


Detecting Bullying Traces in Tweets

Application of Natural Language Processing for social studies

By Niti Mishra K.C.




Why are we interested?

- Many scholars interested in bullying studies are blocked due to data scarcity
 - In general, social media such as Twitter afford a context for cyberbullying to take place.
 - Thus it provides a unique vantage point for these scholars, as it's data can reveal how individuals are representing bullying online in real time.
 - In recent years, the automatic detection of “aggressive behavior” in social media is gaining a lot of attention
- 



Our Objective

- Contribute to present evidences that social media, with appropriate natural language processing techniques, can be a valuable and abundant data source for the study of bullying in both digital and real world.
- 



Data



- Twitter Streaming API
- Collecting tweets since August, 2019 – July 2020
 - 7712.200 mean tweets per day
 - 15.130 lexical diversity
- Keywords:
 - Primary: bullied, bully, bullying, cyberbullying, cyberbullied, cyberbully
 - Secondary: 95 words → Forced, exclusion, hitting, shove, harass, etc.
 - Removed: Trump, @realdonaldtrump, white house @whitehouse, white house @potus, potus @flotus flotus, president
- Full tweets (extended characters)
- 5000 tweets labelled
 - 2 human coders
 - Interrater agreement: 81%
 - 'no': 2838, 'yes': 2077

Example Tweet

- `{'id': 1161690137647955968, 'full_tweet': '@shake1n1bake @benshapiro it's amazing how much of conservative ideology can ultimately be boiled down to a bully grabbing your arm and hitting you in the face with your own fist and yelling "stop hitting yourself! stop hitting yourself!"', 'bullying_trace': 'no'}`
- `{'id': 1162088209104494592, 'full_tweet': "#pmddhour\nnon a few occasions i have received online bullying on social media when talking about pmdd.\n\nit was horrible, awful what some people were saying to me. it came from people i've never meet or some friend of a friend!", 'bullying_trace': 'yes'}`



Pre-processing



During Collection

- Removed:
 - Retweets
 - Non-English tweets
 - Quoted tweets
 - URLs
 - Tweets with ≥ 6 hashtags

During Modelling

- nltk's TweetTokenizer
- Anonymize user
- Hashtags compound words treated as single token
- No stop-words removal
- Emoticons treated as token
- [Github Link](#)



Classifiers



- LogisticRegression (TruncatedSVD)
- **LogisticRegression**
- SGDClassifier (TruncatedSVD)
- SGDClassifier
- MultinomialNB
- GaussianNB (TruncatedSVD)



RESULTS

nGrams

Unigram
(1,1) g



Unigram +
Bigram
(1,2)g

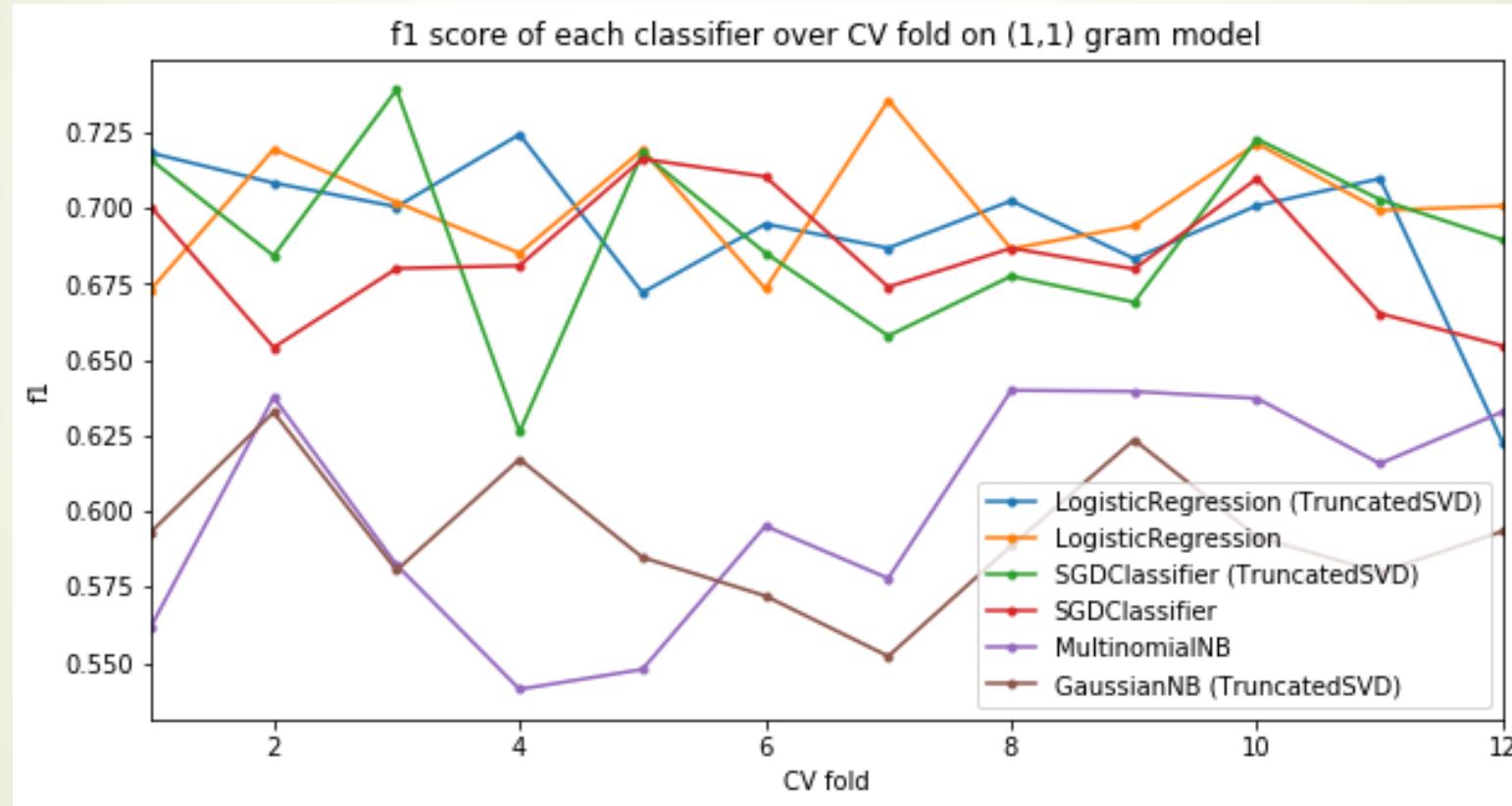


Mean scores over 12-fold CV

	name	accuracy	precision	recall	f1	time
0	LogisticRegression (TruncatedSVD)	0.699277	0.697521	0.699277	0.693680	5.375905
1	LogisticRegression	0.705392	0.704359	0.705392	0.700845	0.981322
2	SGDClassifier (TruncatedSVD)	0.693996	0.693436	0.693996	0.690724	5.208292
3	SGDClassifier	0.684431	0.686420	0.684431	0.684394	0.981969
4	MultinomialNB	0.647431	0.670695	0.647431	0.600744	0.954380
5	GaussianNB (TruncatedSVD)	0.591661	0.616627	0.591661	0.592434	4.990253

	name	accuracy	precision	recall	f1	time
0	LogisticRegression (TruncatedSVD)	0.696644	0.695134	0.696644	0.692732	24.827597
1	LogisticRegression	0.708251	0.708811	0.708251	0.704358	1.502113
2	SGDClassifier (TruncatedSVD)	0.685256	0.698703	0.685256	0.682554	25.422912
3	SGDClassifier	0.695835	0.701849	0.695835	0.697216	1.523191
4	MultinomialNB	0.629895	0.694227	0.629895	0.549975	1.461503
5	GaussianNB (TruncatedSVD)	0.600608	0.603561	0.600608	0.525036	26.287931

Performance in CV mode



[Github Link to other results](#)



Discussions/Challenges

- File encoding problem with emojis
- Switching between csv and xlsx → build app for labelling tweet (spacy)
- Getting the code to preprocessor to run such that it includes rest of the items of the tweet dictionary instead of just the tokenized tweets
- Github!!!!



Questions?