GOALS

1. COMPARE NB, LSTM, SVM ON EMAIL DATASET
2. EXPLORE DIFFERING TOKENIZATION OPTIONS TO IMPROVE RESULTS
3. DETERMINE SAMPLE SIZE FOR GOOD RESULTS

EXPERIMENTS

BASELINE NB / LSTM / SVM

EXP A: Changing SAMP SIZE OF TRAIN SET (PRETTY SIMILAR >5000)

EXP B: Change "IMPORTANT WORDS" (15-50 SIMILAR)

EXP C: USE POS/NER TAGS (70% ACC)

WHAT'S NOVEL

DATASET

TOKEN DEF

CURRENT TOKEN DEFINITION

CHANGE 1-9 → <DIGIT> TAG

REMOVE NON @/#/$/%/!/&/& PUNCTUATION

REMOVE CASE + MANDATORY D/C DATASET

OUTLINE

① INTRO

NEED: SPAM STILL A PROBLEM → LEVCHENKO

SPAM CLASS AN EVER EVOLVING ~~AREA~~ FIELD → HAS TO KEEP UP "ADVERSARIAL"

NOVELTY OF RESULTS — WHY INTERESTING TO OTHERS

WHY INTERESTING TO ME?

② RELATED WORK

PANTEL → SUMMARIZE + WHAT WAS NEW → NEW = NB / SPAMMERS

PAUL GRAHAM → NEW = BETTER TOKENIZATION FOR NB

~~ABROUTSOPOULOS~~ DRUCKER → DEMONSTRATED SVM > NB

BASYAR → ~~BASYAR~~ 1ST LSTM FOR SPAM CLASS

③ GOALS

PRJ A → "IMPROVE CLASS TECHNIQUES"
- VARY SIZE OF TRAIN CORPUS
- DEPLOY NLP TECHNIQUES (POS/NER)

PRJ B → MINIMUM TRAIN SIZE
- → IMPORTANT WORDS
- → NLP TECHNIQUES
- → LSTM NN

④ MODELS → SVM + NB
DATASET → DESCRIBE SOURCE, DESCRIBE PREPROCESSING
EXPERIMENTS → STATE PROBLEM/HYP, DESCRIBE EXP,
DESCRIBE RESULTS, INTERPRET RESULTS

⑤ FUTURE WORK → LSTM, LARGER DATASET / PERSONALIZED

⑥ BIBLIOGRAPHY

PROJECT C REQUIREMENTS

① 2 MEETINGS ✓

② WRITTEN NOTES ✓

③ CODE / DATA / ANALYSIS / README → IN GITHUB

④ FINAL REPORT (4-8 PGS)

↳ ACM Conference Template

↳ TITLE AUTHOR + "MENTOR:" + URL

SEC #1 → INTRO: NEED, WHAT HAS BEEN DONE, NOVELTY + NEW RESULTS, WHY WORK = INTERESTING

SEC #2 → RELATED WK SUMMARY → SUMMARIZE + WHY ITS DIFFERENT

↳ INCLUDE CITES

SEC #3 → GOALS FROM PRJ A/B, IF THEY ✓, WHY DIDN'T ✓

SEC #4 → DISCUSSION OF WORK → GOAL / METHODS / DATA / RESULTS

SEC #5 → FUTURE WORK

SEC #6 → BIBLIOGRAPHY

Related Work

PANTEL ET AL. (NB)

[redacted] (NB)

GRAHAM 2002 (NB)

GRAHAM 2003 (NB)

SINGH ET AL (SVMs)

RAHMAN ET AL (LSTMs)

~~[redacted]~~ MAMA ET AL (NB/SVM)

EXP A/B/C

STATE

HYPOTHESIS: There is a main truth

EXPERIMENT

RESULTS

INTERPRET

- 38 PTS TOTAL

PROJECT B FEEDBACK

- 10 PTS
  - IMPROVE NOTES + MEET w/ PROF

- 9 PTS
  - INTRO - INCLUDE WHY INTERESTING TO FIELD/STUDENT + WHY NOVEL
- 3 PTS
  - ALIGN CITATIONS w/ BIBLIOGRAPHY

- 8 PTS
  - RELATED WORK - ORDER BY PAPER NOT FEATURE
    - ADD MORE THOROUGH RW SECTION
    - MAYBE LSTM / OTHER NN INFO FOR STATE OF ART
    - FIND ADDITIONAL PAPERS FROM GRAHAM / PANTEL TL
  - DATASET ✓

- 5 PTS
  - WRITEUP SUGGESTIONS
    1. # / HIERARCHY TO SECTIONS
    2. OVERVIEW TO SUMMARIZE METHODS → BREAKDOWN FROM THERE
    3. FIGURE OUT HOW TO EXPAND BEYOND TOKENIZATION
    4. BREAK UP METHODOLOGY / DATASET / RESULTS
    5. TITLE EXP. APPROPRIATELY
    6. USE ACM / LaTeX template