Institute of Policy Studies

ENGAGING MINDS, EXCHANGING IDEAS

‘Digital Frontiers Series: 1’

Seminar on
“Assessing the Rationality of Political Online Space: Man and Machine”

Wednesday, 11 February 2015

Conference Room, Level 1, Oei Tiong Ham Building

Lee Kuan Yew School of Public Policy
National University of Singapore

Possibilities of Machine Classification

Professor Lim Ee-Peng

Professor of Information Systems & Director
Living Analytics Research Centre
School of Information Systems
Singapore Management University

SMU
SINGAPORE MANAGEMENT UNIVERSITY

LARC
LIVING ANALYTICS RESEARCH CENTRE

Carnegie Mellon University

Analyzing Political Online Space: Man and Machine

(Part 2: The Machine Perspective)

Ee-Peng LIM
Co-Director, LARC

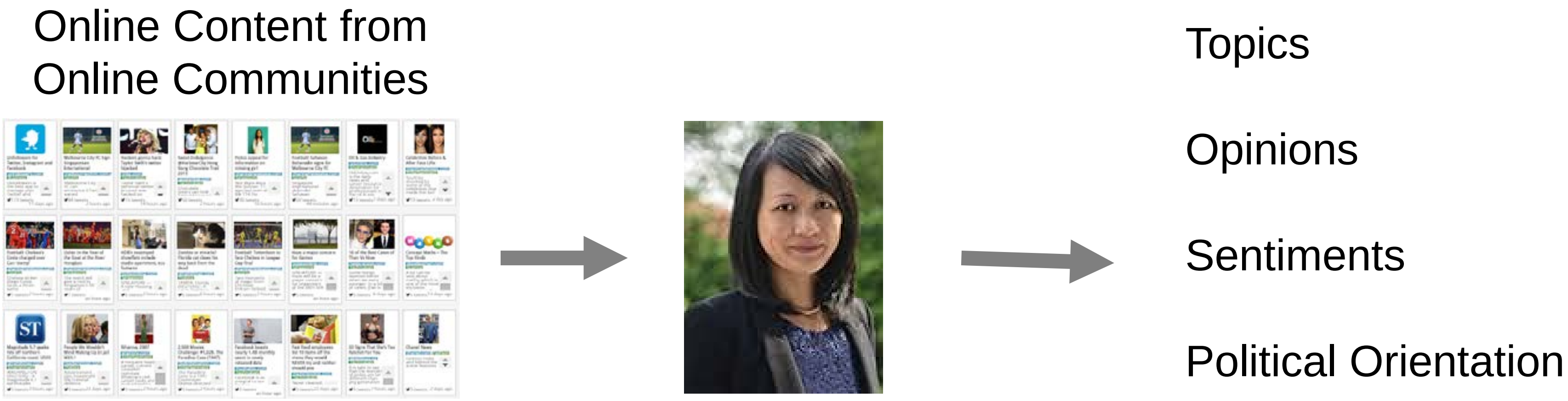
Collaboration

Aek Palakorn

Carol Soon

Philips Kokoh Praseyto

How do we analyze online content?



Online Content from Online Communities

Topics

Opinions

Sentiments

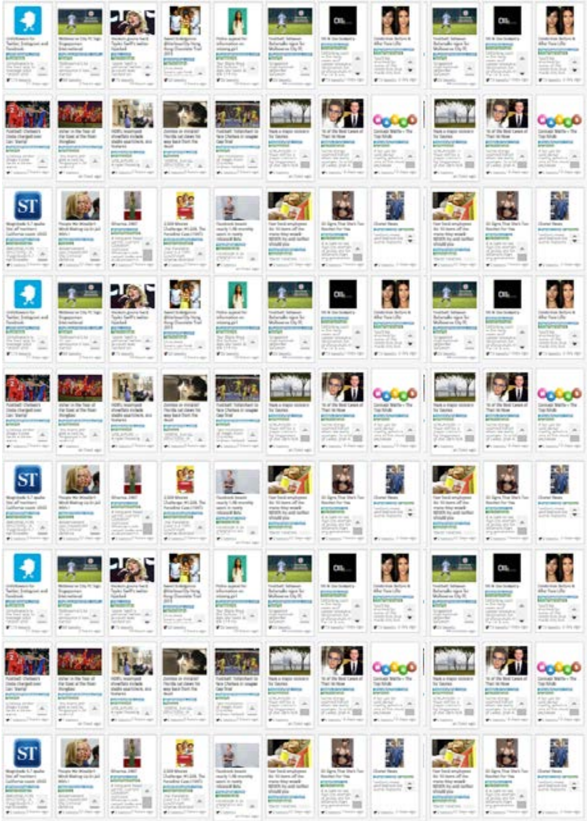
Political Orientation

Human can give accurate reading of natural language text BUT with caveats:

- **Familiar with the relevant domain**
- **Trained in making good judgment**

Human approach is not scalable

Online Content from
Online Communities



Topics

Opinions

Sentiments

Political Orientation

Can machine come to rescue?

Answer: YES but with caveats

- **Machine is good at mechanical activities**
- **Machine needs training**
- **Machine may not perform as accurate as human**

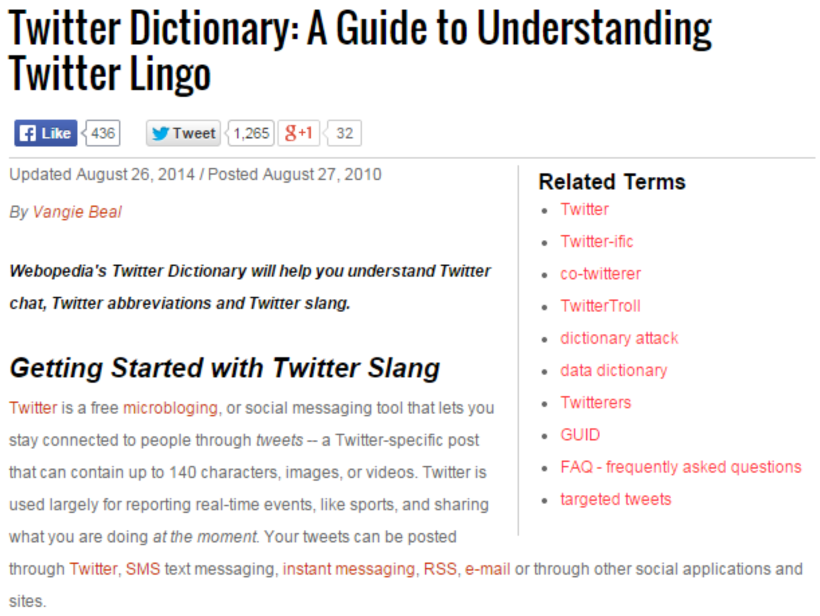
Technical Issues

- Data crawling
  - Blogs
  - Social media: Facebook, Twitter, Instagram
- Heterogeneous data
  - Long text vs short text
  - Directory/ad page vs content page
  - Multi-page content

Technical issues

- Dynamic data
  - SG users generate about
    1.5M to 2M Twitter messages/day
    40M Twitter messages/month
    3.2M URLs in Twitter messages/month

- Social media lingo
  - Twitter Dictionary
    http://www.webopedia.com/
  - Hashtags
    http://hashtagdictionary.com



Twitter Dictionary: A Guide to Understanding Twitter Lingo

Updated August 26, 2014 / Posted August 27, 2010

By Vangie Beal

Webopedia's Twitter Dictionary will help you understand Twitter chat, Twitter abbreviations and Twitter slang.

Getting Started with Twitter Slang

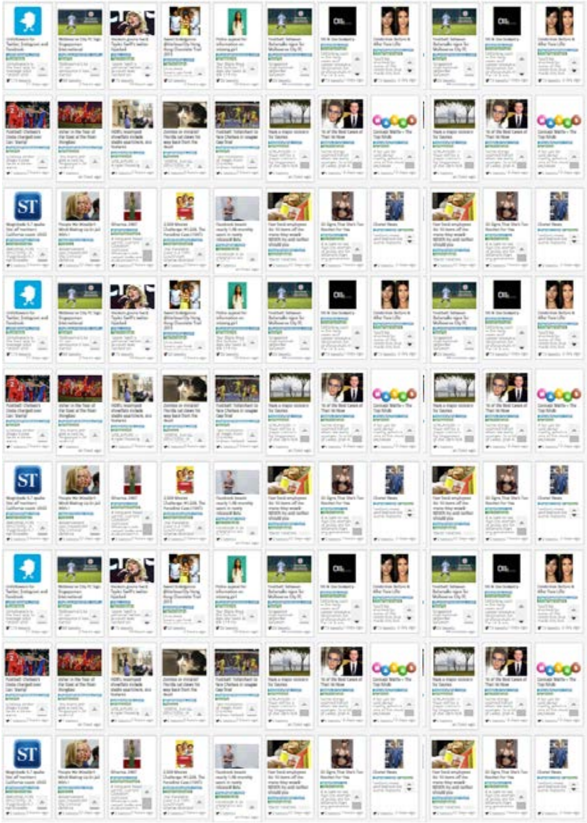
Twitter is a free microbloging, or social messaging tool that lets you stay connected to people through *tweets* -- a Twitter-specific post that can contain up to 140 characters, images, or videos. Twitter is used largely for reporting real-time events, like sports, and sharing what you are doing *at the moment*. Your tweets can be posted through Twitter, SMS text messaging, instant messaging, RSS, e-mail or through other social applications and sites.

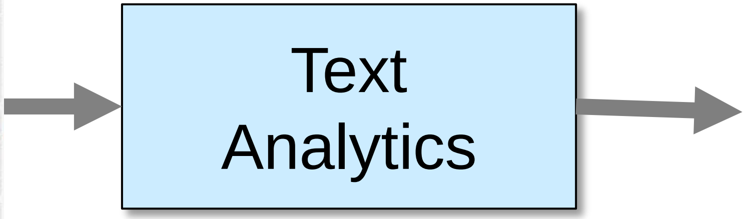
Related Terms

- Twitter
- Twitter-ific
- co-twitterer
- TwitterTroll
- dictionary attack
- data dictionary
- Twitterers
- GUID
- FAQ - frequently asked questions
- targeted tweets

Text Analytics for Online Content

Online Content from
Online Communities





Text Analytics

Topics

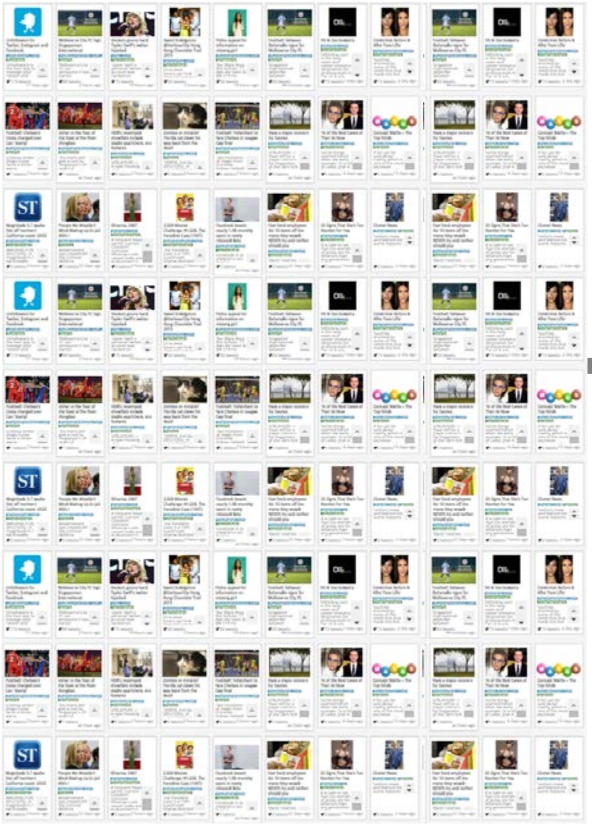
Opinions

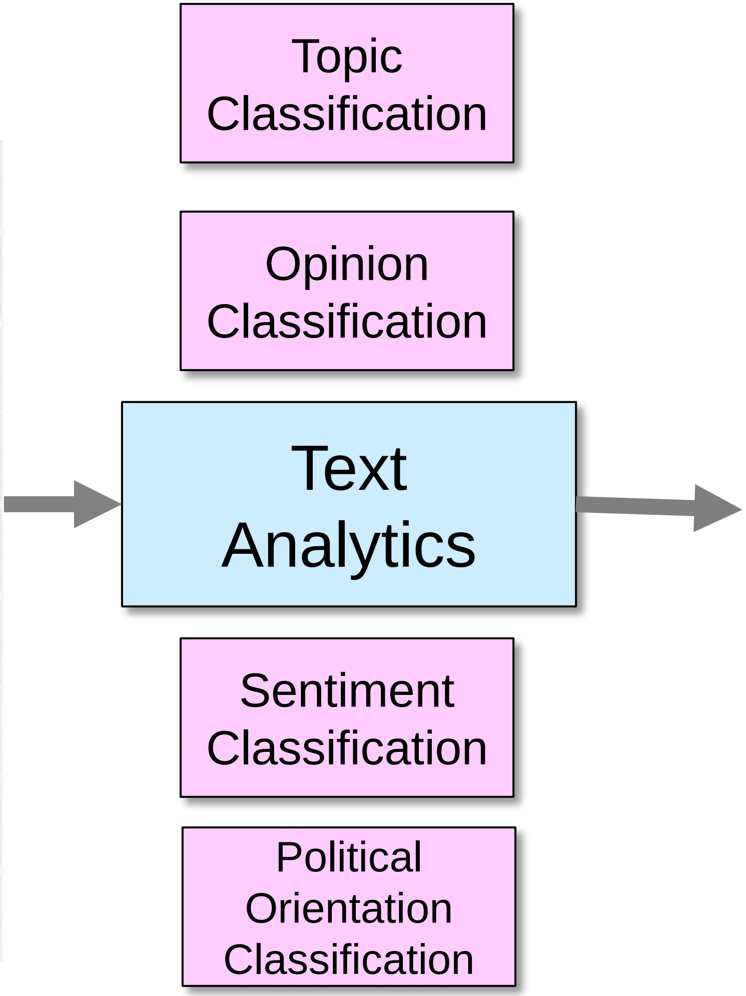
Sentiments

Political Orientation

Text Classification for Online Content

Online Content from Online Communities





Topic Classification

Opinion Classification

Text Analytics

Sentiment Classification

Political Orientation Classification

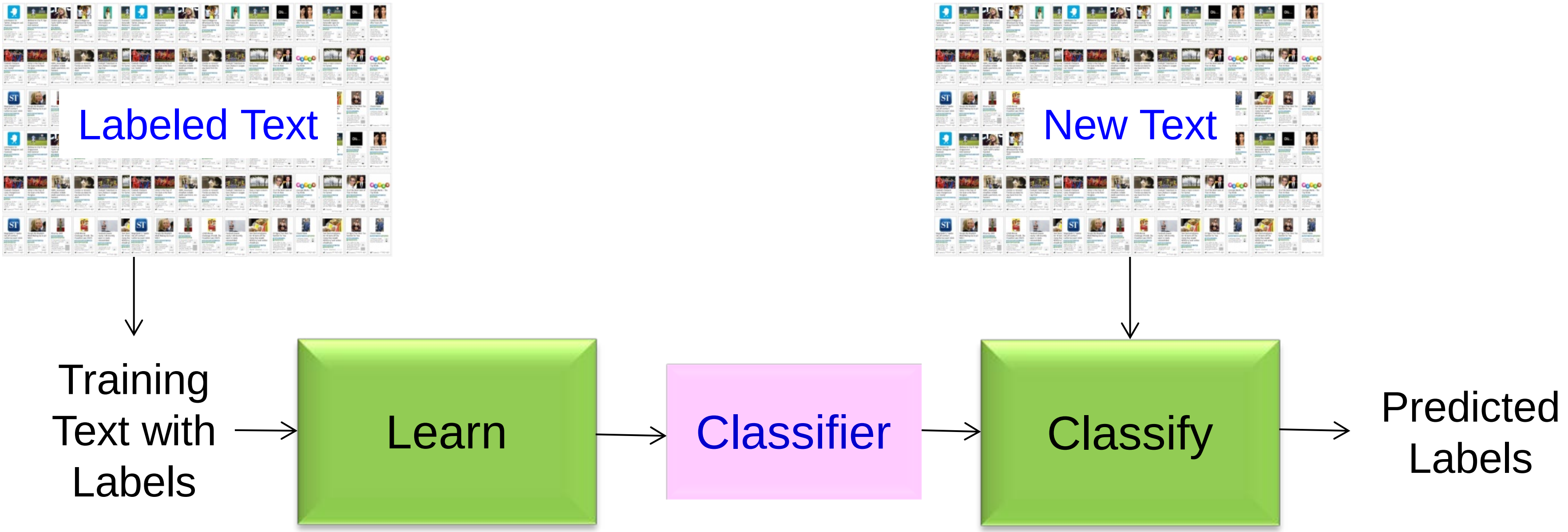
Topics

Opinions

Sentiments

Political Orientation

How does text classification work?



Labeled Text

New Text

Training Text with Labels → Learn → Classifier → Classify → Predicted Labels

Labels

- Topics
- Opinions
- Sentiments
- Political Orientation

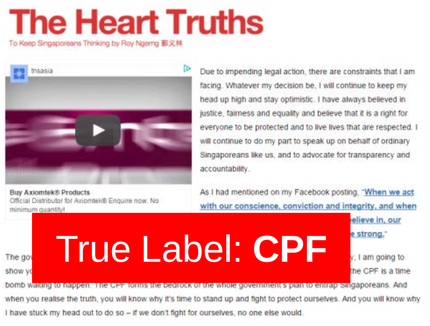
Type of Classifiers

- Rule based classifier
- Naïve Bayes classifier
- Logistic Regression
- Support Vector Machine
- Stochastic Gradient Descent

Accuracy of Classifier

LARC
LIVING ANALYTICS RESEARCH CENTRE

- Text with true labels



The Heart Truths

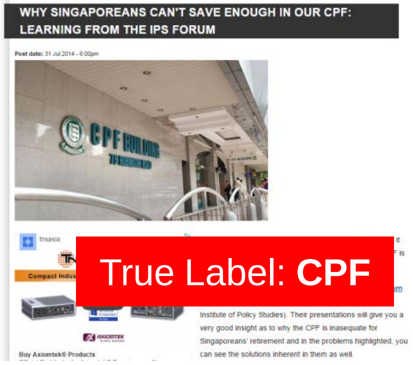
To Keep Singaporeans Thinking by Roy Ngerng 鄞义林

Due to impending legal action, there are constraints that I am facing. Whatever my decision be, I will continue to keep my head up high and stay optimistic. I have always believed in justice, fairness and equality and believe that it is a right for everyone to be protected and to live lives that are respected. I will continue to do my part to speak up on behalf of ordinary Singaporeans like us, and to advocate for transparency and accountability.

As I had mentioned on my Facebook posting, "**When we act with our conscience, conviction and integrity, and when**
elieve in, our
e strong."

The go
, I am going to
show y
the CPF is a time
bomb waiting to happen. The CPF forms the bedrock of the whole government's plan to entrap Singaporeans. And when you realise the truth, you will know why it's time to stand up and fight to protect ourselves. And you will know why I have stuck my head out to do so – if we don't fight for ourselves, no one else would.

True Label: **CPF**



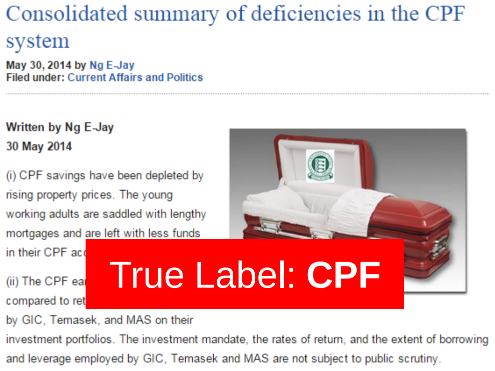
WHY SINGAPOREANS CAN'T SAVE ENOUGH IN OUR CPF: LEARNING FROM THE IPS FORUM

Post date: 31 Jul 2014 - 6:00pm

it
F is
m

Institute of Policy Studies). Their presentations will give you a very good insight as to why the CPF is inasequate for Singaporeans' retirement and in the problems highlighted, you can see the solutions inherent in them as well.

True Label: **CPF**



Consolidated summary of deficiencies in the CPF system

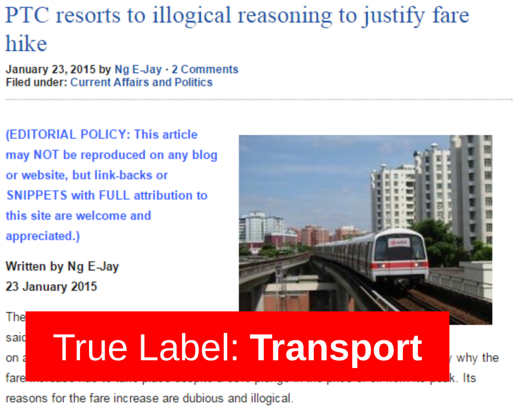
May 30, 2014 by Ng E-Jay
Filed under: Current Affairs and Politics

Written by Ng E-Jay
30 May 2014

(i) CPF savings have been depleted by rising property prices. The young working adults are saddled with lengthy mortgages and are left with less funds in their CPF ac

(ii) The CPF ea
compared to ret
by GIC, Temasek, and MAS on their investment portfolios. The investment mandate, the rates of return, and the extent of borrowing and leverage employed by GIC, Temasek and MAS are not subject to public scrutiny.

True Label: **CPF**



PTC resorts to illogical reasoning to justify fare hike

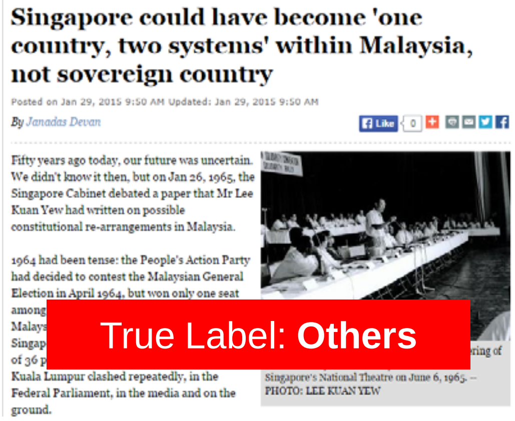
January 23, 2015 by Ng E-Jay · 2 Comments
Filed under: Current Affairs and Politics

(EDITORIAL POLICY: This article may NOT be reproduced on any blog or website, but link-backs or SNIPPETS with FULL attribution to this site are welcome and appreciated.)

Written by Ng E-Jay
23 January 2015

The
said
on a
why the
fare
k. Its
reasons for the fare increase are dubious and illogical.

True Label: **Transport**



Singapore could have become 'one country, two systems' within Malaysia, not sovereign country

Posted on Jan 29, 2015 9:50 AM Updated: Jan 29, 2015 9:50 AM

By Janadas Devan

Like 0

Fifty years ago today, our future was uncertain. We didn't know it then, but on Jan 26, 1965, the Singapore Cabinet debated a paper that Mr Lee Kuan Yew had written on possible constitutional re-arrangements in Malaysia.

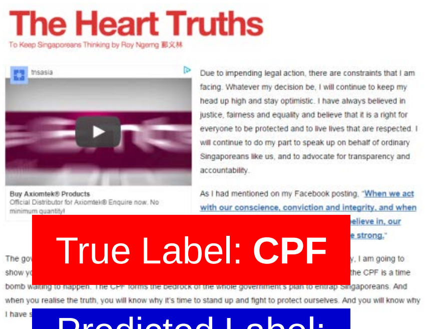
1964 had been tense: the People's Action Party had decided to contest the Malaysian General Election in April 1964, but won only one seat
among
Malays
Singap
of 36 p
Kuala Lumpur clashed repeatedly, in the Federal Parliament, in the media and on the ground.

ring of
Singapore's National Theatre on June 6, 1965. – PHOTO: LEE KUAN YEW

True Label: **Others**

Accuracy of Classifier

- Text with predicted labels assigned by classifier



The Heart Truths

To Keep Singaporeans Thinking by Roy Ngerng 鄞义林

tnsasia

Buy Axiomtek® Products
Official Distributor for Axiomtek® Enquire now. No minimum quantity!

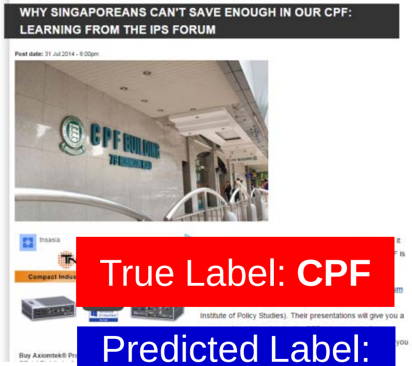
Due to impending legal action, there are constraints that I am facing. Whatever my decision be, I will continue to keep my head up high and stay optimistic. I have always believed in justice, fairness and equality and believe that it is a right for everyone to be protected and to live lives that are respected. I will continue to do my part to speak up on behalf of ordinary Singaporeans like us, and to advocate for transparency and accountability.

As I had mentioned on my Facebook posting, "When we act with our conscience, conviction and integrity, and when believe in, our e strong."

The go y, I am going to show yo the CPF is a time bomb waiting to happen. The CPF forms the bedrock of the whole government's plan to entrap Singaporeans. And when you realise the truth, you will know why it's time to stand up and fight to protect ourselves. And you will know why I have s

True Label: **CPF**

Predicted Label: **CPF**



WHY SINGAPOREANS CAN'T SAVE ENOUGH IN OUR CPF: LEARNING FROM THE IPS FORUM

Post date: 31 Jul 2014 - 6:00pm

tnsasia

Compact Indus

Buy Axiomtek® Pr

it

F is

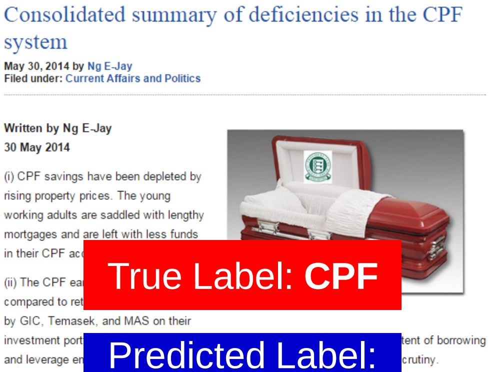
um

Institute of Policy Studies). Their presentations will give you a

you

True Label: **CPF**

Predicted Label: **CPF**



Consolidated summary of deficiencies in the CPF system

May 30, 2014 by Ng E-Jay
Filed under: Current Affairs and Politics

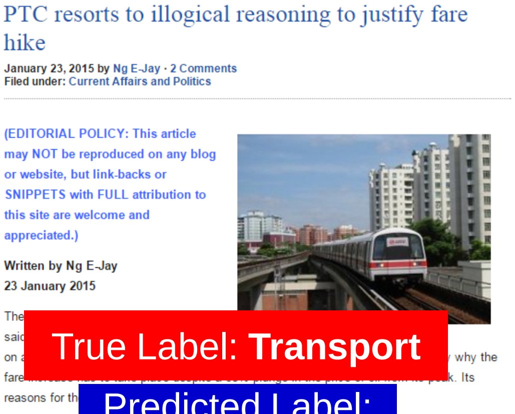
Written by Ng E-Jay
30 May 2014

(i) CPF savings have been depleted by rising property prices. The young working adults are saddled with lengthy mortgages and are left with less funds in their CPF ac

(ii) The CPF ea compared to ret by GIC, Temasek, and MAS on their investment port tent of borrowing and leverage en crutiny.

True Label: **CPF**

Predicted Label: **NOT CPF**



PTC resorts to illogical reasoning to justify fare hike

January 23, 2015 by Ng E-Jay · 2 Comments
Filed under: Current Affairs and Politics

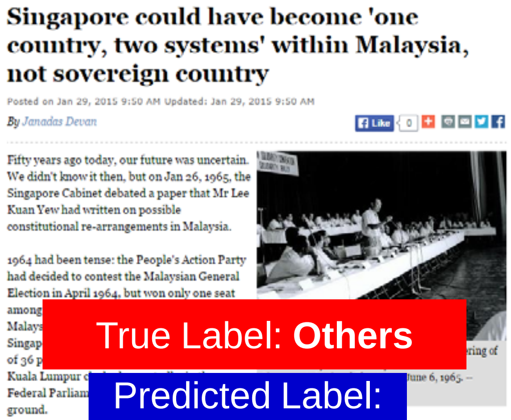
(EDITORIAL POLICY: This article may NOT be reproduced on any blog or website, but link-backs or SNIPPETS with FULL attribution to this site are welcome and appreciated.)

Written by Ng E-Jay
23 January 2015

The said on a why the fare k. Its reasons for th

True Label: **Transport**

Predicted Label: **CPF**



Singapore could have become 'one country, two systems' within Malaysia, not sovereign country

Posted on Jan 29, 2015 9:50 AM Updated: Jan 29, 2015 9:50 AM

By Janadas Devan

Fifty years ago today, our future was uncertain. We didn't know it then, but on Jan 26, 1965, the Singapore Cabinet debated a paper that Mr Lee Kuan Yew had written on possible constitutional re-arrangements in Malaysia.

1964 had been tense: the People's Action Party had decided to contest the Malaysian General Election in April 1964, but won only one seat among Malays Singap of 36 p Kuala Lumpur c Federal Parliam ground.

ering of June 6, 1965. –

True Label: **Others**

Predicted Label: **CPF**

- Accuracy measures:

$$\text{Precision} = \frac{2}{4} = \frac{1}{2} \qquad \text{Recall} = \frac{2}{3} \qquad \text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = 0.57$$

Topic Classification

- Multi-class classification
  - Given a blog post, assign it with one of the following topic labels.
- Topic labels:
  - Censorship
  - CPF
  - Foreigners
  - Inequality
  - LGBT

Topic Classification

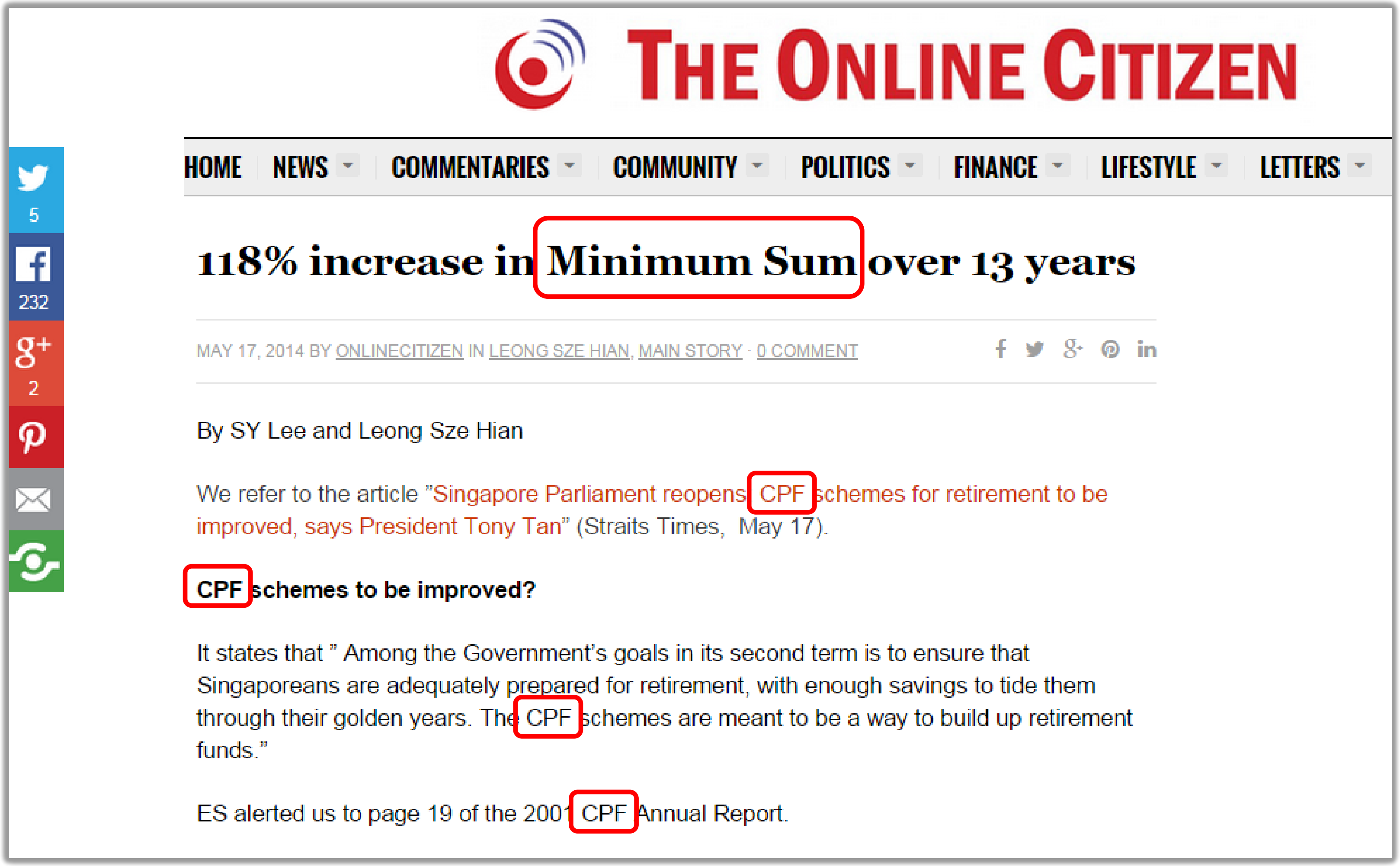
- Dataset
  - Select any posts containing specific keywords (in the table below) from The Online Citizen and The Real Singapore blogs.
  - Exclude short posts: word count ≤ 50

Topics	**Keywords (Noisy Labels)**	**# Instances**
Censorship	Censorship, freedom of speech, freedom of expression	1,189
CPF	CPF, central provident fund, retirement	3,605
Foreigners	Anti-foreigners, influx of foreigners, xenophobia, xenophobic	502
Inequality	Income distribution, income inequality, gini coefficient	212
LGBT	377a, gay rights	213

5721 blog posts

Example Blog Post

Topic Label: CPF



118% increase in Minimum Sum over 13 years

MAY 17, 2014 BY ONLINECITIZEN IN LEONG SZE HIAN, MAIN STORY - 0 COMMENT

By SY Lee and Leong Sze Hian

We refer to the article "Singapore Parliament reopens CPF schemes for retirement to be improved, says President Tony Tan" (Straits Times, May 17).

CPF schemes to be improved?

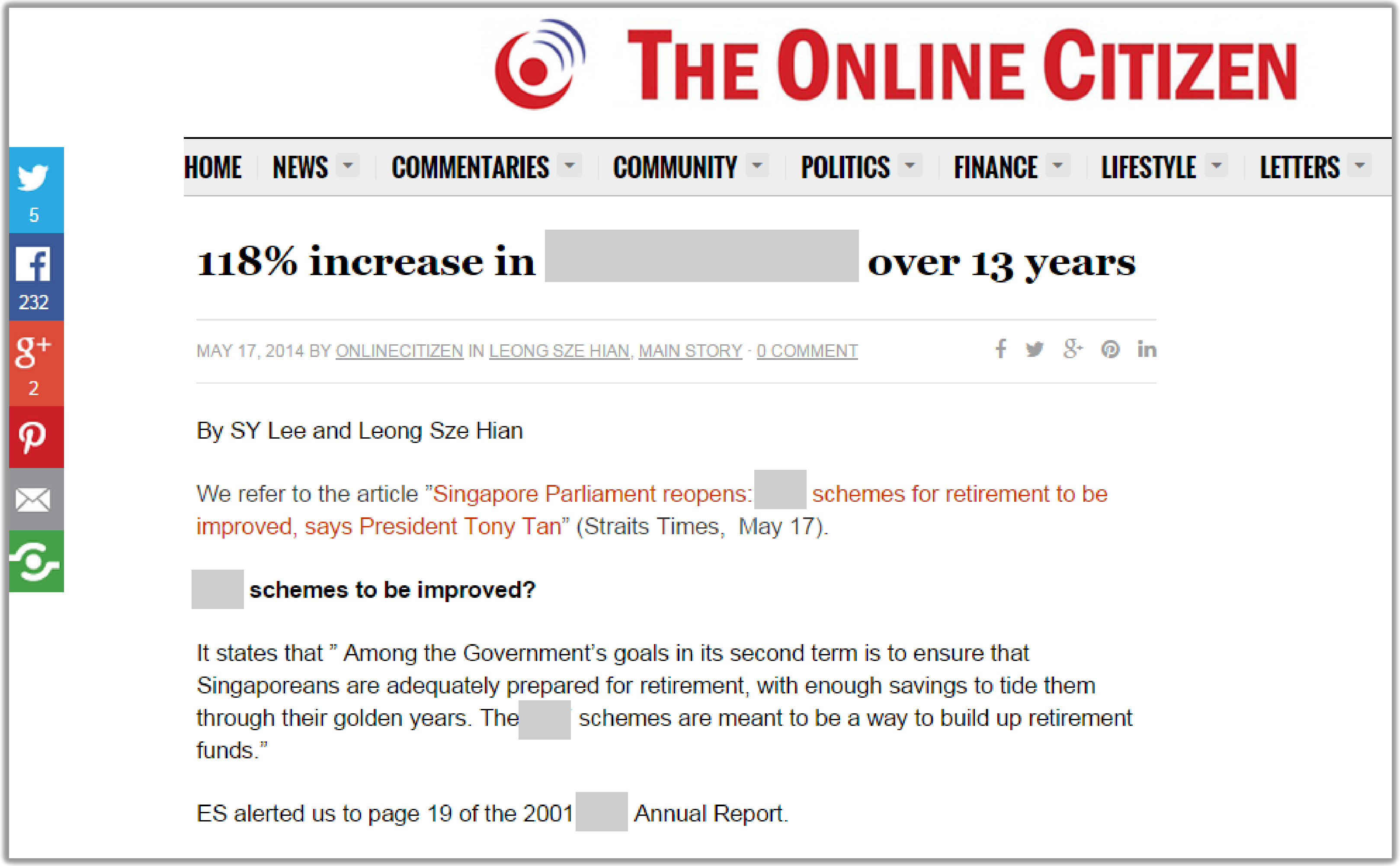
It states that " Among the Government's goals in its second term is to ensure that Singaporeans are adequately prepared for retirement, with enough savings to tide them through their golden years. The CPF schemes are meant to be a way to build up retirement funds."

ES alerted us to page 19 of the 2001 CPF Annual Report.

Example Blog Post

LARC LIVING ANALYTICS RESEARCH CENTRE

Predict the Topic Label?



THE ONLINE CITIZEN

HOME | NEWS | COMMENTARIES | COMMUNITY | POLITICS | FINANCE | LIFESTYLE | LETTERS

118% increase in [redacted] over 13 years

MAY 17, 2014 BY ONLINECITIZEN IN LEONG SZE HIAN, MAIN STORY · 0 COMMENT

By SY Lee and Leong Sze Hian

We refer to the article "Singapore Parliament reopens: [redacted] schemes for retirement to be improved, says President Tony Tan" (Straits Times, May 17).

[redacted] schemes to be improved?

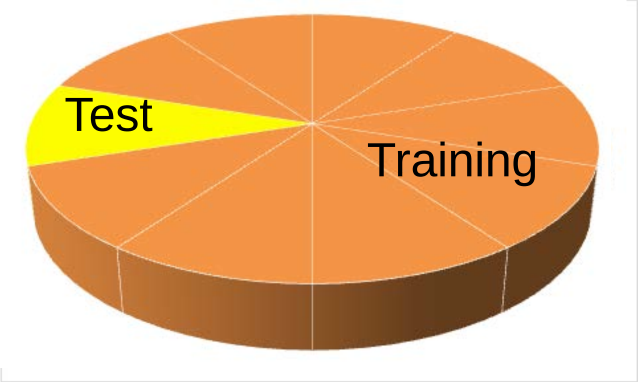
It states that " Among the Government's goals in its second term is to ensure that Singaporeans are adequately prepared for retirement, with enough savings to tide them through their golden years. The [redacted] schemes are meant to be a way to build up retirement funds."

ES alerted us to page 19 of the 2001 [redacted] Annual Report.

Topic Classification
- Experiments

- Preprocessing
  - Remove noisy labels from blog posts before training
    E.g., Censorship, freedom of speech, freedom of expression are assumed to be dropped by blog posts about Censorship
  - Remove common stop words

- Results
  - Labeled data: 5 classes, 5721 blog posts
  - Features: Word unigrams, TF-IDF weighting
  - 10-Fold cross validation
  - SGD (loss=hinge, alpha = 0.0001):
    F1 = 0. 84
  - Logistic regression (penalty=L1, C=7):
    F1 = 0.83



Class	Top-50 features
Censorship	coverage, community, freedom, expression, censor, right, source, audience, liberty, jailed, licensing, bold, censored, assembly, allegation, art, imposed, democracy, mda, exercise, violence, free, independent, libel, trs, controversial, military, movie, allowing, freely, censoring, broadcasting, judge, platform, alternative, journalist, debated, breach, constitution, allegedly, material, artist, reflected, democratic, review, produced, islam, oppression, proxy, swearing
CPF	medishield, saving, fund, age, contribution, retire, health, hdb, roy, pension, scheme, ngerng, healthcare, generation, monies, sum, pioneer, ageing, medisave, welfare, gic, appointment, hui, returnourcpf, benefit, withdrawal, ymca, retired, withdraw, lose, insurance, heckling, ratio, dream, parliament, retrenched, investment, loan, breakdown, sport, payment, nparks, batam, ite, board, meeting, pmet, managed, teo, senior
Foreigners	sentiment, nationality, anti, talent, racist, infrastructure, indian, racism, prc, xenophobes, orchard, population, integration, born, singaporean, overcrowding, philippine, clan, accent, mainland, overpopulated, race, policy, deeply, national, threatening, bigot, cultural, hygiene, hatred, migration, insensitive, rhetoric, permanent, island, alarming, migrant, space, rapidly, generous, prs, europe, displacement, independence, intolerance, demonstrated, flame, bred, xenophobe, armed
Inequality	mobility, inequality, measure, growth, income, severe, eatery, disparity, widened, amidst, meritocracy, firm, aside, bargaining, comrade, billionaire, hike, path, globalization, gap, equality, predicted, glcs, equal, rose, rest, extreme, data, respectable, yah, chip, scored, economy, union, business, poverty, fuel, indicator, complain, china, redistribution, debate, crime, innovate, richer, reflect, raffle, enterprise, median, negotiating
LGBT	gay, homosexual, repeal, penal, lgbt, homosexuality, pink, heterosexual, illuminated, constitutional, acceptance, code, law, khong, church, pastor, sexual, conservative, movement, position, sexuality, repealed, partner, retention, marriage, demonstration, belief, challenge, christian, equal, explicitly, nominated, lifestyle, pray, overturn, criminalisation, lawrence, discrimination, diversity, hpb, divided, nmp, faith, moral, retaining, alex, badge, norm, bright, religion

Political Orientation Classification

- Binary classification: Given a blog post, assign it with one of the following labels.
  - Anti-PAP vs pro-PAP (posts in defense of or not criticizing the PAP)

- Dataset Construction
  - Exclude short posts: ≤ 50 words
  - **Anti-PAP posts:**
    PAP posts from `www.theonlinecitizen.com`

    Assumption: Any posts from The Online Citizen containing “PAP” to be anti-PAP.
  - **Pro-PAP posts:**
    All posts from `sghardtruth.com`, `sggeneralelections2016.blogspot.sg`,
  - **Number of posts: Anti-PAP = 421, pro-PAP = 421**

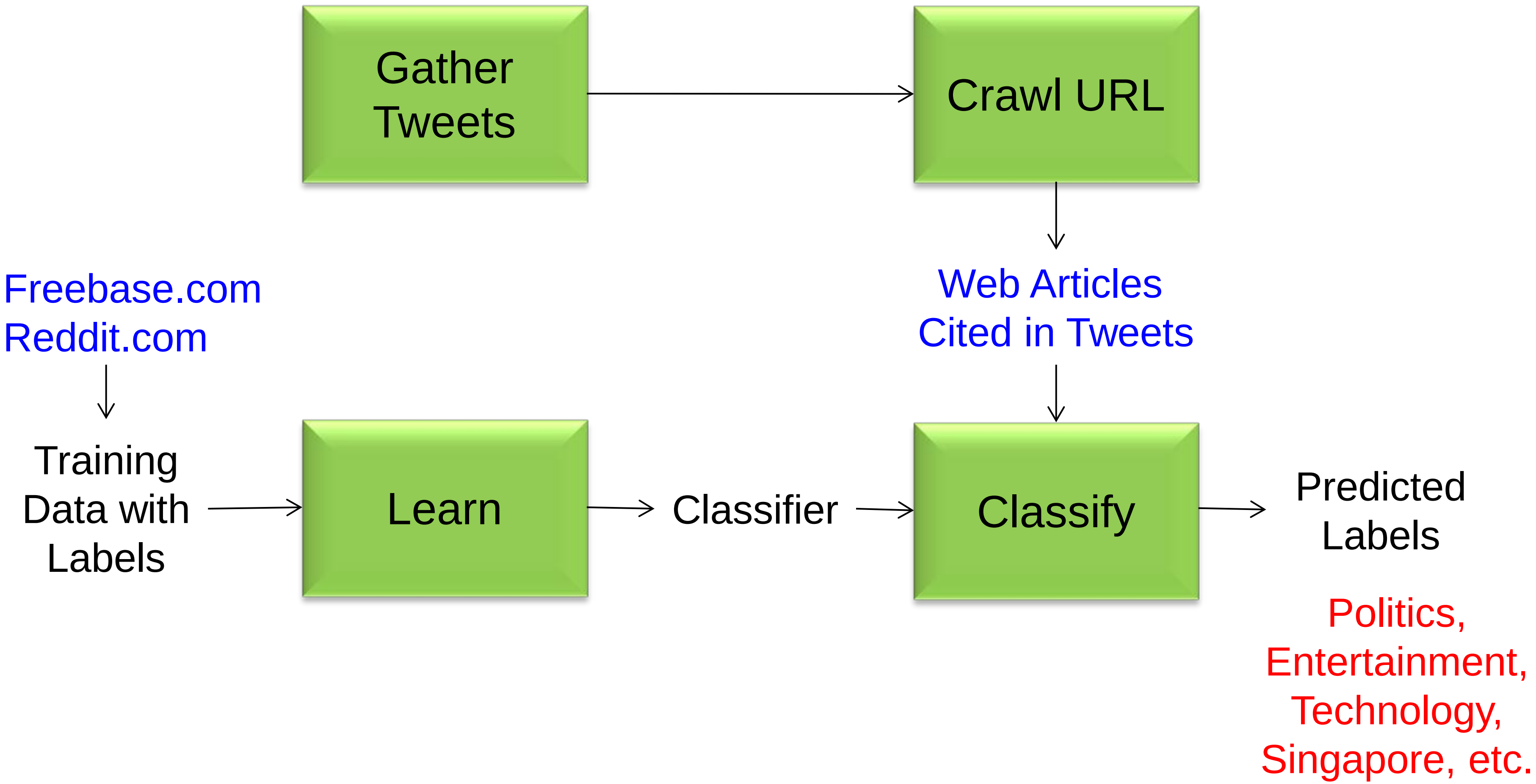
Political Orientation Classification

- 10-Fold cross validation
  - Labeled data:
    Anti-PAP = 421 blog posts, Pro-PAP = 421 blog post
  - Features: Word unigrams, TF-IDF weighting
  - SGD (loss=log, alpha=0.0001):
    F-1 (anti-pap)= 0.83

- Results: 75% training and 25% test
  - SGD (loss=log, alpha=0.0001):
    F-1 (pro-pap)= 0.59
    F-1 (anti-pap)= 0.77
  - Average F-1 = 0.68

- It is harder to predict the Pro-PAP label.

Topic Profiling



Gather Tweets

Crawl URL

Web Articles Cited in Tweets

Freebase.com
Reddit.com

Training Data with Labels

Learn

Classifier

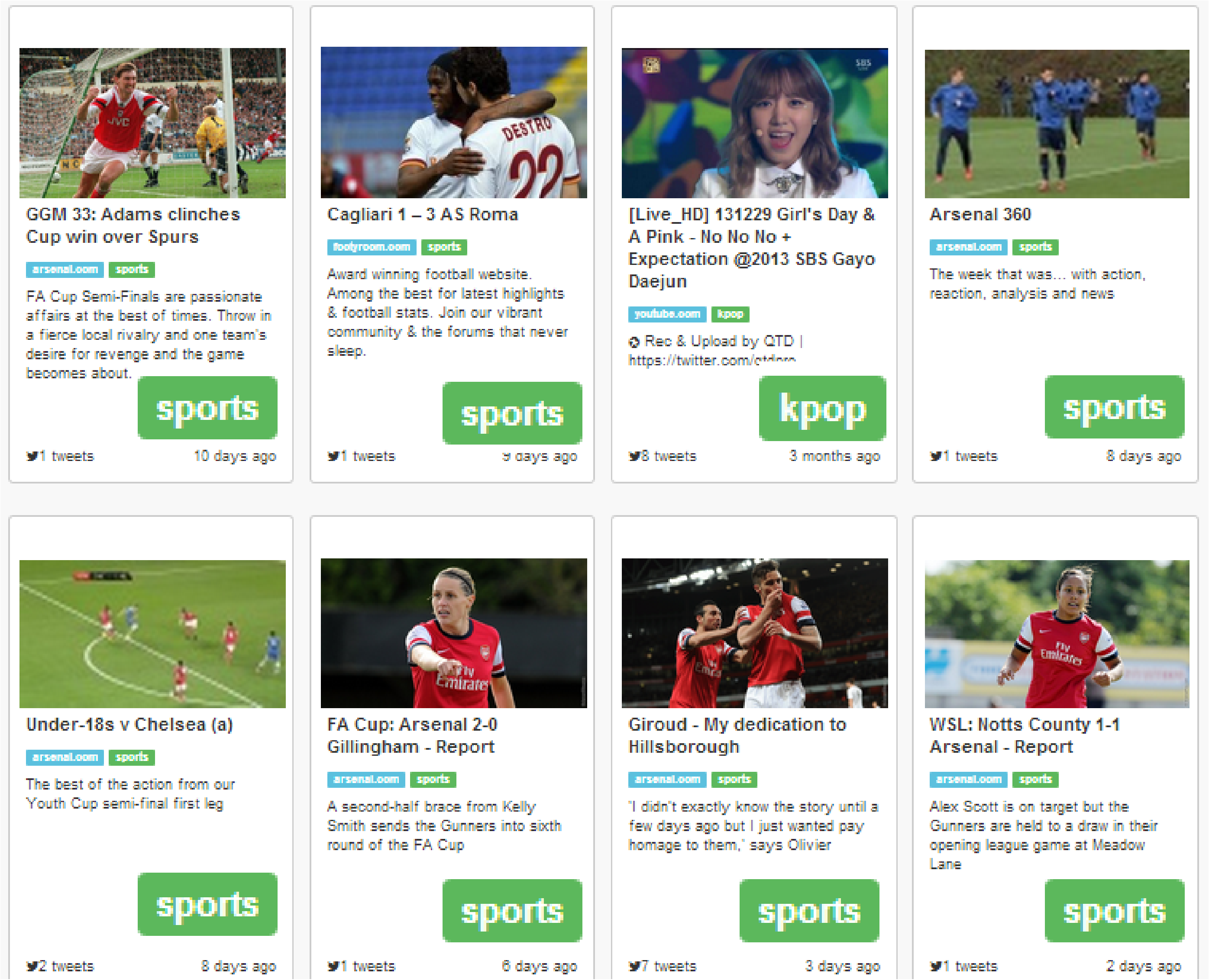
Classify

Predicted Labels

Politics, Entertainment, Technology, Singapore, etc.

Classy (http://research.larc.smu.edu.sg/classy/)

Latest Shared Web Pages



GGM 33: Adams clinches Cup win over Spurs

arsenal.com sports

FA Cup Semi-Finals are passionate affairs at the best of times. Throw in a fierce local rivalry and one team's desire for revenge and the game becomes about.

sports

1 tweets 10 days ago

Cagliari 1 – 3 AS Roma

footyroom.com sports

Award winning football website. Among the best for latest highlights & football stats. Join our vibrant community & the forums that never sleep.

sports

1 tweets 9 days ago

[Live_HD] 131229 Girl's Day & A Pink - No No No + Expectation @2013 SBS Gayo Daejun

youtube.com kpop

Rec & Upload by QTD | https://twitter.com/ctdorn

kpop

8 tweets 3 months ago

Arsenal 360

arsenal.com sports

The week that was... with action, reaction, analysis and news

sports

1 tweets 8 days ago

Under-18s v Chelsea (a)

arsenal.com sports

The best of the action from our Youth Cup semi-final first leg

sports

2 tweets 8 days ago

FA Cup: Arsenal 2-0 Gillingham - Report

arsenal.com sports

A second-half brace from Kelly Smith sends the Gunners into sixth round of the FA Cup

sports

1 tweets 6 days ago

Giroud - My dedication to Hillsborough

arsenal.com sports

'I didn't exactly know the story until a few days ago but I just wanted pay homage to them,' says Olivier

sports

7 tweets 3 days ago

WSL: Notts County 1-1 Arsenal - Report

arsenal.com sports

Alex Scott is on target but the Gunners are held to a draw in their opening league game at Meadow Lane

sports

1 tweets 2 days ago

Conclusion

- Classification is a very useful text analytics technique to complement human labeling efforts.
  - Scalable to classify large number of blog posts.
  - Realtime when integrated with data crawlers (e.g., Classy).
  - Minimum Heisenberg effect compared with user survey
- Examples:
  - Topic classification
  - Political orientation classification
- Classification however are not always easy:
  - Availability of labeled data (noisy labeling by keyword/blog)
  - Imbalanced data (e.g., Pro-PAP blog posts)

Interesting Directions

- Realtime topic directory of blog posts:
  - Realtime crawling of blog posts
  - Realtime classification of blog posts
  - New topics detection
- Research topics:
  - Lurking users in social media
  - Extraction of contextual information of content: people, organizations, locations, date & time
  - Non-English text
  - Human computation: Games for labeling?

Thank you

LARC

LIVING ANALYTICS
RESEARCH CENTRE