From Bias to Opinion: A Transfer-Learning Approach to Sentiment Analysis

Pedro Calais Guerra, Adriano Veloso
Wagner Meira Jr., Virgilio Almeida

Universidade Federal de Minas Gerais, Brazil

InWeb – Brazilian National Institute of Science and Technology for the Web

inweb

August 23, 2011

Social Media and Opiniated Data

Blogs, microblogs and online social networks are flooded with real-time opinions about a multitude of topics:

- Politics - *"candidate X is the best"*
- Sports competitions - *"Team Y is favorite to win today!"*
- Other topics and current "buzz"

Sentiment Analysis

- **Sentiment Analysis** (or opinion mining) aims to interpret text and **predict polarity** of the writer regarding a topic or entity [Pang and Lee, 2008]
- The initial application was on mining product review data [Turney, ACL'02]: *"Cell Phone X is great"*, *"Its interface is awful"*

Sentiment Analysis

- **Sentiment Analysis** (or opinion mining) aims to interpret text and **predict polarity** of the writer regarding a topic or entity [Pang and Lee, 2008]
- The initial application was on mining product review data [Turney, ACL'02]: *"Cell Phone X is great"*, *"Its interface is awful"*

Mining opinions in real-time must face new challenges:

- Dinamicity of discussions
- Lack of labeled textual data

Dinamicity of Discussions and Lack of Labeled Data

- Major events are composed of numerous sub-events [Leskovec et al., KDD'2009]
- Real-time data (e.g., Twitter)
- Huge number of messages (e.g., 100k tweets per day)
- Sparsity of textual content

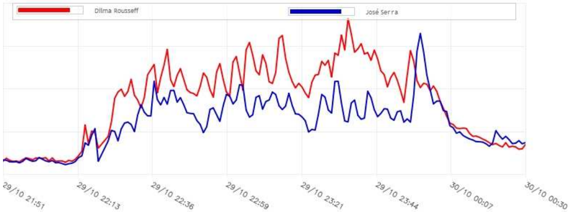


Figure: Twitter traffic related to one candidates' debate in the Brazilian Presidential Elections

Our proposal

What is the time-invariant pattern that does not require significant labeling efforts and supports real-time sentiment analysis?

Our proposal

What is the time-invariant pattern that does not require significant labeling efforts and supports real-time sentiment analysis?
⇒ **User Bias**

Our proposal

What is the time-invariant pattern that does not require significant labeling efforts and supports real-time sentiment analysis?

$\Rightarrow$ **User Bias**

How can we measure user bias in social media, using only social interactions?

Our proposal

What is the time-invariant pattern that does not require significant labeling efforts and supports real-time sentiment analysis?
⇒ **User Bias**

How can we measure user bias in social media, using only social interactions?
⇒ **Endorsements**

Our proposal

What is the time-invariant patterns that does not require significant labeling efforts and supports real-time sentiment analysis?
$\Rightarrow$ **User Bias**

How can we measure user bias in social media, using only social interactions?
$\Rightarrow$ **Endorsements**

How can we use user bias information to analyze sentiment?

Our proposal

What is the time-invariant pattern that does not require significant labeling efforts and supports real-time sentiment analysis?
$\Rightarrow$ **User Bias**

How can we measure user bias in social media, using only social interactions?
$\Rightarrow$ **Endorsements**

How can we use user bias information to analyze sentiment?
$\Rightarrow$ **Transfer Learning**

Bias and opinions

Bias is inherent to most humans [Watson 1991], since they:

- take particular position regarding a subject
- have a personal interest from the arguer in the outcome of the argument or discussion.
- lack proper balance and neutrality in argumentation
- lack proper critical doubt

Bias and opinions

Bias is inherent to most humans [Watson 1991], since they:

- take particular position regarding a subject
- have a personal interest from the arguer in the outcome of the argument or discussion.
- lack proper balance and neutrality in argumentation
- lack proper critical doubt

Bias and opinions are not independent!

- Supporters of a candidate are likely to issue positive opinions on him/her
- Soccer team supporters act similarly

Social Media Endorsements as Evidence of User Bias

Endorsements: interactions through which a user implicitly **agrees** with another user w.r.t. a certain content:

retweet @OfficialMyTeamProfile, @CandidateX...

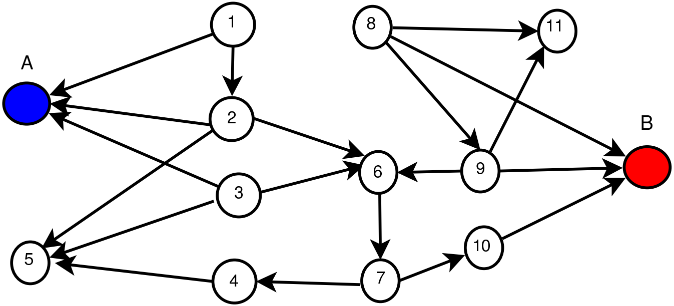
like Democrats, Republicans, San Diego Chargers

Measuring Bias as a Networked Data Classification Problem

Given

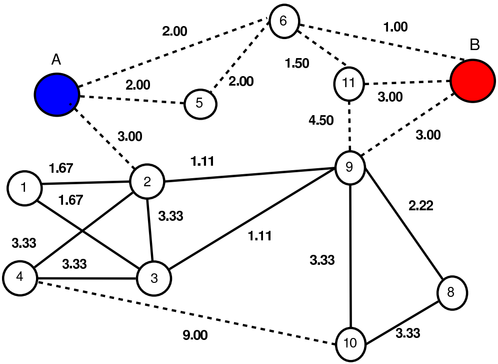
- $G = (V, E)$, *(u,v)* denotes that user u endorsed user v
- K possible sides that a user may have a sentiment about
- Some users ("attractors") have a clear bias w.r.t. a topic

Measuring bias is to estimate $\vec{B_u} = [B_{u1}, \ldots, B_{uK}], \forall u$



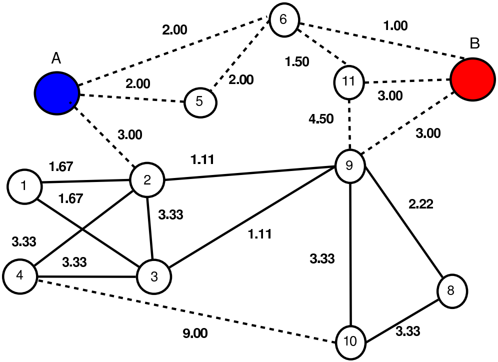
The Opinion Agreement Graph

- $G = (V, E)$, where each vertice is a user and each edge expresses a **global** judgement of the connected users
- **Undirected** edges eases the relational learning process



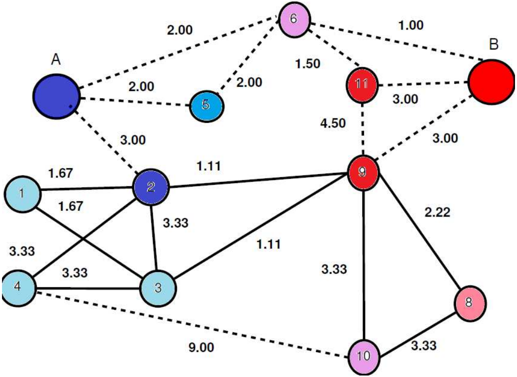
The Opinion Agreement Graph

- **Solid edge**: two users **endorse** the same users
- **Dashed edge**: two users **are endorsed** by the same users
- **Edge weight**: the lift of the size of both sets



Measuring Bias as a Vector of Random Walks

Nodes with similar bias should be closer one from the other!
The proximity of user u w.r.t to side i is a random walk from seed A_i to u: $B_{ui} = RandomWalk(G, A_i, u)$



Datasets

- Brazilian 2010 Presidential Elections
- Brazilian 2010 Soccer League

	Elections-BR	**Soccer**
period	Dec 2009 to Oct 2010	May 2010 to Nov 2010
entities	2 candidates	12 football clubs
# of tweets	7.7 mi	8.8 mi
# of retweets	2.5 mi (32% of tweets)	1.9 mi (21% of tweets)
# of users	1.0 mi	1.6 mi
# of users retweeting	0.5 mi (48% of users)	0.6 mi (36% of users)
# of retweeted users	0.2 mi (18% of users)	0.3 mi (21% of users)

Datasets

- Brazilian 2010 Presidential Elections (**2 labeled nodes**)
- Brazilian 2010 Soccer League (**12 labeled nodes**)

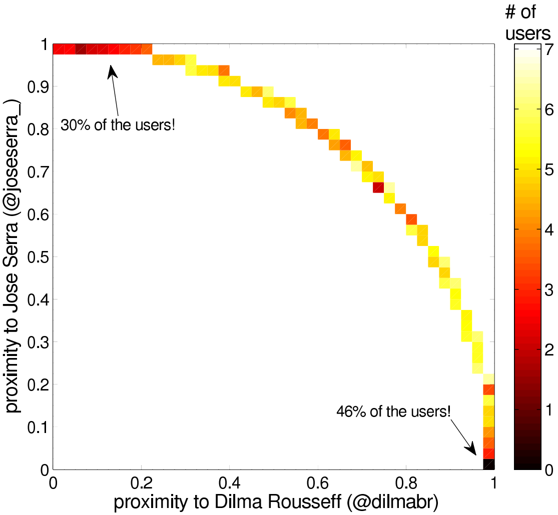
	Elections-BR	**Soccer**
period	Dec 2009 to Oct 2010	May 2010 to Nov 2010
entities	**2 candidates**	**12 football clubs**
# of tweets	7.7 mi	8.8 mi
# of retweets	2.5 mi (32% of tweets)	1.9 mi (21% of tweets)
# of users	1.0 mi	1.6 mi
# of users retweeting	0.5 mi (48% of users)	0.6 mi (36% of users)
# of retweeted users	0.2 mi (18% of users)	0.3 mi (21% of users)

Datasets

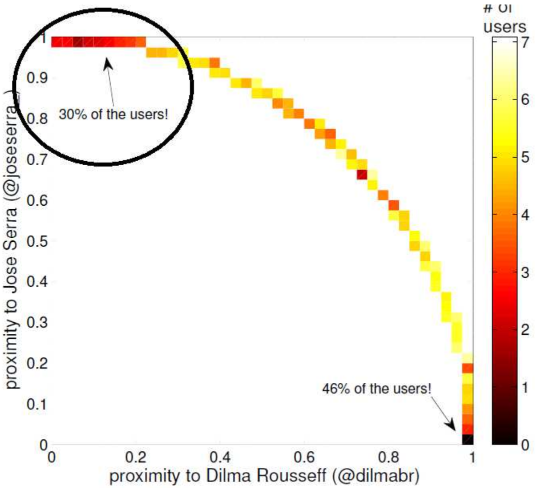
- Brazilian 2010 Presidential Elections
- Brazilian 2010 Soccer League

	Elections-BR	**Soccer**
period	Dec 2009 to Oct 2010	May 2010 to Nov 2010
entities	2 candidates	12 football clubs
# of tweets	**7.7 mi**	**8.8 mi**
# of retweets	**32% of tweets**	**21% of tweets**
# of users	1.0 mi	1.6 mi
# of users retweeting	**(48% of users)**	**(36% of users)**
# of retweeted users	**(18% of users)**	**(21% of users)**

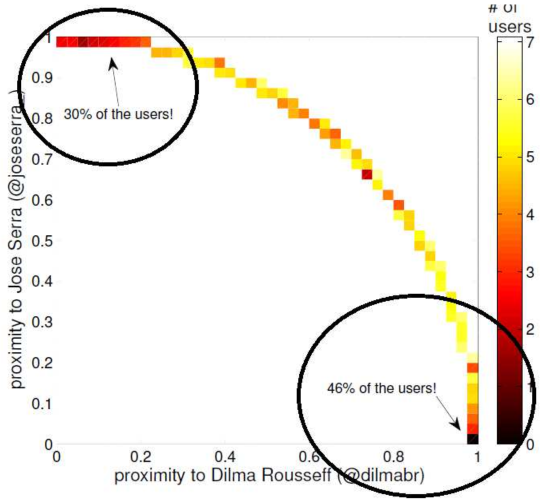
Bias in Elections Discussions



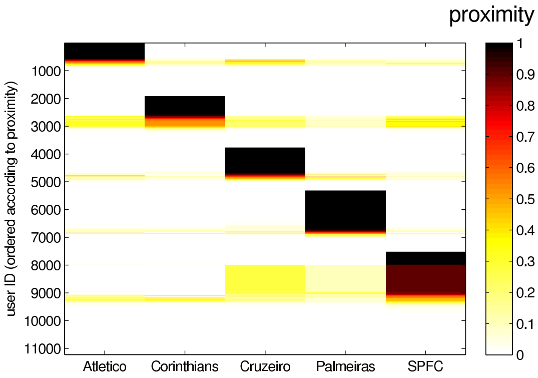
Bias in Elections Discussions



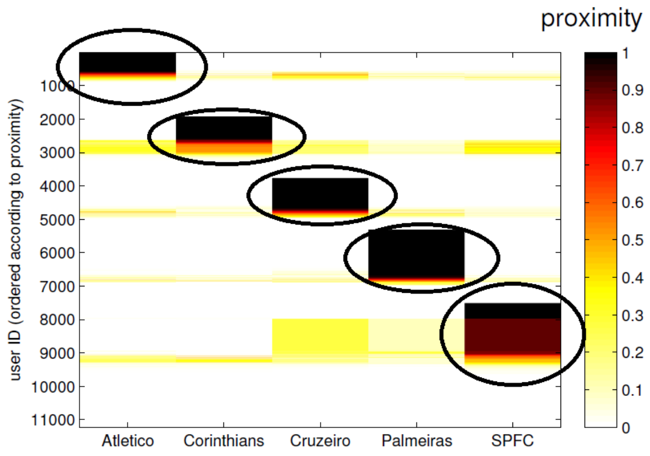
Bias in Elections Discussions



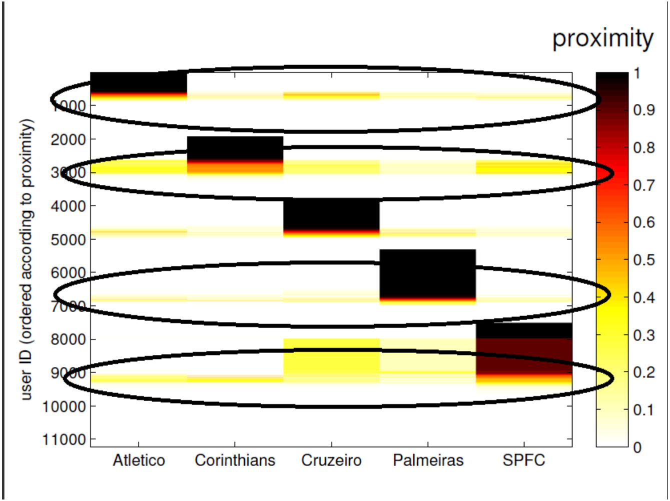
Bias in Soccer Discussions



Bias in Soccer Discussions



Bias in Soccer Discussions



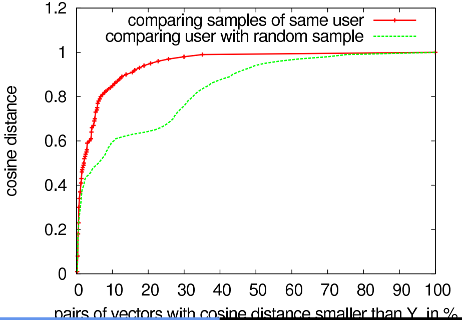
Bias is a consistent pattern!

- We divided retweets into two disjoint sets
- We calculated cosine similarity between the 2 bias vectors for each user

Bias is a consistent pattern!

- We divided retweets into two disjoint sets
- We calculated cosine similarity between the 2 bias vectors for each user

90% of the users have a cosine similarity greater than 0.8 when two samples of data are compared.



Transfer Learning

- Transfer learning solves a target task τ_t, which is **hard**, by transferring knowledge obtained from other **similar tasks** or domains (τ_s) [Pan and Yang, 2010]
- It has been applied to perform sentiment analysis using knowledge from different product domains [Dredze and Pereira, ACL'07; Sindhwani et. al; SDM'2010]
- Example:

  τ_s = mine book review data

  τ_t = mine movie review data

Transferring Bias from Users to Content

Bias of a term with respect to an entity is the **sum vector** of the biases of all users that referred to that term:

$$\vec{B}_{t,e} = \sum_{u \in V} \vec{B}_u \quad (1)$$

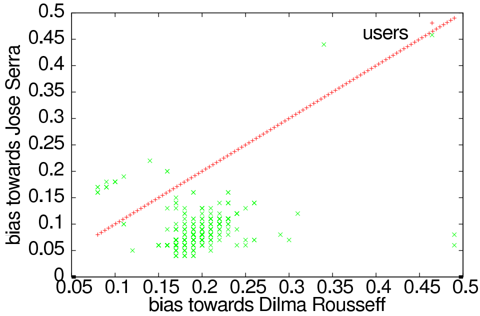
$\Rightarrow$ Enable **quick judgements** of new terms!

Positive polarity of a term is the strength of side (component) **e** in $\vec{B}_{t,e}$:

$$\hat{p}(polarity = +|t, e) = \frac{B_{t,e,e}}{||\vec{B}_{t,e}||} \quad (2)$$

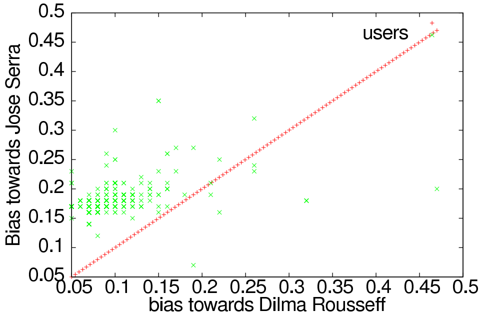
$\vec{B}_{t,e} = [0.7, 0.3] \rightarrow \hat{p}(polarity = +|t, e_1) = 70\%$

#nowDilma is a positive tag for candidate Dilma Rousseff:



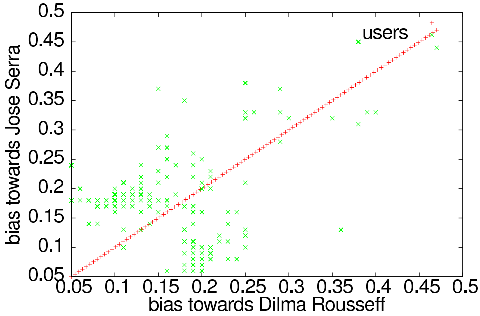
Relationship between terms and user bias

#DilmaLies is a negative tag for candidate Dilma Rousseff:



Relationship between terms and user bias

#2010Elections is a neutral tag:



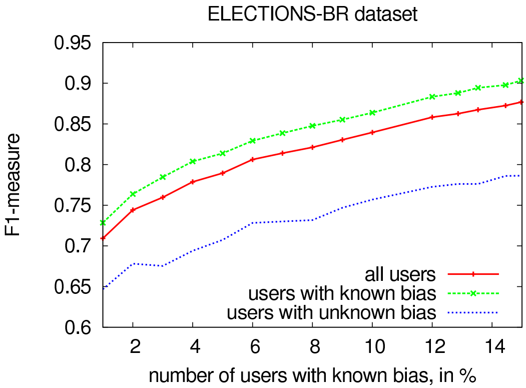
Our sentiment classification model

Just take the term of highest polarity:

$$polarity = argmax(\hat{p}(polarity = x|t)) \tag{3}$$

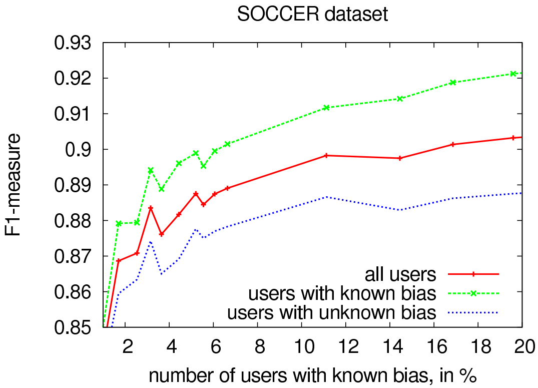
Evaluating the Knowledge Transfer Process - Elections

- ↑ bias, ↑ accuracy
- Knowing bias of 5% of users is enough to get F1 $>$ 0.8



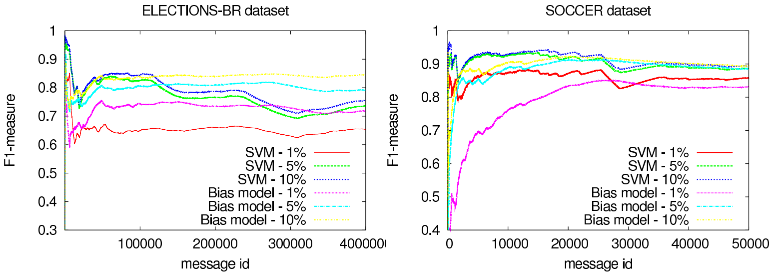
Evaluating the Knowledge Transfer Process - Soccer

- $\uparrow$ bias, $\uparrow$ accuracy
- Knowing bias of 5% of users is enough to get F1 > 0.85

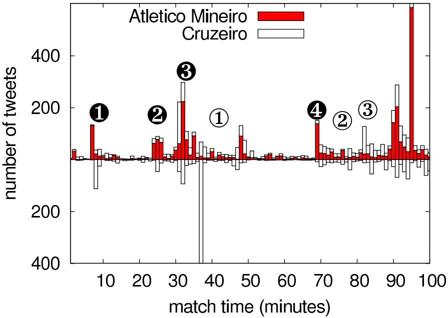


Comparison to SVM

- Competitive to SVM, despite not using labeled textual data
- SVM performance decreases over time, bias-based do not



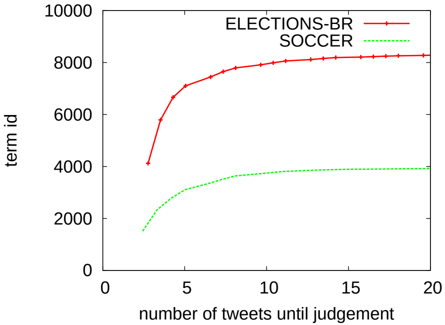
Analyzing a Soccer Match in Real Time



- Positive sentiment coincides with goals
- Peak of positive sentiment for the winner after the match
- Negative opinions are mostly in the losing side

Learning New Terms

⇒ New popular terms are quickly "judged" by users.



Summary

- Novel transfer learning setting to determine sentiment over polemic, heavily-debated topics, in real-time

Summary

- Novel transfer learning setting to determine sentiment over polemic, heavily-debated topics, in real-time
- Few seeds of known bias → propagate through endorsements → propagate user bias across terms

Summary

- Novel transfer learning setting to determine sentiment over polemic, heavily-debated topics, in real-time
- Few seeds of known bias $\rightarrow$ propagate through endorsements $\rightarrow$ propagate user bias across terms
- Little human effort!
- Real-time sentiment analysis grounded on the consistency of the user bias

Future Work

- Integrating user bias into state-of-the-art sentiment analysis algorithms
- Temporal-context based Sentiment Analysis
- Are opinions equally important?
- 2012 US Elections and 2012 Olympics

Questions?

From Bias to Opinion:
A Transfer-Learning Approach to Sentiment Analysis

`www.dcc.ufmg.br/~pcalais`