Text Mining of Social Media: Going beyond the Text and Only the Text

Timothy Baldwin

THE UNIVERSITY OF MELBOURNE

Source(s): `http://people.eng.unimelb.edu.au/tbaldwin/pubs/wnut2015.pdf`

Talk Outline

Introduction

- NLP is inevitably focused on ... text, the whole text and nothing but the text, generally at the sentence or document level, ignoring the greater sentence/document context

Source(s): `http://goo.gl/ZS6A4n`

Introduction

- Meanwhile, the network analysis and data mining communities largely ignore (or use very simple models of) the textual content of nodes, and focus instead on connections between them, in the form of graphs of different types

Source(s): `http://goo.gl/APOQPP`

Thought Experiment I

In the absence of any other textual context, what does *no change* mean?

Thought Experiment I

What if I provide non-textual context?

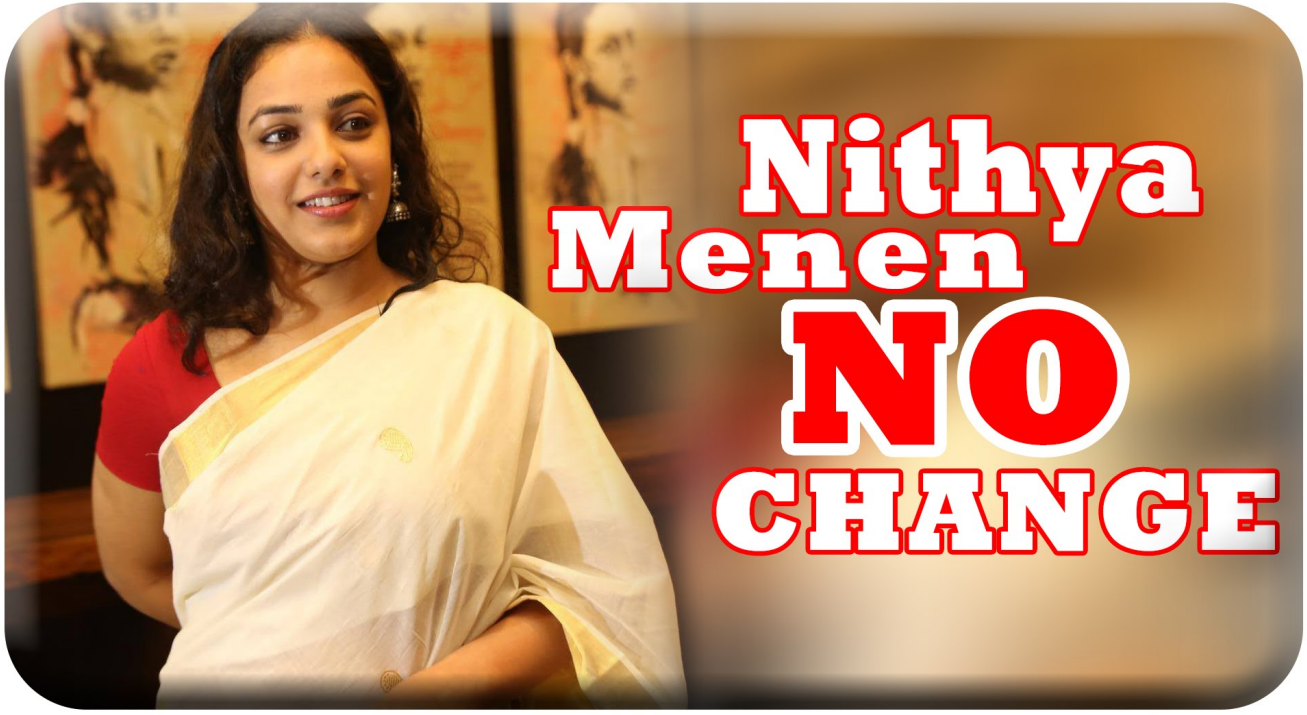
Thought Experiment I

Source(s): `http://40.media.tumblr.com/tumblr_lm6hqnxkHU1qeb9g8o1_500.jpg`

Thought Experiment I

Source(s): `http://i.ytimg.com/vi/5uEwcBT8MFM/maxresdefault.jpg`

Thought Experiment I



Source(s): `http://i.ytimg.com/vi/I0sk0ATOaws.jpg`

Thought Experiment II

Text $\leftrightarrow$ non-text predictive modelling

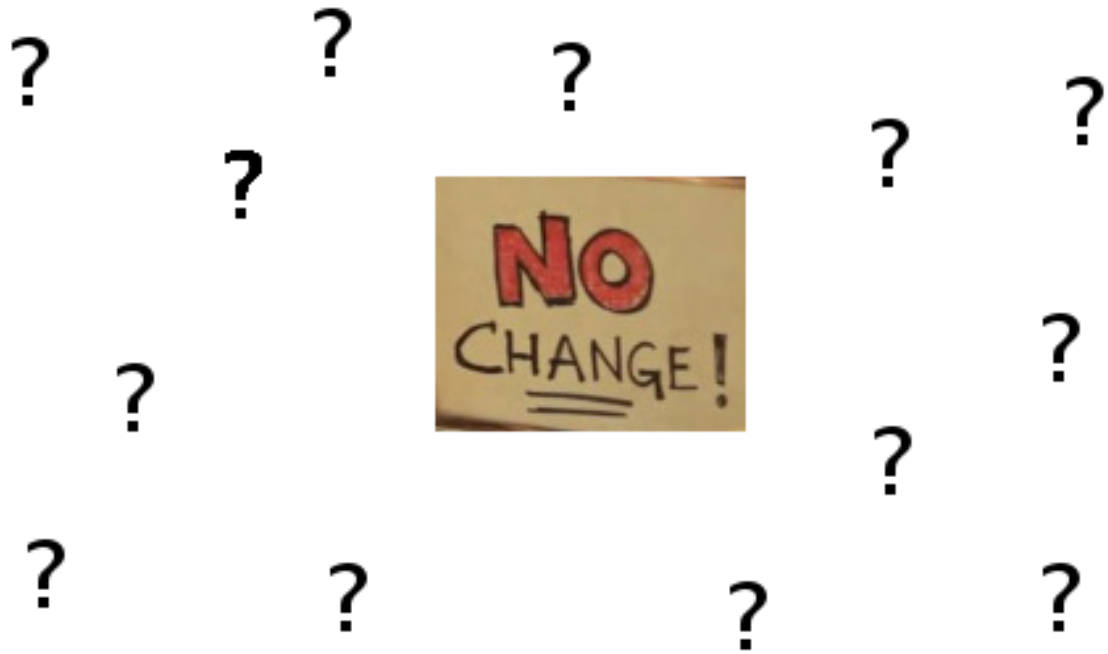
Thought Experiment II

Predict the text:

Source(s): `http://40.media.tumblr.com/tumblr_lm6hqnxkHU1qeb9g8o1_500.jpg`

Thought Experiment II

Predict the image:



Source(s): `http://40.media.tumblr.com/tumblr_lm6hqnxkHU1qeb9g8o1_500.jpg`

Because ... You're all Individuals!

Source(s): https://whatdaddoes.files.wordpress.com/2014/02/life-of-brian.jpg

So What Context are We Talking About?

I used graphical context as illustrative examples, but what I am really talking about is:

- Document metadata:
  - author [Wang et al., 2011, Yogatama et al., 2011, Carter et al., 2013, Lui and Baldwin, 2014]
  - author profile [Eisenstein et al., 2011, Bergsma et al., 2013, Volkova et al., 2013, Hovy, 2015]
  - publisher/host site [Yogatama et al., 2011]
  - timestamp [Yogatama et al., 2011]
  - genre
  - domain [Yogatama et al., 2011]
- Document type
- Document markup
- Position within document [Li et al., 2015]
- Extra-textual content (tables, graphs, etc.)

The Relevance to Social Media

- This is particularly relevant to social media, as a lot of context is immediately accessible, in terms of:
  - likes/favourites
  - user metadata of different types
  - message metadata of different types
  - user "timeline"
  - social network data of each user
  - explicit interactions between users (favourites, mentions, shares/retweets/...)

Broad Aim

- **Aim:** improve NLP through the use of document context, focusing in this talk on:
  - document metadata
  - inter-document graph structure
- **Concerns along the way:**
  - means of extracting context
  - scalability of models
  - model expressivity

Talk Outline

What is User Geolocation?

- Given a set of messages from a user, e.g.:
  - *Waiting for a tram in the rain in Collins St. A more typical Melbourne day today.*
  - *Why you keep me up? I ain't got no worries.*
  - *New Aussie Hip Hop News: The Yarra stinks.*
  - *Just had a rather thrilling albeit bumpy camel ride around Uluru - SO. MUCH. FUN!! Fancy joining me? Enter my comp here*

  predict their "home" (as distinct from "source" or "about") location

What is User Geolocation?

- Given a set of messages from a user, e.g.:
  - *Waiting for a tram in the rain in Collins St. A more typical Melbourne day today.*
  - *Why you keep me up? I ain't got no worries.*
  - *New Aussie Hip Hop News: The Yarra stinks.*
  - *Just had a rather thrilling albeit bumpy camel ride around Uluru - SO. MUCH. FUN!! Fancy joining me? Enter my comp here*

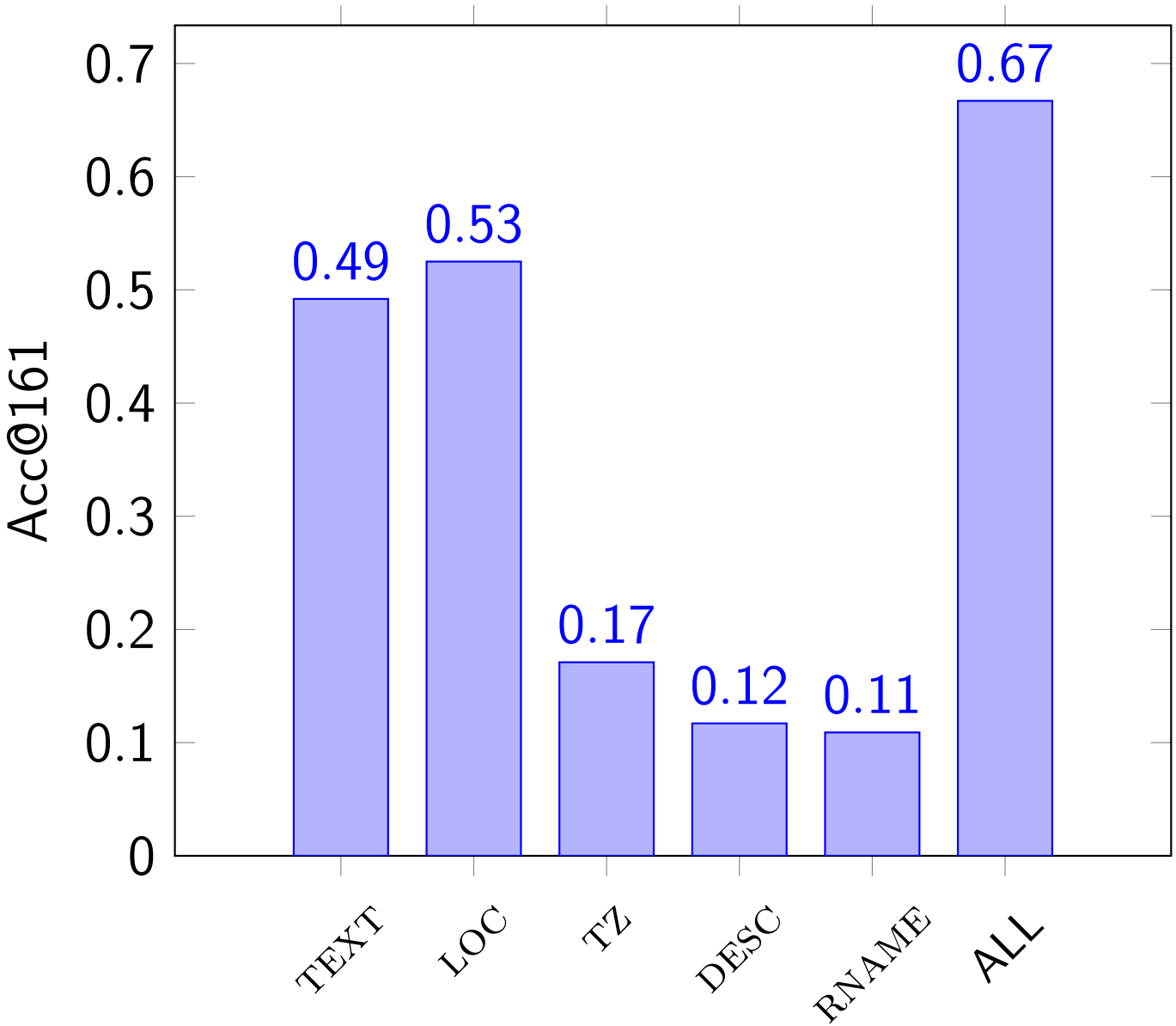
  predict their "home" (as distinct from "source" or "about") location ... Melbourne, AU

Twitter User Geolocation: Overall Methodology

1. Discretise the geolocation class space, e.g. using a k-d tree or via gazetteers [Roller et al., 2012, Han et al., 2012]
2. Classify each user by supervised classification, based on users with geotagged tweets, using either:
   1. the content of the users' messages ("TEXT")
   2. user metadata
      - user-declared location ("LOC")
      - timezone ("TZ")
      - description ("DESC")
      - real name ("RNAME")
   3. a combination of metadata and message content ("ALL")

Dataset = TWITTER-WORLD: around 12M English tweets from 1.4M users based around the world [Han et al., 2012]; user location = centre of the closest city to the centroid of tweets.

User Geolocation: Results



Source(s): Han et al. [2014]

User Geolocation: Findings

- User-declared location more accurate than text of posts from user; little information in other meta-data fields
- The combination of the text and meta-data fields (based on a stacked logistic regression model) is more accurate again

User Geolocation: Findings

- User-declared location more accurate than text of posts from user; little information in other meta-data fields
- The combination of the text and meta-data fields (based on a stacked logistic regression model) is more accurate again
- Important to realise that pre-computing and storing imputed user priors for all users is potentially intractable

User Geolocation: Findings

- User-declared location more accurate than text of posts from user; little information in other meta-data fields
- The combination of the text and meta-data fields (based on a stacked logistic regression model) is more accurate again
- Important to realise that pre-computing and storing imputed user priors for all users is potentially intractable
- ... but also note that all this metadata is in plain sight in the Twitter JSON object, along with the message text, although not necessarily as immediately accessible for other social media sources

  ... BUT still text-based; what about going beyond text?

User Geolocation: Moving Forward

- Much more metadata context that can be integrated, including:
  - hashtag priors [Carter et al., 2013]
  - content similarity
  - language priors [Han et al., 2014]
  - user "popularity"

Talk Outline

Inter-document Graphs

- Various possible sources of inter-document graphs:
  - explicit inter-document interactions (e.g. mentions)
  - explicit author-level interactions (e.g. following)
  - implicit inter-document similarity (e.g. document overlap/similarity)

Inter-document Graphs

- Various possible sources of inter-document graphs:
  - explicit inter-document interactions (e.g. mentions)
  - explicit author-level interactions (e.g. following)
  - implicit inter-document similarity (e.g. document overlap/similarity)
- Various possibilities for graph semantics:
  - directed vs. undirected
  - weighted vs. unweighted
  - single vs. multiple graphs

Network Analytics 101

- One of the core concepts in network analytics is **homophily** — the tendency of individuals to associate and bond with similar others
  - the corollary for network analytics is that strongly connected subgraphs tend to share the same label
  - obvious analogies in clustering and classification, with the main difference being the presence/absence of an explicit graph

Network Analytics 101

- One of the core concepts in network analytics is **homophily** — the tendency of individuals to associate and bond with similar others
  - the corollary for network analytics is that strongly connected subgraphs tend to share the same label
  - obvious analogies in clustering and classification, with the main difference being the presence/absence of an explicit graph
- Sometimes connections actually represent **heterophily**, esp. in adversarial contexts such as debates

Approaches to Network Inference

- Popular approaches to network inference:
  - **label propagation**: nearest neighbour-style iterative semi-supervised approach
  - **collective classification**: combine base and network classifiers to optimise consistency in the network
  - **matrix factorisation**: factorise the matrix into a product of lower-dimensional matrices

Label Propagation

- Given a graph $G = (\mathcal{V}, E, W)$ where $\mathcal{V}$ is the set of nodes with $|\mathcal{V}| = n = n_l + n_u$ (where n_l nodes are labelled and n_u nodes are unlabelled), E is the set of edges, and W is an edge weight matrix.
- Simple iterative algorithm [Zhu and Ghahramani, 2002]:
  1. for each node $u_u^{(i)} \in \mathcal{V}_u$, get the set of labelled neighbours based on E, and label $u_u^{(i)}$ based on the (weighted) median latitude and longitude of the neighbours
  2. repeat until convergence

Label Propagation

- Modified Adsorption [Talukdar and Crammer, 2009]:

$$C(\hat{Y}) = \sum_{l} \Big[\mu_1 (Y_l - \hat{Y}_l)^T S (Y_l - \hat{Y}_l) + \mu_2 \hat{Y}_l^T L \hat{Y}_l + \mu_3 \| \hat{Y}_l - R_l \|^2 \Big]$$

where μ_1, μ_2 and μ_3 are hyperparameters; L is the Laplacian of an undirected graph derived from G; S is a diagonal binary matrix indicating if a node is labelled or not; and R_l is the lth column of matrix R of dimensions $n \times (m+1)$.

Collective Classification

- **Collective classification:** given a network and an object o in the network, use (up to) three types of correlations to infer a label for o:
  1. the correlations between the label of o and its observed attributes
  2. the correlations between the label of o and the observed attributes and labels of nodes connected to o
  3. the correlations between the label of o and the unobserved labels of objects connected to o

Source(s): Sen et al. [2008]

Collective Classification

- Formally, collective classification takes a graph, made up of:
  - nodes $\mathcal{V} = \{V_1, \ldots, V_n\}$
  - edges E
- The task is to label the nodes $V_i \in \mathcal{V}$ from a label set $\mathcal{L} = \{L_1, \ldots, L_q\}$, making use of the graph in the form of a neighborhood function $\mathcal{N} = \{N_1, \ldots, N_n\}$, where $N_i \subseteq \mathcal{V} \setminus \{V_i\}$.

Approaches to Collective Classification

- Two general approaches to capturing the first two correlations:
  - **iterative classification:** bootstrap node labels with a content-only classifier and generate a random ordering over nodes $\mathcal{V}$, then iteratively update estimate of v_i based on the current $\mathcal{N}_i$ and update $\vec{a}_i$ accordingly [**local approach**]
  - **dual classifier + graph inference:** train separate content-only and link classifiers, and use graph inference (mean field, loopy belief propagation, min-cut, etc.) to "smooth" the predictions over the graph [**global approach**]

Source(s): Sen et al. [2008]

User Geolocation: Enter the Network

- The easiest way to generate network for Twitter user geolocation is via @user mentions (e.g. `@eltimster lovin the talk`)
- Question of what to do with user mentions outside the training/dev/test data sample

  (one) solution = collapse edges through out-of-network nodes into direct edges

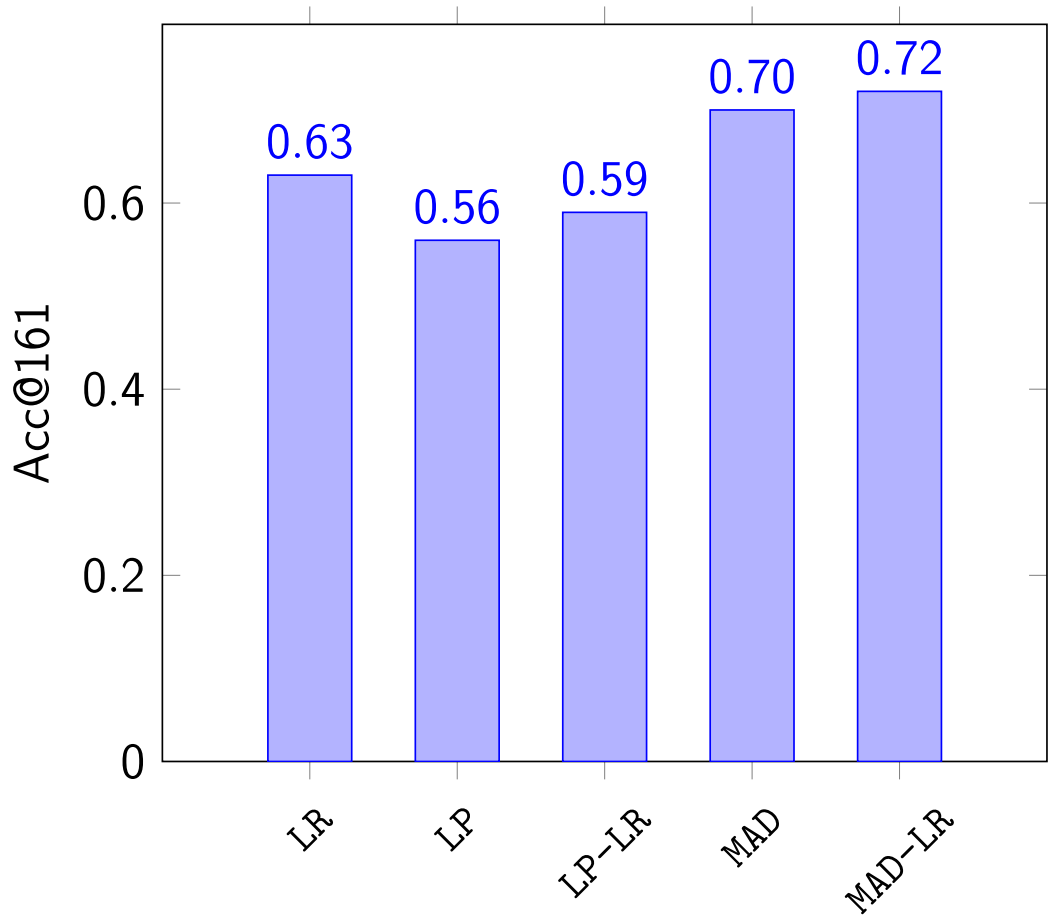
- Weighted, directed graph the most obvious approach, but we have found unweighted, undirected graphs to work best
- Modified Adsorption doesn't scale to well to large, highly-connected graphs, so consider removing edges associated with highly-connected users

Source(s): Jurgens [2013], Rahimi et al. [2015a,b]

User Geolocation: What about the Text?

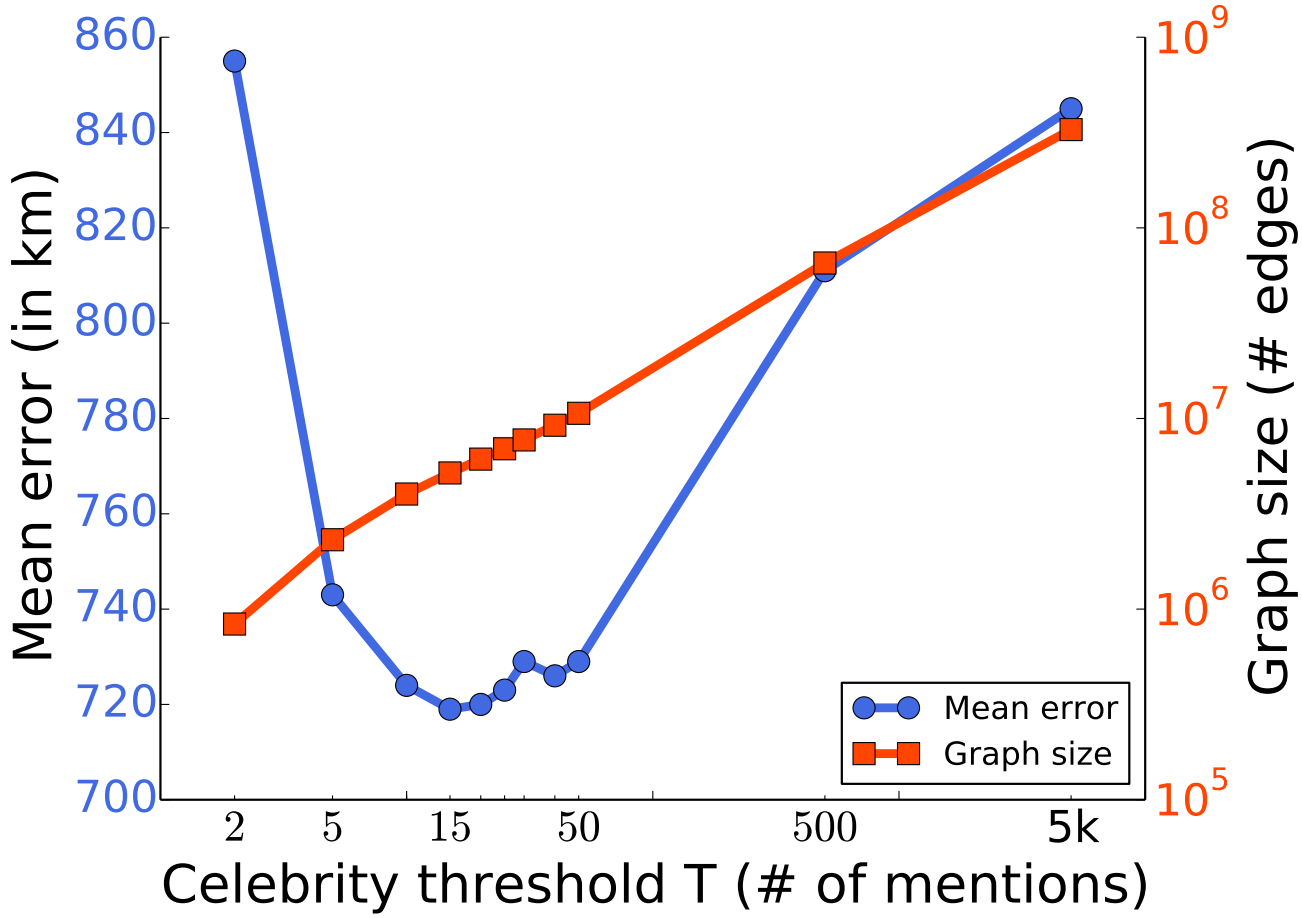
- The easiest way to integrate graph- and text-based classification is to use the text as a source of priors
- **Approach 1:** use pointwise text-based user priors as backoff for disconnected nodes [**post-processing**]
- **Approach 2:** use pointwise text-based user priors as priors for all unlabelled nodes [**pre-processing**]
  - with MAD, we incorporate the priors as "dongle" nodes (uniquely) connected to a given user

User Geolocation: Results



Source(s): Han et al. [2014], Rahimi et al. [2015a,b]

User Geolocation: "Celebrity Nodes"



Source(s): Rahimi et al. [2015a]

User Geolocation: Findings

- Little to separate text- or network-only results ($\texttt{LP} < \texttt{LP} < \texttt{MAD}$)
- Both network-based methods improve with the incorporation of text-based user priors
- In terms of computational efficiency, $\texttt{LP} > \texttt{LR} \gg \texttt{MAD}$
- Removal of highly-connected nodes leads to greater tractability and also better results for $\texttt{MAD}$

User Geolocation: Moving Forward

- Much more network context that can be integrated, including:
  - retweet interaction data
  - time distribution data
  - geographical similarity
  - represent user by cluster of geotagged tweets rather than single node
- More refined analysis of local vs. global "celebrities"
- Matrix factorisation, and other graph inference methods

Vote Prediction: Task

- Given the text for a given speaker in a political debate, e.g.:
  - Blackburn, Marsha (R)
    at this time , i would like to recognize the gentleman from texas (mr. hensarling) who has worked tirelessly not only on budgeting and not only on looking at how we budget , but looking at what ...
  - Hensarling, Jeb (R)
    mr. speaker , unless we enact h.r. 4297 and defeat the democratic substitute , americans will receive a most unwelcome christmas gift from the democrats ...

  predict their vote (FOR or AGAINST)
- User mentions take the form of direct mentions of others in the debate, and are provided as part of the dataset

Source(s): Thomas et al. [2006]

Vote Prediction: Task

- Given the text for a given speaker in a political debate, e.g.:
  - Blackburn, Marsha (R) [for]
    at this time , i would like to recognize the gentleman from texas (mr. hensarling) who has worked tirelessly not only on budgeting and not only on looking at how we budget , but looking at what ...
  - Hensarling, Jeb (R) [for]
    mr. speaker , unless we enact h.r. 4297 and defeat the democratic substitute , americans will receive a most unwelcome christmas gift from the democrats ...

  predict their vote (for or against)
- User mentions take the form of direct mentions of others in the debate, and are provided as part of the dataset

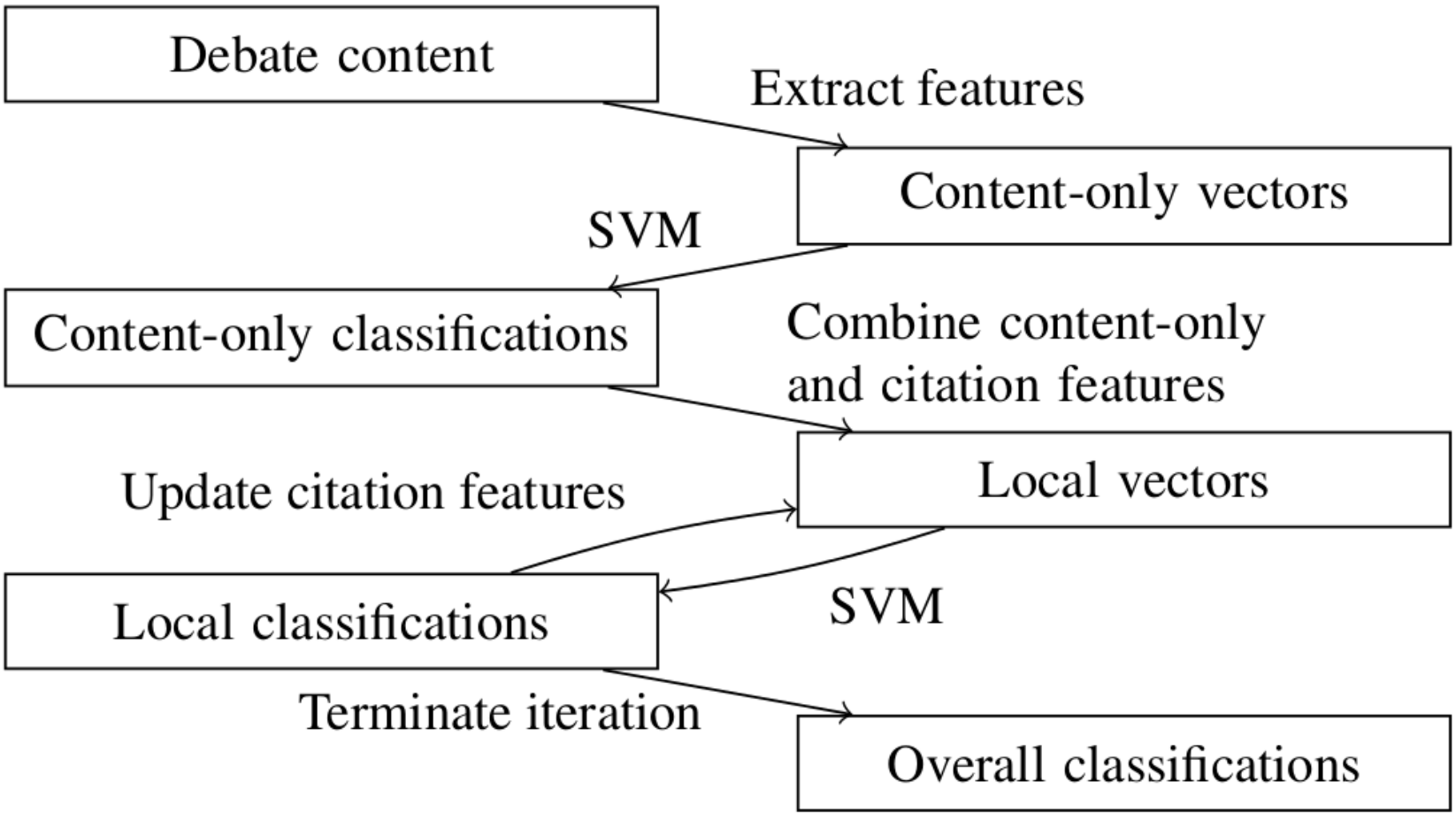
Source(s): Thomas et al. [2006]

Vote Prediction Dataset: ConVote

- **Data:** US congressional debates from 2005 [Thomas et al., 2006]
- Mentions manually tagged in the data

	Total
Tokens	1.2M
Speeches	1699
Debates	53
Average speakers/speeches per debate	32
Average tokens per speech	735
Proportion of For speeches	49%

Vote Prediction: Iterative Classifier

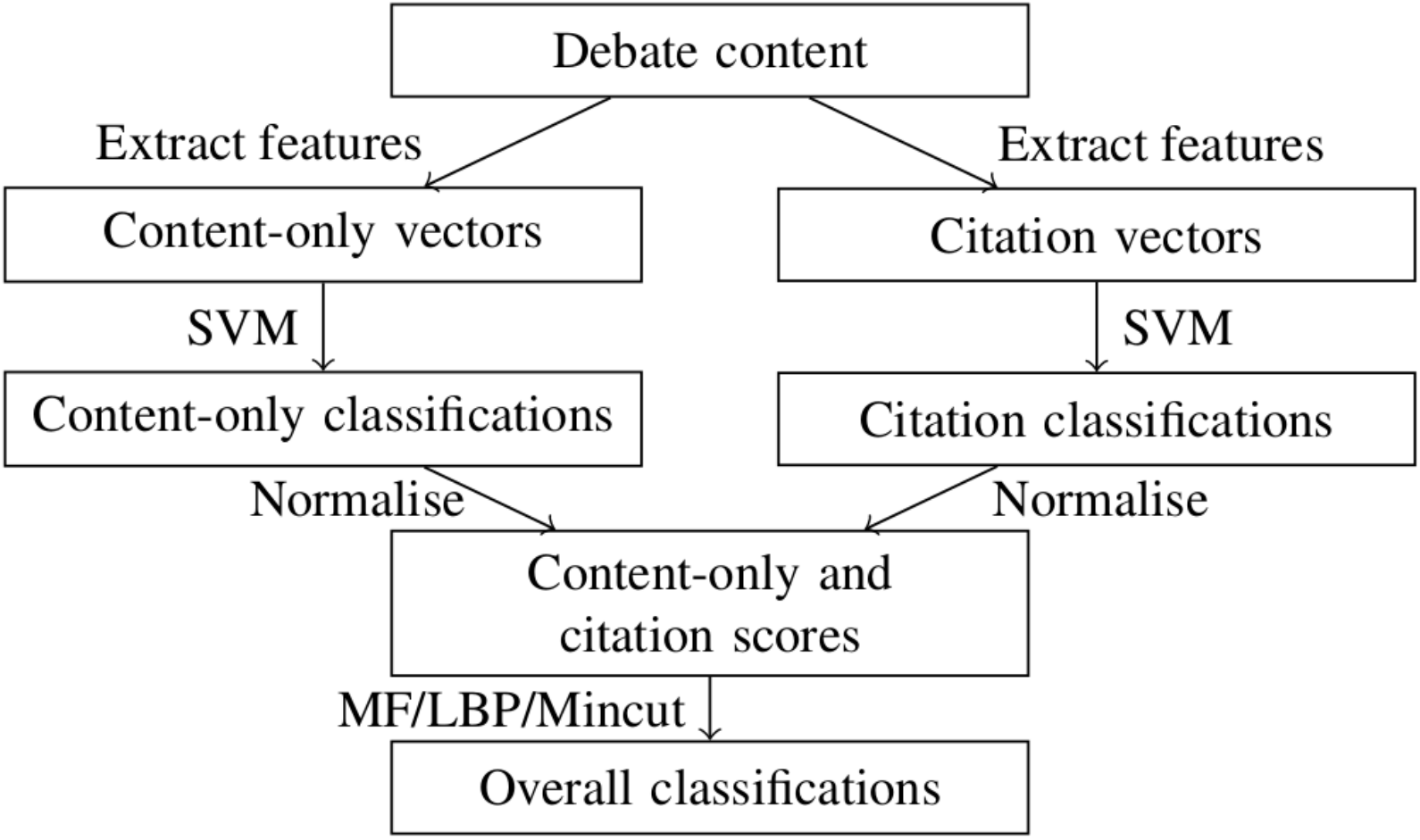


Source(s): Burfoot et al. [2011]

Citation Features: Representing Context

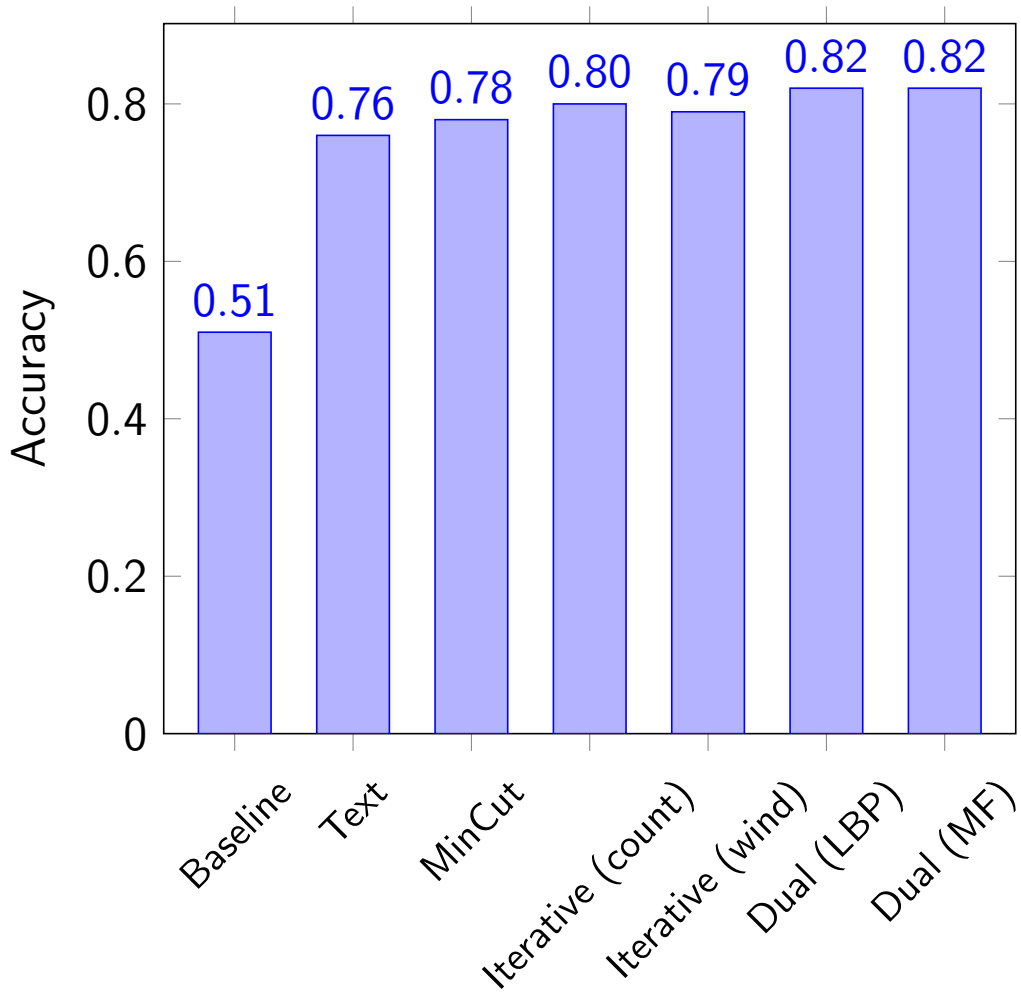
- For iterative classification over ConVote, we experiment with three representations of citation:
  1. **citation count:** what are the counts of nodes in $\mathcal{N}_i$ that have the same/different class to v_i?
  2. **context window:** generate a feature vector $\mathcal{L} \times \mathcal{C}$ over the context windows of each document in $\mathcal{N}_i$

Vote Prediction: Dual Classifier



Source(s): Burfoot et al. [2011]

Vote Prediction: Results



Source(s): Burfoot et al. [2011]

Vote Prediction: Findings

- Once again, text- and graph-based methods produce similar accuracy in isolation
- ... and once again, the combination of the two performs better again
- Dual classifiers tend to do slightly better than iterative classifiers

Source(s): Burfoot et al. [2011]

Vote Prediction: Moving Forward

- Possibility of adding further context, including:
  - joint modelling of stance [Sridhar et al., 2015]
  - joint modelling of sentiment of citation context
  - modelling of speech turn taking/chronology
- Possibility of doing user modelling:
  - cross-debate user modelling of different speakers/parties
  - modelling of speaker "influence"

What about Implicit Graphs?

- What happens if we don't have any explicit mention data to generate our graph from?
- **ANSWER**: we can still generate (complete) graphs through content similarity, e.g. based on text similarity

Implicit Graphs: ConVote

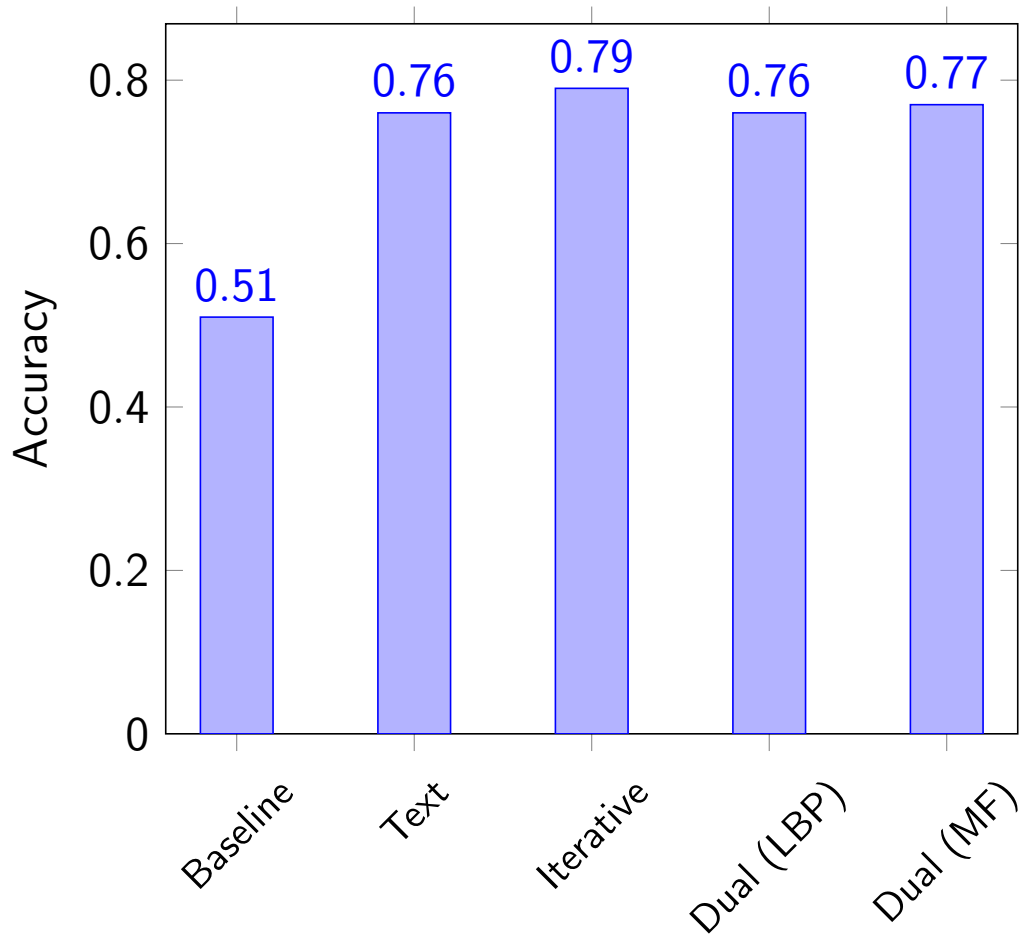
- Generate an undirected weighted graph based on the cosine similarity between each document pair, represented by TF-IDF vectors $\langle ...w_{i,j}, ... \rangle$ for each document d_j where:

$$w_{i,j} = \frac{\mathbf{1}_{d_j}(w_i)}{1 + \log \sum_k \mathbf{1}_{d_k}(w_i)}$$

- For iterative classification, calculate the average similarity score with neighbours of each class

Source(s): Burford et al. [2015]

Implicit Graphs: Results (5-grams)



Source(s): Burford et al. [2015]

Talk Outline

Other Random Ideas

- Expanding our notion of "context" in distributional semantics beyond words to include (at least) the user dimension
- Inference methods for multiple graph overlays
- Expanding our notion of "domain" to include user context

Talk Outline

Conclusions

- There's plenty of context out there in social media, in forms including:
  - user metadata
  - explicit mention graph
  - implicit content similarity graph

Conclusions

- There's plenty of context out there in social media, in forms including:
  - user metadata
  - explicit mention graph
  - implicit content similarity graph
- There's plenty of evidence to suggest that this context has high utility in NLP tasks

Conclusions

- There's plenty of context out there in social media, in forms including:
  - user metadata
  - explicit mention graph
  - implicit content similarity graph
- There's plenty of evidence to suggest that this context has high utility in NLP tasks
- There's also plenty of evidence to suggest that the combination of text and context analysis is a potent combination ... GET TO IT!

References

Shane Bergsma, Mark Dredze, Benjamin Van Durme, Theresa Wilson, and David Yarowsky. Broadly improving user classification via communication-based name and location clustering on Twitter. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2013)*, pages 1010–1019, Atlanta, USA, 2013. URL `http://www.aclweb.org/anthology/N13-1121`.

Clinton Burfoot, Steven Bird, and Timothy Baldwin. Collective classification of congressional floor-debate transcripts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011)*, pages 1506–1515, Portland, USA, 2011.

Clint Burford, Steven Bird, and Timothy Baldwin. Collective document classification with implicit inter-document semantic relationships. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics (*SEM 2015)*, pages 106–116, Denver, USA, 2015.

Simon Carter, Manos Tsagkias, and Wouter Weerkamp. Microblog language identification: overcoming the limitations of short, unedited and idiomatic text. *Language Resources and Evaluation*, 47(1):195–215, 2013.

References

Jacob Eisenstein, Noah A. Smith, and Eric P. Xing. Discovering sociolinguistic associations with structured sparsity. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011)*, pages 1365–1374, Portland, USA, 2011. URL `http://www.aclweb.org/anthology/P11-1137`.

Bo Han, Paul Cook, and Timothy Baldwin. Geolocation prediction in social media data by finding location indicative words. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, pages 1045–1062, Mumbai, India, 2012.

Bo Han, Paul Cook, and Timothy Baldwin. Text-based Twitter user geolocation prediction. *Journal of Artificial Intelligence Research*, 49:451–500, 2014.

Dirk Hovy. Demographic factors improve classification performance. In *Proceedings of the 53nd Annual Meeting of the Association for Computational Linguistics and 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2015)*, pages 752–762, Beijing, China, 2015. URL `http://www.aclweb.org/anthology/P15-1073`.

David Jurgens. That's what friends are for: Inferring location in online social media platforms based on social relationships. In *Proceedings of the 7th International Conference on Weblogs and Social Media (ICWSM 2013)*, pages 273–282, Dublin, Ireland, 2013.

References

Jiwei Li, Thang Luong, and Dan Jurafsky. A hierarchical neural autoencoder for paragraphs and documents. In *Proceedings of the 53nd Annual Meeting of the Association for Computational Linguistics and 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2015)*, pages 1106–1115, Beijing, China, 2015. URL `http://aclweb.org/anthology/P15-1107`.

Marco Lui and Timothy Baldwin. Accurate language identification of Twitter messages. In *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)*, pages 17–25, Gothenburg, Sweden, 2014. URL `http://www.aclweb.org/anthology/W14-1303`.

Afshin Rahimi, Trevor Cohn, and Timothy Baldwin. Twitter user geolocation using a unified text and network prediction model. In *Proceedings of the 53nd Annual Meeting of the Association for Computational Linguistics and 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2015)*, pages 630–636, Beijing, China, 2015a.

Afshin Rahimi, Duy Vu, Trevor Cohn, and Timothy Baldwin. Exploiting text and network context for geolocation of social media users. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics — Human Language Technologies (NAACL HLT 2015)*, pages 1362–1367, Denver, USA, 2015b.

References

Stephen Roller, Michael Speriosu, Sarat Rallapalli, Benjamin Wing, and Jason Baldridge. Supervised text-based geolocation using language models on an adaptive grid. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning 2012 (EMNLP-CoNLL 2012)*, pages 1500–1510, Jeju Island, Korea, 2012. URL `http://www.aclweb.org/anthology/D12-1137`.

Prithviraj Sen, Galileo Mark Namata, Mustafa Bilgic, Lise Getoor, Brian Gallagher, and Tina Eliassi-Rad. Collective classification in network data. *AI Magazine*, 29(3): 93–106, 2008.

Dhanya Sridhar, James Foulds, Bert Huang, Lise Getoor, and Marilyn Walker. Joint models of disagreement and stance in online debate. In *Proceedings of the 53nd Annual Meeting of the Association for Computational Linguistics and 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2015)*, pages 116–125, Beijing, China, 2015. URL `http://www.aclweb.org/anthology/P15-1012`.

Partha Pratim Talukdar and Koby Crammer. New regularized algorithms for transductive learning. In *Proceedings of the European Conference on Machine Learning (ECML-PKDD) 2009*, pages 442–457, 2009.

References

Matt Thomas, Bo Pang, and Lillian Lee. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*, pages 327–335, Sydney, Australia, 2006.

Svitlana Volkova, Theresa Wilson, and David Yarowsky. Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*, pages 1815–1827, Seattle, USA, 2013.

Li Wang, Marco Lui, Su Nam Kim, Joakim Nivre, and Timothy Baldwin. Predicting thread discourse structure over technical web forums. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*, pages 13–25, Edinburgh, UK, 2011.

Dani Yogatama, Michael Heilman, Brendan O'Connor, Chris Dyer, Bryan R. Routledge, and Noah A. Smith. Predicting a scientific communitys response to an article. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*, pages 594–604, Edinburgh, UK, 2011. URL `http://www.aclweb.org/anthology/D11-1055`.

Xiaojin Zhu and Zoubin Ghahramani. Learning from labeled and unlabeled data with label propagation. Technical report, Technical Report CMU-CALD-02-107, Carnegie Mellon University, 2002.