Affective Content Classification using Convolutional Neural Networks

Daniel Claeser

Fraunhofer FKIE, Fraunhoferstrasse 20, 53343 Wachtberg, Germany
daniel.claeser@fkie.fraunhofer.de

Abstract. We present a three-layer convolutional neural network for the classification of two binary target variables 'Social' and 'Agency' in the HappyDB corpus exploiting lexical density of a closed domain and a high degree of regularity in linguistic patterns. Incorporating demographic information is demonstrated to improve classification accuracy. Custom embeddings learned from additional unlabeled data perform competitive to established pre-trained models based on much more comprehensive general training corpora. The top-performing model achieves accuracies of 0.90 for the 'Social' and 0.875 for the 'Agency' variable.

Keywords: Convolutional Neural Networks · Unsupervised Learning · GloVe · FastText.

1 Introduction

The CL-Aff Shared Task [1], held as a part of the Affective Content Analysis workshop at AAAI 2019, invited participants to analyze and classify the contents of HappyDB [2], a corpus of 100,000 'Happy Moments'. Subtask 1 consisted of classifying contents with respect to two binary variables, 'Agency' and 'Social', with 'Agency' indicating whether the author of a happy moment was in control of events and 'Social' indicating whether additional people were explicitly or implicitly involved. In addition, an open-ended second subtask invited participants to share insights from the corpus with respect to 'ingredients of happiness'.

To the best of the author's knowledge, no similar shared task or challenge has previously been proposed, and while there has been extensive research on sentiment and affect analysis, the task at hand is very specific and its scope is limited to pre-classified data describing 'happy moments'. The task at hand could therefore not be approached with established techniques for sentiment or polarity analysis. It was rather considered a classification task aiming for the detection of semantic ('Social' variable) and syntactic ('Agency' variable) patterns, with both implicit and explicit concepts present in the data.

In recent years, embedding-based deep learning techniques have gained momentum superseding conventional machine learning techniques in a broad range of linguistic tasks, currently constituting the absolute majority of publications at the four major venues of computational linguistics [5]. The use of neural networks employing the technique of vector embeddings seemed a natural choice

given the need to extend the language model to abstract concepts beyond the lexical surface structure.

2 Dataset

A comprehensive description of the dataset provided along with informative basic statistics can be found in the original HappyDB paper ([2]). The following section describes some additional insights into the data structure that proved to be relevant for classification approach and performance.

2.1 Analyzed subsets

It was quickly noted that 95.2% of the provided happy moments were tagged as submitted from just two countries, United States (8378 or 79.3%) and India (1674 or 15.9%), while the remainder of the corpus of just 508 happy moments was distributed among 69 other countries. In light of this uneven distribution and the resulting challenges for claiming statistically significant insights on this data, only the subsets from the aforementioned two countries were considered for further evaluation and additional classification experiments.

2.2 Duplicates

While the authors of HappyDB took basic cleaning and quality assurance measures with respect to misspellings and removal of non-informative entries, the corpus contains a considerable proportion of duplicates.

While the corpus contains 1,674 entries with the country tag 'IND', a manual inspection of those moments revealed the presence of a high number of duplicates. After removing exact literal duplicates, the subset was 391 entries lighter, leaving 1,283 entries. Removing punctuation to further catch small variations in otherwise identical utterances, like the college example in table 1, left 1,246 unique entries, reducing the number of available examples for training and evaluation of the classifier by more than 25%.

Occurrences	**Duplicate**
126	i went to temple
100	i went to shopping
15	i went to college.
15	i went to college
13	the day with my wife
12	my boy friend love feeling
10	when i am getting ready to [...]*

Table 1. Example duplicates for country code 'IND'

The seven most common duplicate entries alone make up 391 (23.3%) of all moments with country tag 'IND'. Note that the entry *"i went to college"* occurs with and without full stop 15 times each. Additionally, the majority of these duplicates were submitted along with contradicting demographic information. While sentences like *"i went to college"* might indeed have been submitted by multiple participants, more distinct duplicates like the irregular pattern second to the bottom or complex utterances like the example at the bottom (shortened from originally *"when i am getting ready to go to my office my parents send off with cute smile and say have a nice day and take care"*) were almost certainly submitted multiple times by the same worker. Even the cleaned-up subset still contains several very similar complex utterances. Undeniably, the presence of such a high proportion of duplicates in one category has a considerable distorting effect on training and evaluation of a classifier.

The situation was far less critical for the 'USA' subset of the corpus with 208 duplicates amounting to less than 2.5% of entries in the corpus. The overall duplicate ratio over the entire corpus was 6.2%

Only the cleaned up versions of the 'USA' and 'IND' subsets were considered for further analysis and training the classifiers.

2.3 Lexical, syntactic and idiomatic properties

The material provided by participants from the US and India differed from each other in several linguistic dimensions. The exact linguistic background of individual authors remained unclear as both countries are polyglot, however, it seems reasonable to assume the majority of US participants to be native speakers of English or highly fluent in the language. The vast majority of authors submitting from India is in contrast assumed to use English as a second language, with a more diverse linguistic background than US participants. Assuming a descriptive rather than prescriptive point of view, it is not of particular interest whether particular patterns in the Indian subset might be considered correct or appropriate by native and proficient speakers of English as long as they are distinct and reproducible enough for a classifier to learn. The intuition that patterns in this subset might be distinct enough for the classifier to benefit from learning them separately was proven correct experimentally.

American and Indian submissions differed considerably with respect to syntactic patterns to start with: While statements from US authors contained 13.52 tokens on average per sentence with a standard deviation of 6.78, Indian statements contained 12.71 tokens on average with a considerably higher standard deviation of 10.59 caused e.g. by a larger proportion of particularly long statements. While the authors were originally instructed to state complete sentences, the level of compliance varied between the two groups, with e.g. US authors starting 8.4% of sentences by a gerund form compared to 5.7% of Indian authors. Tables 2 and 3 show the most common trigrams starting sentences from the two different groups, demonstrating US authors use a considerably higher share of idiomatic expressions such as "i got to" and framing expressions such as "an event that [made me happy]" and "i was happy", marked in bold. The

Indian statements might in that light tentatively be characterized as being more straightforward. Additional differences involve Indians using simple and progressive present substituting simple past more often than US authors and a higher rate of omission of particles such as prepositions. Indian statements were lexically more dense with a types to tokens ratio of 9.67 compared to 8.00 in US statements.

Occurrence	**Trigram**	**Relative**	**Cumulated**
216	**i was happy**	2.65	2.65
206	i went to	2.52	5.17
188	i was able	2.30	7.47
178	**i got to**	2.18	9.65
177	i got a	2.17	11.82
144	i had a	1.76	13.59
92	i bought a	1.13	14.71
71	i received a	0.87	15.58
71	**i found out**	0.87	16.45
67	**an event that**	0.82	17.28
56	i made a	0.69	17.96
51	i watched a	0.62	18.59
43	i went on	0.53	19.11
43	i found a	0.53	19.64
42	i ate a	0.51	20.15
40	i went out	0.49	20.64
34	**it made me**	0.42	21.06
33	my wife and	0.40	21.47
30	my husband and	0.37	21.83
30	**i took my**	0.37	22.20

Table 2. 20 most common trigrams at sentence beginning, USA

2.4 Syntactic patterns

Participants in the crowdfunding process creating the HappyDB corpus were explicitly asked to state moments that made them happy in single full sentences. While not all participants submitted strictly complied to those instructions, the overwhelming majority of statements are in the form of full declarative sentences. Syntax in the corpus can thus be regarded fixed and discarded as distinct piece of information in the classification process.

Occurrence	Trigram	Relative	Cumulated
53	i went to	4.31	4.31
20	i got a	1.63	5.93
20	i bought a	1.63	7.56
15	i went for	1.22	8.78
14	**my happiest moment**	1.14	9.92
10	yesterday i went	0.81	10.73
10	i met my	0.81	11.54
9	me and my	0.73	12.28
9	i was very	0.73	13.01
9	in the past	0.73	13.74
8	when i am	0.65	14.39
8	last month i	0.65	15.04
8	i got my	0.65	15.69
7	my best friend	0.57	16.26
7	i purchased a	0.57	16.83
7	i had a	0.57	17.40
6	we bought a	0.49	17.89
6	the day i	0.49	18.37
6	bought a new	0.49	18.86
5	we went to	0.41	19.27

Table 3. 20 most common trigrams at sentence beginning, India

3 Experiments and results

3.1 Basic considerations and setup

Given the almost uniform syntactic structure of the corpus with respect to declarative sentences, a convolutional neural network was determined to be an appropriate architecture rather than a time-step based approach: Considering syntax more or less fixed relieves the classifier of the effort to interpret the complete input as sequences and allows to focus on detecting the presence or absence of features relating to agency or social participation in the utterance. Two binary classifiers were trained to address each variable separately. A large search space of configurations was explored, yielding the following configuration with the best performance in terms of accuracy: Two convolutional layers with 128 filters each with a step size of 5 and a dense layer with 128 units. Applying dropout of 10 and 20 % yielded slight but statistically insignificant improvements. Batch sizes were iterated in steps of 8, 16, 32, 64 and multiples of 64 up to 1024, with medium batch-sizes of around 384 performing best in the vast majority of configurations.

Table 4 shows overall results in the best-performing configurations with the architecture described above.

As higher dimensional embeddings consistently outperformed low-dimensional models, only the 300 dimensional models were considered for further experiments.

3.2 Pre-trained and customized embeddings

Three major groups of pre-trained embeddings were used for the initialization layer of the neural network: FastText by Facebook AI [4], GloVe by Stanford University [3] and custom FastText embeddings trained on the joint set of labeled and unlabeled HappyDB data provided by the task's authors.

To assess the degree to which the supplied labeled and unlabeled HappyDB data were able to reflect syntactic and semantic relations of the domain in comparison to broader knowledge of predefined embeddings as distributed by the authors of FastText and GloVe, FastText embeddings of different dimensionality and with both available approaches, CBOW and SkipGram, were trained and evaluated as displayed in Table 4.

3.3 Constructing two binary classifiers

Based on aforementioned considerations, one binary classifier was constructed for each dependent variable, 'Agency' and 'Social', each with the target values 'yes' or 'no' as labeled in the training data.

Embedding	**Dim's**	**Accuracy A**	**Accuracy S**	**F1 A**	**F1 S**
GloVe, 6B	300	0.868	0.887	0.835	0.885
GloVe, 840B	300	0.871	0.8975	0.841	0.894
FastText, Wiki-News	300	**0.875**	**0.900**	0.842	0.888
FastText, Wiki-News Subword	300	0.87	0.8925	0.839	0.889
FastText Crawl	300	0.872	0.896	0.842	0.892
FastText Crawl Subword	300	0.871	0.896	0.840	0.892
FastText, Wikipedia	300	0.873	0.898	0.841	0.896
FastText, HappyDB, Skip	300	0.874	0.894	0.843	0.889
FastText, HappyDB, CBOW	300	0.869	0.889	0.838	0.884
GloVe, Twitter	300	0.871	0.894	0.840	0.891
GloVe6B	200	0.87	0.885	0.840	0.882
FastText, HappyDB, Skip	200	0.873	0.896	0.842	0.892
FastText, HappyDB, CBOW	200	0.869	0.885	0.839	0.880
FastText, HappyDB, Skip	100	0.872	0.895	0.840	0.891
FastText, HappyDB, CBOW	100	0.868	0.882	0.838	0.879
GloVe, Twitter	100	0.871	0.894	0.837	0.890
GloVe, 6B	100	0.867	0.881	0.821	0.878
GloVe, 6B	50	0.862	0.879	0.818	0.876
GloVe, Twitter	50	0.868	0.871	0.821	0.867
FastText, HappyDB, CBOW	50	0.863	0.871	0.830	0.867
FastText, HappyDB, Skip	25	0.862	0.877	0.832	0.873
GloVe, Twitter	25	0.863	0.871	0.832	0.868

Table 4. Accuracy, Macro F1 Agency, Social (abbreviated A, S) 10-fold cross-validated

3.4 Training classifiers on four classes

Table 5 shows the uneven distribution of the two variables and their co-occurrences in the corpus, illuminating some basic connections in agreement with the psychological findings quoted by the authors of HappyDB: A majority of 73.8% of happy moments involves active participation or control by the author. Within these moments, an absolute majority of 54.4% involves no other people than the acting authors themselves. In turn, within the 26.2% of moments with no active participation of the author, the probability is 74.9% that other people are involved, reflecting the intuition that in most instances, something, or somebody, needs to cause the happiness after all. This connection raised interest in the performance of a classifier considering each combination of the two variables a distinct class, thus forming four classes "Agency no, social no", "Agency yes, social no", "Agency no, social yes" and "Agency yes, social yes". While there is apparently a strong conditional probability of "Social: yes" given "Agency: no", the significantly lowered number of samples was expected to cause a drop in performance, especially with only 693 samples for the "Agency no, social no" class, an assumption that was confirmed by the experimental results as displayed below.

	Social no	**Social yes**	**Sum**
Agency no	693	2071	2764
Agency yes	4242	3554	7796
Sum	4935	5625	10560

Table 5. Distribution of classes and co-occurrence of target variables

Embedding	**Agency**	**Social**	**Both**
FastText, HappyDB, 300d, SkipGram	0.771	0.820	0.691
Glove6B, 300d,	0.753	0.801	0.690
FastText, Wiki-News	0.803	0.822	0.686

Table 6. Results of initial experiments with four classes

The results of the three top-performing high-dimensional configurations instantly affirmed those expectations and ceased interest in further experiments: Combining the two variables into four categories decreased the performance even when evaluating only for one variable per category well below the achievable results in the binary setting.

3.5 Training separate classifiers by countries

The presence of aforementioned distinct syntactic and lexical characteristics in the two largest groups by country inspired the question whether classification performance would benefit from training separate classifiers for each group. Since only the USA and India subsets contained more than 1000 samples, the exploration was limited to those subsets. Three separate classifiers were trained, one for 'IND' and 'USA' each with 1246 samples (which equals the number of available samples for 'IND' to receive a balanced setting) each and one with the 1246 split between the two countries proportionally in alignment to the original full training corpus.

Embedding	**Acc USA**	**Acc IND**	**Acc Mixed**
GloVe840	0.849	0.857	0.844
GloVe6b	0.849	0.854	0.850
FastText Crawl	0.852	0.859	**0.856**
FastText Crawl Subword	0.852	**0.863**	0.843
FastText Wiki-News	0.857	0.857	0.845
FastText Wiki-News Subword	0.842	0.845	0.829
FastText Wikipedia	0.850	0.848	0.855
FastText, HappyDB, CBOW	0.856	0.859	0.854
FastText, HappyDB, Skip	**0.860**	0.857	0.842

Table 7. Accuracy for 'Agency' with USA and IND trained separately and jointly

The results show a modest but statistically significant (confidence level 0.95) improvement for both language groups with the moments submitted under country code IND benefitting considerably stronger. We suggest this might be an effect of more compact syntax patterns (see above).

Embedding	**USA**	**IND**	**Mixed**
GloVe840	**0.895**	0.866	**0.838**
GloVe6b	0.880	0.859	0.821
FastText Crawl	0.894	0.863	0.830
FastText Crawl Subword	0.867	0.855	0.820
FastText Wiki-News	0.896	0.859	0.824
FastText Wiki-News Subword	0.843	0.847	0.823
FastText Wikipedia	0.880	0.851	0.823
FastText HappyDB, CBOW	0.890	**0.868**	0.832
FastText HappyDB, Skip	0.887	0.864	0.829

Table 8. Accuracy for 'Social' with USA and IND trained separately and jointly

The picture is even clearer for the Social variable. For both variables, the separated classifiers achieve better performance than their combined average. However, the degree of convergence for this phenomenon towards larger training sets has not been investigated.

3.6 Classification by concepts

The authors of HappyDB report on successful efforts to categorize the corpus by a set of crowd-sourced category labels. Additionally, they identified a set of *concepts* or *topics* of happy moments in a seemingly rather intuitive and subjective way. To apply a limited test of replicability to this set of topics, a classifier with the aforementioned architecture was trained on a subset of the corpus consisting of happy moments labeled with exactly one concept, limited to concepts with more than 1000 labeled examples, which were *careeer*(1280), *entertainment* (1135), *family* (1259) and *food* (1007).

Embedding	**Dimensions**	**Accuracy**
GloVe6b	300	0.908
GloVe840	300	0.913
FastText, Wiki-News	300	0.912
FastText, Wiki-News Subword	300	0.884
FastText, Crawl	300	**0.916**
FastText, Crawl Subword	300	0.899
FastText, Wikipedia	300	0.908
FastText, HappyDB, Skip	300	0.915
FastText, HappyDB, CBOW	300	0.892
GloVe, Twitter	200	0.910
GloVe6B	200	0.901
FastText, HappyDB, Skip	200	0.914
FastText, HappyDB, CBOW	200	0.893
FastText, HappyDB, Skip	100	0.911
FastText, HappyDB, CBOW	100	0.889
GloVe, Twitter	100	0.901

Table 9. Classification by concepts, four most common single labels

4 Conclusion

We introduced a rather simplistic architecture to classify the HappyDB contents with respect to the two binary variables 'Agency' and Social. HappyDB prove to be a high-quality linguistic resource with a high degree of replicability in terms of machine learning and classification as proven by experimental results for both the target variables defined by the Shared Task and the ability to reproduce the concepts introduced by the HappyDB authors. We observe that while embeddings

trained only on HappyDB without any external world knowledge supplied cannot statistically significantly outperform established general purpose embeddings such as FastText and Glove trained on Wikipedia and crawled web content, they appear to be almost competitive utilizing a database of not even 20,000 types as opposed to up to 2 million types in the pre-trained embeddings. We observe no particular social media benefit for embeddings in accordance to the assumption that most statements were given in a rather formal register as intended by the corpus' authors. Classification appears to benefit from taking linguistic backgrounds of different groups of authors into account, and we recommend cleaning the corpus from remaining duplicates to avoid distortions.

5 Acknowledgement

I would like to express my gratitude to Fahrettin Gökgöz and Albert Pritzkau of Fraunhofer FKIE and Maria Jabari of University of Bonn for their expertise and insights supporting the system design and dataset analysis.

References

1. Jaidka, K., Mumick, S., Chhaya, N., Ungar, L.: The CL-Aff Happiness Shared Task: Results and Key Insights. In: Proceedings of the 2nd Workshop on Affective Content Analysis @ AAAI (AffCon2019). Hawaii (2019)
2. Asai, A., Evensen, S., Golshan, B., Halevy, A., Li, V., Lopatenko, A., Xu, Y.: HappyDB: A Corpus of 100,000 Crowdsourced Happy Moments. In: Proceedings of LREC 2018. European Language Resources Association (ELRA), Miyazaki, Japan (2018)
3. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543) (2014)
4. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL) (2017)
5. Young, T., Hazarika, D., Poria, S., Cambria, E.: Recent trends in deep learning based natural language processing. ieee Computational intelligenCe magazine, 13(3), 55-75. (2018)