Identifying Twitter Spam by Utilizing Random Forests

Humza Haider
Division of Science and Mathematics
University of Minnesota Morris

2017-04-15

Introduction

- Top social media platforms

Introduction

- Top social media platforms
- 500 million tweets per day

Introduction

- Top social media platforms
- 500 million tweets per day
- Attracts spammers and malicious users

Introduction

- Top social media platforms
- 500 million tweets per day
- Attracts spammers and malicious users
- Twitter spam: Any unsolicited, repeated actions that negatively impact other users

Introduction

- Top social media platforms
- 500 million tweets per day
- Attracts spammers and malicious users
- Twitter spam: Any unsolicited, repeated actions that negatively impact other users
- How can we identify spammers?

Introduction

- Top social media platforms
- 500 million tweets per day
- Attracts spammers and malicious users
- Twitter spam: Any unsolicited, repeated actions that negatively impact other users
- How can we identify spammers?
  - Manual classification

Introduction

- Top social media platforms
- 500 million tweets per day
- Attracts spammers and malicious users
- Twitter spam: Any unsolicited, repeated actions that negatively impact other users
- How can we identify spammers?
  - Manual classification
  - URL blacklisting

Introduction

- Top social media platforms
- 500 million tweets per day
- Attracts spammers and malicious users
- Twitter spam: Any unsolicited, repeated actions that negatively impact other users
- How can we identify spammers?
  - Manual classification
  - URL blacklisting
  - **Machine learning classification**

Outline

- Decision Trees

- Decision Trees
  - Machine learning technique for classification

- Decision Trees
  - Machine learning technique for classification
  - Classifies an observation based on features available in a dataset

URL	Account Age	Reported	**Class**
No	Old	Yes	**Not Spam**
No	Old	Yes	**Not Spam**
No	Old	No	**Not Spam**
No	New	No	**Not Spam**
Yes	New	Yes	**Spam**
No	New	Yes	**Spam**
No	Old	Yes	**Spam**
Yes	New	No	**Not Spam**
⋮	⋮	⋮	⋮

Background

Decision Trees

URL	Account Age	Reported	**Class**
No	Old	Yes	**Not Spam**
No	Old	Yes	**Not Spam**
No	Old	No	**Not Spam**
No	New	No	**Not Spam**
Yes	New	Yes	**Spam**
No	New	Yes	**Spam**
No	Old	Yes	**Spam**
Yes	New	No	**Not Spam**

Background

Decision Trees

32% Spam
68% Not Spam

<u>8 Tweets</u>
3 Spam
5 Not Spam

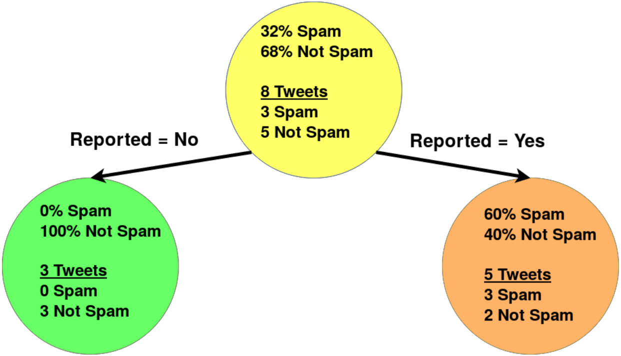
Background

Decision Trees

URL	Account Age	**Reported**	**Class**
No	Old	**Yes**	**Not Spam**
No	Old	**Yes**	**Not Spam**
No	Old	**No**	**Not Spam**
No	New	**No**	**Not Spam**
Yes	New	**Yes**	**Spam**
No	New	**Yes**	**Spam**
No	Old	**Yes**	**Spam**
Yes	New	**No**	**Not Spam**

Background

Decision Trees



Background

Decision Trees

URL	Account Age	Reported	**Class**
No	Old	Yes	**Not Spam**
No	Old	Yes	**Not Spam**
~~No~~	~~Old~~	~~No~~	~~**Not Spam**~~
~~No~~	~~New~~	~~No~~	~~**Not Spam**~~
Yes	New	Yes	**Spam**
No	New	Yes	**Spam**
No	Old	Yes	**Spam**
~~Yes~~	~~New~~	~~No~~	~~**Not Spam**~~

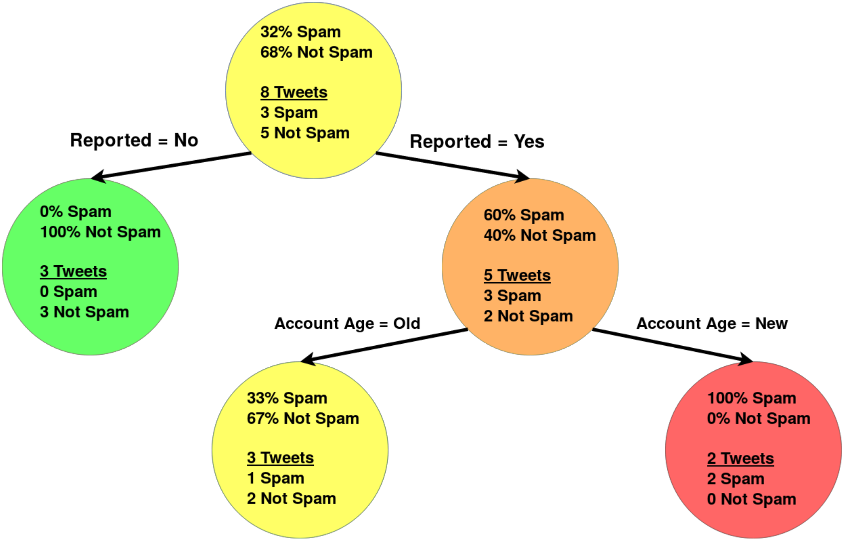
Background

Decision Trees

URL	**Account Age**	Reported	**Class**
No	**Old**	Yes	**Not Spam**
No	**Old**	Yes	**Not Spam**
~~No~~	~~**Old**~~	~~No~~	~~**Not Spam**~~
~~No~~	~~**New**~~	~~No~~	~~**Not Spam**~~
Yes	**New**	Yes	**Spam**
No	**New**	Yes	**Spam**
No	**Old**	Yes	**Spam**
~~Yes~~	~~**New**~~	~~No~~	~~**Not Spam**~~

Background

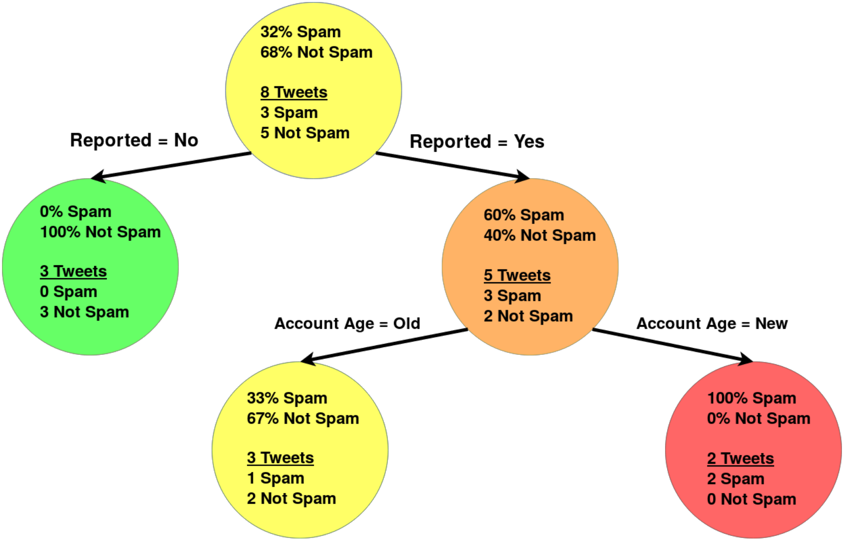
Decision Trees



URL	Account Age	Reported	**Class**
No	Old	Yes	**Not Spam**
No	Old	Yes	**Not Spam**
~~No~~	~~Old~~	~~No~~	~~**Not Spam**~~
~~No~~	~~New~~	~~No~~	~~**Not Spam**~~
~~Yes~~	~~New~~	~~Yes~~	~~**Spam**~~
~~No~~	~~New~~	~~Yes~~	~~**Spam**~~
No	Old	Yes	**Spam**
~~Yes~~	~~New~~	~~No~~	~~**Not Spam**~~

Background

Decision Trees



Background

Decision Trees

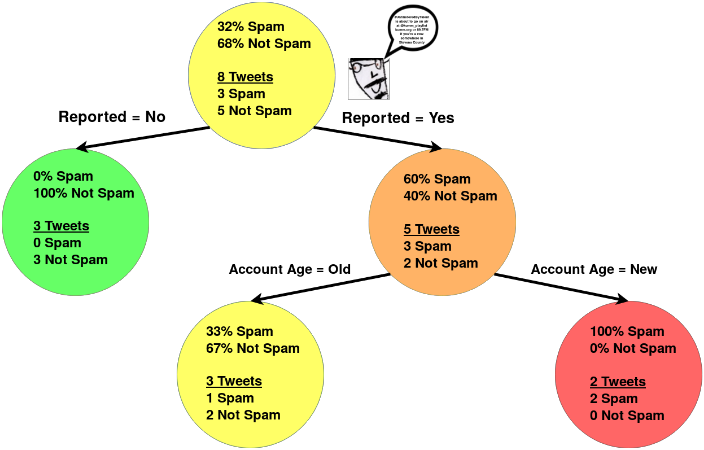
Nic McPhee @NicMcPhee · Jan 29

#UnhinderedByTalent is about to go on air at @kumm_playlist kumm.org or 89.7FM if you're a cow somewhere in Stevens County.

URL	Account Age	Reported	**Class**
Yes	Old	Yes	TBD

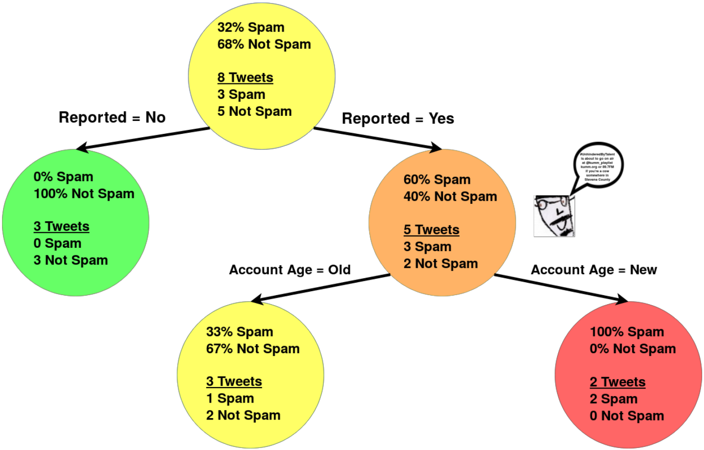
Background

Decision Trees



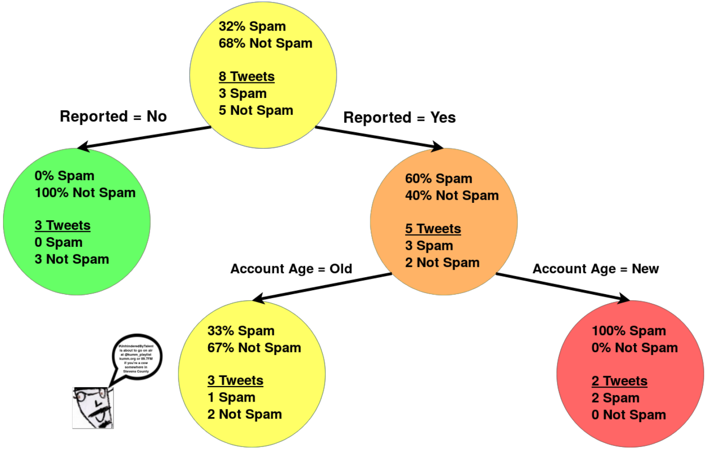
Background

Decision Trees



Background

Decision Trees



- How are splits decided?

- How are splits decided?
  - Entropy

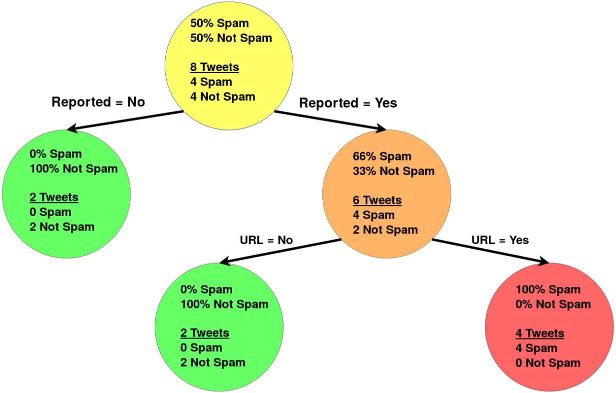
- How are splits decided?
  - Entropy
  - Information Gain

- How are splits decided?
  - Entropy
  - Information Gain
- Trees seem pretty neat! Why do I need a whole forest?

- How are splits decided?
  - Entropy
  - Information Gain
- Trees seem pretty neat! Why do I need a whole forest?
  - Disagreement in decisions between different trees

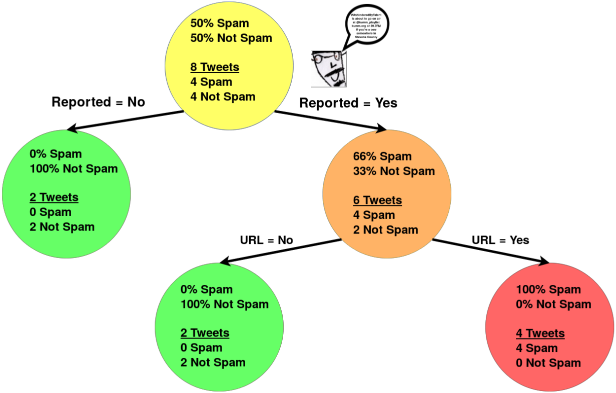
Background

Decision Trees



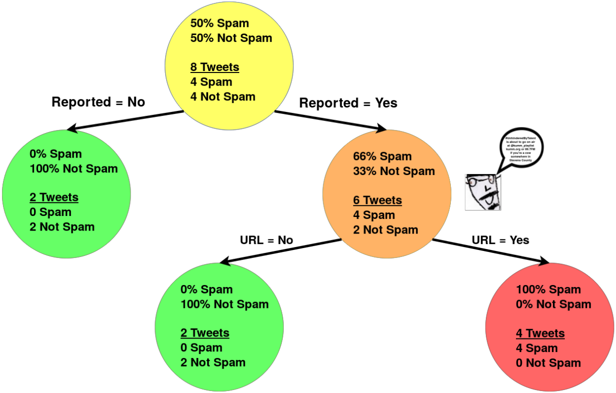
Background

Decision Trees



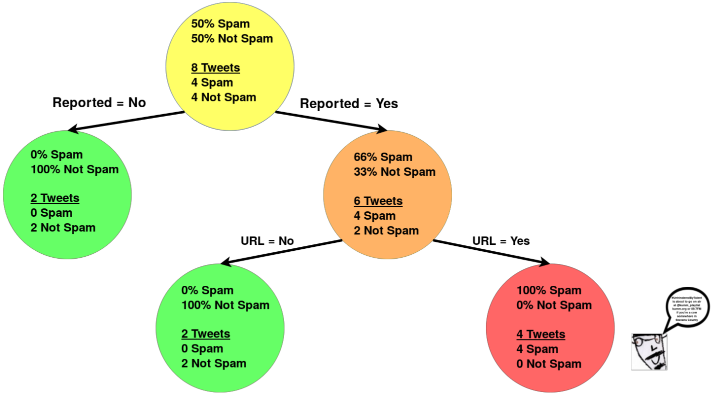
Background

Decision Trees



Background

Decision Trees



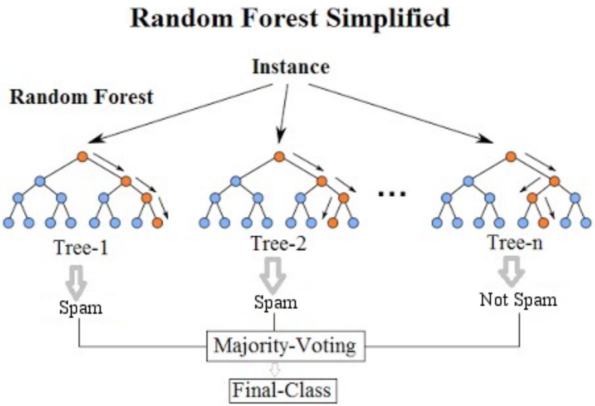
- How do we handle disagreement?

- How do we handle disagreement?
  - Train many trees on samples of the data (**Bagging**)

- How do we handle disagreement?
  - Train many trees on samples of the data (**Bagging**)
  - Don’t let trees access all the features (**Feature Bagging**)

- How do we handle disagreement?
    - Train many trees on samples of the data (**Bagging**)
    - Don't let trees access all the features (**Feature Bagging**)
- After we make a bunch of trees, how do we combine them?

- How do we handle disagreement?
  - Train many trees on samples of the data (**Bagging**)
  - Don’t let trees access all the features (**Feature Bagging**)
- After we make a bunch of trees, how do we combine them?
  - Majority vote



Source: https://i.ytimg.com/vi/ajTc5y3OqSQ/hqdefault.jpg

- How do we evaluate a random forest's performance?

- How do we evaluate a random forest's performance?

Truth

Prediction	Spam	Not Spam
Spam	True Positive	False Positive
Not Spam	False Negative	True Negative

Accuracy

- "How many tweets were correctly identified?"

		Truth: Spam	Truth: Not Spam
Prediction	Spam	True Positive	False Positive
Prediction	Not Spam	False Negative	True Negative

Accuracy

- "How many tweets were correctly identified?"

		Truth: Spam	Truth: Not Spam
Prediction	Spam	**True Positive**	False Positive
Prediction	Not Spam	False Negative	**True Negative**

Accuracy

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision (p)

- “How good is our spam prediction?"

		Truth: Spam	Truth: Not Spam
Prediction	Spam	True Positive	False Positive
	Not Spam	False Negative	True Negative

Precision (p)

- "How good is our spam prediction?"

Truth

Prediction		Spam	Not Spam
	Spam	**True Positive**	**False Positive**
	Not Spam	**False Negative**	**True Negative**

Precision

$$\text{Precision} = \frac{TP}{TP + FP}$$

<u>Recall</u> (r)

- “How much spam was identified?"

		Truth	
		Spam	**Not Spam**
Prediction	**Spam**	**True Positive**	**False Positive**
	Not Spam	**False Negative**	**True Negative**

Recall (r)

- “How much spam was identified?"

		Truth	
		Spam	**Not Spam**
Prediction	**Spam**	**True Positive**	**False Positive**
	Not Spam	**False Negative**	**True Negative**

Recall

$$\text{Recall} = \frac{TP}{TP + FN}$$

- F-measure (F)
  - Harmonic mean of Precision and Recall
  - Equally weights both Precision and Recall

F-Measure

$$\text{F-measure} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Outline

Methods
Tweet and User Content Features

- Chen et al. identify tweets, as opposed to users

Methods

Tweet and User Content Features

- Chen et al. identify tweets, as opposed to users
- Utilized 12 features directly accessible from a tweet

- Chen et al. identify tweets, as opposed to users
- Utilized 12 features directly accessible from a tweet
  - 6 user features

- Chen et al. identify tweets, as opposed to users
- Utilized 12 features directly accessible from a tweet
  - 6 user features
  - 6 tweet features

Methods
Tweet and User Content Features

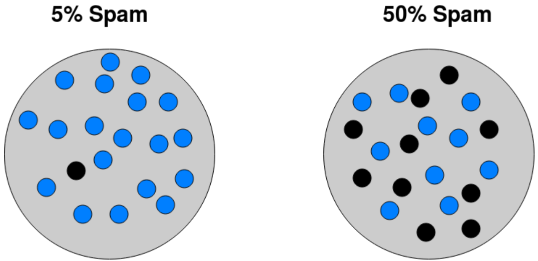
- Chen et al. identify tweets, as opposed to users
- Utilized 12 features directly accessible from a tweet
  - 6 user features
  - 6 tweet features
- User Features
  - Age in days of account
  - Number of followers, followees
  - Number of tweets

Methods

Tweet and User Content Features

- Chen et al. identify tweets, as opposed to users
- Utilized 12 features directly accessible from a tweet
  - 6 user features
  - 6 tweet features
- User Features
  - Age in days of account
  - Number of followers, followees
  - Number of tweets
- Tweet Features
  - Number of hashtags (#)
  - Number of mentions
  - Number of URLs

- Two sets of testing data.



Outline

Methods
Geo-Tagged Features

- So, what is a geo-tagged tweet?

- So, what is a geo-tagged tweet?

Morris Weather
@MorrisMNWeather

Follow

The weather is boring. 50ºF and Light Rain. #MorrisMNWeather

6:02 PM - 9 Apr 2017 from Morris, MN

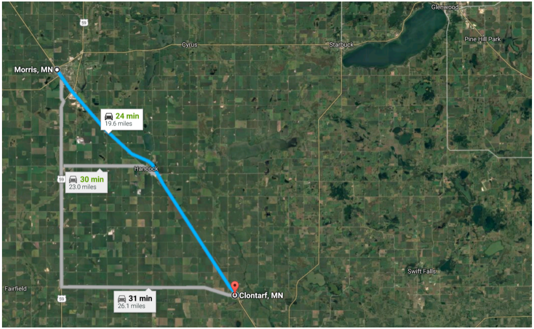
- Guo and Chen identify non-personal users
  - Spammers
  - Bots
  - Business accounts

- Guo and Chen identify non-personal users
  - Spammers
  - Bots
  - Business accounts
- Features:

Methods
Geo-Tagged Features

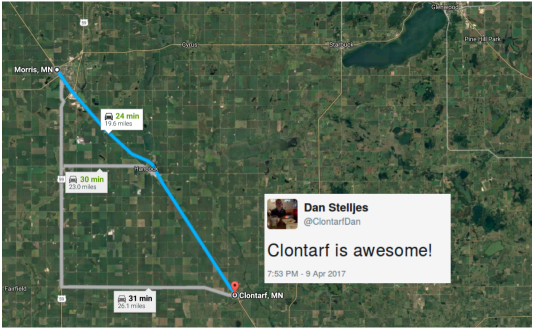
- Guo and Chen identify non-personal users
  - Spammers
  - Bots
  - Business accounts
- Features:
  - Tweeting Speed

Methods
Geo-Tagged Features



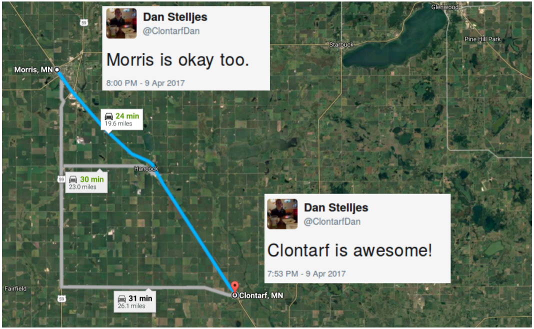
- 19.6 Miles

Methods
Geo-Tagged Features



- 19.6 Miles
- From Clontarf at 7:53 PM

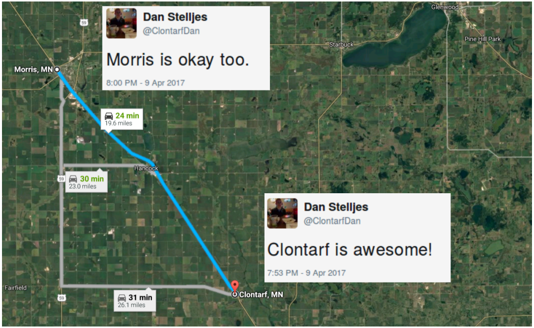
Methods
Geo-Tagged Features



- 19.6 Miles
- From Clontarf at 7:53 PM
- From Morris at 8:00 PM

Methods

Geo-Tagged Features



- 19.6 Miles
- From Clontarf at 7:53 PM
- From Morris at 8:00 PM
- Tweeting speed = $\frac{19.6 \text{ miles}}{7 \text{ minutes}}$ = 2.8 miles per minute (168 MPH)

- Features:

Methods
Geo-Tagged Features

- Features:
  - Max Speed

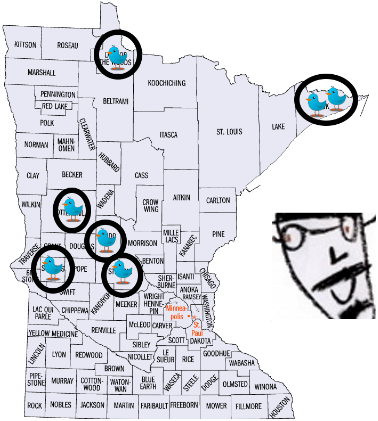
Methods
Geo-Tagged Features

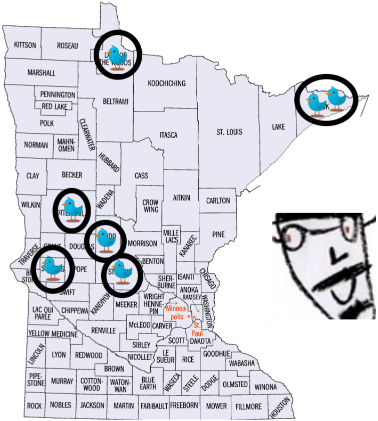
- Features:
  - Max Speed
  - Mean Speed

- Features:
  - Max Speed
  - Mean Speed
  - Max Distance (connected to Max Speed)

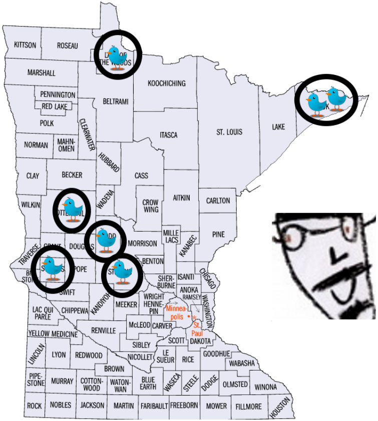
- Features:
    - Max Speed
    - Mean Speed
    - Max Distance (connected to Max Speed)
    - Mean number of times a user exceeds 90 MPH per month

- Features:
  - Max Speed
  - Mean Speed
  - Max Distance (connected to Max Speed)
  - Mean number of times a user exceeds 90 MPH per month
- County based features


KITTSON
ROSEAU
MARSHALL
KOOCHICHING
PENNINGTON
RED LAKE
BELTRAMI
POLK
CLEARWATER
ITASCA
ST. LOUIS
LAKE
NORMAN
MAHN-OMEN
HUBBARD
CLAY
BECKER
CASS
WILKIN
WADENA
CROW WING
AITKIN
CARLTON
MILLE LACS
PINE
MORRISON
KANABEC
TRAVERSE
BENTON
POPE
SWIFT
SHER-BURNE
ISANTI
CHISAGO
ANOKA
RAMSEY
WASHINGTON
WRIGHT
HENNE-PIN
Minnea-polis
St. Paul
LAC QUI PARLE
CHIPPEWA
MEEKER
McLEOD
CARVER
RENVILLE
YELLOW MEDICINE
SCOTT
DAKOTA
SIBLEY
LINCOLN
LYON
REDWOOD
NICOLLET
LE SUEUR
RICE
GOODHUE
WABASHA
BROWN
PIPE-STONE
MURRAY
COTTON-WOOD
WATON-WAN
BLUE EARTH
WASECA
STEELE
DODGE
OLMSTED
WINONA
ROCK
NOBLES
JACKSON
MARTIN
FARIBAULT
FREEBORN
MOWER
FILLMORE
HOUSTON



- Number of times a user crosses county borders per month



- Number of times a user crosses county borders per month
- Mean number of counties a user has been to per month

Outline

Methods
Time Features

- Washha et al. classify spammers on a user level

Methods
Time Features

- Washha et al. classify spammers on a user level
- Motivated to use time since altering time dependent features is a challenge.

- Washha et al. classify spammers on a user level
- Motivated to use time since altering time dependent features is a challenge.
- Features that spammers can easily manipulate:

- Washha et al. classify spammers on a user level
- Motivated to use time since altering time dependent features is a challenge.
- Features that spammers can easily manipulate:
  - Number of URLs

Methods
Time Features

- Washha et al. classify spammers on a user level
- Motivated to use time since altering time dependent features is a challenge.
- Features that spammers can easily manipulate:
  - Number of URLs
  - Number of Hashtags

- Washha et al. classify spammers on a user level
- Motivated to use time since altering time dependent features is a challenge.
- Features that spammers can easily manipulate:
  - Number of URLs
  - Number of Hashtags
  - Including Geo-tags

Features

- Differences in Account Age

Features

- Differences in Account Age
  - Spammers have multiple accounts

Features

- Differences in Account Age
  - Spammers have multiple accounts
  - Likely to be made at the same time

Features

- Differences in Account Age
  - Spammers have multiple accounts
  - Likely to be made at the same time
  - Followers

Features

- Differences in Account Age
    - Spammers have multiple accounts
    - Likely to be made at the same time
    - Followers
    - Followees

Methods
Time Features

Features

- Differences in Account Age
    - Spammers have multiple accounts
    - Likely to be made at the same time
    - Followers
    - Followees
    - Bi-directional relationships

Features

- Differences in Account Age
  - Spammers have multiple accounts
  - Likely to be made at the same time
  - Followers
  - Followees
  - Bi-directional relationships
- Time weighted correlations:

Features

- Differences in Account Age
  - Spammers have multiple accounts
  - Likely to be made at the same time
  - Followers
  - Followees
  - Bi-directional relationships
- Time weighted correlations:
  - URLs

Features

- Differences in Account Age
  - Spammers have multiple accounts
  - Likely to be made at the same time
  - Followers
  - Followees
  - Bi-directional relationships
- Time weighted correlations:
  - URLs
  - Mentions

Features

- Differences in Account Age
  - Spammers have multiple accounts
  - Likely to be made at the same time
  - Followers
  - Followees
  - Bi-directional relationships
- Time weighted correlations:
  - URLs
  - Mentions
  - Hashtags

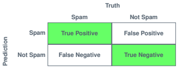
Features

- Differences in Account Age
  - Spammers have multiple accounts
  - Likely to be made at the same time
  - Followers
  - Followees
  - Bi-directional relationships
- Time weighted correlations:
  - URLs
  - Mentions
  - Hashtags
- Tweet similarity weighted by time

Outline

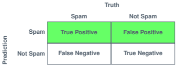
Results

- Accuracy:



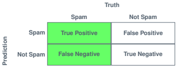
$\frac{TP+TN}{TP+FP+TN+FN}$ "How many tweets were correctly identified?"

- Precision (p):



$\frac{TP}{TP+FP}$ "How good is our spam prediction?"

- Recall (r):



$\frac{TP}{TP+FN}$ "How much spam was identified?"

- F-measure (F): $2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$

Model Results of the Three Studies

Study	% Spam	p	r	F	Accuracy
User/Tweet Features: I	50.0%	0.929	0.943	0.936	0.936
User/Tweet Features: II	5.0%	0.929	0.407	0.566	0.978
Geo-tagged Features	21.4%	0.959	0.959	0.958	0.959
Time Features	46.9%	0.932	0.931	0.931	0.931

Conclusion

- Classification via random forest
- Recall (r) may drop when test set contains a low proportion of spam
  - Future work: Apply this finding to geo-tagged tweets and time features
- Future spam classification by Twitter: Random forests?

Acknowledgments

- Peter Dolan
  - For acting as my advisor for this research project and his continued friendship for the past few years
- Elena Machkasova
  - For the invaluable advice and insightful comments throughout the entire senior seminar course
- Jacob Opdahl
  - For the exceptional revisions and comments he made as my alumni review

References I

- Fabricio Benevenuto, Gabriel Magno, Tiago Rodrigues, and Virgilio Almeida, *Detecting spammers on twitter*, Collaboration, electronic messaging, anti-abuse and spam conference (CEAS), vol. 6, 2010, p. 12.
- Leo Breiman, *Random forests*, Machine learning **45** (2001), no. 1, 5–32.
- Chao Chen, Jun Zhang, Xiao Chen, Yang Xiang, and Wanlei Zhou, *6 million spam tweets: A large ground truth for timely twitter spam detection*, 2015 IEEE International Conference on Communications (ICC), IEEE, 2015, pp. 7065–7070.
- Diansheng Guo and Chao Chen, *Detecting non-personal and spam users on geo-tagged twitter network*, Transactions in GIS **18** (2014), no. 3, 370–384.

References II

- Miroslav Kubat, *An introduction to machine learning*, 1st ed., Springer Publishing Company, Incorporated, 2015.
- Nikita Spirin and Jiawei Han, *Survey on web spam detection: Principles and algorithms*, SIGKDD Explor. Newsl. **13** (2012), no. 2, 50–64.
- Claude Elwood Shannon, *A mathematical theory of communication*, ACM SIGMOBILE Mobile Computing and Communications Review **5** (2001), no. 1, 3–55.
- Igor Santos, Igor Miñambres-Marcos, Carlos Laorden, Patxi Galán-García, Aitor Santamaría-Ibirika, and Pablo García Bringas, *Twitter content-based spam filtering*, pp. 449–458, Springer International Publishing, Cham, 2014.

References III

Kurt Thomas, Chris Grier, Dawn Song, and Vern Paxson, *Suspended accounts in retrospect: An analysis of twitter spam*, Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference (New York, NY, USA), IMC '11, ACM, 2011, pp. 243–258.

Mahdi Washha, Aziz Qaroush, and Florence Sedes, *Leveraging time for spammers detection on twitter*, Proceedings of the 8th International Conference on Management of Digital EcoSystems (New York, NY, USA), MEDES, ACM, 2016, pp. 109–116.