1. Logistic regression with words and char n-grams

<https://www.kaggle.com/tunguz/logistic-regression-with-words-and-char-n-grams>

It combines char features and word features to fit logistic regression. Result: 0.9789

2. NBSVM

Replace SVM with logistic regression. Result: 0.9772

The original algorithm considers 1~2 ngrams and punctuation. Without punctuation, the results become worse.

Text substitute:

1. low frequency

2. consist word in list, or high overlapping

Spell check: fq<30 + in / ratio>=0.5:

Train[110098]:

Original:

Hey f\*\*king swine \n\nDoing bad? Suffering a lot?

Corrected:

hey going die doing bad? suffering a lot?

Train[20945]:

Original:

U ANNOYIN BIITCH.. WHY U KEEP ON REVERTIN WAT I JUS DID? SMH LOSER ASS CRACKA WHO WANNA BE PUERTO RICAN

Corrected:

u going bitch why u keep on revertin wat i jus did? shit loser ass cracka who wanna be puerto rican

Idea: too many words repeating in a comment may be non-sense.