Social Media Writing Style Fingerprint

Himank Yadav, Juliang Li

1.
Overview

Overview

- Humans have the cognitive ability to differentiate between writing styles of various authors.
- Previous work has been done on authorship attrition for books and long texts.
- We focus on shorter writing samples gathered through social media

2.
Motivation

Sahil Dhanju
September 22 ·

I'm happy to announce that I have accpeted a full-time job at McDonalds to flip burgers and will start immediately after finishing my Computer Science Degree at Texas A&M. I know my family and friends are proud, and honestly now that I think about it, I could not have found a better job. Actually, based on my grade in Operating Systems and Calculus, I'm really suprised that I even made it this far. Drop by to say hi if you are in the area!

Love Comment

You, Juliang Li, Denise Irvin and 489 others

Tyler Durden I think you are right about how you couldn't have done any better. Well done Sahil.
Like · Reply · September 22 at 6:32pm

Jay Khatri lol did you get hacked again man?
Like · Reply · September 22 at 9:18pm

Write a comment...

Sahil Dhanju
September 24 ·

My account got hacked by someone that obviously knows me but I have no idea who they are. I luckily am not going to work at McDonald's.

Sad Comment

You, Juliang Li, Victoria Wei and 35 others

Alexandra Neikirk I was so proud
Like · Reply · September 24 at 9:52pm

Danny Byrne Rip the profile pic
Like · Reply · September 25 at 8:17am

Write a comment...

Applications

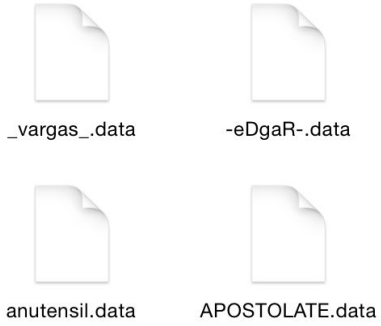
- Detect social media hacking activity.
- Establish the credibility of a source.
- Identifying anonymous negative phenomenon (bullying)

3.
Data

Collection & Cleaning

- Comments v.s. Posts bias
- VRCkid (aka Sahil) and other top redditors that he interacts with
- PRAW - Python Reddit API Wrapper
- Removed empty comments and edge cases

Sample Data



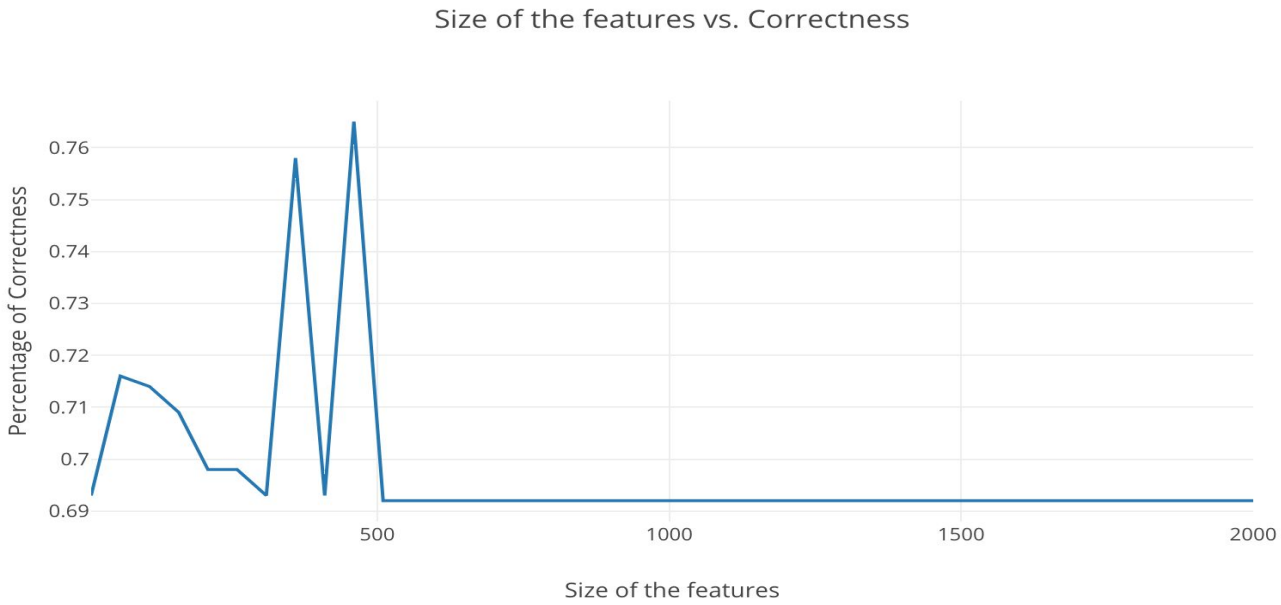

```
[
    "This is an authorized meme. ",
    "Hurricane pike is instant at least so the sto
    "Commenting on this late but I'm a senior CS m
    "This post was removed due to breaking our giv
]
```

4.
Method

Word Frequency

- Counts the usage of vocabulary
- Decide features (bag of word)
- Use logistic regression to train and test
- 76% accuracy

Performance



Lexical KMeans

- Using unsupervised learning to capture distinctiveness in sentence structure.
- Tokenize words and sentences, focus on lexical features and punctuation style of text
- Word density, vocabulary diversity and punctuation placement
- Use clustering to find natural groupings, predict using a KMeans cluster ~ 69%

Character N-gram

- Views the text as a sequence of characters
- Use N-gram to extract data
- Select top N-gram
- 77% Accuracy

Performance

Gram-size/ features size	100	200	300	400	500
1	0.763	0.763	0.764	0.766	0.763
2	0.728	0.73	0.741	0.772	0.772
3	0.722	0.731	0.732	0.729	0.728
4	0.699	0.689	0.691	0.730	0.741
5	0.694	0.696	0.728	0.718	0.729
6	0.684	0.694	0.687	0.707	0.697
7	0.689	0.681	0.688	0.681	0.692
8	0.691	0.684	0.674	0.695	0.698
9	0.703	0.694	0.688	0.686	0.684
10	0.702	0.699	0.696	0.691	0.69

Author Verification for Short Messages

- Supervised learning combined with n-gram stylometric analysis
- Split data into training and verification, compute a specific threshold for the given user, derive user profile by extracting n-grams.
- Divide the verification user data into p blocks of characters of the same size
- Calculate the percentage of unique n-gram shared by blocks and the training set
- Block p is said to be a genuine sample of user if the percentage of unique n-grams shared by a block is greater than threshold value specified for the user

Parts of Speech

- Analyze Syntactic information
- Author may subconsciously make similar phrase structure.
- Pick the most frequently used structures as features

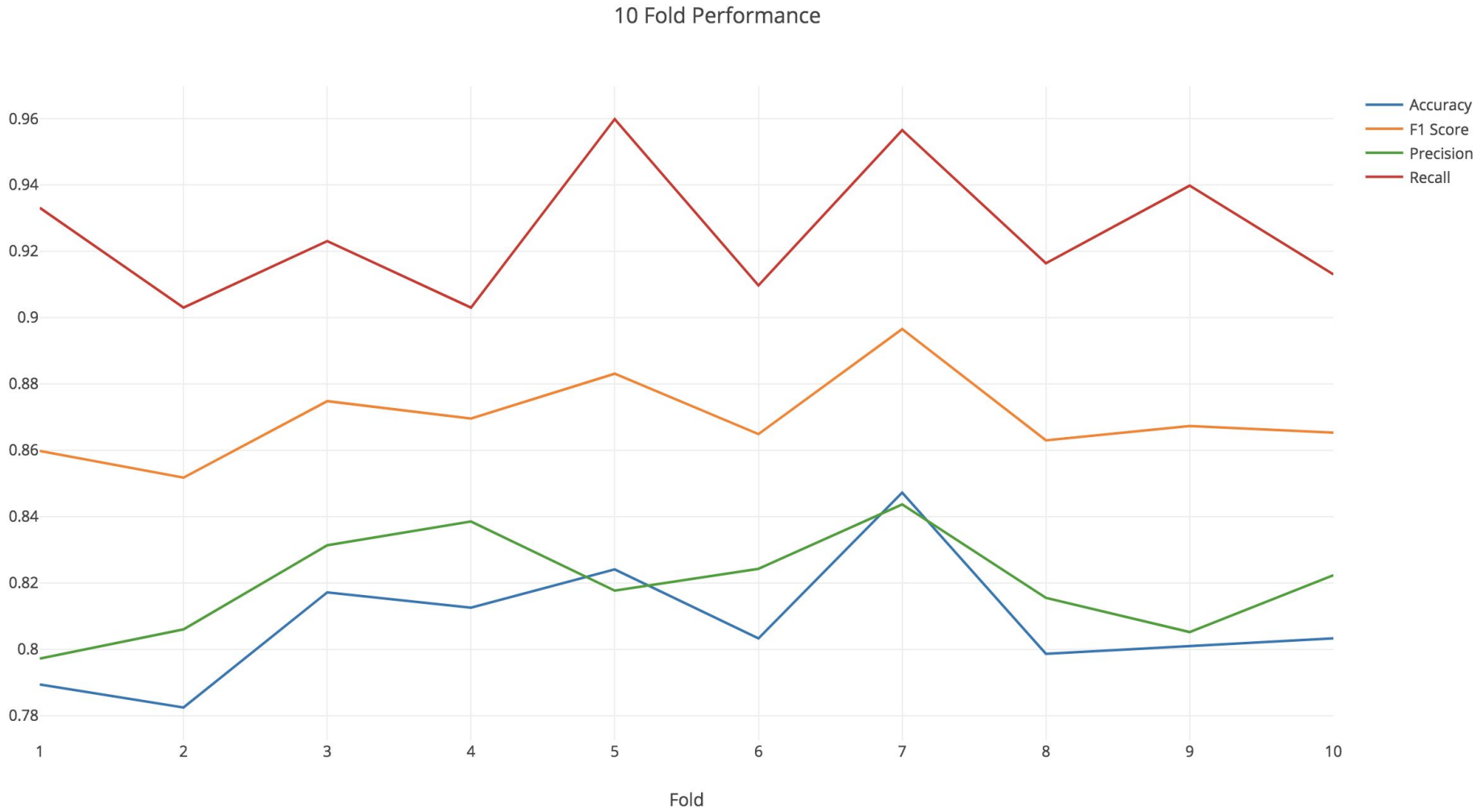
Performance

Gram size	Top 1 frequency	Top 2 frequency	Top 3 frequency
1	NN	IN	DT
2	DT, NN	IN, DT	PRP, VBP
3	IN, DT, NN	DT, JJ, NN	DT, NN, IN
4	NN, IN, DT, NN	IN, DT, NN, "."	IN, DT, JJ, NN

Feature size/gram size	1	2	3	4	5
30	0.729	0.681	0.689	0.691	0.689
80	0.732	0.698	0.675	0.691	0.686
130	0.714	0.676	0.678	0.694	0.682
180	0.725	0.694	0.679	0.675	0.685
230	0.725	0.722	0.694	0.679	0.680
280	0.725	0.719	0.692	0.685	0.674
330	0.738	0.704	0.698	0.676	0.685
380	0.725	0.716	0.696	0.690	0.684
430	0.724	0.720	0.701	0.678	0.680
480	0.727	0.716	0.699	0.690	0.680
530	0.719	0.718	0.685	0.679	0.682
580	0.720	0.717	0.700	0.688	0.68

Master Classifier

- Multi-layered classifier simulating a neural net
- 5 input nodes where the final layer uses majority vote from the middle layer vector to classify
- Focus on strengths and weakness for each of the 5 classifiers
- Accuracy ~80%


10 Fold Performance
Accuracy
F1 Score
Precision
Recall
0.96
0.94
0.92
0.9
0.88
0.86
0.84
0.82
0.8
0.78
1
2
3
4
5
6
7
8
9
10
Fold

5.
Future Work

Extension

- Expand features to include semantic analysis and other external non-language data points
- Explore different social media datasets
- Build tooling to measure real world success

Questions?