Fake Account Filter

Project Deliverable 2, MAIS 202

1. Problem statement

I'm making an Instagram fake/spam account classifier using a dataset[1] from Kaggle. I'll try out multiple classification models with the help of scikit-learn libraries like support vector machines, random forests and naive bayes.

2. Data Preprocessing

This dataset was created by someone who classified all the samples by hand which means it could contain some human error in the classes. The ~700 data points each have eleven features and and split in two classes (fake or not). For now, I've only had the time to split the data into sets of vectors with each component as features. I have also standardized the data since I found that my SVM model would not work properly as the value did not converge. However, I have noticed that some features seem to have very little variance, meaning that it barely affects the predictions. Similarly, others seem to have a higher impact on the results. So, I was thinking of weighting the features according to their "importance" (criteria by which this is determined to come…) Then, because the data was already split into training and test sets, I just ran my first models with those. However, I think I should merge that data and split it again to obtain a validation set.

3. Machine learning model

I have tested SVM, RFC, and Bernoulli's Naive Bayes from scikit-learn's library on my datasets and SVM seems to be the most promising model. The RFC model tends to overfit and naive bayes is less accurate than SVM according to the training and test scores. After training the model, I

[1] Bakhshandeh, B. (n.d.). Instagram fake spammer genuine accounts. Retrieved from Kaggle: https://www.kaggle.com/free4ever1/instagram-fake-spammer-genuine-accounts

tested it with some data from accounts that I know are real and it worked. However, I need to test with some fake accounts too. I also have this idea of integrating something that would analyze the biography of each account to find suspicious words (like in assignment #2), but I don't have this data available in this dataset so I need to do more research.

4. Preliminary results

From the SVM model from scikit-learn, the scores are around 91% for the training set and 87% for the test set. However, I feel like the variance in the data is too small to really capture the problem. I will have to find other ways manipulate my features to get better results.

5. Next steps

I need to do more research about the issue of filtering fake accounts on social media and how other people approached it to see how I can improve my model. Given that this is still a problem that companies have to this day, I know that it's not as easy as it can seem.