Generating Document Embeddings for Humor Recognition using Tensor Decomposition

Andrew Cattle[1[0000−0002−0133−1237]], Zhenjie Zhao[1[0000−0002−4396−9191]], Evangelos Papalexakis[2[0000−0002−3411−8483]], and Xiaojuan Ma[1[0000−0002−9847−7784]]

[1] Hong Kong University of Science and Technology
Clear Water Bay, Kowloon
Hong Kong
acattle@connect.ust.hk, {zzhaoao, mxj}@cse.ust.hk
[2] University of California, Riverside
900 University Ave
Riverside, CA 92521
USA
epapalex@cs.ucr.edu

Abstract. This paper details our submission to the HAHA 2019 (5) group task on humor recognition. We propose a novel humor recognition system based on tensor embeddings, capable of being trained without the need for any external training corpora. Our model achieves an F1-score of 0.736 on a binary humor classification task and a root-mean-squared-error of 0.963 on a humor scoring task, both using a Spanish-language Twitter corpus (2). While our experiments are performed on Spanish documents, our approach is truly language agnostic and can be applied to any language with minimal adaptation.

Keywords: Computational Humor · Humor Recognition · Tensor Decomposition

1 Introduction

Humor is an integral part of human interaction. It can be used to defuse tense situations, increase likeability, or even for pure entertainment. As such, the automatic recognition of humor represents an important step for natural human-computer interaction (18). While early works tended to approach humor recognition as a binary classification task (11; 21), fine-grained humor evaluation has gained recent attention (3; 14; 18).

This paper details our submission to the HAHA 2019 (5) group task on humor recognition. We propose a novel humor recognition system based on tensor embeddings, capable of being trained without the need for any external training

corpora. While most humor recognition works focus on English corpora, HAHA 2019 utilizes a Spanish-language corpora (2). As such, our model prioritizes language-agnosticism and thus can be applied to any language with minimal adaptation. The code used to create our model is available for download[1].

2 Related Work

2.1 Humor Recognition

Although humor recognition has typically been framed as a binary classification task (1; 11; 21), recent works have moved beyond mere humor detection and toward humor evaluation by framing humor recognition as a relative ranking task (3; 14; 18). Generally, these works have focused on documents generated around a common prompt such as cartoon captions (18) or Twitter Hashtag Wars (3; 14).

Humor is a complex phenomenon, incorporating aspects of phonology, style, semantics, and word-choice. As such, existing humor recognition works have tended to use a variety of features designed to capture this complexity. (1) extract acoustic features from sitcom audio tracks while (6) incorporate "phonetic embeddings" generated using a character-to-phoneme LSTM encoder-decoder. (11) look for alliteration, rhyming, negative sentiment, and adult slang to aid humor recognition while (17) add emotional scenarios. Inspired by incongruity theory (16), several works attempt to measure the amount of incongruity in a document using various lexical similarity metrics (4; 18; 21). Other works represent semantics using word (1; 6) or document (21) embeddings as model inputs. (15) apply document centrality as defined by the graph-based text summarization model LexRank (7). A more basic approach is to use word frequency (11) and n-gram probability (20) as indications of humor.

The majority of humor recognition works focus on English language humor. Given the dependence of many models on language dependant resources, adapting these models for non-English contexts can be challenging. (21) and (1) make use of the semantic ontology WordNet (12). Similarly, (6) train their phonetic embeddings on the CMU pronouncing dictionary (10). More fundamentally, many existing systems achieve their results using high quality, large scale pre-trained word embedding models such as Google word2vec[2] or Stanford GloVe [3]. As such, adapting such models for use with other languages would be dependant on the availability of equivalent resources.

2.2 Tensor Embeddings

State-of-the-art document embedding approaches like doc2vec (9) or sent2vec (13) are capable of encoding the meaning of a document but often require large

[1] https://github.com/acattle/HumourTools/
[2] https://code.google.com/archive/p/word2vec/
[3] https://nlp.stanford.edu/projects/glove/

amounts of training data. By comparison, tensor decomposition is capable of generating low-rank embeddings of sentences that capture the similarity of contextual patterns without the need for large training corpora (8).

Such "tensor embeddings", combined with a label propagation technique, have been shown effective in semi-supervised fake news detection (8). The fact that high quality embeddings can be created from relatively small corpora makes tensor embeddings an attractive option for training models for low-resource languages where large scale corpora may not be available.

3 Method

3.1 Tensor Embeddings

For each document in the corpus we compute a tensor embedding based on word co-occurrence. That is, for a corpus $\mathcal{D} = \{s_1, s_2, \ldots, s_D\}$ with D sentences, we first extract a vocabulary $w_1, w_2, \ldots, w_V$, where V is the number of words. For each sentence s in $\mathcal{D}$, we count the word-word co-occurrences within a small window H. This results in a frequency matrix $\mathbf{W}_s \in \mathbb{Z}^{V \times V}$, where $\mathbb{Z}$ denotes the set of integers. In particular, $\mathbf{W}_s(i, j)$ indicates the frequency that word w_i and w_j co-occur in s within the window H. This allows us to encode the lexical patterns of s in $\mathbf{W}_s$. All $\mathbf{W}_s$ are then stacked, creating a three-dimensional tensor $\boldsymbol{\mathcal{W}} \in \mathbb{Z}^{V \times V \times D}$. The objective of tensor decomposition is to find an approximation $\hat{\boldsymbol{\mathcal{W}}}$ of $\boldsymbol{\mathcal{W}}$ such that:

$$\hat{\boldsymbol{\mathcal{W}}} = \sum_{r=1}^{R} \mathbf{v}_r \otimes \mathbf{v}_r \otimes \mathbf{d}_r, \tag{1}$$

where $\mathbf{v}_r \in \mathbb{R}^V$, $\mathbf{d}_r \in \mathbb{R}^D$, R is the pre-defined rank parameter, and $\otimes$ is the outer product, namely, $\mathbf{v}_r \otimes \mathbf{v}_r \otimes \mathbf{d}_r$ being a three-dimensional tensor, and

$$\mathbf{v}_r \otimes \mathbf{v}_r \otimes \mathbf{d}_r(i, j, k) = \mathbf{v}_r(i) \cdot \mathbf{v}_r(j) \cdot \mathbf{d}_r(k). \tag{2}$$

In particular, $\mathbf{C} = [\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_R] \in \mathbb{R}^{D \times R}$, where the s-row of $\mathbf{C}$ is the embedding vector of sentence s.

The low-rank sentence embeddings are calculated using the alternating least squares method of the CANDECOMP/PARAFAC tensor decomposition implementation in Matlab tensor toolbox [4] (19). H was set as 5 and R set as 50.

3.2 Humor Recognition

After extracting tensor embeddings for each document, we performed a simple baseline experiment using the (2) corpora. We trained both classification and regression models in a supervised manner on the tensor embeddings corresponding to the documents in the training set along with their labels. These models were

[4] https://www.tensortoolbox.org/

Table 1. Humor Recognition performance by model. Accuracy, Precision, Recall, and F1-score describe binary classification performance (Task 1). Root-mean-squared-error describes humor scoring performance (Task 2).

Model	Acc	P	R	F1	RMSE
Embedding Only	0.736	0.683	0.602	0.640	0.963
Label Prop	0.595	0.400	0.076	0.128	—
Enhanced	0.726	0.672	0.580	0.623	0.958

then used to predict the labels and scores of documents in the test set as part of HAHA 2019's (5) binary classification and humor scoring tasks, respectively.

Second, inspired by (8)'s work on fake news detection, we performed applied label propagation (22), a semi-supervised labeling technique, to the extracted tensor embeddings. In this method, labels are propagated to their neighbors in a weight average way.

Third, inspired by (15), we compute a lexical centrality feature. While (15) uses a graph-based approach, we instead take a vector-space approach. Operating on the assumption that funnier documents are more central, we calculate a tensor embedding centroid as the average of all the tensor embeddings. The Euclidean distance of each tensor embedding from the centroid is then taken as an indication of humor. While this distance-based metric is a type of humor score in-and-of-itself, it is not appropriate for the HAHA 2019 (5) humor scoring task. While the (2) dataset rates each document on a five point scale, with funnier tweets receiving higher scores, this distance-based metric has no upper bound and is inversely proportional to the perceived humor. Instead, we trained further so-called "enhanced" classifier and regression models on a combination our lexical centrality, label propagation, and tensor embeddings. We hoped that the addition of the label propagation and lexical centrality features would help capture complex pattern that may not be extracted from the raw tensor embeddings.

Except where otherwise noted, all results reported in this paper were obtained using the Random Forest classifier and regressor implementations in scikit-learn[5] with default parameters. These results are comparable to results we obtained using Support Vector Machine and Extra Trees models. Label propagation was also performed using the implementation in scikit-learn with default parameters.

4 Results and Discussion

The performance of our humor recognition models is shown in Table 1. Despite the simplicity of our models, we are able to achieve reasonable performance. Our tensor embedding only models outperforms both our label propagation and enhanced models across the board. Despite the performance of a similar model

[5] https://scikit-learn.org/

on fake news detection (8), our label propagation model performs surprisingly poorly; correctly identifying less than 10% of true positives in the test set.

It is important to note at while the (2) dataset is in Spanish, our tensor embeddings, and thus all of our models, are completely language agnostic. Because our embeddings are generated only from the dataset itself, without the need for any external training corpora, our models can be readily applied to any corpus regardless of language. The only requirement is for reliable word segmentation, which may be an issue for languages with optional or inconsistent whitespace (e.g. Chinese).

Another advantage of our approach is its relative simplicity. The most computationally expensive aspect of our model is computing the tensor decomposition. This makes approach quite scalable when compared with complex sequenced-based neural models such as those used by (1) or (6) and better-suited than such systems for small datasets.

Part of the reason for the poor label propagation performance may be the (2) dataset's unbalanced nature. Of the 24,000 training examples, only 9,253 have a positive label. Since our label propagation system employed a K Nearest Neighbors kernel (the default in scikit-learn), the larger number of negative humor labels may have overwhelmed any positive labels. A different choice of kernel or further hyperparameter tuning may lead to better results.

The semi-supervised nature of label propagation may further explain this lack of performance. Given the difficulty in obtaining reliable humor judgments, it should come as no surprise that most humor datasets tend to be relatively small. Some datasets, like Pun of the Day (21), are as small as a few thousand documents. With so few training examples, fully supervised approaches run the risk of either failing to extract meaningful patterns or placing too much emphasis on patterns in the training set that may not generalize. Thus, we expected a semi-supervised approach, like label propagation, would help mitigate these risks. However, the (2) dataset used in this paper is relatively large for a humor dataset, containing 24,000 training examples and 6,000 test, diminishing these advantages.

The performance of label propagation may also be affecting the performance of the enhanced model. Another potential limitation the enhanced model is our lexical centrality feature. One major difference between the (2) dataset and the dataset used by (15) is that (15) compared documents generated in response to a common prompt (i.e. captions submitted to the same New Yorker Cartoon Caption Contest). By comparison, (2) uses tweets sampled from Twitter with no regard for common prompts. As such, our centrality assumption may not hold. One possible improvement would be to cluster the tweets and computing multiple centroids. Unfortunately, we were unable to run a test using lexical centrality only due to limitations imposed as part of the HAHA 2019 (5) group task in terms of time and numbers of submissions.

Another potential area for improvement is related to the hyper parameters used to generate our tensor embeddings. The window size of 5 and tensor rank of 50 was chosen empirically due to their high performance on smaller, English-

language datasets. The larger size of the (2) dataset may allow for higher ranks while differences between English and Spanish may favor different window sizes.

5 Conclusion

In this paper we have shown that tensor embeddings are capable of producing reasonable performance for both binary humor classification and humor scoring. Furthermore, we identify some key advantages of this tensor embedding approach including its simplicity and language agnosticism. Finally, we offer several potential avenues for improving the performance of our models including better hyperparamter tuning.

Bibliography

[1] Bertero, D., Fung, P.: A long short-term memory framework for predicting humor in dialogues. In: 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016-Proceedings of the Conference. pp. 130–135 (2016), https://pdfs.semanticscholar.org/015c/26ca824f9d20b6523f44e6b9dc3b3dd65b2d.pdf

[2] Castro, S., Chiruzzo, L., Rosá, A., Garat, D., Moncecchi, G.: A crowd-annotated spanish corpus for humor analysis. In: Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media. pp. 7–11 (2018)

[3] Cattle, A., Ma, X.: Effects of semantic relatedness between setups and punchlines in twitter hashtag games. In: Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES). pp. 70–79. The COLING 2016 Organizing Committee, Osaka, Japan (Dec 2016), https://www.aclweb.org/anthology/W16-4308

[4] Cattle, A., Ma, X.: Recognizing humour using word associations and humour anchor extraction. In: Proceedings of the 27th International Conference on Computational Linguistics. pp. 1849–1858. Association for Computational Linguistics, Santa Fe, New Mexico, USA (Aug 2018), https://www.aclweb.org/anthology/C18-1157

[5] Chiruzzo, L., Castro, S., Etcheverry, M., Garat, D., Prada, J.J., Rosá, A.: Overview of HAHA at IberLEF 2019: Humor analysis based on human annotation. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019). CEUR Workshop Proceedings, CEUR-WS, Bilbao, Spain (9 2019)

[6] Donahue, D., Romanov, A., Rumshisky, A.: HumorHawk at SemEval-2017 task 6: Mixing meaning and sound for humor recognition. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). pp. 98–102. Association for Computational Linguistics, Vancouver, Canada (Aug 2017). https://doi.org/10.18653/v1/S17-2010, https://www.aclweb.org/anthology/S17-2010

[7] Erkan, G., Radev, D.R.: Lexrank: Graph-based lexical centrality as salience in text summarization. Journal of Artificial Intelligence Research **22**, 457–479 (2004)

[8] Hosseinimotlagh, S., Papalexakis, E.E.: Unsupervised content-based identification of fake news articles with tensor decomposition ensembles. In: WSDM 2018 Workshop on Misinformation and Misbehavior Mining on the Web (MIS2) (2018)

[9] Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32. pp. II–1188–II–1196.

ICML'14, JMLR.org (2014), http://dl.acm.org/citation.cfm?id=3044805.3045025

[10] Lenzo, K.: The CMU pronouncing dictionary (1998), http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[11] Mihalcea, R., Strapparava, C.: Making computers laugh: Investigations in automatic humor recognition. In: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing. pp. 531–538. Association for Computational Linguistics (2005)

[12] Miller, G.A.: WordNet: a lexical database for English. Communications of the ACM **38**(11), 39–41 (1995)

[13] Pagliardini, M., Gupta, P., Jaggi, M.: Unsupervised Learning of Sentence Embeddings using Compositional n-Gram Features. In: NAACL 2018 - Conference of the North American Chapter of the Association for Computational Linguistics (2018)

[14] Potash, P., Romanov, A., Rumshisky, A.: SemEval-2017 task 6: #HashtagWars: Learning a sense of humor. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). pp. 49–57. Association for Computational Linguistics, Vancouver, Canada (Aug 2017). https://doi.org/10.18653/v1/S17-2004, https://www.aclweb.org/anthology/S17-2004

[15] Radev, D.R., Stent, A., Tetreault, J.R., Pappu, A., Iliakopoulou, A., Chanfreau, A., de Juan, P., Vallmitjana, J., Jaimes, A., Jha, R., Mankoff, R.: Humor in collective discourse: Unsupervised funniness detection in the new yorker cartoon caption contest. CoRR **abs/1506.08126** (2015)

[16] Raskin, V.: Semantic theory of humor. In: Semantic Mechanisms of Humor. Springer (1985). https://doi.org/10.1007/978-94-009-6472-3_4, http://dx.doi.org/10.1007/978-94-009-6472-3_4

[17] Reyes, A., Rosso, P., Buscaldi, D.: From humor recognition to irony detection: The figurative language of social media. Data & Knowledge Engineering **74**, 1–12 (2012)

[18] Shahaf, D., Horvitz, E., Mankoff, R.: Inside jokes: Identifying humorous cartoon captions. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1065–1074. KDD '15, ACM, New York, NY, USA (2015). https://doi.org/10.1145/2783258.2783388, http://doi.acm.org/10.1145/2783258.2783388

[19] Sidiropoulos, N.D., Lathauwer, L.D., Fu, X., Huang, K., Papalexakis, E.E., Faloutsos, C.: Tensor decomposition for signal processing and machine learning. IEEE Transactions on Signal Processing **65**(13), 3551–3582 (July 2017). https://doi.org/10.1109/TSP.2017.2690524

[20] Yan, X., Pedersen, T.: Duluth at semeval-2017 task 6: Language models in humor detection. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). pp. 384–388. Association for Computational Linguistics, Vancouver, Canada (August 2017), http://www.aclweb.org/anthology/S17-2064

[21] Yang, D., Lavie, A., Dyer, C., Hovy, E.: Humor recognition and humor anchor extraction. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 2367–2376. Association for Computational Linguistics, Lisbon, Portugal (Sep 2015). https://doi.org/10.18653/v1/D15-1284, https://www.aclweb.org/anthology/D15-1284

[22] Zhou, D., Bousquet, O., Lal, T.N., Weston, J., Schölkopf, B.: Learning with local and global consistency. In: Proceedings of the 16th International Conference on Neural Information Processing Systems. pp. 321–328. NIPS'03, MIT Press, Cambridge, MA, USA (2003), http://dl.acm.org/citation.cfm?id=2981345.2981386