Introducing gobbli

Deep learning with text doesn’t have to be scary

Jason Nance
Data Scientist
RTI International
Using a State of the Art Model for Text Classification

Line Count: 0 101 108 142 167 187
Best Scores on DBpedia Classification Benchmark*


* [https://paperswithcode.com/sota/text-classification-on-dbpedia](https://paperswithcode.com/sota/text-classification-on-dbpedia)
The Problem: Hard-Coding for Benchmark Problems

<table>
<thead>
<tr>
<th>Rank Name</th>
<th>Model</th>
<th>URL Score</th>
<th>CoLA SST-2</th>
<th>MRPC</th>
<th>STS-B</th>
<th>QQP MNLI-m</th>
<th>MNLI-mm</th>
<th>QNLI</th>
<th>RTE</th>
<th>WNL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Microsoft D365 AI &amp; UMD Adv-RoBERTa (ensemble)</td>
<td>88.8</td>
<td>88.0</td>
<td>96.8</td>
<td>93.1/90.8</td>
<td>92.4/92.2</td>
<td>74.6/90.3</td>
<td>91.1</td>
<td>90.7</td>
<td>98.8</td>
</tr>
<tr>
<td>2</td>
<td>Facebook AI RoBERTa</td>
<td>88.5</td>
<td>67.8</td>
<td>96.7</td>
<td>92.3/89.8</td>
<td>92.2/91.9</td>
<td>74.3/90.2</td>
<td>90.8</td>
<td>90.2</td>
<td>99.9</td>
</tr>
<tr>
<td>3</td>
<td>XLNet Team XLNet-Large (ensemble)</td>
<td>88.4</td>
<td>67.8</td>
<td>96.8</td>
<td>93.0/90.7</td>
<td>91.6/91.1</td>
<td>74.2/90.3</td>
<td>90.2</td>
<td>89.8</td>
<td>98.6</td>
</tr>
<tr>
<td>4</td>
<td>Microsoft D365 AI &amp; MSR AI MT-DNN-ensemble</td>
<td>87.6</td>
<td>68.4</td>
<td>96.5</td>
<td>92.7/90.3</td>
<td>91.9/90.7</td>
<td>73.7/89.9</td>
<td>87.9</td>
<td>87.4</td>
<td>95.0</td>
</tr>
<tr>
<td>5</td>
<td>GLUE Human Baselines GLUE Human Baselines</td>
<td>87.1</td>
<td>66.4</td>
<td>97.8</td>
<td>86.3/80.8</td>
<td>92.7/92.6</td>
<td>59.5/80.4</td>
<td>92.0</td>
<td>92.8</td>
<td>91.2</td>
</tr>
<tr>
<td>6</td>
<td>王栋 AI ALICE large ensemble (Alibaba DAMO NLP)</td>
<td>87.0</td>
<td>89.2</td>
<td>95.2</td>
<td>92.6/90.2</td>
<td>91.1/90.6</td>
<td>74.4/90.7</td>
<td>88.2</td>
<td>87.9</td>
<td>95.7</td>
</tr>
<tr>
<td>7</td>
<td>Stanford Hazy Research Snorkel MeTaL</td>
<td>83.2</td>
<td>63.8</td>
<td>96.2</td>
<td>91.5/88.5</td>
<td>90.1/89.7</td>
<td>73.1/89.9</td>
<td>87.6</td>
<td>87.2</td>
<td>93.9</td>
</tr>
<tr>
<td>8</td>
<td>XLM Systems XLM (English only)</td>
<td>83.1</td>
<td>82.9</td>
<td>95.6</td>
<td>90.7/87.1</td>
<td>88.8/88.2</td>
<td>73.2/89.8</td>
<td>89.1</td>
<td>88.5</td>
<td>94.0</td>
</tr>
<tr>
<td>9</td>
<td>Zhuosheng Zhang SemBERT</td>
<td>82.9</td>
<td>62.3</td>
<td>94.6</td>
<td>91.2/88.3</td>
<td>87.6/86.7</td>
<td>72.8/89.8</td>
<td>87.6</td>
<td>86.3</td>
<td>94.6</td>
</tr>
<tr>
<td>10</td>
<td>Danqi Chen SpanBERT (single-task training)</td>
<td>82.8</td>
<td>84.3</td>
<td>94.8</td>
<td>90.9/87.9</td>
<td>89.9/89.1</td>
<td>71.9/89.5</td>
<td>88.1</td>
<td>87.7</td>
<td>94.3</td>
</tr>
<tr>
<td>11</td>
<td>Kevin Clark BERT + BAM</td>
<td>82.3</td>
<td>61.5</td>
<td>95.2</td>
<td>91.3/88.3</td>
<td>88.6/87.9</td>
<td>72.5/89.7</td>
<td>86.6</td>
<td>85.8</td>
<td>93.1</td>
</tr>
<tr>
<td>12</td>
<td>Nitish Shirish Keskar Span-Extractive BERT on STILTs</td>
<td>82.3</td>
<td>83.2</td>
<td>94.5</td>
<td>90.6/87.6</td>
<td>89.4/89.2</td>
<td>72.2/89.4</td>
<td>86.5</td>
<td>85.8</td>
<td>92.5</td>
</tr>
<tr>
<td>13</td>
<td>Jason Phang BERT on STILTs</td>
<td>82.0</td>
<td>62.1</td>
<td>94.3</td>
<td>90.2/86.6</td>
<td>88.7/88.3</td>
<td>71.9/89.4</td>
<td>86.4</td>
<td>85.6</td>
<td>92.7</td>
</tr>
<tr>
<td>14</td>
<td>廖亿 RGLM-base (Huawei Noah's Ark Lab)</td>
<td>81.0</td>
<td>55.1</td>
<td>94.2</td>
<td>90.7/87.7</td>
<td>89.5/88.7</td>
<td>72.2/89.4</td>
<td>85.6</td>
<td>85.1</td>
<td>92.1</td>
</tr>
<tr>
<td>15</td>
<td>Jacob Devlin BERT: 24-layers, 16-heads, 1024-hidden</td>
<td>80.5</td>
<td>60.5</td>
<td>94.9</td>
<td>89.3/85.4</td>
<td>87.6/86.5</td>
<td>72.1/89.3</td>
<td>86.7</td>
<td>85.9</td>
<td>92.7</td>
</tr>
</tbody>
</table>

https://gluebenchmark.com/leaderboard/
gobbli: A Uniform Interface for Text Deep Learning Models
gobbli: Library Design

Users → Experiment Results → Experiment API

Task API

Training → Prediction → Embeddings

Docker

BERT → fastText → ...

gobbli: Additional Features

Data Augmentation

- Word2Vec
- WordNet
- BERT Masked Language Model
gobbli: Additional Features

Document Windowing

BERT

P(pos) = 0.1

P(pos) = 0.1
P(pos) = 0.2
P(pos) = 0.9
P(pos) = 0.8

P(pos) = 0.5
gobbli: Benefits and Drawbacks

+ Cross-platform
+ Abstracts dependency management

- Latency/overhead

+ Parallel/distributed training

- Experiment API only
Example Experiment Results
Example Experiment Results: Metrics

**Metrics:**
---

Weighted F1 Score: 0.8806791429898766
Weighted Precision Score: 0.8806909370983464
Weighted Recall Score: 0.88068
Accuracy: 0.88068

**Classification Report:**
------------------------

<table>
<thead>
<tr>
<th></th>
<th>precision</th>
<th>recall</th>
<th>f1-score</th>
<th>support</th>
</tr>
</thead>
<tbody>
<tr>
<td>neg</td>
<td>0.88</td>
<td>0.88</td>
<td>0.88</td>
<td>12500</td>
</tr>
<tr>
<td>pos</td>
<td>0.88</td>
<td>0.88</td>
<td>0.88</td>
<td>12500</td>
</tr>
<tr>
<td>accuracy</td>
<td></td>
<td></td>
<td>0.88</td>
<td>25000</td>
</tr>
<tr>
<td>macro avg</td>
<td>0.88</td>
<td>0.88</td>
<td>0.88</td>
<td>25000</td>
</tr>
<tr>
<td>weighted avg</td>
<td>0.88</td>
<td>0.88</td>
<td>0.88</td>
<td>25000</td>
</tr>
</tbody>
</table>
Example Experiment Results: Plot
**True Class:** comp.os.ms-windows.misc

**Predicted Class:** sci.med (Probability: 0.97)

**Text:**
“My wife is a physiotherapist and she is looking for some cliparts of skeleton and male/female body. We're currently using Windows Draw which can import all kind of graphic formats. Therefore, anything will do. Please advise ...”
Initial open source release on GitHub
- https://github.com/RTIInternational/gobbli/
- Models implemented: BERT, MT-DNN, USE, fastText, pytorch_transformers (XLNet, XLM, BERT, RoBERTa)

Next steps:
- Support multilabel classification
- Helper module for downstream tasks using embeddings
- Helper module for exploratory descriptives
- Other bug fixes/enhancements requested by the community
Jason Nance
Data Scientist
RTI International
jnance@rti.org