Spam Detection by Hashtag Relationships

CS 533 Final Presentation

Yalım Doğan, M.Enes Karaca, Soner Koç

29 April 2016

Presentation Outline

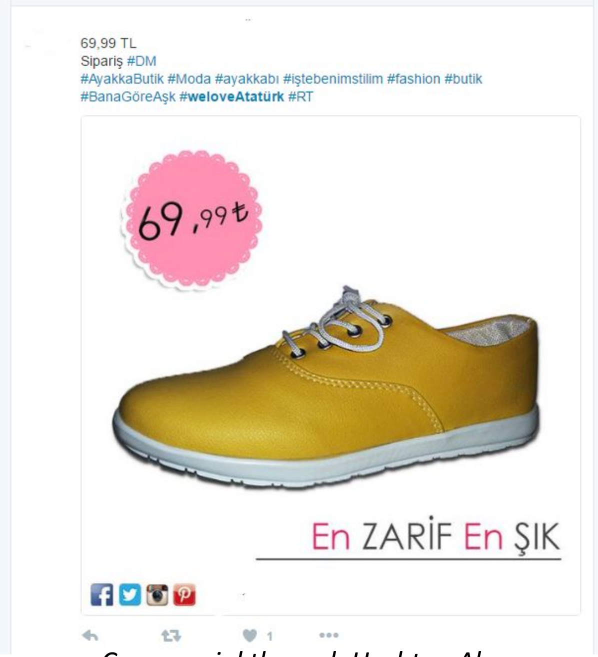
- Problem Description
- Motivation
- Methodology
- Results
- Future Works
- References

Problem Description

- The popularity of Twitter attracts spammers for advertising, propaganda, adult content etc.
- The tweets are considered as spam for following reasons:
  - Containing more than specific number of hashtags (can be considered Hashtag Abuse) and some of them are unrelated.
  - Containing links generated by URL shorteners. Considered as a strong indication.
  - Includes words(key stops) that are considered as spam (advertisement and adult content related)
  - Posted by different(or same) users in same time period (in the same minute at shortest) which are not by Retweeting.

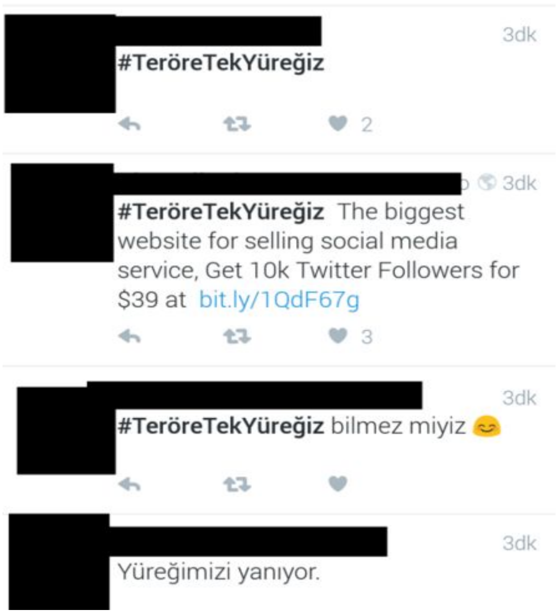
Problem Description

Examples



69,99 TL
Sipariş #DM
#AyakkaButik #Moda #ayakkabı #iştebenimstilim #fashion #butik
#BanaGöreAşk #weloveAtatürk #RT

Commercial through Hashtag Abuse



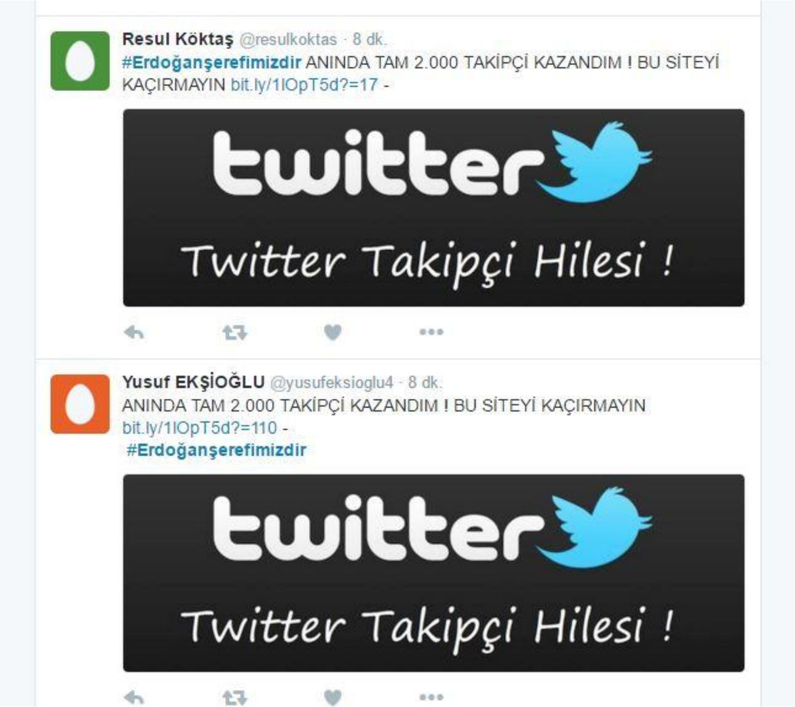
[redacted] 3dk
#TeröreTekYüreğiz
2

[redacted] 3dk
#TeröreTekYüreğiz The biggest website for selling social media service, Get 10k Twitter Followers for $39 at bit.ly/1QdF67g
3

[redacted] 3dk
#TeröreTekYüreğiz bilmez miyiz 😊

[redacted] 3dk
Yüreğimizi yanıyor.

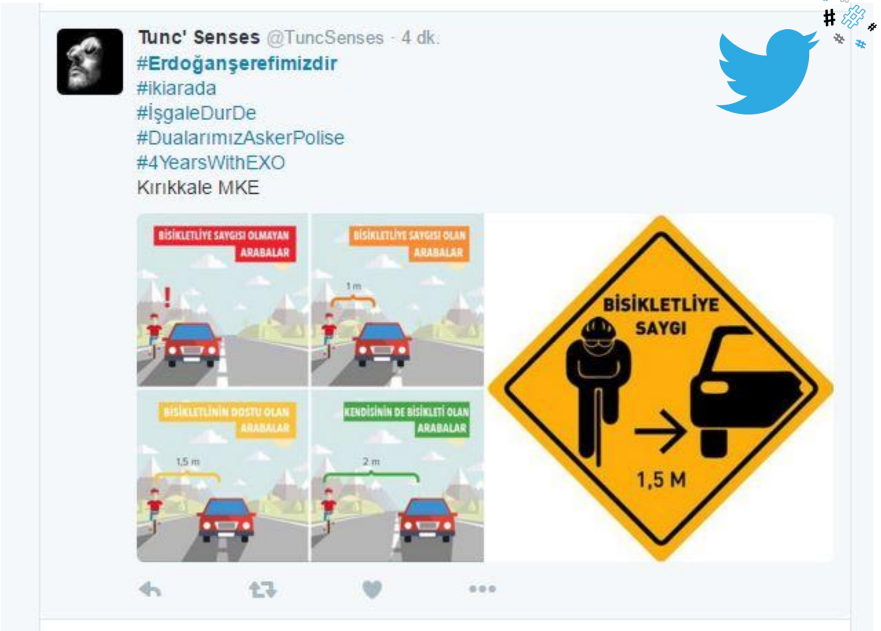
A Tweet Containing Spam Related Words Together with shortened URL



Resul Köktaş @resulkoktas · 8 dk.
#Erdoğanşerefimizdir ANINDA TAM 2.000 TAKİPÇİ KAZANDIM ! BU SİTEYİ KAÇIRMAYIN bit.ly/1lOpT5d?=17 -

Yusuf EKŞİOĞLU @yusufeksioglu4 · 8 dk.
ANINDA TAM 2.000 TAKİPÇİ KAZANDIM ! BU SİTEYİ KAÇIRMAYIN bit.ly/1lOpT5d?=110 -
#Erdoğanşerefimizdir

Attempt for gaining more followers



Tunc' Senses @TuncSenses · 4 dk.
#Erdoğanşerefimizdir
#ikiarada
#İşgaleDurDe
#DualarımızAskerPolise
#4YearsWithEXO
Kırıkkale MKE

BİSİKLETLİYE SAYGISI OLMAYAN ARABALAR
BİSİKLETLİYE SAYGISI OLAN ARABALAR
1 m
BİSİKLETLİNİN DOSTU OLAN ARABALAR
1,5 m
KENDİSİNİN DE BİSİKLETİ OLAN ARABALAR
2 m
BİSİKLETLİYE SAYGI
1,5 M

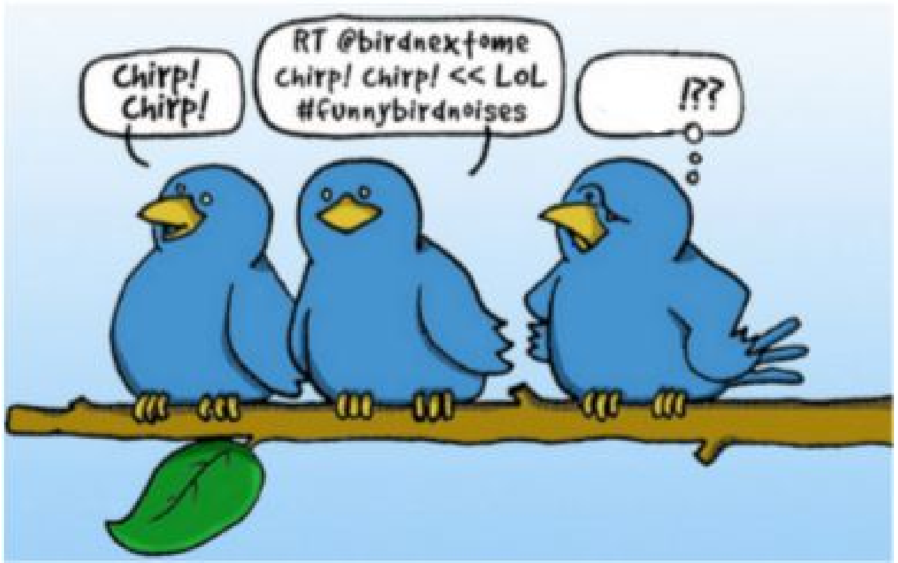
Hashtag Abuse

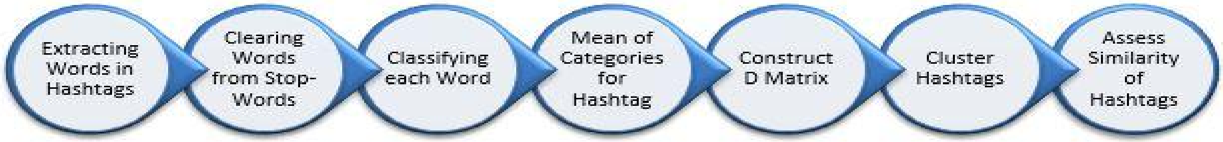
Motivation

- There are several methods proposed in order to fight spamming which consider the common elements of spam tweets.
  - URL's posted in tweets [1]
  - Number of Hashtags (**BUT** doesn't consider the relation between them)
  - If the tweets are duplicate (Not RT)
  - Usage of spam words [2]
- We propose a system that checks the relationship between hashtags in tweets, in addition to the above parameters.
- In this way, we are aiming to improve the accuracy of spam detection on Twitter.

Methods

- Tweet Link Status
- Spam and Non-Spam Words
- Number of Hashtags
- Duplicate Tweets
- Hashtag Abuse Detection
  - Relation Between Hashtags
    - Classification of Hashtags
    - Clustering of Hashtags




Extracting Words in Hashtags → Clearing Words from Stop-Words → Classifying each Word → Mean of Categories for Hashtag → Construct D Matrix → Cluster Hashtags → Assess Similarity of Hashtags

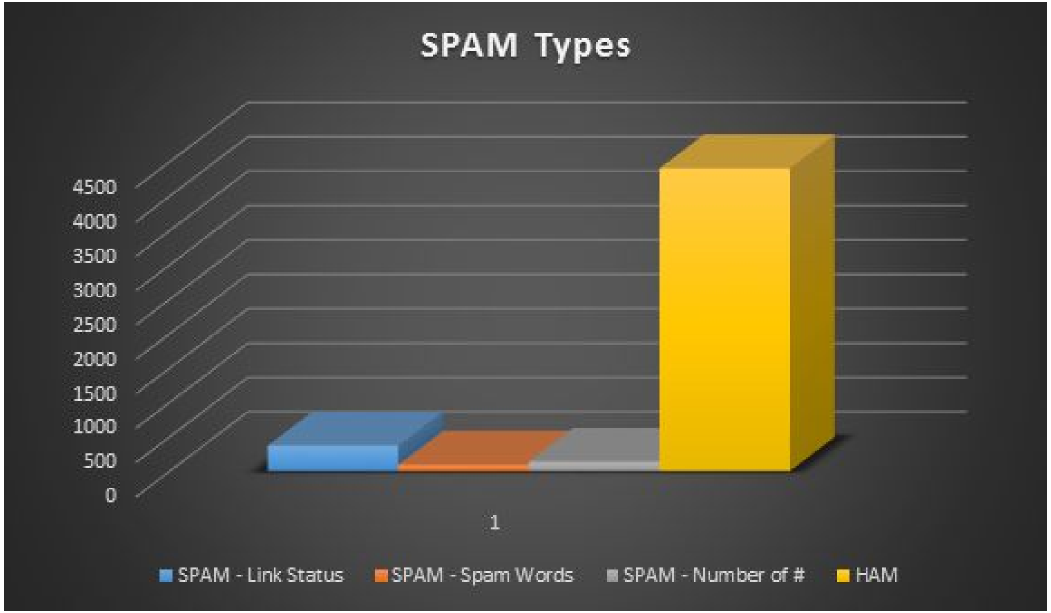
Methods

Tweet Link Status

- 371 tweets contain bad links
- 4629 tweets have no bad links

Bad links can be considered as advertisement such as: Amazon, eBay.

371 out of 5000 tweets contain bad links. (%7.42)



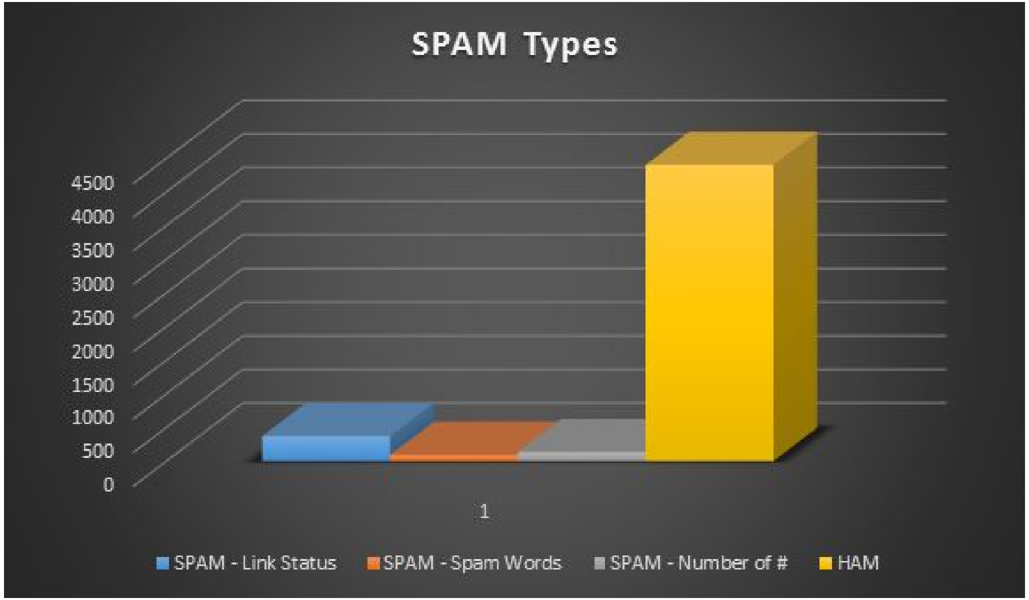
Methods

Spam and Non-Spam Words

- List of 278 spam words
- 92 spam tweets
- 4537 non-spam tweets

Spam words such as:

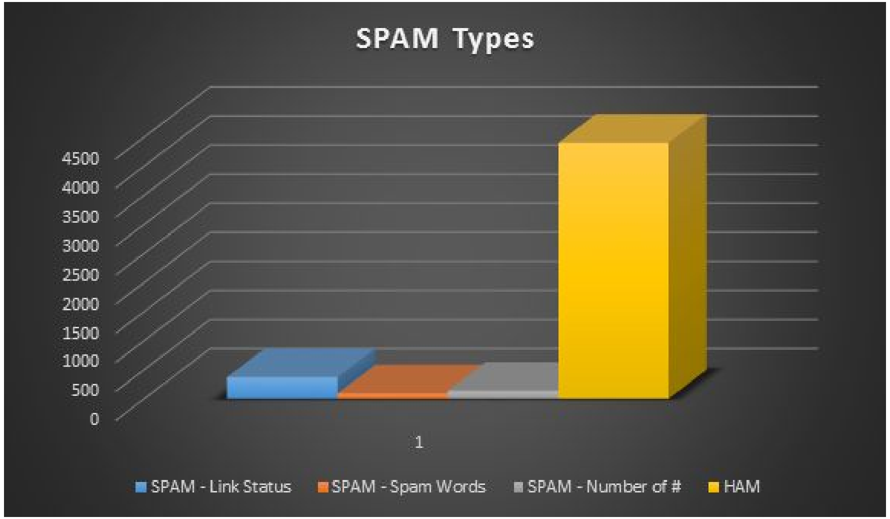
Percentage is: %1.98

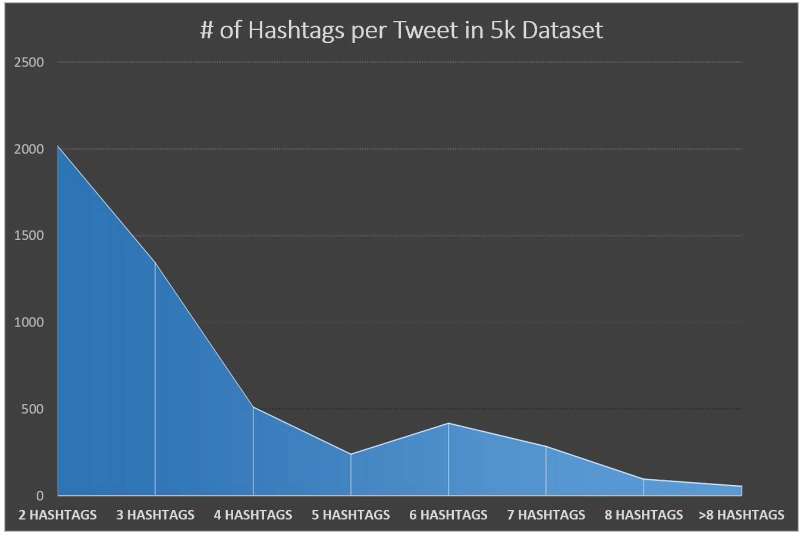


Methods

Number of Hashtags

- 136 tweets contain more than 7 hashtags (%2.99)
- 4401 tweets contain less than 7 & more than 2 hashtags





- #Selenators #TaylorSwift #encuesta #DemocracyIn5Words #like4like #likeforlike #twitter #retwitthis #BirdieSanders
- #ioho20anni #DemocracyIn5Words #BatmanvSuperman #BirdieSanders #ENGvSL #Amici15 #Elite8 #saturdaykitchen #Pasqua#10YearsOfAmazingPhil
- #BernieSanders #askjack #GERENG #WAcaucus #BirdieSanders #DemocracyIn5Words #BatmanvSuperman #Elite8 #FinalFour
- #GERENG #WAcaucus #DemocracyIn5Words #BatmanvSuperman #BirdieSanders #DubaiWorldCup #GoodFriday #AIRMAXDAY #doac
- #Love #GERENG #WAcaucus #askjack #DemocracyIn5Words #BirdieSanders #BatmanvSuperman #DCRebirth #Elite8
- Cual creen q es mejor? #encuesta #twitter #retwitthis #like #DemocracyIn5Words #BatmanvSuperman #like4like #likes(Que prefieren?
- #twitter #instagram #likebackteam #likealways #encuesta #DemocracyIn5Words #BatmanvSuperman #Elite8
- #askjack #GERENG #DemocracyIn5Words #BatmanvSuperman #indie #singer #atlanticrecords #bmth #bvb #cte #om&amp;m #bands

Methods

Hashtag Abuse Detection

Two Categories
Name “Donald”
Belongs

- Relation Between Hashtags
  - Classification of Hashtags

WEKA
The University
of Waikato

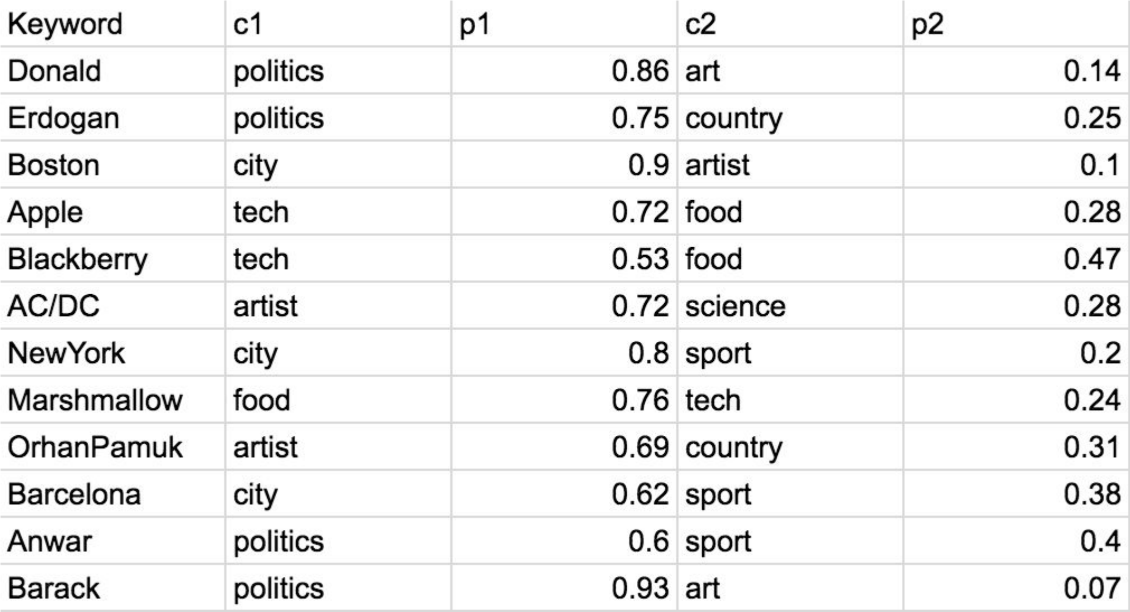
    - SVM algorithm is used with weighted indexes (JAVA WEKA Tool)
    - Hashtags are classified according to a prepared dictionary
    - Each hashtag is weighted for each category in dictionary
    - Dictionary of categories is updated with popular hashtags regularly
    - Determination of the mean weight of hashtag for each category

Two Categories
Name “Lincoln”
Belongs

Methods

Hashtag Abuse Detection

- Generation of dataset of Weights on Hashtags
  - As an illustration, “Apple” has weight of 0.72 on category “tech” and weight of 0.28 on category “food”



Keyword	c1	p1	c2	p2
Donald	politics	0.86	art	0.14
Erdogan	politics	0.75	country	0.25
Boston	city	0.9	artist	0.1
Apple	tech	0.72	food	0.28
Blackberry	tech	0.53	food	0.47
AC/DC	artist	0.72	science	0.28
NewYork	city	0.8	sport	0.2
Marshmallow	food	0.76	tech	0.24
OrhanPamuk	artist	0.69	country	0.31
Barcelona	city	0.62	sport	0.38
Anwar	politics	0.6	sport	0.4
Barack	politics	0.93	art	0.07

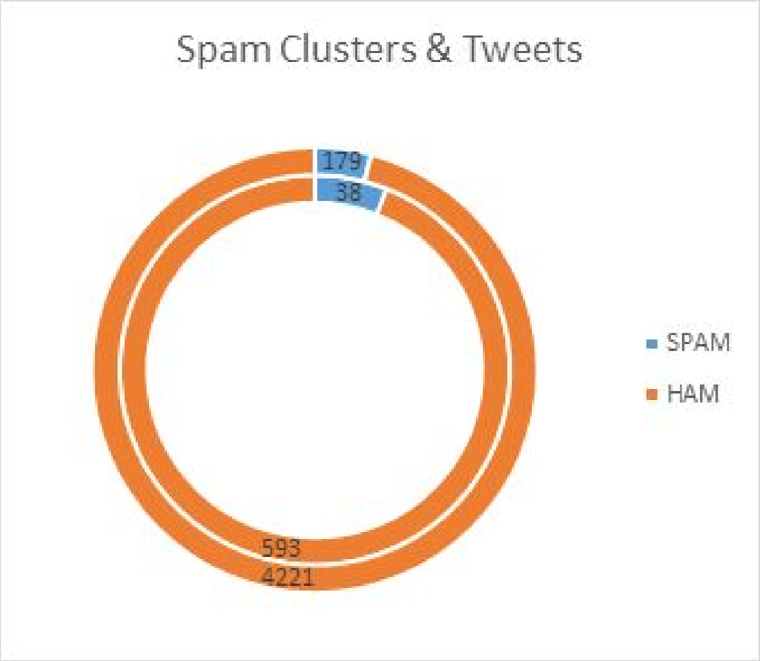
Methods

Hashtag Abuse Detection

- Clustering of Hashtags has been done by
  - Selecting a random tweet from duplicate cluster (prepared beforehand)
    - Duplicate cluster contains tweet that have more than 0.6 similarity
    - For efficiency, only one element is enough as most of them are identical
  - Extract hashtags’ terms used. Remove stop words. ~~Apply stemming~~
    - Stemming created problems so it is omitted
  - Using weights of each term as category, create a D matrix as
    - Each hashtag is a document, each term is a CATEGORY weight (sum = 1)
  - Calculate number of clusters. Define a threshold for it.
- What is the threshold ?
  - I have tried (#of hashtags/2) and nc = 1.

Results

- It appears that because SVM is not trained enough to cover too diverse categories yet, and limit of nc is high: there were very small amount of spam detected.
- What happened with nc > 1 is SPAM ?
    - Improvement !
- Precision is about %88 which indicates system have few amount of FP.
- Recall is about %48, half of the spam clusters are not found.
- Need to consider improving SVM (or selecting another approach) to recognize much more terms and obtain better identification of spam tweets.
- **Maximum tweets in a cluster is 450, minimum is 1.**
- **There are SPAM 38 cluster (179 tweets) , out of 631 clusters (4400 tweets)**



Max cluster: RT @JimKilbane: #DemocracyIn5Words hereIT Goes (#VoteOutGOP at will) because we need vote out the 1% #vote

New Future Works

Throughout our research, we found out new kind of approaches for these kind of problems which will be brand new approach for Spam related Twitter research topics

- Approximation algorithms, Count-Min Sketch etc. for increasing accuracy with tremendous data
- Deep Learning algorithms with online learning

References

- [1] F. Benevenuto, G. Magno, T. Rodrigues, V. Almeida, “Detecting Spammers on Twitter”, Electronic messaging, Anti-Abuse and Spam Conference (CEAS), 2010.
- [2] A Collection of 14 Million Tweets for HashtagOriented Spam Research. In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '15).
- [3] Wang, A.H.: Detecting spam bots in online social networking sites: A machine learning approach. In: Foresti, S., Jajodia, S. (eds.) Data And Applications Security and Privacy XXIV.LNCS, vol. 6166, pp. 335-342. Springer, Heidelberg (2010)

#Thank#You#For#Your#Listening