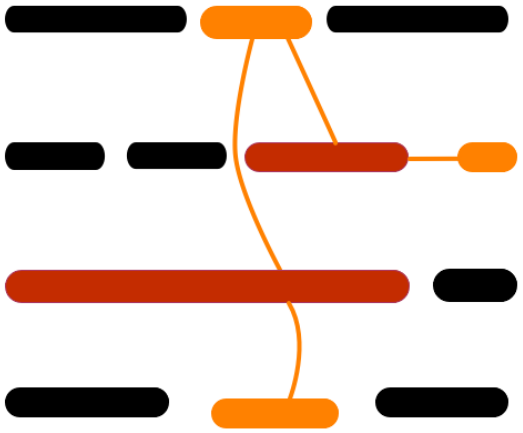
Inter-sentence Relation Extraction for Associating Biological Context with Events in Biomedical Texts

Paul Hein

Motivation



"Context such as cell type is the big thing missing in reading."

– Mike Bolt, Chicago biologist CURE, March 21

Background: biological relations

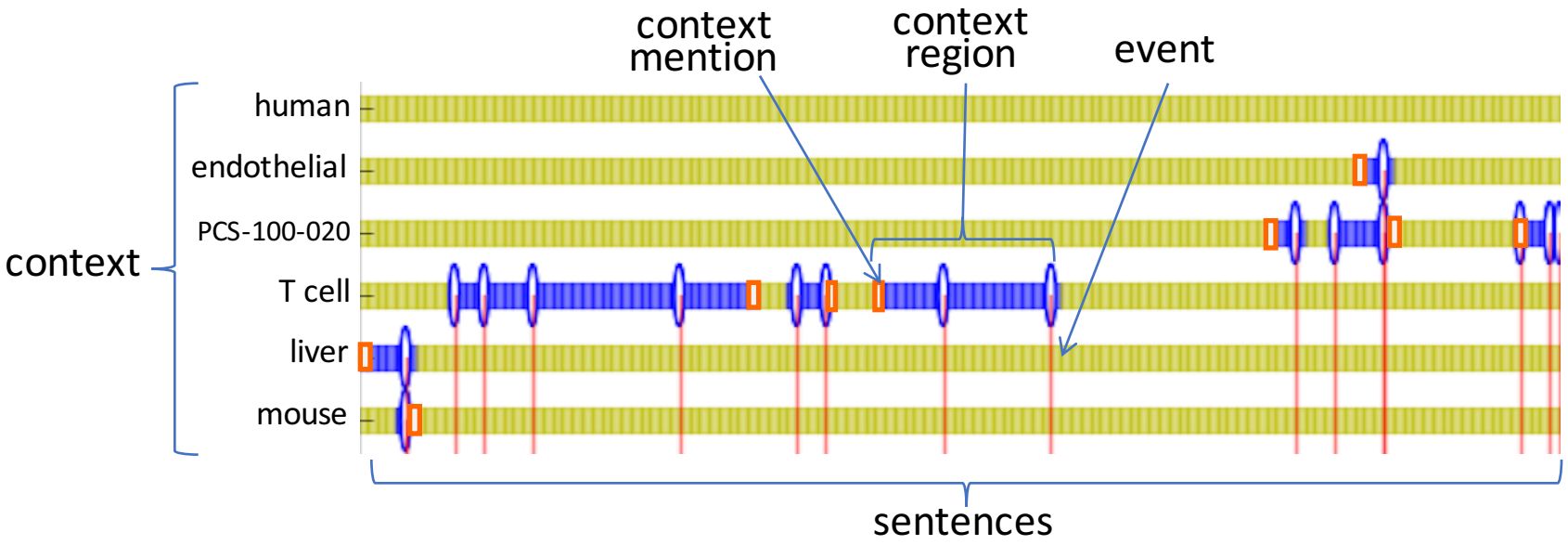
Biological context examples

- Species: Human, yeast, mouse
- Organ: Liver, lung
- Tissue Type: lymphoid, embryo
- Cell type: t-cell, endothelial
- Cell line: human artery
- Sub-cellular: endosome, nucleus

Biological event examples

- Release of RAC1
- Rac1 activation
- translocation of Rac1/RhoGDI

Background: visualizing the problem



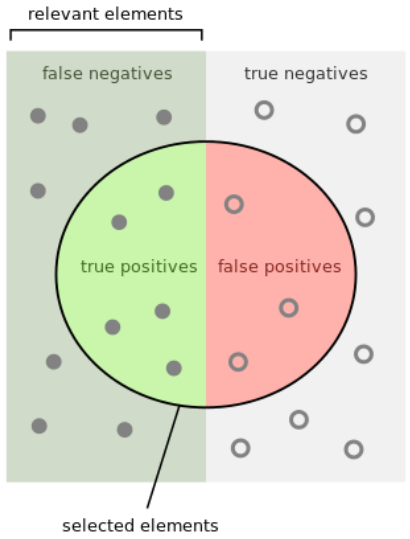
- Multiple context can hold over a particular section of text
- Events commonly occur in sentences that do not include the correct context mention

Background: linguistic features

TABLE I
CLASSIFICATION FEATURES

Name	Details
General features	
Sentence distance	No. of sentences separating the event and context mentions
Dependency distance	No. of edges separating the mentions within dependency graph
Context type frequency	No. of context mentions of the same type
Is context closest	Indicates whether the context mention is the *closest* one to the event
Φ features	
Is sentence *first person* Is sentence *past tense* Is sentence *present tense*	An instance for each: event and context mentions
Syntactic features	
Event spanning dependency bigrams	Sequence of dependency bigrams spanning from event mention
Negated event mention	Indicates whether a *neg* dependency is within 2 degrees in dep. graph
Context spanning dependency bigrams	Sequence of dependency bigrams spanning from context mention
Negated context mention	Indicates whether a *neg* dependency is within 2 degrees in dep. graph

Experimental design: measurements

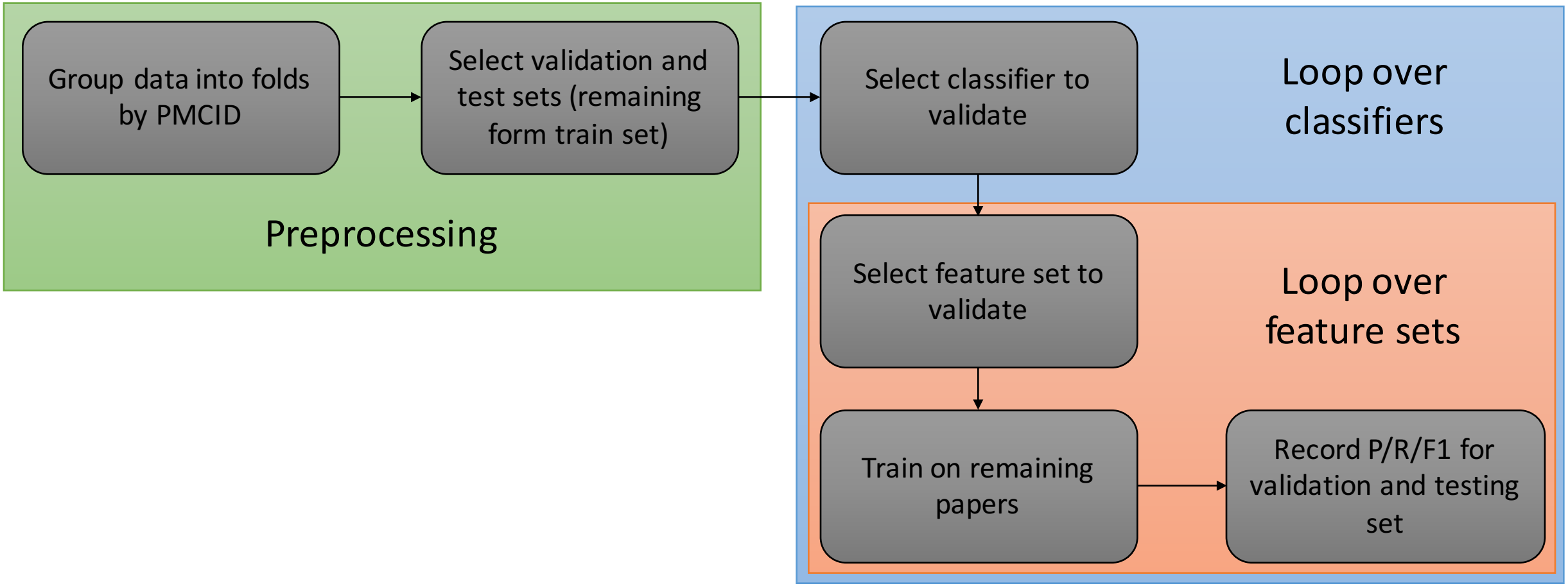


$$F_1 = \frac{2}{\frac{1}{\mathrm{recall}} + \frac{1}{\mathrm{precision}}} = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}.$$

Experimental design: data formatting

- Events are key
- Class discrepancy between positive and negative labels
- Correcting for annotation style

Experimental design: CV procedure



Classification: our baseline classifier

Algorithm 1 Deterministic context baseline

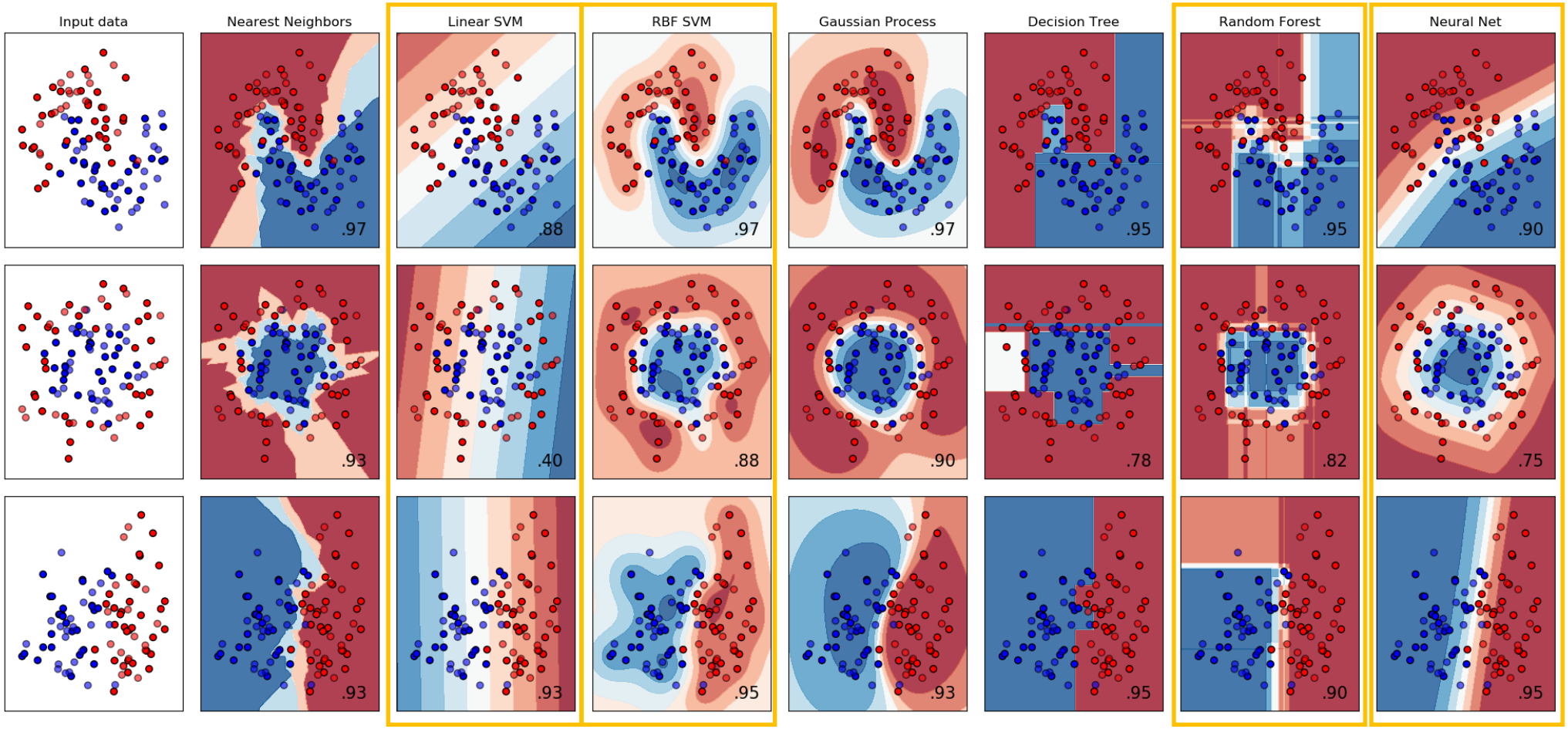
function IsContext(evt, ctx, k)
  $evtSen = getSentenceIx(evt)$
  $ctxSen = getSentenceIx(ctx)$
  $interval = [evtSen - k, evtSen + k]$
  if $ctxSen \in interval$ **then**
    return $True$
  else
    return $False$
  end if
end function

Classification: classifier list

- *Logistic Regression* with ℓ^2 regularization penalty
- *Support Vector Machines* with the following kernels:
  - Linear
  - Polynomial
  - Gaussian
- *Random Forest*
- *Feed-Forward Neural Network*

Classification: Advanced classifier comparison

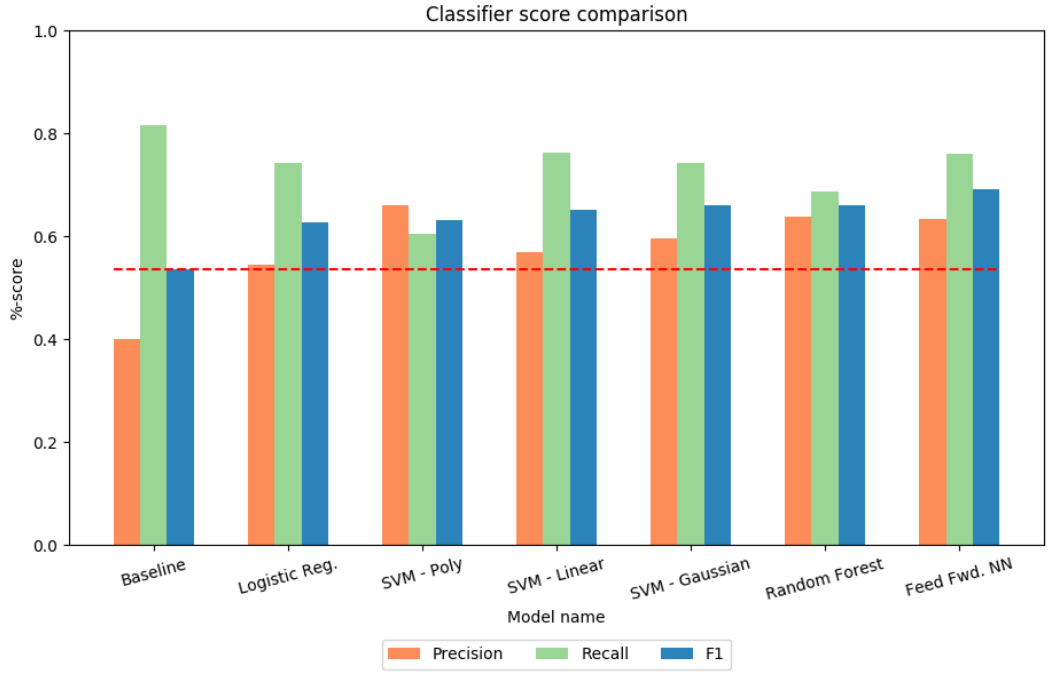
Images and data taken from Scikit-learn



Results: classifier scores

TABLE III
BEST SCORES PER CLASSIFYING ALGORITHM

Algorithm	Precision	Recall	F1
Baseline	0.399	**0.816**	0.536
Logistic Reg.	0.543	0.743	0.628
SVM - Poly	**0.660**	0.605	0.631
SVM - Linear	0.568	0.763	0.651
SVM - Gaussian	0.595	0.742	0.660
Random Forest	0.637	0.686	0.661
Feed Fwd. NN	0.633	0.761	**0.691**

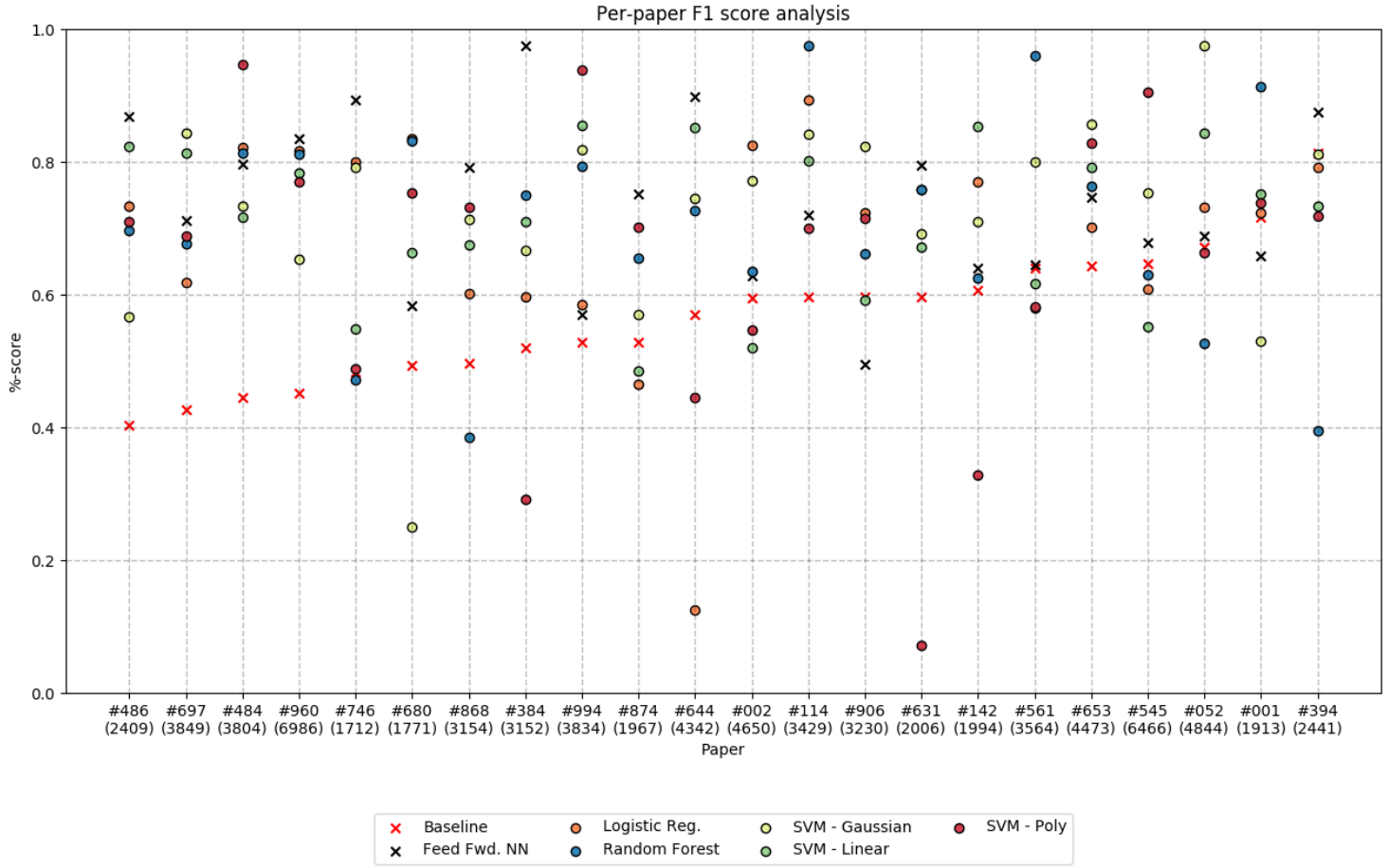


Results: optimal feature sets

TABLE II
OPTIMAL FEATURE SUBSET PER ALGORITHM

Algorithm	Feature Subset
SVM - Poly	Is ctx. closest, Is ctx. closest, Dependency dist., Ctx. sent. past tense, Context freq., Sentence dist., Evt. sent. present tense, Ctx. sent. present tense, Negated evt. mention, Ctx. spanning dep. bigrams
SVM - Gaussian	Evt. sent. present tense, Context freq., Ctx. sent. present tense, Sentence dist., Ctx. sent. present tense, Is ctx. closest, Dependency dist., Evt. sent. present tense, Is ctx. closest, Ctx. spanning dep. bigrams
SVM - Linear	Ctx. sent. present tense, Sentence dist., Negated evt. mention, Evt. sent. present tense, Context freq., Ctx. sent. present tense, Ctx. sent. past tense, Is ctx. closest, Evt. sent. past tense, Is ctx. closestDependency dist., Ctx. spanning dep. bigrams
Logistic Reg.	Ctx. sent. present tense, Evt. sent. present tense, Context freq., Ctx. sent. present tense, Ctx. sent. past tense, Is ctx. closest, Evt. sent. past tense, Is ctx. closestDependency dist., Evt. sent. present tense, Ctx. spanning dep. bigrams
Random Forest	Ctx. sent. present tense, Negated ctx. mention, Sentence dist., Negated evt. mention, Evt. sent. present tense, Context freq., Ctx. sent. present tense, Ctx. sent. past tense, Is ctx. closest, Is ctx. closestDependency dist., Ctx. spanning dep. bigrams
Feed Fwd. NN	Ctx. sent. present tense, Negated evt. mention, Evt. sent. present tense, Context freq., Ctx. sent. present tense, Ctx. sent. past tense, Is ctx. closest, Evt. sent. past tense, Is ctx. closestDependency dist., Ctx. spanning dep. bigrams

Results: per-paper analysis



Future work

- Neural model creation to boost F1 score
- Statistical significance testing of comparative results between classifiers
- Polarity mismatch testing