# Gathering and Classification of Sports Injury Data

Presented By-
Jhagrut Lalwani,
Joe Datz

# Problem Statement

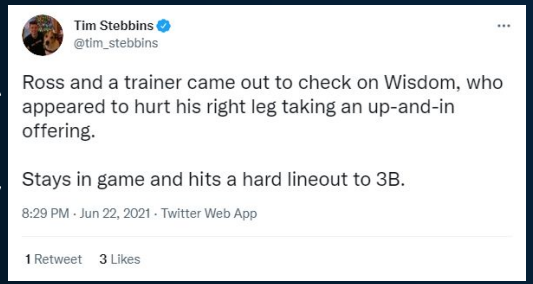
We want to create a function:

$$H(X) \rightarrow Y$$

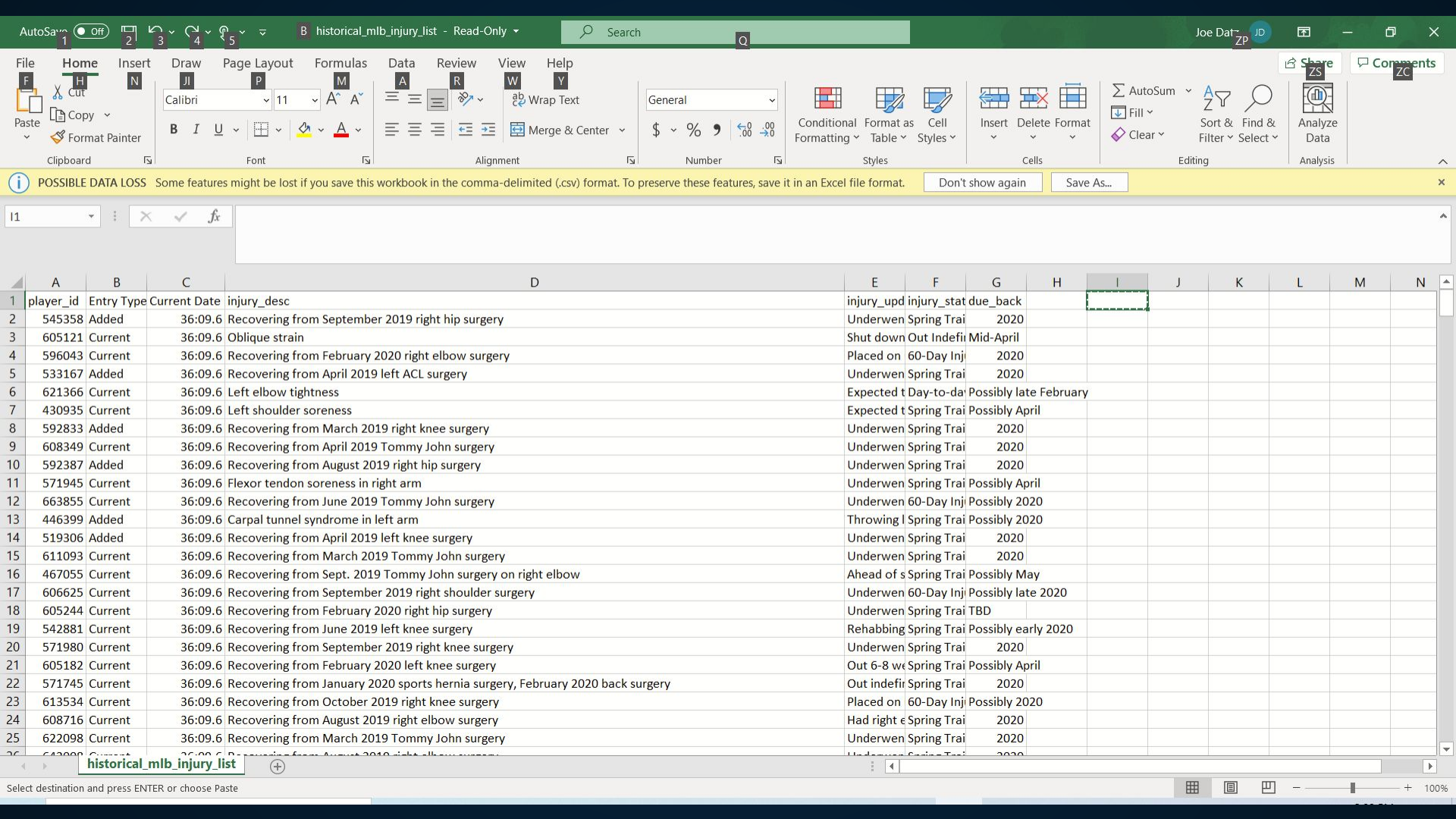
Where X is our dataset (text) and Y is our target variable (1 for injury report, 0 for not an injury report).

H( 

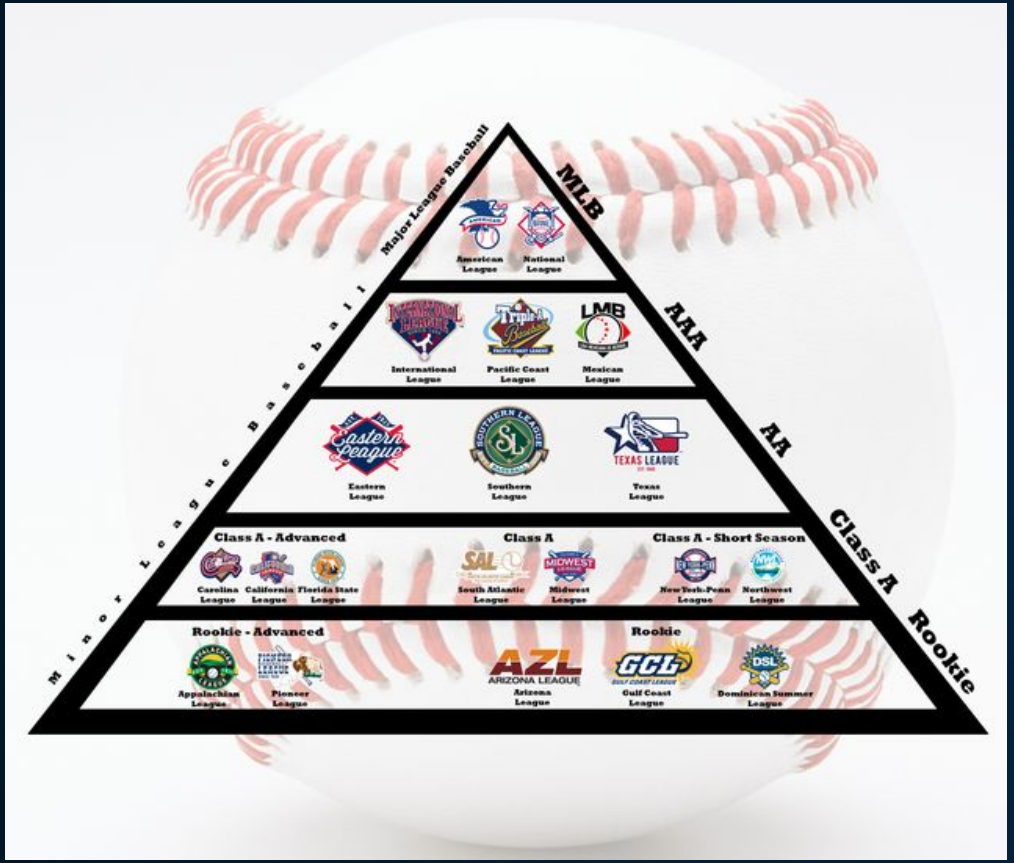
) → 0

H( 

) → 1



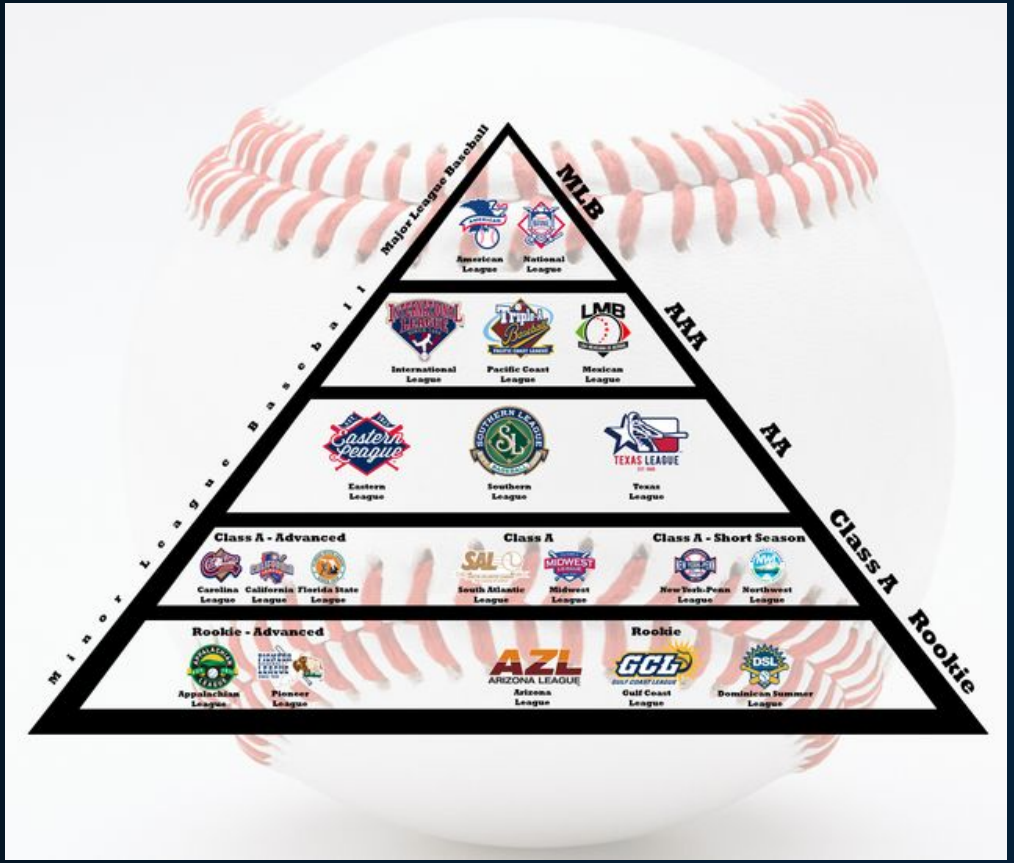
| player_id | Entry Type | Current Date | injury_desc | injury_upd | injury_stat | due_back |
|---|---|---|---|---|---|---|
| 545358 | Added | 36:09.6 | Recovering from September 2019 right hip surgery | Underwen | Spring Trai | 2020 |
| 605121 | Current | 36:09.6 | Oblique strain | Shut down | Out Indefi | Mid-April |
| 596043 | Current | 36:09.6 | Recovering from February 2020 right elbow surgery | Placed on | 60-Day Inj | 2020 |
| 533167 | Added | 36:09.6 | Recovering from April 2019 left ACL surgery | Underwen | Spring Trai | 2020 |
| 621366 | Current | 36:09.6 | Left elbow tightness | Expected t | Day-to-da | Possibly late February |
| 430935 | Current | 36:09.6 | Left shoulder soreness | Expected t | Spring Trai | Possibly April |
| 592833 | Added | 36:09.6 | Recovering from March 2019 right knee surgery | Underwen | Spring Trai | 2020 |
| 608349 | Current | 36:09.6 | Recovering from April 2019 Tommy John surgery | Underwen | Spring Trai | 2020 |
| 592387 | Added | 36:09.6 | Recovering from August 2019 right hip surgery | Underwen | Spring Trai | 2020 |
| 571945 | Current | 36:09.6 | Flexor tendon soreness in right arm | Underwen | Spring Trai | Possibly April |
| 663855 | Current | 36:09.6 | Recovering from June 2019 Tommy John surgery | Underwen | 60-Day Inj | Possibly 2020 |
| 446399 | Added | 36:09.6 | Carpal tunnel syndrome in left arm | Throwing l | Spring Trai | Possibly 2020 |
| 519306 | Added | 36:09.6 | Recovering from April 2019 left knee surgery | Underwen | Spring Trai | 2020 |
| 611093 | Current | 36:09.6 | Recovering from March 2019 Tommy John surgery | Underwen | Spring Trai | 2020 |
| 467055 | Current | 36:09.6 | Recovering from Sept. 2019 Tommy John surgery on right elbow | Ahead of s | Spring Trai | Possibly May |
| 606625 | Current | 36:09.6 | Recovering from September 2019 right shoulder surgery | Underwen | 60-Day Inj | Possibly late 2020 |
| 605244 | Current | 36:09.6 | Recovering from February 2020 right hip surgery | Underwen | Spring Trai | TBD |
| 542881 | Current | 36:09.6 | Recovering from June 2019 left knee surgery | Rehabbing | Spring Trai | Possibly early 2020 |
| 571980 | Current | 36:09.6 | Recovering from September 2019 right knee surgery | Underwen | Spring Trai | 2020 |
| 605182 | Current | 36:09.6 | Recovering from February 2020 left knee surgery | Out 6-8 we | Spring Trai | Possibly April |
| 571745 | Current | 36:09.6 | Recovering from January 2020 sports hernia surgery, February 2020 back surgery | Out indefi | Spring Trai | 2020 |
| 613534 | Current | 36:09.6 | Recovering from October 2019 right knee surgery | Placed on | 60-Day Inj | Possibly 2020 |
| 608716 | Current | 36:09.6 | Recovering from August 2019 right elbow surgery | Had right e | Spring Trai | 2020 |
| 622098 | Current | 36:09.6 | Recovering from March 2019 Tommy John surgery | Underwen | Spring Trai | 2020 |




# Problem Statement

It is only at the top of this pyramid where data on injury information is publicly available. At all other levels, MLB teams have their data private.

We'd like to use twitter so that we may greatly expand the usable injury data for analysis.




# Problem Statement

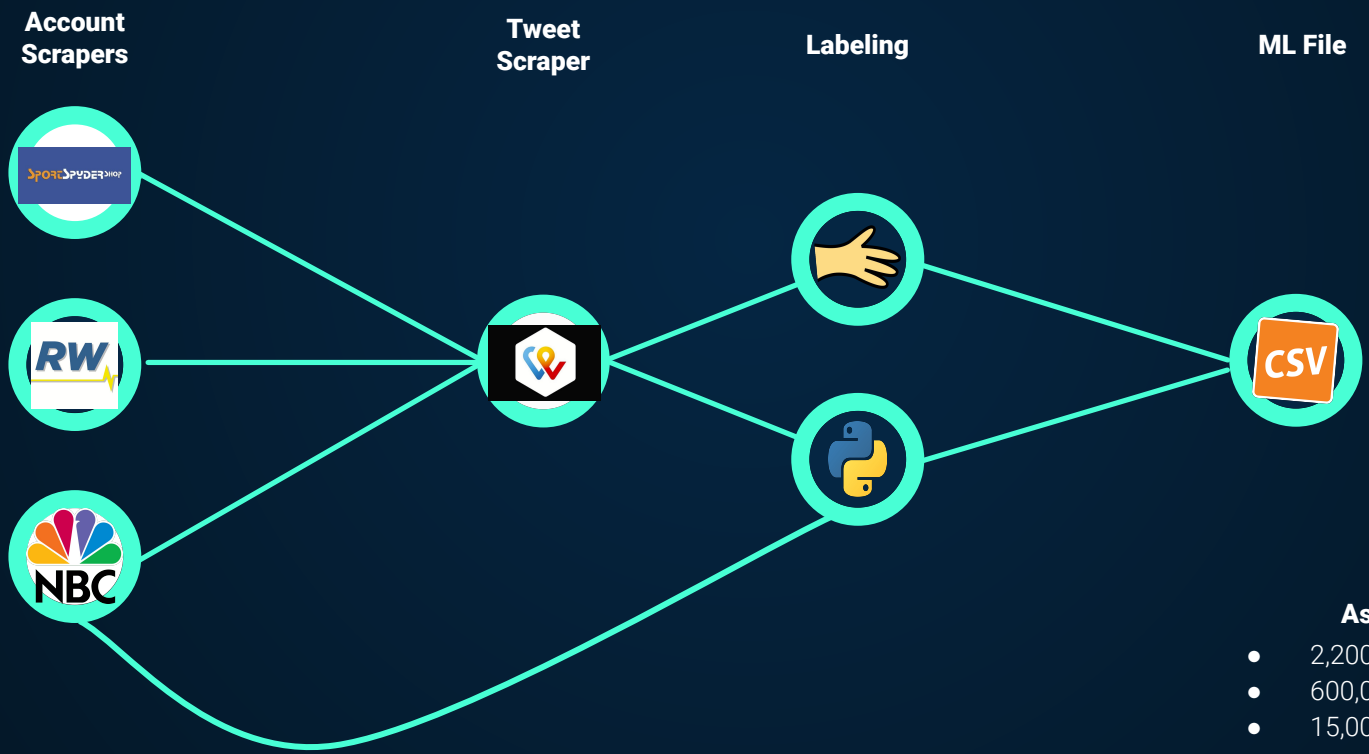
It is only at the top of this pyramid where data on injury information is publicly available. At all other levels, MLB teams have their data private.

We'd like to use twitter so that we may greatly expand the usable injury data for analysis.

Other important considerations:

- Utilize both labeled and unlabeled data. **(Semi-Supervised / Multi-View Learning)**
- Each cycle of gathering tweets and evaluating them preferably stays under 24 hours.

# WEB SCRAPING PIPELINE




**As Of This Writing:**

- 2,200 Twitter Accounts
- 600,000 Tweets
- 15,000 Unique Labels

# Results with Standard Models

Each model had:

- 8-fold Cross Validation
- Stratified Sampling over same dataset
- Stop word removal and word stemming

| Model | Data Type | Results |
|---|---|---|
| kNN | TF-IDF | [[4060 92]<br>[ 103 329]] |
| Bernoulli NB | Boolean | [[3972 180]<br>[ 43 389]] |
| Multinom. NB | Count | [[3896 256]<br>[ 34 398]] |
| Logistic Regression | TF-IDF | [[3997 155]<br>[ 47 385]] |
| Random Forest | Boolean | [[3929 223]<br>[ 68 364]] |
| Random Forest | TF-IDF | [[3939 213]<br>[ 72 360]] |
| SVM | TF-IDF | [[4092 60]<br>[ 87 345]] |

# Results with Standard Models

Each model had:

- 8-fold Cross Validation
- Stratified Sampling over same dataset
- Stop word removal and word stemming

A variant of PCA called IPCA was also experimented with to see how a Dense NN would work on the data. However, it failed to apply to our use case.

| Model | Data Type | Results |
|---|---|---|
| kNN | TF-IDF | [[4060 92]<br>[ 103 329]] |
| Bernoulli NB | Boolean | [[3972 180]<br>[ 43 389]] |
| Multinom. NB | Count | [[3896 256]<br>[ 34 398]] |
| Logistic Regression | TF-IDF | [[3997 155]<br>[ 47 385]] |
| Random Forest | Boolean | [[3929 223]<br>[ 68 364]] |
| Random Forest | TF-IDF | [[3939 213]<br>[ 72 360]] |
| SVM | TF-IDF | [[4092 60]<br>[ 87 345]] |

# Results with Standard Models

Best: **Multinomial Naive Bayes**

Reason: Lowest false negative rate (34)

In our use case, we would check the output of all our tweets the model output as a 1 (true positives and false positives) before adding to our injury record.

The False Negatives would be lost in the tens of thousands of 0 outputs, so we'd like to avoid this.

| Model | Data Type | Results |
|---|---|---|
| kNN | TF-IDF | [[4060 92]<br>[ 103 329]] |
| Bernoulli NB | Boolean | [[3972 180]<br>[ 43 389]] |
| Multinom. NB | Count | [[3896 256]<br>[ 34 398]] |
| Logistic Regression | TF-IDF | [[3997 155]<br>[ 47 385]] |
| Random Forest | Boolean | [[3929 223]<br>[ 68 364]] |
| Random Forest | TF-IDF | [[3939 213]<br>[ 72 360]] |
| SVM | TF-IDF | [[4092 60]<br>[ 87 345]] |

# Results with Neural Networks

Each model had:

- Random samples of 6000 non-injury tweets to prevent imbalance issues
- Stop word removal and word stemming
- Stratified train-test split on injury_report.

| Model | Epochs | Results | Sens. |
|---|---|---|---|
| XLNet | 5 | [[1485 32]<br>[ 28 316]] | 0.9212 |
| DistilBERT | 3 | [[1491 26]<br>[ 30 312]] | 0.9122 |
| RoBERTa | 3 | [[1472 48]<br>[ 22 225]] | 0.9109 |
| XLM-RoBERTA | 5 | [[1475 46]<br>[ 69 270]] | 0.7964 |

```
DistilBertConfig {
  "_name_or_path": "distilbert-base-uncased",
  "activation": "gelu",
  "architectures": [
    "DistilBertForMaskedLM"
  ],
  "attention_dropout": 0.1,
  "dim": 768,
  "dropout": 0.1,
  "hidden_dim": 3072,
  "initializer_range": 0.02,
  "max_position_embeddings": 512,
  "model_type": "distilbert",
  "n_heads": 12,
  "n_layers": 6,
  "pad_token_id": 0,
  "qa_dropout": 0.1,
  "seq_classif_dropout": 0.2,
  "sinusoidal_pos_embds": false,
  "tie_weights_": true,
  "transformers_version": "4.9.2",
  "vocab_size": 30522
}
```

```
XLNetConfig {
  "_name_or_path": "xlnet-base-cased",
  "architectures": [
    "XLNetLMHeadModel"
  ],
  "attn_type": "bi",
  "bi_data": false,
  "bos_token_id": 1,
  "clamp_len": -1,
  "d_head": 64,
  "d_inner": 3072,
  "d_model": 768,
  "dropout": 0.1,
  "end_n_top": 5,
  "eos_token_id": 2,
  "ff_activation": "gelu",
  "initializer_range": 0.02,
  "layer_norm_eps": 1e-12,
  "mem_len": null,
  "model_type": "xlnet",
  "n_head": 12,
  "n_layer": 12,
  "pad_token_id": 5,
  "reuse_len": null,
  "same_length": false,
  "start_n_top": 5,
  "summary_activation": "tanh",
  "summary_last_dropout": 0.1,
  "summary_type": "last",
  "summary_use_proj": true,
  "task_specific_params": {
    "text-generation": {
      "do_sample": true,
      "max_length": 250
    }
  },
  "transformers_version": "4.9.2",
  "untie_r": true,
  "use_mems_eval": true,
  "use_mems_train": false,
  "vocab_size": 32000
}
```

```
RobertaConfig {
  "_name_or_path": "roberta-base",
  "architectures": [
    "RobertaForMaskedLM"
  ],
  "attention_probs_dropout_prob": 0.1,
  "bos_token_id": 0,
  "eos_token_id": 2,
  "gradient_checkpointing": false,
  "hidden_act": "gelu",
  "hidden_dropout_prob": 0.1,
  "hidden_size": 768,
  "initializer_range": 0.02,
  "intermediate_size": 3072,
  "layer_norm_eps": 1e-05,
  "max_position_embeddings": 514,
  "model_type": "roberta",
  "num_attention_heads": 12,
  "num_hidden_layers": 12,
  "pad_token_id": 1,
  "position_embedding_type": "absolute",
  "transformers_version": "4.9.2",
  "type_vocab_size": 1,
  "use_cache": true,
  "vocab_size": 50265
}
```

```
XLMRobertaConfig {
  "_name_or_path": "xlm-roberta-base",
  "architectures": [
    "XLMRobertaForMaskedLM"
  ],
  "attention_probs_dropout_prob": 0.1,
  "bos_token_id": 0,
  "eos_token_id": 2,
  "gradient_checkpointing": false,
  "hidden_act": "gelu",
  "hidden_dropout_prob": 0.1,
  "hidden_size": 768,
  "initializer_range": 0.02,
  "intermediate_size": 3072,
  "layer_norm_eps": 1e-05,
  "max_position_embeddings": 514,
  "model_type": "xlm-roberta",
  "num_attention_heads": 12,
  "num_hidden_layers": 12,
  "output_past": true,
  "pad_token_id": 1,
  "position_embedding_type": "absolute",
  "transformers_version": "4.9.2",
  "type_vocab_size": 1,
  "use_cache": true,
  "vocab_size": 250002
}
```

# Results with Neural Networks

DistilBERT: faster inference speed with slightly poor prediction metrics

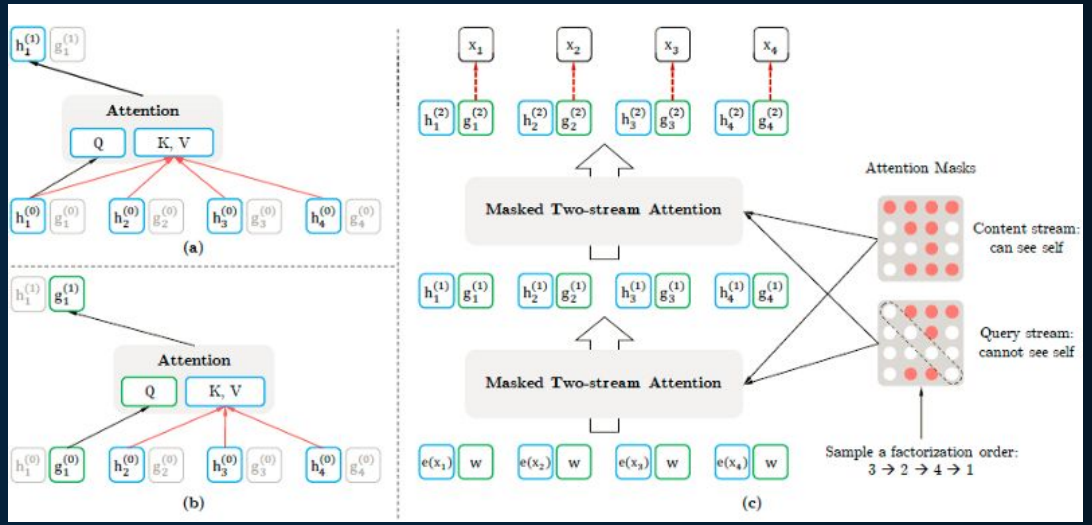
XLNet: permutation based training handles dependencies well

| Models | Sensitivity | Accuracy | Precision | mcc | Selectivity | F1 score |
|---|---|---|---|---|---|---|
| **XL-Net** | 0.9186 | 0.9678 | 0.979 | 0.8935 | 0.979 | 0.9133 |
| RoBERTa | 0.911 | 0.9604 | 0.8242 | 0.8437 | 0.9684 | 0.8564 |
| Distil-BERT | 0.9123 | 0.9699 | 0.923 | 0.8992 | 0.9829 | 0.9176 |
| LSTM | 0.8863 | 0.9731 | 0.965 | 0.909 | 0.9927 | 0.924 |
| GRU | 0.8338 | 0.9613 | 0.9502 | 0.8676 | 0.9901 | 0.8882 |
| XLM-RoBERTa | 0.7965 | 0.9382 | 0.8544 | 0.7877 | 0.9698 | 0.8244 |

# Results with Neural Networks

DistilBERT: faster inference speed with slightly poor prediction metrics

XLNet: Uses a modified language model training objective which learns conditional distributions for all permutations of tokens in a sequence. Permutation based training handles dependencies well



# Thank you!