Genetic and Evolutionary Feature Selection (November 2018)

Sarp Aykent, Jordan Cox, and Blake Schilleci

*Abstract*—**Genetic algorithms are extremely beneficial when masking features to optimize our machine learning algorithms. Improving the baseline accuracy is the main goal when using genetic algorithms on top of the machine learning algorithms. Creating a population of Feature Masks is what we undertook and used to improve accuracy when predicting authors. This paper explains the processes, algorithms, and optimizations we used in order to improve our baseline accuracy.**

# INTRODUCTION

Authorship attribution is always being looked at for improvements. Having flawless accuracy is the overall goal for this assignment. This is a feasible goal due to the face that our data set is not as large as others. Finding authors to write about Mississippi State football is difficult, but this simply means that our accuracies from the baseline results should improve dramatically after implementing genetic algorithms. Improving authorship attribution from machine learning algorithms comes from masking features whilst using the genetic algorithms. We chose three genetic algorithms to use for our optimizations: Steady-State Genetic Algorithm, Elitist Genetic Algorithm, and an Estimation of Distribution Genetic Algorithm.

# Methodology

When running the machine learning algorithms in our past assignment, we noticed that the computation with 95 features can take some time, and since there are so many features, they may not all be relevant. To alleviate the mishaps that may come along with having so many features, we used several genetic and evolutionary algorithms to find the best features. These algorithms are Steady-State Genetic Algorithm, Elitist Genetic Algorithm, and Estimation of Distribution Algorithm. They all essentially do the same thing, but go about it in slightly different ways.

The Steady State Genetic Algorithm starts with an initial population of randomly generated individuals. Then, uses tournament selection to select two parents from the k-best parents. We then use crossover to create a child and mutate it. With this child, we replace it with the worst fit child in the population even if the new child’s fitness is worse than the worst fit in the population. For Steady State Genetic Algorithm we used stylometry-based feature vector[[1]](#footnote-1).

The Elitist Genetic Algorithm starts with an initial population of randomly generated individuals. It saves the best fit individual to the next population. Then, chooses two random individuals in the population to create the children to fill up the rest of the population for the next generation.

The Estimation of Distribution Algorithm starts with a randomly generated initial population. We then select the *n*-best individuals as parents to generate our children. Since we chose an elitist approach for doing this, we saved the best individual in each generation and saved them to the next generation. The rest of the next generation was filled with our children generated from the *n*-best parents and mutated.

# Experiment

We wrote our genetic algorithms in Python, and we gave it two  datasets to traverse, our CASIS-25 dataset and our SEC Sports Writer Dataset (Mississippi State)[1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36]. We began with three baseline algorithms and used a different genetic algorithm for each in order to optimize accuracy. These genetic algorithms are simply masking the features in order for the accuracy to increase for each iteration. Through our experiment there will be 5000 evaluations of the individuals with 10 overall iterations from each method working on the CASIS-25 and SEC Sports Writer data sets. This is to test the average accuracy improvements from the baseline accuracies.

For this experiment we were given specific parameters to begin with. The population size would consist of 25 individuals. The crossover that we would be using was Uniform crossover. Any other crossover would work, but this allowed for consistency. The mutation rate that we were required to use was 0.05. This simply states the probability being used throughout our experiment. As stated above, the number of evaluations for each individual is 5000. And the evaluation function will be testing accuracy rather than error rate. As a whole, the Genetic and Evolutionary Feature Selection will be ran 10 times for each algorithm and on each data set and record average and best accuracy for each run.

Optimizing the genetic algorithms is actually really simple when you change the given parameters. Affecting the mutation rate, population, and even the number of evaluations will help the accuracy plateau and become better and more consistent. The optimization that we used for Steady State Genetic Algorithm is basically precomputing the effect of masking each bit individually. Based on their effects for accuracy, we have a modified portion(11%) of the initial population. Bits that increase accuracy are set 0 and vice versa. Every single individual is forced to be unique. The optimization we used for Elitist GA was increasing the mutation rate to 0.1 and making the parents the five best in each population. It still remained to be Elitist in the end because on each generation we only kept the best individual and replaced every other individual with a new child. The optimization we used for Estimation of Distribution Algorithm was changing the number of parents to 17 and only using the best 17 individuals as parents for each generation. We also kept all the parents into the next generation and replaced the individuals who failed to become parents due to their lower fitness with new children.

# Results

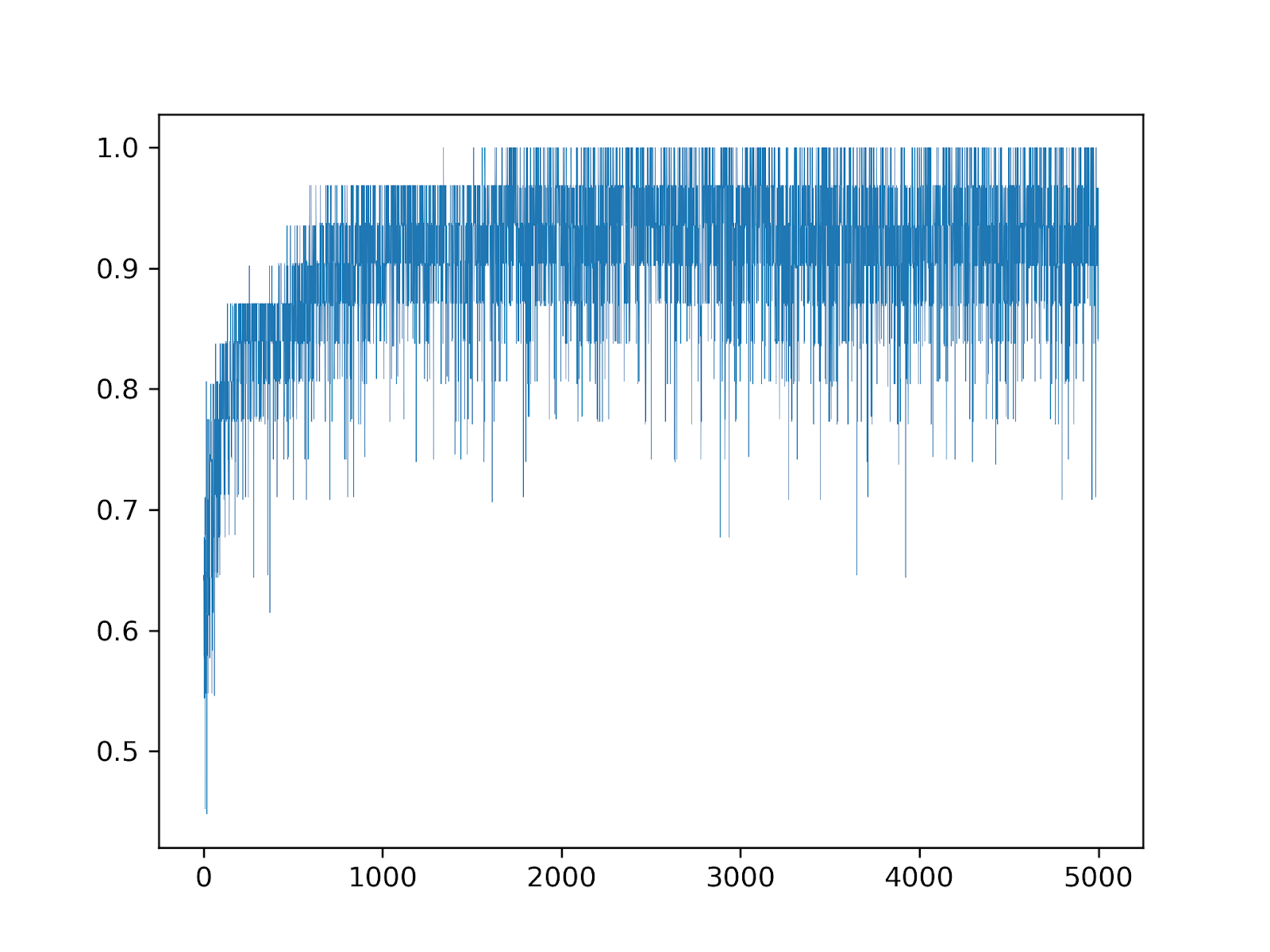


Fig. 1. This is an accuracy plot of our optimized Estimation of Distribution Algorithm predicting authors from our SEC dataset. As you can see, it depicts how the accuracy converges as the number of evaluations increases. This specific figure converges to 100 percent accuracy around the 3000th evaluation, so this also shows us that any more evaluations is redundant.

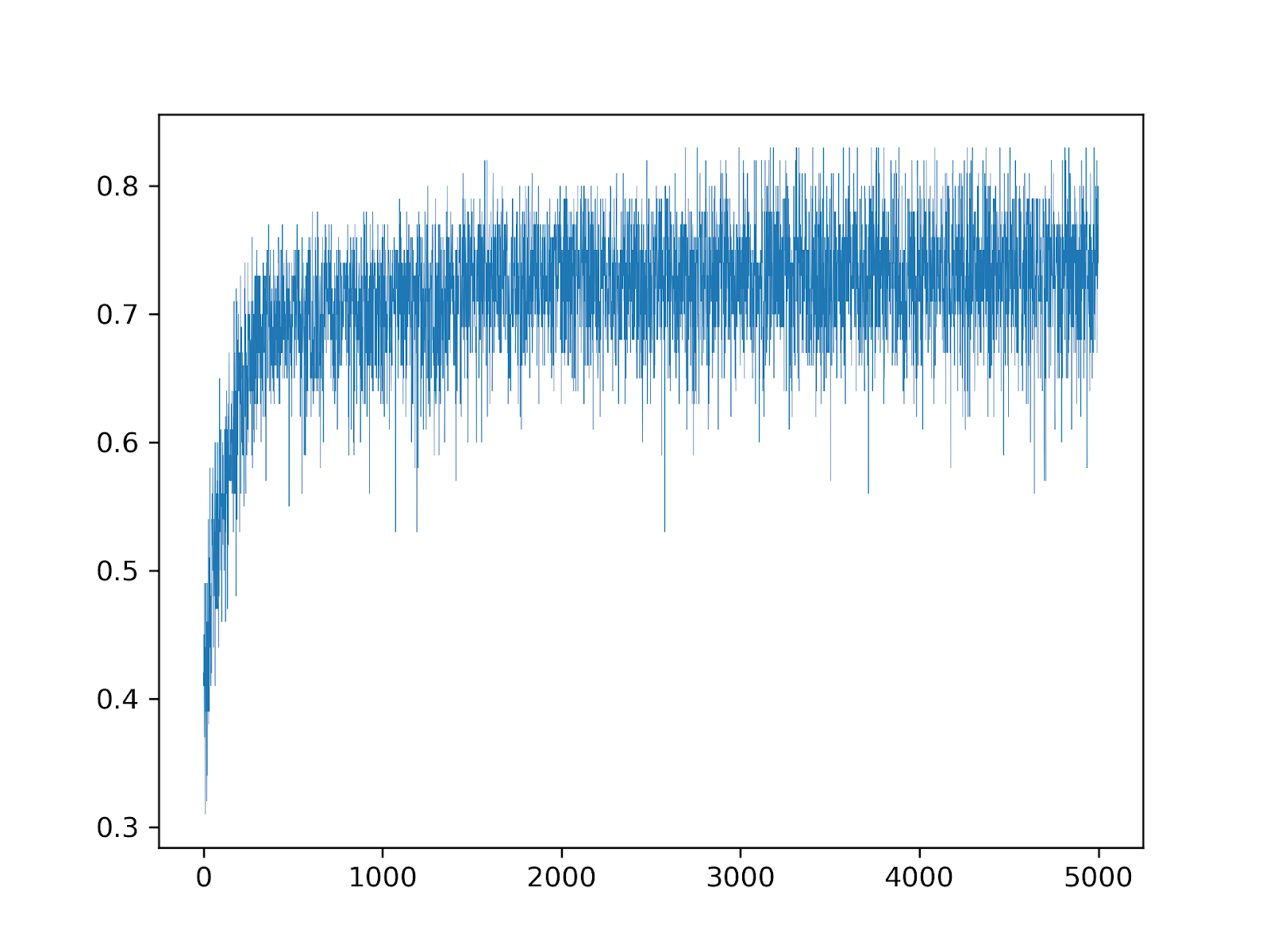


Fig. 2. This is an accuracy plot of our optimized Estimation of Distribution Algorithm predicting authors from our CASIS dataset.

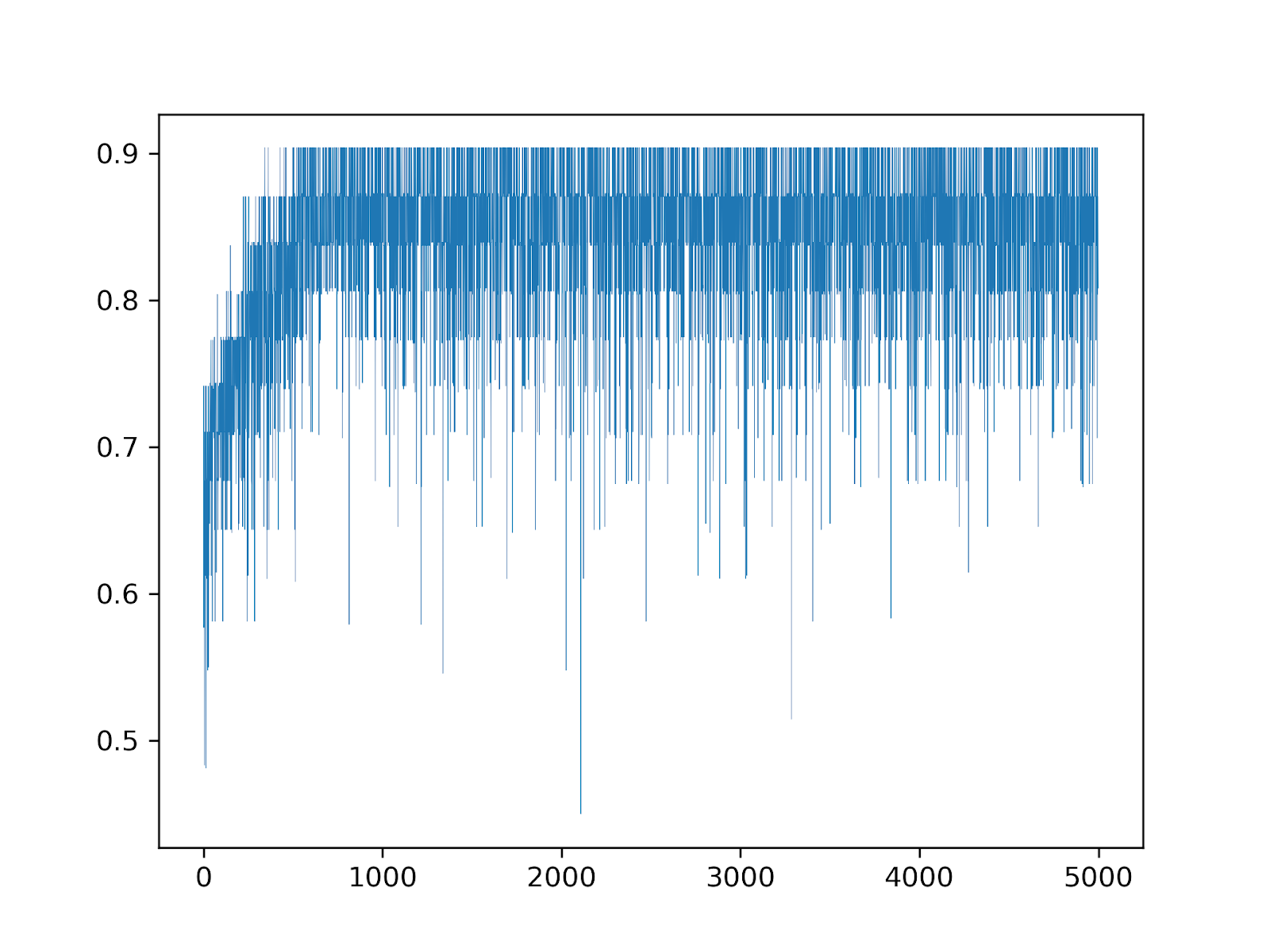


Fig. 3. This is an accuracy plot of our optimized Steady State GA predicting authors from our SEC dataset.

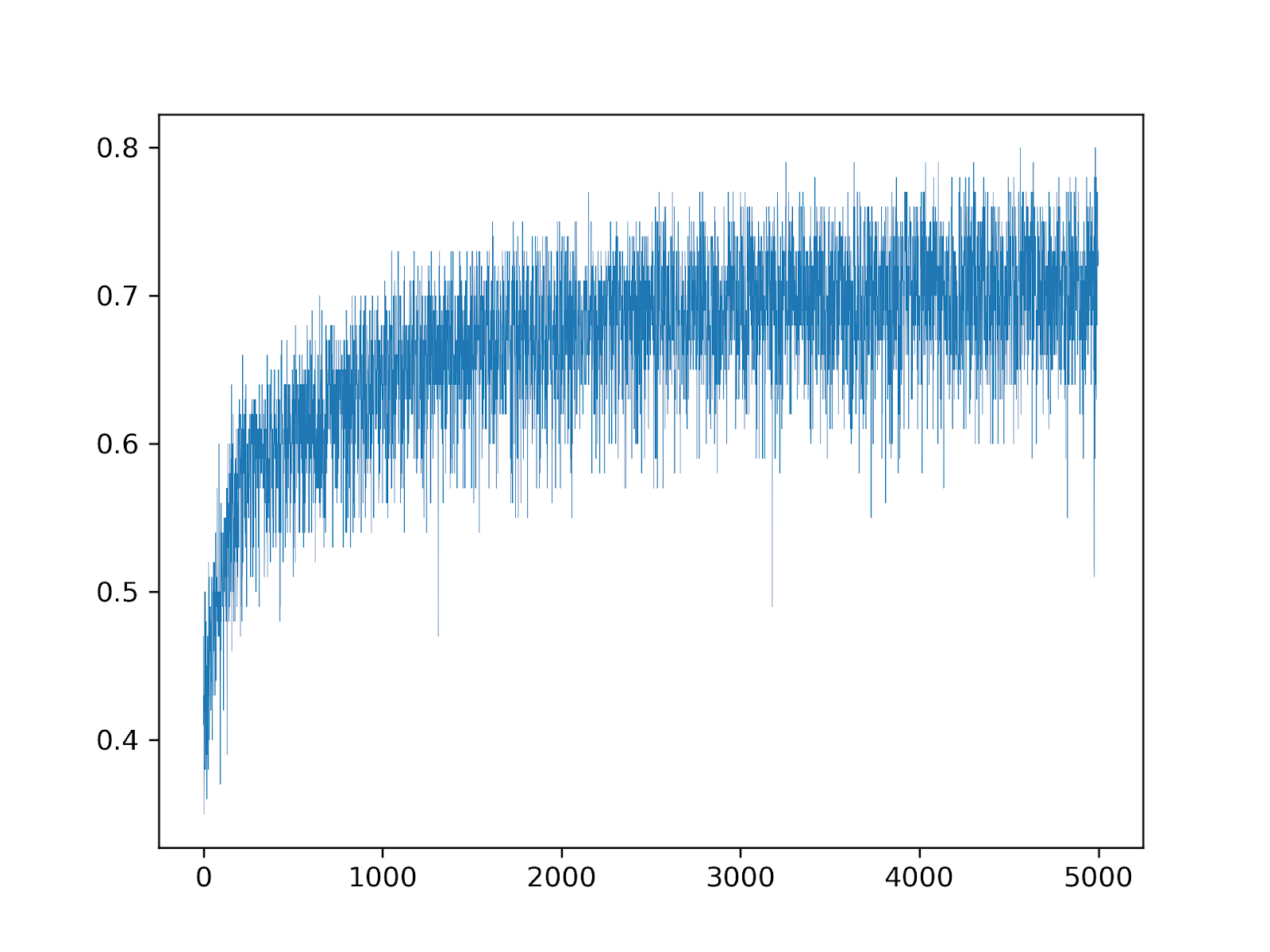


Fig. 4. This is an accuracy plot of our optimized Steady State GA predicting authors from our CASIS dataset.

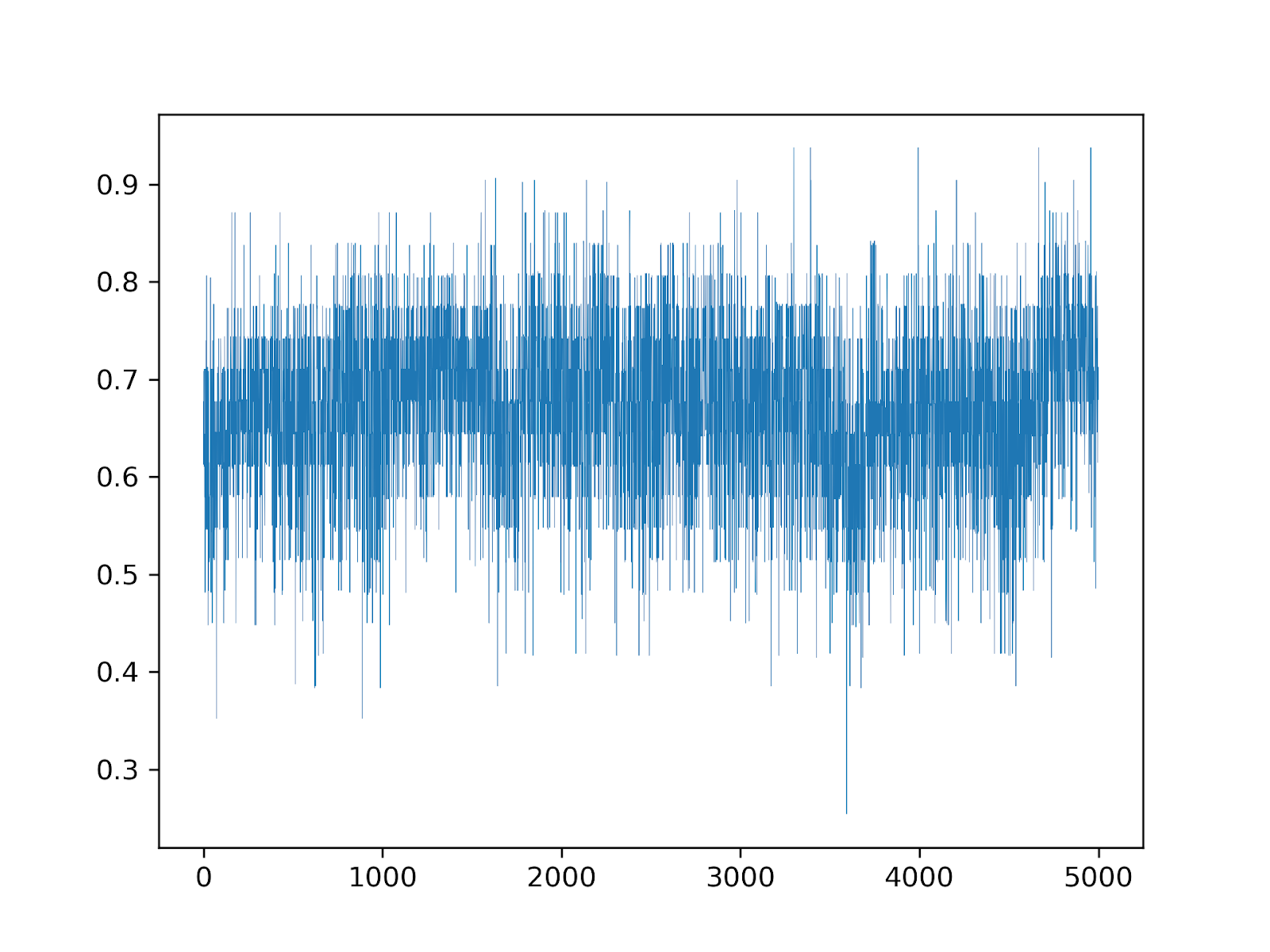


Fig. 5. This is an accuracy plot of our optimized Elitist GA predicting authors from our SEC dataset. As you can see, the Elitist GA has a much wider range of accuracies as the iterations increase because you only save one best individual for each generation.

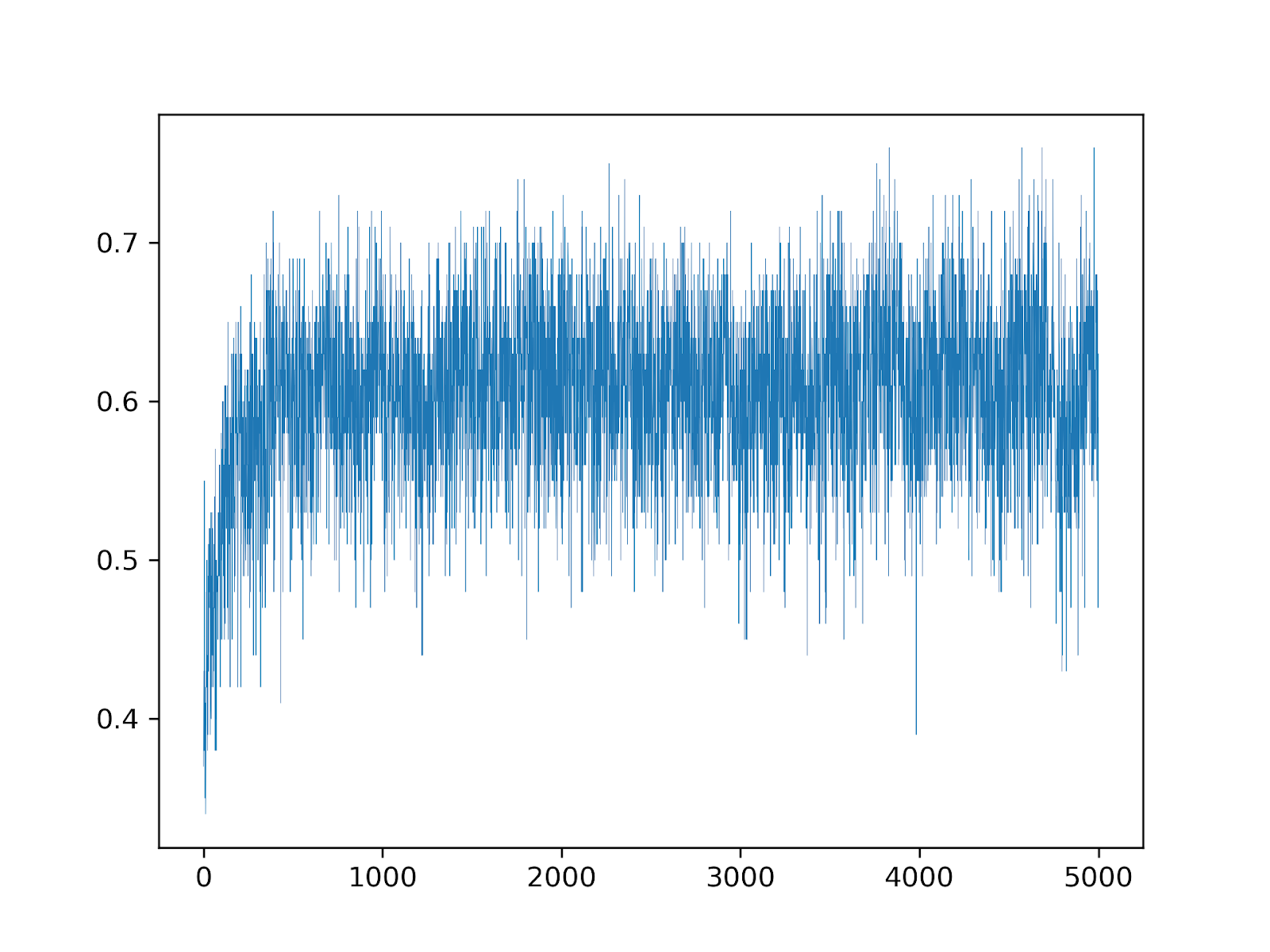


Fig. 6. This is the accuracy plot of our optimized Elitist GA predicting authors on our CASIS dataset.

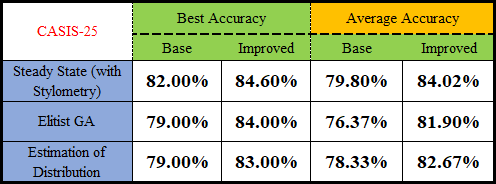


Fig. 7. This shows the accuracies for each Genetic Algorithm when it involves the CASIS-25 dataset. As desired, we found improvements in each algorithm due to optimization. Each algorithm was able to improve around 5%.

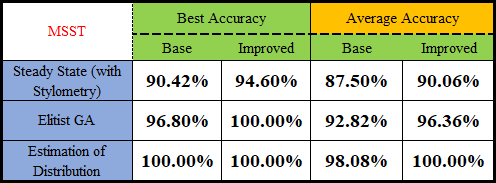


Fig. 8. This shows the accuracies for each Genetic Algorithm when it involves the SEC Sports Writer dataset. As desired, we found improvements with this dataset as well due to optimization. Due to the Estimation of Distribution GA already being at 100% best accuracy, there were no improvements needed to be made. Each algorithm proved to be extremely accurate when dealing with the SEC dataset.

# Breakdown of Work

**Sarp Aykent:** Programmed Steady-State GA and collected data related to said algorithm. Co-wrote paper.

**Jordan Cox:** Programmed Estimation of Distribution Algorithm and collected data related to said algorithm. Co-wrote paper.

**Blake Schilleci:** Programmed Elitist GA and collected data related to said algorithm. Co-wrote paper.

References

[1] A. Press. “No. 18 Mississippi State tops SFA 63-6 in Moorhead's debut.” *ESPN,* 1 Sept. 2018. [Online]. www.espn.com.

[2] A. Press. “Fitzgerald leads no. 18 miss st to 31-10 rout of K-State.” *ESPN,* 8 Sept. 2018. [Online]. www.espn.com.

[3] A. Press. “No. 18 Mississippi State breezes by SF Austin.” *SEC Sports,* 8 Sept. 2018. [Online]. www.secsports.com.

[4] A. Press. “No. 18 Mississippi State stifles Kansas State 31-10.” *SEC Sports,* 8 Sept. 2018. [Online]. www.secsports.com.

[5] H. Cloud. “Defensive dog collars, slam dunks, are all part of a culture change.” *The Reflector,*12 Sept. 2018. [Online]. www.reflector-online.com.

[6] H. Cloud. “Bulldog offense lifts off, defense smothers in first win for Moorhead.” *The Reflector,*2 Sept. 2018. [Online]. www.reflector-online.com.

[7] J. Coleman. “Bulldogs becoming disruptive on defense.” *Bulldogs Becoming Disruptive on Defense | Starkville Daily News*, 2 Sept. 2018. [Online]. Starkvilledailynews.com.

[8] T. Horka. “Running dack Kylin Hill leads Mississippi State to 31-10 victory over Kansas State.” *The Clarion Ledger*, The Clarion-Ledger, 8 Sept. 2018. [Online]. www.clarionledger.com.

[9] T. Horka. “Thompson, Hill lead Mississippi State to 63-6 stomping of Stephen F. Austin.” *The Clarion Ledger*, The Clarion-Ledger, 2 Sept. 2018. [Online]. www.clarionledger.com.

[10] B. Hudson. “MSU's defense shows versatility in dominating Kansas State.” *The Commercial Dispatch*, 10 Sept. 2018. [Online]. www.cdispatch.com.

[11] B Hudson. “Columbus' Hill has rousing start to season.” *The Commercial Dispatch*, 2 Sept. 2018. [Online]. www.cdispatch.com.

[12] D. Skretta. “Thrill from Hill: sophomore running back produces in MSU victory.” *Bulldogs Becoming Disruptive on Defense | Starkville Daily News*, 8 Sept. 2018. [Online]. starkvilledailynews.com.

[13] D. Brandt. “Fitzgerald, No. 16 Mississippi State Rout La.-Lafayette.” *ESPN*, ESPN Internet Ventures, 16 Sept. 2018, www.espn.com/college-football/recap?gameId=401012278.

[14] D. Brandt. “Snell Leads Kentucky Past No. 14 Mississippi State 28-7.” *ESPN*, ESPN Internet Ventures, 23 Sept. 2018, www.espn.com/college-football/recap?gameId=401012286.

[15] D. Brandt. “Mullen, Florida Beat No. 23 Mississippi State 13-6.” *ESPN*, ESPN Internet Ventures, 30 Sept. 2018, www.espn.com/college-football/recap?gameId=401012298.

[16] D. Brandt. “Fitzgerald Leads Miss St to 23-9 Win over No. 8 Auburn.” *ESPN*, ESPN Internet Ventures, 7 Oct. 2018, www.espn.com/college-football/recap?gameId=401012304.

[17] J. Coleman. “A-Train Gets Back on Track for No. 14 MSU.” *A-Train Gets Back on Track for No. 14 MSU | Starkville Daily News*, 16 Sept. 2018, starkvilledailynews.com/node/124495.

[18] G. Graves. “Bad Night: Wildcats Put Damper on Plans of Bulldogs.” *Bad Night: Wildcats Put Damper on Plans of Bulldogs | Starkville Daily News*, 23 Sept. 2018, starkvilledailynews.com/node/124552.

[19] J. Coleman. “Searching for Answers: Bulldogs Left Scratching Their Heads as Former Coach Leaves with a Win.” *Searching for Answers: Bulldogs Left Scratching Their Heads as Former Coach Leaves with a Win | Starkville Daily News*, 30 Sept. 2018, starkvilledailynews.com/node/124604.

[20] J. Coleman. “Sweet Relief: MSU Finds End Zone, Victory over Auburn.” *Sweet Relief: MSU Finds End Zone, Victory over Auburn | Starkville Daily News*, 7 Oct. 2018, starkvilledailynews.com/node/124650.

[21] B. Hudson. “Fitzgerald Looks Dominant as No. 16 Bulldogs Slice and Dice Ragin' Cajuns.” *The Commercial Dispatch*, www.cdispatch.com/sports/article.asp?aid=68524.

[22] B. Hudson. “Kentucky Upsets No. 14 MSU.” *The Commercial Dispatch*, www.cdispatch.com/sports/article.asp?aid=68661.

[23] B. Hudson. “MSU's Offense Struggles in Loss to Florida.” *The Commercial Dispatch*,

www.cdispatch.com/sports/article.asp?aid=68856.

[24] B. Hudson. “Running Games Propels MSU Past No. 8 Auburn.” *The Commercial Dispatch*,

www.cdispatch.com/sports/article.asp?aid=69007.

[25] H. Cloud. “Wilcox Honored, MSU Rolls to Victory in Third Game of Season.” *The Reflector*, 19

Sept. 2018, www.reflector-online.com/sports/article\_c849b1bc-b977-11e8-9f39-3faf1e8637ad.html.

[26] H. Cloud. “Bulldogs Knock down No. 9 Tigers.” *The Reflector*, 8 Oct. 2018, www.reflector-

online.com/sports/article\_bf4c2cc4-c9ef-11e8-8bb1-8b906830601d.html.

[27] H. Cloud. “Moving on from Kentucky, MSU Focused on First Home SEC Game.” *The Reflector*, 27

Sept. 2018, www.reflector-online.com/sports/article\_a15ade6a-c25e-11e8-bb32-ab7b74da8d2a.html.

[28]H. Cloud. “Mullen's Return Proves Unhappy for Bulldogs.” *The Reflector*, 30 Sept. 2018, www.reflector-online.com/sports/article\_36296bcc-c469-11e8-af8e-eb779ea01df7.html.

[29] H. Cloud. “Making a statement in win over Aggies,” The Reflector, 30-Oct-2018. [Online]. Available: http://www.reflector-online.com/sports/article\_6767fa6e-daf8-11e8-bf0b-7b9c303aff93.html. [Accessed: 04-Nov-2018].

[30] H. Cloud, “MSU falls silently as offense dies in Death Valley,” The Reflector, 22-Oct-2018. [Online]. Available: http://www.reflector-online.com/sports/article\_af69ae1c-d4e4-11e8-ac39-5b3a9e9c1c4d.html. [Accessed: 04-Nov-2018].

[31] B. Hudson. “LSU delivers Death Valley demolition,” The Dispatch, 22-Oct-2018. [Online]. Available: http://www.cdispatch.com/sports/article.asp?aid=69313. [Accessed: 04-Nov-2018].

[32] B. Hudson. “Fitzgerald (4 TDs) leads MSU past No. 16 Texas A&M,” The Dispatch, 30-Oct-2018. [Online]. Available: http://www.cdispatch.com/sports/article.asp?aid=69500. [Accessed: 04-Nov-2018].

[33] J. Coleman. “Going about business: MSU's Fitzgerald silences critics, leads MSU to big home win,” Starkville Daily News, 22-Oct-2018. [Online]. Available: http://www.starkvilledailynews.com/node/124802. [Accessed: 04-Nov-2018].

[34] J. Coleman. “Bright spot: Despite loss, MSU defense continues to play well,” Starkville Daily News, 22-Oct-2018. [Online]. Available: http://www.starkvilledailynews.com/node/124756. [Accessed: 04-Nov-2018].

[35] B. Martel. “Defense leads No. 5 LSU past No. 22 MSU, 19-3,” ESPN, 20-Oct-2018. [Online]. Available: http://www.espn.com/college-football/recap?gameId=401012315. [Accessed: 04-Nov-2018].

[36] D. Brandt. “Mississippi State upsets No. 16 Texas A&M, 28-13,” ESPN, 27-Oct-2018. [Online]. Available: http://www.espn.com/college-football/recap?gameId=401012321. [Accessed: 04-Nov-2018].

[37]A. Narayanan, H. Paskov, N. Z. Gong, J. Bethencourt, E. Stefanov, E. C. R. Shin, and D. Song, *On the Feasibility of Internet-Scale Author Identification - IEEE Conference Publication*. [Online]. Available: https://ieeexplore.ieee.org/document/6234420. [Accessed: 04-Nov-2018].

1. We computed our features in the same way as [37] [↑](#footnote-ref-1)