Detecting Spammers on Twitter

Fabrício Benevenuto

(joint work with Gabriel Magno, Tiago Rodrigues, and Virgílio Almeida)

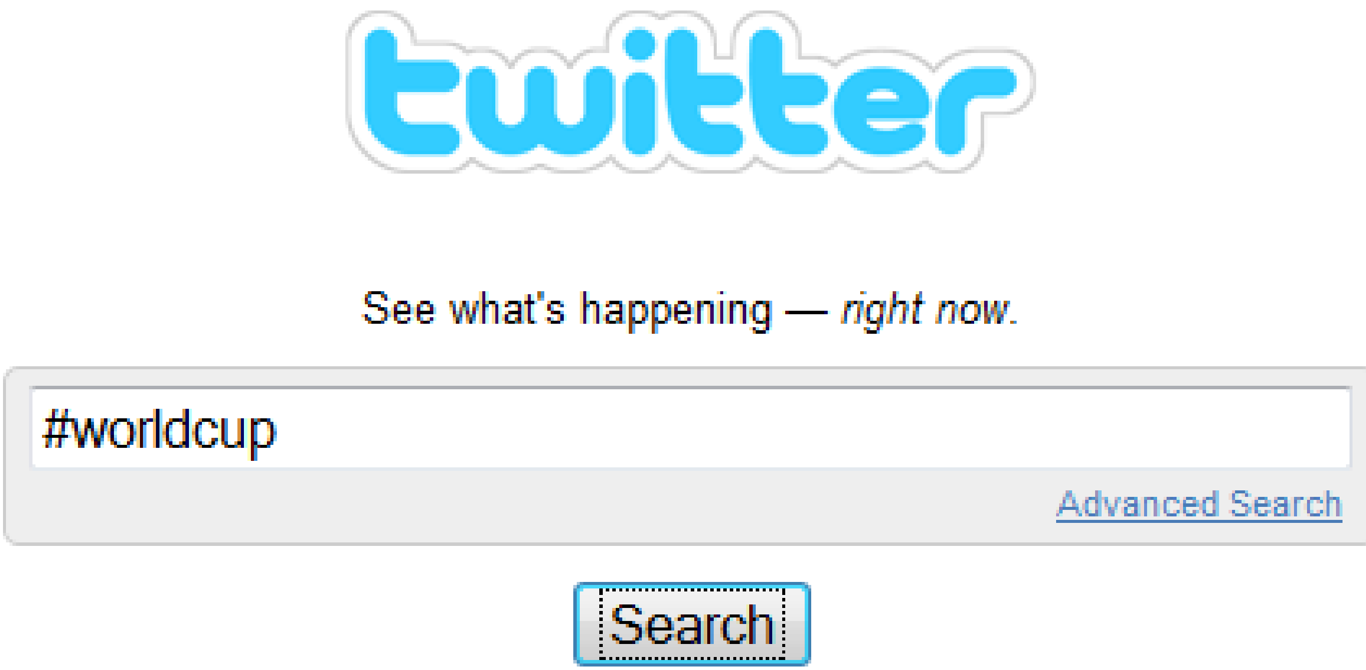
Federal University of Minas Gerais (UFMG) - Brazil

New forms of Spam

- Online social networks
  - Very popular
  - Easy to create content

- New forms of spam
  - Unsolicited invitations on Facebook
  - Vandalism on Wikipedia
  - Video spam on YouTube [SIGIR'09]
  - Spam on Twitter

Spam on Twitter



Spam on Twitter

Users post URLs unrelated to content

Negative impact of spam

- Pollute real time search
- Interfere on mining tools and statistics about trendings and events on Twitter
- Consume user and system resources
- Waste human attention

Goal and methodology

- Goal: Detect spammers on Twitter

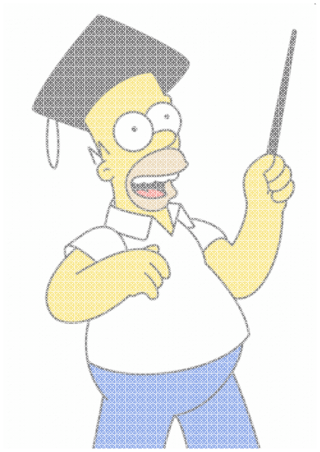
- 4-step approach

  1. Collect data from Twitter

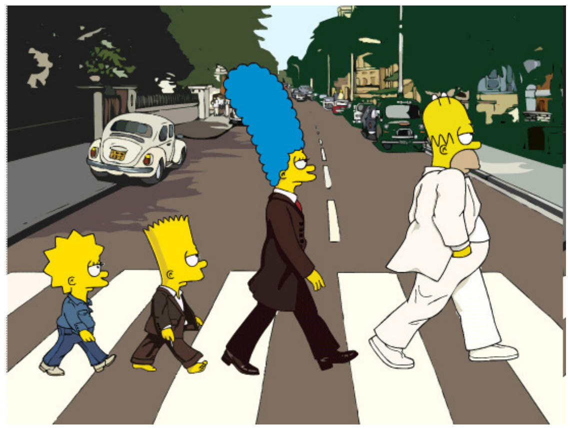
  2. Manually create a collection of users labeled as spammers or non-spammers

  3. Identify attributes able to distinguish spammers from non-spammers users

  4. Classification approach to detect spammers



Part1.
Motivation
& Problem



Part2.
4-step
approach

Part3.
Experimental
results

Step1. Collecting Twitter

- Crawls subject to rate-limiting
  - Twitter provide us a white list for 58 machines at MPI-SWS
  - Rate limit = 20,000 requests/hour.

- Inspected all user IDs collecting
  - User profile information
  - List of followers and followees
  - All tweets they posted

- In total we collected **54,981,152** users, **1,963,263,821** unique links, and**1,755,925,520** tweets

Step 2. Labeled collection

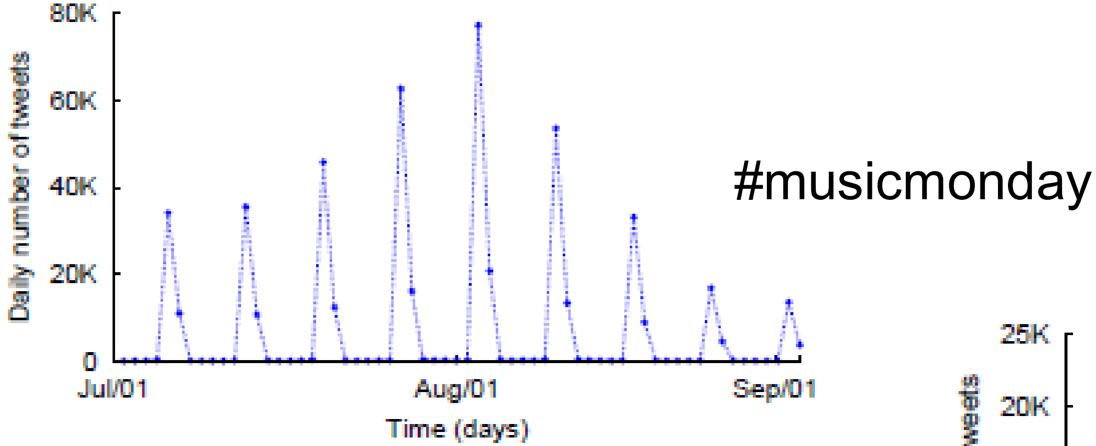
- Desired properties

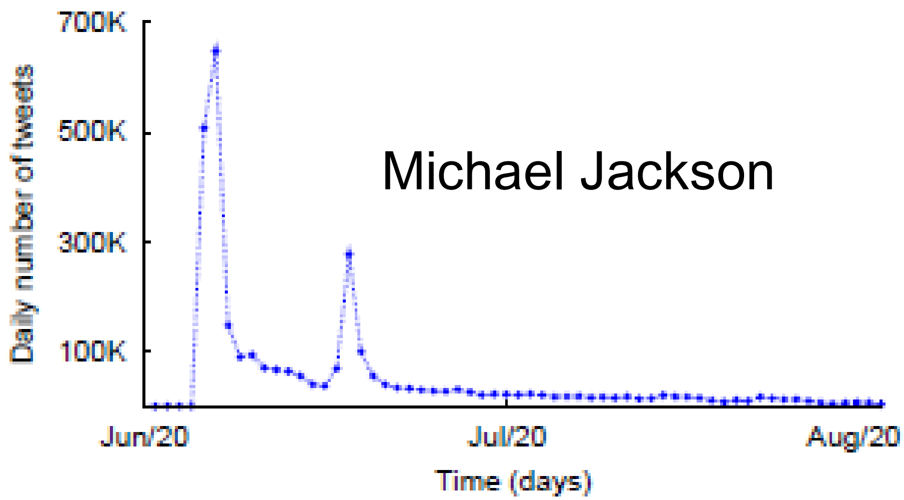
  1) Have a significant number of spammers and non-spammers

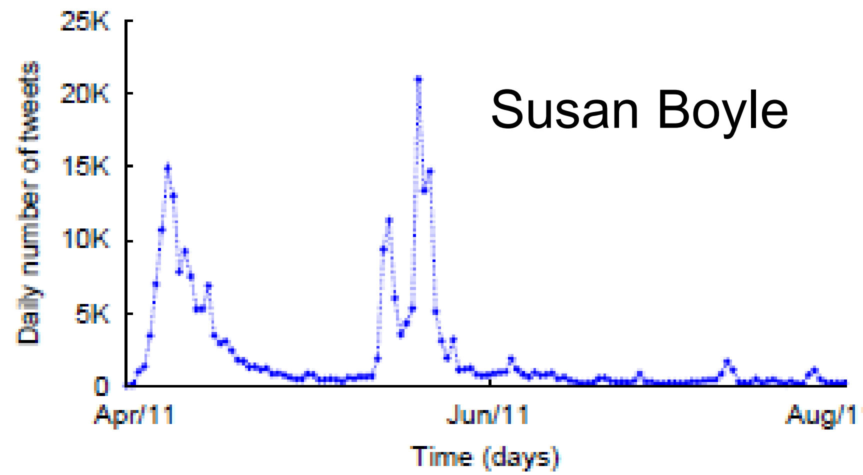
  2) Include spammers who are aggressive in their strategies

  3) Choose users randomly and not based on their characteristics

Step2. Labeled Collection







Focus on popular events of 2009

Step2. Labeled Collection

- Volunteers analyze tweets of randomly selected users that post to the three trending topics analyzed
  - Development of a Web system to ease the process
  - Each user is analyzed by at least two volunteers
  - Agreement in 99% of cases
- 8,207 users were analyzed out of which 355 are spammers

Labeled collection:
355 spammers + 710 non-spammers = 1,065 users

Step3. Attributes

- **User behavior** (total = 23)
  - number of followers
  - number of followees
  - number of tweets
  - age of the user account
  - etc.

- **Tweet content** (total = 39)
  - Fraction of tweets with spam words
  - Fraction of tweets with URL
  - etc.

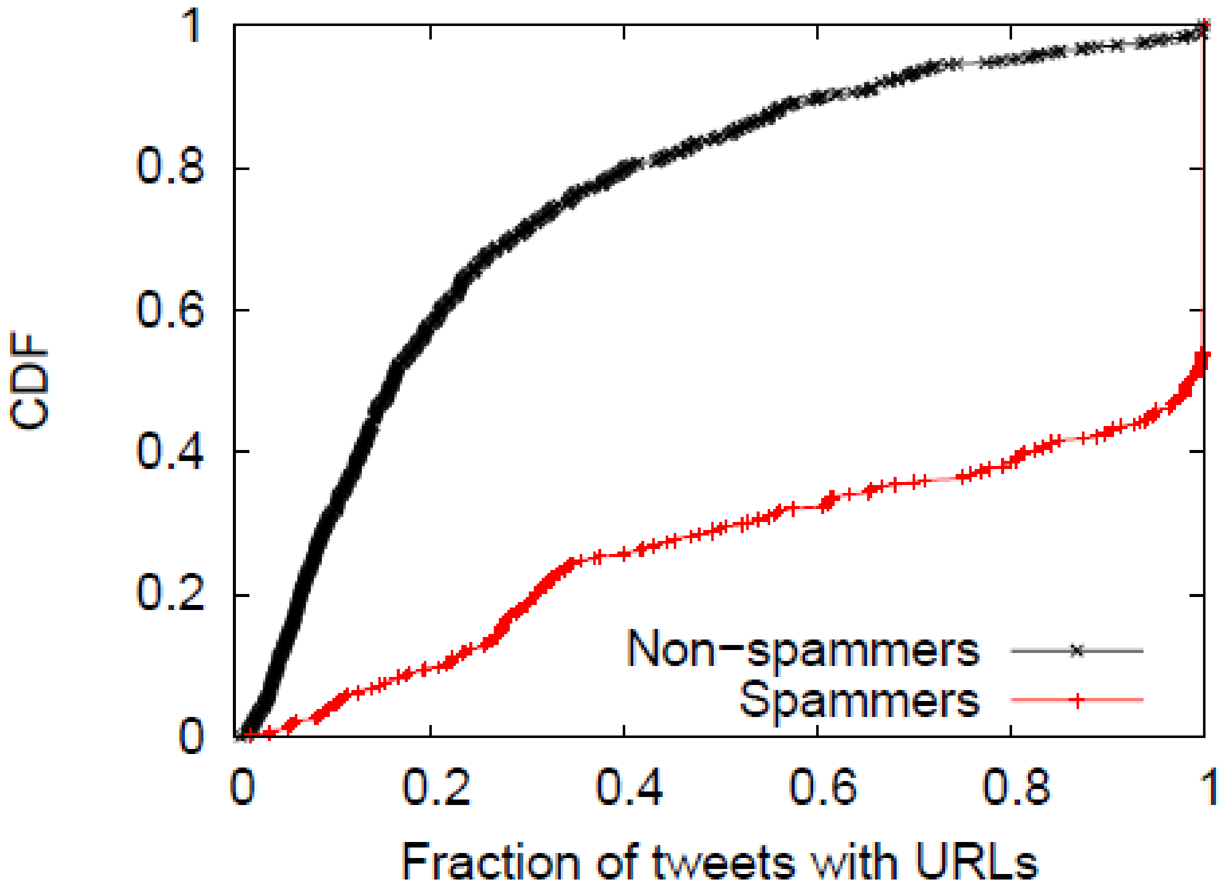
Importance of the attributes

Position	χ^2 ranking
1	fraction of tweets with URLs
2	age of the user account
3	average number of URLs per tweet
4	fraction of followers per followees
5	fraction of tweets the user had replied
6	number of tweets the user replied
7	number of tweets the user receive a reply
8	number of followees
9	number of followers
10	average number of hashtags per tweet

Attributes from content and user behavior are both important

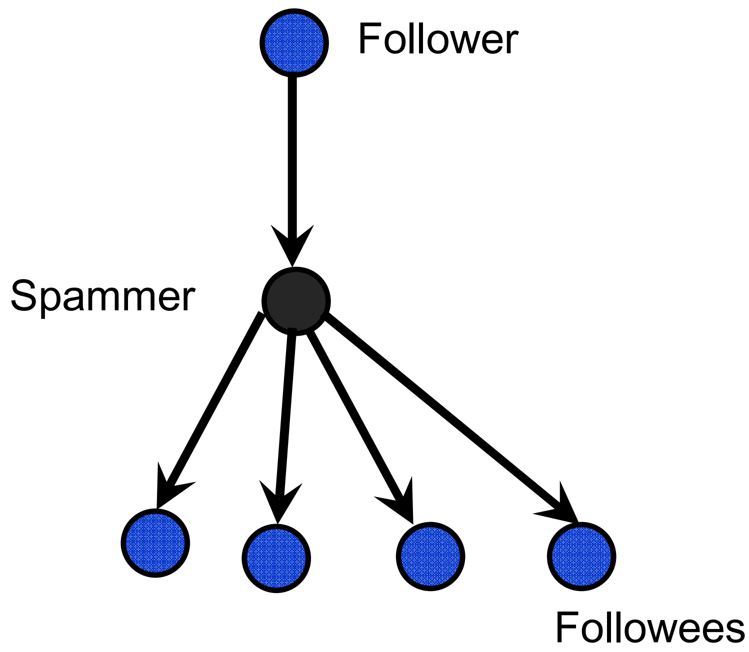
	Tweet content	User behavior
Top 10	4	6
Top 20	10	10
Top 30	17	13
Top 40	23	17
Top 50	31	19
Top 62	39	23

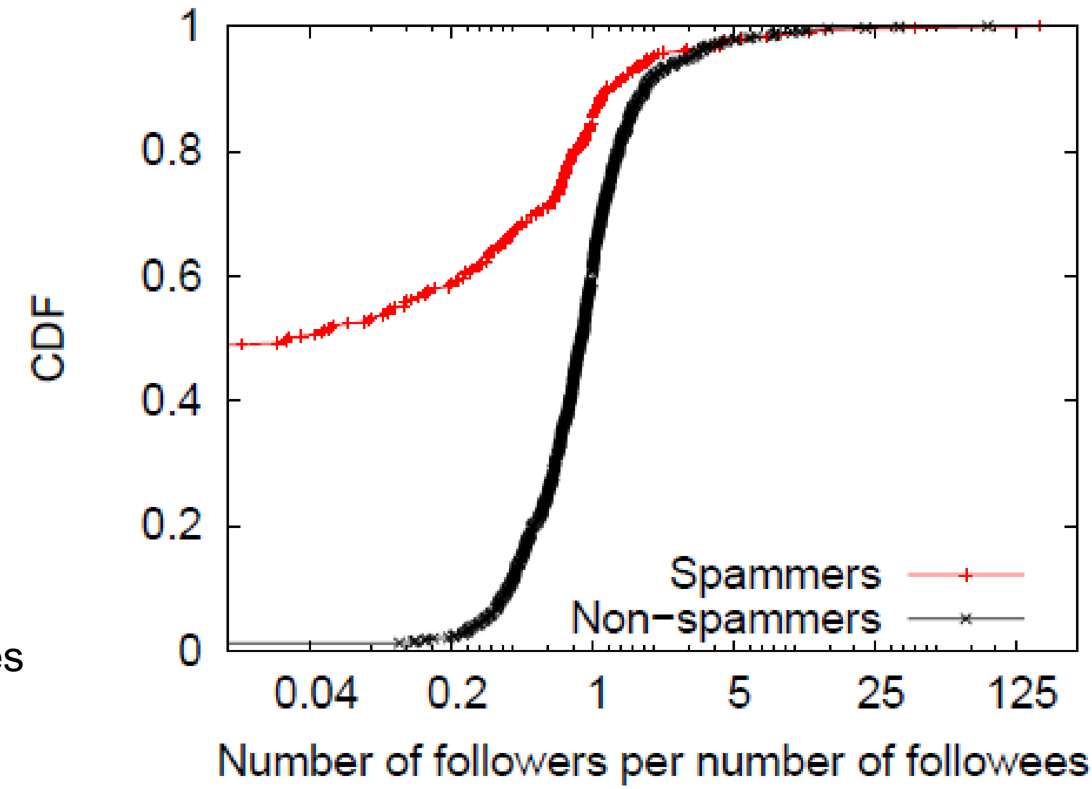
Distinguishing classes of users (1)



Spammers post most of the tweets containing URLs

Distinguishing classes of users (2)



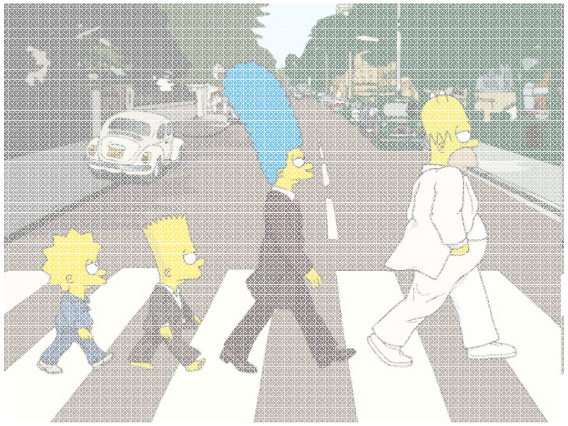


Spammers have less followers than followees

Step4. Classification approach

- SVM (Support vector machine) as classifier
- Use all attributes
- 5-fold cross validation

Part1.
Motivation
& Problem



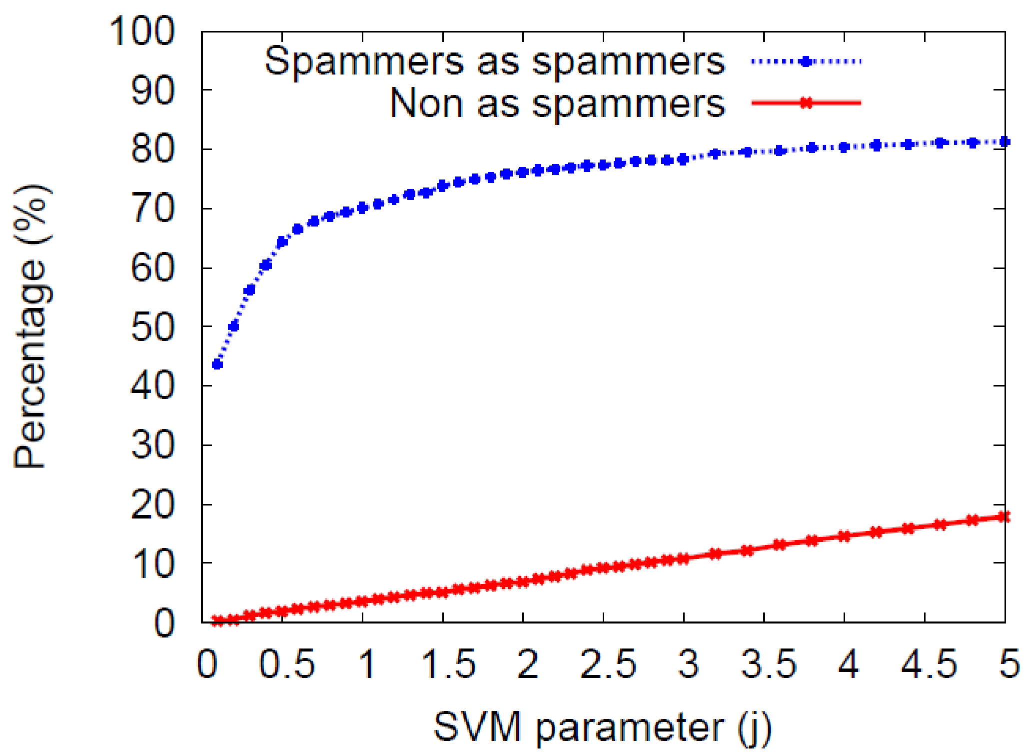
Part2.
4-step
approach

**Part3.
Experimental
results**

Classification results

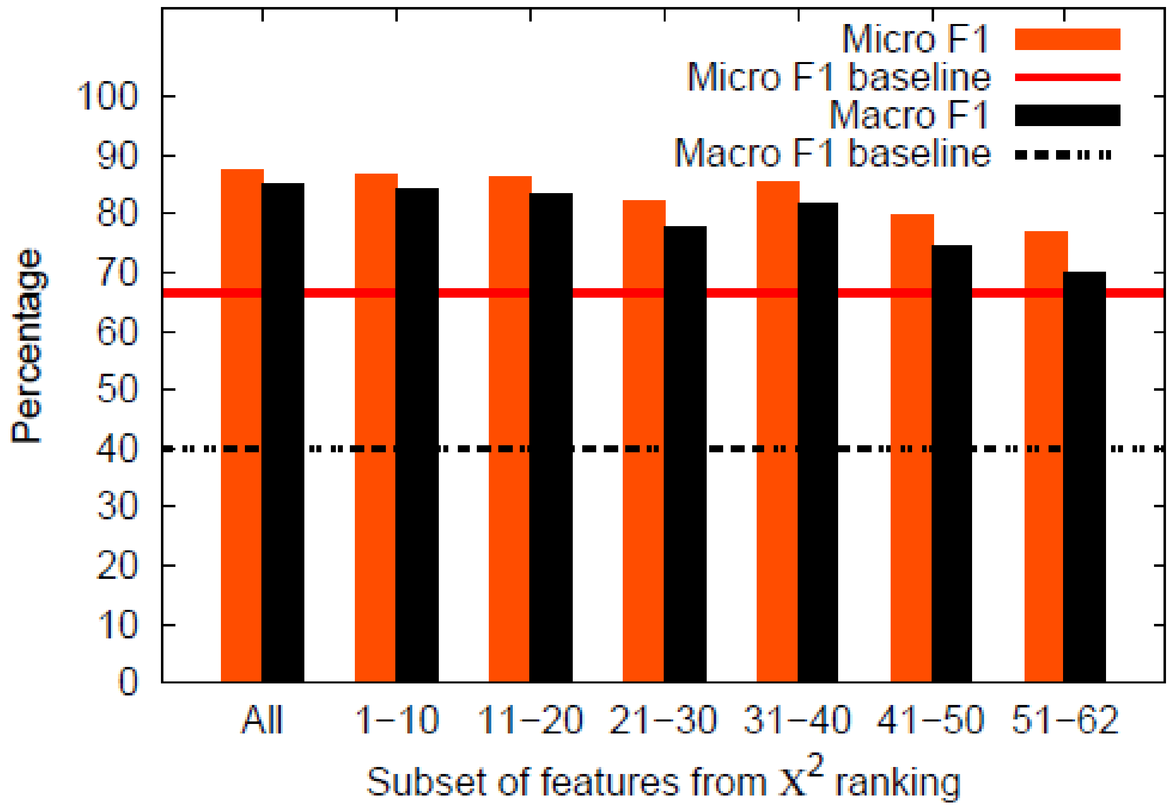
		Predicted	
		Spammer	Non-spammers
True	Spammer	70.1%	29.9%
	Non-spammer	3.6%	96.4%

88% of accuracy



- **J = 0.1:** correctly classify 44% spammers, misclassifying <0.3% non-spammers
- **J = 3:** correctly classify 81% spammers, paying the cost of misclassifying 18% non-spammers

Reducing the attribute set



Different subsets of features can obtain competitive results

Detecting tweets instead of users

		Predicted	
		Spam	Non-spam
True	Spam	**78.5%**	21.5%
	Non-spam	92.5%	**7.5%**

84.5% of accuracy

Good results, but tweet content is easy to be faked

We can still obtain good results classifying users even if we disregard content attributes

		Predicted	
		Spammer	Non-spammers
True	Spammer	**69.7%**	30.3%
	Non-spammer	4.3%	**95.7%**

Conclusions

- We propose a mechanism to detect spammers on Twitter
  - Twitter dataset and labeled collection
    - Publicly available (soon) at www.dcc.ufmg.br/~fabricio
  - Attribute identification
  - Classification approach
    - Correctly identified majority of spammers
    - Different subsets of features can obtain competitive results
    - Detection of spam also works, but attributes are easier to be faked

Questions?

fabricio@dcc.ufmg.br

http://www.dcc.ufmg.br/~fabricio