Explaining the Stars: Weighted Multiple-Instance Learning for Aspect-Based Sentiment Analysis

Nikolaos Pappas and Andrei Popescu-Belis

Idiap Research Institute, Martigny, Switzerland

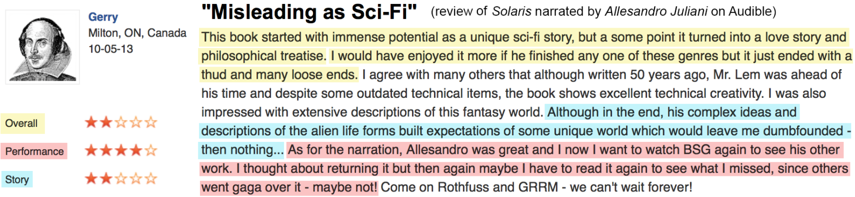
EMNLP 2014, Doha, Qatar

October 26, 2014

Aspect-based sentiment analysis

Fine-grained sentiment analysis i.e. determining opinions expressed on different aspects of products:

- **review segmentation**
  detect which sentences refer to which aspect (discovered or fixed)
- **aspect-rating (or sentiment) prediction**
  estimate sentiment towards each aspect (unsupervised, supervised)
- **review summarization**
  create summary of aspect-sentiments with representative sentences



Gerry
Milton, ON, Canada
10-05-13

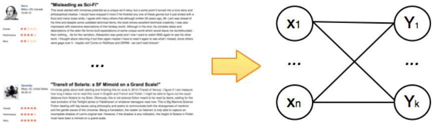
Overall	★★☆☆☆
Performance	★★★★☆
Story	★★☆☆☆

"Misleading as Sci-Fi" (review of *Solaris* narrated by *Allesandro Juliani* on Audible)

This book started with immense potential as a unique sci-fi story, but a some point it turned into a love story and philosophical treatise. I would have enjoyed it more if he finished any one of these genres but it just ended with a thud and many loose ends. I agree with many others that although written 50 years ago, Mr. Lem was ahead of his time and despite some outdated technical items, the book shows excellent technical creativity. I was also impressed with extensive descriptions of this fantasy world. Although in the end, his complex ideas and descriptions of the alien life forms built expectations of some unique world which would leave me dumbfounded - then nothing... As for the narration, Allesandro was great and I now I want to watch BSG again to see his other work. I thought about returning it but then again maybe I have to read it again to see what I missed, since others went gaga over it - maybe not! Come on Rothfuss and GRRM - we can't wait forever!

The problem: aspect-rating prediction

- typically formulated as traditional supervised multi-label learning: given $\mathcal{D} = \{(x_i, y_i) \mid i = 1 \ldots m\}$, $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}^k$, find $\Phi_k : \mathcal{X} \rightarrow \mathcal{Y}_k$
- representations x_i for sentiment analysis:
  - feature engineering (bow, n-grams, topic models and more)
  - feature learning (neural networks)



Gerry
Milton, ON, Canada
10-06-13

Overall ★★☆☆☆
Performance ★★★★☆
Story ★★☆☆☆

"Misleading as Sci-Fi"

This book started with immense potential as a unique sci-fi story, but a some point it turned into a love story and philosophical treatise. I would have enjoyed it more if he finished any one of these genres but it just ended with a thud and many loose ends. I agree with many others that although written 50 years ago, Mr. Lem was ahead of his time and despite some outdated technical items, the book shows excellent technical creativity. I was also impressed with extensive descriptions of this fantasy world. Although in the end, his complex ideas and descriptions of the alien life forms built expectations of some unique world which would leave me dumbfounded - then nothing... As for the narration, Alessandro was great and I now I want to watch BSG again to see his other work. I thought about returning it but then again maybe I have to read it again to see what I missed, since others went gaga over it - maybe not! Come on Rothfuss and GRRM - we can't wait forever!

...

[illegible]
Mesa, AZ, United States
06-06-12

Overall ★★★★★
Performance ★★★★☆
Story ★★★★★

"Transit of Solaris: a SF Mimoid on a Grand Scale!"

I'm kinda giddy about both starting and finishing this on June 5, 2012 (Transit of Venus). I figure if I can measure how long it takes me to read this novel in English and French and Polish, I might be able to figure out the exact distance from Solaris to my Brain. Obviously, this is not science fiction meant to be read by teens, waiting for the next evolution of the Twilight series or Fablehaven or whatever teenagers read now. This is Big Mamma Science Fiction dealing with big issues using philosophy and poetry to communicate both the strangeness of mankind and the gentle waves of the Universe. Being a translation, the reader (or listener) is only able to capture an incomplete shadow of Lem's original text. However, if the shadow is any indication, the height of Solaris in Polish must have been a mimoid on a grand scale.

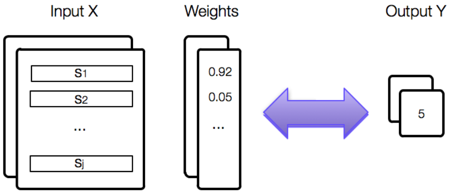
→ treat a text globally and ignore the weak nature of the labels

→ suffer polymorphism and part-whole ambiguities (feeble to noise)

→ offer few or no means for interpretation (how to explain the stars?)

Proposed solution

1. aspect-rating prediction as multiple-instance learning problem
2. hypothesize that text is composed by several parts (sentence-level or paragraph-level) which have unequal contribution to its rating
3. an efficient model to learn to predict contributions and ratings



s1 (w=0.92): Great talk, I really felt inspired.

s2 (w=0.05): However, I think our worst risk is not technological, but political.

...

Outline of the talk

1. Motivation
2. Multiple-instance learning
3. The proposed model
4. Experiments
5. Conclusion

Outline of the talk

1 Motivation

2 Multiple-instance learning

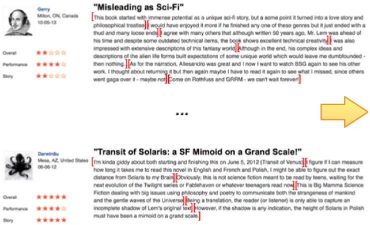
3 The proposed model

4 Experiments

5 Conclusion

Multiple-instance learning (MIL)

- each text is a *bag* described by many data points or *instances*: given $\mathcal{D} = \{(b_{ij}, y_i) \mid i = 1 \dots n, j = 1 \dots n_i\}$, $b_{ij} \in \mathbb{R}^d$ and $y_i \in \mathbb{R}^k$, find $\Phi_k : \mathcal{B} \overset{?}{\rightarrow} \mathcal{X} \rightarrow \mathcal{Y}_k$, where $\mathcal{X} = \{x_{ik}\}$, $x_{ik} \in \mathbb{R}^d$ is unknown
- instances b_{ij} are represented as before but on different levels: paragraph-level, <u>sentence-level</u> or phrase-level



Gerry
Milton, ON, Canada
10-05-13

Overall ★★☆☆☆
Performance ★★★★☆
Story ★★☆☆☆

"Misleading as Sci-Fi"

[This book started with immense potential as a unique sci-fi story, but a some point it turned into a love story and philosophical treatise][I would have enjoyed it more if he finished any one of these genres but it just ended with a thud and many loose ends][I agree with many others that although written 50 years ago, Mr. Lem was ahead of his time and despite some outdated technical items, the book shows excellent technical creativity][I was also impressed with extensive descriptions of this fantasy world][Although in the end, his complex ideas and descriptions of the alien life forms built expectations of some unique world which would leave me dumbfounded - then nothing..][As for the narration, Allesandro was great and I now I want to watch BSG again to see his other work. I thought about returning it but then again maybe I have to read it again to see what I missed, since others went gaga over it - maybe not][Come on Rothfuss and GRRM - we can't wait forever!]

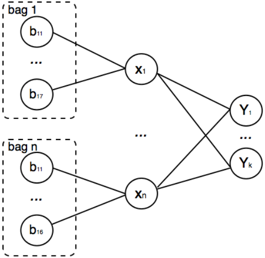
...

DarwinBu
Mesa, AZ, United States
06-06-12

Overall ★★★★★
Performance ★★★★☆
Story ★★★★★

"Transit of Solaris: a SF Mimoid on a Grand Scale!"

[I'm kinda giddy about both starting and finishing this on June 5, 2012 (Transit of Venus)][I figure if I can measure how long it takes me to read this novel in English and French and Polish, I might be able to figure out the exact distance from Solaris to my Brain][Obviously, this is not science fiction meant to be read by teens, waiting for the next evolution of the Twilight series or Fablehaven or whatever teenagers read now][This is Big Mamma Science Fiction dealing with big issues using philosophy and poetry to communicate both the strangeness of mankind and the gentle waves of the Universe][Being a translation, the reader (or listener) is only able to capture an incomplete shadow of Lem's original text][However, if the shadow is any indication, the height of Solaris in Polish must have been a mimoid on a grand scale.]



Flexible (uncovers structure) and **cheaper** (operates on coarse labels).

MIL assumptions

1. **Aggregated instances**: sum or average instances

$$f \leftarrow D_{agg} = \{(x_i, y_i) \mid i = 1, \ldots, m\}$$
$$\hat{y}(B_i) = f(x_i) = f(mean(\{b_{ij} \mid wj = 1, \ldots, n_i\})) \tag{1}$$

2. **Instance-as-example**: each instance is labeled by its bag's label

$$f \leftarrow D_{ins} = \{(b_{ij}, y_i) \mid j = 1, \ldots, n_i; i = 1, \ldots, m\}$$
$$\hat{y}(B_i) = mean(\{f(b_{ij}) \mid j = 1, \ldots, n_i\}) \tag{2}$$

3. **Prime instance**: a single instance is responsible for its bag's label

$$\forall i \ b_i^p = \underset{j}{\operatorname{argmax}} |y_i - f(b_{ij})|$$
$$f \leftarrow D_{pri} = \{(b_i^p, y_i) \mid i = 1, \ldots, m\}$$
$$\hat{y}(B_i) = mean(\{f(b_{ij}) \mid j = 1, \ldots, n_i\}) \tag{3}$$

Weighted-MIL assumptions

4. **Instance relevance**: each instance contributes unequally to its bag's label

   - (Wagstaff 2007) applied to crop yield modeling
   - (Zhoua 2009) treats instances in an non-i.i.d. way that exploits relations among instances
   - (Wang 2011) defines instance-specific distance which is derived by comparisons with training data (it is not directly learned)

→ no model to estimate instance relevances of unseen bags

→ prohibitive complexity for large feature spaces or number of bags

→ most works have focused on classification

Outline of the talk

1. Motivation
2. Multiple-instance learning
3. The proposed model
4. Experiments
5. Conclusion

Proposed model: main idea and assumption

A new weighted multiple-instance learning model for text regression tasks:

- models both instance relevances and target ratings (applicable to prediction and interpretable)
- learns an optimal method to aggregate instances, rather than a pre-defined one (less simplified than previous assumptions)
- supports high dimensional spaces as required for text (computationally efficient)

Assumption: the point x_i is a convex combination of the points in the bag, in other words B_i is represented by the weighted average of its instances b_{ij}

$$x_i = \sum_{j=1}^{n_i} \psi_{ij} b_{ij} \text{ with } \psi_{ij} \geq 0 \; \forall i, j \text{ and } \sum_{j=1}^{n_i} \psi_{ij} = 1 \qquad (4)$$

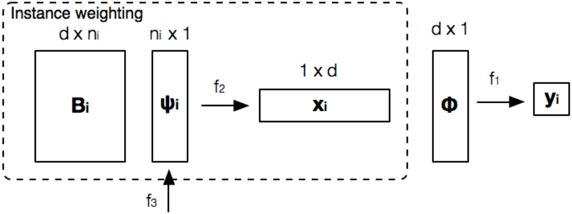
Proposed model: optimization objectives

- **RLS objectives:**

$$\psi_1, \ldots, \psi_m, \Phi = \underset{\psi_1,\ldots,\psi_m,\Phi}{arg\ min} \sum_{i=1}^{m} \left(\left(y_i - \Phi^T (B_i \psi_i) \right)^2 + \epsilon_1 ||\psi_i|| \right) + \epsilon_2 ||\Phi||^2$$

$$O = \underset{O}{arg\ min} \sum_{i=1}^{N} \sum_{j=1}^{n_i} \left(\psi_{ij} - O^T b_{ij} \right)^2 + \epsilon_3 ||O||^2$$

$$\text{subject to: } \psi_{ij} \geq 0 \ \forall i, j \text{ and } \sum_{i=1}^{n_i} \psi_{ij} = 1 \ \forall i. \qquad (5)$$



Learning with alternating steps

- inspired by alternating projections (Wagstaff'07), proceeds as follows:
    - → for each bag optimize f1 model for the instance weights s.t constraints (keep f2 fixed)
    - → optimize f1 model for the regression hyperplane (keep f1 fixed)
    - → optimize f3 model by keeping the other two fixed

```
Initialize($\psi_1, \ldots, \psi_N, \Phi, X$)
**while** not converged **do**
    **for** $B_i$ in $B$ **do**
        $\psi_i = cRLS(\Phi^T Bi, Y_i, \epsilon_1)$ # $f_1$ model
        $x_i = B_i \psi_i^T$
    **end for**
    $\Phi = RLS(X, Y, \epsilon_2)$ # $f_2$ model
**end while**
$\Omega = RLS(\{b_{ij} \forall i, j\}, \{\psi_{ij} \forall i, j\}, \epsilon_3)$ # $f_3$ model
```

Outline of the talk

1 Motivation

2 Multiple-instance learning

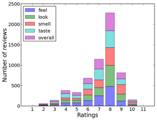
3 The proposed model

4 **Experiments**

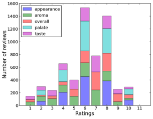
5 Conclusion

Datasets

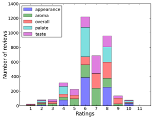
	Bags	Inst.	Dim.	Aspect ratings
BeerAdvocate	1,200	12,189	19,418	feel, look, smell, taste, overall
RateBeer (ES)	1,200	3,269	2,120	appearance, aroma, overall, palate, taste
RateBeer (FR)	1,200	4,472	903	appearance, aroma, overall, palate, taste
Audiobooks	1,200	4,886	3,971	performance, story, overall
Toys & Games	1,200	6,463	31,984	educational, durability, fun, overall
TED comments	1,200	3,814	957	sentiment (polarity)
TED talks	1,200	11,993	5,000	unconvincing, fascinating, persuasive, ingenious, long-winded, funny, inspiring, jaw-dropping, courageous, beautiful, confusing, obnoxious



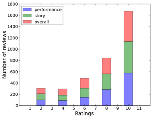
a) BeerAdvocate



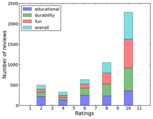
b) RateBeer (ES)



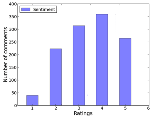
c) RateBeer (FR)



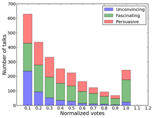
d) Audiobooks



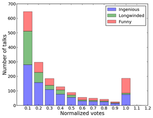
e) Toys & Games



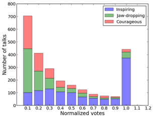
f) TED comments



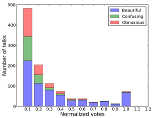
g) TED talks (classes 1 to 3)



h) TED talks (classes 3 to 6)



i) TED talks (classes 6 to 9)



j) TED talks (classes 9 to 12)

Experiments: aspect-rating prediction

	Review labels									
	BeerAdvocate		**RateBeer (ES)**		**RateBeer (FR)**		**Audiobooks**		**Toys & Games**	
Model \ Error	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**
AverageRating	14.20	3.32	16.59	4.31	12.67	2.69	21.07	6.75	20.96	6.75
Aggregated (ℓ_1)	13.62	3.13	15.94	4.02	12.21	2.58	20.10	6.14	20.15	6.33
Aggregated (ℓ_2)	14.58	3.68	14.47	3.41	12.32	2.70	19.08	5.99	18.99	5.93
Instance (ℓ_1)	12.67	2.89	14.91	3.54	11.89	2.48	20.13	6.17	20.33	6.34
Instance (ℓ_2)	13.74	3.28	14.40	3.39	11.82	2.40	19.26	6.04	19.70	6.59
Prime (ℓ_1)	12.90	2.97	15.78	3.97	12.70	2.76	20.65	6.46	21.09	6.79
Prime (ℓ_2)	14.60	3.64	15.05	3.68	12.92	2.98	20.12	6.59	20.11	6.92
Clustering (ℓ_2)	13.95	3.26	15.06	3.64	12.23	2.60	20.50	6.48	20.59	6.52
APWeights (ℓ_2)	**12.24**	**2.66**	**14.18**	**3.28**	**11.37**	**2.27**	**18.89**	**5.71**	**18.50**	**5.57**
vs. SVR (%)	*+16.0*	*+27.7*	*+2.0*	*+3.8*	*+7.6*	*+15.6*	+1.0	*+4.5*	*+2.6*	*+6.0*
vs. Lasso (%)	*+10.1*	*+15.1*	*+11.0*	*+18.4*	*+6.8*	*+11.8*	*+6.0*	*+6.9*	*+8.1*	*+11.9*
vs. 2^{nd} (%)	+3.3	+7.8	+1.5	*+3.3*	*+3.7*	*+4.9*	+1.0	*+4.5*	*+2.6*	*+6.0*

Table : Performance of aspect rating prediction (the lower the better) in terms of MAE and MSE (× 100) with 5-fold cross-validation. All scores are averaged over all aspects in each dataset. The scores of the best method are in **bold** and the second best ones are underlined.

Experiments: aspect-rating prediction (2/2)

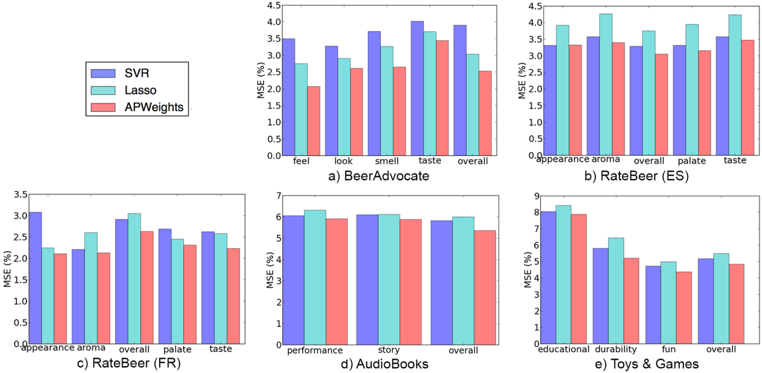


Figure : MSE scores of SVR, Lasso and APWeights for each aspect over the five review datasets.

Experiments: sentiment and emotion prediction

	Sent. Labels		Emo. labels	
	TED comm.		**TED talks**	
Model \ Error	**MAE**	**MSE**	**MAE**	**MSE**
AverageRating	19.47	5.05	17.86	6.06
Aggregated (ℓ_1)	17.08	4.17	15.98	5.03
Aggregated (ℓ_2)	16.88	4.47	15.24	4.97
Instance (ℓ_1)	17.69	4.37	16.48	5.30
Instance (ℓ_2)	16.93	4.24	16.10	5.57
Prime (ℓ_1)	17.39	4.37	15.98	5.78
Prime (ℓ_2)	18.03	4.91	16.74	5.94
Clustering (ℓ_2)	17.64	4.34	17.71	6.02
APWeights (ℓ_2)	**15.91**	**3.95**	**15.02**	**4.89**
APW vs SVR (%)	*+5.7*	*+11.5*	+1.5	+1.6
APW vs Lasso (%)	*+6.8*	+5.3	*+6.0*	*+2.9*
APW vs 2^{nd} (%)	*+5.7*	+5.3	+1.5	+1.6

Table : MAE and MSE (× 100) on sentiment and emotion prediction with 5-fold c.-v. Scores on TED talks are averaged over the 12 emotions.

- similar results are obtained with more sophisticated features (BOW tf-idf)

Examples: sentiment prediction

Sentences per comment	$\hat{\psi}_i$	$\hat{y}_i$	y_i
"Very brilliant and witty, as well as great improvisation."	0.64		
"I enjoyed this one a lot."	0.36	5.0	5.0
"That's great idea, I really like it!"	0.56	4.2	4.0
"I can't wait to try it, but first thing, I need a house with big windows, next year, maybe I can do that."	0.44		
"Unfortunately countries are not led by gifted children."	0.48		
"They are either dictated by the most extreme personalities who crave nothing but power or managed by politicians who are voted in by a far from gifted population."	0.52	2.4	2.0
"I am very disappointed by this, smug, cliched and missing so much information as to be almost (...)"''	0.43		
"No mention of ship transport lets say 50% of all material transport, no mention of rail transport, (...)"	0.29		
"I am sorry to be so negative, this just sounds like a sales pitch that he has given too many times (...)."	0.28	1.8	1.0

Table : Predicted sentiment for TED comments: y_i is the actual sentiment, $\hat{y}_i$ the predicted one, and $\hat{\psi}_i$ the estimated relevance of each sentence.

Examples: emotion prediction

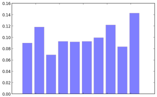
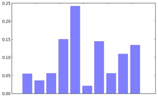
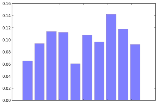
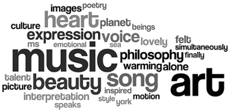
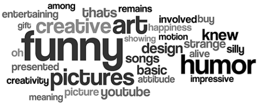
Class	Top comment per talk (according to weights ψ_i)	$\hat{\psi}_i$ distribution
beautiful	"The beauty of the nature. It would be more interesting just integrates his thought and idea into a mobile device, like a mobile, so we can just turn on the nature gallery in any time. The paintings don't look incidental but genuinely thought out, random perhaps, but with a clear grand design behind the randomness. Drawing is an art where it doesn't (...)"	
funny	"Funny story, but not as funny as a good 'knock, knock' joke. My favorite knock-knock joke of all time is Cheech & Chong's 'Dave's Not Here' gag from the early 1970s. I'm still waiting for someone to top it after all these years. [Knock, knock] 'Who is it?' the voice of an obviously stoned male answers from the other side of a door, (...)"	
courageous	"I was a soldier in Iraq and part of the unit represented in this documentary. I would question anyone that told you we went over there to kill Iraqi people. I spent the better part of my time in Iraq protecting the Iraqi people from insurgents who came from countries outside of Iraq to kill Iraqi people. We protected families men, women, and (...)"	

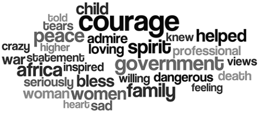
Table : Top comments for correctly predicted emotions in four TED talks and their distribution of weights.



b) beautiful



c) funny



d) courageous

Outline of the talk

1. Motivation
2. Multiple-instance learning
3. The proposed model
4. Experiments
5. **Conclusion**

Conclusion and perspectives

1. we proposed a promising MIR model for text regression tasks
   - models aspect ratings and instance contributions
   - discovers structure of labeled and unlabeled texts
2. first results on multi-aspect sentiment analysis based on MIR
   - competitive results with respect to SOA
   - instance relevance performs better than all other assumptions
   - interpretable output

Future work

→ test on sentence-level sentiment classification
→ experiment with other model settings, regularization and features
→ investigate instance weights for other NLP tasks (summaries, segmentation)

Thank you! Any questions or comments?