Identifying spam in Social Listening data

Project developed as a part of Data Science Fundamentals SCC 460

Team #7

Oleksa Stepaniuk

Adi Karri

Mansoor Muneer Reehana

Chia-Yen Chiang

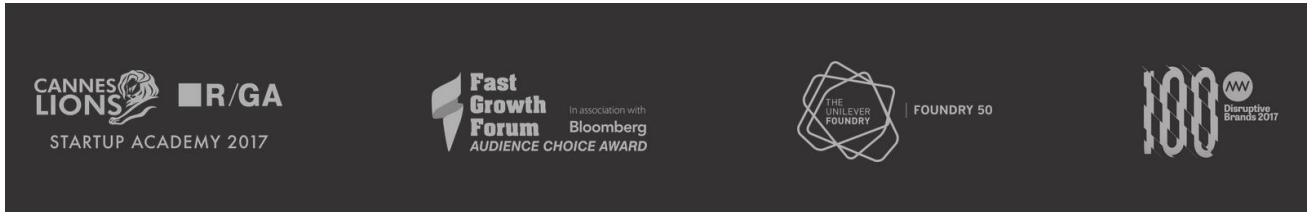
Leke Omogbeja

Mathematics & Statistics | Lancaster University

Motivation. Company

Relative INSIGHT

- Founded in 2012
- Offices in Lancaster and London
- Based on research conducted in Lancaster University for crime detection
- Business model: understand how people talk and write about different topics and provide this information to product makers and marketing agencies



Motivation. Problem

- Data is provided by the client but also web-scrapping is used
- Source of data – forums and social media posts collected by social intelligence companies such as Brandwatch and Crimson Hexagon
- Problem: client pays for each observation that has a keyword in it. As a result providers are not interested in cleaning data
- Social media information is especially problematic
- Analysis based on the data with spam can provide distorted results

Motivation. Why Twitter

- Instagram has almost no text in messages
- Facebook provides a limited access to user posts after the recent sequence of scandals (for example Cambridge Analytica). Only last the two months can be accessed.

Objective and Research question

Objective: identify spam in the Twitter data.

Spam in the context of study:

1. Irrelevant tweets
2. Spam or abusive tweets as defined by Twitter
3. Tweets of companies/professionals (e.g. bloggers, etc.)

Research question: how to identify tweets from each of the three groups

Division of labor

Oleksa: merging files, processing data, processing text, writing a web scrapper

Adi: performing web-scrapping, exploratory analysis, logistic regression

Mansoor: literature review, business requirement gathering

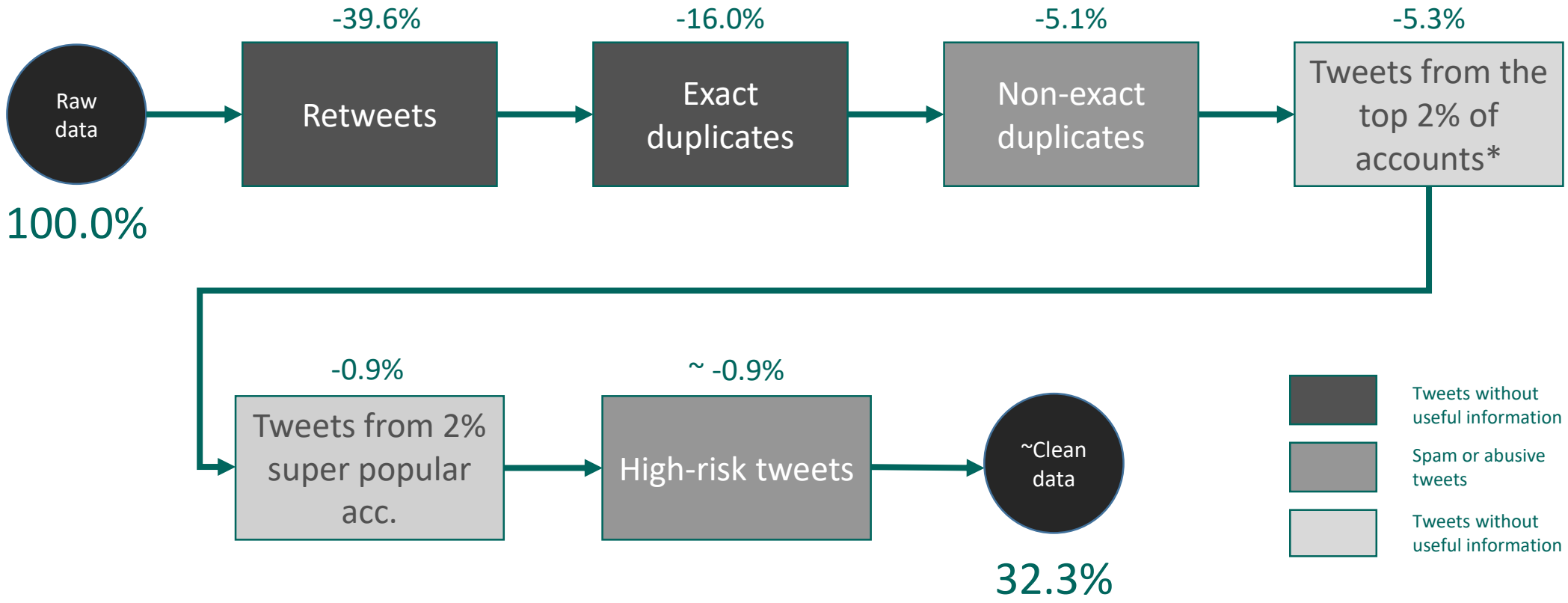
Chia-Yen: processing text research

Leke: writing report Introduction

How to identify spam

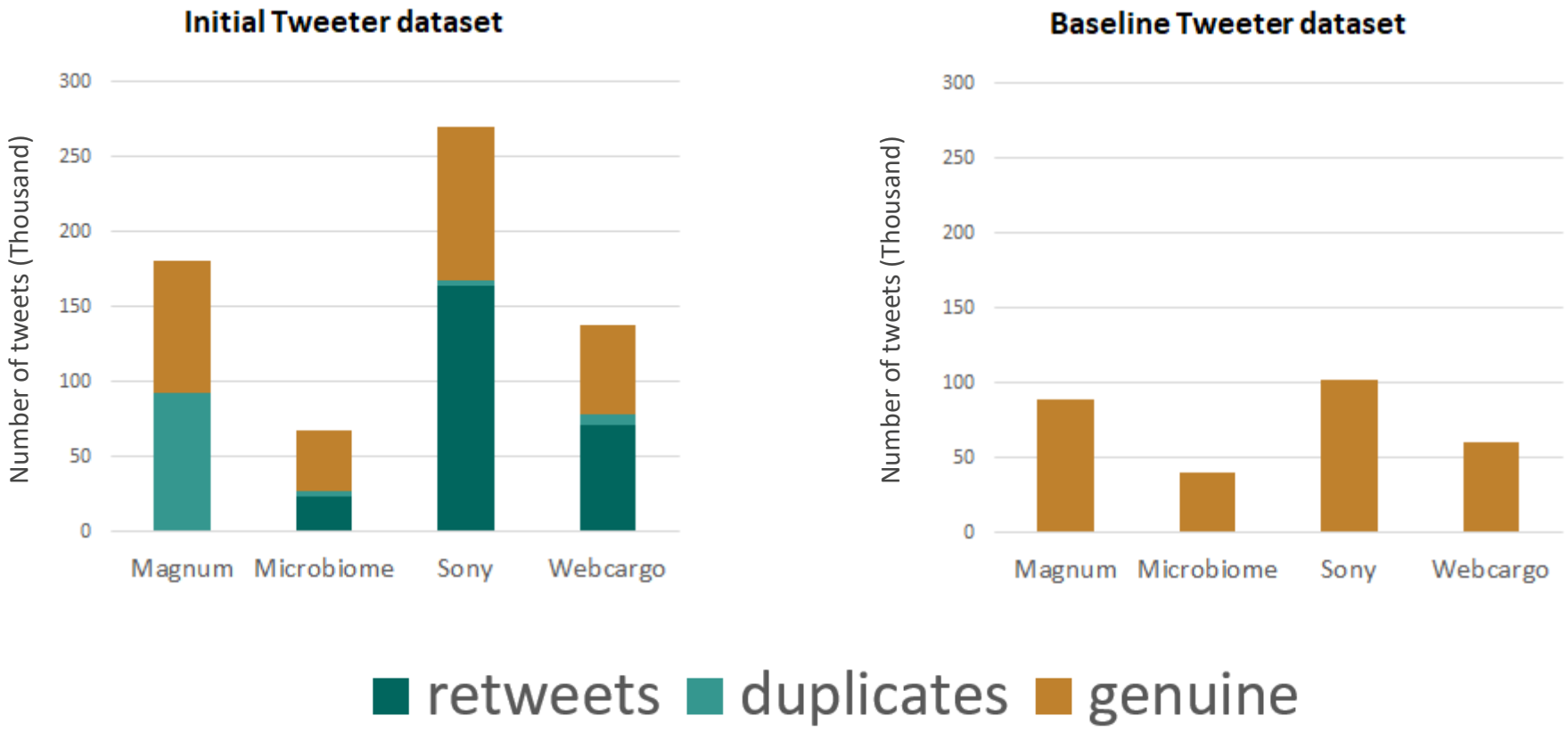
- Use labels provided by users and moderators (Heymann et al., 2007)
- Classify messages manually (Zinman and Donath, 2007)
- Create fake accounts to attract spam (Stringhini et al., 2010)
- Track links to spam and fishing websites (Thomas et al., 2011)
- Analyze account activity (Varol et al., 2017)

Research strategy



* Based on the number of tweets in the database per account after removing retweets and exact duplicates. In our case threshold is > 5 tweets

1. Retweets and exact duplicates



-39.6%, -16.0% observations

2. Non-exact duplicates

Procedure for identifying non-exact duplicates:

1. Change tweets text to lowercase
2. Remove mentions of other accounts (e.g., @john_snow)
3. Remove hashtags
4. Remove links
5. Remove punctuation (e.g., ?!,.•…@%*, etc.)
6. Remove digits

-5.1% observations

2. Non-exact duplicates. Example

Account	Before processing the text
_vegan__vegan_	Please RT? #vegetarian #vegan #healthyfood #recipes Gut-Loving Ginger Mango Green Smoothie dlvr.it/PwDj5s pic.twitter.com/I71INnvhVo
_vegan_recipes_	Please RT? #vegetarian #vegan #healthyfood #recipes Gut-Loving Ginger Mango Green Smoothie dlvr.it/PwDZr3 pic.twitter.com/pXuKygpxVt
veganrecipes	Please RT? #vegetarian #vegan #healthyfood #recipes Gut-Loving Ginger Mango Green Smoothie dlvr.it/Pw4R2K pic.twitter.com/YZyVS64rLE
_veganrecipes_m	Please RT? #vegetarian #vegan #healthyfood #recipes Gut-Loving Ginger Mango Green Smoothie dlvr.it/Pw6HDl pic.twitter.com/g6nu1lfq5e
_vegetarianfood	Please RT? #vegetarian #vegan #healthyfood #recipes Gut-Loving Ginger Mango Green Smoothie dlvr.it/Pw4k50 pic.twitter.com/deSQxsdnAp

After processing the text
please rt gut-loving ginger mango green smoothie

3. Tweets from the top 2% of accounts

On average: 1.44 tweets per account

* After removing retweets and exact duplicates

3. Tweets from the top 2% of accounts

83%

of accounts have only one tweet in the dataset

0.5%

of accounts have more than 12 tweets in the dataset

3. Tweets from the top 2% of accounts

850 tweets about Sony

-5.3% observations

4. Popular accounts

- **98%** of accounts have less than 23,000 followers
- **98%** of accounts have less than 40,000 visits per tweet

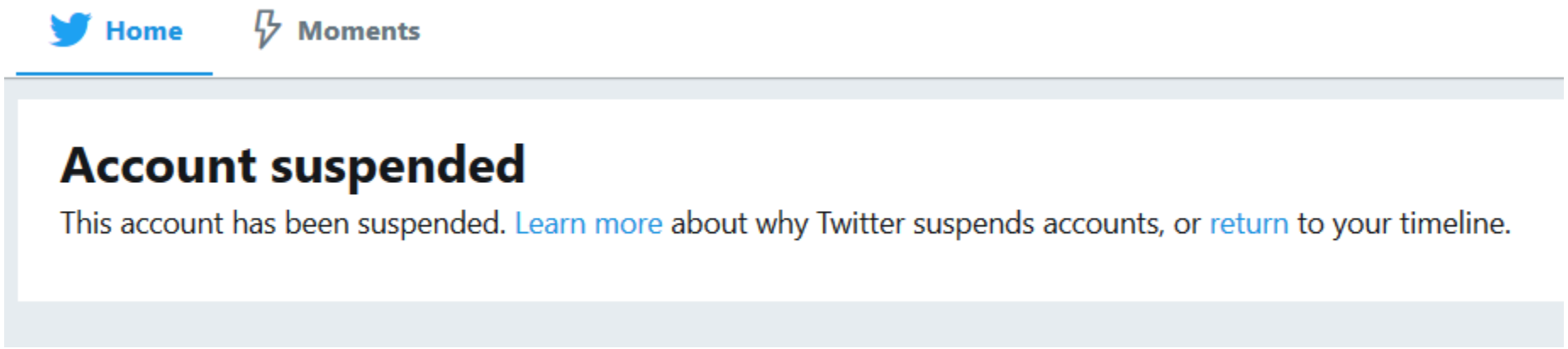
Accounts with largest # of followers	Accounts with largest # of visits
New York Times	New York Times
NASA	Ash Joey
National Football League	NASA
Reuters	National Football League
Play Station	Reuters

-0.9% observations

5. High-risk tweets

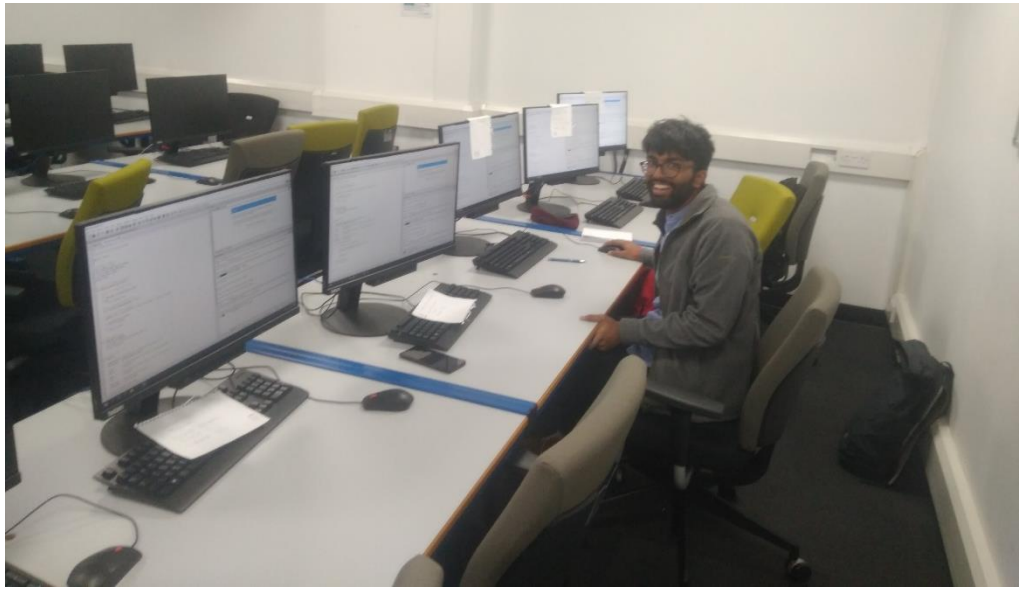
Source: https://blog.twitter.com/

5. High-risk tweets



5. High-risk tweets

- Spam identifier = 1 if account is suspended
  = 0 if not suspended



5. High-risk tweets

3.6%

of accounts in the dataset are suspended

4.7%

of tweets are from the suspended accounts

6. Modeling. Spam identifier

Spam

- Tweets from suspended accounts

OR

- Non-exact duplicates (one instance per sequence of duplicates)

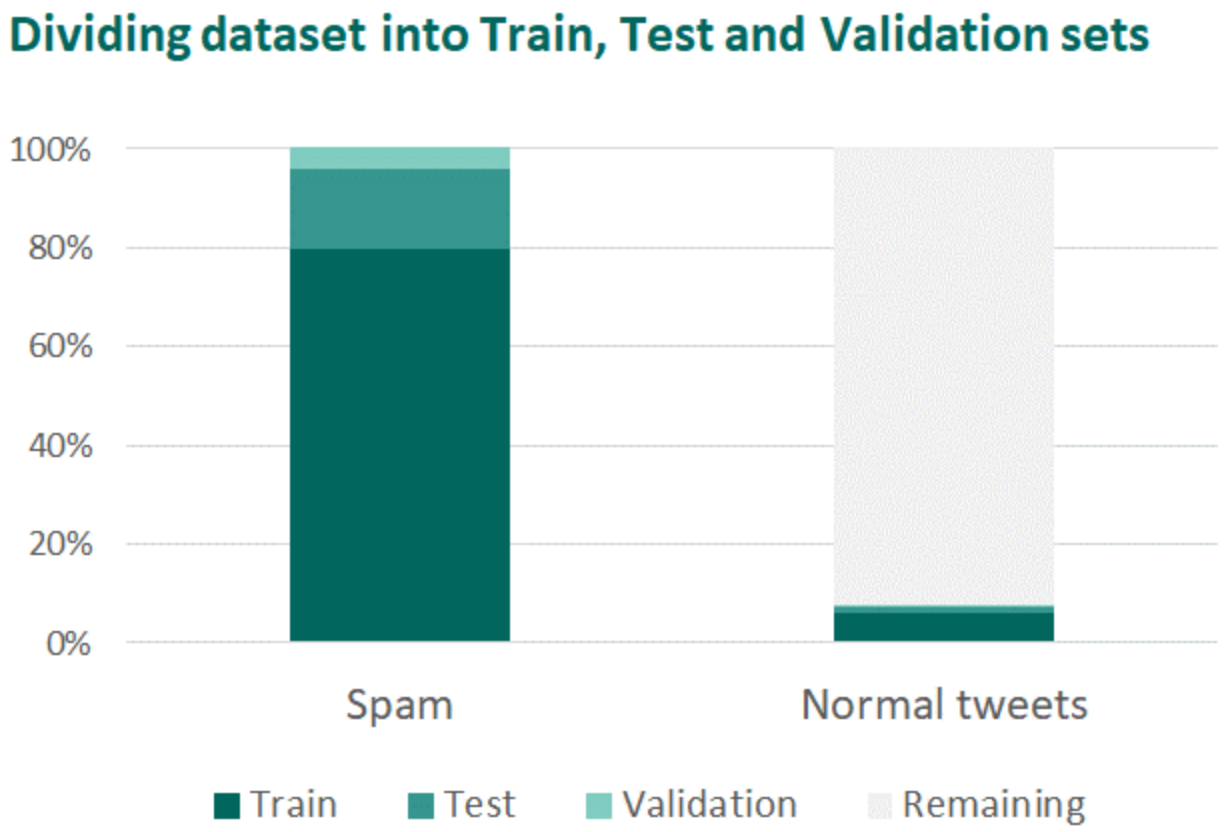
15,179 tweets

Normal tweets

- Tweets from accounts that are not suspended and still exist

222,962 tweets

6. Modeling. Class imbalance



6. Modeling. 37 independent variables

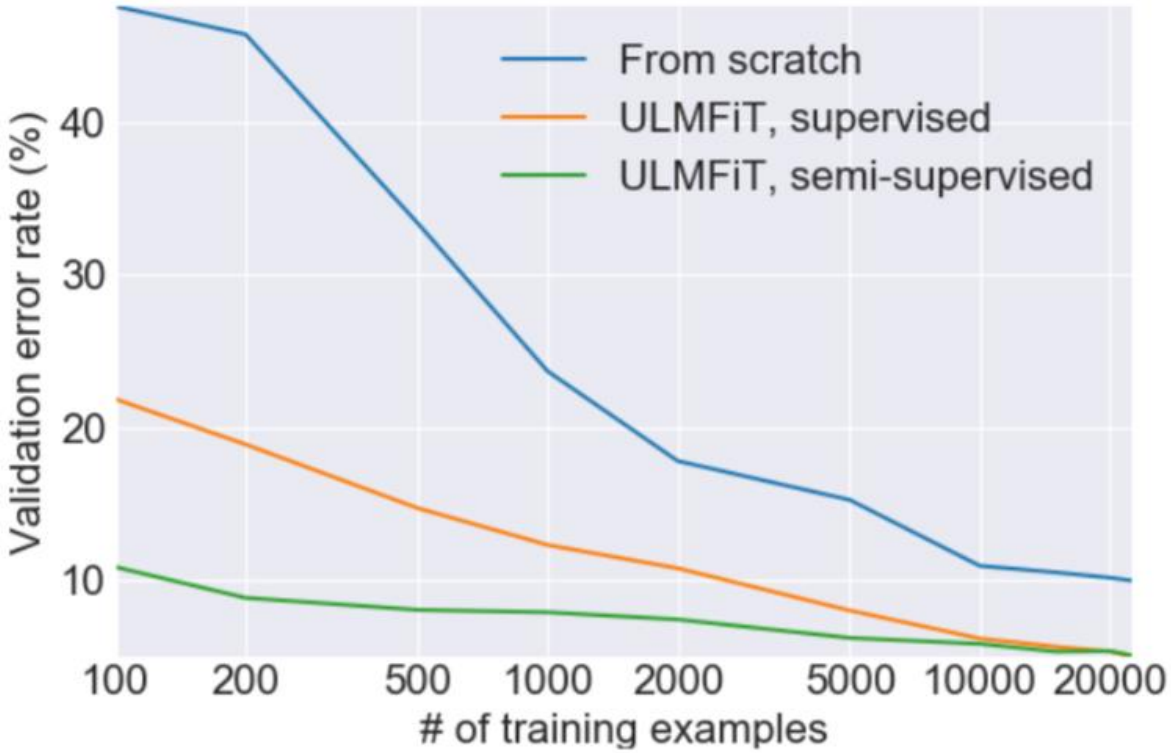
Describing text	Describing text 2	Describing author	Describing account	Other metrics
picture	Sentiment	tweets_in_database	Twitter.Tweets	Engagement.Score
hashtags	word_good	name_2_w	Twitter.Followers	Impact
mentioned	word_like	profession	Twitter.Following	Kred.Influence
links_number	word_new	interest	Twitter.Retweets	Kred.Outreach
links_twitter	word_sex	position	Twitter.Verified	Reach
links_facebook	word_woman	Region		Impressions
links_youtube		Gender		mozRank.Score
links_instagram		Account.Type		
links_other				
Thread.Entry.Type				
Twitter.Reply.Count				

Variables provided only by the Brandwatch

7. Modeling strategy

- One of the most cited papers in the field of email spam detection are two essays by Paul Graham 'A Plan for Spam‘ (2002) and 'Better Bayesian Filtering‘ (2003) that propose to use Naive Bayes classifier
- There are recent successes of applying neural networks to spam detection. Google claims that their accuracy is 99.9%
- We will build a **logistic regression** as a baseline model and **neural network** based on the recent paper by researchers from FastAi group: *Howard, J. and Ruder, S. (2018). Universal Language Model Fine-tuning for Text Classification*

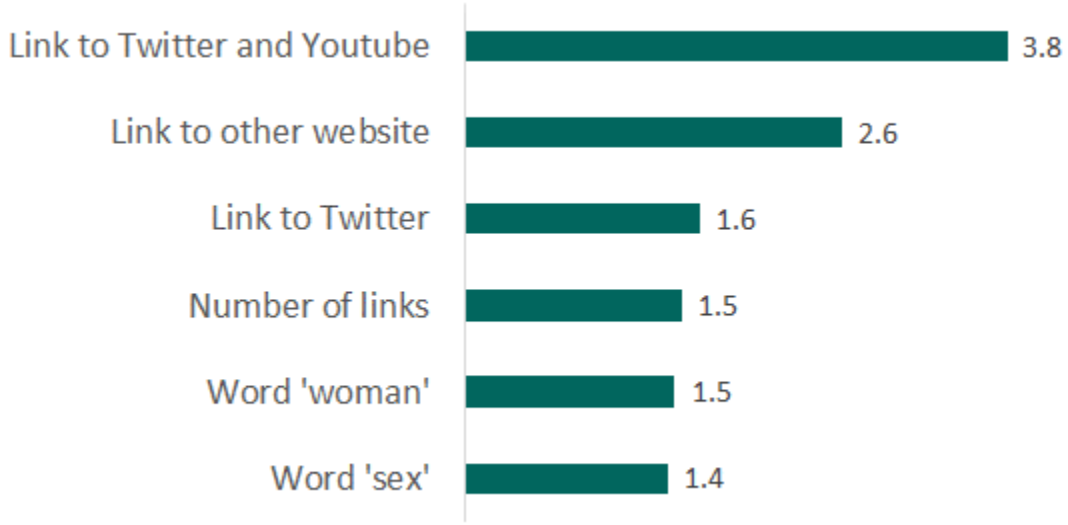
7. Modeling strategy



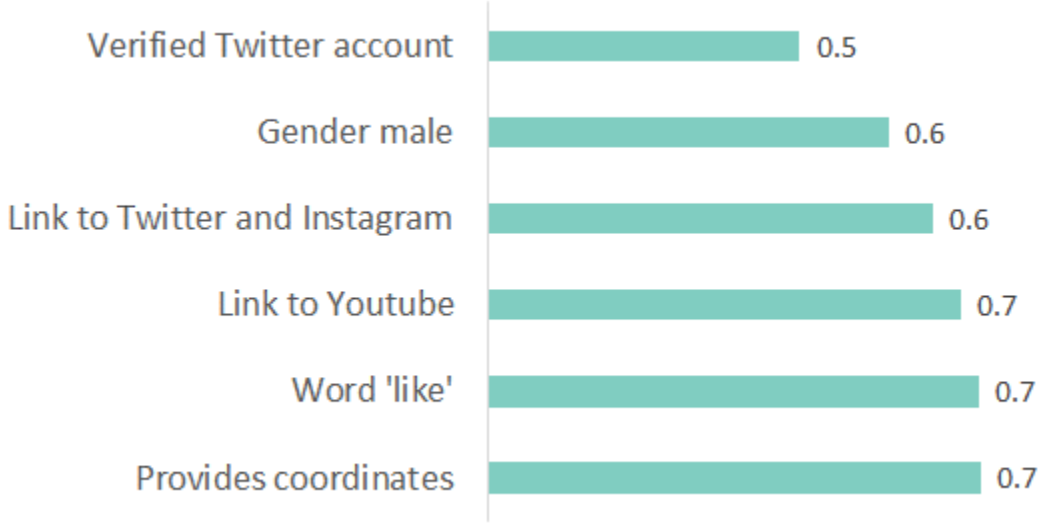
Source: http://nlp.fast.ai/

8. Logit. Variables with strongest impact

Increase probability of spam



Decrease probability of spam



$$\text{Unit of measurement – odds ratio} = \frac{\Pr(tweet\ is\ spam)}{\Pr(tweet\ is\ not\ spam)}$$

8. Logit. Train and test sets

Train set

Model

Data		Spam	Not Spam
	Spam	32.0%	18.0%
	Not Spam	11.8%	38.2%

Test set

Model

Data		Spam	Not Spam
	Spam	31.7%	18.3%
	Not Spam	11.4%	38.6%

% of correctly identified spam: Train – 64.1%, Test – 63.4%

9. Intermediate conclusions

Research question: how to identify tweets from each of the three groups.

1. By identifying retweets, exact duplicates and tweets without any information after text processing we are able to detect majority of tweets from the first group
2. By identifying non-exact duplicates we are able to remove a large proportion of spam messages. This will be further improved after completing modeling stage
3. Identification of ‘professional’ tweets relies on simple ad-hoc rules

10. Suggestions for further research

1. Build a model for identifying ‘professional’ tweets
2. Use ensemble approach to use all observations from the non-spam group in the spam detection model

References

Cs.cmu.edu. (2018). *Zinman and Donath, CEAS 2007 - ScribbleWiki: Analysis of Social Media*. [online] Available at: https://www.cs.cmu.edu/~wcohen/10-802/fixed/Zinman_and_Donath%2C_CEAS_2007.html [Accessed 6 Dec. 2018].

Graham, P. (2002). *A Plan for Spam*. [online] Paulgraham.com. Available at: http://www.paulgraham.com/spam.html [Accessed 6 Dec. 2018].

Graham, P. (2003). *Better Bayesian Filtering*. [online] Paulgraham.com. Available at: http://www.paulgraham.com/better.html [Accessed 6 Dec. 2018].

Howard, J. and Ruder, S. (2018). *Universal Language Model Fine-tuning for Text Classification*. [online] Arxiv.org. Available at: https://arxiv.org/abs/1801.06146 [Accessed 6 Dec. 2018].

Heymann, P., Koutrika, G. and Garcia-Molina, H. (2007). Fighting Spam on Social Web Sites: A Survey of Approaches and Future Challenges. *IEEE Internet Computing*, 11(6), pp.36-45.

Stringhini, G., Kruegel, C. and Vigna, G. (2010). Detecting spammers on social networks. In: *26th Annual Computer Security Applications Conference,ACSAC*. [online] New York, pp.1-9. Available at: https://dl.acm.org/citation.cfm?id=1920263 [Accessed 6 Dec. 2018].

Thomas, K., Grier, C., Ma, J., Paxson, V. and Song, D. (2011). Design and evaluation of a real-time url spam filtering service. In: *IEEE Symposium on Security and Privacy*. [online] IEEE, pp.447-462. Available at: https://www.ieee-security.org/TC/SP2011/PAPERS/2011/paper028.pdf [Accessed 6 Dec. 2018].

Varol, O., Ferrara, E., A. Davis, C., Menczer, F. and Flammini, A. (2017). Online Human-Bot Interactions: Detection, Estimation, and Characterization. In: *The International AAAI Conference on Web and Social Media (ICWSM)*. [online] Available at: https://arxiv.org/abs/1703.03107 [Accessed 6 Dec. 2018].

OTHER SLIDES

Remove non-exact duplicates. Example

Account	Before processing the text	After processing the text
Food Blogger	6 #Foods That Are Good for Gut #Health ift.tt/2yimoyn pic.twitter.com/amtRhMztfP	that are good for gut
	6 #Foods That Are Good for Gut #Health ift.tt/2fv3GYS pic.twitter.com/lChRq7D8GJ	that are good for gut
	6 #Foods That Are Good for Gut #Health ift.tt/2wpggj6 pic.twitter.com/xM4dAkvQFO	that are good for gut
	6 #Foods That Are Good for Gut #Health ift.tt/2xS85zs pic.twitter.com/VWaOCvJZ4i	that are good for gut
	6 #Foods That Are Good for Gut #Health ift.tt/2fAqFSi pic.twitter.com/iuMGffGrhN	that are good for gut
	6 #Foods That Are Good for Gut #Health ift.tt/2xL555Q pic.twitter.com/mMi5EbP1p7	that are good for gut
	6 #Foods That Are Good for Gut #Health ift.tt/2xQAjul pic.twitter.com/vkVlw5mNCw	that are good for gut
	6 #Foods That Are Good for Gut #Health ift.tt/2kdEyLo pic.twitter.com/abgXDCtHJK	that are good for gut
	6 #Foods That Are Good for Gut #Health ift.tt/2xCTAzS pic.twitter.com/wxgDbjcS98	that are good for gut
	6 #Foods That Are Good for Gut #Health ift.tt/2hxbrSj pic.twitter.com/v8W947MOcf	that are good for gut
	6 #Foods That Are Good for Gut #Health ift.tt/2yzQp8J pic.twitter.com/99n5e7mDC5	that are good for gut
	6 #Foods That Are Good for Gut #Health ift.tt/2yacjmU pic.twitter.com/hlQFHaKSv9	that are good for gut
	6 #Foods That Are Good for Gut #Health ift.tt/2fYpYmu pic.twitter.com/CErNAXlqyN	that are good for gut
	6 #Foods That Are Good for Gut #Health ift.tt/2fJbWIz pic.twitter.com/PI0wU4VLg6	that are good for gut

Remove tweets from the top 2% of accounts

There is no objective criteria to separate professionals tweeting about certain topic and regular people that are passionate about topic.

We decided to remove tweets from top 2% of accounts that have the highest number of tweets in the database:

- **2%** of accounts have more than 5 tweets in the database*
- **5.3%** of tweets belong to these accounts

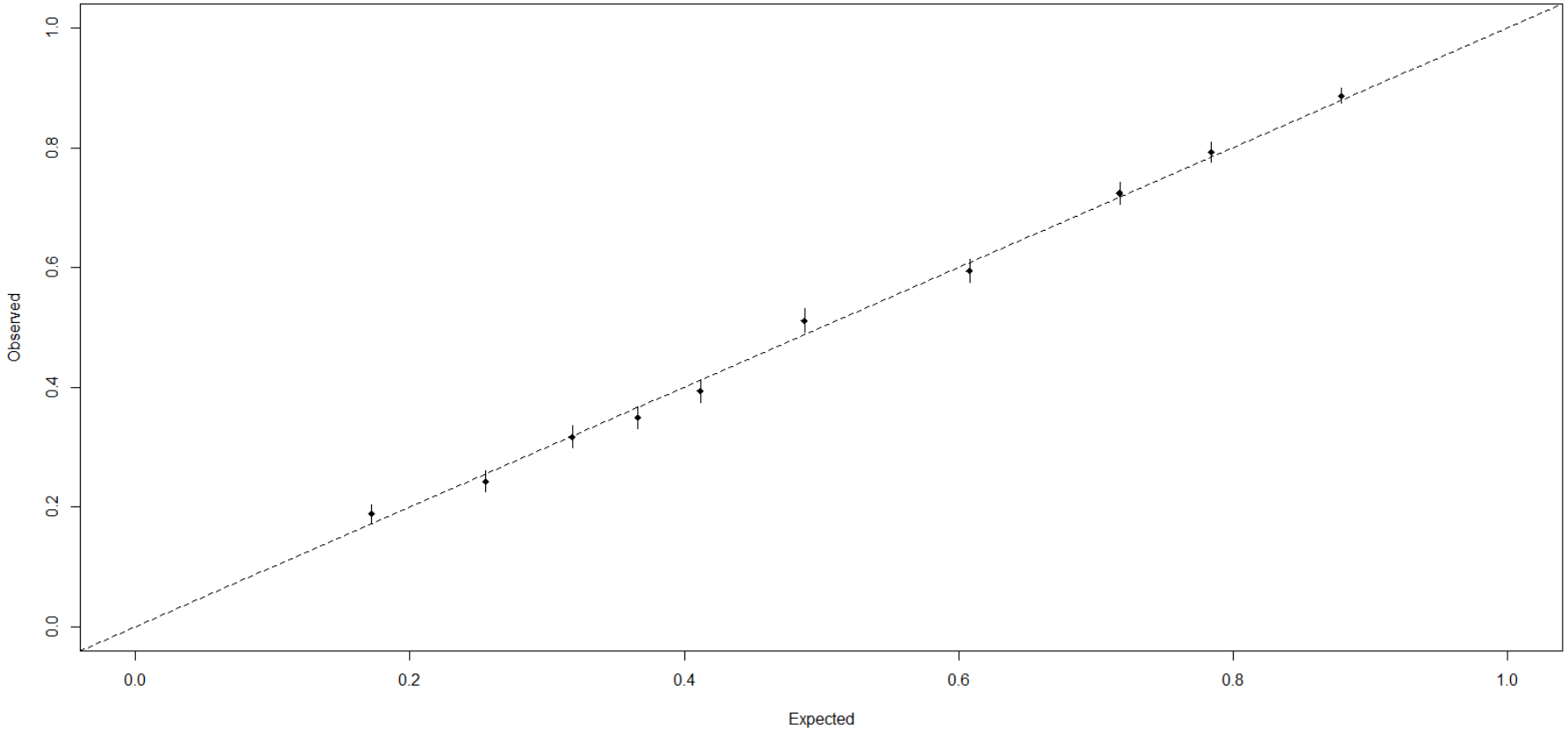
High-risk tweets

	# of accounts	# of tweets	tweets per account
nonexistent	21 541	28 477	1.32
normal	172 969	248 940	1.44
suspended	7 324	13 808	1.89
Total	201 834	291 225	1.44
% of suspended	3.6%	4.7%	-

Regions

1. EastAsia (8 countries + Mongolia)
2. SouthEastAsia (12 countries)
3. CentralAsia (5 countries)
4. SouthAsia (6 countries + Pakistan)
5. MiddleEast (19 countries + Israel + Turkey + Egypt)
6. EasternEurope (22 countries + Finland)
7. NorthAmerica (3 countries + Puerto Rico + Greenland)
8. CentralAmericaCaribbean (21 countries)
9. SouthAmerica (12 countries)
10. SubSaharanAfrica (48 countries)
11. Oceania (16 countries)

8. Exploratory analysis. Logistic regression



Hosmer-Lemeshow value = 22.5

Associated p value = 0.95

Spam identifier

- **What if actual spam account was not suspended by Twitter?**

Such risk exists. However, we should take into account that our data was generated in 2016-2017. Thus for at least two years both Twitter users and Twitter algorithms were analyzing tweets of this account and trying to determine whether it is spam.

We cannot aim at absolute objectivity but we can aim at the level of efficiency of existing Twitter algorithms for identifying spam.

Spam identifier

- **Why train a model based on suspensions when we can just check suspension itself for each account at any moment?**

Yes, we can check suspension at every moment. However, in the real world situation Relative Insight is more likely to work with recent data than with data that is three years old. As a result spam accounts can also be newly created and not classified as spam yet.

The aim of building a model is to be able classify account/tweet as spam **before** it will be done by Twitter.

Spam identifier

- **Will we have a “clean” dataset?**

No, but it will be much cleaner.

Other

- **Online Human-Bot Interactions: Detection, Estimation, and Characterization**

https://arxiv.org/abs/1703.03107