Twitter Opinion Topic Model: Extracting Product Opinions from Tweets by Leveraging Hashtags and Sentiment Lexicon

Kar Wai Lim, Wray Buntine

3 October 2014

CIKM 2014

Outline

- Motivation
- Background and Previous Work
- Twitter Opinion Topic Model
- Experiments

Online Reviews

- Abundant

Online Reviews

- But...
  - Fake reviews are everywhere too.
    - Consumers trust reviews more than ads.
    - 1 star increase in Yelp = 5-9% increase in revenue *
    - 1 bad review => 30 customers loss
  - Cheaper compared to advertising.
  - Estimated about 30% of online reviews are fake.

* from Luca (2011)

Alternatives

- Opinions from Social Media
  - usually meant for friends and family.
  - Hence are usually truthful opinions.
  - People more willing to post social update than write a proper review.
  - Less targeted by malicious companies due to lower reach.

Problems

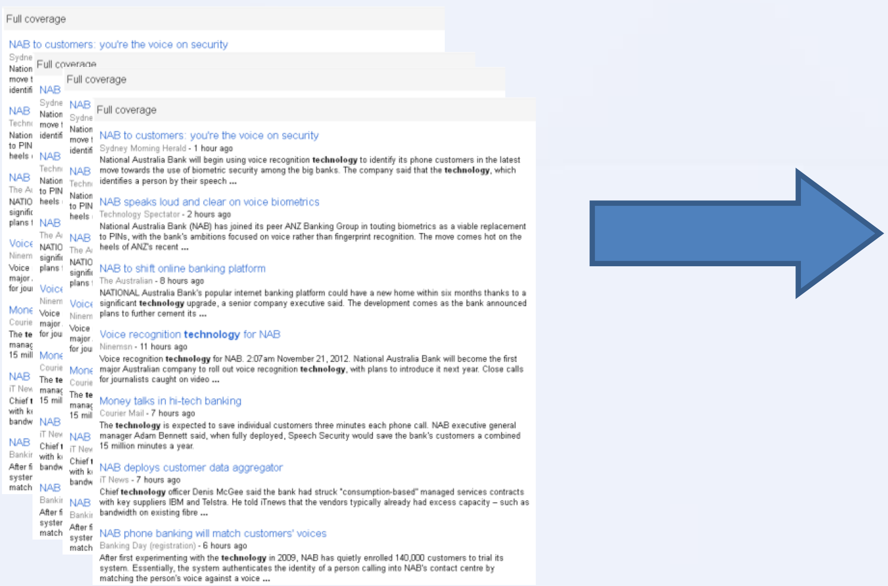
- Social updates tend to be short with little details.
- Improper language makes it harder to analyse with existing NLP approach.
- Sarcasm:

Outline

- ~~Motivation~~
- Background and Previous Work
- Twitter Opinion Topic Model
- Experiments

Background

- Aspect-based opinion mining



Example:

- Target: Sony, Microsoft, Nintendo...
- Aspect: Game consoles
  - PS4 – impressive...
  - XboxOne – cool...
  - Gameboy – retro...

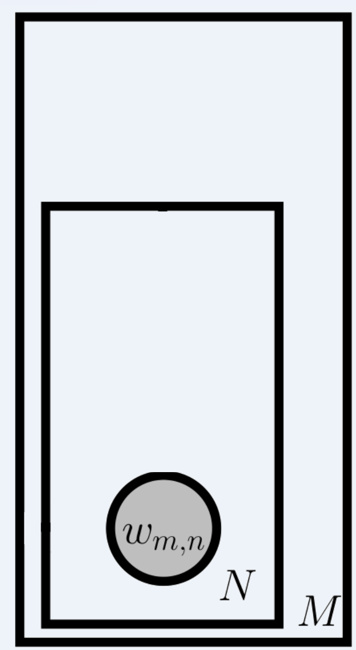
Background

- LDA-based models considered state-of-the-art for aspect-based opinion mining (Moghaddam, 2012).
- LDA is the simplest Bayesian topic model.
- Topic Model
  - assigns a categorical label (topic) to each word in each document.
  - Allow us to analyse the words of each topic,
  - and also topic composition of each document.

LDA

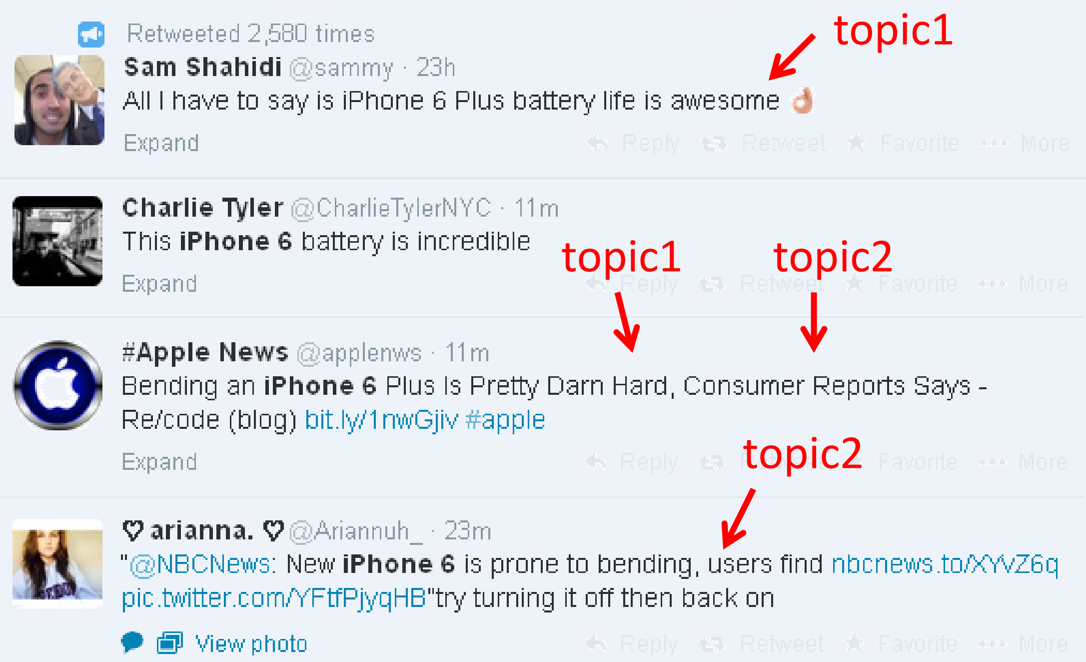
- LDA models the words in each document.
  - In our case a document is a tweet.

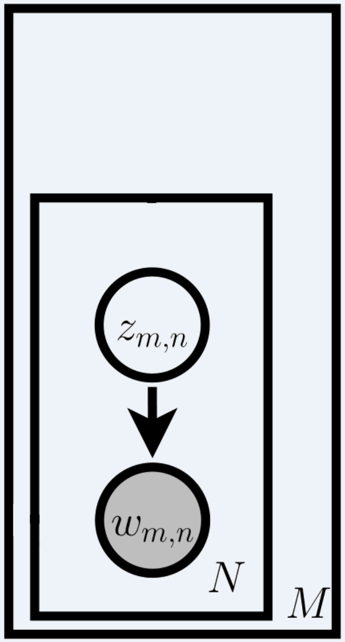
Plate Notation



LDA

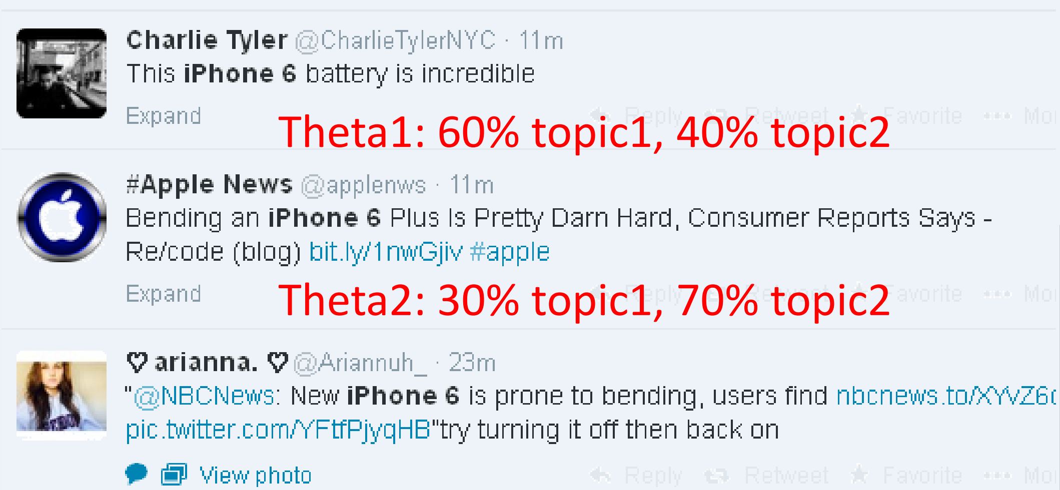
- LDA assigns a topic label to each word.
  - Topic label is latent (unobserved).

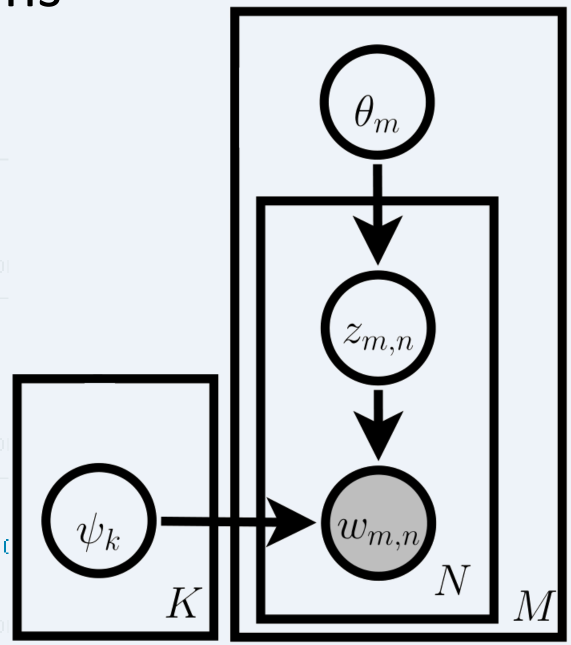




LDA

- Words and topics are generated from probability distributions.
  - Theta : document-topic distributions



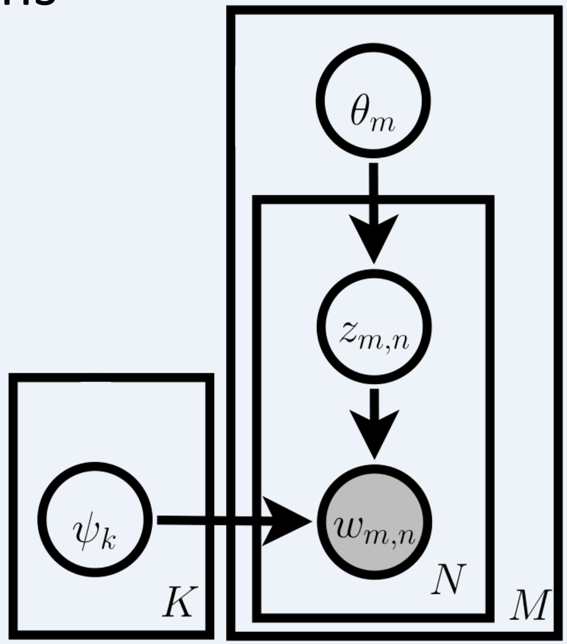


LDA

- Words and topics are generated from probability distributions.
  - Theta : document-topic distributions
  - Psi : topic-word distributions

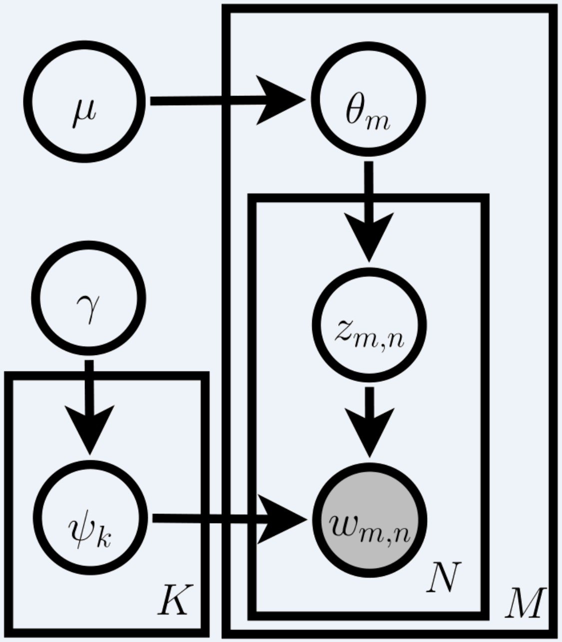
Psi1 : 10% “awesome”, 3% “hard”, etc...

Psi2 : 20% “users”, 5% “consumer reports”, etc...



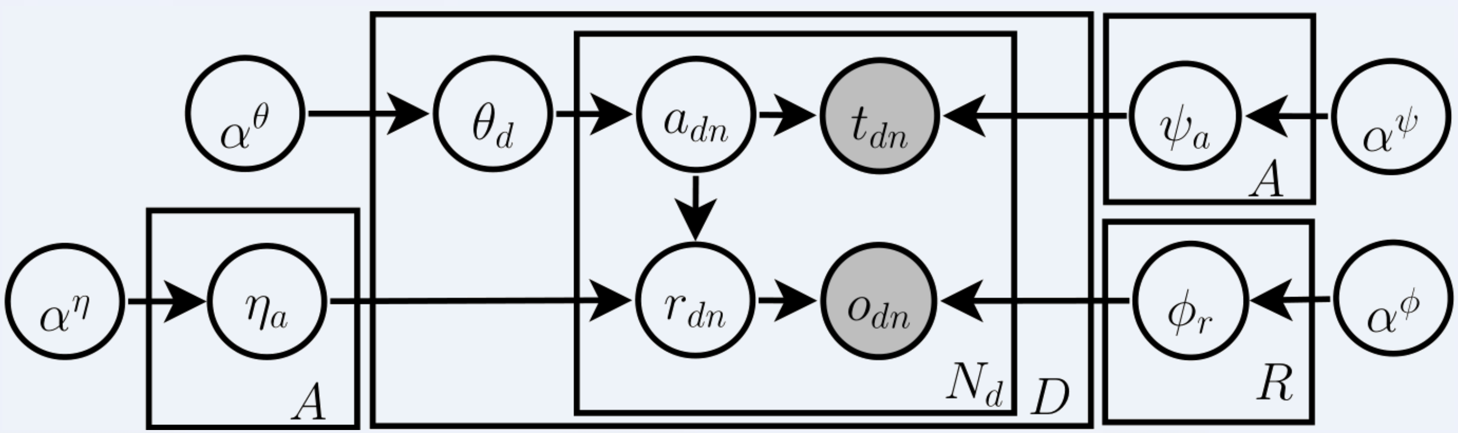
LDA

- Probability distributions are assigned Dirichlet priors (for LDA).
  - Can use other priors:
    - Hierarchical Dirichlet
    - Hierarchical Dirichlet Process
    - Pitman-Yor Process
- A flexible prior is important for learning.



ILDA

- Interdependent LDA (ILDA)
  - Extension of LDA for aspect-based opinion mining.



ILDA

- ILDA separates “target” and “opinion” words.

a : aspect/topic

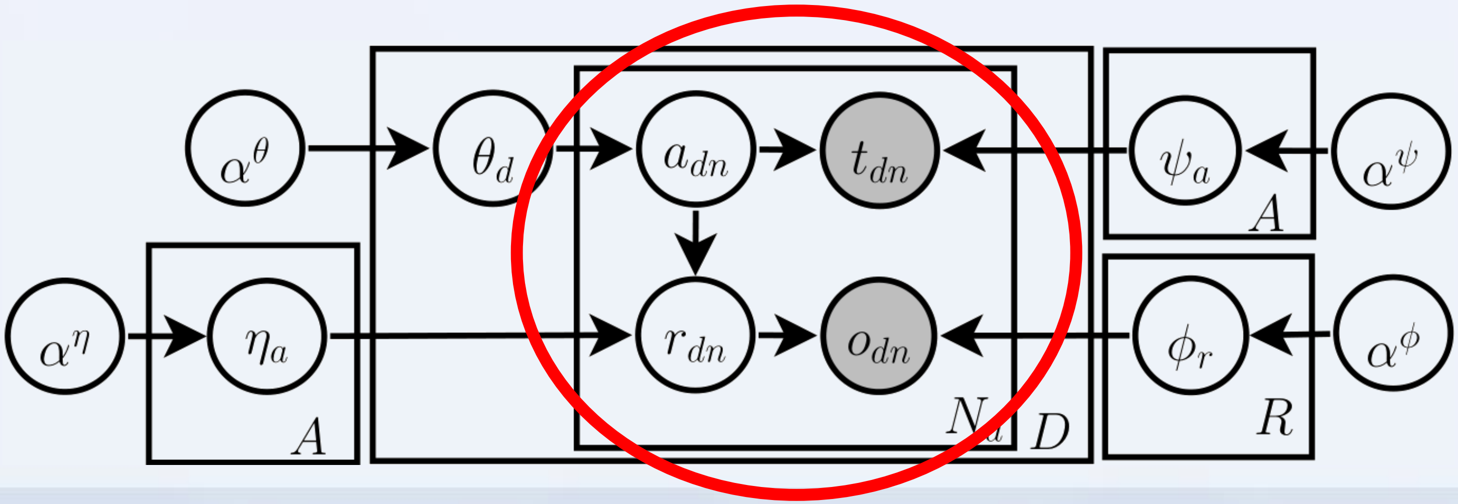
r : sentiment/rating

latent

t : target

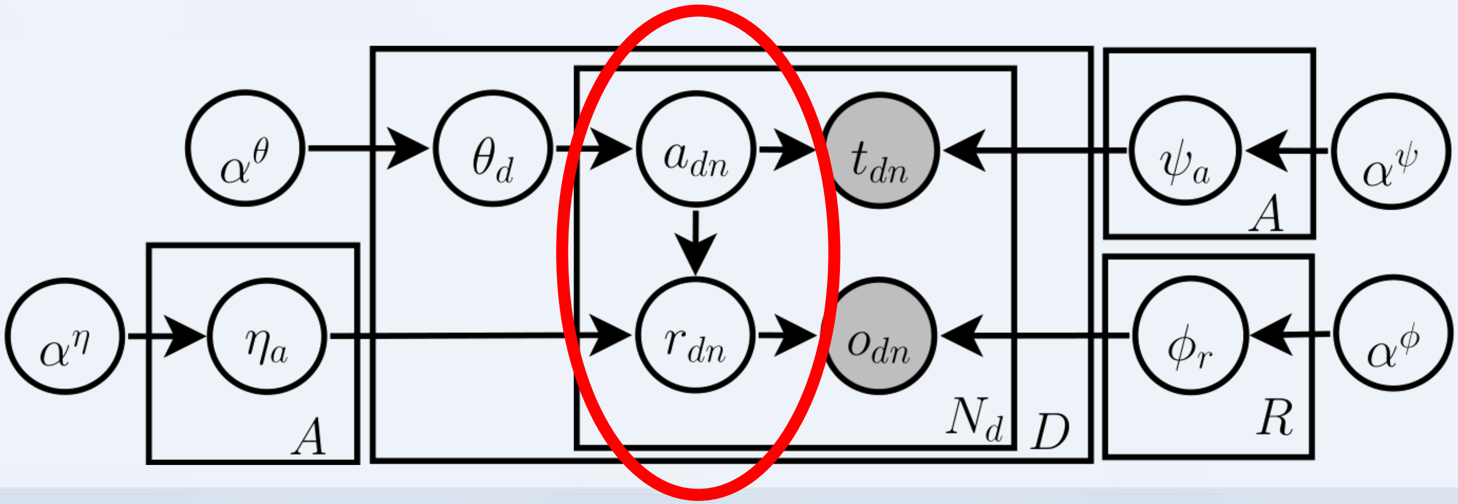
o : opinion

observed



ILDA

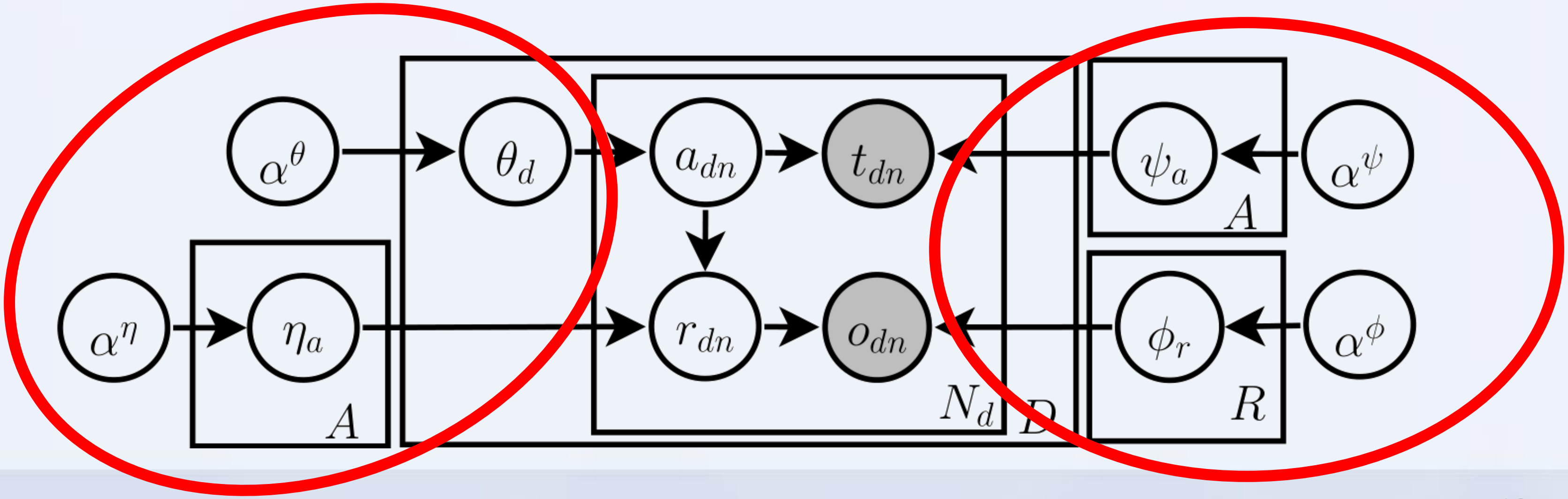
- ILDA models the sentiments of each aspects.
  - What is the proportion of positive sentiment for aspect “mobile phone”?



ILDA

- Theta : document-topic distributions
- Psi : aspect-target distributions
- Phi : sentiment-opinion distributions
- Eta : aspect-sentiment distributions

Alphas are
the priors



ILDA

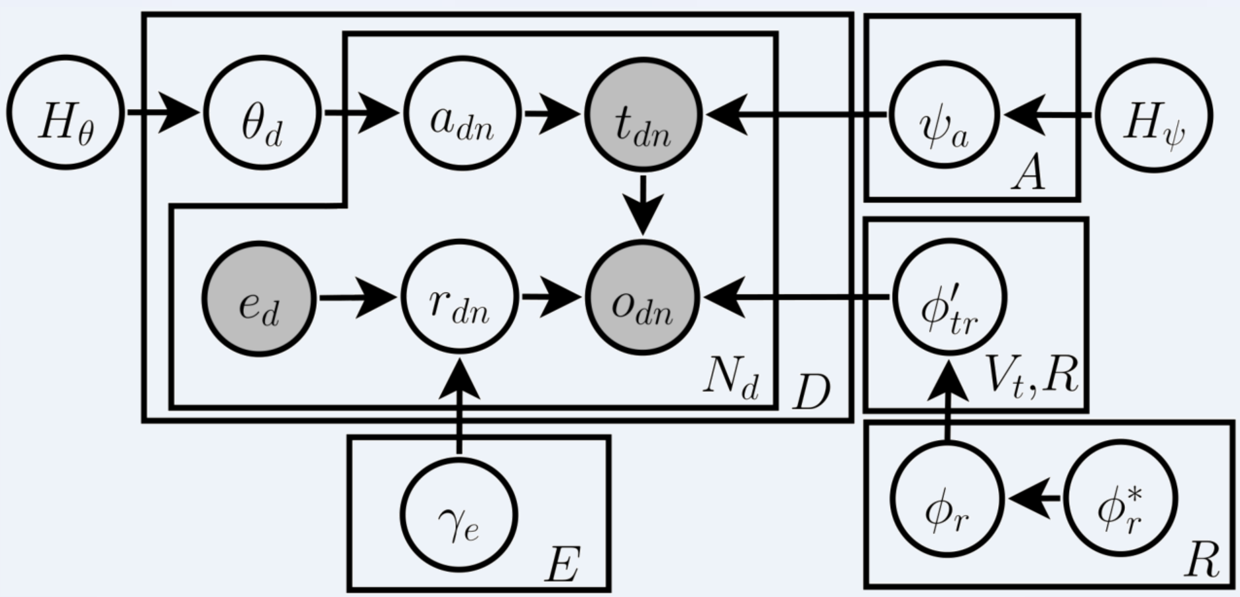
- Problem with ILDA
  - Sentiment/Rating is arbitrary.
    - Need to manually inspect and give them positive/neutral/negative labels.
  - Does not consider target-opinion interaction directly.
    - eg: "*short* camera quality" is plausible in ILDA.

Outline

- ~~Motivation~~
- ~~Background and Previous Work~~
- Twitter Opinion Topic Model
- Experiments

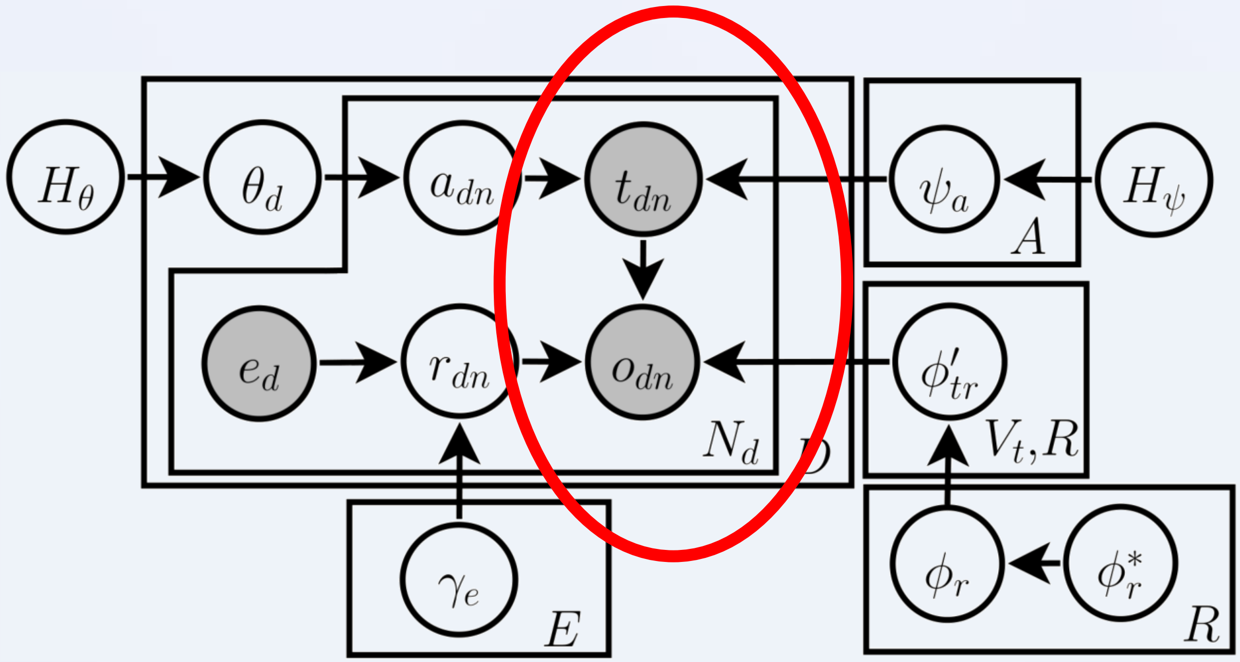
Twitter Opinion Topic Model (TOTM)

- Designed to extract opinions from tweets.
- Use state-of-the-art Bayesian non-parametric modelling - Hierarchical Pitman-Yor process



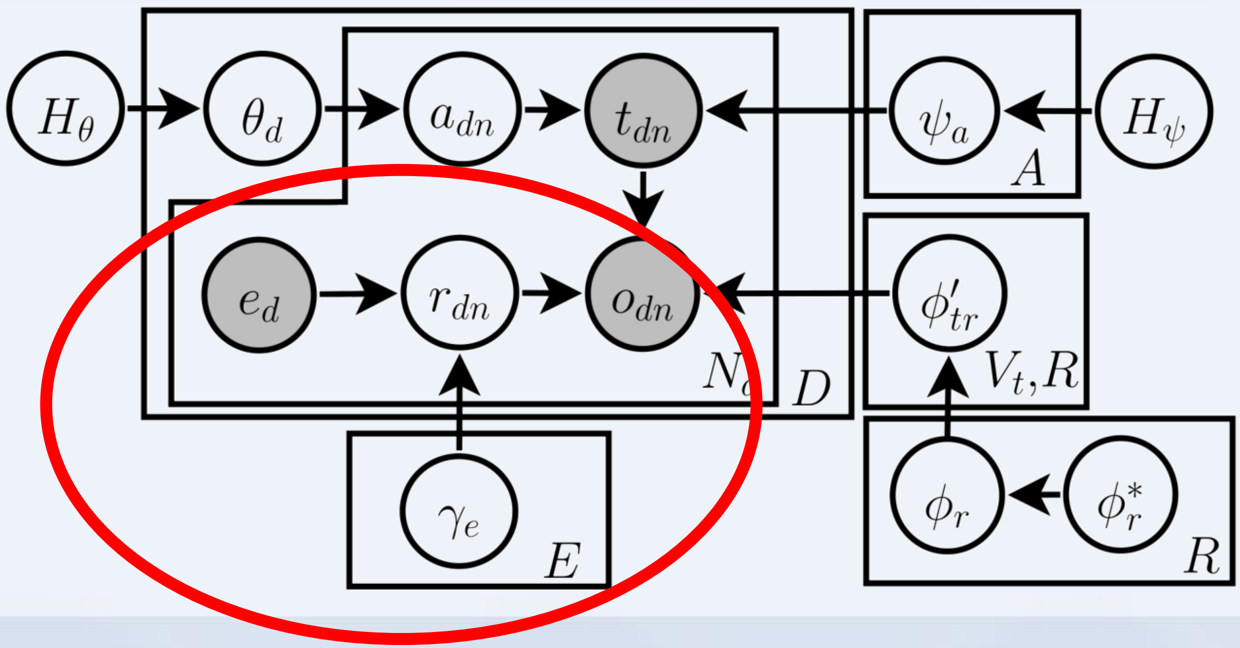
Twitter Opinion Topic Model (TOTM)

- Model target-opinion interaction directly.
  - Tasty burger is more likely than friendly burger.



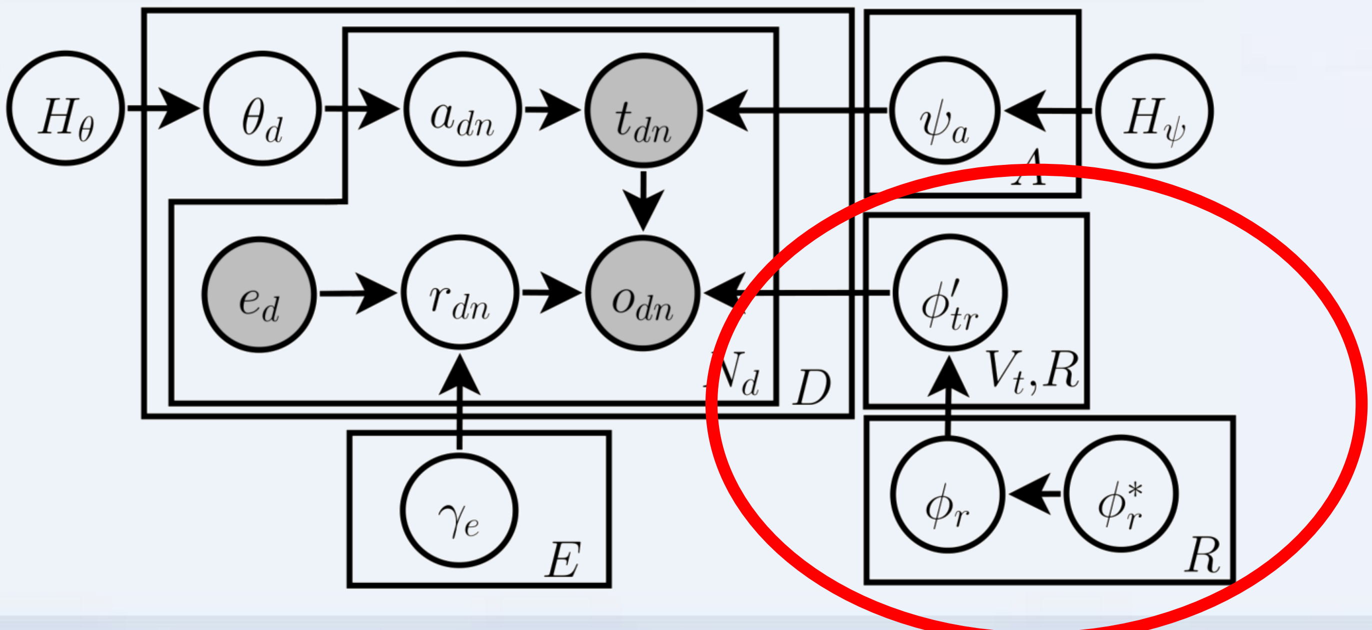
Twitter Opinion Topic Model (TOTM)

- Makes use of emoticons to learn sentiment.
  - Positive opinions tend to come with positive emoticon ☺



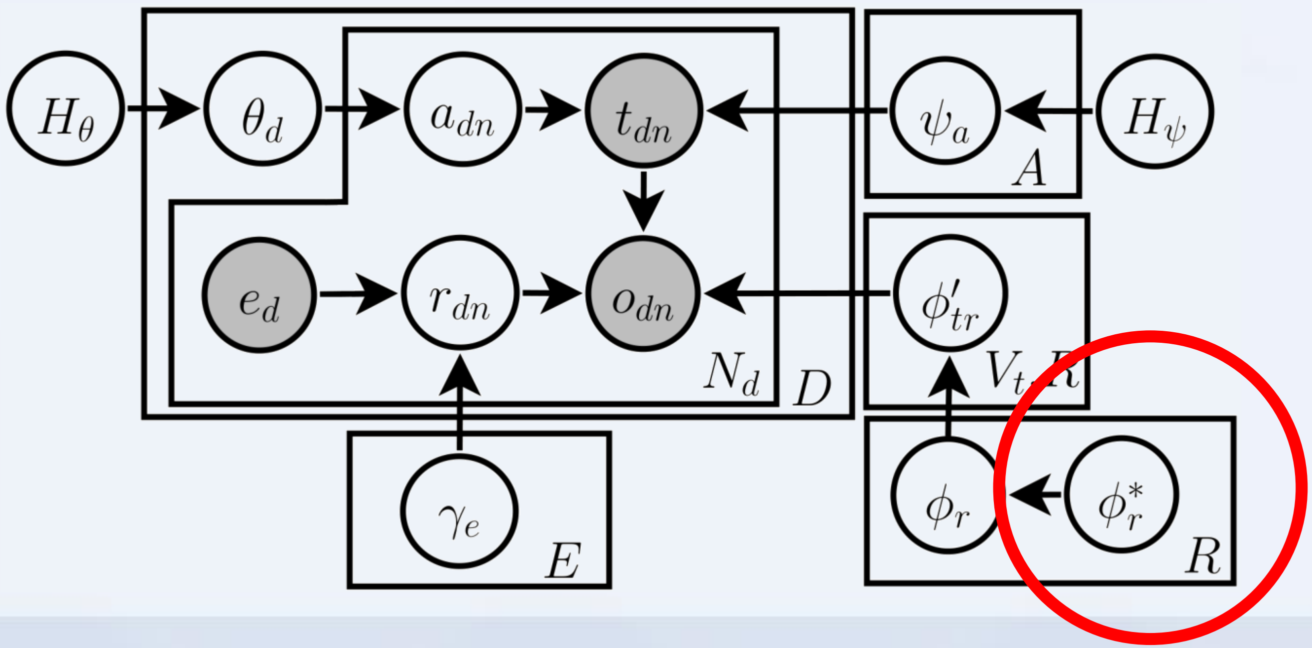
Twitter Opinion Topic Model (TOTM)

- Hierarchical priors for opinion words.
  - Model both target-specific and general sentiment-opinion distributions.



Twitter Opinion Topic Model (TOTM)

- Can use existing sentiment lexicon as prior.
  - We use SentiStrength and MPQA lexicons.



Sentiment Prior Formulation

- TOTM uses a tuneable parameter b that control the strength of the sentiment lexicon:

$$\phi^*_{rv} \propto (1+b)^{X_{rv}}$$

- X = the sentiment score for sentiment r
  – Higher value means stronger sentiment.
- Easy to differentiate => simple to learn b.

Sentiment Prior Formulation

- How to formulate X?

$$X_{rv} = \begin{cases} S_v & \text{if } \ r = 1 \ \ \text{(positive)} \\ -|S_v| & \text{if } \ r = 0 \ \ \text{(neutral)} \\ -S_v & \text{if } \ r = -1 \ \ \text{(negative)} \end{cases}$$

- S = sentiment score from lexicon
  - Assumed positive sentiment => positive S
  - Negative sentiment => negative S

Example

- For the word “happy”:
  - SentiStrength score, S = +2
  - So X = +2 for positive sentiment

    X = -2 for neutral sentiment

    X = -2 for negative sentiment
- Hence it is a priori more likely for “happy” to be given a positive sentiment.

Training TOTM

- Collapsed Gibbs Sampling for Hierarchical PYP
  - Probability distributions are integrated out.
  - Store information as counts, like LDA.
  - Algorithm consists of decrementing and incrementing the counts.
  - More details in the paper.

Learning Hyperparameters

- For PYP hyperparameters
  - Use auxiliary variable sampler (Teh, 2006).
- For tuneable sentiment strength parameter b
  - Use gradient ascent:
    - new b = old b + gradient × learning rate
    - Gradient $l'(b) = \frac{1}{(1+b)} \sum_r \sum_v c_{rv} \underline{\left(X_{rv} - \mathbb{E}_{\phi_r}[X_r]\right)} + \rho'(b)$
  - Quite intuitive:
    - Increase b if the sentiment score is greater than expected.

Outline

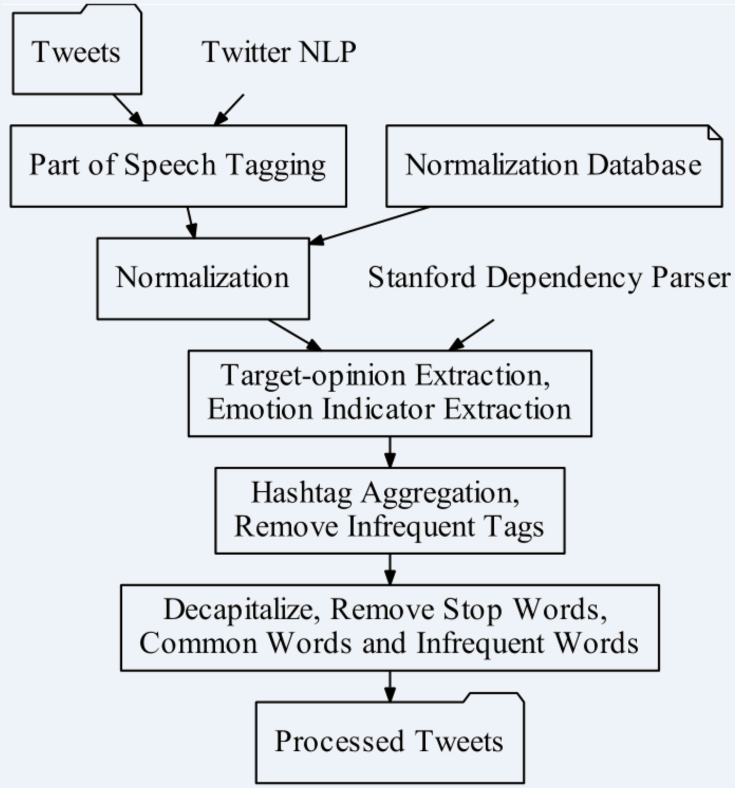
- ~~Motivation~~
- ~~Background and Previous Work~~
- ~~Twitter Opinion Topic Model~~
- Experiments

Data

- Use 3 corpus:
  - From Twitter 7 dataset (Yang & Leskovec, 2011)
    - Query 9 millions tweets on Electronic Products.
    - Non-English tweets are removed.
    - Tweets containing URL are removed.
  - Sentiment 140 tweets (Go et al., 2009)
    - 1.6 millions tweets annotated using emoticons.
  - SemEval tweets (Nakov et al, 2013)
    - 6322 tweets annotated by humans (Mechanical Turk).

Data Preprocessing for TOTM

- Convert raw tweets to target-opinion pairs.



Data Preprocessing for TOTM

- Part-of-Speech Tagging
  - with TwitterNLP (Owoputi, 2013).
  - State-of-the-art for tweets.
  - Also tokenise the tweets.

@user yep , the quality of the new iPhone is really good :) ! i like it .

@ ! , D N P D A ^ V R A E , O V O ,

Noun Adjective Proper Noun

Data Preprocessing for TOTM

- Normalisation
  - Use a conversion dictionary from Han et al. (2012).
  - But do not convert proper noun (iPhone $\nrightarrow$ phone).
  - Examples:
    - amaaazing $\longrightarrow$ amazing
    - nite $\longrightarrow$ night
    - 2morrow $\longrightarrow$ tomorrow

Data Preprocessing for TOTM

- Target-opinion extraction
  - Use Stanford Dependency parser (De Marneffe et al., 2006).
  - Convert relations to target-opinion pairs.
  - Rules:

$$amod(N, A) \rightarrow < N, A >$$
$$acomp(V, A) + nsubj(V, N) \rightarrow < N, A >$$
$$cop(A, V) + nsubj(A, N) \rightarrow < N, A >$$
$$< h_1, m > + conj\_and(h_1, h_2) \rightarrow < h_2, m >$$
$$< h, m_1 > + conj\_and(m_1, m_2) \rightarrow < h, m_2 >$$
$$< h, m > + neg(m, not) \rightarrow < h, not + m >$$
$$< h, m > + nn(h, N) \rightarrow < N + h, m >$$
$$< h, m > + nn(N, h) \rightarrow < h + N, m >$$

Data Preprocessing for TOTM

- Extract positive or negative emoticons
  - Use both eastern and western smileys:

	Eastern	**Western**
Positive	**^_^ (^u^)**	**:) =-)**
Negative	**<(` ^´)> T_T**	**:@ :'(**

  - Use strong sentiment words
    - Such as “happy”, “sad”, etc.

Data Preprocessing for TOTM

- Aggregate tweets based on hashtags
  - Word co-occurrence to be used by topic model.
  - Give a different way to view the results.
- Remove stop words, common words and rare words.
  - These words are of less interest.
  - eg: “he”, “she”, misspellings, etc...

Experiments

- Compare TOTM with 2 baselines:
  - ILDA as mentioned previously.
  - LDA-DP
    - Vanilla LDA but apply ad hoc modification to the prior following He (2012).
    - Set ϕ^*_{rv} to 0.9 if sentiment for word v is the same as r, else set to 0.05 .

Experiments

- Quantitative Evaluations
  - Perplexity
  - Sentiment classification
  - Sentiment prior evaluation
- Qualitative Evaluations
  - Inspecting word distributions
  - Comparing opinions
  - Opinions extraction

Perplexity

- Commonly used to evaluate topic models.
- Negatively related to the log likelihood of observed words.
  - So lower perplexity is better.

Log likelihood for words in the test set

$$\mathrm{perplexity}(\mathbf{W}) = \exp\left(-\frac{\sum_{d=1}^{D}\log P(\vec{w}_d)}{\sum_{d=1}^{D} N_d}\right)$$

Normaliser (Number of words)

Perplexity

- Results

	Target	Opinion	Overall
LDA-DP	N/A	510.15 ± 0.08	N/A
ILDA	594.81 ± 13.61	519.84 ± 0.43	556.03 ± 6.22
TOTM	592.91 ± 13.86	$\mathbf{137.42} \pm 0.28$	$\mathbf{285.42} \pm 3.23$

- Significant improvement on opinion words
  – since TOTM model target-opinion interaction directly, i.e. better prediction for opinion words.

Sentiment Classification

- Evaluate on annotated tweets.
- Predict sentiment by selecting the polarity that has higher likelihood given the sentiment-word distributions.

$$\text{polarity}(d) = \underset{r=\{-1,1\}}{\text{argmax}} \prod_i \phi_{r,o_{di}}$$

Sentiment Classification

- Results

Sent140 Tweets	Accuracy	Precision	Recall	F1-score
LDA-DP	57.3	56.1	90.1	69.2
ILDA	54.1	56.9	55.3	55.9
TOTM	**65.0**	**61.7**	**90.2**	**73.3**
SemEval Tweets	Accuracy	Precision	Recall	F1-score
LDA-DP	52.1	65.0	58.3	61.4
ILDA	46.8	60.7	53.6	56.3
TOTM	**73.3**	**84.0**	**74.9**	**79.0**

- TOTM performs best in sentiment classification.

Evaluating Sentiment Prior

- Use SentiWordNet to evaluate the learned sentiment-opinion distributions.
- SentiWordNet gives positive affinity and negative affinity for each word, eg:
  - “Active” -> positive 0.5 , negative 0.125
  - “Supreme” -> positive 0.75 , negative 0
- So can calculate both positivity and negativity of an opinion word distribution.

Evaluating Sentiment Prior

- Evaluation metric
  - Sentiment score – expected sentiment under an opinion word distribution.

opinion word distribution

$$Score(\phi_r, Z) = E_{\phi_r}[Z] = \sum_{v=1}^{V_o} Z_v \phi_{rv}$$

Z = positive or negative affinity from SentiWordNet

Evaluating Sentiment Prior

- Results

	Electronic Product Tweets		*Sent140 Tweets*		*SemEval Tweets*	
	Negativity	Positivity	Negativity	Positivity	Negativity	Positivity
No lexicon	17.82 $\pm$ 1.26	17.39 $\pm$ 0.45	22.63 $\pm$ 0.96	32.31 $\pm$ 1.98	15.24 $\pm$ 1.45	21.03 $\pm$ 3.85
MPQA	**23.91** $\pm$ 0.49	31.96 $\pm$ 0.09	24.10 $\pm$ 0.49	**42.65** $\pm$ 1.02	16.88 $\pm$ 0.31	29.47 $\pm$ 0.99
SentiStrength	23.19 $\pm$ 0.08	**35.69** $\pm$ 0.33	**24.29** $\pm$ 1.07	41.26 $\pm$ 1.53	**16.94** $\pm$ 0.78	**32.17** $\pm$ 2.07

  - No lexicon = use only emoticons
  - SentiStrength is slightly better than MPQA lexicon.
  - Sentiment lexicon gives significant improvement.

Experiments

- ~~Quantitative Evaluations~~
  - ~~Perplexity~~
  - ~~Sentiment classification~~
  - ~~Sentiment prior evaluation~~
- Qualitative Evaluations
  - Inspecting word distributions
  - Comparing opinions
  - Opinions extraction

Inspecting Word Distributions

- We can inspect aspect-target word distributions to see if the target words are correctly clustered.
  - Some examples:

Aspects (a)	Target Words (t)
Camera	camera, pictures, video camera, shots
Apple iPod	ipod, ipod touch, songs, song, music
Android phone	android, apps, app, phones, keyboard
Macbook	macbook, macbook pro, macbook air
Nintendo games	nintendo, games, game, gameboy

  - Target words are closely related.

Inspecting Word Distributions

- Similarly, we can inspect the opinion word distributions.
  – TOTM allows in depth analysis by looking at opinion word distributions for a particular target.

Target (t)	+/−	Opinions (o)
phone	−	**dead** damn stupid bad crazy
	+	**mobile smart** good great f***ing
battery life	−	terrible poor bad horrible **non-existence**
	+	good **long** great **7hr ultralong**
game	−	**addictive** stupid free **full addicting**
	+	great good awesome favorite **cat-and-mouse**
sausage	−	silly **argentinian cold** huge stupid
	+	**hot grilled** good **sweet** awesome

* Words in **bold** are more specific and can only describe certain targets.

Comparing Opinions

- Aggregating tweets using hashtags allows additional analysis.
  - We inspect hashtags that correspond to electronic companies such as #sony, #canon, #samsung...

Comparing Opinions

- A snapshot

Brands	Sentiment	Aspects / Targets' Opinions		
		Camera	Phone	Printer
Canon	–	*camera* → expensive small bad *lens* → prime cheap broken		*printer* → obscure violent digital *scanner* → cheap
	+	*camera* → great compact amazing *pictures* → great nice creative		*printer* → good great nice *scanner* → great fine
Sony	–	*camera* → big crappy defective *lens* → vertical cheap wide	*phone* → worst crappy shittest *battery life* → low	*printer* → stupid
	+	*photos* → great lovely amazing *camera* → good great nice	*phone* → great smart beautiful *reception* → perfect	
Samsung	–	*camera* → digital free crazy *shots* → quick wide	*phone* → stupid bad fake *battery life* → solid poor terrible	*scanner* → worst
	+	*camera* → gorgeous great cool *pics* → nice great perfect	*phone* → mobile great nice *service* → good sweet friendly	

Opinions Extraction

- Finally, TOTM allows us to query tweets that correspond to certain opinions.
  - Example: query opinions on iPhone

Positive	Negative
RT @user : the iPhone is so awesome!!! Emailing, texting, surfing the sametime! — Can do all tgat while talkin on the phone?...	@user awww thx! I can't send an email right now bc my iPhone is stupid with sending emails. Lol but I can tweet or dm u?
Ahhh! Tweeting on my gorgeous iPhone! I missed you! hehe am on my way home, put the kettle on will you pls :)	It would appear that the iPhone, due to construction, is weak at holding signal. Combine that with a bullshit 3G network in Denver.
Thanks @user for the link to iPhone vs Blackberry debate. I got the iPhone & it's just magic! So intuitive!	@user @user Ah, well there you go. The iPhone is dead, long live Android! ;)
Finally my fave lover @user has Twitter & will be using it all the time with her cool new iPhone :)	@user Finally eh? :D I think iphone is so ugly x.x

Major Contributions

- Introduce TOTM for aspect-based opinion mining on tweets.
  - Makes use of emoticons and hashtags on tweets.
- Novel way of incorporating sentiment prior information into topic model.
  - Simple to implement and allow automatic learning of hyperparameters.

Thanks!

References

- M. De Marneffe, B. MacCartney, and C. Manning. Generating typed dependency parses from phrase structure parses. In *LREC*, pages 449–454, 2006.
- A. Go, R. Bhayani, and L. Huang. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, pages 1–12, 2009.
- B. Han, P. Cook, and T. Baldwin. Automatically constructing a normalisation dictionary for microblogs. In *EMNLP-CoNLL*, pages 421–432. ACL, 2012.
- M. Luca. Reviews, reputation, and revenue: The case of Yelp.com. *Working Paper, Harvard Business School*, 2011.
- S. Moghaddam and M. Ester. On the design of LDA models for aspect-based opinion mining. In *CIKM*. ACM, 2012.

References

- P. Nakov, Z. Kozareva, A. Ritter, S. Rosenthal, V. Stoyanov, and T. Wilson. SemEval-2013 task 2: Sentiment analysis in Twitter. In *Workshop on Semantic Evaluation*, 2013.
- O. Owoputi, B. O'Connor, C. Dyer, et al. Improved part-of-speech tagging for online conversational text with word clusters. In *NAACL-HLT*, pages 380–390, 2013.
- Y. W. Teh. A Bayesian interpretation of interpolated Kneser-Ney. *Tech Report A2/06, NUS*, 2006.
- J. Yang and J. Leskovec. Patterns of temporal variation in online media. In *WSDM*, pages 177–186, 2011.