Learning from Microblog data – methods and challenges

Thashmee Karunaratne

Departement of Computer and Systems Sciences

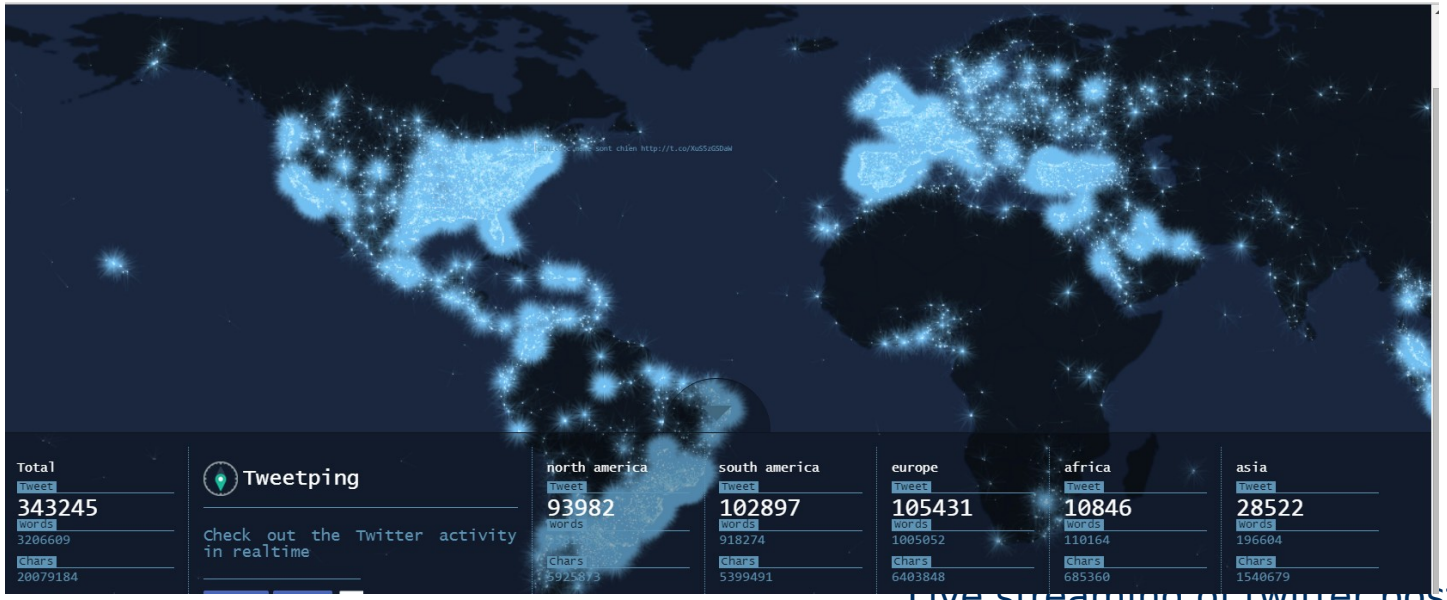
Stockholm University

Microblog data

Twitter posts, forums, Tumblr tumbles etc.

- free text
- creative language “microbloggers are typically exempted from all of the established protocols for writing”
- highly contextualized
- informal
- limited length (mostly 140 characters)

Twitter



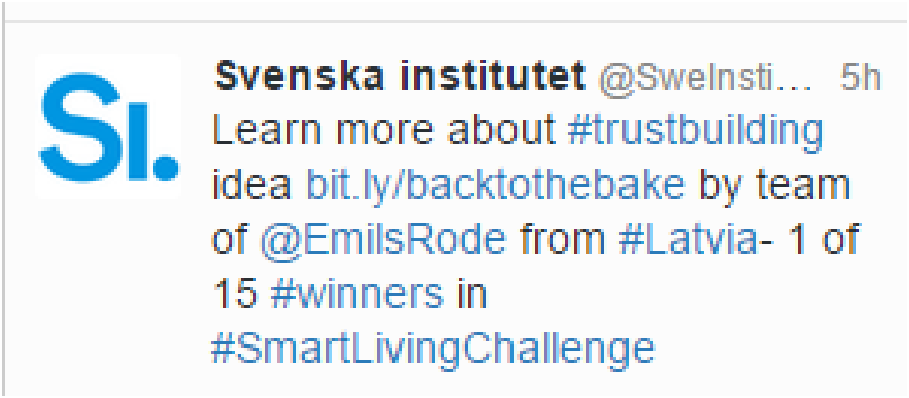
Live streaming of twitter posts (http://tweetping.net/)

Popular

- openness
- clean and well-documented API
- good developer tooling
- tweets happen at the "speed of thought”
- available for consumption as they happen in near real time

Stockholm University

Twitter feeds



https://tweetdeck.twitter.com/#

- Words (and numbers)
- Emoticons :), : (etc…
- Hashtags
- Negations
- @, RT, d

Stockholm University

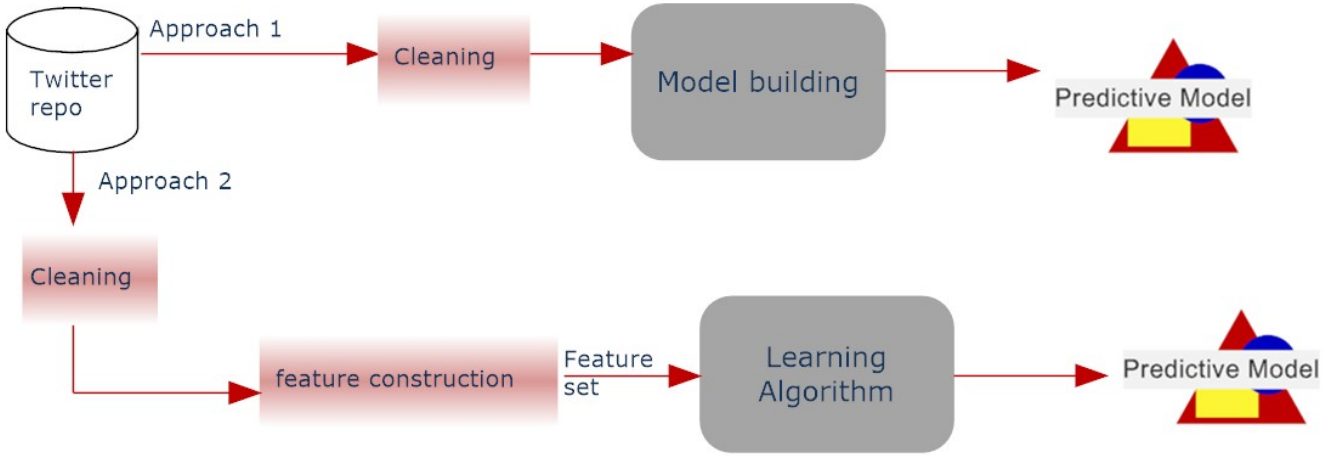
Twitter data

- Sentiment analysis
  - Supervised (manually labelled)
  - Distantly supervised(noisy labelling with emoticons)
  - Unsupervised (Clustering)
- Ranking popular topics
  - Clustering
  - Pattern discovery (largely on topic detection)

Document classification methods doesn't perform well

- Informal language – cleaning looses information
- Short documents – very sparse feature vectors

Supervised Sentiment analysis



1. Probabilistic Learning Models – assign probability to word sequences (Liu et.al, 2013)
2. Using feature set
   - unigrams, bigrams and emotions (":)" and ":(" etc.) (Go et.al, 2009)
   - Binary weighting of words in specified dictionary of +ve –ve emotions (Twitratr, 2013)

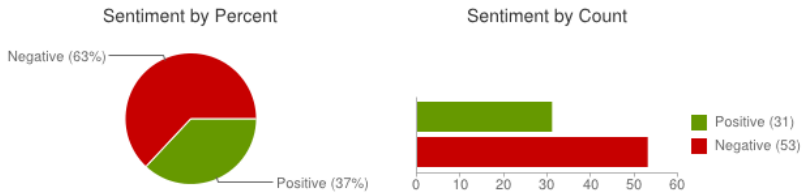
Stockholm University

Sentiment140

Tweet 763 Like 700 +1 180

iphone English Search

Sentiment analysis for iphone



Tweets about: iphone

Jahmya: I regret getting the **iPhone** 6 in gold because everyone has the **iPhone** 6 or 6 plus in gold and I hate having the same thing as everyone.
Posted: 24 seconds ago

donald_i: @mgdm samsung smart TV, then **iphone**. Might be due to rubbish broadband, but not impressed it can manage the ads without difficulty!
Posted: 24 seconds ago

janocurlyharry: @godsgirl8494 no that hasn't happened to my **iPhone**
Posted: 26 seconds ago

KAPITALOFFICIAL: Phones going off. To much music biz, on my brain and after watching that Itv program I need to chill from this **Iphone**. Siri got me like ??
Posted: 27 seconds ago

harryarmpitz: @STYLATORARMY if you follow me ill set all the **iPhone** wallpapers at best buy to a picture of your profile ????
Posted: 32 seconds ago

godsgirl8494: My **iPhone** 6 is slower than my **iPhone** 5 and I have no apps or anything on It. & it takes forever to charge. Is this normal ?

http://www.sentiment140.com/search?query=iphone&hl=en

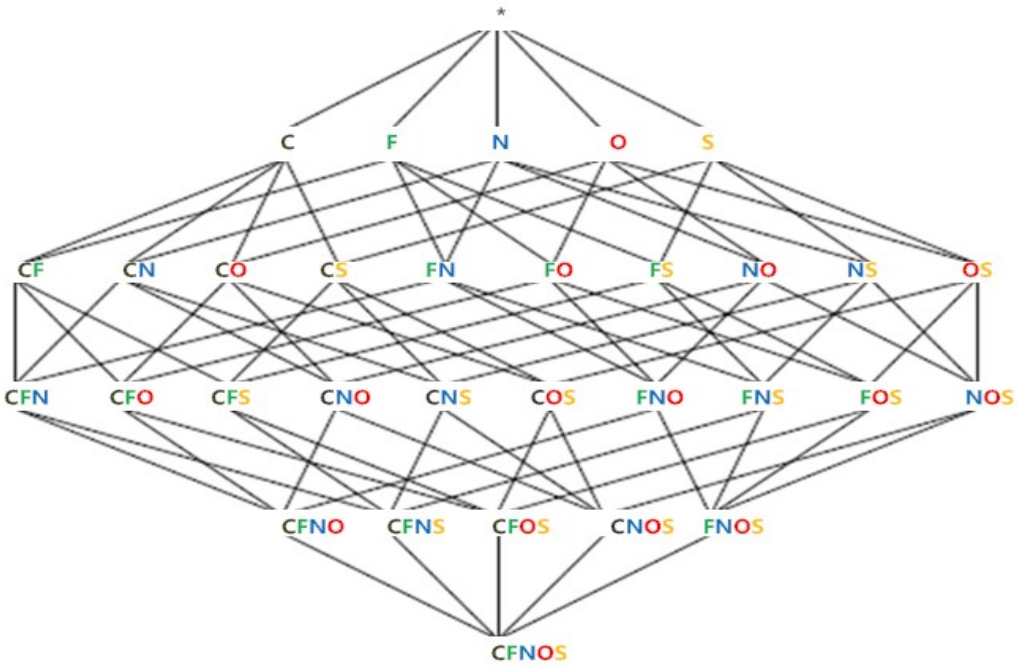
Stockholm University

Other patterns as feature sets

- Itemsets (entities)
- Sequences (entities with precedence order)
- Tree structures (hierarchical association among entities)
- Graphs (arbitrary association among entities)

Itemset and itemset mining

e.g, with 4 items – C, F, N, O, S



Pattern Selection

- Frequency (maximal, closed)
- Use of constraints to select which itemsets are to be selected

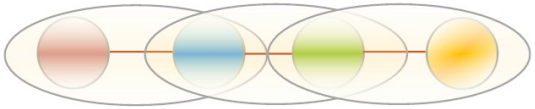
  (info gain, chi-square gini index, accuracy, laplace, convex hull etc.)

Stockholm University

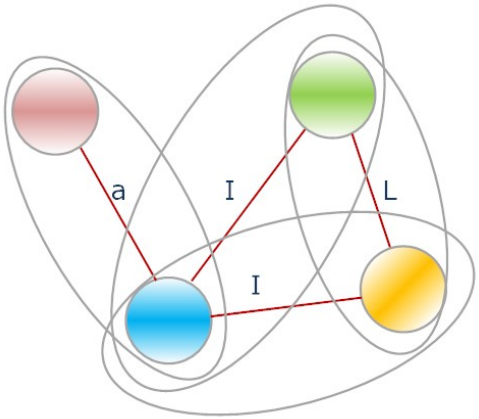
Representations

1. Each entity (words, numbers, emoticons, hashtags,@) in the twitter message as an item

2. Twitter message as a sequence – include n-grams



3. As a graph where nodes are entities and edges are relations between entities
   - Adjecent (a)
   - In the sentence (i)
   - In another sentence (n)
   - Link (l)



Simplifying complex structures – Edge List

Edge List

The edge list L of an undirected graph $G = (V, E, \lambda, \mu)$, where for any $e_k = (v_i, v_j) \in E$, there exists $e_k = (v_j, v_i)$, is defined as

$$L = \begin{cases} (\lambda(v_i), \lambda(v_j), \mu(e_k)) & \text{for all } v_i, v_j \in V \text{ such that } e_k = (v_i, v_j) \in E \\ & \text{and } \lambda(v_i) \prec \lambda(v_j) \\ \lambda(v_i) & v_i \text{ is not connected to any other node in the} \\ & \text{graph (no edges)} \end{cases}$$

We define each unit $(\lambda(v_i), \lambda(v_j), \mu(e_k))$, or, $(\lambda(v_i)$ (if v_i has no edges), to

Stockholm University

Efficiency and effectivenes

- Shown before:
  - Efficiency of the predictive models can be improved by using computationally less demanding forms than graphs, without affecting the predictive performance of resulting models

Thashmee Karunaratne and Henrik Boström. Can frequent itemset mining be efficiently and effectively used for learning from graph data? In *11th International Conference on Machine Learning and Applications*, pages 409-414. 2012. IEEE Computer Society.

Stockholm University

Experiments

Sentiment analysis datasets – manually labelled

Dataset	No. of Tweets	#Negative	#Neutral	#Positive	#Mixed	#Other	#Irrelevant
STS-Test	498	177	139	182	-	-	-
HCR	2,516	1,381	470	541	-	45	79
OMD	3,238	1,196	-	710	245	1,087	-
SS-Twitter	4,242	1,037	1,953	1,252	-	-	-
Sanders	5,513	654	2,503	570	-	-	1,786
GASP	12,771	5,235	6,268	1,050	-	218	-
WAB	13,340	2,580	3,707	2,915	-	420	3,718
SemEval	13,975	2,186	6,440	5,349	-	-	-

-only positive, negative and neutral instances

Pattern mining

- Represent the tweets as pattern sets
- Pattern mining
  - Constraint based method (CP)
  - Top k patterns
  - Different constraints: https://dtai.cs.kuleuven.be/CP4IM/download.php
- Patterns as features to any Machine Learning algorithm
- Baselines to compare
  - unigram, bigram based feature sets
  - Twitratr features

Thank you!

Stockholm University