CS 1671 Human Language Technologies (2/20/2025 11:59 pm)

Homework 2: Text Classification

Bridget Brinkman

Diplomacy by Maxime Bonzi, Flickr, License CC BY 2.0

Part 1: Feature-based logistic regression models

- **Q:** Table of 5-fold cross-validation performance scores for all models trained on each set of features, with the average accuracy, precision, recall, f1-score for positive (lying) class across the 5 folds.

model	accuracy	precision	recall	f1-score	
unigram	0.609034	0.613021	0.591486	0.601528	
tf-idf 1-1 gram	0.639718	0.630294	0.676682	0.65187	
tf-idf 1-2 gram	0.635359	0.614254	0.73684	0.668966	
tf-idf 1-3 gram	0.622196	0.593505	0.788192	0.676164	
tf-idf 1-4 gram	0.602772	0.572096	0.822028	0.674344	
tf-idf 1-1 gram w/ sentiment	0.578966	0.573458	0.620314	0.595626	
tf-idf 1-2 gram w/ sentiment	0.616548	0.600224	0.700519	0.64615	
tf-idf 1-3 gram w/ sentiment	0.606542	0.59284	0.685464	0.63549	
tf-idf 1-4 gram w/ sentiment	0.607157	0.596869	0.664151	0.628114	
the ngram range represents the ngram features included in the model:					
1-1 gram includes only unigrams					
1-2 gram includes unigrams and bigrams					
1-3 gram includes unigrams, bigrams, trigrams					
1-4 gram includes unigrams, bigrams, trigrams, and four-word sequences					

- **Q:** For each feature or change in the input text processing describe your motivation for including the feature. Discussion of results: Did it improve performance or not? (Either result is fine. It is not necessary to beat logistic regression with unigram features)
    - **A:** Tf-idf unigram:
        - The motivation for adding tf-idf to the base unigram model was to give more common words less weight. This was done with the assumption common words do not have as much influence on the information we need to differentiate truth from lies, having our model focus on informative words.
        - This improved the model quite a bit by ~3% accuracy, ~2% precision, ~8% recall, ~5% f1 score.
    - **A:** Tf-idf ngram:
        - The motivation for adding n-gram features to the tf-idf model is to see whether certain combinations of words would lead to better predictions of detecting lies. This is because certain chains of words might be correlated with lying (i.e., “trust me,” “I never said,” “I promise you that”). These combinations might help the model make better predictions.
        - This model did not produce any improvements over the unigram model ☹

- **A:** Tf-idf ngram w/ sentiment:
    - The motivation for adding sentiment to the tf-idf n-gram model was to see whether different sentiments correlate more with lies. Sentiment was added using nltk's vader_lexicon and nltk's SentimentIntensityAnalyzer.
    - This model worsened performance.

- For a feature-based model of your choice:
    - **Q:** Extract and discuss the most informative features that are most strongly positively and negatively associated with deception. Report the 5 features with the highest weights and 5 features with the lowest (negative) weights. Discuss how these may or may not make sense for this task. You may adapt code provided by the instructor, use another source online, or write your own. Give specific informative features, such as particular words (e.g. "actually") for bag-of-words features, instead of sets of features like "tf-idf unigram features".

A: Top Features Strongly Associated with Deception (Lying):

i	2.19847243
,	1.94667783
to	1.68908197
.	1.15046664
and	1.11039527
alright	1.03346022
for	0.90819765
i can	0.89275154
done	0.85340319
going	0.83352557
think	0.81353538
okay	0.74385089
it	0.73588274
, and	0.72127591
i think	0.72010593
attack	0.71251994
going to	0.67538605
. i	0.67497796
ok	0.66493821
, i	0.65405011

A: Top Features Strongly Associated with Truthfulness (Not Lying):

?	-1.4457022
what	-0.8750062
!	-0.7843394
f	-0.6812359
hmmm	-0.5795142
did	-0.5178023
dude	-0.4909214
thanks	-0.4865222
yeah	-0.4832036
can you	-0.4697258
! !	-0.4494848
you can	-0.4416467
makes	-0.4319405
no worries	-0.4260832
worries	-0.4260832
thanks !	-0.4105281
what are	-0.3954448
rgr	-0.3779688
doing	-0.3735135
that 's	-0.3729973

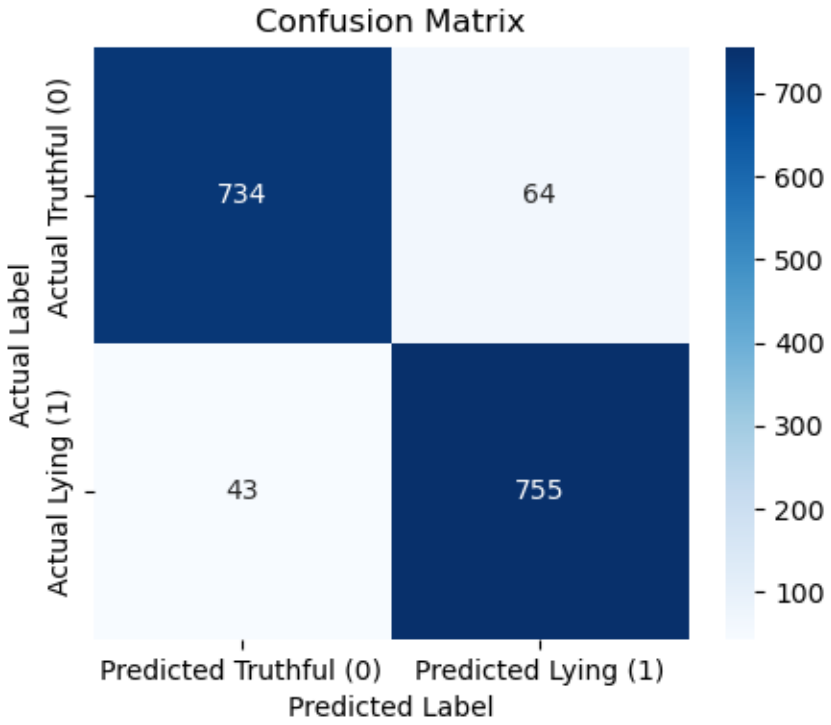
- **A:** I chose the tf-idf 1-3 gram model to extract the most informative features because it achieved the highest f1 score.

- **A:** Lying: Some patterns that I recognized in the features for lying were self-centeredness. The word “i” shows up in 5/20 of the top features. This makes sense because narcissism relates to deception, and I can reason that a player using “i” frequently may be saying self-centered or self-serving phrases. Other words that stick out to me are “and,” “alright,” “done,” “going,” “think,” “okay,” “going to.” Again, these seem self-centered to me, I can think of ways a player might craft sentences about what they’re going to do, what they think someone else is going to do (blame), what they’re thinking (self-centered, egotistical, power dynamics/control), or chaining together many things with “and” like the player is making many points, possibly over-explaining the state of the game to try to deceive other players with their reason, knowledge, and perception. Punctuation associated with lying included periods and commas, which to me have a more serious or straightforward tone.

- **A:** Truth: Some patterns that I recognized in the features for truth were a lot of casual and polite words such as “hmm,” “dude,” “thanks,” “yeah,” “can you,” “no worries,” “thanks!” “rgr.” Punctuation associated with truthfulness included question mark and exclamation points, and

exclamation points show up in 3/20 top results. These conclusions lead me to think there is more emotion (positive emotion, questions being asked) involved with truthfulness, more casualness, and politeness.

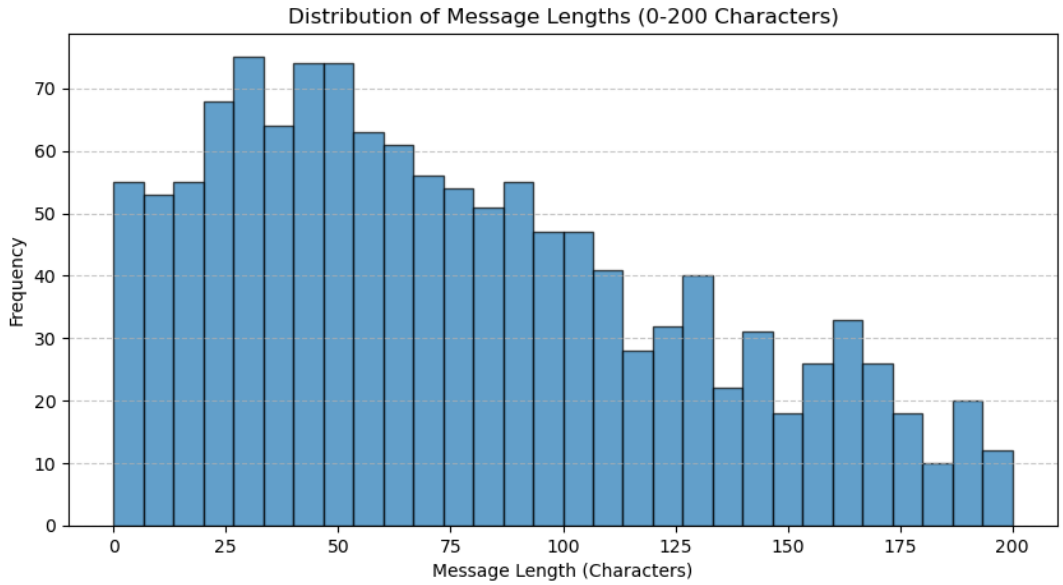
- **Q:** Do an error analysis. From one of the cross-validation runs, provide a confusion matrix on the test fold, not on training folds. You can also choose to create a separate development set from the data and provide a confusion matrix from that. Sample examples from both false negatives and false positives and present a few of them in the report. Do you see any patterns in these errors? How might these errors be addressed with different features or if the system could understand something else? (You don't have to implement these, just speculate.)
- **A:** Confusion Matrix:



Sample False Positives (Predicted as Lying, Actually Truthful)

France has asked my opinion on it, and I haven't given it yet. To my estimation things look a lot better for me if you don't end up there: I don't want to see England in Mao, and I don't want to see you snagging pieces of the north.
Btw just sent this to England "One of us has to move into it by the end of the year. Germany is building to move against me, Turkey is still a pain in the butt, and Italy is getting testy. I need your help to crack this dude, and I can't be running damage control for the board on my own"
What I was/am trying to avoid is a scenario in which you (a) work with Germany and (b) send way more strength to the Med than you can actually use. I was concerned, since I had continually offended you, that you were going to go full-on spite against me! Since it would appear that you're not going to do that, ok. As iron sharpens iron, I will fight you in the Med over the remaining dots! Perhaps we'll both learn something! And while I was genuinely willing to hold what I currently have, I will tell you the truth that I am going to continue the attack!
Of course, I don't think we should be fighting(at least this early in the game), and that it would be mutually unproductive. I agree with a DMZ in Piedmont and Gulf of Tyrolia, and we might as well extend it to Western Mediterranean and North Africa. But I hope you understand how unreasonable for you to expect me not to move the Spain, as it is a typical France acquisition. Also I can't build a fleet in Piedmont... Anyways, Good luck, and have fun!
Move into Brest with the fleet in Picardie, and set yourself to three-way draw and we'll have it.

- **A:** The main pattern I see in these messages are that they are long. I've never played diplomacy, so I don't know what is standard. I plotted the distribution of message lengths (next page) in the dataset to gather more information (limiting the range from 0 to 200, although there were message lengths above 200, they were infrequent). These messages seem to be of standard or most likely length, so I don't think my model had issues with context. I notice a lot of neutral diplomacy-relevant words, like "DMZ," "Gulf of Tyrolia," "Spain," "Western Mediterranean," "Germany," "England," "Mao." I am wondering if these words are equally likely in truth and lies, thus not giving insight. It would be interesting to see the prediction of some of these messages, to see how "wrong" my model got them, in order to identify specific features so that I can tweak the model.



Sample False Negatives (Predicted as Truthful, Actually Lying)

Ok. Try again next turn
Yes, that's what I'm hoping for.
Deal!
Btw have you talked to Italy?
I haven't really heard much of anything from Germany.

- **A:** The main pattern I see is that these messages are extremely short. With limited words, my model probably didn't have great context. I think broader context (like message history or other meta-game features) is necessary. Knowing a player's current positioning or ranking in the game, knowing what they've said before, the language and words they've used before (to detect changes in language) would help my model accurately classify these sentences. I think it's a balancing act, where you don't want too much or too little context. If you're analyzing too many features, you may get confused

and weights or determining factors might get blurred. Too few features, and you don't have sufficient knowledge to be sure with your prediction. At the end of the day it's all probability, and we hope to have high probability!

Part 2: Submit your classifier in the class challenge

- Username: BridgetBrinkman
- hw2_brb351_kaggle.py
    - unigram tf-idf model trained on the entire diplomacy_cv.csv dataset