IST707 Data Analytics

Lie detection

Scenario:

We have collected a collection of customer reviews, some are true some are fake, and we are going to test different algorithms for fake review detection.

This data set has sentiment label for each review, and we will try to predict that as well.

For both tasks, Naïve Bayes and SVM got implemented. Here we are going to explore and report the results.

1. Different preprocessing methods – e.g., with or without stop-words, lemmatization, reducing the specific tokens you've used to maximize information gain.

Different preprocessing methods got used to maximize information gain. Here, we just mentioned the methods that increased the F1 scores. (All the methods that are not mentioned below, had a detrimental effect on the F1 value)

Transformation:

Words not transformed to lower case:

General Strategy	Parameter settings	Sentiment			Lie		
		F1	Precision	Recall	F1	Precision	Recall
SVM		0.793	0.793	0.793	0.478	0.586	0.552
Naïve Bayes		0.805	0.805	0.805	0.595	0.605	0.598

Words transformed to lower case:

General Strategy	Parameter settings	Sentiment			Lie		
		F1	Precision	Recall	F1	Precision	Recall
SVM		0.793	0.793	0.793	0.519	0.526	0.529
Naïve Bayes		0.805	0.805	0.805	0.574	0.577	0.575

Tokenization:

White spaces not removed:

General Strategy	Parameter settings	Sentiment			Lie		
		F1	Precision	Recall	F1	Precision	Recall
SVM		0.793	0.793	0.793	0.519	0.526	0.529
Naïve Bayes		0.805	0.805	0.805	0.574	0.577	0.575

White spaces removed:

General Strategy	Parameter settings	Sentiment			Lie		
		F1	Precision	Recall	F1	Precision	Recall
SVM		0.793	0.793	0.793	0.609	0.609	0.609
Naïve Bayes		0.805	0.805	0.805	0.552	0.553	0.552

Normalization:

Porter Stemmer not implemented:

General Strategy	Parameter settings	Sentiment			Lie		
		F1	Precision	Recall	F1	Precision	Recall
SVM		0.793	0.793	0.793	0.609	0.609	0.609
Naïve Bayes		0.805	0.805	0.805	0.552	0.553	0.552

Porter Stemmer implemented:

General Strategy	Parameter settings	Sentiment			Lie		
		F1	Precision	Recall	F1	Precision	Recall
SVM		0.793	0.793	0.793	0.620	0.643	0.632
Naïve Bayes		0.805	0.805	0.805	0.552	0.553	0.552

Filtering:

Stop words not filtered:

General Strategy	Parameter settings	Sentiment			Lie		
		F1	Precision	Recall	F1	Precision	Recall
SVM		0.793	0.793	0.793	0.620	0.643	0.632
Naïve Bayes		0.805	0.805	0.805	0.552	0.553	0.552

Stop words filtered:

General Strategy	Parameter settings	Sentiment			Lie		
		F1	Precision	Recall	F1	Precision	Recall
SVM		0.793	0.793	0.793	0.632	0.634	0.632
Naïve Bayes		0.805	0.805	0.805	0.678	0.678	0.678

There are some words like would, could, should, even, ever, said that do not provide useful information. All these words are stored in a text file and were removed during the data preprocessing step.

1. With either category (lie / sentiment) does it help to include the other category as a feature? For the lie feature

When we did not include the sentiment as a feature while detecting lies:

General Strategy	Parameter settings	Sentiment			Lie		
		F1	Precision	Recall	F1	Precision	Recall
SVM					0.723	0.725	0.724
Naïve Bayes					0.676	0.680	0.678

When we included the sentiment as a feature while detecting lies:

General Strategy	Parameter settings	Sentiment			Lie		
		F1	Precision	Recall	F1	Precision	Recall
SVM					0.586	0.586	0.586
Naïve Bayes					0.665	0.668	0.667

Therefore, including the sentiment as a feature while detecting lies did not help us.

When we did not include the lies as a feature while predicting sentiment:

General Strategy	Parameter settings	Sentiment			Lie		
		F1	Precision	Recall	F1	Precision	Recall
SVM		0.793	0.793	0.793			
Naïve Bayes		0.805	0.805	0.805			

When we included the lies as a feature while predicting sentiment:

General Strategy	Parameter settings	Sentiment			Lie		
		F1	Precision	Recall	F1	Precision	Recall
SVM		0.804	0.806	0.805			
Naïve Bayes		0.805	0.805	0.805			

Therefore, including the lies as a feature while predicting sentiment did not help us.

2. Use a topic model & sentiment analysis module to generate additional features and use these in combination with / instead of raw tokens.

Sentiment analysis:

We used the sentiment analysis package to generate sentiment analysis instead of the existing column. However, we don't use sentiment analysis to predict sentiment.

When sentiment analysis was not implemented

General Strategy	Parameter settings	Sentiment			Lie		
		F1	Precision	Recall	F1	Precision	Recall
SVM		0.559	0.564	0.563	0.644	0.645	0.644
Naïve Bayes		0.538	0.540	0.540	0.653	0.657	0.655

When Vader sentiment analysis got implemented (All the other sentiment analysis methods had a detrimental effect on the F1 score)

General Strategy	Parameter settings	Sentiment			Lie		
		F1	Precision	Recall	F1	Precision	Recall
SVM		0.793	0.793	0.793	0.632	0.634	0.632
Naïve Bayes		0.805	0.805	0.805	0.678	0.678	0.678

Topic Modeling:

When Topic modeling was not implemented:

General Strategy	Parameter settings	Sentiment			Lie		
		F1	Precision	Recall	F1	Precision	Recall
SVM		0.804	0.806	0.805	0.542	0.551	0.552
Naïve Bayes		0.805	0.805	0.805	0.445	0.450	0.448

When Latent Semantic Indexing topic modeling got implemented:

General Strategy	Parameter settings	Sentiment			Lie		
		F1	Precision	Recall	F1	Precision	Recall
SVM		0.793	0.793	0.793	0.632	0.634	0.632
Naïve Bayes		0.805	0.805	0.805	0.689	0.690	0.690

For each task and feature set (sentiment classification / lie detection), use the Rank module (Gini and Information Gain Ratio) to rank the features and list top 10 features from each method. Based on these top features, can you understand what patterns the classifiers have learned from the data?

When all the features were selected

General Strategy	Parameter settings	Sentiment			Lie		
		F1	Precision	Recall	F1	Precision	Recall
SVM		0.758	0.760	0.759	0.469	0.469	0.471
Naïve Bayes		0.828	0.828	0.828	0.574	0.577	0.575

When just the top 5 features were selected regarding information gain, information gain ratio, and Gini decrease.

General Strategy	Parameter settings	Sentiment			Lie		
		F1	Precision	Recall	F1	Precision	Recall
SVM		0.826	0.836	0.828	0.723	0.725	0.724
Naïve Bayes		0.805	0.805	0.805	0.676	0.680	0.678

We got the best F1 score when we selected the top five features which conclude: Topic 1, Topic 3, Topic 5, Topic 6, and neutral (from sentiment analysis).

The top 10 features conclude: Topic 1, Topic 2, Topic 3, Topic 5, Topic 6, Topic 9, Topic 10, neutral (from sentiment analysis), Positive (from sentiment analysis), and positivity number (pos-neg).

Topic 1: wa, food, restaur, thi, order, went, place, veri, minut, would
Topic 2: wa, thi, place, food, restaur, order, call, like, price, came
Topic 3: order, food, minut, ice, fri, wait, would, overal, best, veggi
Topic 4: place, restaur, thi, waiter, never, wa, us, want, best, serv
Topic 5: minut, us, order, thi, menu, restaur, wait, 15, best, ice
Topic 6: one, restaur, best, serv, food, view, dinner, ever, ice, boat
Topic 7: thi, food, best, waiter, veri, view, went, salad, fri, place
Topic 8: went, order, ice, servic, like, one, serv, good, friend, also
Topic 9: went, call, food., tri, smell, peopl, type, person., dinosaur, crowded,
Topic 10: good, waiter, us, order, veri, even, ever, tabl, wait, thi

The classifier has identified that topics 5 and 6 were the most useful topics (concerning information gain, information gain ratio, and Gini decrease).

In a write-up following the table, please explain the rationale for the strategies you have chosen, including the theoretical foundation for your choice. Also explain your parameter tuning approach for every attempted strategy. Where did you start? How much of a difference did parameter tuning make? Why?

Lie detection:

We started at an F1 score of 0.478 for lie detection using SVM. After preprocessing the data, tuning the parameters, and selecting the features, an F1 of 0.723 got achieved for lie detection using SVM. This means about a 25 percent improvement which is significant.

For Naïve Bayes, we started at an F1 score of 0.445 in lie detection. After preprocessing the data, tuning the parameters, and selecting the features, an F1 of 0.689 got achieved. This means about a 24 percent improvement which is significant.

Sentiment prediction:

For sentiment prediction, initially we got an F1 score of 0.559 using SVM. After preprocessing the data, tuning the parameters, and selecting the features, an F1 of 0.826 got achieved for sentiment prediction using SVM. This means about a 27 percent improvement which is significant.

For Naïve Bayes, we started at an F1 score of 0.538 in sentiment prediction. After preprocessing the data, tuning the parameters, and selecting the features, an F1 of 0.805 got achieved. This means about a 26 percent improvement which is significant.

Compare performance difference in sentiment classification and lie detection, and tell us which task is harder, and try to explain why.

It seems that sentiment prediction is by far easier than lie detection. Sentiment prediction was also less vulnerable to data preprocessing than lie detection.

Lies are typically difficult to detect, this is because only small and unreliable differences exist between genuine and fabricated statements.

Try to get the best results you can.

Finally, the following table indicates the best result we have achieved.

General Strategy	Parameter settings	Sentiment			Lie		
		F1	Precision	Recall	F1	Precision	Recall
SVM		0.826	0.836	0.828	0.723	0.725	0.724
Naïve Bayes		0.805	0.805	0.805	0.676	0.680	0.678