Explicit Document Modeling using Weighted Multiple-Instance Learning

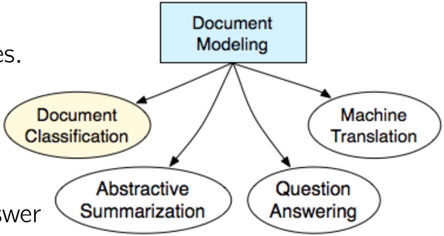
Nikolaos Pappas

October 11, 2016

Document modeling

"Representing the intrinsic relations of words or sentences and the semantic content of a document."

- Classification
  - Predict one or more categories.
- Summarization
  - Generate a summary.
- Question answering
  - Collect relevant facts and answer comprehension questions.



Goal of this talk

→ Present a mechanism which learns to focus on relevant regions.
→ Assess its merits on aspect rating prediction (EMNLP 2014).
→ Compare this mechanism to humans (SocialNLP@EMNLP 2016).

Example: aspect rating prediction of reviews

audible
an amazon company

Overall quality: poor [2/5]

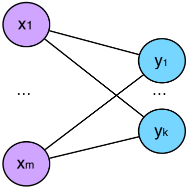
"Misleading as Sci-Fi" (review of *Solaris* narrated by *Allesandro Juliani* on Audible)

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 Lem was ahead of his time and despite some outdated technical items, the book shows excellent technical creativity.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!

Story: poor [2/5]

Narration: good [4/5]

Problem formulation

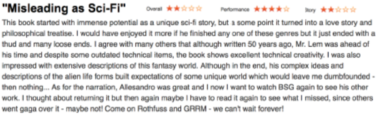


Given $\mathcal{D} = \{(x_i, y_i), \mid i = 1 \ldots m\}$, find $\Phi_k : \mathcal{X} \rightarrow \mathcal{Y}_k$

- The $x_i \in \mathbb{R}^d$ represents a review
- The $y_i \in \mathbb{R}^k$ are the k target aspect ratings

Challenges

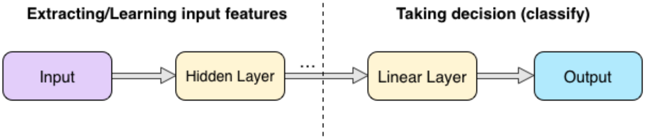
→ What input features to use?
→ How can we deal with the <u>weak</u> relation of the input to target labels?


"Misleading as Sci-Fi" Overall Performance Story

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!

Feature engineering and learning

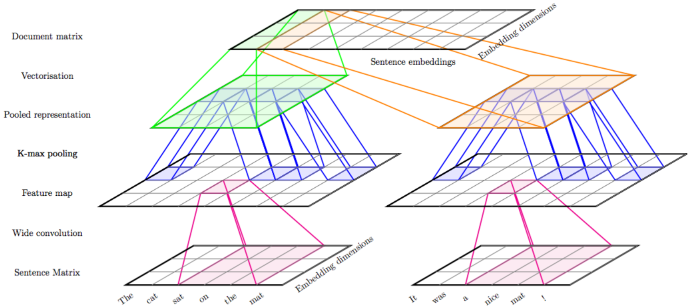
- BOW, n-grams, topic models and others (Pang and Lee, 2005), (Titov and McDonald, 2008), (Zhu et al., 2012)
- Autoencoders, convolutional or recursive NNs (Maas et al., 2011), (Mikolov et al., 2013), (Mesnil et al., 2014), (Tang et al., 2015)
- Train on segmented text i.e. sentences of each particular aspect or structured learning to capture label relations (McAuley et al., 2012)



Limitations

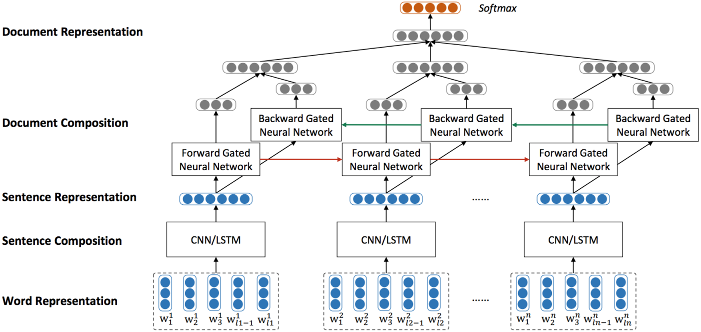
→ Treat the text globally and ignore the weak nature of labels
→ Make simplistic assumptions when aggreagating or pooling features
→ Offer few means for model interpretation

Convolutional networks



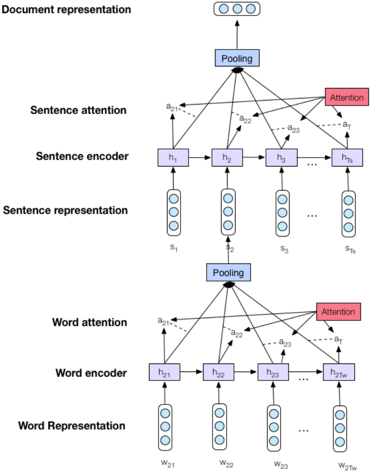
"Modelling, visualising and summarising documents with a single convolutional neural network", Misha Denil, Alban Demiraj, Nal Kalchbrenner, Phil Blunsom, Nando de Freitas, CoRR, 2014. (Denil et al., 2014)

Recursive networks



"Document Modeling with Gated Recurrent Neural Network for Sentiment Classification", Duyu Tang, Bing Qin, Ting Liu, EMNLP, 2015. (Tang et al., 2015)

Attention networks (cutting edge)



$$u_{it} = \tanh(W_w h_{it} + b_w)$$

$$\alpha_{it} = \frac{\exp(u_{it}^\top u_w)}{\sum_t \exp(u_{it}^\top u_w)}$$

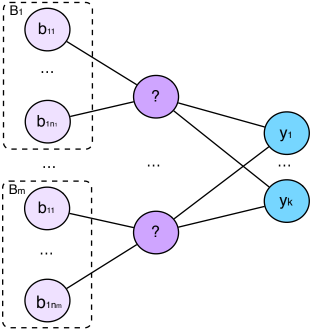
$$s_i = \sum_t \alpha_{it} h_{it}.$$

Improves many NLU tasks

- reading comprehension
- question answering
- machine translation
- document classification

"Hierarchical Attention Networks for Document Classification", Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J. Smola, Eduard H. Hovy, NAACL, 2016. (Yang et al., 2016)

Multiple-instance learning



Given $\mathcal{D} = \{(b_{ij}, y_i) \mid j = 1 \dots n_i\}^m$,
find $\Phi_k : \mathcal{B} \xrightarrow{?} \mathcal{X} \rightarrow \mathcal{Y}_k$

- The bag B_i is a review represented by n_i instances b_{ij}, its sentences
- The labels $y_i \in \mathbb{R}^k$ are the aspect ratings of the review
- The exemplar (representation) $x_i \in \mathbb{R}^d$ of B_i is initially unknown

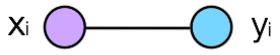
Advantages

→ Several input assumptions (Aggregated, Instance, Prime, Clustering)
→ Subsumes traditional supervised regression (Aggregated)
→ Better suited for weak labels, interpretable and flexible

Structural assumptions

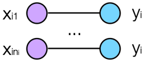
1. **Aggregated instances**: sum or average instances

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



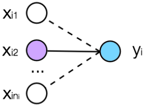
2. **Instance-as-example**: instances inherit bag labels

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



3. **Prime instance**: a single instance is selected

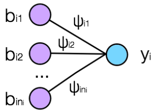
$$f \leftarrow D_{pri} = \{(b_i^p, y_i) \mid i = 1, \ldots, m\}$$
$$\hat{y}(B_i) = mean(\{f(b_{ij}) \mid j = 1, \ldots, n_i\})$$



Instance relevance (weighted average)

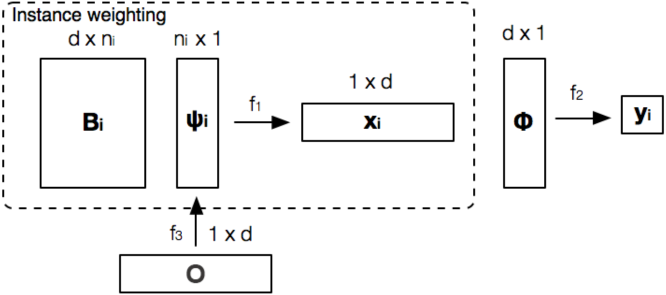
Inspired from method proposed by Wagstaff and Lane (2007):

$$x_i = \sum_{j=1}^{n_i} \psi_{ij} b_{ij},\ \psi_{ij} \geq 0 \ \text{ and } \sum_{j=1}^{n_i} \psi_{ij} = 1$$



1. Models both instance weights and target labels
    - Target labels model: $\hat{y}_i = f(\Phi, B_i) = \Phi^T(B_i\psi_i)$
    - Instance weights model: $\hat{\psi}_i = f(O, B_i)O^T B_i$
2. Defines loss based on regularized least squares
    - Supports large datasets and high dimensionality $\mathcal{O}(md^2)$
    - Adapts to domain data through regularization

Optimization objectives



- Target label model $g(f_1, f_2)$:

$$\mathcal{L}(\Psi, \Phi) = \sum_{i=1}^{m} \left(y_i - \Phi^T(B_i\psi_i)\right)^2 + \Omega(\Psi, \Phi) \ s.t.$$
$$\psi_{ij} \geq 0 \quad \text{and} \quad \sum_{j=1}^{n_i} \psi_{ij} = 1$$

- Instance weights model f_3:

$$\mathcal{L}(O) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left(\psi_{ij} - O^T b_{ij}\right)^2 + \Omega(O)$$

Learning parameters consecutively

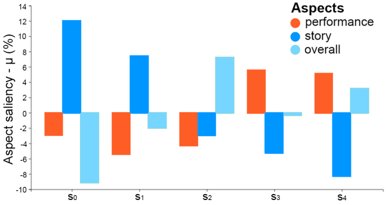
Alternating projections

1. `Until converged`
   1.1 `Optimize weights` Ψ_i `(keep` Φ `fixed)`
   1.2 `Optimize coefficients` Φ `(keep` Ψ `fixed)`
2. `Optimize coefficients` O

Testing on unseen bags

Predicts the bag's label $\hat{y}_i$ and its instance weights $\hat{\psi}_i$

$$\hat{y}_i = \Phi^T B_i' \hat{\psi}_i = \Phi^T B_i' (O^T B_i')$$



MIR weights for a book review.

Learning parameters jointly

Based on stochastic gradient descent

$$\sigma(B_i, O) = P(\psi = i|x) = \frac{e^{(O^T B_i)}}{\sum_{k=1}^{n_i} e^{(O^T B_{ik})}}$$

$$O, \Phi = \underset{O,\Phi}{arg\ min} \sum_{i=1}^{m} (y_i - \Phi^T(B_i \cdot \sigma(B_i, O)))^2 + \Omega(\Phi, O)$$

- Preserves constraints of instance relevance assumption
- Achieves similar performance to alternating projections
- Makes the learning procedure more scalable

Shared material

$\rightarrow$ Code: wmil, wmil-sgd

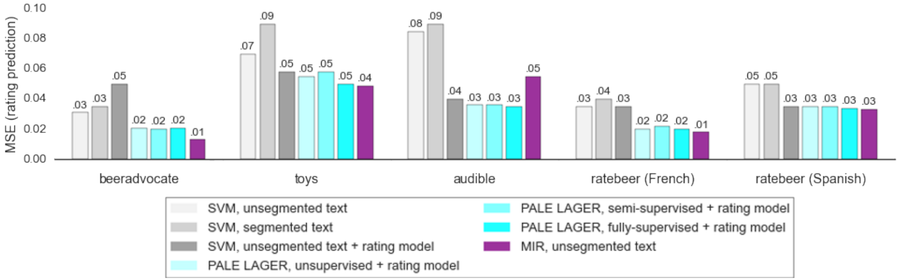
`https://github.com/nik0spapp/`

Datasets, protocol and metrics

Data	Bags	Instances	BOW Dim.	Labels
BeerAdvocate	1,586,259	16,883,058	19,418	5 aspects
Toys & Games	373,974	2,105,647	31,984	4 aspects
Audible	10,989	44,487	3,971	3 aspects
RateBeer (FR)	17,998	105,569	903	5 aspects
RateBeer (ES)	1,259	3,511	2,120	5 aspects
TED comments	1,200	3,814	957	1 sentiment
TED talks	1,203	12,023	5,000	14 emotions

- Two series of experiments
  - Comparison with previous studies (train/test on uniform split)
  - Effects of model design choices (5-fold c-v on subsets of same size)
- Parameters optimized on a subset of the training data
- Error metric on numerical prediction $\frac{1}{k}\sum_{i=1}^{k}(y_i - \hat{y}_i)^2$

Performance on aspect rating prediction



- Weighted MIR achieves lower error than:
  - Methods trained with segmented text (SVM, PALE LAGER[1])
  - Structured learning methods (Structured SVM, PALE LAGER[1])

[1] Graphical model proposed in (McAuley et al., 2012).

Comparison of structural assumptions

Mean Squared Error x 100 (%)

Methods	**beeradvocate**	**toys**	**audible**	**ratebeer-fr**	**ratebeer-sp**
Aggregated MIR	3.68	5.93	2.70	5.99	3.41
Instance MIR	3.28	6.59	2.40	6.04	3.39
Prime MIR	3.64	6.92	2.98	6.59	3.68
Clustering MIR	3.26	6.52	2.60	6.48	3.64
Weighted MIR	**2.66**	**5.57**	**2.27**	**5.71**	**3.28**

- Strong supervision (Aggregated) is not the optimal assumption
- Instance relevance mechanism is superior to other alternatives
    - All regions are useful but to a different extent
    - The relevance of each region depends on the task

Independence from the feature space

	BOW		TF-IDF		word2vec	
Model \ Error	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**
Aggregated (ℓ_1)	17.08	<u>4.17</u>	16.59	<u>3.97</u>	16.03	3.84
Aggregated (ℓ_2)	<u>16.88</u>	4.47	<u>16.25</u>	4.16	<u>14.62</u>	<u>3.30</u>
Instance (ℓ_1)	17.69	4.37	18.11	4.50	16.37	3.86
Instance (ℓ_2)	16.93	4.24	16.88	4.23	15.60	3.67
Prime (ℓ_1)	17.39	4.37	17.72	4.43	16.13	3.89
Prime (ℓ_2)	18.03	4.91	17.10	4.29	15.71	3.72
Ours (ℓ_2)	**15.97**	**3.97**	**15.36**	**3.63**	**14.25**	**3.29**

- Our mechanism is beneficial regardless of the input features
- This suggests that it may be combined with feature learning
    - Recent studies confirm this idea! (e.g. attention networks)

Comparing mechanism to humans

- Capture human attention to sentences when attributing categories (aspect ratings) to documents (audiobook reviews)
  - How much does each sentence explain the given aspect rating?
  - 100 reviews, 1,662 sentences and 3 aspects, 1-5 scale
- Main goal:
  - Train a document attention model with weak labels (50k reviews)
  - Compare the attention mechanism to humans on a test set

Shared material

→ Human attention in document classification dataset

`https://www.idiap.ch/paper/hatdoc`

Crowdsourcing task

Read the highlighted sentence from the review of the audiobook Ghost of a Potion: Magic Potion Mystery Series #3 by user Mario:

> *My problem with the first two books has been Carly and Dylan's relationship because they all but ignored the reason it ended in the first place; Dylan's Mama.* ***Since it was one the main plot points I have no real complaints about it now.*** *Unlikely. I have read the previous two books in the series and while I like the well enough there not the kind of stories I would listen to again. Carla Mercer-Meyer is a good narrator but she is just not as good as other "southern" narrators I have listen to before. It's hard to really enjoy a performance when you know there is someone who could have done a better job. Not laugh or cry but a few of the twist did surprise me. If you enjoyed the first two books there is no reason you won't enjoy this one. My favorite character is still Delia and the blooming friendship that is developing between her and Carly.*

Question:

How much does the highlighted sentence explain a **Story** aspect rating of **3 out of 5** (neutral) ?

- ◯ Not at all ◯ A little ◯ Moderately ◯ Rather well ◯ Very well

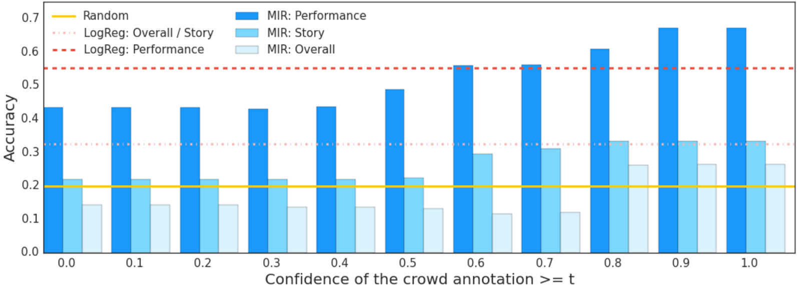
Examples: positive review

Ove. (5/5)	Perf. (5/5)	Story (5/5)	Document (id=969066)
0.45	0.56	0.18	Narrated by one of my favorite narrators, Scott Brick, I found this offering by Harlan Coben to be one of their best - for them both.
0.18	0.22	0.36	I found it very difficult to "put this down".
0.36	0.22	0.45	It is one of those no-brainer 5 star thillers!

Examples: negative review

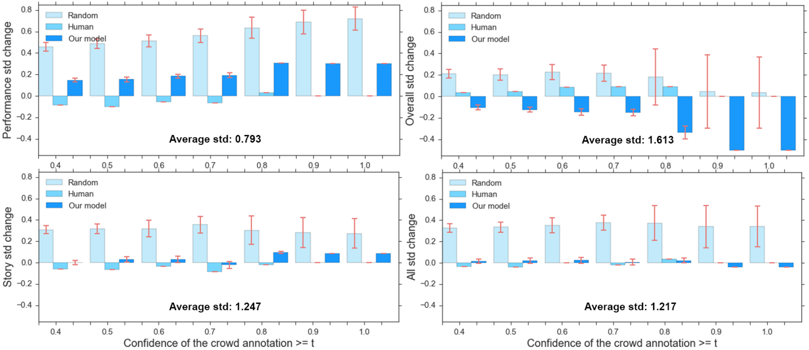
Ove. (2/5)	Perf. (3/5)	Story (1/5)	Document (id=319628)
0.14	0.07	0.07	This little pamphlet essentially advises you to be mindful of what you are feeling.
0.36	0.14	0.29	That's always good advice, but this presentation is poor: Very little advice or examples on how to put his idea into practice, very repetitive (all this info could have been on 1 page - in fact, he sums it up on an index card that he suggests you write up), and for some odd reason he insults The Affordable Care Act, out of nowhere.
0.21	0.21	0.21	If the author put some meat into this, it might have been a more helpful purchase.
0.21	0.29	0.21	When listening I felt like I was sitting at a timeshare sales pitch in exchange for free ski lift tickets.
0.07	0.29	0.21	Try Pema Chodron (any book) or the RAIN meditation by Tara Brach .

Human attention prediction (exact match)



- Positive correlation between machine and human attention, especially for sentences with high human agreement
- Best accuracy on *performance* aspect (least ambiguous)
- Compares favorably to LogReg (oracle)

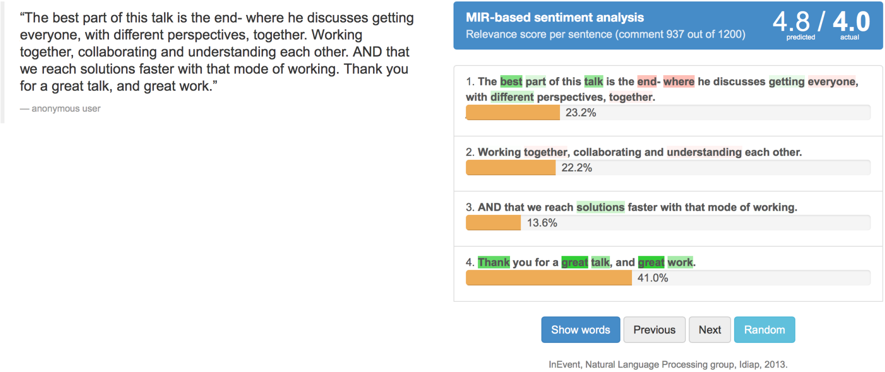
Reliability analysis



- Consistently outperforms ‘Random’ for all aspects and levels
- Comparable results to qualified humans for *Performance* and *Overall*

Demo: sentiment prediction

MIR results over 1,200 TED comments: sentiment ratings



Demo: emotion-based recommendation

MIR results over 1,000 TED transcripts: emotion ratings (12 dim.)

Recommended talks
Top 8 based on selected similarity

1. **Patricia Kuhl: The linguistic genius of babies**
2. **Richard Dawkins: Why the universe seems so strange**
3. **Penelope Boston says there might be life on Mars**
4. **Juan Enriquez: The next species of human**
5. **Ron Eglash: The fractals at the heart of African designs**
6. **Sebastian Seung: I am my connectome**
7. **VS Ramachandran: The neurons that shaped civilization**
8. **John Delaney: Wiring an interactive ocean**

Emo-based ▾ | Get a random talk | How it works

InEvent, Natural Language Processing group, Idiap, 2013.

Stefano Mancuso: The roots of plant intelligence

unconvincing	13%	fascinating	91%
persuasive	27%	ingenious	31%
longwinded	9%	funny	14%
inspiring	52%	jaw-dropping	25%
courageous	11%	beautiful	19%
confusing	5%	obnoxious	5%

Introduction

Background and motivation

Supervised learning

Related work

Explicit Document Modeling

Multiple-instance learning

Structural assumptions

Instance relevance mechanism

Experiments

Aspect-based rating prediction

Comparing mechanism to humans

Qualitative results (demos)

Conclusion

Conclusion

- Document modeling benefits from a weakly supervised objective
- MIL improves accuracy and captures structural information
  - Learns to focus on relevant parts of the input (assumptions)
  - Provides meaningful and interpretable weights
  - Equivalent to NN attention mechanisms
- Extensions:
  → Attention with external knowledge (‘memory’)
  → Multiple passes of attention (‘reasoning’)
  → Other modalities (visual, acoustic)

Thank you!

Acknowledgments

Publications

1. Nikolaos Pappas and Andrei Popescu-Belis, Human versus Machine Attention in Document Classification, *Proceedings of EMNLP 2016 SocialNLP workshop (SocialNLP@EMNLP)*, Austin, Texas, 2016.
2. Nikolaos Pappas, Learning Explainable User Preference and Sentiment for Information Filtering, *cole Polytechnique Fdrale de Lausanne, PhD thesis*, 2016.
3. Nikolaos Pappas and Andrei Popescu-Belis, Adaptive Sentiment-Aware One-Class Collaborative Filtering, *Expert Systems with Applications (ESWA)*, 43(1):23–41, 2015.
4. Nikolaos Pappas and Andrei Popescu-Belis, Combining Content with User Preferences for Non-Fiction Multimedia Recommendation: A Study on TED Lectures, *Multimedia Tools and Applications (MTAP), Special Issue on Content Based Multimedia Indexing*, 74(4):1175–1197, 2015.
5. Nikolaos Pappas and Andrei Popescu-Belis, Explaining the Stars: Weighted Multiple-Instance Learning for Aspect-Based Sentiment Analysis, *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 455–466, Doha, Qatar, 2014.
6. Nikolaos Pappas and Andrei Popescu-Belis, Sentiment Analysis of User Comments for One-Class Collaborative Filtering over TED Talks, *36th ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), Short papers*, pages 773–776, Dublin, Ireland, 2013.
7. Nikolaos Pappas and Andrei Popescu-Belis. Combining Content with User Preferences for TED Lecture Recommendation, *11th International Workshop on Content Based Multimedia Indexing (CBMI)*, Veszprém, pages 47–52, Hungary, 2013

References I

Misha Denil, Alban Demiraj, Nal Kalchbrenner, Phil Blunsom, and Nando de Freitas. Modelling, visualising and summarising documents with a single convolutional neural network. *CoRR*, abs/1406.3830, 2014.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 142–150, Portland, OR, USA, 2011.

Julian McAuley, Jure Leskovec, and Dan Jurafsky. Learning attitudes and attributes from multi-aspect reviews. In *Proceedings of the 12th IEEE International Conference on Data Mining*, ICDM '12, pages 1020–1025, Brussels, Belgium, 2012. doi: 10.1109/ICDM.2012.110.

Grégoire Mesnil, Tomas Mikolov, Marc'Aurelio Ranzato, and Yoshua Bengio. Ensemble of generative and discriminative techniques for sentiment analysis of movie reviews. *CoRR*, abs/1412.5335, 2014. URL http://arxiv.org/abs/1412.5335.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In Christopher Burges, Lon Bottou, Max Welling, Zoubin Ghahramani, and Kilian Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119, 2013.

Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 115–124, Ann Arbor, Michigan, 2005. doi: 10.3115/1219840.1219855.

Duyu Tang, Bing Qin, and Ting Liu. Document modeling with gated recurrent neural network for sentiment classification. In *EMNLP*, 2015.

Ivan Titov and Ryan McDonald. Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pages 111–120, Beijing, China, 2008. doi: 10.1145/1367497.1367513.

Kiri L. Wagstaff and Terran Lane. Salience assignment for multiple-instance regression. In *ICML 2007 Workshop on Constrained Optimization and Structured Output Spaces*, Corvallis, OR, USA, 2007.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J. Smola, and Eduard H. Hovy. Hierarchical attention networks for document classification. In *HLT-NAACL*, 2016.

Jingbo Zhu, Chunliang Zhang, and Matthew Y. Ma. Multi-aspect rating inference with aspect-based segmentation. *IEEE Trans. on Affective Computing*, 3(4):469–481, 2012. ISSN 1949-3045. doi: 10.1109/T-AFFC.2012.18.