Using Support Vector Machine Ensembles for Target Audience Classification on Twitter

Siaw Ling Lo, **Raymond Chiong** and David Cornforth

PLoS ONE, vol. 10(4), e0122855, April 2015

THE UNIVERSITY OF
NEWCASTLE
AUSTRALIA

The Power of Social Media

In business

- Nearly 80% of consumers would more likely be interested in a company due to its brand's presence on social media[1]
- 77% of the Fortune 500 companies have active Twitter accounts and 70% of them maintain an active Facebook account to engage with their potential customers[2]

[1] Internet Advertising Bureau (IAB) , UK
[2] The University of Massachusetts Dartmouth

With over 1.3 billions active social media users …

how can a company find ***prospective customers*** in the increasingly ***crowded social space***?

Hypothesis

- The content of a Twitter account owner can be used to identify a target audience.
- Twitter users interested in the content posted by an owner -> they choose and take action to follow the account owner -> contents shared should be similar
- Hence, these followers are more likely to comprise the target audience compared to others who are not sharing similar contents.

Twitter and samsungsg

- **Twitter**
  - open and real-time
  - data can be extracted through APIs
- Data (tweets) from samsungsg (the account owner) and its list of followers were extracted from the same period of time.

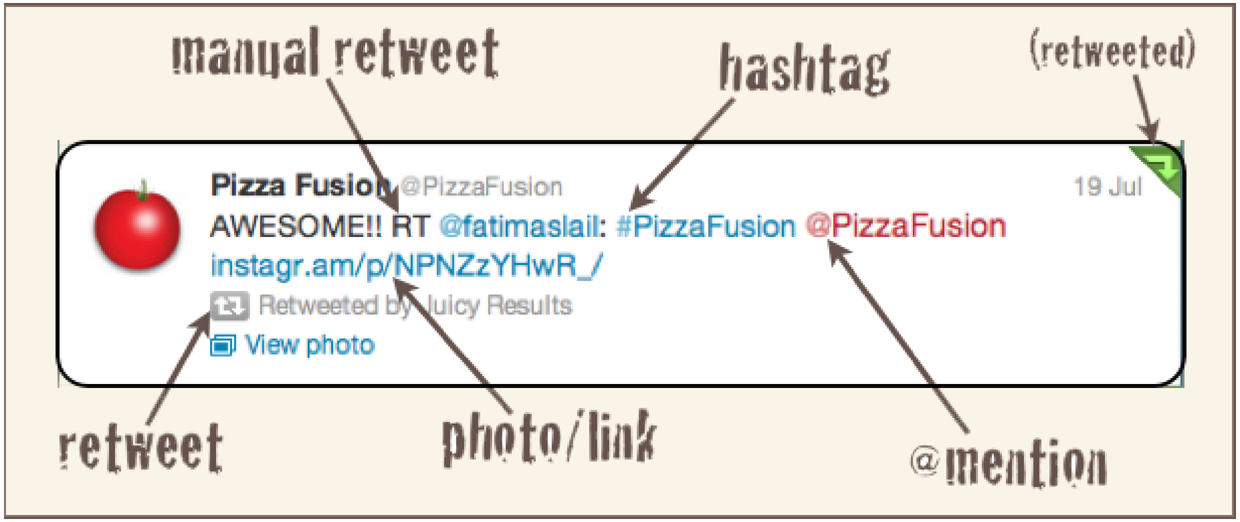
Challenges

- Data privacy – Twitter (open and real-time) instead of Facebook
- Vast amount of data to identify relevant contents.
- Twitter content or Tweet - 140 characters
  - informal languages mix with linguistic variations where localised expression is commonly used
  - purposely misspelt words or repetitions of punctuation signs for emphasis (e.g., “perrrrfeeect” or “!!!!!”)

Challenges

* Special characters used in a tweet:

  RT, #hashtag, @username, link, emoticon



Challenges

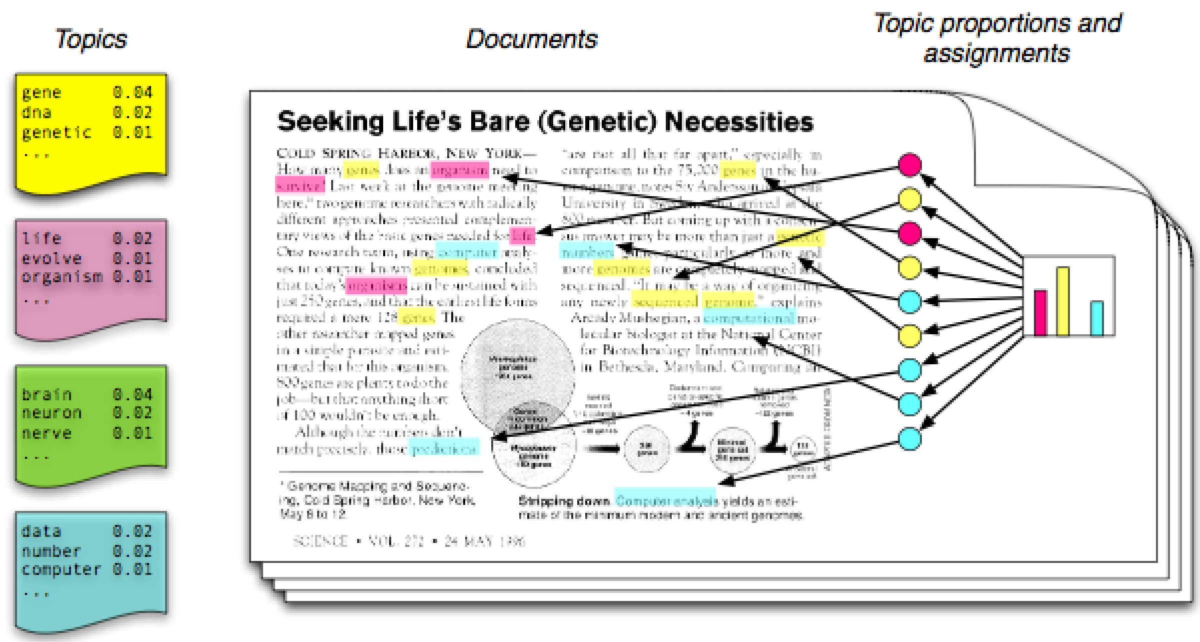
- Vast amount of tweets
  - Assumption: Find those who share similar information as the account owner

  - Supervised learning through annotated training datasets
    - Account owner => positive training data
    - Negative training data?
      - Learn from the contents of individual followers
      - Data imbalance issues

Proposed Approach

* The use of both unsupervised and supervised learning methods for target audience classification on Twitter with minimal annotation efforts
  * [Unsupervised] Twitter Latent Dirichlet Allocation (LDA): topic domains discovery from the contents shared by followers
  * [Supervised] SVM Ensembles: supervised models using the contents from the different account owners of topics identified

LDA



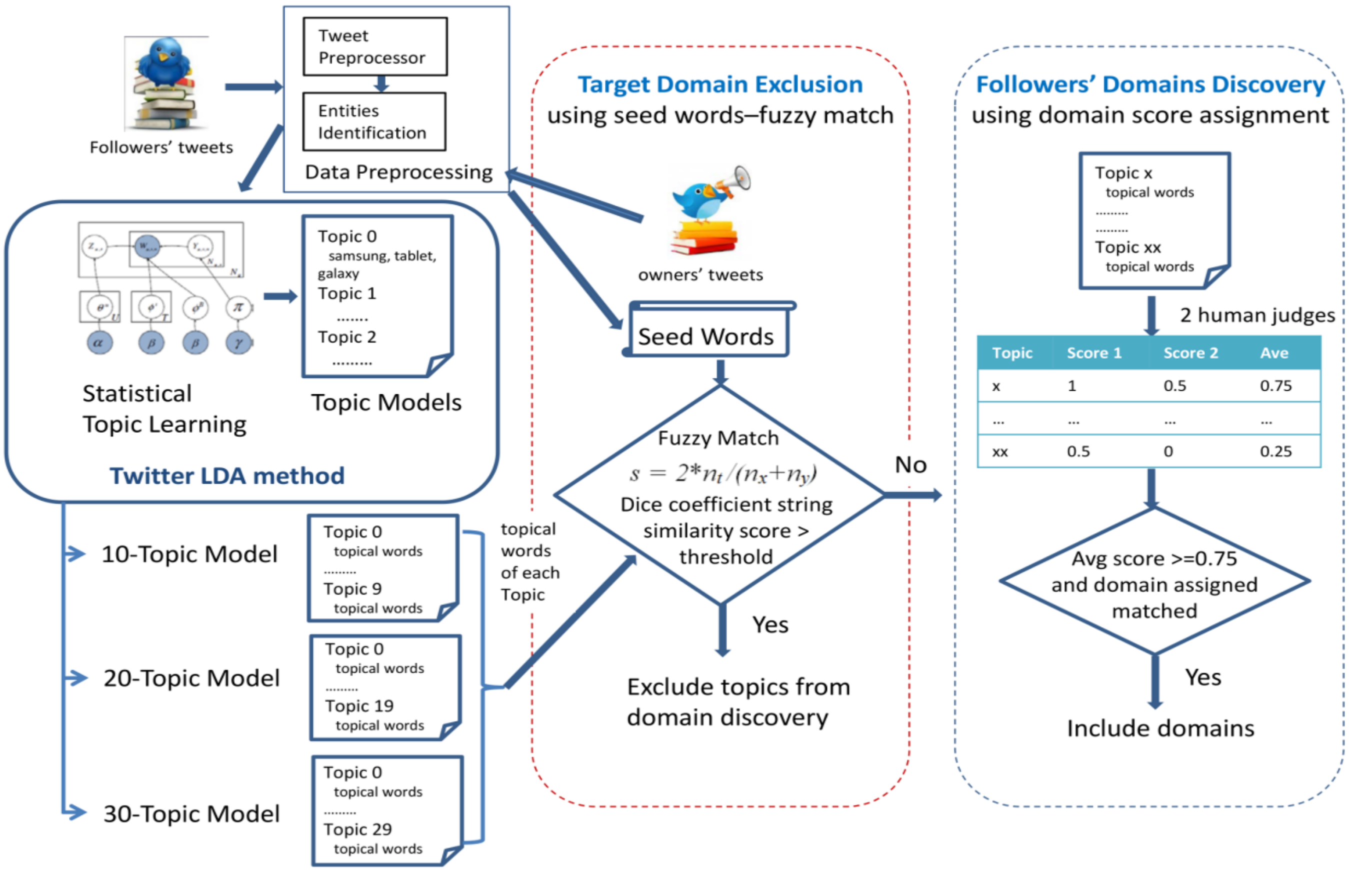
- Each **topic** is a distribution over words
- Each **document** is a mixture of corpus-wide topics
- Each **word** is drawn from one of those topics

LDA is an unsupervised approach in identifying hidden "topics" in the documents, where a topic is a subject like "genetic" or "computer".

Twitter LDA

* Twitter LDA is a an enhanced version of LDA to address the noisy nature of tweets where it handles background words specific to tweets
* Original LDA treats each word as a topic and hence may not work well with Twitter as tweets are short and each tweet is likely a topic
* Instead of combining tweets as a topic, it treats each tweet as a single topic

Followers' domains discovery using Twitter LDA



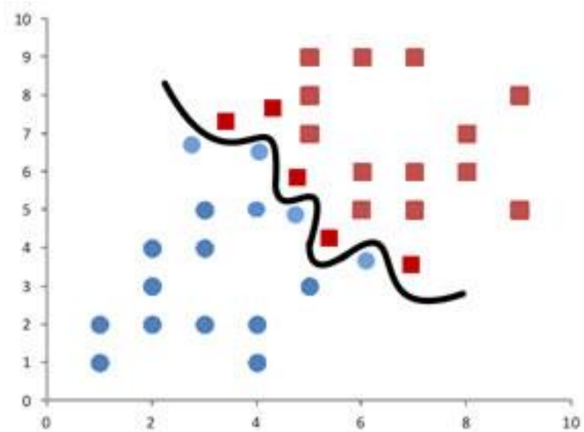
Followers’ Domains Discovery

- 60 topics groups -> exclude from Seed Words – Fuzzy Match
- 2 human judges annotations with scores
- Eight domains with average score of 0.75 and above

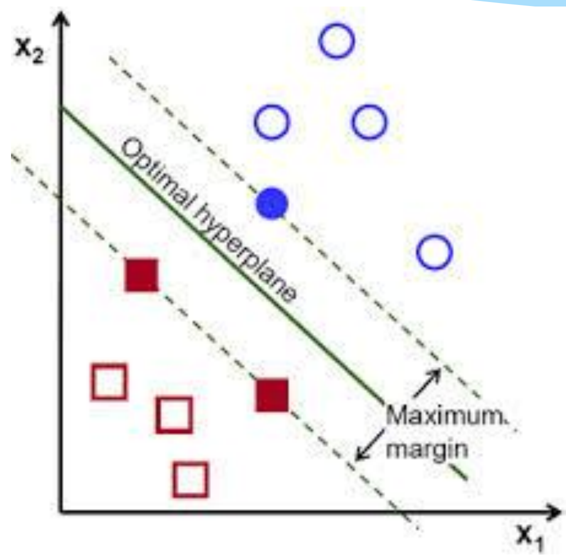
Topic Model	Topic Group Id	Annotated Domain	Topical words	Average Score
10	Topic 9	Daily musing	love, people, life, god, things, feel	1
20	Topic 6	Food	singapore, food, lunch, dinner, coffee, tea, chicken	1
	Topic 7	Football, English premier league (EPL)	united, Manchester, league, Chelsea, david, goal	1
	Topic 8	Daily musing	people, love, life, things, god, feel	1
	Topic 12	Singapore related	singapore, airport, points, club, changi	0.75
	Topic 0	Daily musing	happy, video, birthday, love, mothers	0.75
30	Topic 10	Daily musing	day, good, happy, morning, mothers, birthday, dinner	1
	Topic 15	Daily musing	time, work, sleep, school, long	1
	Topic 18	Daily musing	people, life, love, happy, things, god	1
	Topic 28	Football, EPL	chelsea, league, united,match, madrid	1
	Topic 1	Social media marketing	social, media, marketing, twitter, facebook, business	0.75
	Topic 14	Music	singapore concert, tour, fans, tickets, album	0.75
	Topic 16	Transport	singapore, mrt, blk, bus, wifi	0.75
	Topic 25	News	indonesia, model, tokyo, festival	0.75

SVM model

Input space



Feature space

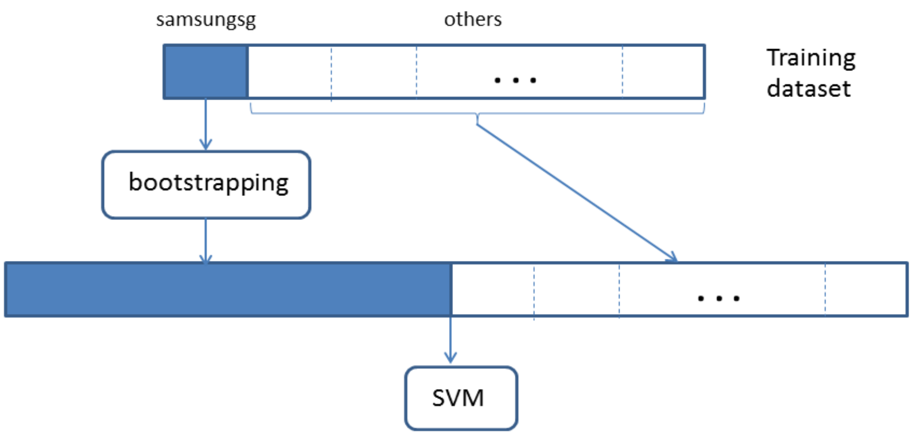


- Supervised learning approach for two or multi-class classification
- It separates a given known set of {+1, -1} labelled training data via a hyperplane that is maximally distant from the positive and negative samples respectively.

SVM Ensembles

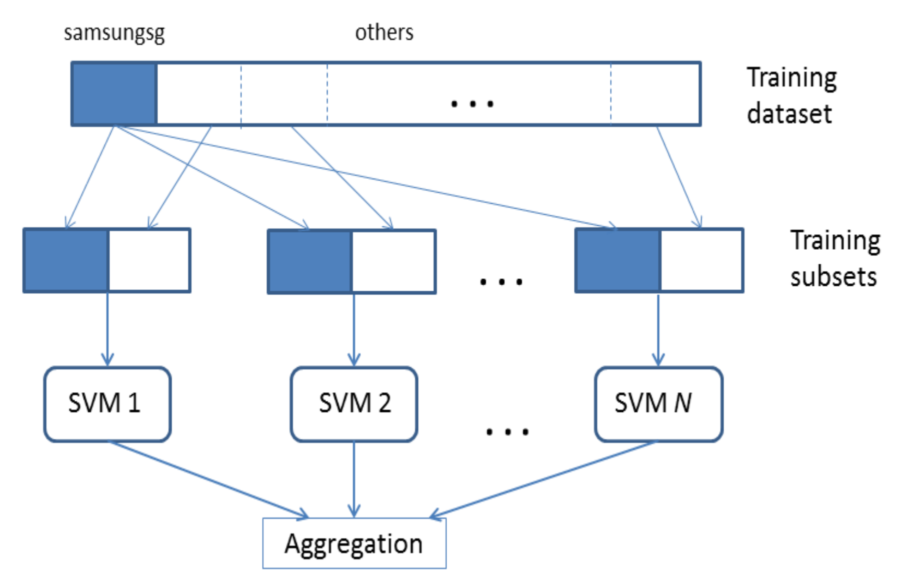
* Data imbalance issues
  * positive dataset – account owner
  * Negative dataset – other domains discovered from followers (extracted from identified account owners)
* Two approaches
  * Bootstrapping using a single SVM model
  * Ensembles using multiple SVM models

SVM Ensembles



Method	Training dataset	Configuration
SVM with bootstrap sampling	samsungsg (1978) and others (1978)	1 SVM model

SVM Ensembles



Method	Training dataset		Configuration
SVM with 10 random sampling with majority vote	samsungsg (200)	others (~200) x 10	10 SVM models
SVM with majority vote	samsungsg (200)	10 others	10 SVM models
SVM with bagging	samsungsg (200)	others (1978)	10 SVM models
SVM with stacking	samsungsg (200)	10 others	10 SVM models with Naïve Bayes (kernel) as the tier two classifier

Experimental Setup

- Data Collection
  - Time of tweets : 2 Nov 2012 to 3 Apr 2013.
  - The most recent 200 tweets by the account owner (samsungsg)
  - For each of the followers, Twitter API is used to extract their past 100 tweets, giving a total of 187,746 records, and 2,449 unique users having at least 5 tweets
- Twitter Search API is used

Performance Metrics

$$precision = TP/(TP + FP)$$

$$recall \text{ or } True\ Positive\ Rate\ (TPR) = TP/(TP + FN)$$

$$True\ Negative\ Rate\ (TNR) = TN/(FP + TN)$$

$$F\ measure = 2 \times \frac{precision \times recall}{precision + recall}$$

$$G\ mean = \sqrt{TPR \times TNR}$$

where TP, TN, FP and FN represent the true positive, true negative, false positive and false negative respectively.

Testing Datasets

* Contents of 300 followers (which were randomly sampled) were manually annotated
* 1239 features
  * Term frequency with word stemming
* 124,462 records were used

Experimental Results

* Representative Target Topical Words

Topic Model	Topic Group Id	Topical words
10	Topic 1	samsung, galaxy, phone, iphone, app, mobile
	Topic 8	singapore, android, ipad, Samsung, sg
20	Topic 9	tv, led, Samsung, contest, giveaway
	Topic 10	galaxy, Samsung, android, tablet, sony, xperia
	Topic 16	samsung, galaxy, android, phone, mobile, iphone, app
30	Topic 0	samsung, galaxy, android, phone, note, iphone, htc
	Topic 2	tv, Samsung, led, video, review, hd
	Topic 12	android, touch, tablet, pc
	Topic 17	galaxy, Samsung, video
	Topic 23	app, google, ipad, android, iphone

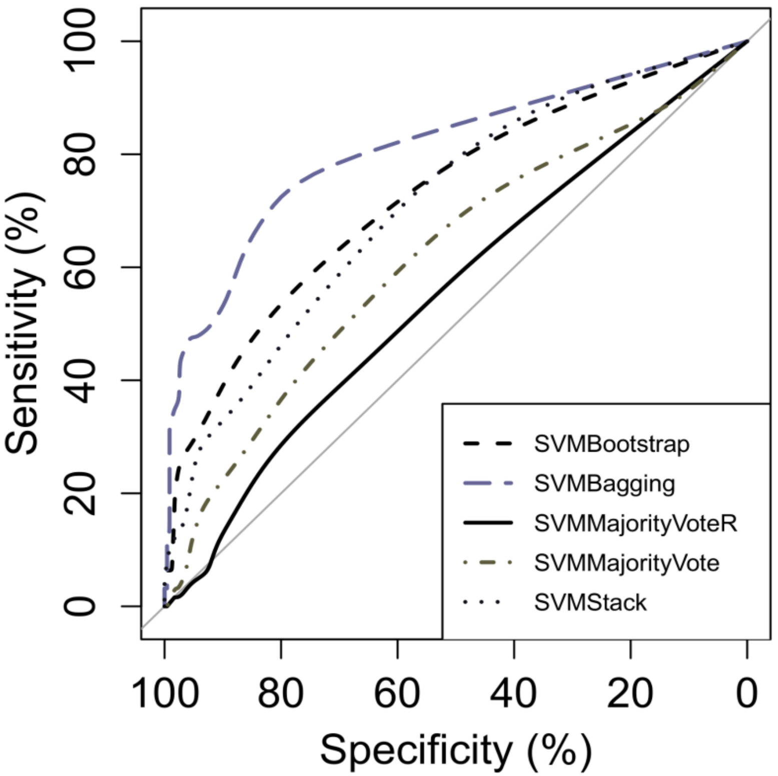
Experimental Results

- Training Performance of Various SVM Ensembles
  - 10 fold cross-validation
  - Bootstrapping method – best result
  - Random sampling – worst result

Method	Recall	Precision	F measure	G Mean
SVM with bootstrapping sampling	1	0.98	0.99	0.99
SVM with 10 random sampling with majority vote	0.31	0.46	0.37	0.54
SVM with majority vote	0.84	0.38	0.52	0.85
SVM with bagging	0.69	0.97	0.80	0.83
SVM with stacking	0.96	0.90	0.93	0.95

Experimental Results

* ROC curves of various SVM ensembles on the testing dataset



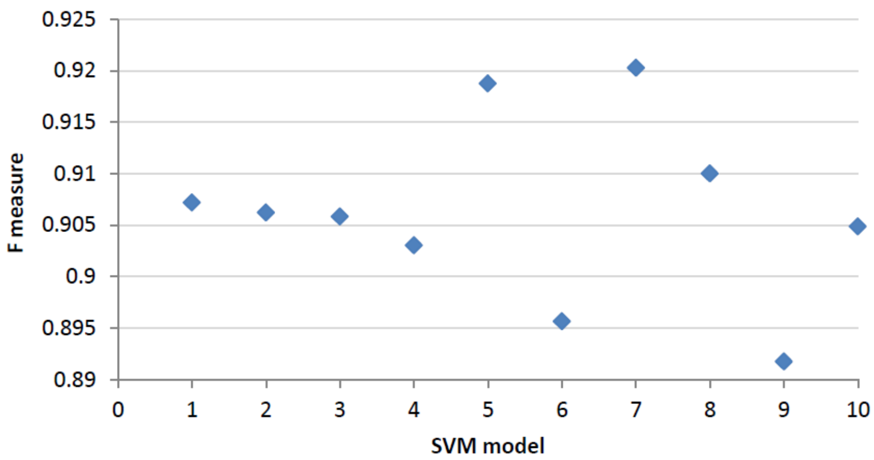
Experimental Results

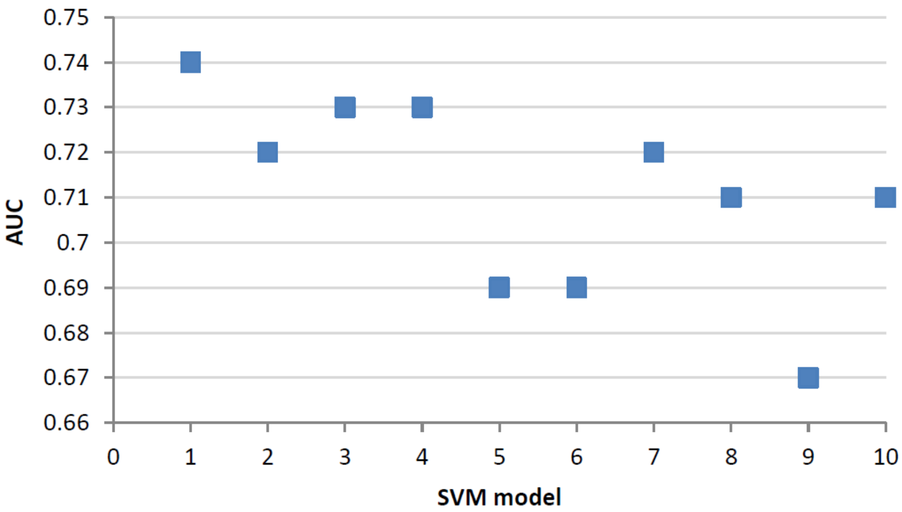
* Results of various SVM ensembles on the testing dataset
  * The SVM ensemble with bagging performs the best
  * The bootstrapping method is the next best performer, followed by the stacking method.
  * Both majority vote methods do not perform as well with the random sampling method obtaining only an AUC value of 0.62

Method	AUC	Time taken (s)
SVM with bootstrapping sampling	0.76	1932±61
SVM with 10 random sampling with majority vote	0.62	722±29
SVM with majority vote	0.64	723±16
SVM with bagging	0.89	482±22
SVM with stacking	0.73	629±25

Discussion

* Inconsistency from 10 SVM models through random sampling:





Advantages of using an ensemble method is to minimise the risk of choosing a particularly poor performing classifier from the list of randomly generated models

Discussion

* G mean is a good indicator to assess an ensemble’s performance.
* While majority vote methods have lower F measure scores, SVM majority vote that uses the dataset from each of the 10 account owners (instead of random sampling) has a higher G mean.
* This implies that the method has a more balanced combination and hence is not biased towards any class. As a result, it has performed better in classifying the testing dataset.

Discussion

- SVM ensemble using bagging does not perform as well in the training dataset but generalise well in the testing dataset
  - Statistical and computational reasons

Conclusion

- Using unsupervised (Twitter LDA) and supervised (SVM ensembles) learning methods, it is possible to automatically classify and identify a target audience from a list of followers of a Twitter account
- Account owners' tweets can be used as the training dataset in an ensemble system for classifying the target audience with minimal annotation efforts
- A novel way of constructing the training dataset from various account owners for ensemble learning, actionable insights can be uncovered to assist in making better decisions for any company

Ongoing/Future Projects

* Development of new approaches for online topic detection using SenticNet
* Intelligent dictionary generation for financial news analysis based on physiological measures (e.g., heart rates, skin conductance, pupil diameters) and sentiment analysis

optimizationBenchmarking.org

* In this talk, I have discussed a *machine learning* problem.
  1. We have done a set of experiments and tested different methods to tackle this problem.
  2. We compared the results of the different methods.
  3. We presented the results in diagrams and tables.
* This is a very typical way of doing research in our domain.
* But it is also cumbersome and there is always a risk of making mistakes (statistical soundness, typos in values, ...).
* With the *optimizationBenchmarking.org evaluator*, we hope to make things easier for researchers.

optimizationBenchmarking.org

- The *optimizationBenchmarking.org evaluator* is a tool that
  - can read experimental results (log files) produced by either optimisation or machine learning processes
  - produce human-readable reports either in HTML or LaTeX (compiled to PDF), which contain performance results and comparisons of different algorithms
- Currently available as the alpha version 0.8.3 at http://www.optimizationBenchmarking.org/

optimizationBenchmarking.org

Experi-ments

- Manually done
- Several algorithms
- Several instances
- Several runs
- 1 log file per run

Evaluator

- Reads in log files
- Performs user-defined evaluations
- Produces report

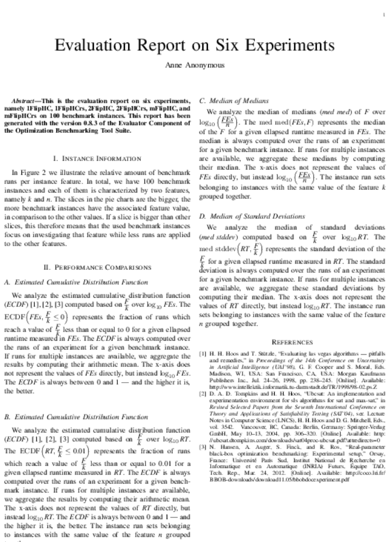
Reports

- Contains comp-arisons, diagrams, tables, and conclusions
- In ‘publishable‘ format

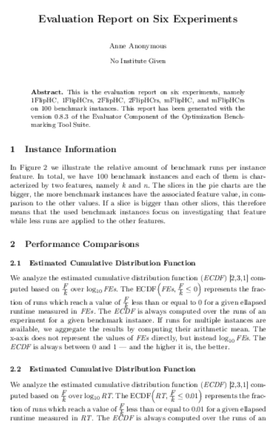
*Currently, the selection is quite limited: This is work in progress, more diagrams and evaluation modules will be added in the coming versions

optimizationBenchmarking.org

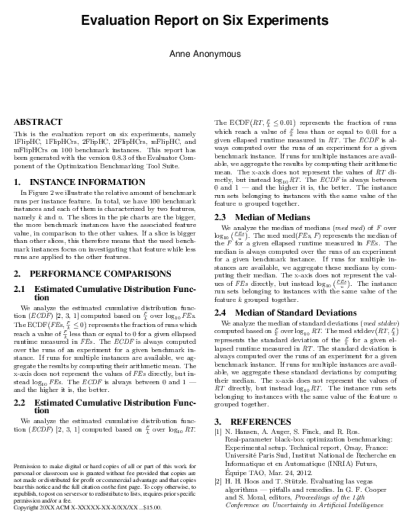
* Reports are generated for different formats and document classes



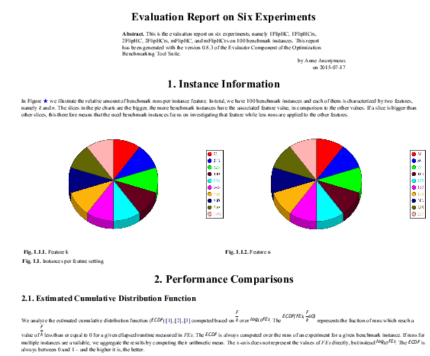
IEEE Transactions



Springer LLNCS



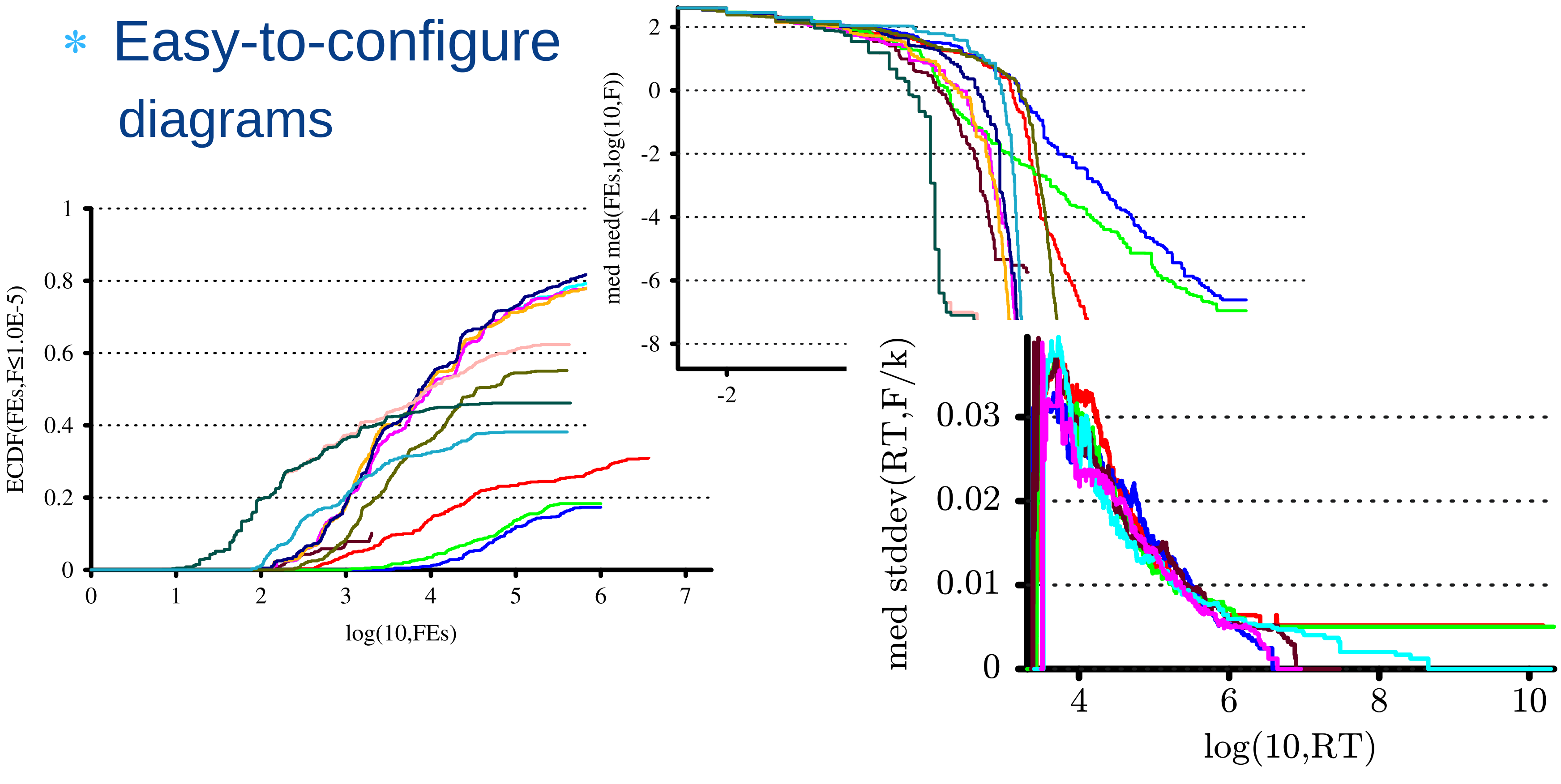
SigAlternate



XHTML

optimizationBenchmarking.org

* Easy-to-configure diagrams



optimizationBenchmarking.org

- Goal: Drastically reduce time needed to analyse experimental data
- Easier to understand relationships between algorithm parameters, instance features, and performance
- Easier to compare different algorithms and setups
- Reduce chance of making statistical mistakes
- Provide figures (and conclusions/text building blocks) that can directly be included into publications

Thank You

siawling.lo@uon.edu.au
Raymond.Chiong@newcastle.edu.au