Final Project

Overview

- A large-scale content project.
- We will provide some subjects about natural language processing for your choice.
- We need a lot of communication in every stage to accomplish this project.

Task list

- Stock Price Movement Prediction
- Image Caption Generation
- Dialogue System
- Style Transfer
- Summarization
- Recommender system
- Intent detection

Stock Price Movement Prediction

- Mining textual documents and time series concurrently, such as predicting the movements of stock prices based on the contents of the news articles, is and emerging topic in data mining and text mining community.

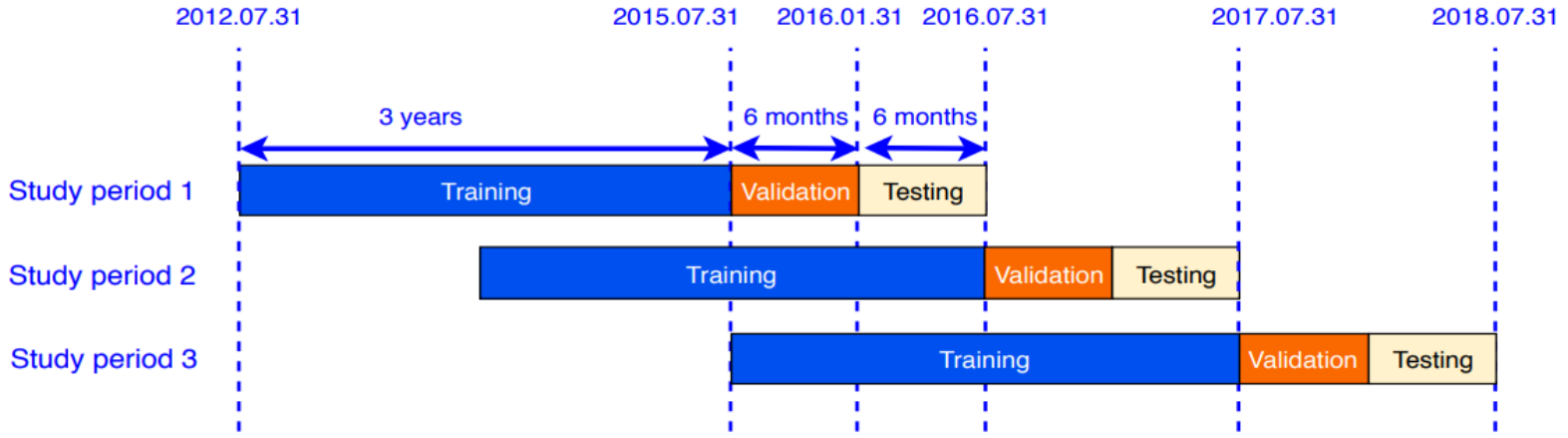


Figure 2. Study periods.

Dataset

- https://www.kaggle.com/aaron7sun/stocknews

	Date	Open	High	Low	Close	Volume	Adj Close
0	2016-07-01	17924.240234	18002.380859	17916.910156	17949.369141	82160000	17949.369141
1	2016-06-30	17712.759766	17930.609375	17711.800781	17929.990234	133030000	17929.990234
2	2016-06-29	17456.019531	17704.509766	17456.019531	17694.679688	106380000	17694.679688
3	2016-06-28	17190.509766	17409.720703	17190.509766	17409.720703	112190000	17409.720703
4	2016-06-27	17355.210938	17355.210938	17063.080078	17140.240234	138740000	17140.240234
5	2016-06-24	17946.630859	17946.630859	17356.339844	17400.750000	239000000	17400.750000
6	2016-06-23	17844.109375	18011.070312	17844.109375	18011.070312	98070000	18011.070312
7	2016-06-22	17832.669922	17920.160156	17770.359375	17780.830078	89440000	17780.830078
8	2016-06-21	17827.330078	17877.839844	17799.800781	17829.730469	85130000	17829.730469
9	2016-06-20	17736.869141	17946.359375	17736.869141	17804.869141	99380000	17804.869141
10	2016-06-17	17733.439453	17733.439453	17602.779297	17675.160156	248680000	17675.160156
11	2016-06-16	17602.230469	17754.910156	17471.289062	17733.099609	91950000	17733.099609
12	2016-06-15	17703.650391	17762.960938	17629.009766	17640.169922	94130000	17640.169922
13	2016-06-14	17710.769531	17733.919922	17595.789062	17674.820312	93740000	17674.820312
14	2016-06-13	17830.500000	17893.279297	17731.349609	17732.480469	101690000	17732.480469

	Date	News
0	2016-07-01	A 117-year-old woman in Mexico City finally re...
1	2016-07-01	IMF chief backs Athens as permanent Olympic host
2	2016-07-01	The president of France says if Brexit won, so...
3	2016-07-01	British Man Who Must Give Police 24 Hours' Not...
4	2016-07-01	100+ Nobel laureates urge Greenpeace to stop o...
5	2016-07-01	Brazil: Huge spike in number of police killing...
6	2016-07-01	Austria's highest court annuls presidential el...
7	2016-07-01	Facebook wins privacy case, can track any Belg...
8	2016-07-01	Switzerland denies Muslim girls citizenship af...
9	2016-07-01	China kills millions of innocent meditators fo...
10	2016-07-01	France Cracks Down on Factory Farms - A viral ...
11	2016-07-01	Abbas PLO Faction Calls Killer of 13-Year-Old ...
12	2016-07-01	Taiwanese warship accidentally fires missile t...
13	2016-07-01	Iran celebrates American Human Rights Week, mo...
14	2016-07-01	U.N. panel moves to curb bias against L.G.B.T....
15	2016-07-01	The United States has placed Myanmar, Uzbekist...
16	2016-07-01	S&P revises European Union credit rating t...
17	2016-07-01	India gets $1 billion loan from World Bank for...
18	2016-07-01	U.S. sailors detained by Iran spoke too much u...
19	2016-07-01	Mass fish kill in Vietnam solved as Taiwan ste...

Image Caption Generation

- http://cocodataset.org/
- Image caption, automatically generating natural language descriptions according to the content observed in an image, is an important part of scene understanding, which combines the knowledge of computer vision and natural language processing.

A man is skate boarding down a path and a dog is running by his side.
A person riding a skate board with a dog following beside.
This man is riding a skateboard behind a dog.

Dialogue Generation

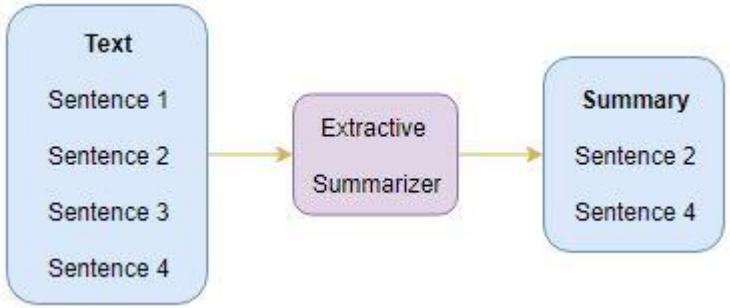
- Dialogue is notoriously hard to evaluate. Past approaches have used human evaluation.
  - Task-oriented systems
  - Non-task-oriented systems

Style Transfer without parallel data

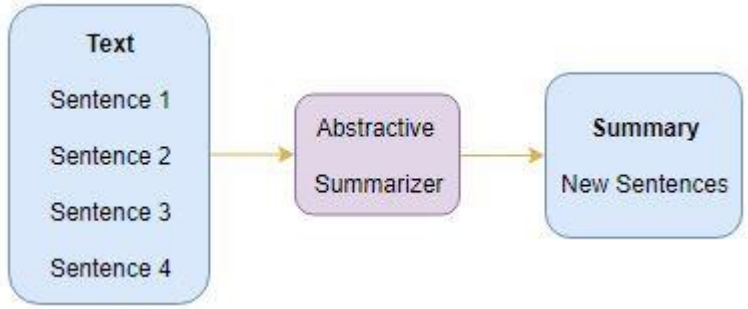
Model	Positive To Negative
Original	the happy hour crowd here can be fun on occasion.
Hu et al. (2017)	the shame hour crowd sucked can be sloppy on occasion.
Li et al. (2018)	crowd here can be ok on occasion.
Shen et al. (2017)	the wait hour , won't be happy at home down.
Prabhumoye et al. (2018)	the worst service at the food is the worst.
Our model(content preserving)	the unhappy hour crowd here can be disappointed on occasion.
Original	the menudo here is perfect.
Hu et al. (2017)	the terrible here is awkward.
Li et al. (2018)	sadly the menudo here is inedible.
Shen et al. (2017)	the family here is an understatement.
Prabhumoye et al. (2018)	the fare is horrible.
Our model(content preserving)	the menudo here is nasty.
Original	the service was excellent and my hostess was very nice and helpful.
Hu et al. (2017)	the awful was nasty and my hostess was very nasty and helpful.
Li et al. (2018)	what was too busy to my hostess than nothing to write trash.
Shen et al. (2017)	the service was dirty and the office was very nice and helpful and.
Prabhumoye et al. (2018)	the service is ok and the food wasn't even and they were halls.
Our model(content preserving)	the service was terrible and my hostess was very disappointing and unhelpful.

Summarization

- Summarization is the task of producing a shorter version of one or several documents that preserves most of the input's meaning.



Extractive Summarization



Abstractive Summarization

- https://www.analyticsvidhya.com/blog/2019/06/comprehensive-guide-text-summarization-using-deep-learning-python/

Intent detection

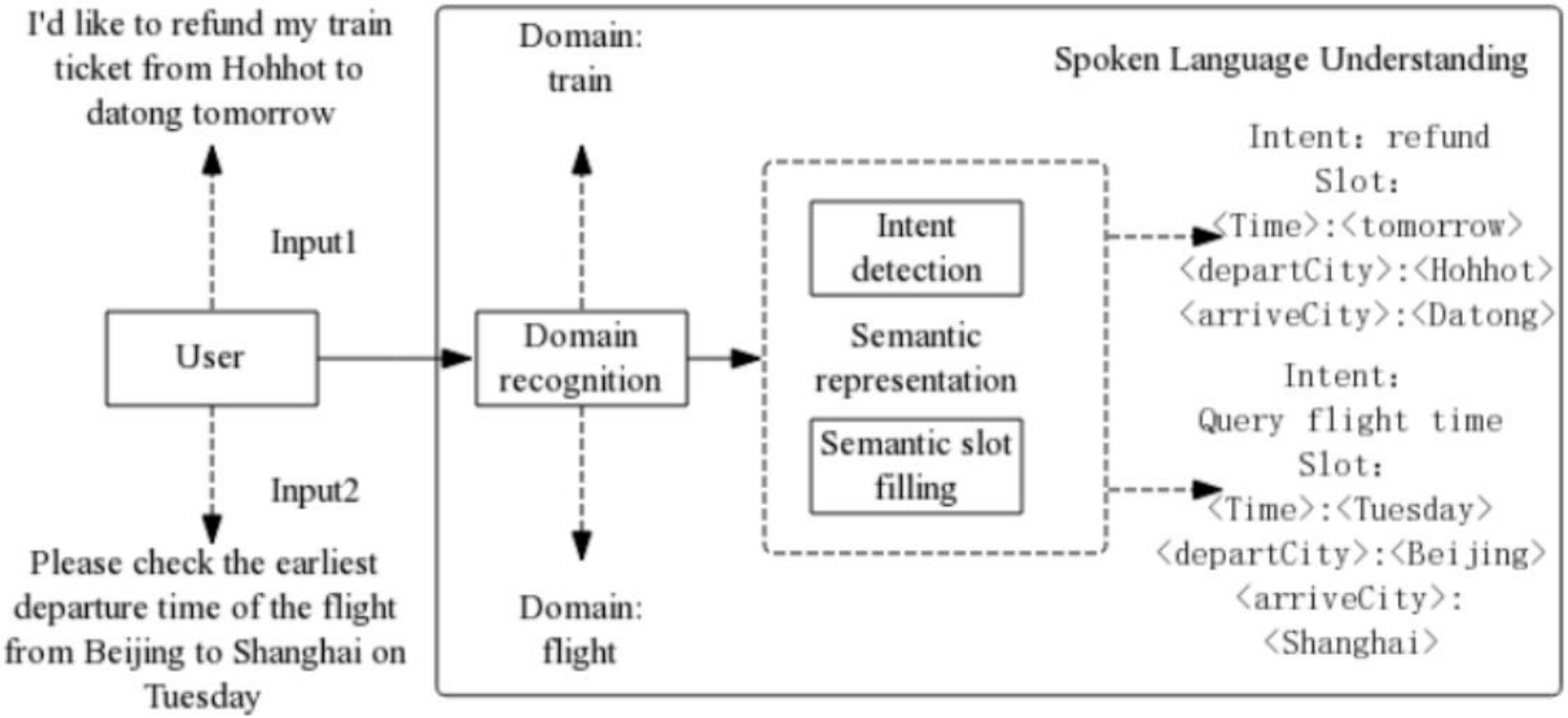


Figure 2. An instance diagram of intent detection.

Deep Learning Techniques for NLP Downstream Tasks

Problem definition

- Text classification
- Not enough data
- Class imbalance
- Data augmentation

What do we do???

Approaches

- Oversampling techniques
- Sampling from unobserved data
- Generative

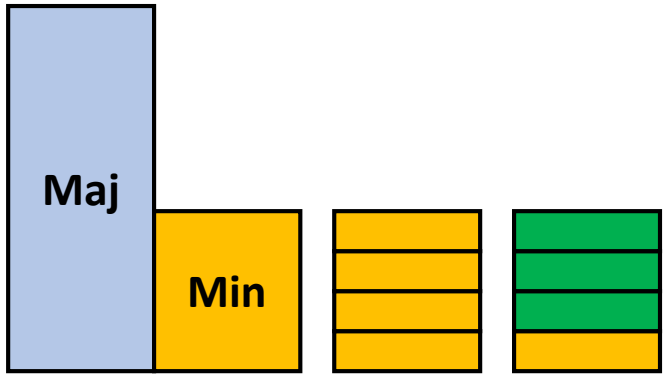
- Synthetic data vs Real data
- Unobserved data vs Observed data

Approaches

- **Oversampling techniques**
- Sampling from unobserved data
- **Generative**

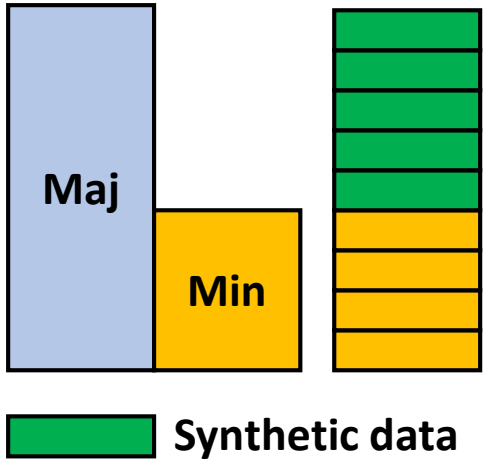
- **Synthetic data vs Real data**
- Unobserved data vs Observed data

Experiments



Synthetic data vs Real data

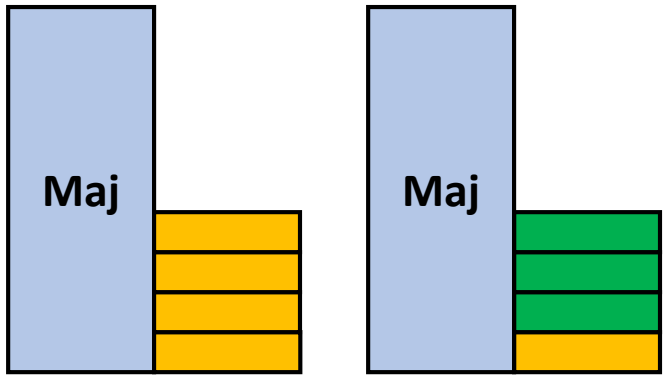
Synthetic data generator: LSTM, SeqGAN, VAE
Baselines: Real data, SMOTE
Classifiers: Linear SVM, NN
Evaluation metrics: F1, G-Mean
Datasets: Quora, Amazon, Yelp

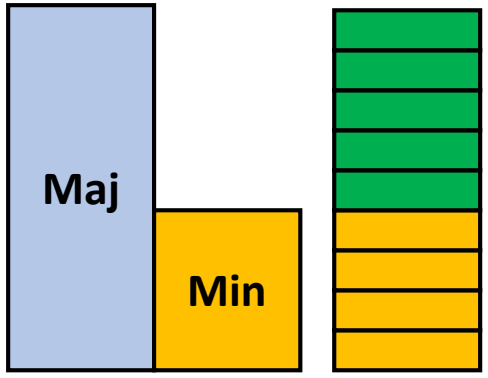


Synthetic data for balanced dataset

Synthetic data generator: LSTM, SeqGAN, VAE
Baselines: Real data, SMOTE
Classifiers: Linear SVM, NN
Evaluation metrics: F1, G-Mean
Datasets: Quora, Amazon, Yelp

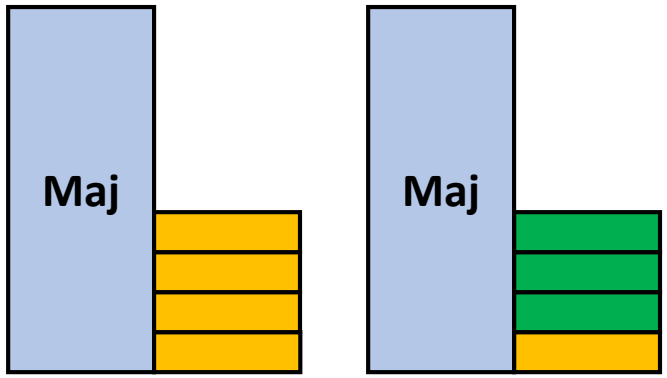
Experiments

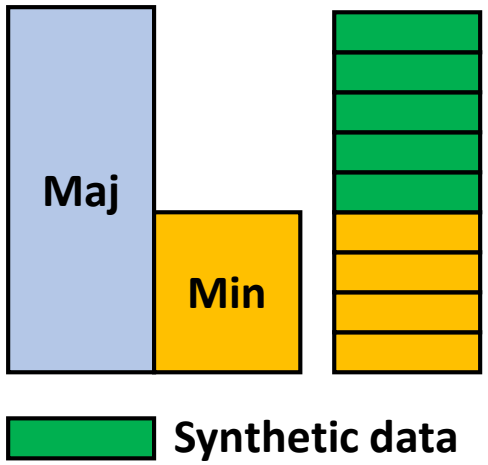


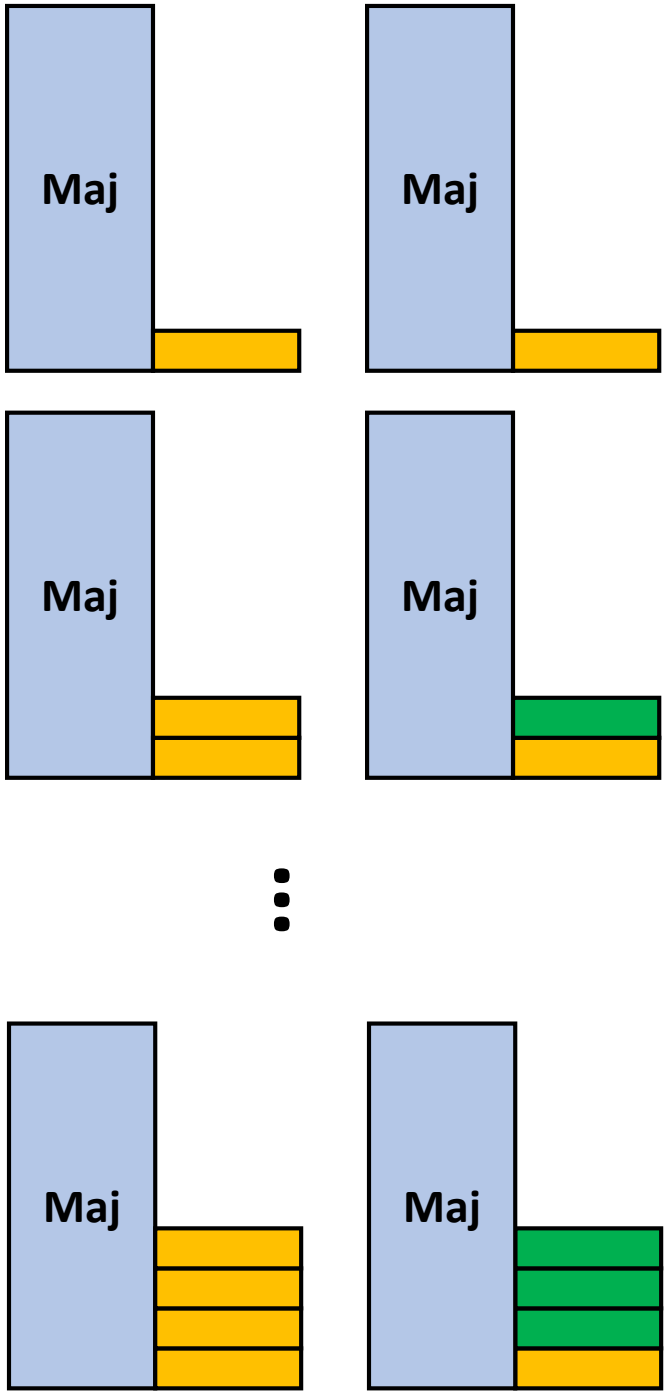


Synthetic data

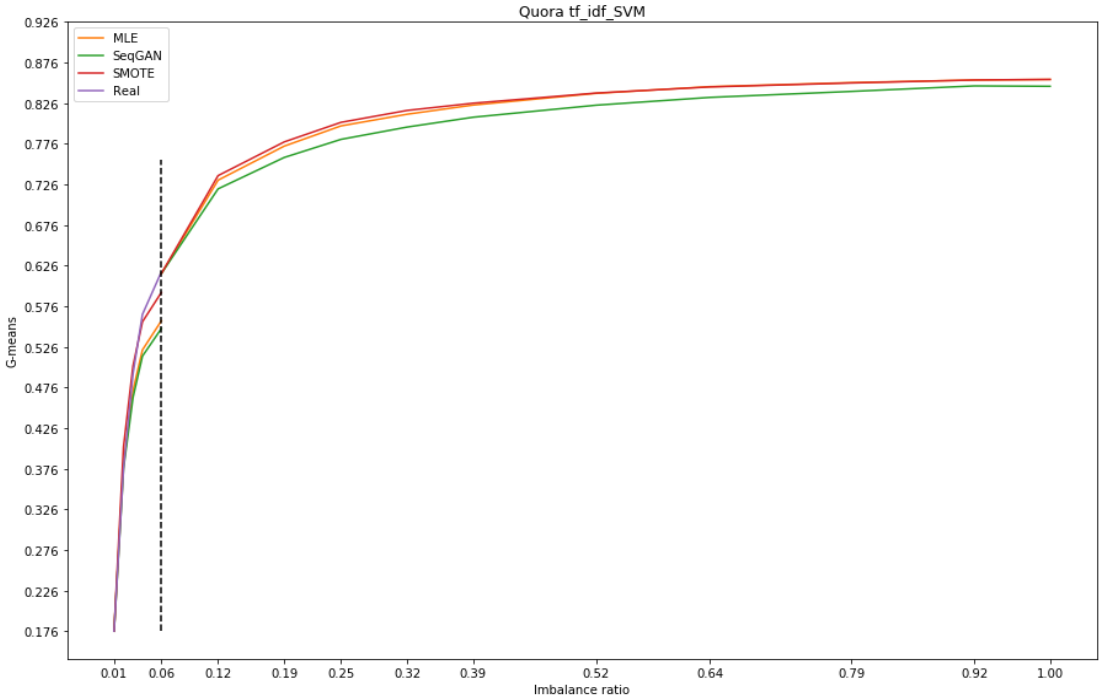
Experiments

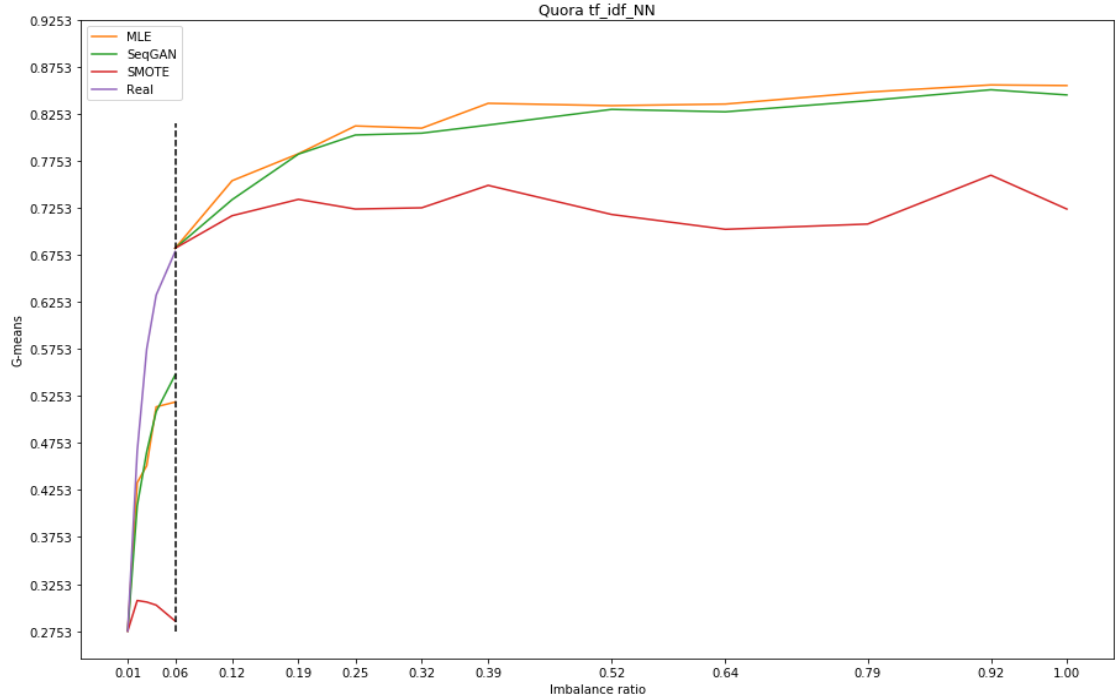




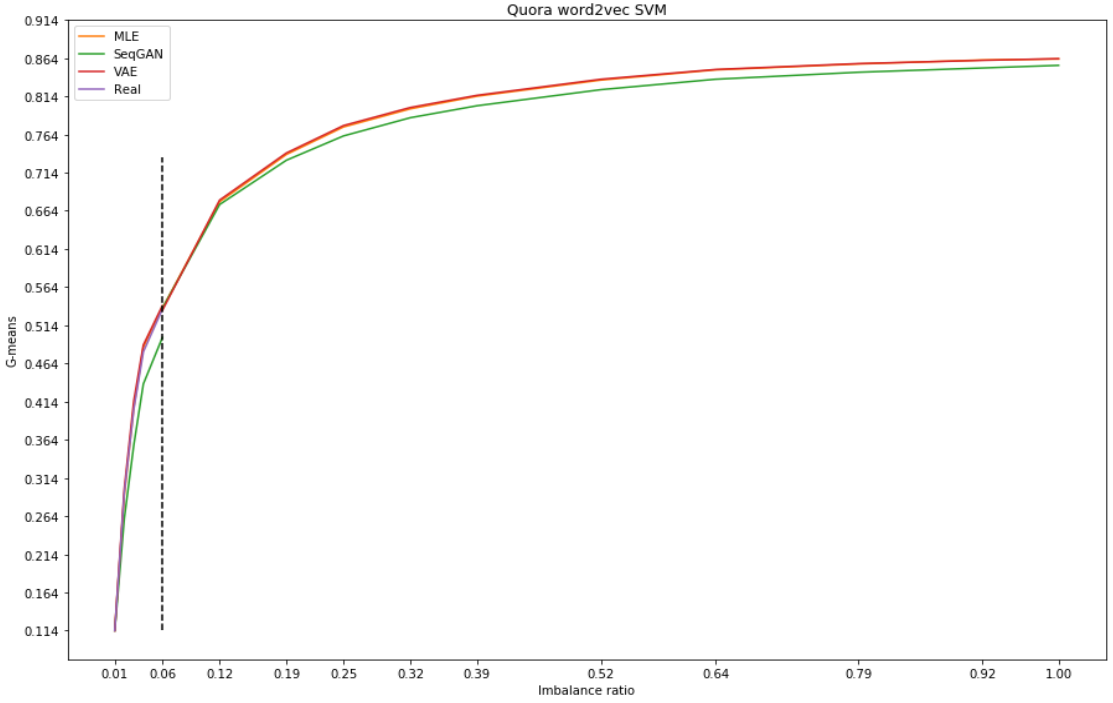


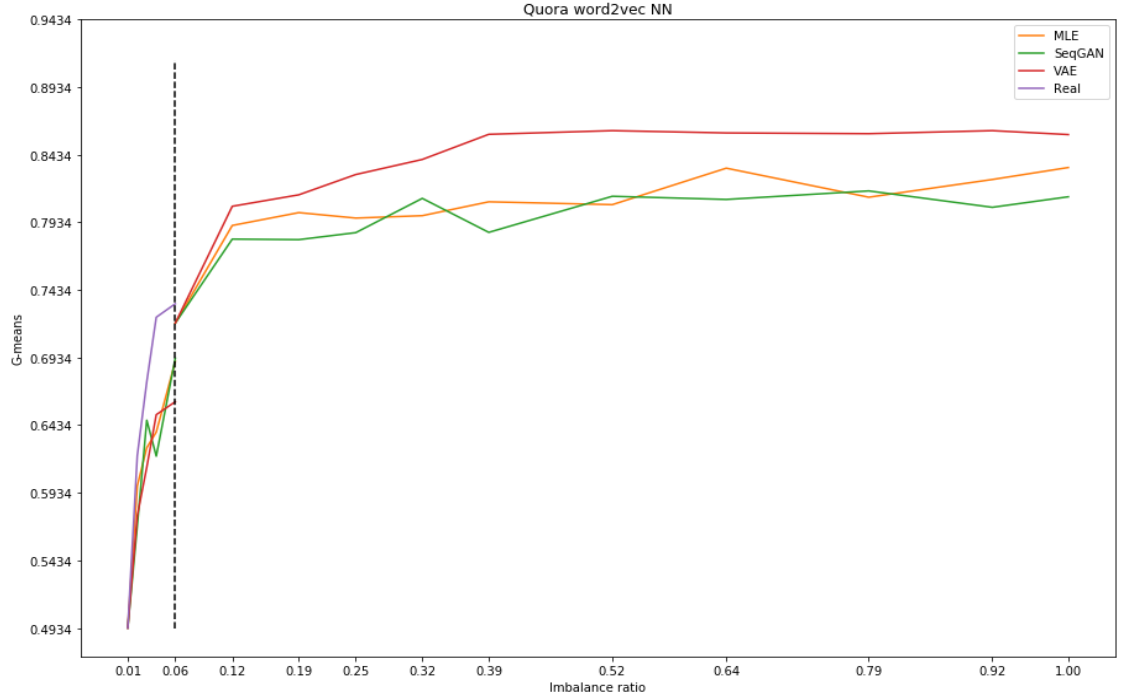
Results

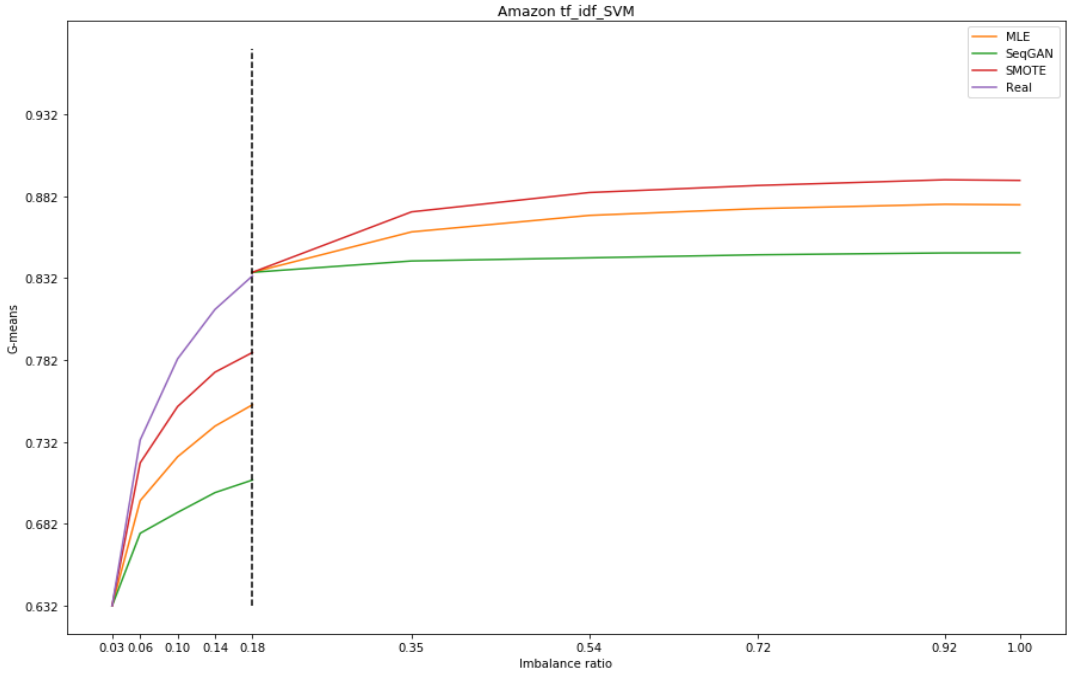

Quora tf_idf_SVM
MLE
SeqGAN
SMOTE
Real
G-means
Imbalance ratio

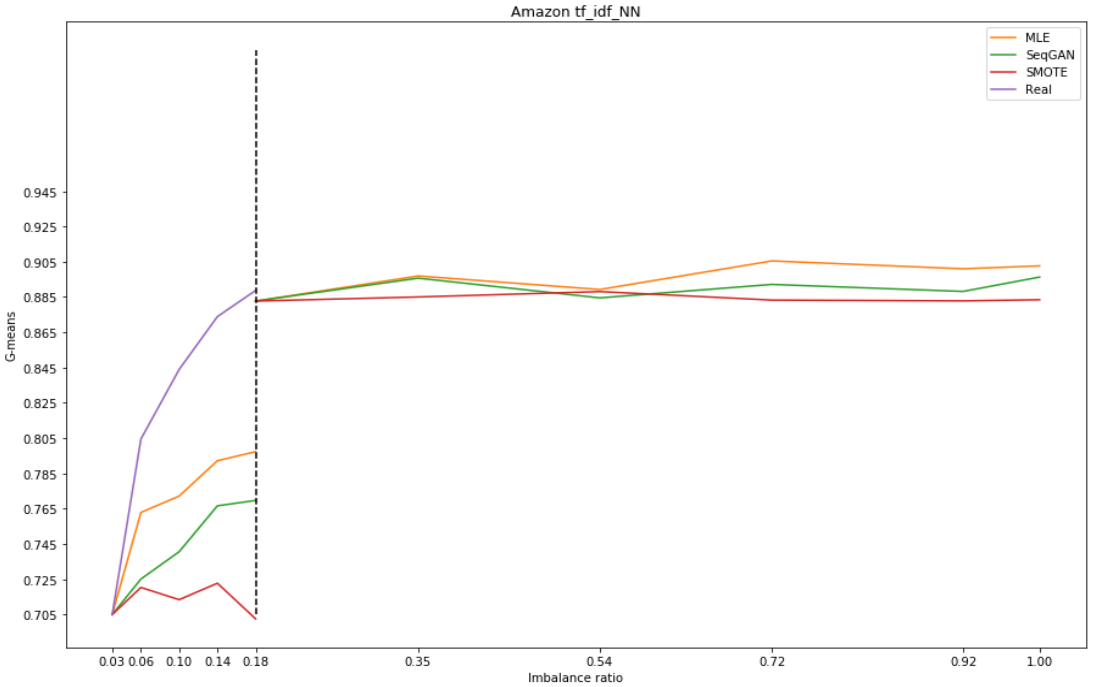

Quora tf_idf_NN
MLE
SeqGAN
SMOTE
Real
G-means
Imbalance ratio

Quora

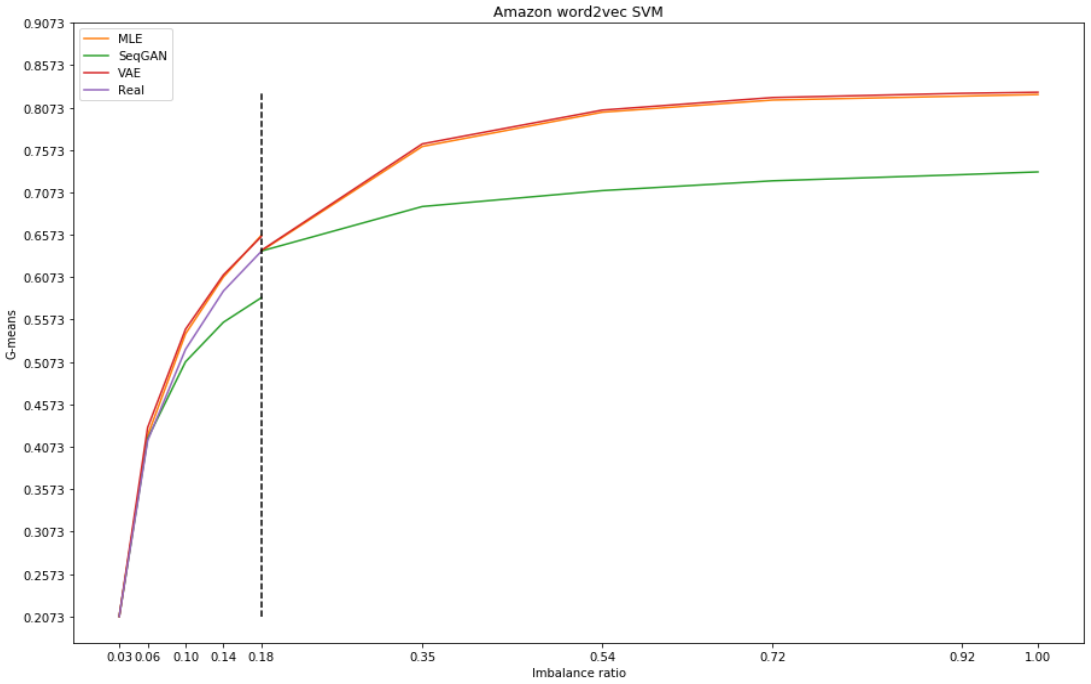

Quora word2vec SVM
MLE
SeqGAN
VAE
Real
G-means
Imbalance ratio


Quora word2vec NN
MLE
SeqGAN
VAE
Real
G-means
Imbalance ratio

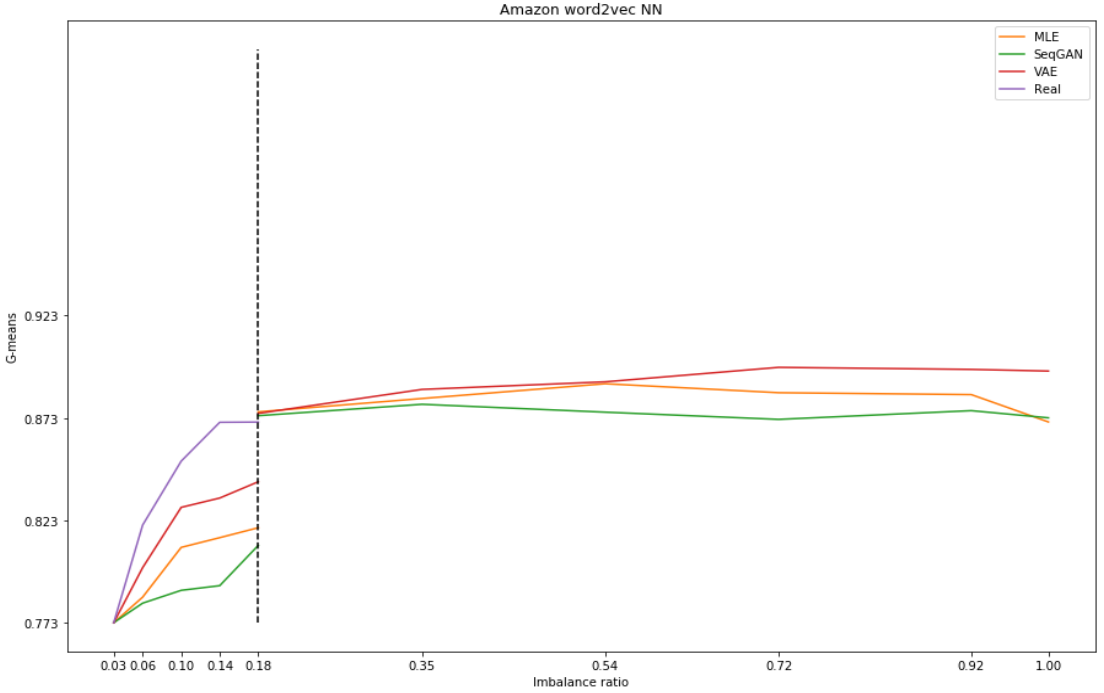


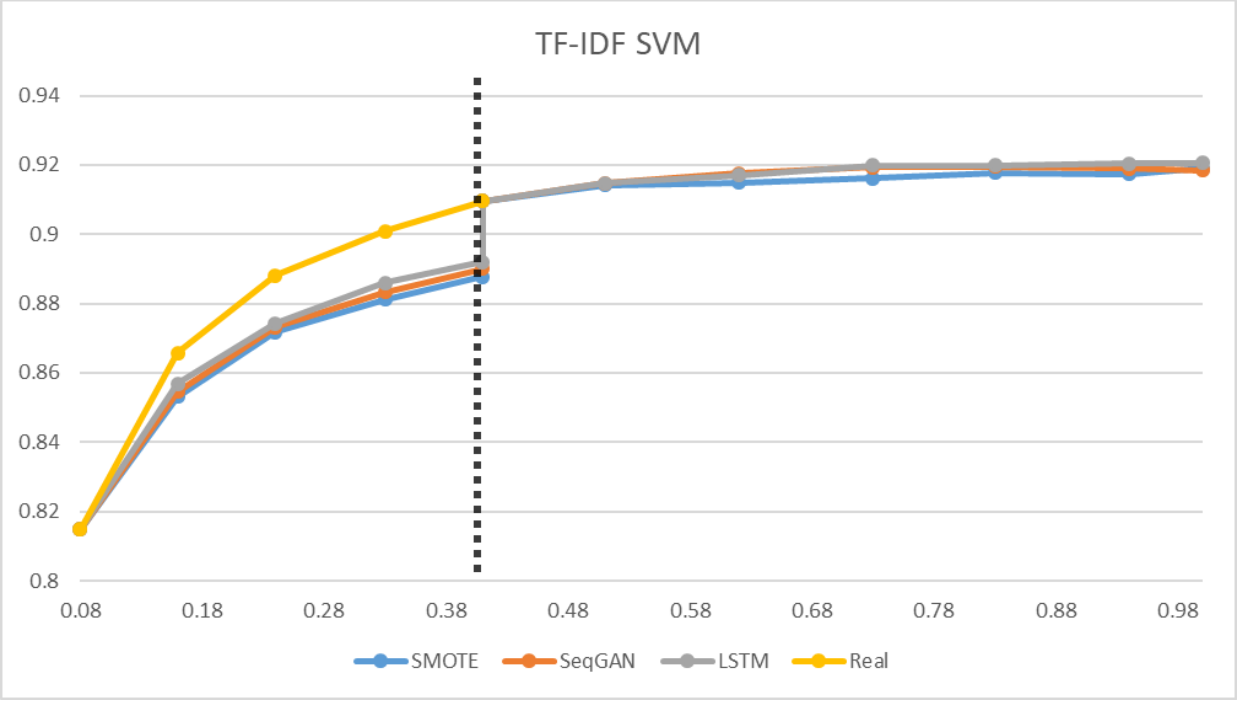


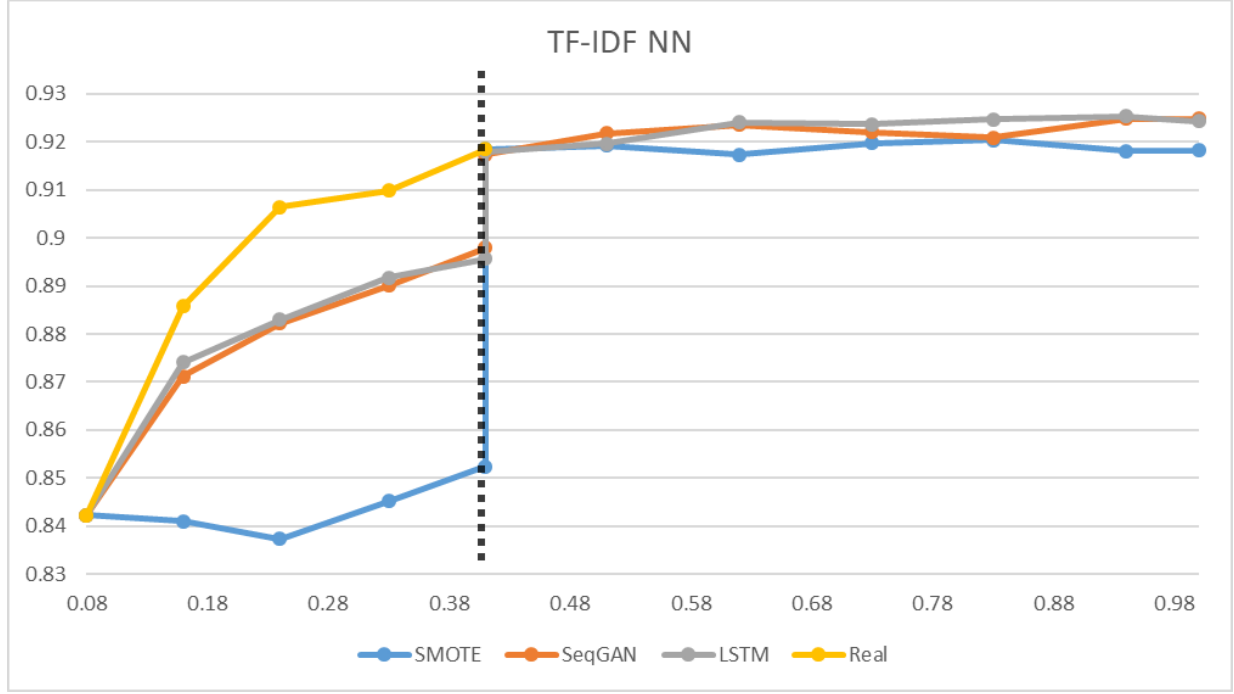
Amazon



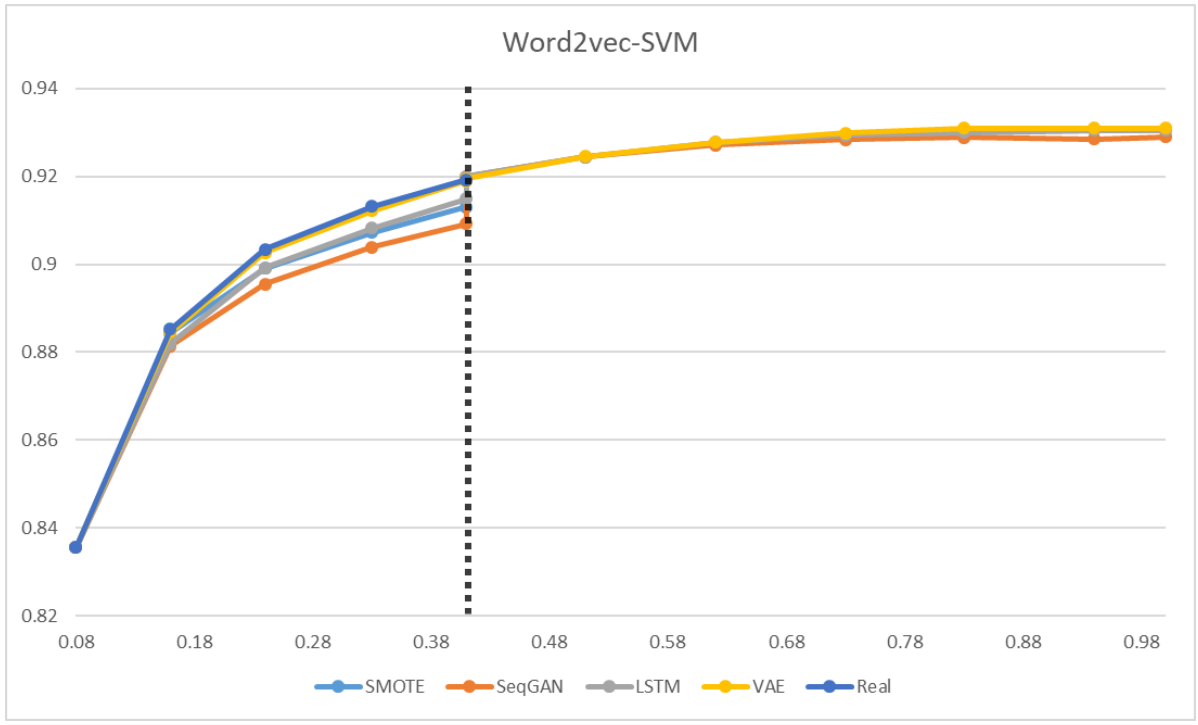
Note: SMOTE is missing for W2V

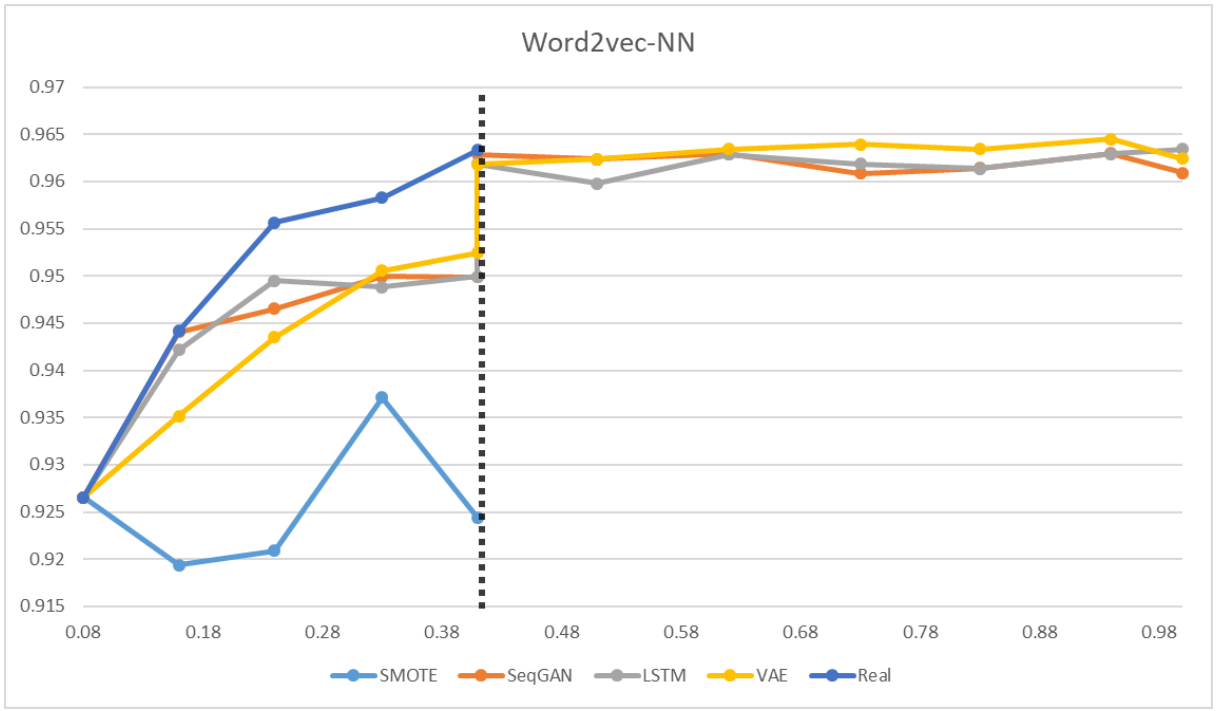



TF-IDF SVM
0.94
0.92
0.9
0.88
0.86
0.84
0.82
0.8
0.08
0.18
0.28
0.38
0.48
0.58
0.68
0.78
0.88
0.98
SMOTE
SeqGAN
LSTM
Real


TF-IDF NN
0.93
0.92
0.91
0.9
0.89
0.88
0.87
0.86
0.85
0.84
0.83
0.08
0.18
0.28
0.38
0.48
0.58
0.68
0.78
0.88
0.98
SMOTE
SeqGAN
LSTM
Real

Yelp


Word2vec-SVM
0.94
0.92
0.9
0.88
0.86
0.84
0.82
0.08
0.18
0.28
0.38
0.48
0.58
0.68
0.78
0.88
0.98
SMOTE
SeqGAN
LSTM
VAE
Real


Word2vec-NN
0.97
0.965
0.96
0.955
0.95
0.945
0.94
0.935
0.93
0.925
0.92
0.915
0.08
0.18
0.28
0.38
0.48
0.58
0.68
0.78
0.88
0.98
SMOTE
SeqGAN
LSTM
VAE
Real

Summary of results

- W2V + NN > TF-IDF + SVM
- In many cases, synthetic data was almost as good as real data for the downstream task.
- Generative methods were superior than SMOTE.

Thoughts?

- Why real data gave us better results?
  - Vocabulary coverage is better.
  - Quality of synthetic data is not monitored.
  - What defines quality of a piece of text??
- Classification performance gained marginally improvement after the dashed line... Why?