Understanding Happiness by Using a Crowd-sourced Database with Natural Language Processing

Yixia Wang  Giacomo Spigler
Tilburg University
y.wang_1@tilburguniversity.edu, g.spigler@uvt.nl

Abstract
In this thesis, there are three studies that delve into a database to know more about Happiness. This crowd-sourced database is called Happy DB, which consists of more than 100,000 descriptions of happy moments collected using Amazon’s Mechanical Turk. Today we focus on the first study where we conduct two classification tasks on the said database to learn two critical concepts of happiness, agency and sociality. We apply the state-of-art word embedding algorithm BERT to transform all happy moments to context-sensitive representations and then feed them to a one-layer LSTM. We found that the proposed setup improves performance compared to the previous works.

Dataset Description
For these classification tasks, there is already a corpus called CL-Aff corpus[4] where ‘sociality’ and ‘agency’ are annotated as an extension to the original happy DB database. ‘Sociality’ refers to feeling happy in the presence of others vs alone, while ‘agency’ denotes whether the happy moment refers to the participant who reported it or to other people.

Previous Work
Previous work explored diverse methods involving supervised and semi-supervised learning. Various embedding algorithms were also employed including word2vec [6], GloVe [7], ELMo [8] and word embeddings pre-trained on WikiText-103 corpus[14].

Our Methods
We develop eight baseline models and one-layer LSTM paired with GloVe as a benchmark, to which our main model, one-layer LSTM accompanied by BERT, is compared.

Main Model: LSTM with BERT
Why LSTM?
1. LSTM addresses problems where other neural networks could not have effectively tackled.
2. It retains the characteristics of varying-length inputs.
3. It includes a short-term component and also a long-term memory component.
4. Since invented, LSTM has documented outstanding performance in tasks such as text generation, text synthesis and sentiment analysis, to name a few.

In order to achieve a better outcome, we trailed several rounds and the hyper-parameter values that contribute to the best performer in both tasks are marked in bold.

Baseline Models
Eight baseline models are developed by applying traditional machine learning algorithms, including Support Vector Machine(SVM), Random Forest, Logistic Regression(Log Reg) and Naive Bayes. Each of them is implemented with two sets of word embedding algorithms and they are: Bag of Words(BOW) and Bag of Words with a TF - IDF transformation(BOW tf-idf).

On top of these machine learning approaches, LSTM paired with GloVe is also used as a benchmark. GloVe is an unsupervised learning algorithm for distributed word representation, whose performance we want to compare to that of BERT.

Results
The results presented here suggest that the proposed model with BERT outstrips the other two in both tasks and in all evaluation methods. It achieves 86.42% accuracy, 90.42% F1 score and 91.41% area under curve(auc) in the agency classification task. As to the sociality task, as the samples are more balanced, the results are better. Our model with BERT hit 97.11% in auc. Its F1 score is also the highest, boasting 93.49%. The model also achieved 93% in accuracy, outperforming the previous best model(ELMo+LSTM).

As to baseline models, the linear SVM with BOW as the word representation method performs well on the first task, achieving a score of 90.49%. Logistic Regression outperforms others, reaching 80.65% of accuracy in agency classification.

Table 1. Accuracy of the proposed LSTM+BERT model on Agency and Sociality classification.

<table>
<thead>
<tr>
<th>Models</th>
<th>Accuracy</th>
<th>Agency</th>
<th>Sociality</th>
<th>AUC</th>
<th>Agency</th>
<th>Sociality</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>ELMo + LSTM (publication)</td>
<td>92.00%</td>
<td>None</td>
<td>90.00%</td>
<td>90.42%</td>
<td>None</td>
<td>91.41%</td>
<td>None</td>
</tr>
<tr>
<td>GloVe + LSTM</td>
<td>96.11%</td>
<td>96.90%</td>
<td>95.70%</td>
<td>95.30%</td>
<td>96.55%</td>
<td>96.41%</td>
<td>95.41%</td>
</tr>
</tbody>
</table>

Table 3. Best performing baseline models on Sociality and Agency classification respectively.

Reference