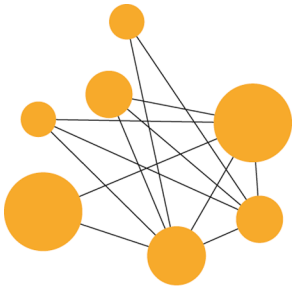
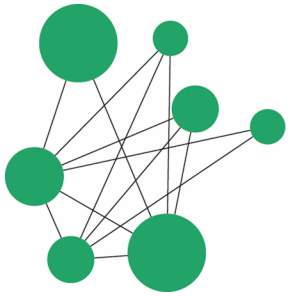
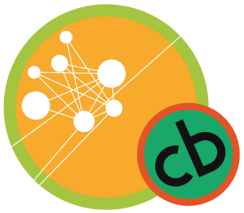
Reviewing Natural Language Processing Research: Discussion/Results/Methods sections

Kevin Cohen,
Karën Fort,
Aurélie Névéol,
Margot Mieskes,
Anna Rogers



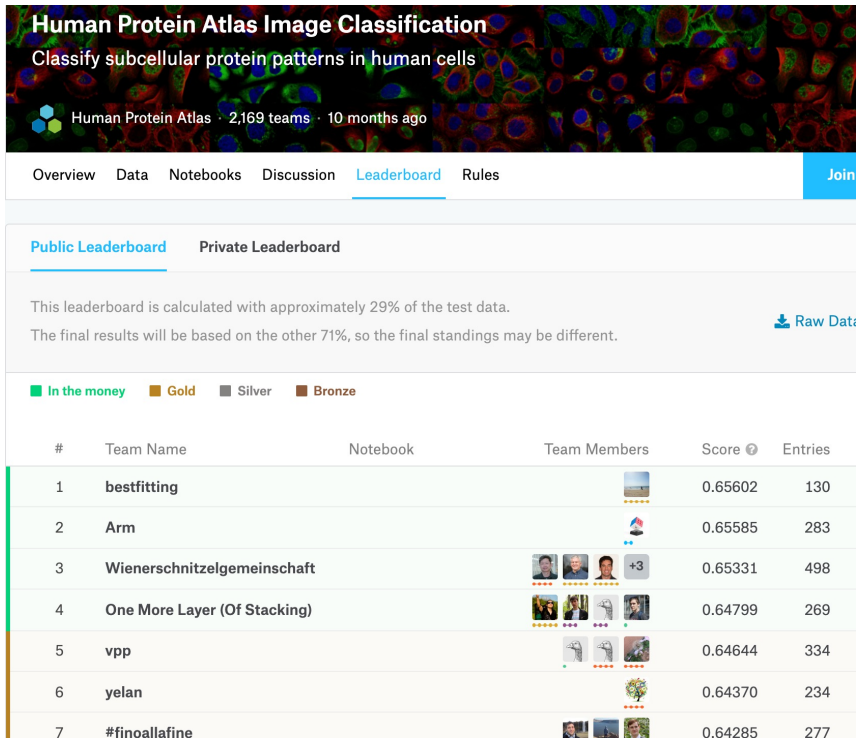


Karën Fort, Margot Mieskes, Aurélie Névéol, and Anna Rogers



Two techniques of reviewing: Kaggle versus science/engineering

Leaderboard model



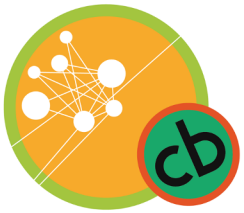
Smarter people than me have given this example—Ken Church, Christopher Manning, Bonnie Webber?

Science/engineering

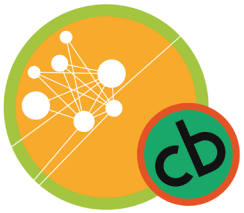
Table 5. **Classification using Gene Ontology concepts** Five-fold cross validation performance of five binary classifiers when providing Gene Ontology concepts as features. Results from both unbalanced and balanced training sets are shown. The highest F-measure is bolded. The baselines provided are OneR (one-node decision tree), Naive Bayes, and randomly assigning classes (median of 5 random assignments).

Classifier	GOA curated P/R/F	NLP abstracts P/R/F	NLP full-text P/R/F
Unbalanced Training			
Random			
OneR			
Naive Bayes			
Random Forest			
SMO			
LibSVM			
Balanced Training			
Random			
OneR			
Naive Bayes			
Random Forest			
SMO			
LibSVM			

Funk, Christopher S., Lawrence E. Hunter, and K. Bretonnel Cohen. "Combining heterogenous data for prediction of disease related and pharmacogenes." In *Pac. Symp. Biocomp. 2014*, pp. 328-339.

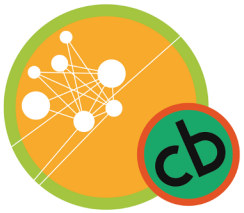


EXERCISE:
IS THE CONCLUSION OF MY EXPERIMENT SUPPORTED BY THE METHODOLOGY AND THE OUTCOMES?



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

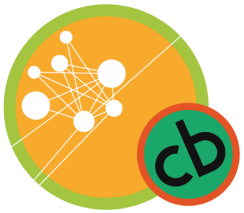
0.85



- **Task**: Completely novel document classification target
- **Data**: 100 positive instances, 100 controls
- **Method**: Bag of words, three classifiers

0.85

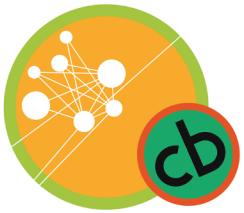
CONCLUSION: State-of-the-art results on a novel classification task.



- **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

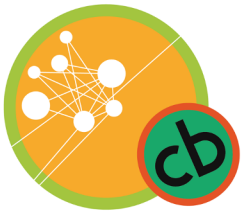
- 100 positive instances, **extracted from PDFs**
- 100 controls, text **extracted from PubMedCentral XML**

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

Checklist for *Discussion* section

…à la recherche des points forts (merci, Karën)

- ❑ There is a conclusion
- ❑ Conclusion is supported by methodology and outcomes
- ❑ Claims do not over-extrapolate from results
- ❑ Claims are not *overly* hedged



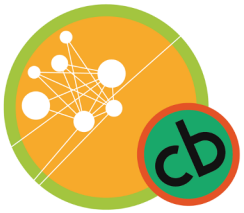
Step One: Is there a conclusion?

Conclusions

- "Best" system can change completely depending on the amount of training data. (Banko and Brill)
- Choice of tokenizer alone can move a system from the bottom third of TREC performers to the top third of TREC performers. (Trieschnigg et al.)
- Using pretrained word embeddings requires that data have the same tokenization. (Goldberg)

Not conclusions

- We achieved state-of-the-art performance on this task.
- We report the HyperCool system.
- Our contribution is the first combination of a convolutionally recurrent neural network with bidirectional long short-term memory and attention with softmax and forward-backward least squares estimation of 229 hyperparameters and…



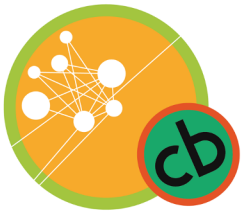
If there is no conclusion…

"The better the numbers are, the more important it is to reject the paper. We can't afford papers that report results without insights."

-- Kenneth Church

If there is a conclusion…

- Conclusion is supported by methodology and outcomes
- Claims do not over-extrapolate from results
- Claims are not *overly* speculative



Three common overgeneralizations

- From English to natural language
- From the Wall Street Journal to English
- From my amount of training data to *any* amount of training data

Adapted from Emily Bender because I can't find her tweet on this topic

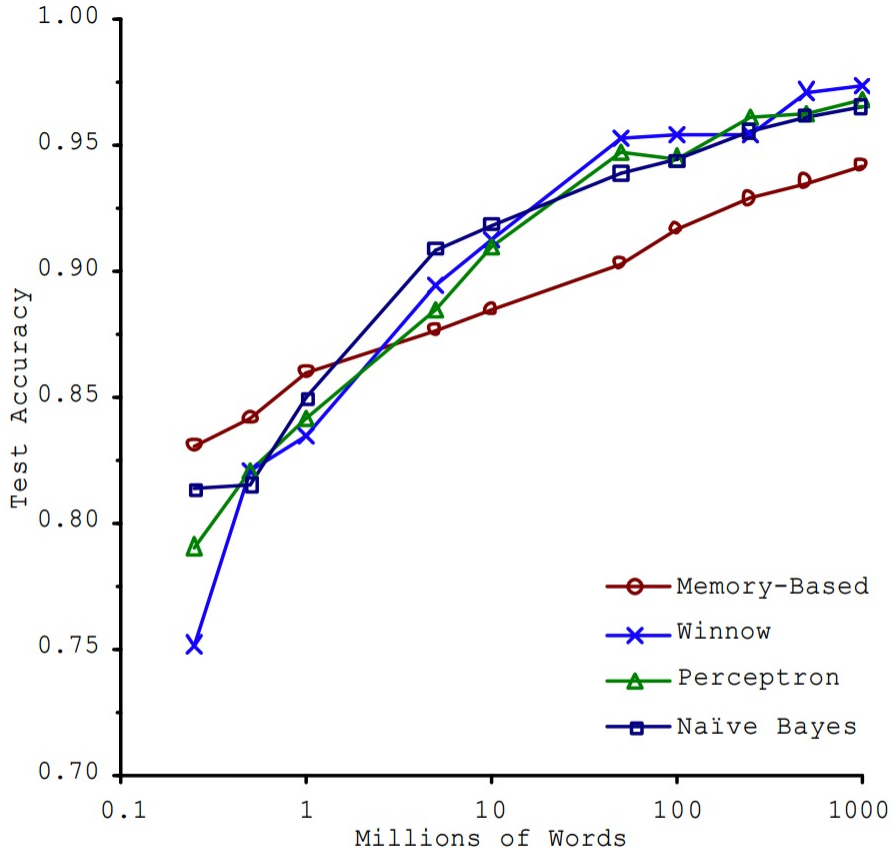
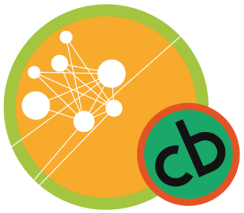


Figure 1. Learning Curves for Confusion Set Disambiguation

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

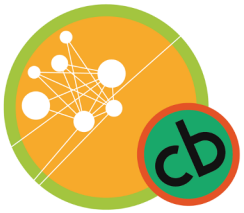
For more on how to think about *Conclusions* sections, see this work on spin…

- 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.



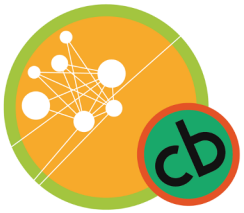
Parenthèse : listes de verification mènent facilement aux patrons d'écriture

- Code is/is not publicly available. This increases/reduces the potential impact of the work.
- The experimental set-up looks at many/multiple/only a small number of/only one variable that could reasonably be expected to affect the results. This is a big strength/serious weakness of the work.



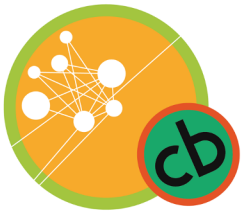
…jusqu'à…

The innovation of the work is asserted to be a unique combination of --, --, and --. However, the paper does not lay out any particular reason why this combination of approaches should be expected to yield any better performance than other, competing combinations of approaches. Without this level of analysis or understanding, the potential impact of the work is limited. **The paper would benefit from** providing a rationale for why this combination of approaches might be interesting in some specific way.



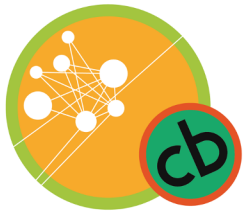
A checklist for the *Results* section

- ❑ Parametric statistics used only with normally distributed data
- ❑ Measures of dispersion are reported
- ❑ Graphs:
  - ❑ Axes are not truncated
  - ❑ Axes are labelled
- ❑ Error analysis is present
- ❑ Error analysis is non-trivial
- ❑ Alternative analyses are considered

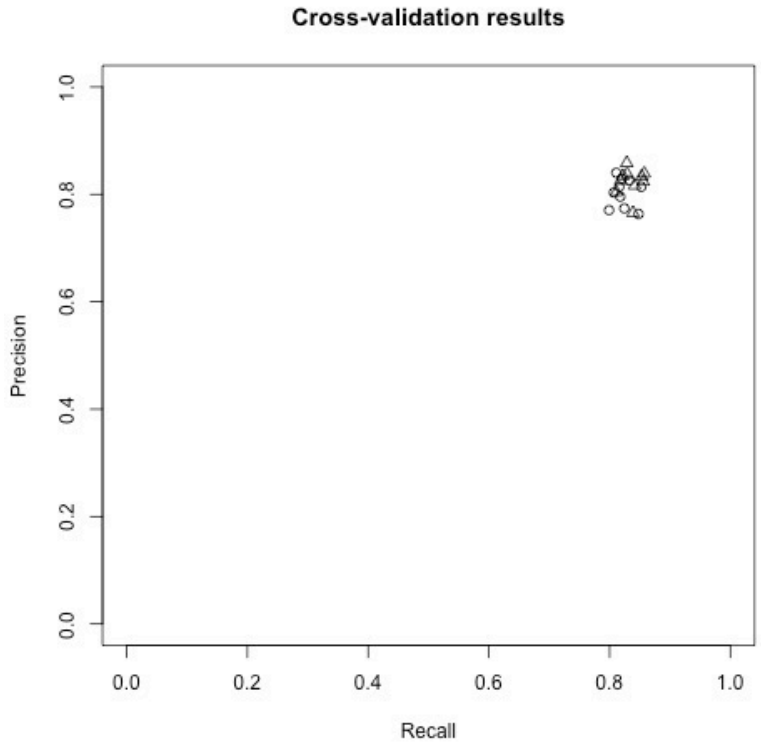


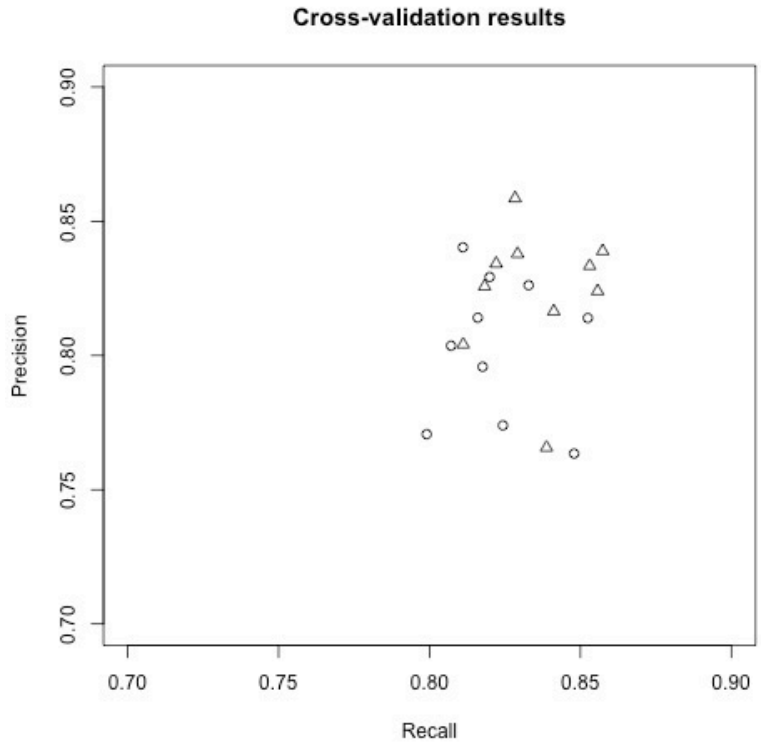
For example: left to right, top to bottom, axes to quadrants

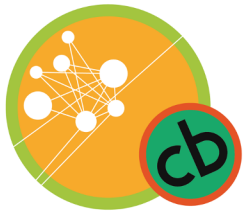
DEVELOP A SYSTEM FOR LOOKING AT GRAPHS



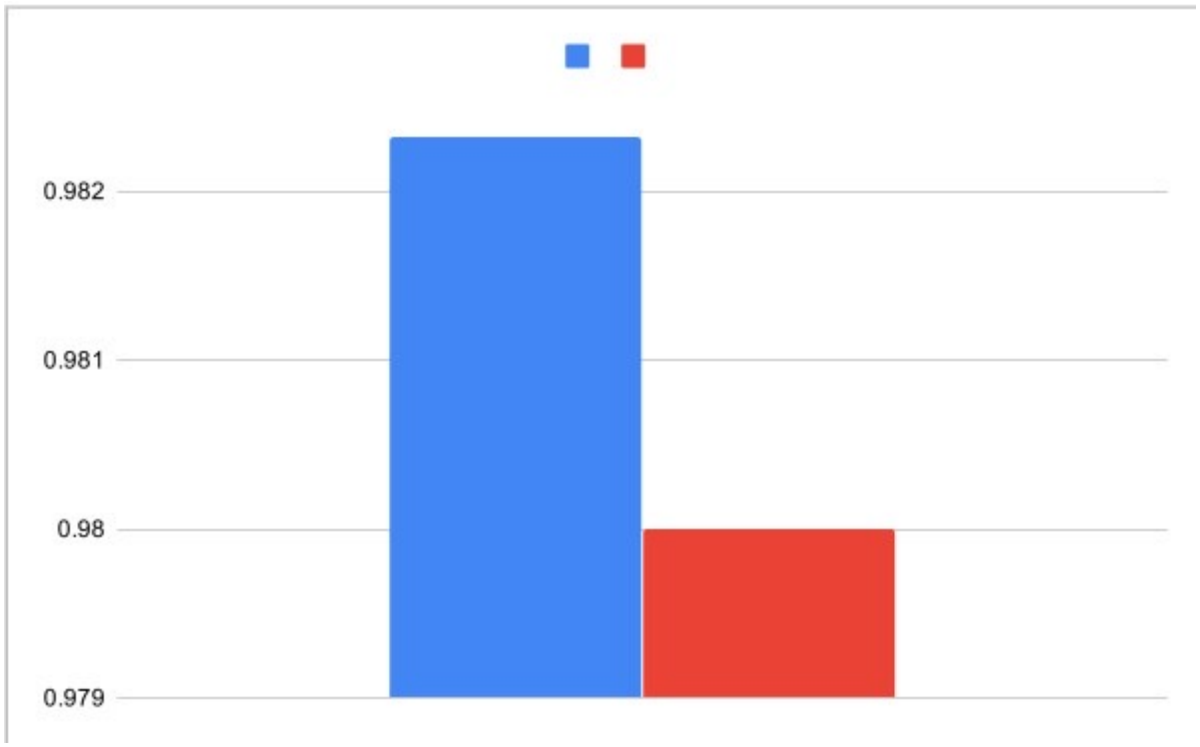
Left to right, top to bottom, axes to quadrants

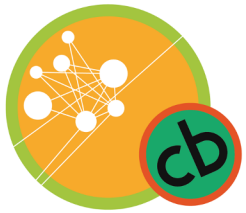






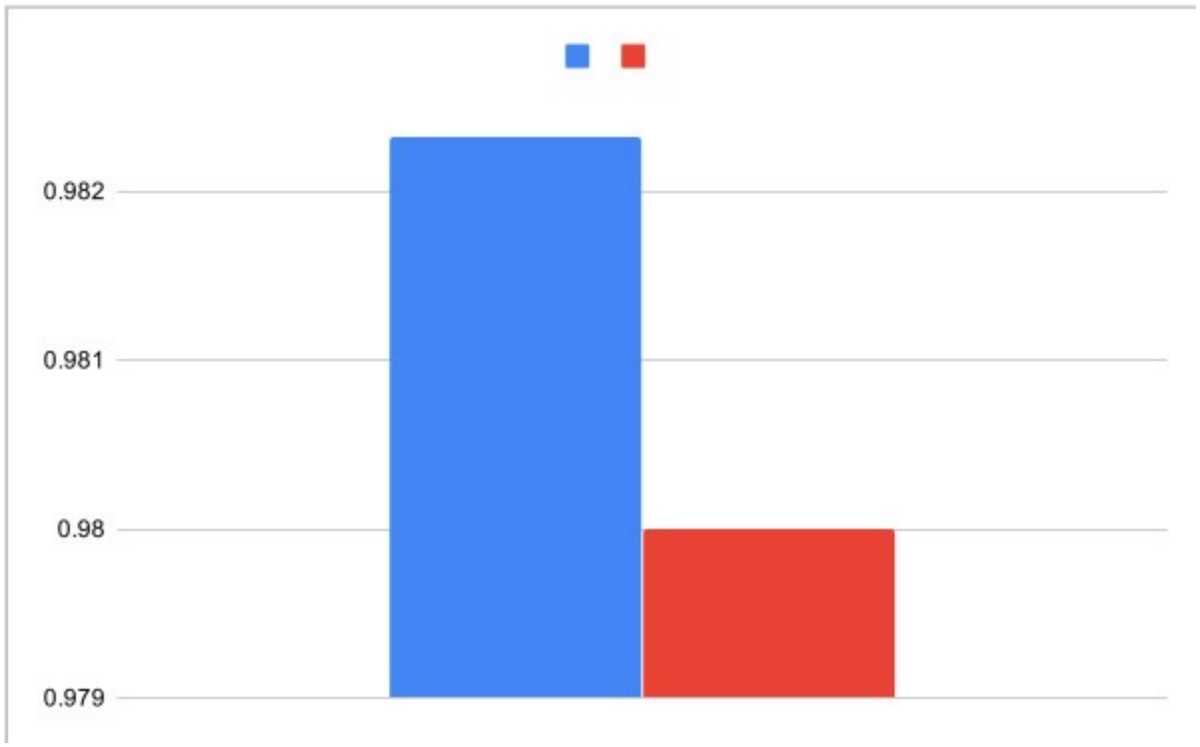
Exercise:
Is this figure good, bad, or mediocre? Why?



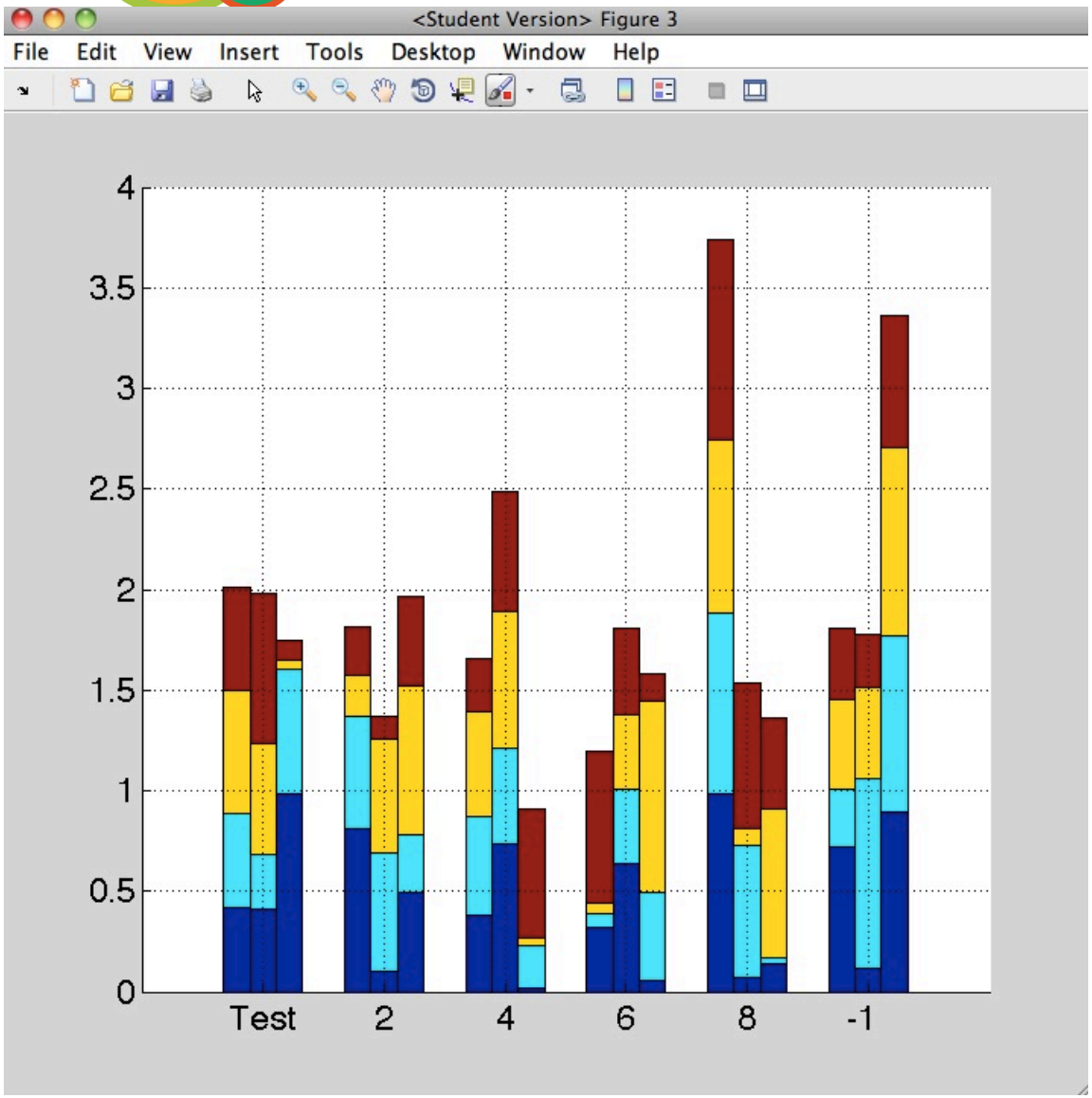


Exercise:

What kind of claim does this figure *not* support well?



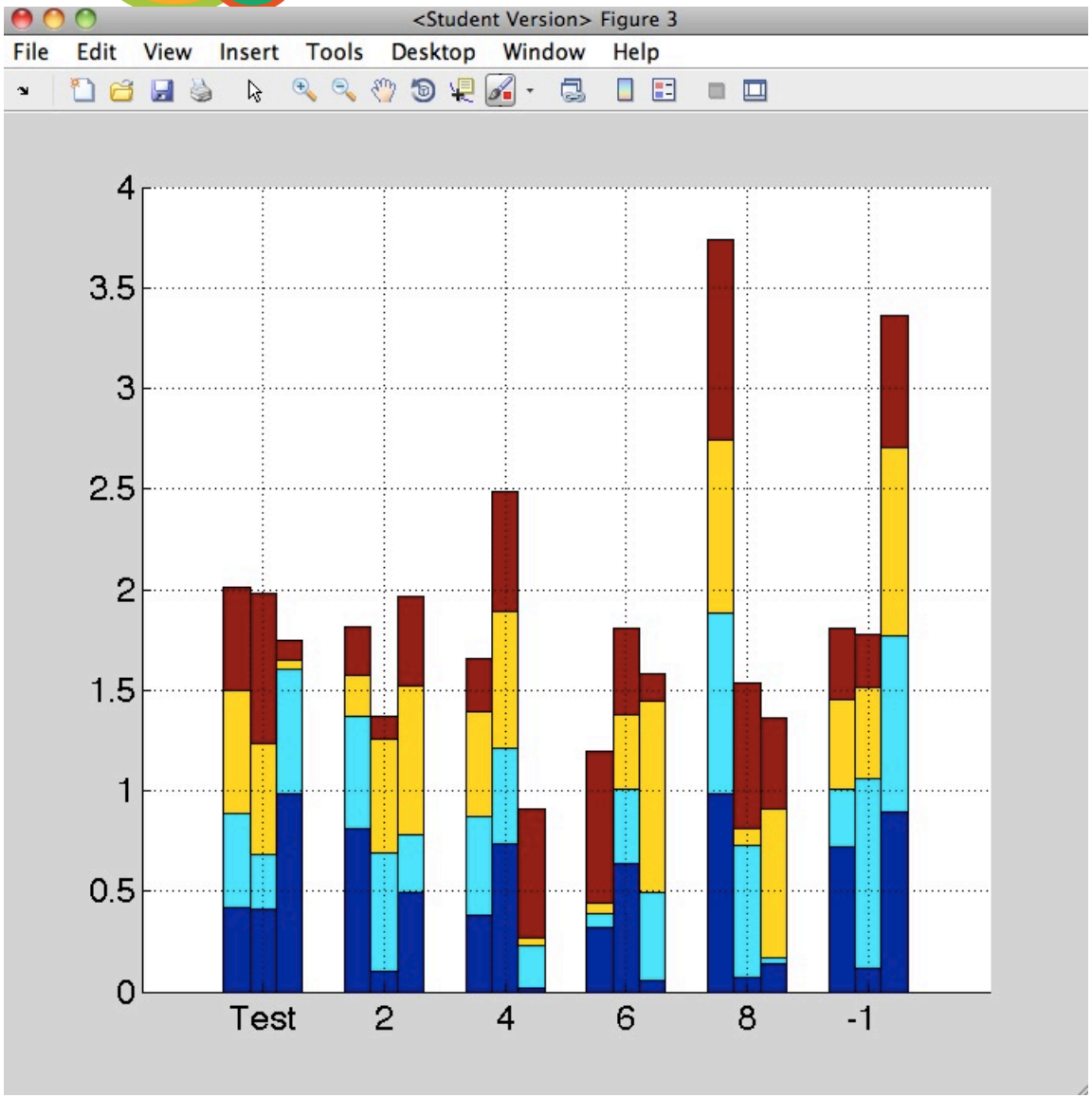
Clarity counts! Exercise…



How many independent variables are being graphed?

https://www.mathworks.com/matlabcentral/fileexchange/32884-plot-groups-of-stacked-bars

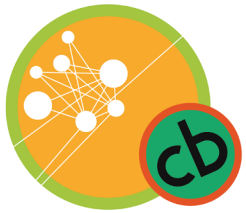
Exercise…



How many independent variables are being graphed?

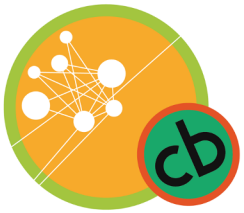
1. Group (Test/2/4/etc.)
2. 3 bars in each group
3. Stack within each bar

https://www.mathworks.com/matlabcentral/fileexchange/32884-plot-groups-of-stacked-bars



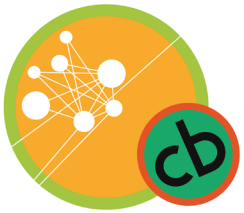
Antske Fokkens et al.

AS A COMMUNITY, WE NEED TO KNOW WHERE OUR APPROACHES FAIL, AS MUCH—IF NOT MORE—AS WHERE THEY SUCCEED



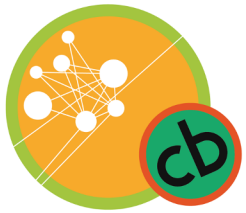
…and this work on what results mean

- Hand, David J. "Classifier technology and the illusion of progress." *Statistical Science* (2006): 1-14.
- 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).



Checklist for the Methods section

- ❑ No confounds
- ❑ No confirmation bias
- ❑ Complete documentation
- ❑ Metric is appropriate for the task
- ❑ Baselines are non-trivial
- ❑ Ablation studies include the single-feature condition (Holte algorithm)
- ❑ Code and data are available
- ❑ Manipulation experiments or observation studies



(xvi) Ties—how handled: Fokkens et al. (2013) describe trying to replicate an important study on semantic similarity measures and discovering that in order to reproduce the original findings, they had to handle ties in the probability scores for part-of-speech tags. This required contacting the original author. Kwong and Tsou (2005) found handling to ties to be a major contributor to performance in Chinese semantic role labelling. Avramidis (2012) presents an extensive treatment of the problem of handling ties in machine translation.

(xvii) Rounding—to how many decimal places: Fokkens et al. (2013) conducted a study of named entity reco
heavily-cited results required
inal researchers had rounded

Offspring from Reproduction Problems:
What Replication Failure Teaches Us

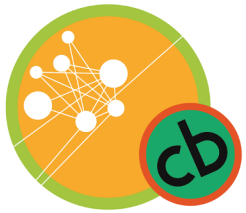
Antske Fokkens and **Marieke van Erp**
The Network Institute
VU University Amsterdam
Amsterdam, The Netherlands
{a.s.fokkens,m.g.j.van.erp}@vu.nl

Marten Postma
Utrecht University
Utrecht, The Netherlands
martenp@gmail.com

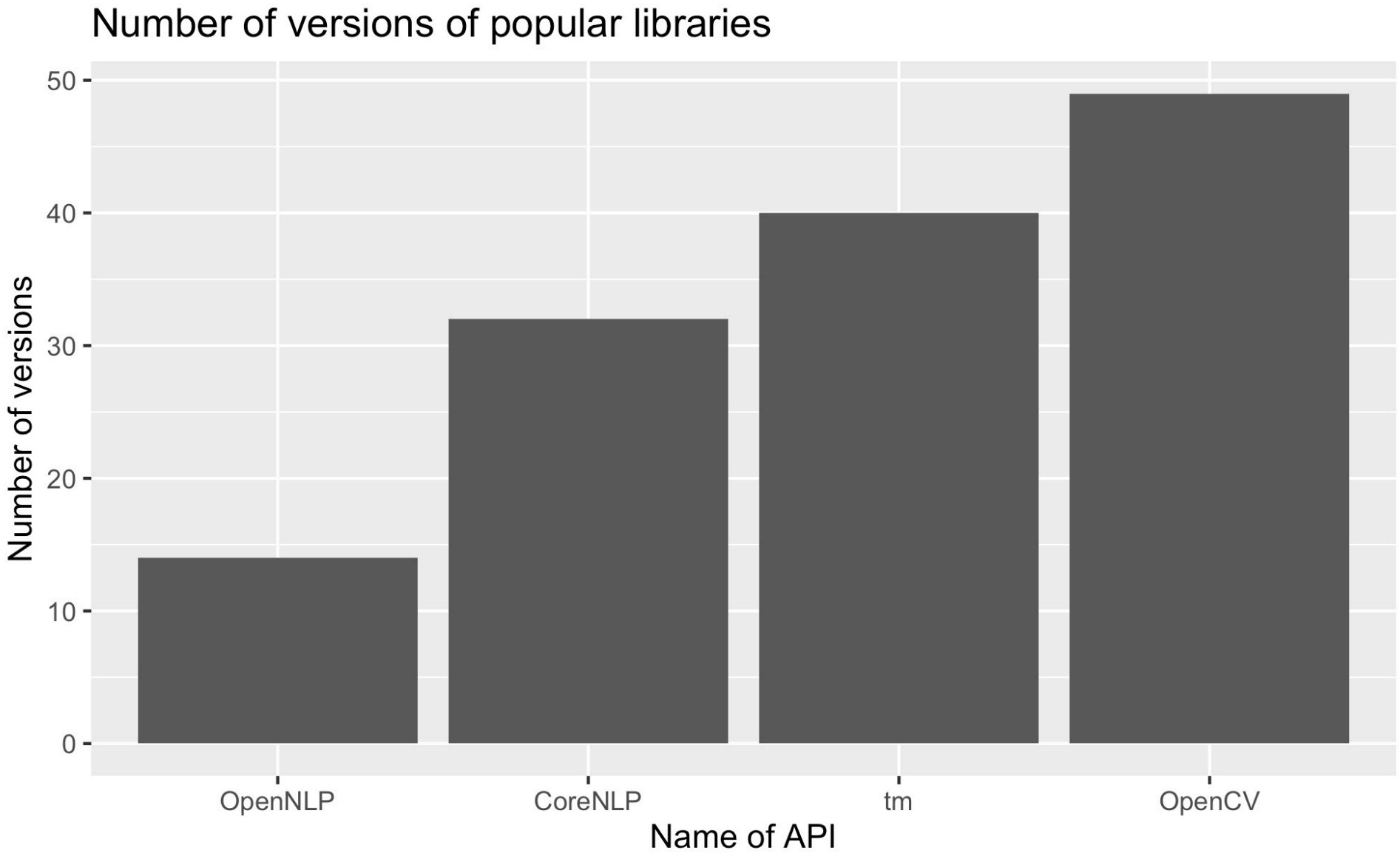
Ted Pedersen
Dept. of Computer Science
University of Minnesota
Duluth, MN 55812 USA
tpederse@d.umn.edu

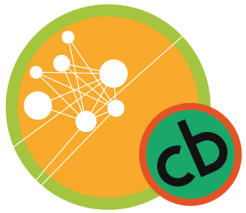
Piek Vossen
The Network Institute
VU University Amsterdam
Amsterdam, The Netherlands
piek.vossen@vu.nl

Nuno Freire
The European Library
The Hague, The Netherlands
nfreire@gmail.com



Complete documentation: Names of tools without version numbers do not tell you anything





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*.

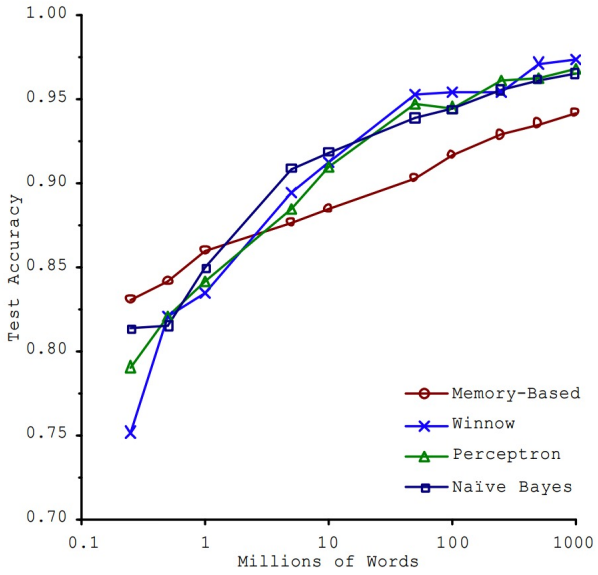
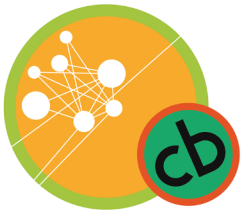


Figure 1. Learning Curves for Confusion Set Disambiguation

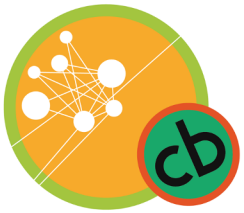
Banko, Michele, and Eric Brill. "Scaling to very very large corpora for natural language disambiguation." ACL 2001.



Dependent variable: document class
Confounding variable: document source

- 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**



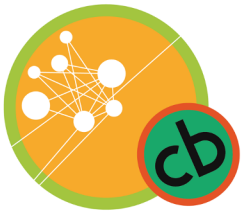
Exercise: Do the methodology and outcomes support the conclusion?

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

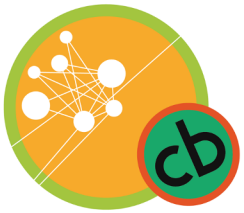
- Compared:
  1. An SVM classifier with bag

Conclusion

- Neural networks are the best approach to

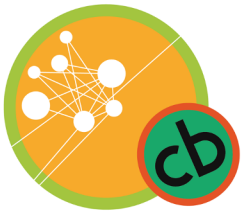
No. The paper compared two feature sets. The fact that one feature set was partially produced by a neural network does not tell us anything about neural networks.

classifier with bag of words as the only feature



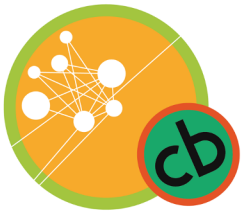
Which paper gets a higher score for methodology?

Evaluated by cross-validation	Evaluated against different dataset
Compared own system against published systems	Optimized parameter settings for own system and for baseline systems
Tried three tokenizers and reported performance using one	Tried three tokenizers and reported performance for each one
Gives version numbers for tokenizer and stemmer used	Tells you that data was tokenized and stemmed
Baseline: random class assignment	Baseline: single feature



Facteur confondant

CONFOUND: MORE THAN ONE EXPLANATION FOR THE OUTCOME



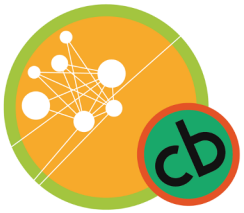
Hypothesis of both papers: Neural networks perform better than naïve Bayesian classifiers. Which paper tests it best?

Paper 1

- Held constant:
  - NLTK sentence segmenter
  - NLTK tokenizer with GENIA model
- Varied between runs:
  - Caret neuralnet() classifier
  - Caret nb() classifier

Paper 2

		F-measure
Zigglebottom (2008)	NB	
Our system	NN	



Hypothesis: Neural networks perform better than naïve Bayesian classifiers.

Paper 1

- Held constant:
  - NLTK sentence segmenter
  - NLTK tokenizer with GENIA model
- Varied between runs:
  - Caret neuralnet() classifier
  - Caret nb() classifier

Paper 2

		F-measure
Zigglebottom (2008)	NB	
Our system	NN	

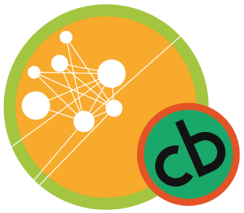
Paper 1 tests the hypothesis best.

Paper 1: Only difference is the classifier.

Paper 2: *One* difference is the classifier.

Biais de confirmation *ou* biais de confirmation d'hypothèse

CONFIRMATION BIAS: FINDING WHAT YOU EXPECTED TO FIND



Claim: Medusa transducers outperform several baselines

Introduction

- Medusa transducers are really great (Frankenfurter 2012). They have been getting great scores (Rocky 1980). Really, the best scores (Magenta 2016). It would be good to try a Medusa transducer for the task of -----.

Results

	A metric
Random class assignment	A number
Medusa transducer + CNN	A number
Medusa transducer + CNN + BFMTV	A number
Medusa transducer + CNN + BFMTV + SCNF	A number
Medusa transducer + CNN + BFMTV + SCNF + DGSE	A number
Medusa transducer + CNN + BFMTV + SCNF + DGSE + RER-B	A number
Our system (Medusa transducer + CNN + BFMTV + SCNF + DGSE + RER-B + BD + CDD + CDI + AOP)	**A number**

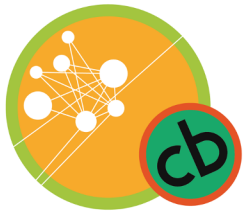
Claim: Medusa transducers outperform several baselines

Introduction

- Medusa transducers are really great (Frankenfurter 2012). They have been getting great scores (Rocky). Really, the best scores (Magenta 2016). It would be good to try a Medusa transducer for the task of -----.

Consider confirmation bias when the only apparent motivation for the approach is previous high performance.

Ra	
M	
M +	
M +	
M +	
M + +	
O tr BF	
RER-B + BD + CDD + CDI + AOP)	

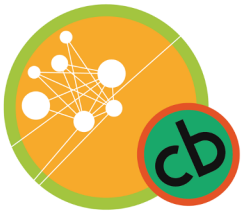


Claim: Medusa transducers outperform several baselines

Consider confirmation bias when there is no obvious pattern or hypothesis behind the differences between the baselines and ”our” system.

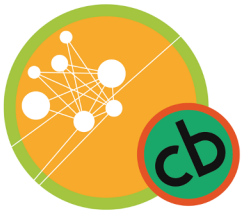
Results

	A metric
Random class assignment	A number
Medusa transducer + CNN	A number
Medusa transducer + CNN + BFMTV	A number
Medusa transducer + CNN + BFMTV + SCNF	A number
Medusa transducer + CNN + BFMTV + SCNF + DGSE	A number
Medusa transducer + CNN + BFMTV + SCNF + DGSE + RER-B	A number
Our system (Medusa transducer + CNN + BFMTV + SCNF + DGSE + RER-B + BD + CDD + CDI + AOP)	**A number**



Exercise: Which gives the least suspicion of confirmation bias?

1. Bidirectional long short-term memory has been shown to be effective for parsing tasks, but has not been applied to product reviews about tomatoes.
2. An alternative explanation of the results is that the systems are differentially sensitive to large number of rare event (LNRE) distributions.
3. Our approach always out-performs competing approaches.



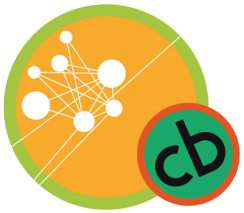
Exercise: Do the methodology and outcomes support the conclusion?

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

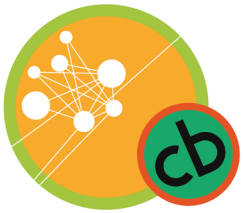
- Compared:
  1. An SVM classifier with bag

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.

classifier with bag of words as the only feature



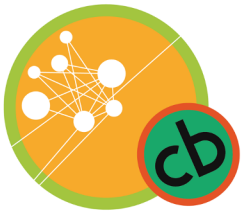
Exercise: Do the methodology and outcome support the conclusion?

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?

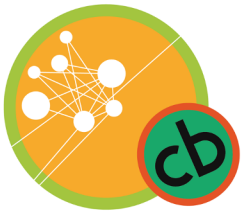
Methodology/Outcome

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

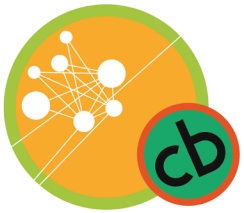
No. The paper provides about as much detail as an advertisement. Under-described methods and unavailable code mean that the paper does not actually describe the methods and the outcome is not valid.



Why **doesn't** this experiment support the following analysis: *BOW had no effect on performance…?*

Ablation study results

Feature set	F_1
Word embeddings	0.90
Word embeddings + Bag of words	0.90



Why is ablation study 2 better than ablation study 1?

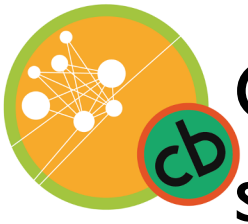
Ablation study 1

Feature set	F_1
Word embeddings	0.90
Word embeddings + Bag of words	0.90

Ablation study 2

Feature set	F_1
Word embeddings	0.90
Bag of words	0.90
Word embeddings + Bag of words	0.90

"Algorithme de Holte"
Holte (1993) Very simple classification rules perform very well on most commonly-used datasets. *Machine Learning* 11:63-91



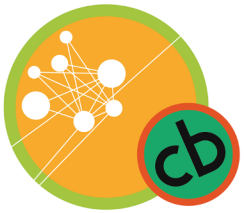
Only one of these supports a claim of superior performance by "our" system?

System	F-measure, 10-fold x-validation
Able (2020)	0.90 (sd +/- 0.02)
Ours	0.93 (sd +/- 0.07)

System	F-measure, 10-fold x-validation
Able (2020)	0.90
Ours	0.93 (sd +/- 0.07)

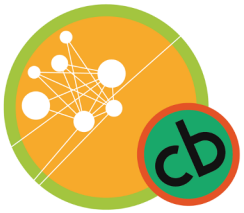
System	F-measure, 10-fold x-validation
Able (2020)	0.90
Ours	0.93

System	F-measure, 10-fold x-validation
Able (2020)	0.90 (sd +/- 0.02)
Ours	0.93



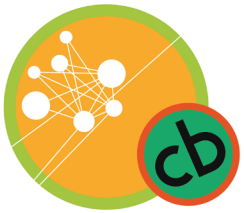
For more on how to think about Methods, see…

- 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.
- Cohen, Paul (1995) Empirical Methods for Artificial Intelligence. MIT Press.
- Resnik, P. and Lin, J., 2010. Evaluation of NLP systems. Ch. 11 of The handbook of computational linguistics and natural language processing, 57.
- 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).
- Hirst, Graeme, Yaroslav Riabinin, and Jory Graham. "Party status as a confound in the automatic classification of political speech by ideology." 2010.



…and other minor points

SIGNIFICANCE AND INNOVATION



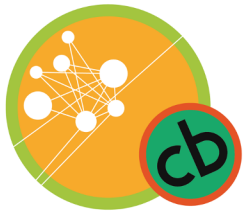
Which paper can you expect to have the highest impact? Why?

Paper 1

- Relevant to a problem outside of NLP
- Data and code are available on GitHub
- Experiments show under what conditions the system does **not** perform well

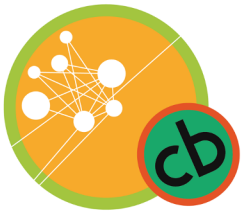
Paper 2

- Reports state-of-the-art performance on a task



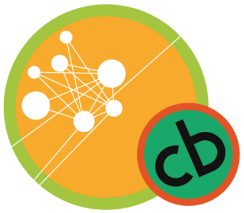
Significance as impact

Higher impact	Lower impact
Interest does not depend on performance	Interest is limited to performance
Multiple metrics are evaluated	Single metric only is evaluated
Data is publicly available	Data is not publicly available
Code is publicly available	Code is not publicly available
Conclusions based on a single dataset	Conclusions based on multiple datasets
Conclusions based on a single task	Conclusions based on multiple tasks
Relevance beyond language processing	Relevance limited to language processing



Source: Kenneth Church

THE BETTER THE NUMBERS ARE, THE MORE IMPORTANT IT IS TO REJECT THE PAPER. WE CAN'T AFFORD PAPERS THAT REPORT RESULTS WITHOUT INSIGHTS.



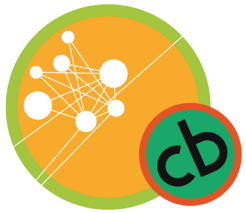
Which paper can you expect to still be read three years from now?

Paper 1

- Relevant to a problem outside of NLP
- Data and code are available on GitHub
- Experiments show under what conditions the system does not perform well

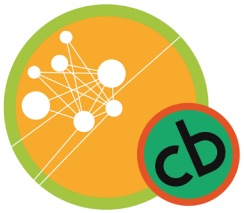
Paper 2

- Reports state-of-the-art performance on a task
- Outperforms a system from five years ago by quite a bit
- Uses a currently very popular approach



Antske Fokkens et al.

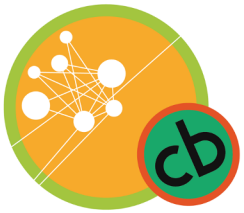
AS A COMMUNITY, WE NEED TO KNOW WHERE OUR APPROACHES FAIL, AS MUCH—IF NOT MORE—AS WHERE THEY SUCCEED



Which of these papers makes the best innovation claim?

Paper A: No previous work has tried the combination of bidirectional long short-term memory in a recurrently convolutional and convolutedly recurrent neural network with attention, a part-of-speech tagger, structural parsing, dependency parsing, semantic role labelling, word sense disambiguation, WordNet, VerbNet, FrameNet, PropBank, and no manually engineered features.

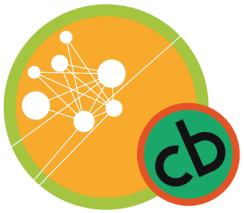
Paper B: The most novel aspect of this work is its application to a new domain. This is important because as others have shown, the domain has linguistic features that typically pose challenges for domain portability (Delta 2018, Echo 2019, Foxtrot 2019). These include aspects of its morphology (Alpha 2012), syntax (Bravo 2013), and phonology (Charlie 2014).



Which of these papers makes the best innovation claim?

Paper A: No previous work has tried the combination of bidirectional long short-term memory in a recurrently convolutional and convolutedly recurrent neural network with attention, a part-of-speech tagger, structural parsing, dependency parsing, semantic role labelling, word sense disambiguation, WordNet, VerbNet, FrameNet, PropBank, and no manually engineered features.

Paper B: The most novel aspect of this work is its application to a new domain. **This is important because** as others have shown, the domain has linguistic features that typically pose challenges for domain portability (Delta 2018, Echo 2019, Foxtrot 2019). These include aspects of its morphology (Alpha 2012), syntax (Bravo 2013), and phonology (Charlie 2014).



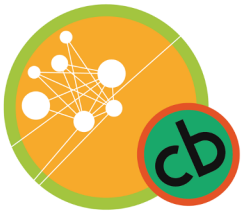
Which paper gets a higher score for its discussion of related work?

Paper 1

- Able (2015) used naive Bayes with binary feature selection by chi-square. Baker (2016) applied a support vector machine with a kernel optimized to take advantage of Universal Dependencies, while Gamma (2017) used a recurrently convolutional neural network with bi-directional long short-term memory and attention trained on cumulative random samples rather than the usual straight cumulative sampling.

Paper 2

- Able (2015) used naive Bayes. Baker (2016) applied a support vector machine, while Charlie (2017) used a recurrently convolutional neural network with bi-directional long short-term memory and attention. These three papers share a reliance on supervised approaches. What they all omit is any attempt to take advantage of anything that we know about how *humans* learn to do the task in question.



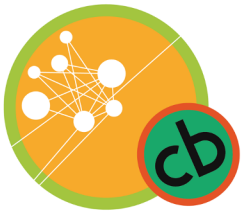
Which paper gets a higher score for its discussion of related work?

Paper 1

- Able (2015) used naive Bayes with binary feature selection by chi-square. Baker (2016) applied a support vector machine with a kernel optimized to take advantage of Universal Dependencies, while Gamma (2017) used a recurrently convolutional neural network with bi-directional long short-term memory and attention trained on cumulative random samples rather than the usual straight cumulative sampling.

Paper 2

- Able (2015) used naive Bayes. Baker (2016) applied a support vector machine, while Charlie (2017) used a recurrently convolutional neural network with bi-directional long short-term memory and attention. **These three papers share** a reliance on supervised approaches. **What they all omit** is any attempt to take advantage of anything that we know about how *humans* learn to do the task in question.



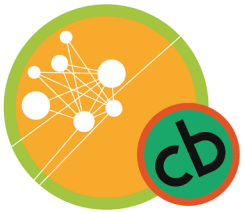
Which paper will probably get a higher score for innovation?

Paper 1

- Able (2015) used naive Bayes with binary feature selection by chi-square. Baker (2016) applied a support vector machine with a kernel optimized to take advantage of Universal Dependencies, while Gamma (2017) used a recurrently convolutional neural network with bi-directional long short-term memory and attention trained on cumulative random samples rather than the usual straight cumulative sampling.

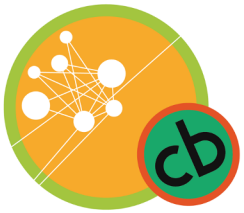
Paper 2

- Able (2015) used naive Bayes. Baker (2016) applied a support vector machine, while Charlie (2017) used a recurrently convolutional neural network with bi-directional long short-term memory and attention. **These three papers share** a reliance on supervised approaches. **What they all omit** is any attempt to take advantage of anything that we know about how *humans* learn to do the task in question.



Source: Christopher Manning

THINK ABOUT PROBLEMS, ARCHITECTURES, COGNITIVE SCIENCE, AND THE DETAILS OF HUMAN LANGUAGE... RATHER THAN JUST CHASING STATE-OF-THE-ART NUMBERS.



How to differentiate between "significance" and "innovation"

Significance: What you did

- Requires less training data than competing approaches
- Tells us something about how language works
- Stable over wide range of parameter settings and datasets
- Simpler than competing approaches
- Solves an important problem
- More interpretable than competing approaches

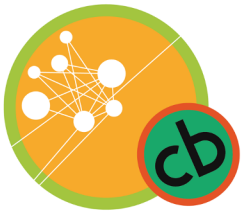
--Anna Rogers, https://hackingsemantics.xyz/2020/reviewing-models/#solution-guidelines-on-what-constitutes-an-acceptable-contribution

Innovation: How you did it

- New domain
- New task
- New metric
- Technique/model from another field
- By addressing confounds in previous work
- More stringent evaluation

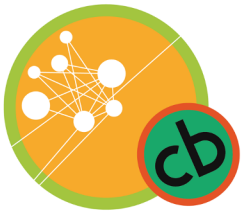
For more on how to think about *Conclusions* sections, see this work on spin…

- 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…

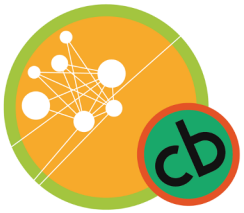
- Cohen, K. Bretonnel (2021) "Writing about data science research. " Cambridge University Press.
- Hirst, Graeme, Yaroslav Riabinin, and Jory Graham. "Party status as a confound in the automatic classification of political speech by ideology." 2010.
- Kumar, Sachin and Wintner, Shuly and Smith, Noah 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).

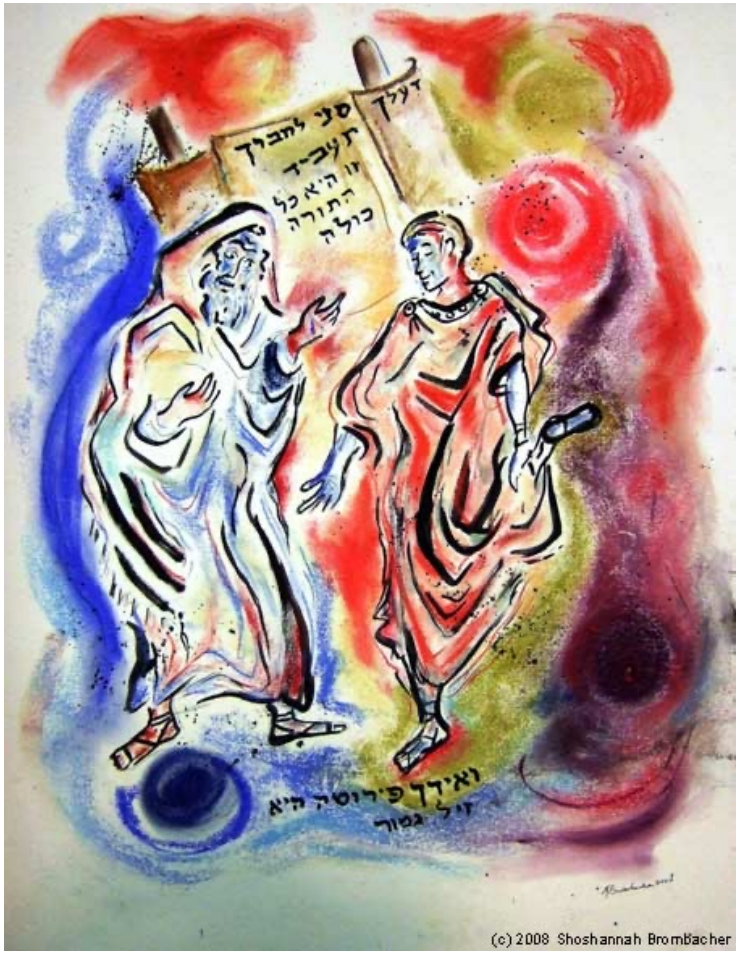
Karën Fort, Margot Mieskes, Aurélie Névéol, and Anna Rogers



Pierre Zweigenbaum

FAIS LA RELECTURE QUE TU VOUDRAIS QUE L'ON TE FASSE.

"…all the rest is commentary. Now go and study." – Hillel l'Ancien



Artist: Shoshannah Brombacher