Yu et al. (2008) all of the above for related tasks of movie review and political party classification. We use SVM^{light} (Joachims, 1999) for SVM and WEKA toolkit (Witten and Frank, 2005; Hall et al., 2009) for both version of Naive Bayes. Parameter optimization for all SVM models is performed using grid search on the training data separately for each partition into train and test data.[6]

5 Results

Table 2 summarizes the cross-validation results for the four datasets discussed above. Notably, the SVM-BOOL model is either the best or not significantly different from the best performing model, although the competitors use more detailed textual information, namely, the count of each word's appearance in the text, either raw (NB-COUNT), normalized (SVM-NORMF), or combined with document frequency (SVM-TFIDF).

Table 2: Classification accuracy. Scores significantly different from the best performance ($p_{2t}<0.05$ on paired t-test) are given an asterisk.

Data	NB		SVM		
	BOOL	COUNT	BOOL	NORMF	TFIDF
PBA	*0.93	0.96	0.96	0.96	0.97
DP-4	0.82	0.82	0.83	0.82	0.72[7]
DP-10	*0.88	*0.93	0.98	*0.97	*0.97
BL	0.89	0.88	0.89	0.86	0.84
BL-I	0.68	0.66	0.73	0.65	0.65

We conclude that there is no evidence for the relevance of the frequency composition of the text for perspective classification, for all levels of venue- and topic-control, from the tightest (PBA debates) to the loosest (Western vs Arab authors on Middle Eastern affairs). This result is a clear indication that perspective classification is quite different from text categorization by topic, where count-based features usually perform better than boolean features. On the other hand, we have not

[4]We excluded Israeli, Turkish, Iranian, Pakistani writers as not clearly representing either perspective.

[5]We additionally removed words containing *support*, *oppos*, *sustain*, *overrid* from the PBA data, in order not to inflate the performance on perspective classification due to the explicit reference to the upcoming vote.

[6]Parameter c controlling the trade-off between errors on training data and margin is optimized for all datasets, with the grid $c = \{10^{-6}, 10^{-5}, \ldots, 10^{5}\}$. On the DP data parameter j controlling penalties for misclassification of positive and negative cases is optimized as well ($j = \{10^{-2}, 10^{-1}, \ldots, 10^{2}\}$), since datasets are unbalanced (for example, there is a fold with 27%-73% split).

[7]Here SVM-TFIDF is doing somewhat better than SVM-BOOL on one of the folds and much worse on two other folds; paired t-test with just 4 pairs of observations does not detect a significant difference.

observed that boolean features are reliably better than count-based features, as reported for the sentiment classification task in the movie review domain (Pang et al., 2002).

We note the low performance on BL-I, which could testify to a low degree of lexical consolidation in the Arab vs Western perspectives (more on this below). It is also possible that the small size of BL-I leads to overfitting and low accuracies. However, PBA subset with only 151 items (only 2002 and 2003 speeches) is still 96% classifiable, so size alone does not explain low BL-I performance.

6 Consolidation of perspective

We explore feature redundancy in perspective classification.We first investigate retention of only N best features, then elimination thereof. As a proxy of feature quality, we use the weight assigned to the feature by the SVM-BOOL model based on the training data. Thus, to get the performance with N best features, we take the $\frac{N}{2}$ highest and lowest weight features, for the positive and negative classes, respectively, and retrain SVM-BOOL with these features only.[8]

Table 3: Consolidation of perspective. Nbest shows the smallest N and its proportion out of all features for which the performance of SVM-BOOL with only the best N features is not significantly inferior (p_{1t}>0.1) to that of the full feature set. No-Nbest shows the largest number N for which a model *without* N best features is not significantly inferior to the full model. N=$\{50, 100, 150, \ldots, 1000\}$; for DP and BL-I, additionally N=$\{1050, 1100, ..., 1500\}$; for PBA, additionally N=$\{10, 20, 30, 40\}$.

Data	Nbest		No-Nbest	
	N	%	N	%
PBA	250	2.6%	10	<1%
BL	500	4.9%	100	<1%
DP	100	<1%	1250	5.2%
BL-I	200	2.2%	950	11%

We observe that it is generally sufficient to use a small percentage of the available words to obtain the same classification accuracy as with the full feature set, even in high-accuracy cases such as PBA and BL. The effectiveness of a small subset of features is consistent with the observation in the discourse analysis studies that rivals in long-lasting controversies tend to consolidate their vocabulary and signal their perspective with certain *stigma words* and *banner words*, that is, specific keywords used by a discourse community to implicate adversaries and to create sympathy with own perspective, respectively (Teubert, 2001). Thus, in abortion debates, using *infanticide* as a synonym for abortion is a pro-life stigma. Note that this does not mean the rest of the features are not informative for classification, only that they are redundant with respect to a small percentage of top weight features.

When N best features are eliminated, performance goes down significantly with even smaller N for PBA and BL datasets. Thus, top features are not only effective, they are also crucial for accurate classification, as their discrimination capacity is not replicated by any of the other vocabulary words. This finding is consistent with Lin and Hauptmann (2006) study of perspective vs topic classification: While topical differences between two corpora are manifested in difference in distributions of great many words, they observed little perspective-based variation in distributions of most words, apart from certain words that are preferentially used by adherents of one or the other perspective on the given topic.

For DP and BL-I datasets, the results seem to suggest perspectives with more diffused keyword distribution (No-NBest figures are higher). We note, however, that feature redundancy experiments are confounded in these cases by either a low power of the paired t-test with only 4 pairs (DP) or by a high variance in performance among the 10 folds (BL-I), both of which lead to numerically large discrepancy in performance that is not deemed significant, making it easy to "match" the full set performance with small-N best features as well as without large-N best features. Better comparisons are needed in order to verify the hypothesis of low consolidation.

In future work, we plan to experiment with additional features. For example, Greene and Resnik (2009) reported higher classification accuracies for the DP-4 data using syntactic frames in which a selected group of words appeared, rather than mere presence/absence of the words. Another direction is exploring words as members of semantic fields – while word use might be insufficiently consistent within a perspective, selection of a semantic domain might show better consistency.

[8] We experimented with the mutual information based feature selection as well, with generally worse results.

References

Herald Baayen. 2001. *Word frequency distributions*. Dordrecht: Kluwer.

Marco Baroni and Stefan Evert. 2007. Words and Echoes: Assessing and Mitigating the Non-Randomness Problem in Word Frequency Distribution Modeling. In *Proceedings of the ACL*, pages 904–911, Prague, Czech Republic.

Suma Bhat and Richard Sproat. 2009. Knowing the Unseen: Estimating Vocabulary Size over Unseen Samples. In *Proceedings of the ACL*, pages 109–117, Suntec, Singapore, August.

Stephan Greene and Philip Resnik. 2009. More than Words: Syntactic Packaging and Implicit Sentiment. In *Proceedings of HLT-NAACL*, pages 503–511, Boulder, CO, June.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringe, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: An update. *SIGKDD Explorations*, 11(1).

Marti Hearst. 1997. TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages. *Computational Linguistics*, 23(1):33–64.

Michael Hoey. 1991. *Patterns of Lexis in Text*. Oxford University Press.

Martin Jansche. 2003. Parametric Models of Linguistic Count Data. In *Proceedings of the ACL*, pages 288–295, Sapporo, Japan, July.

Thorsten Joachims. 1999. Making large-scale SVM learning practical. In B. Schlkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT Press.

Wei-Hao Lin and Alexander Hauptmann. 2006. Are these documents written from different perspectives? A test of different perspectives based on statistical distribution divergence. In *Proceedings of the ACL*, pages 1057–1064, Morristown, NJ, USA.

Wei-Hao Lin, Theresa Wilson, Janyce Wiebe, and Alexander Hauptmann. 2006. Which side are you on? Identifying perspectives at the document and sentence levels. In *Proceedings of CoNLL*, pages 109–116, Morristown, NJ, USA.

Andrew McCallum and Kamal Nigam. 1998. A comparison of event models for Naive Bayes text classification. In *Proceedings of AAAI-98 Workshop on Learning for Text Categorization*, pages 41–48, Madison, WI, July.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. In *Proceedings of EMNLP*, Philadelphia, PA, July.

Wolfgang Teubert. 2001. A Province of a Federal Superstate, Ruled by an Unelected Bureaucracy – Keywords of the Euro-Sceptic Discourse in Britain. In Andreas Musolff, Colin Good, Petra Points, and Ruth Wittlinger, editors, *Attitudes towards Europe: Language in the unification process*, pages 45–86. Ashgate Publishing Ltd, Hants, England.

Ian H. Witten and Eibe Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2 edition.

Bei Yu, Stefan Kaufmann, and Daniel Diermeier. 2008. Classifying party affiliation from political speech. *Journal of Information Technology and Politics*, 5(1):33–48.