CSML1010 Final Presentation

Sentiment Analysis with the Sentiment140 Dataset

Request for Feedback

Pete Gray, York U #217653247, Jan 12 2019

Hello,

I am interested in feedback for a few of the items in my presentation.

(Parts where I seek feedback are fleshed out in this .pdf. Other parts, not so much.)

[not much looking for feedback

- **Introduction** (2 minutes)
- Problem Definition (sentiment analysis)
- Dataset (link, 2009, how it was derived, labeled using emoticons, 140 is Tweet size, perfectly balanced, chronologically sequenced, early work https://www-cs.stanford.edu/people/alecmgo/papers/TwitterDistantSupervision09.pdf, etc.]

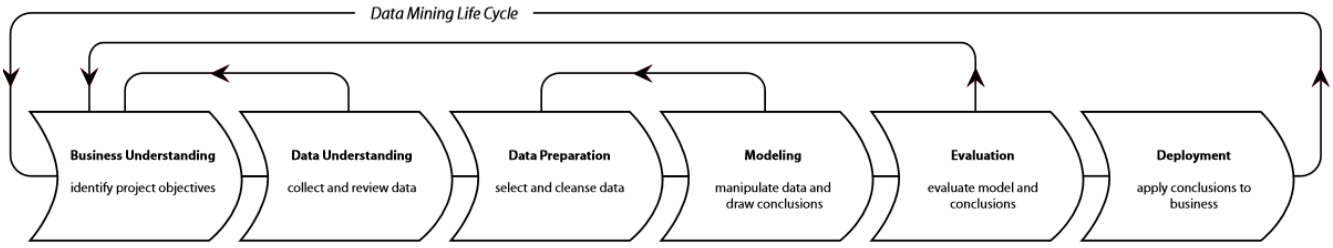
Methodology

[*Seeking feedback* – general idea, and specifics.]

4 minutes

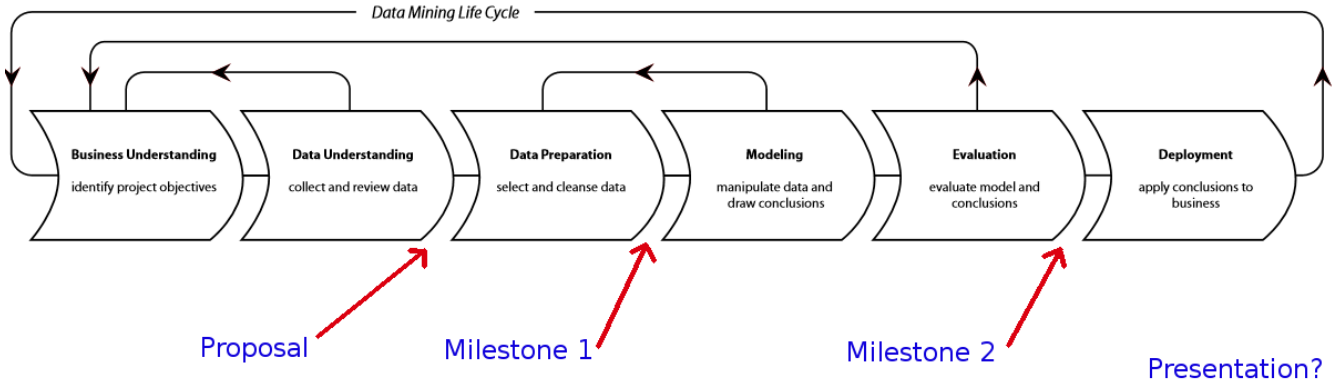
CRISP-DM

This was a roadmap I kept coming back to in Hashmat's class, and it turned out to be a roadmap I kept coming back to in this class as well.



Though, in this class, our primary motivator was not a methodology, but a life-cycle. Or, to belabour the semantics, the life-cycle was our interface with the real

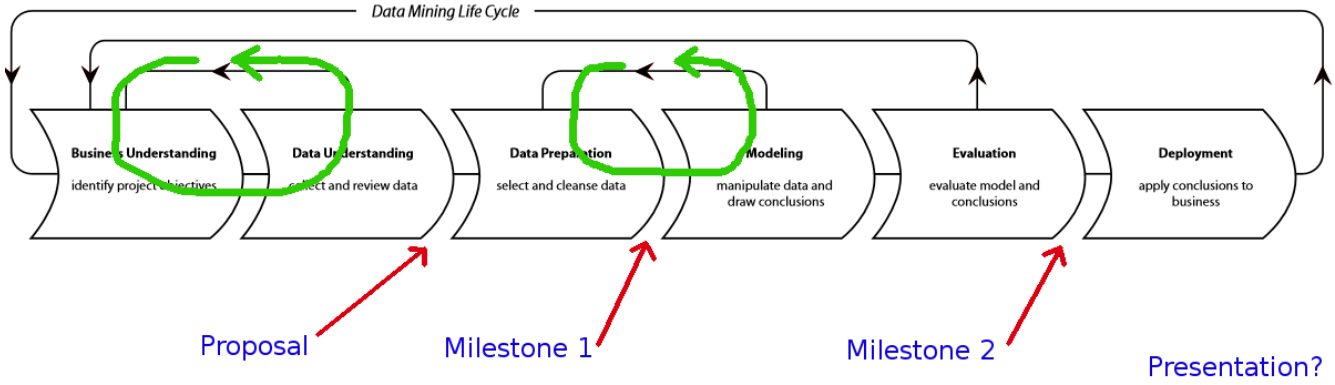
world for which we were undertaking this project, whereas the methodology was more the recipe by which we undertook this project for that real world. The two then line up, and my best interpretation of that looks like this:



A question: We cleaned our data, before we explored our data. Then, at least on my project, we went back to cleaning during data preparation, and again in modeling. This is normal? This is cool? This is part of how the Methodology and the Life-cycle intersect?

A quick description of the things tried at different steps – 3 types of feature engineering, 3 types of feature selection on each, 7 standalone models evaluated with 4 evaluation metrics, 3 types of ensembling with several base models, announce winners (based on evaluations) as Logistic Regression and LinearSVC standalones with TF-IDF vectorizations. Compare and contrast Interpretability of the two.

At any rate, I would like to take a minute or two to talk about those "loops", in the diagram, and in the real-world necessity to go back and improve something along the way:



For example:

- The time I went back to straight string cleaning to remove numbers and usernames because I was getting insanely wide vectorizations full of junk

that slowed down any attempt at feature selection,

- The time I went back to feature selection because only the algorithm I'd benchmarked my features on was giving good results

- The times I doubled back to all sorts of places, trying to achieve better evaluation metric values

- Many of the improvements I got to my models were iterative, and most were achieved by returning to an earlier point in the process

One last question on Methodology – does the presentation count as our "deployment"?

Results

4 minutes

- Plan to show lots of colourful graphs, leaning towards side-by-side comparisons a "good" model (One that performed well across all evaluation metrics, either LinearSVC or Logistic Regression) with a good graph, and one of the other models with a dubious graph.

- Show a couple of tables demonstrating the effect of bigger dataset and more features selected on the various models

- Mention the changes that brought improvement to Ensemble models, even if it didn't make them more accurate than standalones

I don't think I'm looking for much feedback on results – you've seen dozens of my graphs and such – I plan to make it quick, clean, colourful and engaging.

Discussion

4 minutes

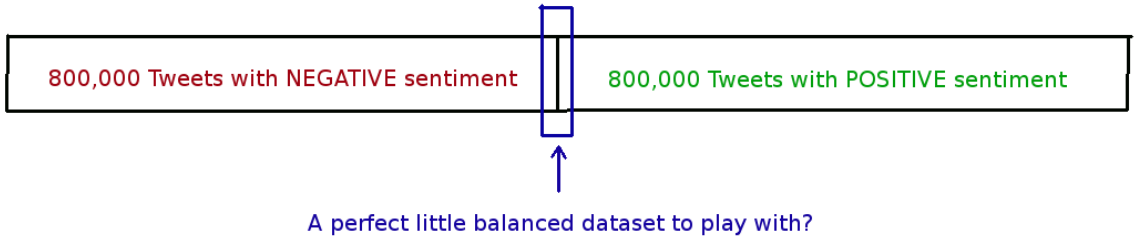
Model performance with improperly prepared data.

Most things I did to "re-prepare" my data (more informed cleaning, more ambitious feature selection) improved the evaluation metrics of my models.

Better preparation, better results, right?

Not so in the case of data sampling.

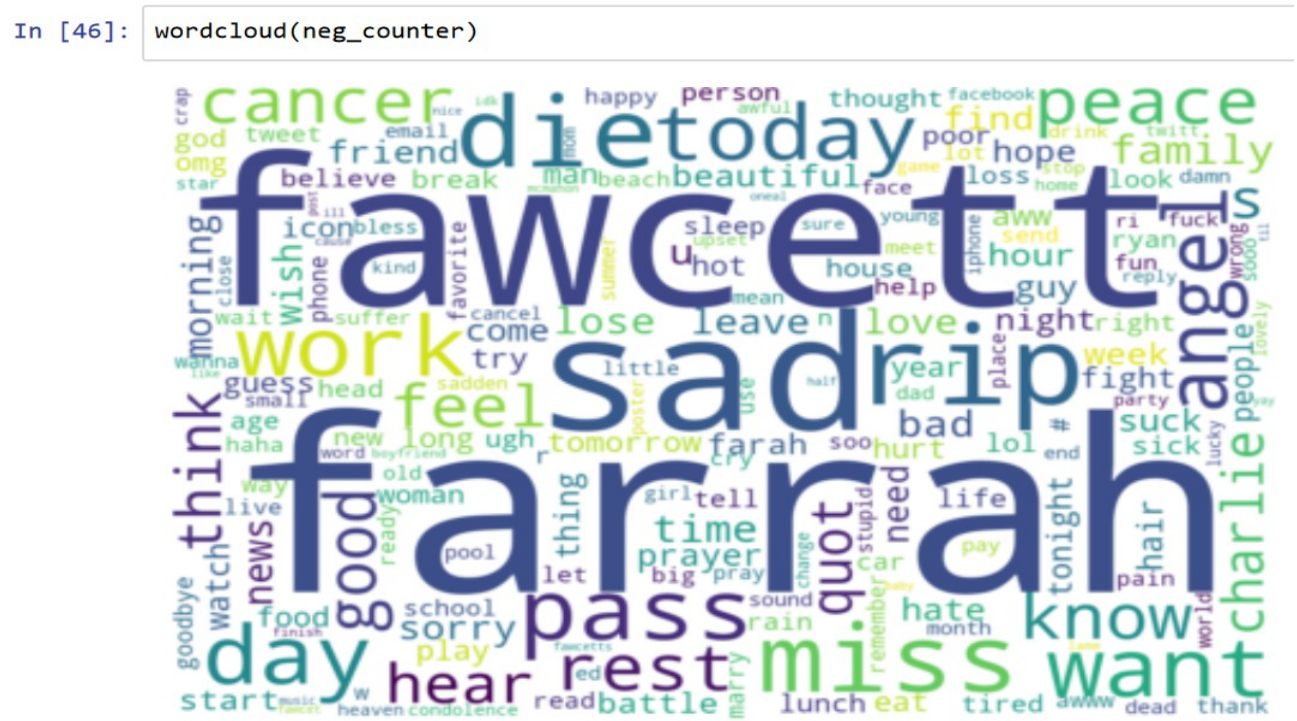
So, my raw dataset had 800,000 negative, and then 800,000 positive Tweets, like this:



"*Can anyone tell me what is wrong with this approach to sampling my dataset?*"

I started out this way, even though I knew it was incorrect – I wanted to do a little "case study" on the effects of doing it wrong, vs. doing it right.

The first sign of trouble came during Data Exploration. I had made Word Clouds of the most significant Positive and Negative words, and found myself looking at this:

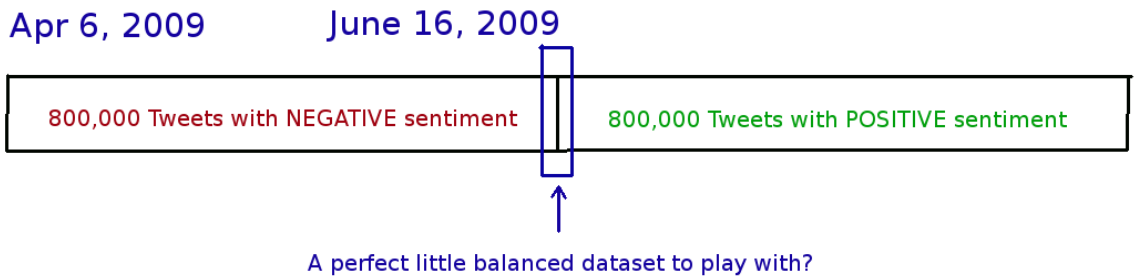


That's pretty odd. And, in case the meaning of that was lost on me, I came upon this again when I was doing feature selection, in particular with my "Bag Of N-Grams" vectorization:

```
actress farrah -       - actress farrah fawcett -       - add thank -       - a
ge sad -       - angel angel -       - angel beloved -       - angel beloved ri
p -       - angel beloved rip pass -       - angel farrah -       - angel farra
h fawcett -       - angel heaven -       - angel quot -       - angel quot canc
er -       - angel quot cancer die -       - angel sad -       - aww farrah -
- awww farrah -       - awww farrah fawcett -       - bad feel -       - bad wa
nt -       - battle cancer -       - battle cancer sad -       - be die -
- be get -       - be use -       - be watch -       - beautiful rip -       -
bed night -       - bed tomorrow -       - believe die -       - believe pass -
- beloved rip -       - beloved rip pass -       - blog post -       - break di
e -       - can believe -       - can believe die -       - can believe pass -
- can go -       - can wait -       - can wait see -       - cancer age -
- cancer die -       - cancer life -       - cancer sad -       - charlie angel
-       - charlie angel farrah -       - charlie angel farrah fawcett -       -
charlie angel quot -       - charlie angel quot cancer -       - charlie angel
quot cancer die -       - charlie angel sad -       - could go -       - could
use -       - couple day -       - day party -       - day sad -       - day ti
me -       - day week -       - die be -       - die feel -       - die miss -
- early go -       - easy make -       - eat want -       - ed mcmahon -
- ed mcmahon farrah -       - ed mcmahon farrah fawcett -       - family friend
-       - family friend sad -       - family sad -       - farah fawcett -
- farrah family -       - farrah faucett -       - farrah fawcet -       - farr
ah fawcett -       - farrah fawcett age -       - farrah fawcett angel -
- farrah fawcett angel beloved -       - farrah fawcett angel beloved rip -
- farrah fawcett battle -       - farrah fawcett battle cancer -       - farrah
fawcett cancer -       - farrah fawcett cancer age -       - farrah fawcett can
cer die -       - farrah fawcett charlie -       - farrah fawcett charlie angel
-       - farrah fawcett day -       - farrah fawcett ed -       - farrah fawce
tt ed mcmahon -       - farrah fawcett hair -       - farrah fawcett hear -
- farrah fawcett hear pass -       - farrah fawcett heart -       - farrah fawc
ett icon -       - farrah fawcett miss -       - farrah fawcett morning -
- farrah fawcett morning die -       - farrah fawcett news -       - farrah faw
cett pass -       - farrah fawcett peace -       - farrah fawcett rest -
- farrah fawcett rest peace -       - farrah fawcett rip -       - farrah fawce
```

"*Can anyone tell me what's going on here?*"

It turns out, that the Tweets are chronological in each of the Positive and Negative parts of the .csv file.



All the Negative Tweets in my poorly sampled small dataset were from June 16, 2009. Around 8:40am, PDT. News of Farrah's passing was just making the rounds. Clearly, my data is inundated by this.

But what effect does this have on the models we are training?

Turns out, this gives them an almost 10% advantage, or more, across the board.

In this table, I compare a few of the evaluation metrics for this non-random dataset, vs a properly sampled dataset. Same number of rows. Same number of features. Huge difference. Never came close, even with matricies 20x as big.

Assorted Evaluation Metrics Without | With Random Sampling of Dataset

(2,000 rows, 2,000 features selected, TF-IDF encoding)

Classifier Type	Accuracy	F1-Score	ROC/AUC	Test Set Errors
Logistic Regression	.77 | .68	.76 | .65	.845 | .740	.17 | .17
LinearSVC	.78 | .68	.78 | .66	.854 | .729	.16 | .16
Random Forest	.71 | .63	.72 | .62	.825 | .709	.27 | .4o

I was curious: What if I trained a model using the Farrah-heavy data, and then ran it on a test set that didn't have that going on? I'd get really nice training and cross-validation results – and then some horrible number on my test set.

While it wouldn't have been too much code to write, I decided it wasn't a good use of my time. No marks to get, no improvements to my overall results. I know I shouldn't use the bad dataset. Belabouring it beyond a quick A/B comparison has no real world value, so I skipped it.

But this little ad hoc case study does demonstrate how much a little bit of thoughtless data sampling can have a profound effect on the performance of our

models.

Conclusions

2 minutes

- Feature Engineering with TD-IDF was superior to Bag-O-Words and Bag-O-N-Grams, at least with the models I was using successfully
- Feature selection with Extra Trees was superior to K-Means, and much easier to run than RFE
- Logistic Regression and Linear Support Vector Classifier had better evaluation metrics than DecisionTree, KNN, RandomForest, or MultinomialNB,
- Important to Sample Data Randomly!
- I did not succeed in getting Ensemble Methods to work better than standalone models

Future Work

2 minutes

I would like to try this Sentiment Classifier on other data. See how well it performs on text from different domains.

I would like to apply a “skilled” General Sentiment Classifier Model to Conversational datasets. We could then engineer features based on a participant's change in sentiment. Did the conversation see an improvement in their positivity, or did negativity increase, as the conversation progressed? If change in sentiment were a feature we could use, we could look into classifying the text of their conversation partner, in terms of it being correlated with an improvement, or a decline, in the user's sentiment.

I would imagine I would want a more sophisticated deep-learning model, using a modern language model for feature engineering, and trained on a broader scope of data, for that future work.

And that's 18 minutes.

I understand we get 20? I'll say my thank yous for attending, or for the advance feedback, as the case may be. And be glad to take any questions you may have.

Thank you so much for everything, see you Friday!