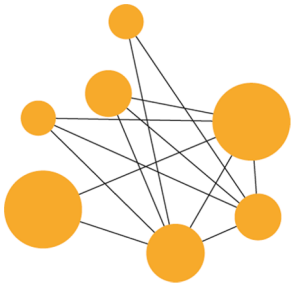
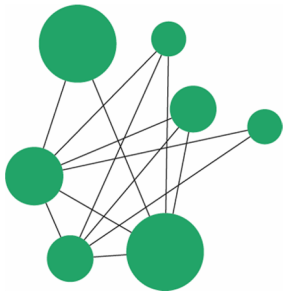
Reviewing Data Science Research: Evaluating Conclusion sections

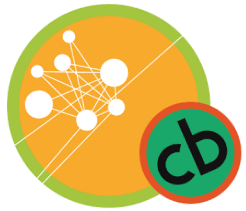
With examples from machine learning and natural language processing

Kevin Bretonnel Cohen
Director, Biomedical Text Mining Group, University of Colorado School of Medicine; Emeritus D'Alembert Chair in Natural Language Processing for the Biomedical Domain, Université Paris-Saclay





kevin.cohen@gmail.com
http://compbio.ucdenver.edu/Hunter_lab/Cohen



Karën Fort, Margot Mieskes, and Aurélie Névéol

Kevin Bretonnel Cohen,
UCSOM

Karën Fort,
Sorbonne Université / Loria

Margot Mieskes,
h_da Darmstadt

Aurélie Névéol,
Université Paris Saclay, CNRS, LIMSI

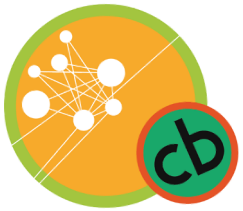
The *Conclusions section* versus *a conclusion*

(Discussion and) Conclusion(s) section

- Review of the paper (question, methods, materials, findings, and meaning)
- "Future work" *or* limitations *or* remaining open questions *or*…
- A conclusion

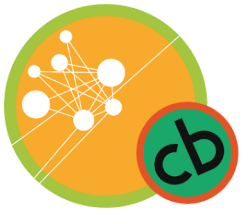
A conclusion

- "the last part of something" (Merriam-Webster)
- "a reasoned judgment" (M-W)
- "the necessary consequence of two or more propositions taken as premise" (Merriam-Webster)
- Finding a paper's conclusion:
  - *We have shown that…*



WE ARE TALKING ABOUT "CONCLUSIONS" SECTIONS BECAUSE THEY ARE A NEAR-UNIVERSAL SECTION OF PAPERS IN OUR FIELD

PRIMARY QUESTION: IS THE CONCLUSION SUPPORTED BY THE METHODOLOGY AND OUTCOMES?



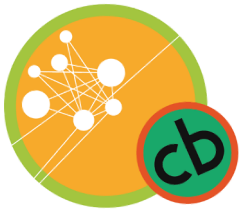
Exercise: Do the methodology and outcomes support the conclusion?

Methodology/Outcome

- Compared:
  1. An SVM classifier with bag of words as the only feature
  2. An SVM classifier with GLOVE word embeddings and bag of words as features
- Outcome: SVM classifier with GLOVE word embeddings as features outperforms SVM classifier with bag of words as the only feature

Conclusion

- Deep learning is the best approach to classification tasks



Exercise: Do the methodology and outcomes support the conclusion?

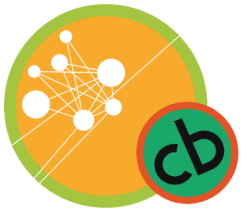
Methodology/Outcome

- Compared:
  1. An SVM classifier with bag classifier with bag of words as the only feature

Conclusion

- Neural networks are the best approach to

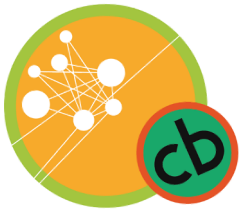
No. The paper compared two feature sets. The fact that one feature set was partially produced by a neural network does not tell us anything about neural networks.



USUALLY SCORED AS PART OF ACCEPTANCE RECOMMENDATION

Typical parts of a (Discussion and) Conclusions section

- Review of the paper:
  - Research question
  - What was done to answer it
  - With what materials
  - What the findings were
  - What the findings mean with respect to the research question
  - Possible alternative explanations
  - What that means with respect to some broader context
- "Future Work"
- A conclusion



Exercise: Do the methodology and outcome support the conclusion?

Methodology/Outcome

1. A specific pipeline is applied to a publicly-available dataset
2. A single value is reported
3. Value is compared to previously reported scores
4. The value is higher than previously reported scores
5. Experimental details are under-described
6. No code available

Conclusion

- We have state-of-the-art performance on this dataset

Exercise: Do the methodology and outcome support the conclusion?

Methodology/Outcome

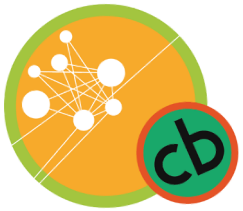
1. A specific pipeline is applied to a publicly-
2. [illegible]
3. [illegible]
4. [illegible]
5. [illegible]
6. No code available

Conclusion

- We have state-of-the-art performance on this

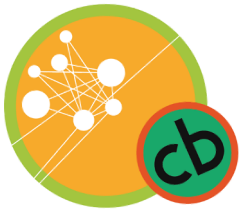
No. The paper provides about as much detail as an advertisement. Under-described methods and unavailable code mean that the paper does not actually describe the methods and the outcome is not valid.

EXTENDED EXERCISE: DO THE METHODOLOGY AND OUTCOME SUPPORT THE CONCLUSION?



- **Task**: Completely novel document classification target
- **Data**: 100 positive instances, 100 controls
- **Method**: Bag of words, three classifiers
- **Result**: Best F-measure = **0.85**

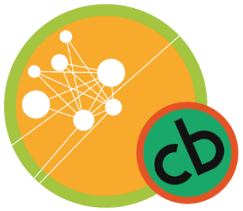
0.85



- **Task**: Completely novel document

No. The methodology did not actually do the task that was intended.

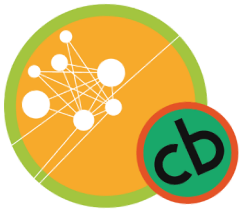
- **Result**: Best F-measure = **0.85**



Here are the materials…

- Task: Completely novel document classification task
- Data: 100 positive instances, 100 controls
- Method: Bag of words, three classifiers
- Result: Best F-measure = 0.85

- 100 positive instances, **extracted from PDFs**
- 100 controls, text **extracted from PubMedCentral XML**



Natural language processing research is rife with confounds

The results show that a small change in tokenization strategy can improve a mediocre 2006 TREC genomics submission (MAP average: 29%) to the top quarter of the submissions (36%-54%). Normalization and splitting compounds to multiple terms shows to be very beneficial for the tested IR models which assume term independence in both queries and documents. We expect that incorporation of proximities of related terms in the retrieval model will even further improve retrieval performance.

Trieschnigg, Dolf, Wessel Kraaij, and Franciska de Jong. "The influence of basic tokenization on biomedical document retrieval." *SIGIR* 2007.

Table 4. Effect of data balance, holding all other factors constant.

POSITIVE INSTANCES	NEGATIVE INSTANCES	*F*-MEASURE
100	100	0.82 ± 0.03
100	200	0.80 ± 0.03
100	300	0.74 ± 0.04
100	400	0.70 ± 0.04

Cohen, K. B., Glass, B., Greiner, H. M., Holland-Bouley, K., Standridge, S., Arya, R., ... Pestian, J., & Glauser, T. (2016). Methodological issues in predicting pediatric epilepsy surgery candidates. *Biomedical Informatics Insights*.

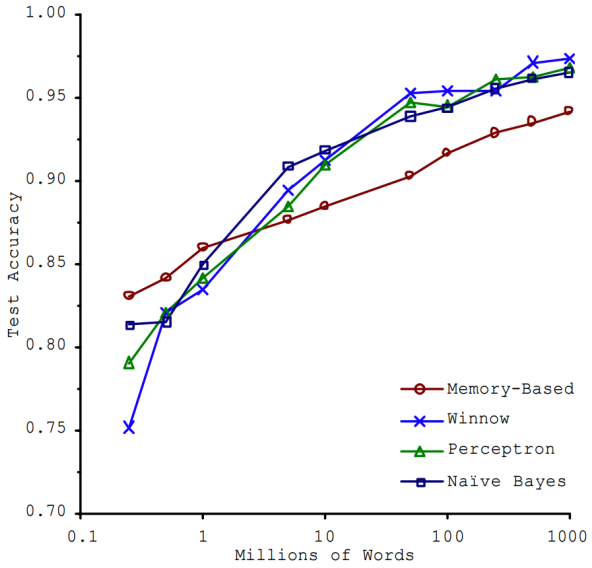
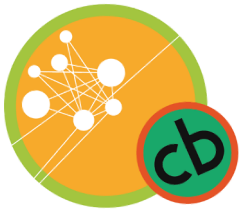


Figure 1. Learning Curves for Confusion Set Disambiguation

Banko, Michele, and Eric Brill. "Scaling to very very large corpora for natural language disambiguation." ACL 2001.

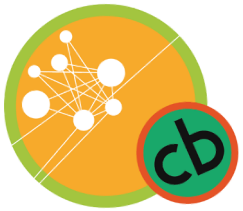
For more on how to think about *Conclusions* sections, see this work on spin…

- Koroleva, Anna. "Vers la detection des affirmations inappropriées dans les articles scientifiques." RECITAL 2017.
- Koroleva, Anna, and Patrick Paroubek. "Automatic detection of inadequate claims in biomedical articles: First steps." MEDA 2017.
- Koroleva, Anna, and Patrick Paroubek. "Annotating Spin in Biomedical Scientific Publications: the case of Random Controlled Trials (RCTs)." *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. 2018.
- Koroleva, A. "Assisted authoring for avoiding inadequate claims in scientific reporting." (2020).
- Lazarus, C., Haneef, R., Ravaud, P., Hopewell, S., Altman, D. G., & Boutron, I. (2016). Peer reviewers identified spin in manuscripts of nonrandomized studies assessing therapeutic interventions, but their impact on spin in abstract conclusions was limited. *Journal of clinical epidemiology*, *77*, 44-51.



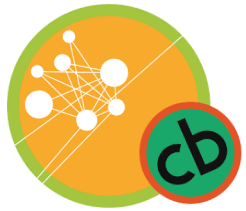
…and this work on confounds in natural language processing…

- Cohen, K. Bretonnel (2021) "Writing about data science research. " Cambridge University Press.
- Hirst, Graeme, Yaroslav Riabinin, and Jory Graham. "Party status as a confound in the automatic classification of political speech by ideology." 2010.
- Kumar, Sachin and Wintner, Shuly and Smith, Noah A. and Tsvetkov, Yulia (2019) "Topics to Avoid: Demoting Latent Confounds in Text Classification." EMNLP.
- Pavalanathan, Umashanthi and Eisenstein, Jacob (2015) "Confounds and Consequences in Geotagged Twitter Data." EMNLP.
- Pryzant, R., Shen, K., Jurafsky, D., & Wagner, S. (2018, June). Deconfounded lexicon induction for interpretable social science. *NAACL*.



…and this work on what results mean

- Cohen, K. Bretonnel (2021) "Writing about data science research." Cambridge University Press.
- Hand, David J. "Classifier technology and the illusion of progress." *Statistical Science* (2006): 1-14.
- Pedersen, Ted. "Empiricism is not a matter of faith." Computational Linguistics 34.3 (2008): 465-470.
- Steedman, Mark. "On becoming a discipline." *Computational Linguistics* 34.1 (2008): 137-144.
- Wu, S., Miller, T., Masanz, J., Coarr, M., Halgrim, S., Carrell, D., & Clark, C. (2014). "Negation's not solved: generalizability versus optimizability in clinical natural language processing." *PLoS ONE*, 9(11).



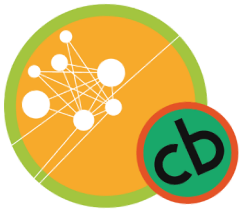
Karën Fort, Margot Mieskes, and Aurélie Névéol

Kevin Bretonnel Cohen,
UCSOM

Karën Fort,
Sorbonne Université / Loria

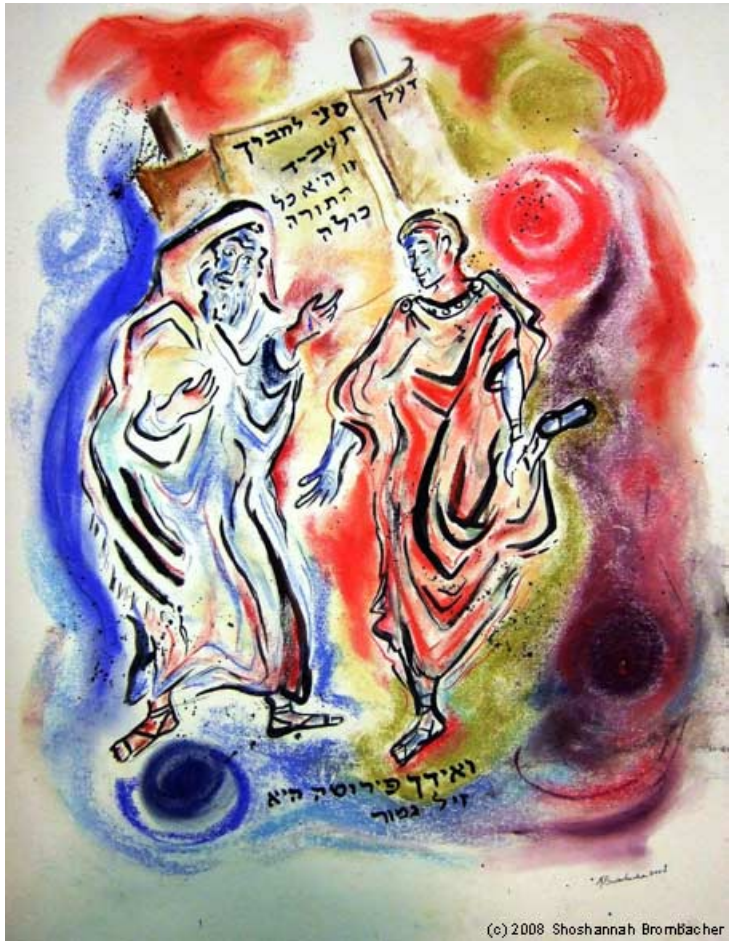
Margot Mieskes,
h_da Darmstadt

Aurélie Névéol,
Université Paris Saclay, CNRS, LIMSI



DESCRIBE A PAPER SUCH THAT THE METHODOLOGY AND OUTCOMES SUPPORT THE CONCLUSION

"…all the rest is commentary. Now go and study."



Artist: Shoshannah Brombacher