



Sentiment Analysis (Opinion Mining) with Machine Learning in GATE



Outline

- What is sentiment analysis?
- Example: consumer reviews
- Using GATE ML for this problem
- Batch Learning configuration
- Building the application and training
- Application and evaluation
- Built-in cross-validation
- Suggestions for further experiments



Sentiment analysis

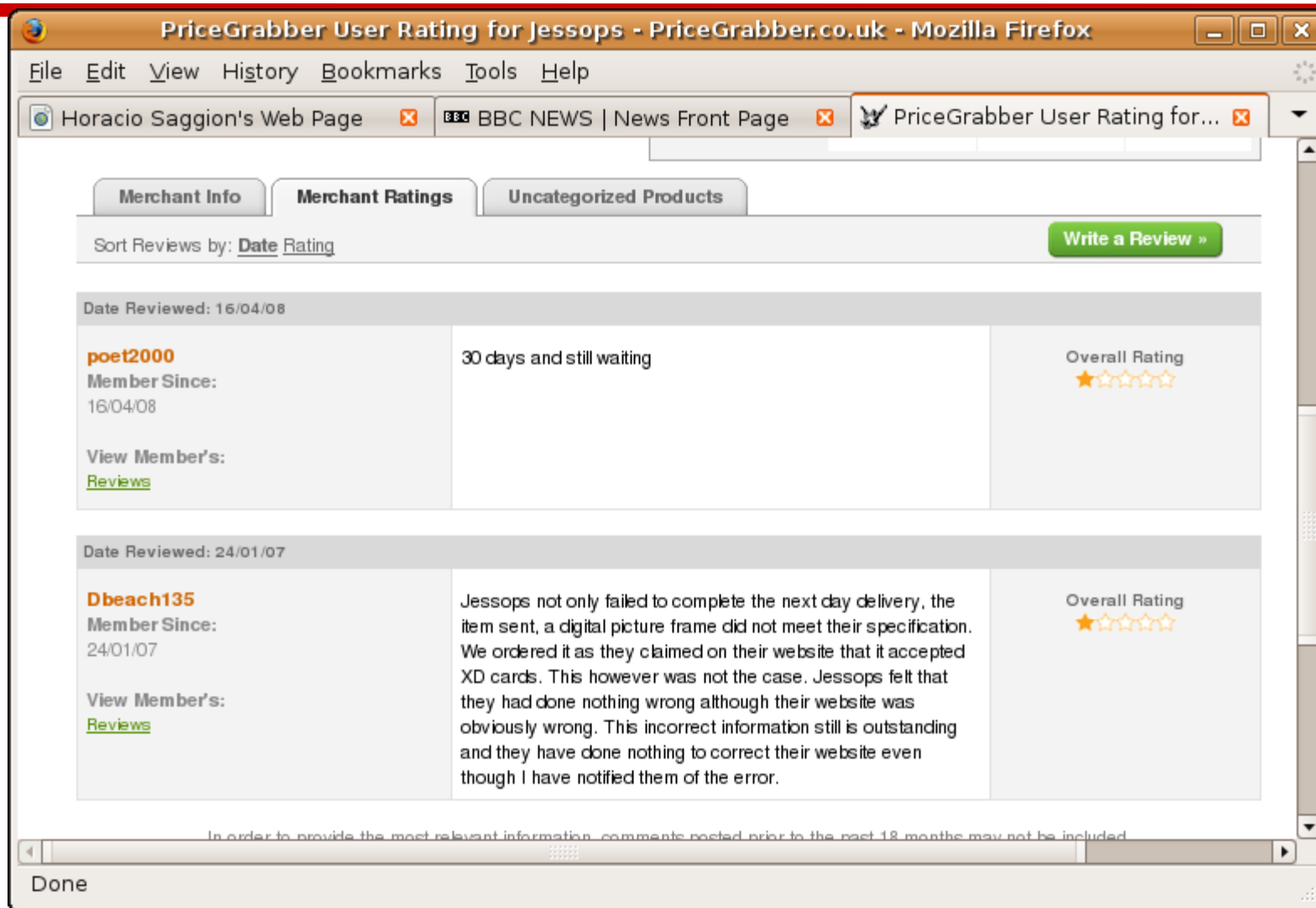
- Applications: business intelligence, opinion mining, reputation monitoring, etc.
- Several problems:
 - identifying opinionated segments of text
 - identifying the opinion-holder
 - classifying the opinion (positive, negative, degree)



Sentiment analysis

- We will work on classifying the opinionated text using ML.
- We want to train a classifier to categorize free text according to the training data. Good examples are consumers' reviews of films, products, and suppliers.

Example: consumer reviews



PriceGrabber User Rating for Jessops - PriceGrabber.co.uk - Mozilla Firefox

File Edit View History Bookmarks Tools Help

Horacio Saggion's Web Page x BBC BBC NEWS | News Front Page x PriceGrabber User Rating for... x

Merchant Info Merchant Ratings Uncategorized Products

Sort Reviews by: Date Rating Write a Review »

Date Reviewed: 16/04/08

poet2000 Member Since: 16/04/08 View Member's: Reviews	30 days and still waiting	Overall Rating ★☆☆☆☆
---	---------------------------	-------------------------

Date Reviewed: 24/01/07

Dbeach135 Member Since: 24/01/07 View Member's: Reviews	Jessops not only failed to complete the next day delivery, the item sent, a digital picture frame did not meet their specification. We ordered it as they claimed on their website that it accepted XD cards. This however was not the case. Jessops felt that they had done nothing wrong although their website was obviously wrong. This incorrect information still is outstanding and they have done nothing to correct their website even though I have notified them of the error.	Overall Rating ★☆☆☆☆
--	---	-------------------------

In order to provide the most relevant information, comments posted prior to the past 18 months may not be included.

Done

Example: consumer reviews

- We have 40 documents containing 552 company reviews. Each review has a 1- to 5-star rating.
- We converted the HTML pages into GATE documents, preprocessed them to label each review with a comment annotation with a rating feature, and saved them as GATE XML to make the corpus for this exercise.

Example: consumer reviews

PriceGrabber User Rating for Jessops - PriceGrabber.co.uk - Mozilla Firefox

File Edit View History Bookmarks Tools Help

Horacio Saggion's Web Page x BBC BBC NEWS | News Front Page x PriceGrabber User Rating for... x

Merchant Info Merchant Ratings Uncategorized Products

Sort Reviews by: Date Rating Write a Review »

Date Reviewed: 16/04/08

poet2000
Member Since:
16/04/08

View Member's:
[Reviews](#)

30 days and still waiting

Overall Rating
★☆☆☆☆

Date Reviewed: 24/01/07

Dbeach135
Member Since:
24/01/07

View Member's:
[Reviews](#)

Jessops not only failed to complete the next day delivery, the item sent, a digital picture frame did not meet their specification. We ordered it as they claimed on their website that it accepted XD cards. This however was not the case. Jessops felt that they had done nothing wrong although their website was obviously wrong. This incorrect information still is outstanding and they have done nothing to correct their website even though I have notified them of the error.

Overall Rating
★☆☆☆☆

In order to provide the most relevant information, comments posted prior to the past 18 months may not be included.

Done



Using GATE ML

- In ML terms:
 - instance = *comment* annotation
 - class = *rating* feature
 - attributes = NLP features of the underlying text
- We will keep the spans of the comment annotations and use ML to classify them with the *rating* feature



Using GATE ML

Annotation Sets Annotations List Annotations Stack Class Co-reference Editor Instance Text

05/11/07

View Member's:
 Reviews quite a trouble free experience. goods arrived in extremely good time - ordered Monday evening arrived Wednesday at 0800 enough said saved a fortune - product arrived as ordered - this is the future. Overall Rating

Date Rev
 shawnfr
 Member S
 26/10/07

View Mem
 Reviews
 so quer
 same day
 delivery and
 is to reject
 my credit

in email saying it shipped Thursday,
 s definitely Saturday delivery - the
 nn say it is on standard working day
 1stAudioVisual should have upgraded it rather than lying to me! My only option
 is to reject the delivery so TNT return it, so now I wait for the refund. I am contacting
 my credit card company and Trading Standards! Do not deal with them! Overall Rating

Date Reviewed: 22/10/07
 robinharris
 Member Since:
 22/08/07

View Member's:
 Reviews Good Company Overall Rating

Date Reviewed: 22/10/07
 lora18147
 Member Since:

Type	Set	Start	End	Id	Features
comment	Key	1400	1468	21639	{rating=5_Star_Review}
comment	Key	1566	1764	21633	{rating=5_Star_Review}
comment	Key	1868	2067	21635	{rating=5_Star_Review}
comment	Key	2165	2675	21617	{rating=1_Star_Review}
comment	Key	2777	2789	21627	{rating=5_Star_Review}
comment	Key	2889	2988	21621	{rating=4_Star_Review}
comment	Key	3090	3462	21665	{rating=1_Star_Review}
comment	Key	3563	3990	21655	{rating=1_Star_Review}
comment	Key	4089	4196	21629	{rating=5_Star_Review}
comment	Key	4293	4362	21651	{rating=5_Star_Review}
comment	Key	4462	4529	21659	{rating=5_Star_Review}
comment	Key	4627	4678	21641	{rating=5_Star_Review}

comment

rating 5_Star_Review

Open Search & Annotate tool

Key
 comment
 Original markups



Batch Learning config

- We will start with the config file **paum.xml** from the Module 11 materials.
- Copy this file to a writable directory, since the Batch Learning PR will need to create a **savedFiles** directory beside and write inside that.



Batch Learning config

- `<PARAMETER`
 `name="thresholdProbabilityClassification"`
 `value="0.5"/>`
 - For this example, this threshold will probably produce a class for each instance
 - Classification problems do not use the other threshold probability parameters



Batch learning config

- `<multiClassification2Binary method="one-vs-others"/>`
 - this is much faster than one-vs-another
- `<ENGINE nickname="PAUM" implementationName="PAUM" options=" -p 50 -n 5 -optB 0.0 "/>`
 - Perceptron with uneven margins
 - default options



Batch learning config

- `<INSTANCE-TYPE>comment</INSTANCE-TYPE>`
- `<ATTRIBUTE>`
 - `<NAME>Class</NAME>`
 - `<SEMTYPE>NOMINAL</SEMTYPE>`
 - `<TYPE>comment</TYPE>`
 - `<FEATURE>rating</FEATURE>`
 - `<POSITION>0</POSITION>`
 - `<CLASS/></ATTRIBUTE>`
- Take comment annotations as instances, and classify them using the rating feature.
- The classes (values of the rating features) form an unordered set (current limitation of the PR).



Batch learning config

- `<NGRAM>`
 - `<NAME>ngram</NAME>`
 - `<NUMBER>1</NUMBER>`
 - `<CONSNUM>1</CONSNUM>`
 - `<CONS-1>`
 - `<TYPE>Token</TYPE>`
 - `<FEATURE>root</FEATURE>`
 - `</CONS-1>`
 - `</NGRAM>`
- Use unigrams of *Token.root* features inside the comment annotations as the instance attributes (bag of words).
- An additional feature in the hands-on file is commented out for you to experiment with later.



Building the application

- Load the ANNIE, Tools, and Learning plugins.
- Create a new corpus called “training” and populate it from the directory **corpora/training** in the Module 11 hands-on material.



Building the application

- Create the following PRs with the default init parameters:
 - Document Reset PR
 - Annotation Set Transfer
 - ANNIE English Tokeniser
 - ANNIE Sentence Splitter
 - ANNIE POS Tagger
 - GATE Morphological Analyser



Building the application

- Create a Batch Learning PR with the configFileURL init parameter set to the **paum.xml** file from the Module 11 materials.
- Create a new Conditional Corpus Pipeline and show it.



Building the application

- Add the PRs to the pipeline as follows
 - Document Reset:
 - setsToKeep = “Key”
 - Annotation Set Transfer:
 - annotationTypes = [comment]
 - copyAnnotations = true
 - inputASName = “Key”
 - outputASName, tagASName, textTagName = “”



Building the application

- Add more PRs to the pipeline
 - English tokeniser
 - Sentence splitter
 - POS tagger
 - Morphological analyser
 - Batch Learning:
 - `inputASName, outputASName = ""`
 - `learningMode = TRAINING`



Training

- This pipeline will
 - leave the Key AS untouched,
 - copy the comment annotations to the default AS to provide the ML instances and classes,
 - run the NLP components to provide ML instance attributes, and
 - train the classifier.



Training

- Run it on the training corpus (this should take less than 1 minute)
- The classifier's model is stored in the **savedFiles** directory beside the **paum.xml** file. The model is stored in text files, but they are not meant to be human-readable.



Application

- Create a “testing” corpus and populate it from the **corpora/testing** directory.
- To apply the classifier, we need to have comment annotations *without* rating features on the default AS. These will give us the instances to classify. A simple JAPE Transducer can do this.



Application

- Create a Jape Transducer from the file *copy_comment_spans.jape* in the Module 11 material.
- Insert the transducer in the pipeline after the AS Transfer PR.
- Set the transducer:
 - `inputASName = "Key"`
 - `outputASName = ""`



Application

- Set the AS Transfer PR's run-mode to “no”
- Set the Batch Learning PR's parameters:
 - inputASName = “”
 - learningMode = APPLICATION
 - outputASName = “Output”
- The classifier will get instances and attributes from the default AS and put instances with classes in the Output AS.



Application

- Set the pipeline to run on the testing corpus, and run it. (This should also take less than a minute.)



Application

Loaded Processing resources

Name	Type

Selected Processing resources

!	Name	
	Document Reset PR_0002A	Docur
	Annotation Set Transfer_0002B	Annot
	JAPE copy_spans	Jape T
	ANNIE English Tokeniser_0002C	ANNIE
	ANNIE Sentence Splitter_0002F	ANNIE
	ANNIE POS Tagger_00033	ANNIE
	GATE Morphological analyser_00032	GATE
	Batch Learning PR_00034	Batch

Run "Batch Learning PR_00034"?

Yes
 No
 If value of feature is

Corpus: testing

Runtime Parameters for the "Batch Learning PR_00034" Batch Learning PR:

Name	Type	Required	Value
inputASName	String		<input type="text"/>
learningMode	RunMode	✓	APPLICATION
outputASName	String		Output

Run this Application



Application & evaluation

- Open a few documents and inspect the annotations:
 - *Key.comment* = the users' ratings
 - default *AS.comment* = the instances with no classes
 - *Output.comment* = the instances with ML classes



Application & evaluation

View Member's:
 Reviews I should have got my act together I placed three seperate orders and could have saved on shipping charges but its a good service overall. thank you Overall Rating

Date Reviewed: 21/12/06
 kennlind50
 Member Since: 12/12/06

View Member's:
 Reviews I HAD SHOPPING FROM SPORTS HQ BEFORE WHIS ORDER,AND I LIKE IT. SHORT SHIPPINGTIME AND THE PRODUCT WAS ALL IN ORDER. Overall Rating

Date Reviewed: 19/12/06
 bergol
 Member Since: 03/12/06

View Member's:
 Reviews Straight forward purchase due to the clear and precise on-screen instructions. In my opinion an excellent site which I will definately be putting on my favourites. Overall Rating

Date Reviewed: 01/12/06
 hobther
 Member Since: 15/11/06

View Member's:
 Reviews although i havent recieved my items as yet so i have no idea what your delivery service is like i hope it is good as if it turns out to be a good service i will recommend your service to all my friends at the golf club where i am a full member hope to be spending a lot more money with you as your products are resonably priced and will suit my friends with low budjets as well many thanks Overall Rating

In order to provide the most relevant information, comments posted prior to the past 18 months may not be included.

Type	Set	Start	End	Id	Features
comment	Output	3854	3969	17079	{prob=0.0, rating=5_Star_Review}
comment		3854	3969	14756	{}
comment	Key	4066	4229	14723	{rating=5_Star_Review}
comment	Output	4066	4229	17080	{prob=0.0, rating=5_Star_Review}
comment		4066	4229	14757	{}
comment	Key	4327	4716	14733	{rating=4_Star_Review}
comment	Output	4327	4716	17081	{prob=0.0, rating=5_Star_Review}
comment		4327	4716	14758	{}

- Sentence
- SpaceToken
- Split
- Token
- comment
- ▼ Key
- comment
- Original markups
- ▼ Output
- comment



Application & evaluation

- Now show the testing corpus and click the Corpus Quality Assurance tab.
- Select
 - Annotation Sets A = *Key*, B = *Output*
 - Annotation Types = *comment*
 - Annotation Features = *rating*
 - F-Score = F1.0-score strict
- Click “Compare”



Application & evaluation

- If every instance has been classified, then the total $P = R = F1$, because every spurious classification is paired with a missing classification
- Use the “Document statistics” sub-pane of Corpus QA to confirm this, and to see how specific documents were annotated



Application & evaluation

- Now change the Measure to “Classification”, select Cohen's Kappa, and click “Compare”.
- In addition to the document statistics with summary lines, there is now a “Confusion Matrices” sub-pane.
- Cells in these tables can be selected with the mouse (or Ctrl-A to select all) and copied with Ctrl-C, so you can paste them in to a spreadsheet.



Built-in cross-validation

- Set the pipeline up almost the same way as for training mode
 - Switch on the AS Transfer PR to copy *Key.comment* to the default AS
 - Switch off the JAPE transducer
 - **but** set the Batch Learning PR differently:
 - inputAS, outputAS = ""
 - learningMode = EVALUATION



Built-in cross-validation

- Create a new corpus “all” and populate it from the **corpora/all** directory (all the documents from training and testing.)
- Run the pipeline on the new corpus. This will take a few minutes.
- This will carry out 5-fold cross-validation as specified in the config file.



Built-in cross-validation

- The config file includes:
 - `<EVALUATION method="kfold" runs="5" ratio="0.66" />`
 - kfold ignores the ratio setting
 - holdout ignores the runs setting
- The Batch Learning PR will automatically split the corpus into 5 parts, and then
 - train on 1,2,3,4; apply on 5; then
 - train on 1,2,3,5; apply on 4; ...
 - train on 2,3,4,5; apply on 1;
 - and average the results.

Built-in cross-validation

GATE

- Cross-validation is a standard way to “stretch” the validity of a manually annotated corpus. The 5-fold averaged result is more significant than the result obtained by training on 80% of the same corpus and testing on 20% once.
- In GATE, you can't use the Corpus QA tool on the result, but you get a detailed statistical report at the end, including P, R, & F1 for each class.



Suggestions for
further experiments...



Suggestions . . .

- The config file URL is an init parameter, but the contents can be re-loaded, so you can
 - use any text editor on the config file, save the changes, and
 - re-initialize the Batch Learning PR to re-load the file with changes.



Suggestions . . .

- Try n-grams where $n > 1$
 - Change `<NUMBER>` in the config
 - Usually this is slower, but sometimes it improves quality
- combining features
 - change `<CONSUM>` in the config to 2 and uncomment the *Token.orth* element
 - this concatenates the features



Suggestions . . .

- Adjust the `thresholdProbabilityClassification`
 - Increasing it may increase precision and decrease recall, and may prevent the classifier from assigning a class to every instance.
 - Decreasing it may increase recall and decrease precision.
 - This is the “pickiness” control of the classifier.



Suggestions . . .

- Try using other features
 - *Token.string*, *Token.category*, or combinations of these with *Token.root* and *Token.orth*
- You could even include other ANNIE PRs in the pipeline and use Lookup or other annotation types.
 - You need to run the same annotation-creating PRs for training and application.
 - If the config file specifies an annotation that is missing in an instance, the ML PR will throw an exception.