BRNIR at the NTCIR-14 finnum task:
Scalable feature extraction technique for numeral classification

Alan Spark, Team Lead at AUTO1 GROUP
Agenda

• Motivation
• Features types
• Extraction pipeline
• Experiment design
• Results
Motivation

• Focus on feature extraction in unsupervised fashion
• Experiments on different features concatenations
• Suggest a feature extraction pipeline
<table>
<thead>
<tr>
<th>Features types</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Topic distribution</strong></td>
<td>a vector with topics distribution of a tweet</td>
<td></td>
</tr>
<tr>
<td><strong>Tickers</strong></td>
<td>multi-label encoding of tickers presented in a tweet</td>
<td></td>
</tr>
<tr>
<td><strong>Tags</strong></td>
<td>multi-label encoding of tags presented in a tweet</td>
<td></td>
</tr>
<tr>
<td><strong>Number properties</strong></td>
<td>a vector encoding a number properties such a value, position &amp; type and other</td>
<td></td>
</tr>
<tr>
<td><strong>Token context</strong></td>
<td>&quot;Bag-of-words&quot; like encoding of tokens neighboring a number.</td>
<td></td>
</tr>
<tr>
<td><strong>Character context</strong></td>
<td>&quot;Bag-of-words&quot; like encoding of characters neighboring a number</td>
<td></td>
</tr>
</tbody>
</table>
Extraction pipeline

Data → Preprocessing

Topic Allocation

Number properties

Value normalisation

without value normalisation

Feature vector
Experiment design

Tweet features
- topic distribution
- tickers
- tags

Number features
- character context
- token context
- number properties

class probabilities
- softmax

ReLu

Dropout

ReLu
Results
Summary and Future work

• unsupervised approaches for feature extraction in application to FinNum task
• methods are parallelizable and meant to be run at scale

• utilize data discovered at preprocessing step
• address natural imbalance
• embedding for all “sparse” features
• experiment with classification models
Thank you
Q&A
Additional plots
$FNKO$ $10$ is a no-brainer. Should trade back to IPO price $12$. Remember, initial range on IPO was $16$ on high end. Quiet period expiry soon.

**target num:** ["10", "12."]
**discovered numbers:** [10, 12, 16]

The approach detects extra numbers in more than 32% of tweets in given corpus
Number of “target numbers” per tweet on the left, Number of unique categories/subcategories per tweet on the right.