Motivation & Objective

**Motivation:**
- News article classification can be subjective
- Alleviate the need to manually categorize news articles

**Goals:**
- Classify large corpus of news articles into seven categories
- Capture the degree to which news articles can be accurately and consistently categorized

---

Data

**News Article Corpus:**
- TagMyNews Dataset ($A^3$ Lab, University of Pisa)
- 32,602 training examples containing the article’s title, a 1-2 sentence article description and a pre-labeled category
- Removed punctuation and capitalized all letters

---

Methodology

**Feature Extraction:**
- Simple cross-validation: 70% Training - 30% Testing
- Used Term Frequency-Inverse Document Frequency (TF-IDF) on each category to extract the $w$-most salient words for each category
- Tested feature sizes of $w = 30, 40, 50, 70, 100$
- Classic TF-IDF had better performance than Sublinear TF scaling

**Classifiers:**
- SVM
  - C-Support Vector Classifier (C-SVC) with a radial basis function, given by $e^{-\gamma d^2}$, where $\gamma$ is a constant
  - Used LIBSVM – one-against-all training model
  - $C$ and $\gamma$ optimized using parameter “grid search” (internal cross validation)
  - Optimal parameter values: $C = 2$ and $\gamma = 2^{-5}$
- Naïve Bayes
  - Applied Laplace Smoothing
- Softmax Regression
  - Applied $p(y=1|x; \Theta) = \Phi$, $\Phi$ is a constant

---

Raw Accuracy:
- Naïve Bayes (with Laplace smoothing): 56.80%
- Softmax Regression: 70.97%
- SVM: 64.88%
- SVM without optimized $C$ and $\gamma$: 57.25%

<table>
<thead>
<tr>
<th>Category</th>
<th>F1-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sports</td>
<td>74.06%</td>
</tr>
<tr>
<td>US</td>
<td>54.60%</td>
</tr>
<tr>
<td>Science &amp; Technology</td>
<td>53.97%</td>
</tr>
<tr>
<td>Business</td>
<td>59.81%</td>
</tr>
<tr>
<td>World</td>
<td>70.17%</td>
</tr>
<tr>
<td>Entertainment</td>
<td>60.67%</td>
</tr>
<tr>
<td>Health</td>
<td>60.33%</td>
</tr>
</tbody>
</table>

---

Conclusions & Future Work

**Conclusions:**
- Softmax Regression is the most accurate classifier
- High computation cost to optimize parameters
- Difficult to classify articles with high accuracy
- Able to classify articles above chance

**Future Work:**
- Assess classifiers when top words are obtained with additional variations of the TF-IDF algorithm
- Continue to improve classifiers and test on others (e.g. Neural Networks)