DATASETS: CSDMC2010, ENRON, SPAMASSASIN, UCI,
KAGGLE PHISING DATA

PROJECT B: CLEAN DATA, CREATE BASELINE NAIVE BAYES

EXP A: VARY TRAINING DATA SIZE BY 2,000 INTERVALS
EXP B: VARY "IMPORTANT WORDS"
EXP C: USE NLP TOOLKIT (POS TAGGING, NER)
EXP D: COMPARE NAIVE BAYES TO LSTM NEURAL NET

DOES IT GENERALIZE?

TEST SET? → TOTALLY NEW(?) OR NEW PORTION OF 5 DATASETS
→ WOULD GUESS IT PERFORMS WORSE BUT MIGHT SPEAK TO IMPORTANCE OF PERSONALIZATION?

Romain Garcia - Next to You
Blake.08 - SUPERMONSTER
TINLICKER - REBIRTH
Eli + Fur - Come Back (?) Around
Michael Caselle - Bells of Konecista
Kabi - Om

4/9

Next Steps:

~~① Fix & Exp A to shorten/efficient code~~

~~② Exp B: Figure out how to track MDIB on top x words~~

~~③ Exp C: Adjust input tokens w/ NER / pos tagging~~

FINISH LSTM + SVM

Experiments + Tests — 4/9

✓ #1 → VARY TRAIN DATA SIZE
✓ #2 → VARY "IMPORTANT WORDS"
✓ #3 → USE POS TAGGING + NER
#4 → PRODUCE CSTM / FIND PAPER + REPRODUCE
#5 → RECREATE PANEL
#6 → RECREATE GRAHAM

FUTURE RESEARCH TO ~~CHANGE~~ ADD/SUB FUTURE TESTS
— → SPAM CLASS STATE OF ART SURVEY
— → LSTM FOR SPAM CLASS

Top - X# Important Words

~~Bucket. dog - Pride~~

FIT MODEL ON ALL WORDS

COMBINE + RANK { GRAB LOG PROBS → CONVERT TO 1 SCORE (THE GREATER (?))
GRAB WORDS ASSOCIATED
~~REFIT of COMBINED~~ FIT NEW MODEL W/ TOP ~~X~~ WORDS
PREDICT

TOKENIZATION PROCESSES

GRAHAM 02

AZ19, DASHES, APOSTROPHES, $

IGNORES "ALL DIGIT" TOKENS

LOWERCASE ONLY

TOP 15 INTERESTING TOKENS USED

GRAHAM 03

KEEPS CASE, ADDS !, ADDS ./, in btwn #'s

ADDS SUBJ TAG