Aspect Based Sentiment Analysis Using Machine Learning

Umair Cheema
Athar Pasha

OVERVIEW

❑ Project Objective

❑ Data Preparation

❑ Exploratory Data Analysis

❑ Methods and Techniques

❑ Results

❑ Q & A

PROJECT OBJECTIVE

❑ Combine Machine Learning and Natural Language Processing to conduct Aspect Based Sentiment Analysis of Customer Reviews on Restaurants.

❑ What is Aspect Based Sentiment Analysis?

Aspect level analysis directly looks at the <u>opinion</u> and its <u>target</u> instead of just looking at document, paragraph, sentence or phrase level sentiment.

ASPECT LEVEL SENTIMENTS

- Example

Seafood platter was delicious but wine options were limited and ridiculously expensive.

Aspect	Polarity
Food quality	Positive
Drinks options	Negative
Drinks Price	Negative

DATA PREPARATION

- Downloaded SemEval 2016 XML with Review, Entity/ Attribute and Sentiment Anotations
- XML to Pandas (Multilabel and Multidimensional transformations)
- Data Cleaning (Removal of punctuations, Case folding, Tokenization etc)

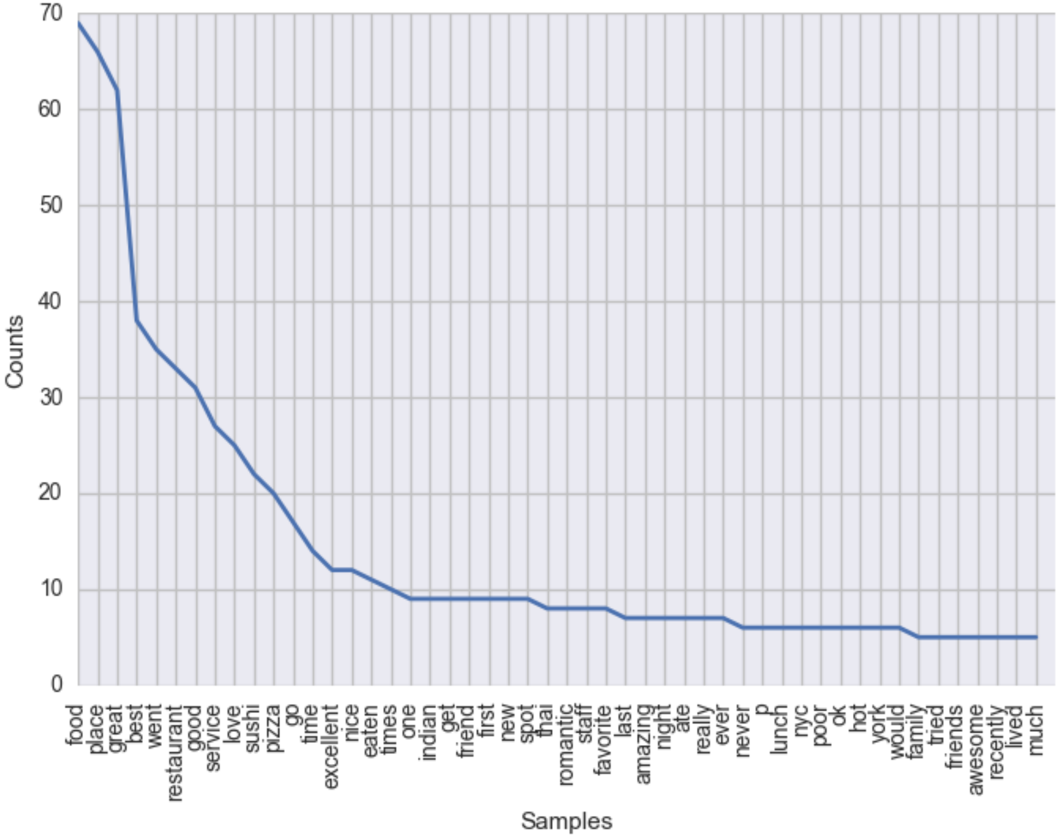
DATA PREPARATION

- ❑ Updated Opensource tool to prepare WordEmbeddings using Yelp data by modifying TensorFlow methods in the new version of Tensorflow

https://github.com/titipata/yelp_dataset_challenge

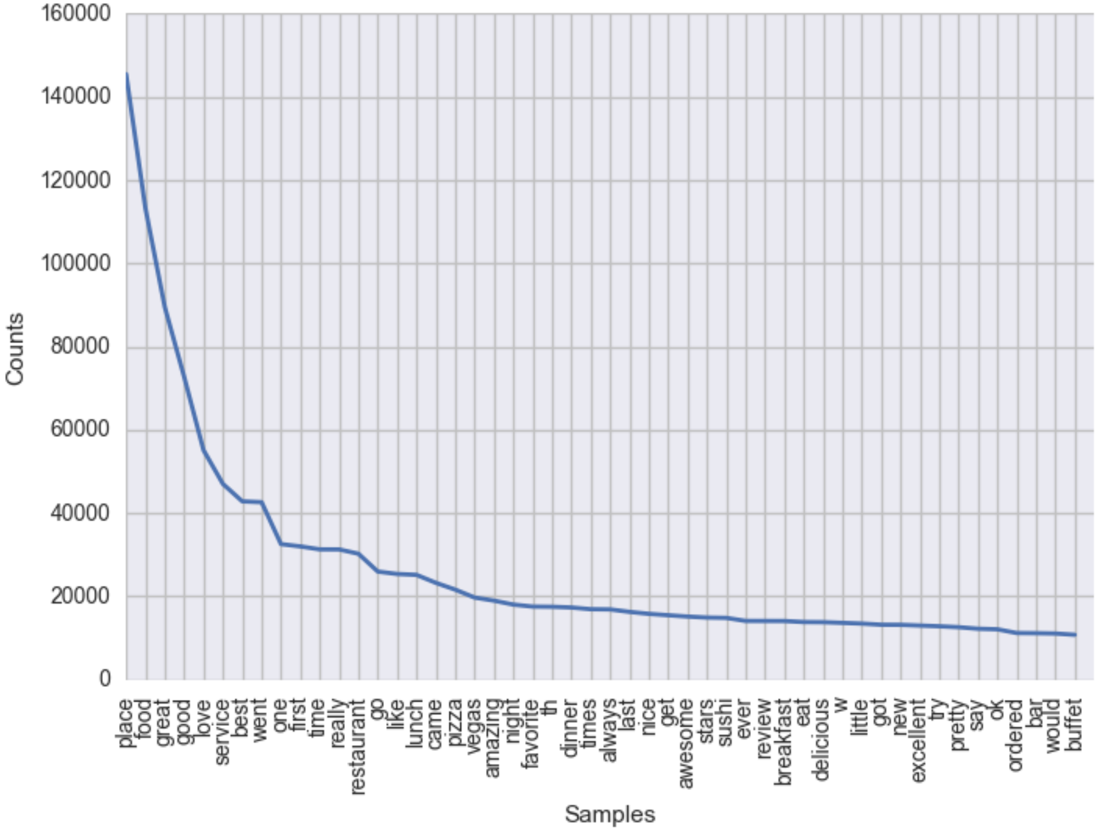
Exploratory Data Analysis

SemEval Dataset



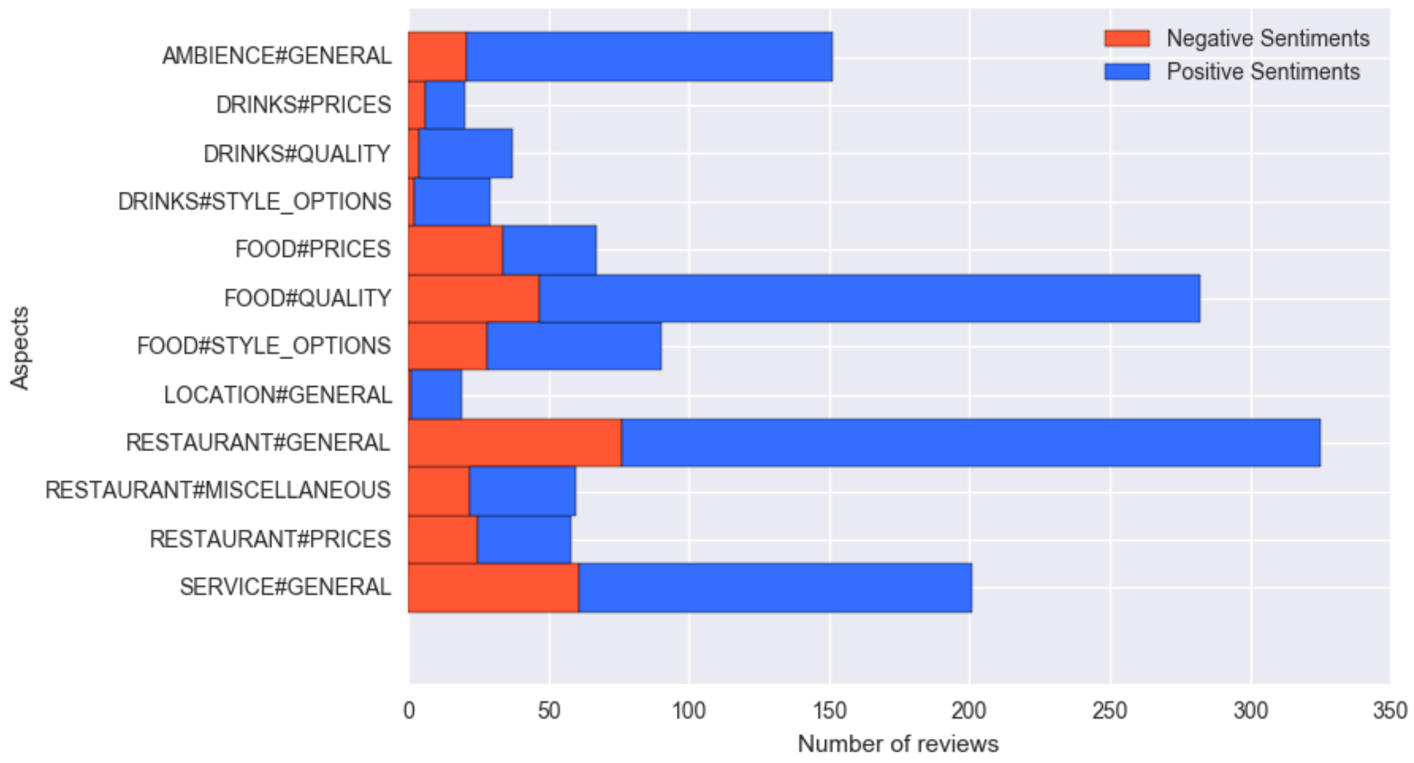
Exploratory Data Analysis

Yelp Dataset



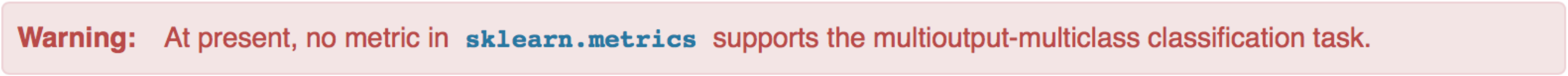
Exploratory Data Analysis

Aspect Sentiment Class Distribution



CHALLENGES

- Only 350 annotated samples to train models
- Imbalanced Classes
- No built-in scikit-learn function available for evaluating Multidimensional classification



Warning: At present, no metric in `sklearn.metrics` supports the multioutput-multiclass classification task.

METHODS AND TECHNIQUES

❑ Implemented Custom F1 micro Scoring Function described in “A MFoM Learning Approach to Robust Multiclass Multi-Label Text Categorization” by Gao et al.

$$F_1^M = 2[\sum_{i=1}^{N} R_i \sum_{i=1}^{N} P_i] / N[\sum_{i=1}^{N} R_i + \sum_{i=1}^{N} P_i]$$

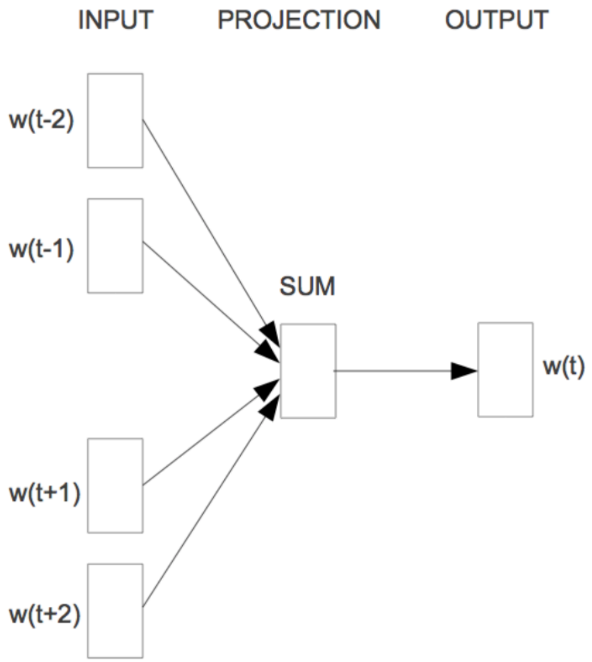
$$F_1^{\mu} = 2\sum_{i=1}^{N} TP_i /[\sum_{i=1}^{N} FP_i + \sum_{i=1}^{N} FN_i + 2\sum_{i=1}^{N} TP_i]$$

METHODS AND TECHNIQUES

- Prepared following features
- Bag of n grams
- POS Tags and Tokens
- Domain Specific CBOW Word Embeddings
- Domain Specific Skip gram Word Embeddings
- Paragraph Vector Models and Inferred Features

METHODS AND TECHNIQUES

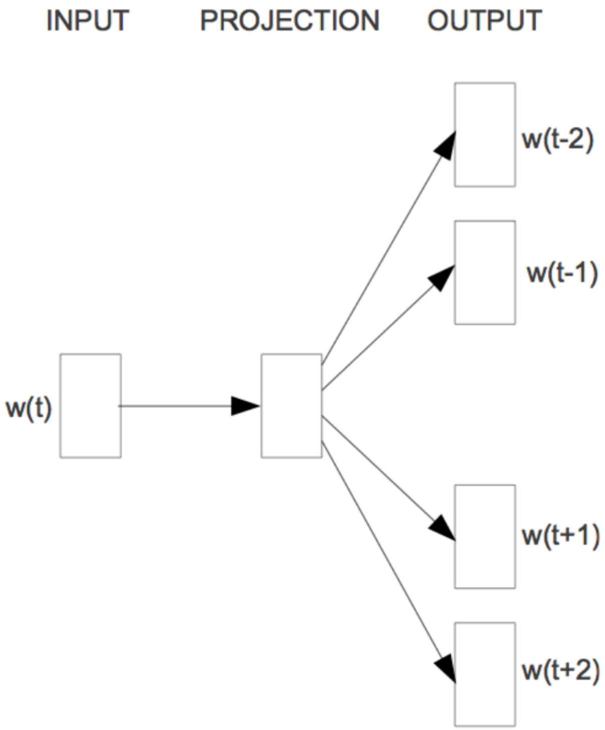
- Word Embeddings (CBOW)



Continuous bag-of-words (Mikolov et al., 2013)

Methods and Techniques

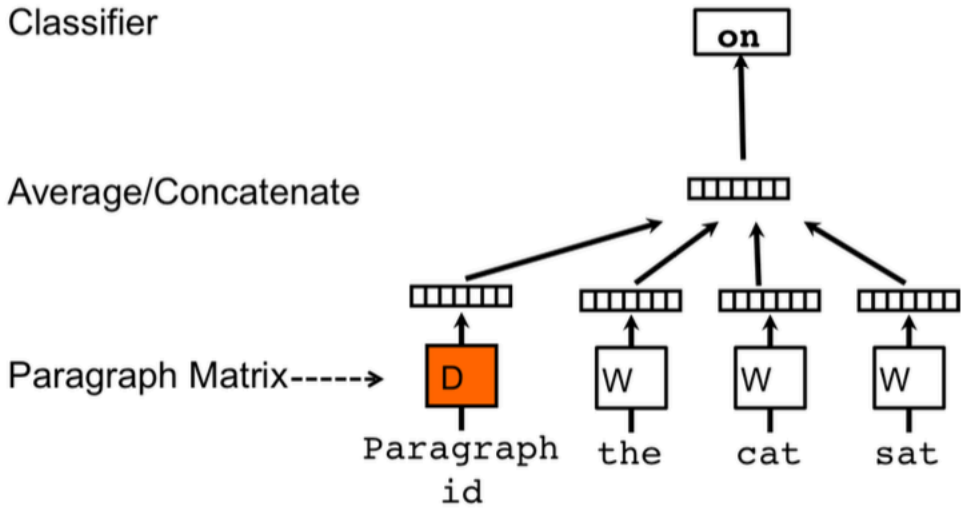
- Word Embeddings (Skip-gram)



Skip-gram (Mikolov et al., 2013)

METHODS AND TECHNIQUES

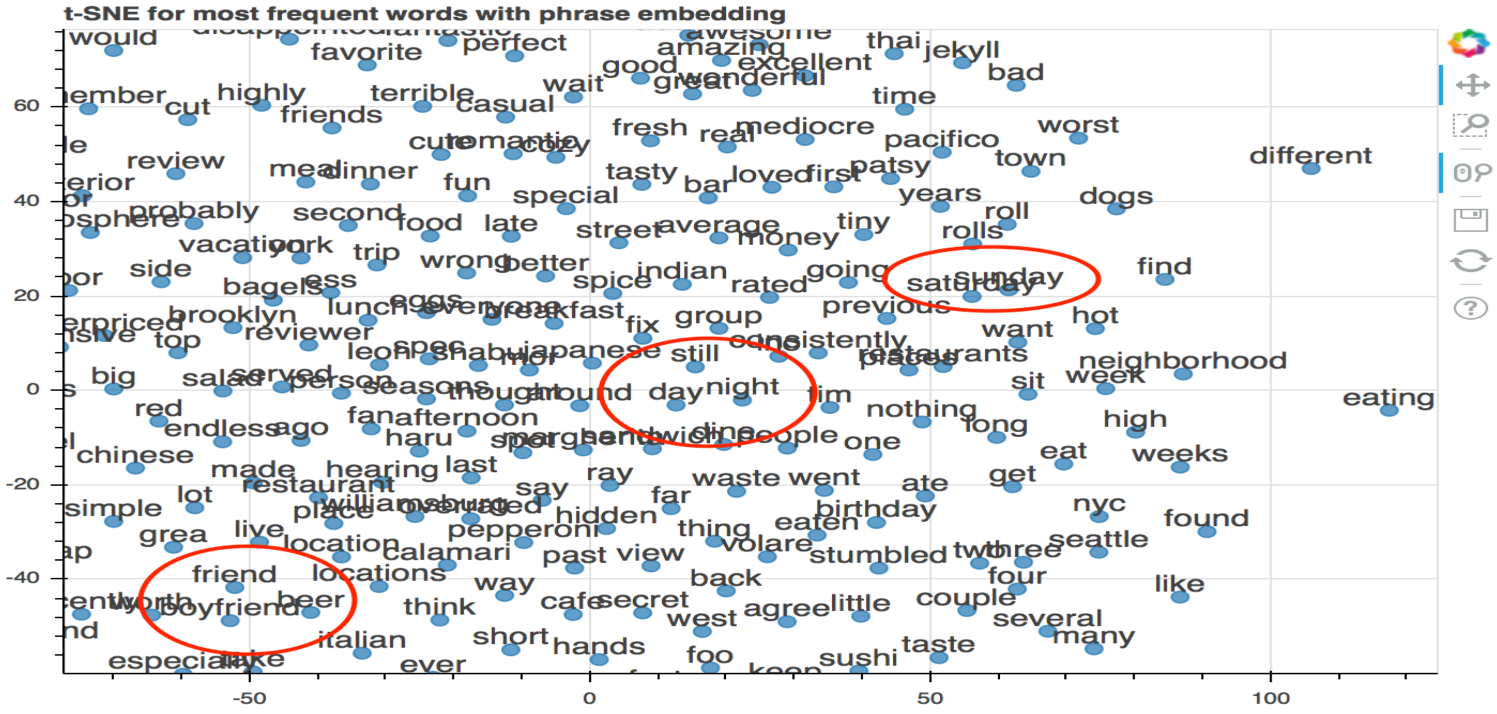
- Paragraph Vectors



(Mikolov et al ,2014)

METHODS AND TECHNIQUES

- t-distributed Stochastic Neighbor Embedding (t-SNE)



METHODS AND TECHNIQUES

- ❑ Classification Algorithms
  - ❑ Support Vector Machines
  - ❑ RandomForests

Results

Method	F1(Aspect)	Polarity
RandomForest (Bag of Words)	0.710	0.84
RandomForest (Bag of Words + POS)	0.732	0.83
RandomForest (Word2Vec CBOW)	0.715	0.871
SVM(Word2Vec CBOW)	0.739	0.913
SVM(Word2Vec Skip gram)	0.752	0.926
SVM(Word2Vec Phrase detection +CBOW)	0.760	0.927
SVM (Doc2Vec)	0.705	0.837

CONCLUSIONS

- Aspect Based Sentiment Analysis is a very challenging Multilabel and Multiclass classification problem.
- Domain specific Word Embeddings is an invaluable tool for converting textual features into Vector Space Model.
- A Multilabel SMOTE oversampling should have been used to balance the class distribution of the labelled datasets.

References

- Bing Liu, 2015. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. 1 Edition. Cambridge University Press.*
- Sebastian Ruder. 2016. *On word embeddings - Part 1. [ONLINE] Available at: http://sebastianruder.com/word-embeddings-1/. [Accessed December 2016]*Methods and Techniques

REFERENCES

- Quoc Le, Tomas Mikolov 2014. *Distributed Representations of Sentences and Documents. Proceedings of the 31st International Conference on Machine Learning, pp. 1188-1196*
- Tomas Mikolov et al 2013. *Efficient Estimation of word representations in vector space* arXiv preprint arXiv:1301.3781

References

- Sheng Gao et al 2004. *A MFoM Learning Approach to Robust Multiclass Multi-Label Text Categorization. Proceedings of the 21st International Conference on Machine Learning, pp. 42*

Q & A