Problem Set 3 - GHOST

Genre, Happening, or Sentiment Tagger

Michael Berezny, Pat Putnam, Sushant Kafle

METHOD: Unused Ideas

- Hand-written features e.g.
  - Sentence length
  - Word presence in existing sentiment corpora
  - POS tagging
  - Punctuation
  - etc.
- Word Vector Representation
  - Used the vector representation of each word in the sentence to get a vector representation of the sentence.
  - Use the sentence vectors to train a prediction model
  - The average accuracy for the sentiment task was around ~58%.

METHOD: Final

- Evaluated a multitude of classifiers against the training data.
- Ranked the models based on the mean accuracy in a k-fold cross-validation task. Selected the top ranked models (n=3).

- Separately tuned the hyperparameters of the models for each task using Grid Search.
- Made use of a Majority Voting classifier to combine predictions from our top ranked models to get our final prediction.

RESULTS: Accuracies

	Genre	Sentiment	Category
Ghost Accuracy	89.42%	74.44%	61.22%
Majority Class Baseline	82.22% (Genre B)	49.20% (Negative)	13.70% (Money Issue)
Probability Baseline	50.00%	33.33%	12.50%

- Majority Class Baseline: The accuracy if we used the most common label every time.
- Probability Baseline: The expected accuracy if we guessed on every label.

RESULTS: Confusion Matrices

Genre	Predicted Genre A	Predicted Genre B	Recall
Actually Genre A	298	45	86.88%
Actually Genre B	159	1427	89.97%
Precision	65.21%	96.94%	

Sentiment	Predicted Negative	Predicted Neutral	Predicted Positive	Recall
Actually Negative	716	8	225	75.45%
Actually Neutral	30	12	21	19.05%
Actually Positive	202	7	708	77.21%
Precision	75.53%	44.44%	74.21%	

RESULTS: Confusion Matrices

Events	Predicted Fear of Physical Pain	Predicted Attending Event	Predicted Communication Issue	Predicted Going to Places	Predicted Legal Issue	Predicted Money Issue	Predicted Outdoor Activity	Predicted Personal Care	Recall
Actually Fear of Physical Pain	**19**	2	5	0	0	1	4	11	**45.24%**
Actually Attending Event	2	**18**	3	6	1	3	7	1	**43.90%**
Actually Communication Issue	3	13	**7**	3	2	0	6	4	**18.42%**
Actually Going to Places	1	3	1	**38**	1	0	0	0	**86.36%**
Actually Legal Issue	0	2	0	1	**38**	1	1	0	**88.37%**
Actually Money Issue	2	4	2	0	2	**36**	1	0	**76.60%**
Actually Outdoor Activity	6	4	1	1	0	0	**29**	2	**67.44%**
Actually Personal Care	0	6	0	5	0	5	4	**25**	**55.56%**
Precision	**57.58%**	**34.62%**	**36.84%**	**70.37%**	**86.36%**	**78.26%**	**55.77%**	**58.14%**	

OBSERVATIONS & DISCUSSION

- GHOST surpassed the baselines of all tasks, especially the non binary classes
- GHOST slightly overpredicted Genre A, likely due to the balanced training data
- GHOST failed to identify most Neutral Sentences with only 19% recall neutral sentiment is generally harder to detect as it is the absence of sentiment
- GHOST was the best at identifying the Legal Issues event with 86% precision and 88% recall, likely due to the identifiable vocabulary of legal topics

OBSERVATIONS & DISCUSSION

- GHOST overpredicted the Attending Events Category the most, likely due similarity with the Going to Places and Outdoor Activity categories
- We initially were interested in deep learning techniques, but ruled these methods out due to the small amount of training data
- Michael developed an extensive set of features
  - Standalone, their accuracy was beaten by the count vectorizer
  - Ran out of time trying to combine custom features with count vecotrizer
- We expected word2vec (more accurately sentence2vec) features to work well on category labeling task - as it had less data per class. But, sadly it didn’t meet our expectations.

CHALLENGES

- **Last minute bug-fix**
  - **Issue:** Found wildly inaccurate output upon first run against test data
  - **Root cause:** Crossed wires in accuracy evaluation and output
  - **Solution:** Minor tweaks to accuracy evaluation and output portions of the code
- **Time-cost of external resources**
  - **Issue:** Planned to use Stanford NER system, extremely long time to evaluate all data points
  - **Root cause:** Spinning up and down the JVM excessively
  - **Solution:** Removed this portion of our feature set
- **Inconsistencies in given data**
  - **Issue:** The training data had some inconsistencies - misspelled categories, and extra spacing
  - **Solution:** Corrected the category names, but missed the extra spaces

TASK DISTRIBUTION

Michael Berezny - 33%	Pat Putnam - 33%	Sushant Kafle - 33%
• Hand rules and extraction • Results analysis • Presentation	• Contributed in the design and implementation of the prediction models.	• Contributed in the design and implementation of the prediction models.

CONCLUSION

- GHOST surpassed both the baselines (Equal Probability Baseline and Majority Class Baseline)
- Using out-of-the-box features and classifiers still lead to a fairly successful model
    - We had hoped to make use of more hand selected features, but we unable to properly incorporate them into our system due
- Our team’s varied backgrounds and experiences helped inform the design of our system

References

Minqing Hu and Bing Liu. "Mining and Summarizing Customer Reviews." Proceedings of the ACM SIGKDD International Conference on Conference on Knowledge Discovery and Data Mining (KDD-2004), Aug 22-25, 2004, Seattle, Washington, USA.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 363-370. http://nlp.stanford.edu/~manning/papers/gibbscrf3.pdf