Innocent Until Proven Guilty: Linguistic Predictors of Court Verdicts

Anastasia Svetlichnaya
CS 229: Machine Learning – Professor Andrew Ng
December 10th, 2010

Abstract

Prediction of court verdicts from linguistic and contextual features of 14,000 trial transcripts is attempted using two models: maximum entropy and support vector machine. Performance of the two classifiers is compared, varying parameters and the set of features. The highest accuracies achieved for the MaxEnt and SVM classifiers are 69.8% and 69.6% respectively, hence MaxEnt is very slightly more successful. The inclusion of contextual features boosts accuracy to 85% on MaxEnt—the particular model set is a good fit for this data set. Classification is approximately 20% better than chance using linguistic features alone, suggesting that it is feasible to predict a court of law's perception of a speaker as guilty or innocent using machine learning methods.

Introduction

Sentiment analysis and automatic identification of lying, sarcasm, and even flirtation have received considerable attention as applications of computational linguistics and natural language processing. Computers can detect these nontrivial aspects of human communication—a speaker's feelings about a subject, higher order intentions like deception or humor—with increasing accuracy. But can machines predict whether a speaker will be perceived as guilty or innocent based on what is said in a court of law? The aim of this supervised learning project is two-fold: (1) develop a classification model tailored to trial verdict prediction on the data set of 1830-1913 London court trials and (2) more generally, explore which linguistic features may bias perception of a speaker towards guilt versus innocence. The training corpus consists of the proceedings of Old Bailey, London's central criminal court. Full transcripts of over 200,000 trials from 1674 to 1913 are available online (www.oldbaileyonline.org). Pre-1830 trials were eliminated to maintain relevancy and the consistency of the vocabulary. The remaining XML-tagged transcripts were parsed into CSV files, producing 38,880 usable trials for a total of over 300,000 utterances. Each utterance is associated with a speaker's name and role and annotated with the lemma and part-of-speech of each word. Each trial is further associated with meta-data including the date and jurisdiction of the court session, the names of all the participants, the specific offense type, and the punishment assigned. While this information does not reveal linguistic style, it is important for contextualizing the proceedings—the prior probability of a guilty versus innocent verdict is expected to depend the leniency of the court members in a certain jurisdiction, the time period, the seriousness of the charge, and so on. Furthermore, classifying based on contextual features only provides a useful baseline for measuring the predictive power of the language alone.

Features

Feature selection for this classification problem is nontrivial and proved rather time-consuming. The categorical contextual features consisted of the trial-level meta-data

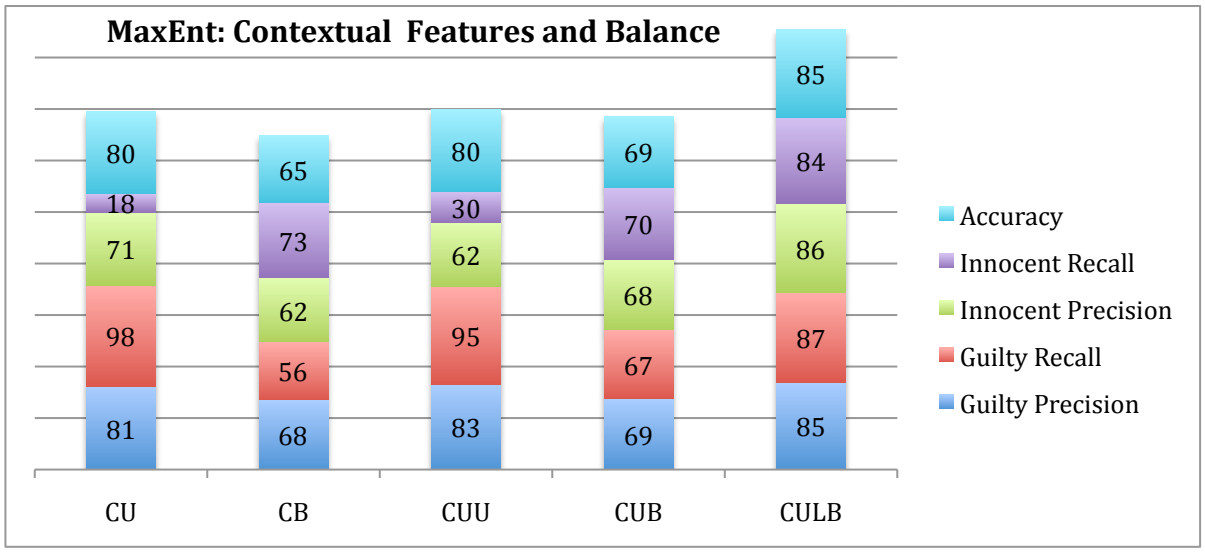
described above. The real-valued linguistic features were chosen based on past findings on lie detection and perception of intention/blame in psychology and linguistics (Adams and Jarvis, 2006; Fausey and Matlock, 2010; Larker and Zakolyukina, 2010; and more). These included passive verbs, imperfective verbs, hypotheticals, and speech acts (e.g. "I swear"). Psychosocial dictionaries—General Inquirer and Linguistic Inquiry and Word Count (Pennebaker et al., 2007)—were used for sentiment analysis (positive, negative, anger) and to construct regular expressions aggregating words and phrases by category or tone (certainty, cognitive mechanisms, money, quantifiers, etc.). Higher-level features included speakers' loquaciousness, average sentence length, and vocabulary sophistication (based on frequency of words used relative to the rest of the trial). A final subset of features came from initial estimates via odds ratios[1] and personal curiosity, such as wondering if the proportion of references to religion affects the verdict.

Each of the features is normalized by the length of the trial or a large constant to keep attributes in the [0,1] range to be compatible with the SVM method in LIBLINEAR (Fan et al., 2008). Since the length of the transcripts in the corpus varies from zero to several thousands of words, this method also avoids overweighting longer trials. To investigate the significance of not only *what* is said but also *who* is speaking, each feature is relativized by the speaker's role (witness, defendant, court, jury, victim, prisoner, prosecuting lawyer, or defending lawyer). This yields 333 features total.

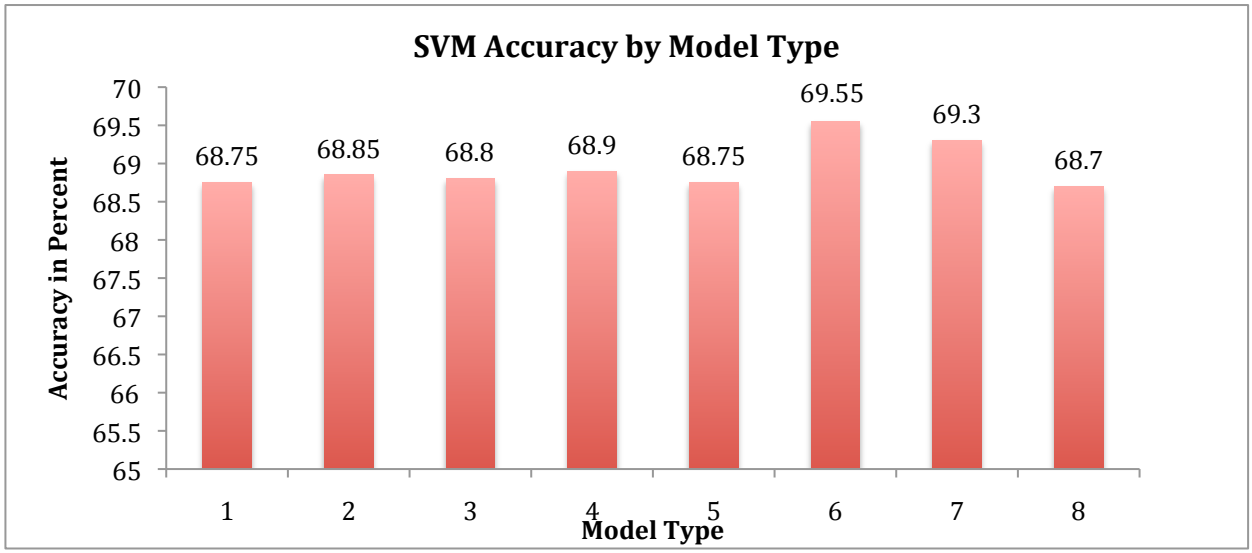
[1]Odds ratios were computed for each word w in the vocabulary of the corpus using the formula [g / (1 – g)] / [ng / (1 – ng)], where g and ng represent the probability (frequency) of w in guilty and innocent trials, respectively. An odds ratio > 1 indicates a bias for a guilty verdict and < 1 for an innocent verdict. The words with the highest and lowest odds ratios were analyzed for patterns to be implemented as features.

Results (number of features = F, train size = TR, test size = TS)

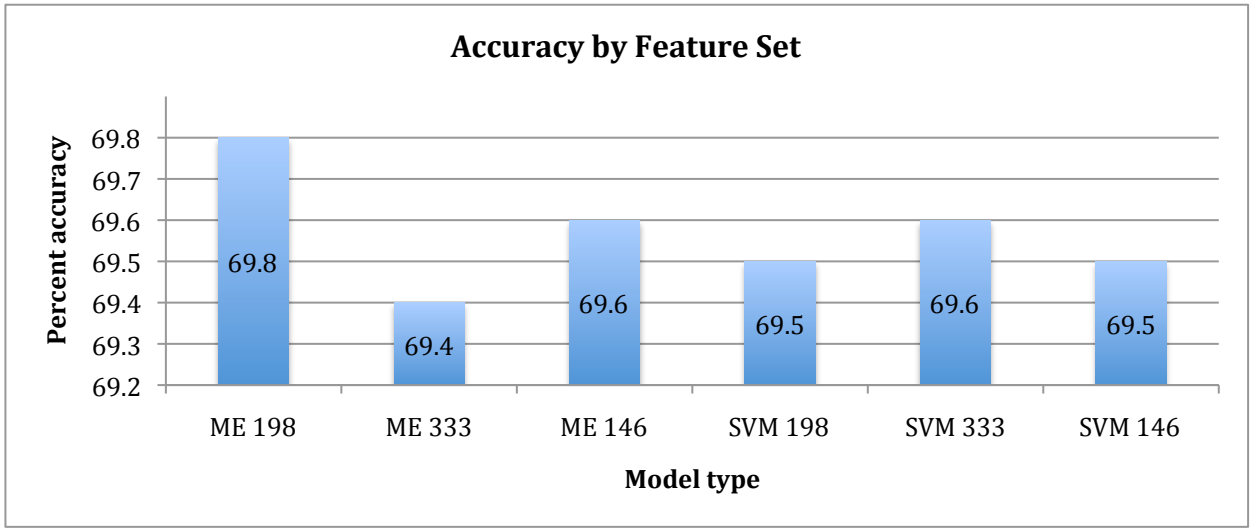
The Stanford Maximum Entropy Classifier was used on the categorical contextual features with the full unbalanced (CU, 6000 innocent, 32,880 guilty) and balanced (CB, TR=12,000) training sets, as well as with unigrams for all the words in the trial (CUU and CUB), with values averaged over five runs. CULB uses *contextual and real-valued linguistic features* on the balanced data set (TR = 14,000, TS= 2,000).



The LIBLINEAR SVM was tested on the real-valued linguistic features only (F=333, TR=14,000, TS=2,000, accuracies averaged over 5 runs). Based on higher accuracy, Model 6 was used for subsequent testing (L1-regularized L2-loss support vector classification).



Both classifiers were then tested with different numbers of features (TR=14,000, TS=2,000, accuracies averaged over 5 runs, columns labeled with classifier and number of features)



The top five predictive features for each class, given by MaxEnt (F = 333).

Innocence		**Guilt**	
Feature	*Weight*	*Feature*	*Weight*
Death – witness	8.8073	Religion - prisoner	4.9586
Death – all	4.9357	Average word length - witness	3.1742
Cognitive verb - all	4.3862	Third person - all	3.0975
Tentative - witness	3.9336	Passive - prisoner	2.9266
Causative - all	3.6278	First person – all	2.9156

Discussion

The linguistic features were written in several stages: the first 22 are all frequency-based (looking for categories of words/tone using regular expressions), the next 7 are similar frequency-based attempts at improving performance, and the final 9 are the higher-level features (the total is 37 features, each relativized for 9 speakers). The additional features slightly improved accuracy on the SVM but decreased it on MaxEnt. Further selection for an optimal feature set with maximum (by eliminating features for which all values were zero, for speakers for whom the feature was not very relevant, and finally the ones with minimal variance i.e. least likely to distinguish trials). This selection did not significantly alter performance in the SVM, which was to be expected, and returned an intermediate value for the accuracy of the MaxEnt classifier. The difference between all six values is rather small, suggesting that the features used are fairly similar—as long as the higher weighted ones remain in the feature set used, accuracy is not greatly affected. The accuracy of 85% on contextual and linguistic features combined, boosted from 69% on either set alone, suggests that this data set was modeled successfully by the classifier. The 20% gain over chance using linguistic features along is promising for applying machine learning to the detection of a speaker's being perceived as guilty or innocent, at least in a court of law.

Future Directions

Additional or more sophisticated features would likely improve performance by capturing the dimensions along which trials with innocent versus guilty verdicts differ most. The features considered in the proposal that were not implemented are parse-based: the frequencies of different parts of speech[1], the frequencies of different verb tenses, higher-level syntax information may help differentiate guilty from innocent verdicts[2]. Unfortunately, the data files containing the full parses of the utterances were too large for feature computation to complete in a reasonable time frame. Latent Dirichlet allocation for topic modeling and a finer-grain sentiment analysis (extremes of positive and negative emotion, sadness, perhaps a carefully selected set of words associated with guilt) may also reveal useful features. Since increasing the size of the training set led to decreased performance (MaxEnt: 69.9% on TR=12,000, TS=4,000 versus 69.5% on TR=14,000, TS=2,000, perhaps due to overfitting), simply using the entire unbalanced data set of 38,880 trials did not appear promising. Keeping the features constant, performance might improve if the training set is pruned (e.g. for longer trials—some have fewer than 3 utterances) rather than increased. Comparing precision and recall on SVM versus MaxEnt while changing feature sets may also elucidate which additional parameters would help.

[1] Preliminary frequency counts by part of speech show that superlative adjectives and the existential "there" bias towards innocence while interjections, superlative adverbs, and coordinating conjunctions bias towards guilt.

[2] For example, psychology research suggests that people are more likely to use simpler past tense constructions ("I went into the house and then I saw the money") when lying.

Acknowledgements

I would like to thank Professor Chris Potts for his support and continued guidance on this project. Earlier work on this data was conducted this summer with the help of the Trial Research Group: Professor Lera Boroditsky, Professor Dan Jurafsky, Eric Action, David Clausen, and Dan Wiesenthal.

References

Adams, S. H. and Jarvis, J. P. (2006). Indicators of veracity and deception: an analysis of written statements made to the police. *Speech, Language, and the Law*. *13(1)*, 1-22.

Fausey, C. M. and Matlock, T. (2010). Can grammar win elections?. *Political Psychology* (in press)

Larker, D. and Zakolyukina, A. (2010). Detecting deceptive discussions in conference calls. Stanford GSB Research Paper No. 2060.

Lin, C. J. et al. (2008). LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, *9.*

Pennebaker, J. W. et al. (2007). The development and psychometric properties of LIWC2007.