Twitter Analytics

A Short Overview of Identifying Spammers on Twitter

Overview of problem

- Twitter is one of the fastest growing social networking sites
- It attracts many spammers due to its popularity
- Spam accounts pose a threat to the legitimate users of the service, as well as Twitter Analytics services

Overview of problem

- Previous studies have showed that 45% of users on social networking sites readily click on links posted by people in their network [Mowbray, The Twittering Machine, 2010]
- Spammers are attracted to Twitter as a tool to send unsolicited messages to legitimate users, post malicious links and hijack trending topics

How much of Twitter is spam?

Forbes | New Posts +10 posts this hour | Popular 5 LinkedIn Strategies | Lists The Most Powerful Peop | Video Hillary Cli

The ONE Stock to Buy in November

Tristan Louis, Contributor
I write about internet & mobile technology.
+ Follow (54) Follow 381

TECH | 4/07/2013 @ 4:58PM | 3,419 views

Twitter's Growing Spam Problem

+ Comment Now + Follow Comments

There's been an increasing amount of reporting about the underground economy that powers the sale of retweets and followers on Twitter, leading to some questions as to how popular the service truly is. With reports

Image via CrunchBase

- Hard to determine; research indicates anywhere from 5-45 %
- According to Microsoft, 97% of all emails are unwarranted

Twitter spammer techniques

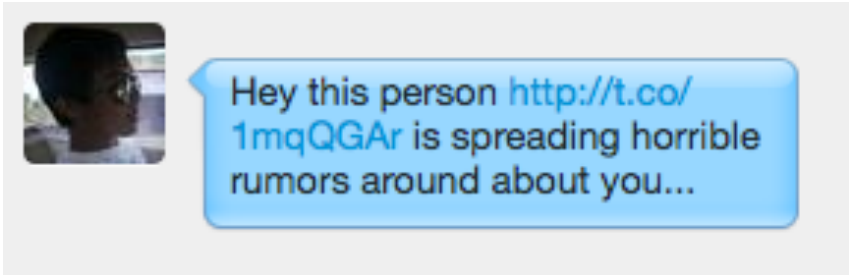
- Hijacking popular topics (i.e. trending topics) and obfuscating web address through shorteners

double_tre3646 **#iranelection** | Download and Watch Public Enemies, Bruno, Ice Age and the other Lastest Movies Online. http://bit.ly/TMD714182 E17C

4 minutes ago from API

Twitter spammer techniques

- Spamming via direct messages (DMs)



Twitter spammer techniques

- Spam networks: Accounts that follow each other
- Mass-following, mass-replying and mass-favoriting to grab attention
- The goal is to get users to click on phishing or malware URLs, or affiliate links
- Grey area: Followback-networks, Viral spam (companies engaging in spammy behavior)

Methodology

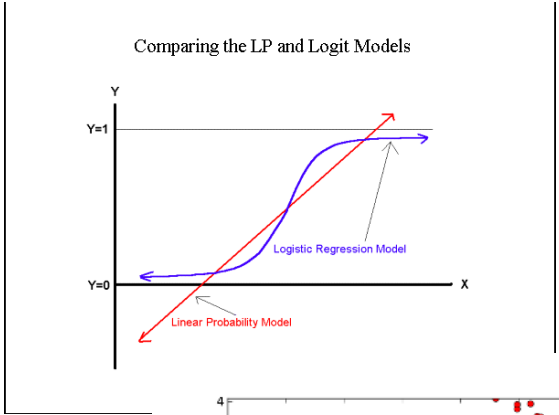
- Collect dataset of legitimate and spam accounts -> 105 accounts in total
- Identify features that are helpful in separating spammers from legitimate users
- Compare various classification algorithms based on the suggested features

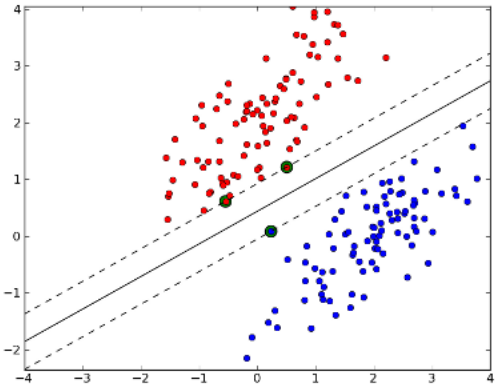
Naive Bayes Classifier

- A very simple classifier that treats each feature as strongly independent (hence *Naive*) and applies Bayes' theorem (hence *Bayes*). Works surprisingly well with small datasets, it is quick and simple.
- But doesn't take into account interactions between features

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

Logistic Regression & SVM

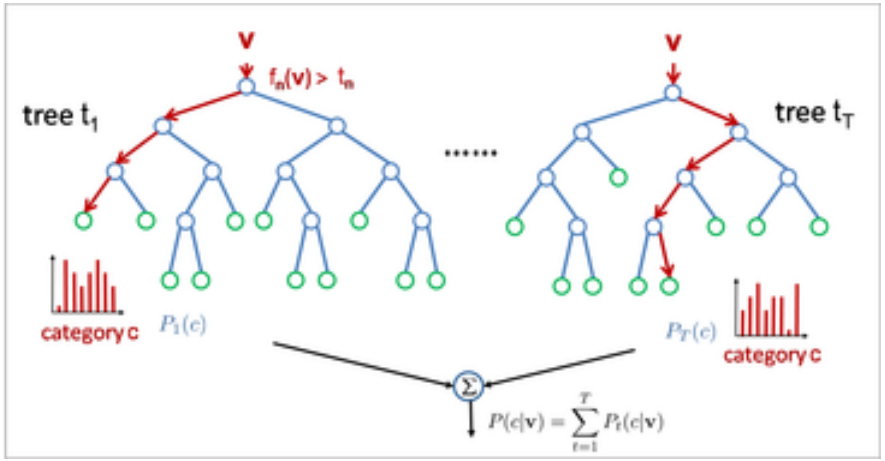




They both separate the plane into two halves, with more or less complex likelihood functions

Logistic regression is similar to linear regression, but maps to between 0 and 1, good for binary classification (spam/ham)

Random Forest



Forms multiple decision trees based on a random sample of features, then takes the average of the outcomes of the trees

Simple, performs well, benchmark for Kaggle competitions

User-based features

1. Followers
2. Following
3. “Friends” (Reciprocated following)
4. Activity = No. of status messages posted

Content-based features from last 300 tweets

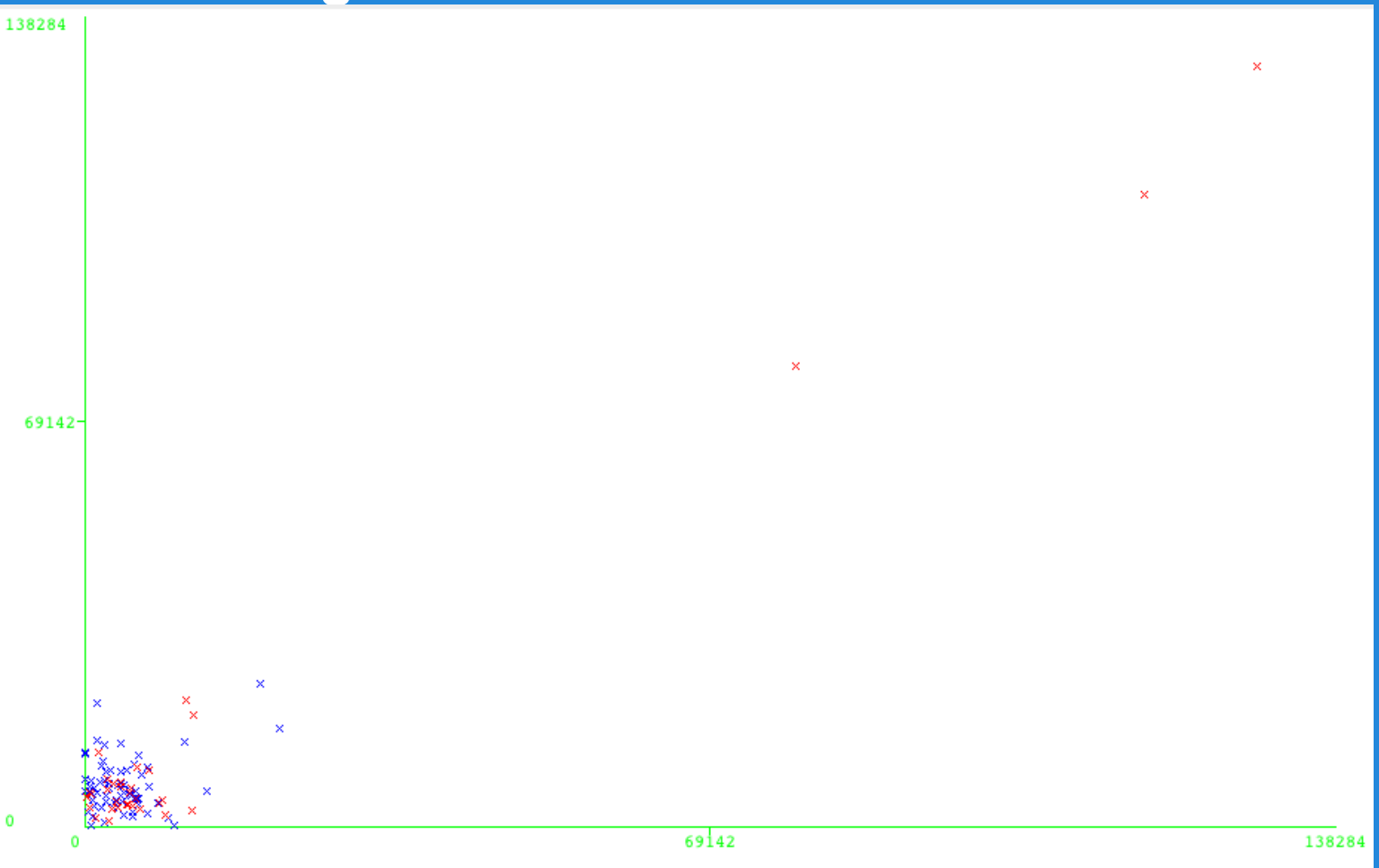
1. Percentage of URLs posted
2. Percentage of mentions
3. Sametext: Number of tweets that are the same message (excluding URLs and hashtags)

Problems with User-based features

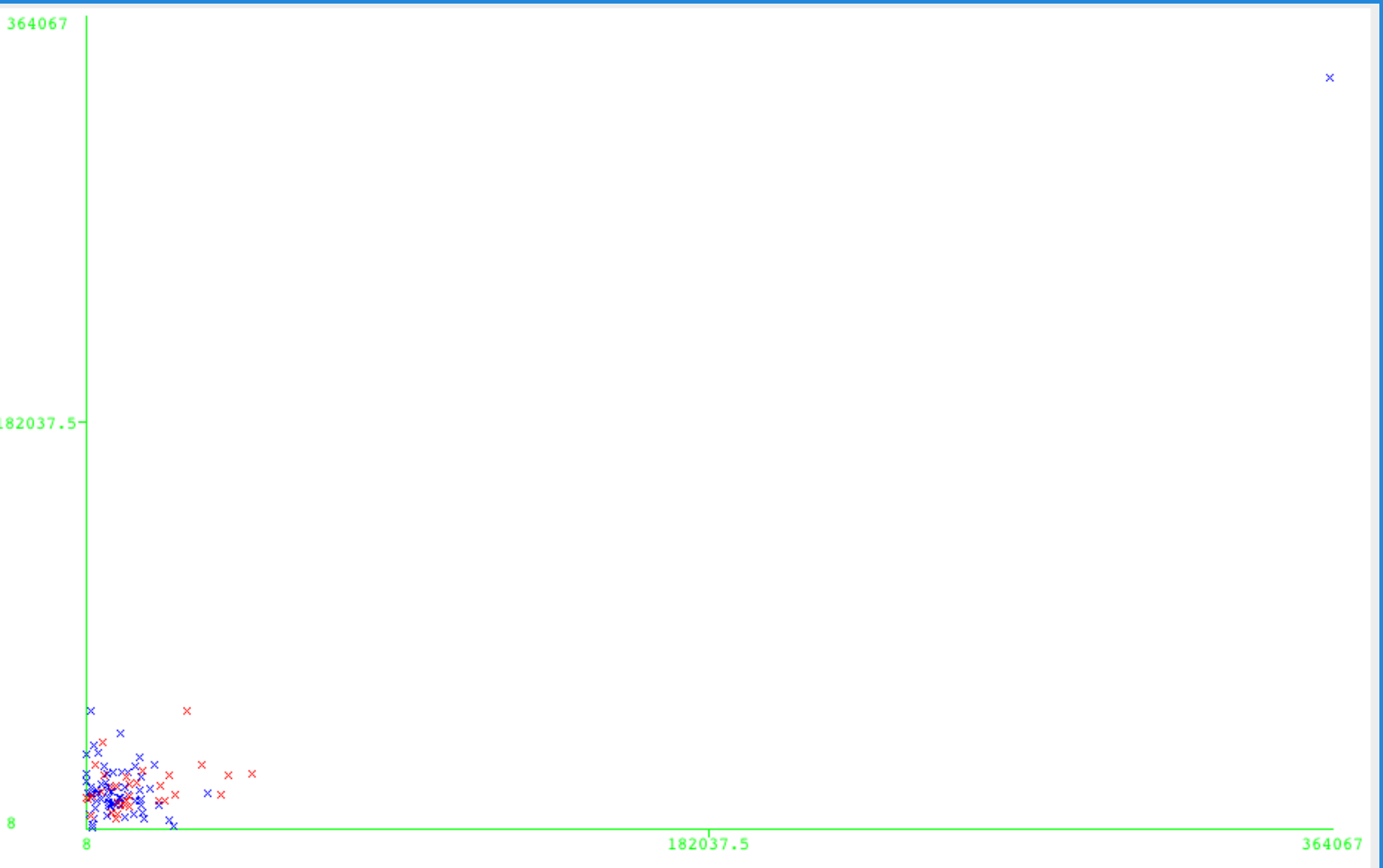
- A lot of spam accounts operate in networks by now; they will have many followers, since everybody in the network will follow each other & ppl are not selective in who they follow back (“courtesy-follow”)
- Number of status messages posted will not be a decisive factor, since there are many legitimate users who post many times

Followers

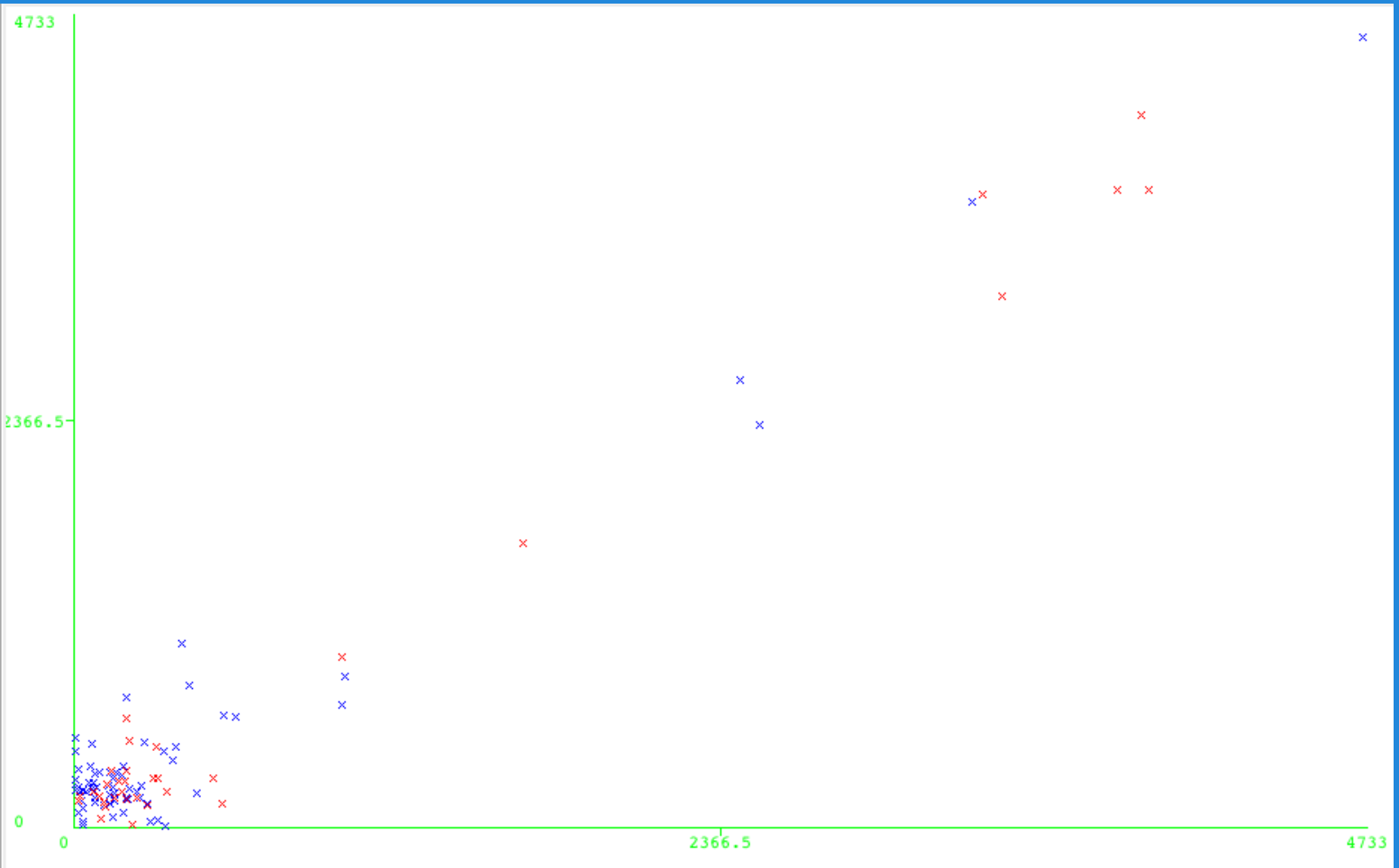
Following



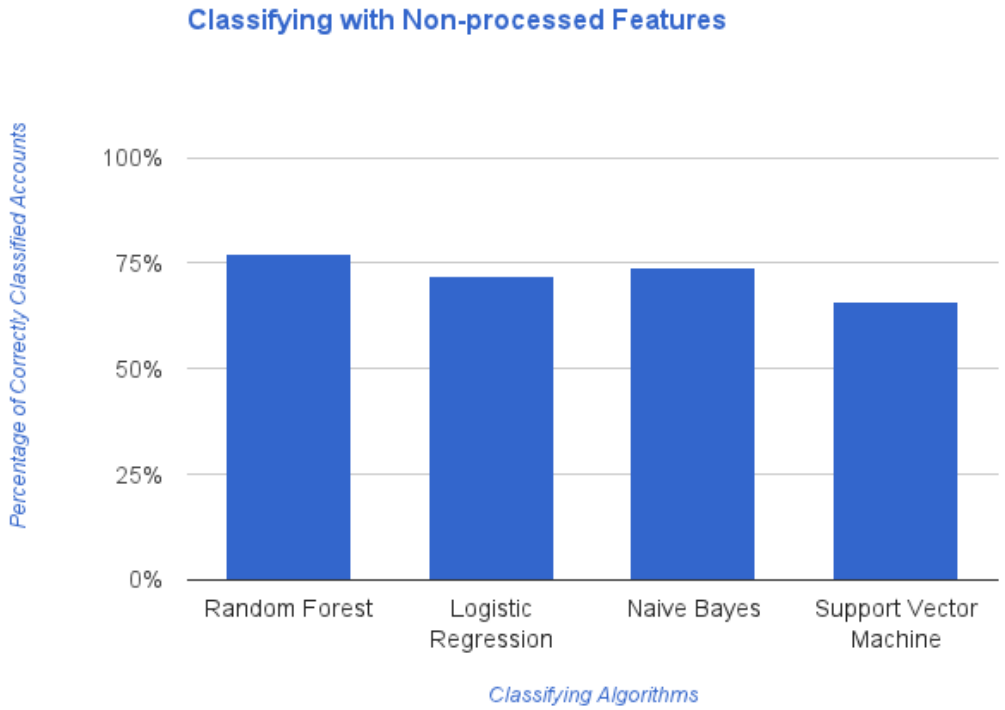
Statuses



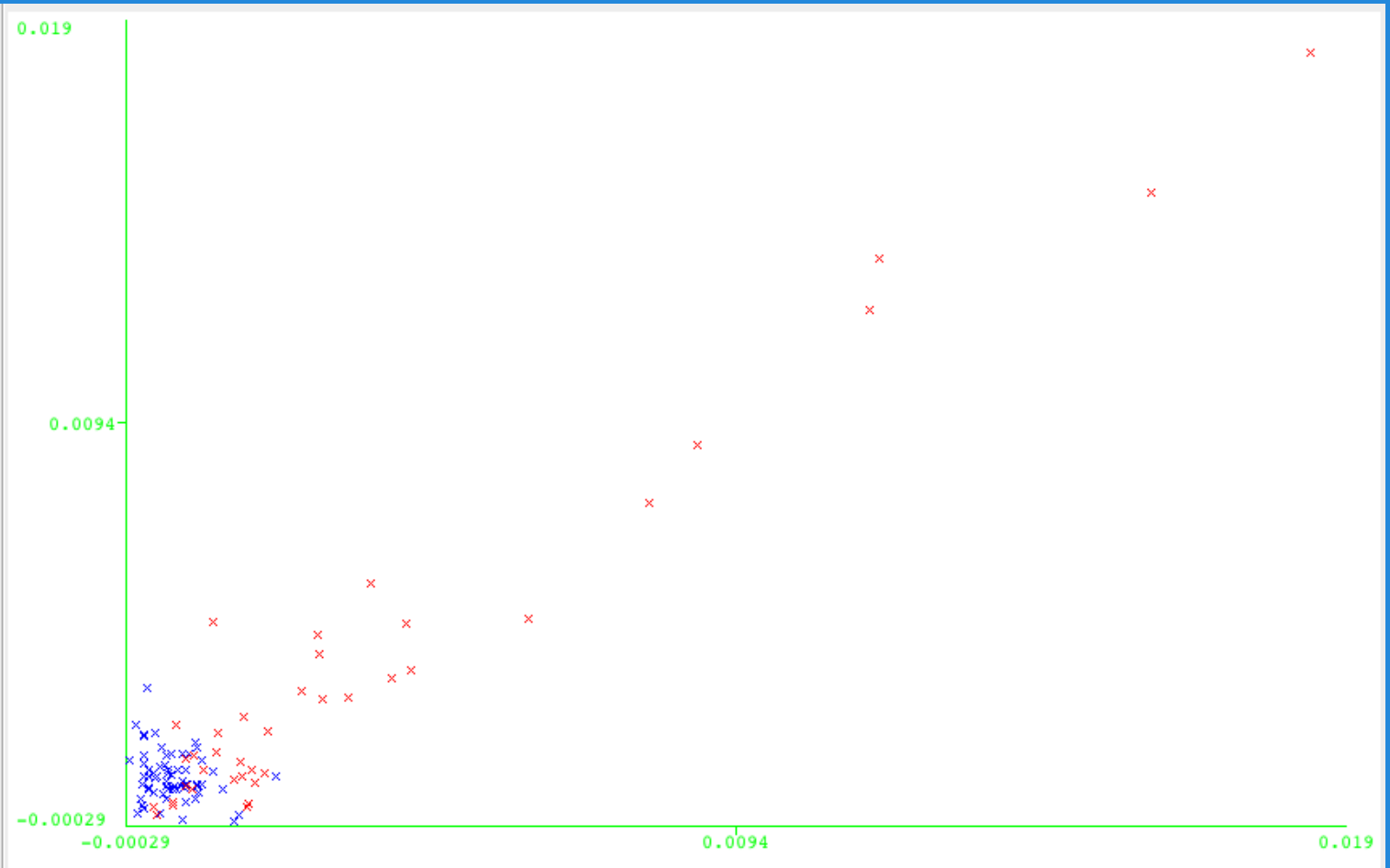
Friends



Classification without pre-processing



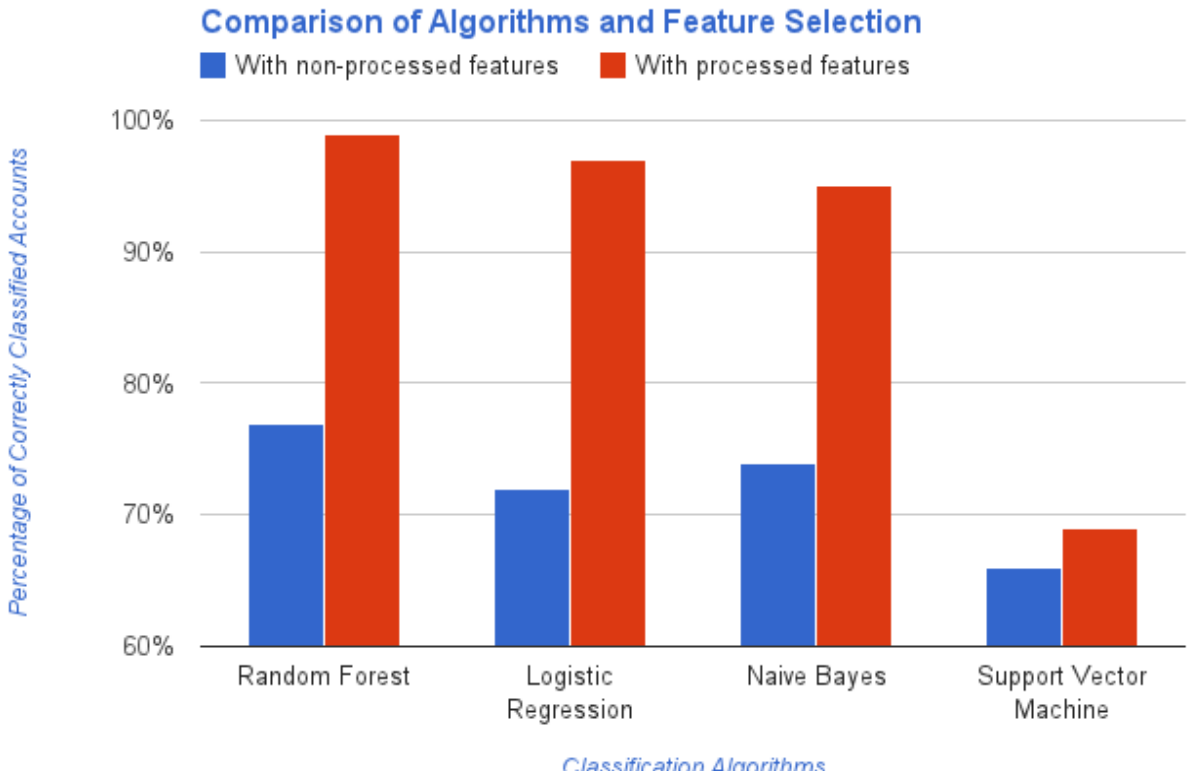
Tweetspeed = number of tweets / account age



Features used in classification

- Followspeed
- Tweetspeed
- Sametext
- Percentage of Links
- Percentage of Mentions
- Followratio

Comparison of Classification Algorithms & Feature Selection



mavenr
connecting knowledge

Ranking of features

1. Speed of tweeting: Account age/amounts of tweets sent
2. Speed of following: Account age/number of ppl followed
3. Number of virtually same tweets (stripped of URLs and hashtags) in last 300 tweets
4. Follow ratio: Followers / Followers + Friends
5. Percentage of links posted in last 300 tweets
6. Percentage of mentions posted in last 300 tweets

Conclusion

- Feature selection and data processing is equally important as algorithm selection (sometimes even more)
- A good starting point in feature selection is visualizing attributes
- Random Forest algorithm + various content and user-based features work remarkably well in identifying spammer activity
- One of the most important features for Twitter spam identification is based on speed or account age

Thank you!

Demet Dagdelen

demet.dagdelen@maven7.com