gundapusunil at SemEval-2020 Task 8: Multimodal Memotion Analysis

Sunil Gundapu
Language Technologies Research Centre
KCIS, IIIT Hyderabad
Telangana, India
sunil.g@research.iiit.ac.in

Radhika Mamidi
Language Technologies Research Centre
KCIS, IIIT Hyderabad
Telangana, India
radhika.mamidi@iiit.ac.in

Abstract

Recent technological advancements in the Internet and Social media usage have resulted in the evolution of faster and efficient platforms of communication. These platforms include visual, textual and speech mediums and have brought a unique social phenomenon called Internet memes. Internet memes are in the form of images with witty, catchy, or sarcastic text descriptions. In this paper, we present a multi-modal sentiment analysis system using deep neural networks combining Computer Vision and Natural Language Processing. Our aim is different than the normal sentiment analysis goal of predicting whether a text expresses positive or negative sentiment; instead, we aim to classify the Internet meme as a positive, negative, or neutral, identify the type of humor expressed and quantify the extent to which a particular effect is being expressed. Our system has been developed using CNN and LSTM and outperformed the baseline score.

1 Introduction

According to Wikipedia article on Internet Memes, "A meme is an idea, behavior, or style that spreads from person to person within a culture often with the aim of conveying a particular phenomenon, theme, or meaning represented by the meme".

Meme is not only about the humorous picture, the Internet culture, or the sentiment that passes along, but also about the richness and distinctiveness of its language: it is often greatly structured with unusual written style. The Internet Memes often include superimposed text description with broken grammars and spellings variations. Nowadays, Internet memes come in almost every form of media, with modish formats continuously expanding. Initially, they work as a medium for humor to be shared, using cultural themes. However, they can also be manipulated to further political ideals, company promotions, and social media marketing.

In this paper, we present a deep learning based multimodal approach for SemEval 2020: Task 8 on Memotion Analysis -The Visuo-Lingual Metaphor! (Sharma et al., 2020). This task has three Subtasks which describe as follows:

(A) Sentiment Classification: From the given input Internet meme extract the sentiment and classify it as a positive, negative, and neutral meme.

(B) Humor Classification: In this sub-task, the system has to recognize the type of humor expressed in a given Internet meme. The humor classes are sarcastic, humorous, offensive, and motivational. A meme can have more than one class.

(C) Scales of Semantic Classes: In the third sub-task, quantify the extent to which a particular effect is being expressed in a given emotion, namely humour, offensive, sarcasm or motivational. Details of such quantification are reported in Table 1.

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 1112–1119
Barcelona, Spain (Online), December 12, 2020.

All our models are trained using only the corpus supplied by Memotion analysis organizers. The evaluation metric for subtask **A** is macro F1 score and macro F1 score for each of the subtasks, and then average for subtask **B** and **C**. Primarily, we started experiments with traditional machine learning algorithms[1] like Support Vector Machine (SVM) and Logistic Regression (LR). Based on the findings from them, we moved to deep learning models[2] like Long Short-Term Memory (LSTM), CNN in search of the better model.

In the next section, we summarize the related work. Section 3 gives details about the dataset. We present the experimental setup, model description, and comparison with baselines & other methods in Section 4. Results and Error analysis constitutes in Section 5. Finally, we conclude this paper in Section 6.

2 Related Work

Sentiment analysis was pioneered for text (Dmitry Davidov et al., 2010; Johan Bollen et al., 2011; Xia Hu et al., 2012) but image-based sentiment analysis has got much less attention compared to text-based sentiment analysis. As Internet memes is a relatively new research topic, our research work is broadly related to studies on predicting the sentiment from the visual imagery in online reviews (Troung et al., 2017). Borth et al. (2013) introduced sentiment analysis on large scale visual sentiment ontology with SentiBank, a system extracting mid-level semantic attributes and from images. To inspect the image posting behavior of social media users Chen et al. (2015) developed the Visual Emotional Latent Dirichlet Allocation (VELDA) model to capture the text-image correlation from multiple modalities: the text, visual and emotional perspective of the picture. And interestingly they found results that 66% of users adding an image to their social media posts.

Recently, Anthony Hu et al. (2018) developed a multimodal sentiment analysis approach that combines text and vision to forecast the emotion word tags attached by users to their Tumblr posts. You et al. (2015) developed a robust Image Sentiment Analysis system by using a CNN on Flickr with domain transfer from Twitter for binary sentiment classification. Benet Sabat et al. (2019) worked on a challenge of automatically detect the hate speech in Internet memes by using visual information. Philipp Blandfort et al. (2019) developed a Multimodal Social Media system to study how public tweets with images posted by teens who mention gang violence on Twitter can be leveraged to automatically discover psycho-social factors and problems.

Few researchers have investigated to automate the Internet meme generation process, while a few others tried to extract its sentiment. To generate popular meme descriptions, William Yang Wang and Miaomiao Wen (2016) proposed a non-paranormal approach by combining visual and textual features. To produce the meme descriptions Peirson et al. (2018) presented an encoder-decoder meme generating system, consisting of a Google's pre-trained Inception-v3 network to generate an image embedding, followed by LSTM model with attention. In this paper, we refer to Wang and Hua's (2014) method to combine textual and vision information, while scaling up the model using effective dropout regularization.

3 Dataset

We used the corpus provided by Task 8 in SemEval 2020. This task is named as "Memotion Analysis". The dataset consists of Internet memes and corresponding text descriptions. For Task A, each meme is labeled into three sentiment classes: Positive, Neutral, or Negative, and for Task B and C, each meme is labeled with several sub-tasks: Humorous, Sarcastic, Offensive, and Motivational. Table 1 shows the detailed labeling of each task.

Table 2 shows the distribution of each task labels in the dataset. Our dataset contained 6990 memes and corresponding descriptions for training and 914 memes for testing. Organizers did not provide any dataset for development so we split the training dataset into the train (85%) and dev (15%).

[1]https://scikit-learn.org/stable/

[2]https://keras.io/

	Humor Classification (Task B)				Semantic Classes (Task C)			
Sentiment Labels (Task A)	Humour	Sarcastic	Offensive	Motivational	Humour	Sarcastic	Offensive	Motivational
Positive	Yes	Yes	Yes	Yes	Funny	General	Offensive	Motivational
Neutral	No	No	No	No	Very Funny	Twisted Meaning	Slight	Not Motivational
Negative					Hilarious	Very Twisted	Very Offensive	
					Not Funny	Not Sarcastic	Hateful Offensive	

Table 1: Labels of each task.

TASKS(in bracket each task labels)	**TRAIN**(in bracket each label counts)	**VALIDATION**
Task A		
(Positive-Neutral-Negative)	(4160-2200-630)	(564-279-71)
Task B		
(Humorous-Not Humorous)	(1650-5340)	(704-210)
(Sarcastic-Not Sarcastic)	(394-5446)	(681-233)
(Offensive-Not Offensive)	(2713-4277)	(545-369)
(Motivational-Not Motivational)	(4524-2466)	(589-325)
Task C		
(Not Funny-Funny-Very Funny-Hilarious)	(1650-2452-2238-650)	(210-315-310-79)
(Not Sarcastic-General-Twisted Meaning-Very Twisted)	(1544-350-1546-394)	(233-444-195-42)
(Not offensive-Slightly offensive-Very Offensive-Hateful Offensive)	(2713-2592-1465-220)	(369-323-198-24)

Table 2: Distribution of labels in each task.

3.1 Pre-processing

While dealing with meme dataset, some of the major challenges faced were word or phrases with multiple spelling variations, short sentences with unclear grammatical structure, memes without any images and only text, URLs, HTML tags, and imbalanced dataset. To tackle some of the above issues we took the following pre-processing steps:

1. **Removal of URLs:** Text/Meme descriptions which are extracted from Internet memes, contain URLs (GrumpyCatPics.com, memegenerator.net, etc). We removed these URLs since it does not contribute towards our goal.

2. **Removal of HTML Tags, Punctuation Marks, Digits, and Non-ASCII Glyphs:** All the HTML tags, punctuation marks, digits, and non-ASCII glyphs in a meme description are removed.

3. **Handling Usernames and Hashtags:** Replaced the usernames with a `USER` tag. Pound (#) sign removed from the hashtag and split into words based on digits and capital letters. Example: #10YearChallenge $\rightarrow$ 10 Year Challenge.

4. **Handling of Word/Phrase Contractions:** Created a word/phrase contractions dictionary which contains around 250 words. By using this dictionary mapped the unusual words/phrases to proper English words. Examples: gng $\rightarrow$ going, ASAP $\rightarrow$ as soon as possible.

5. **Handling of Elongated Words:** By using regex converted the elongated words to standard English words. Examples: Nooooo $\rightarrow$ No, suuuppperrr $\rightarrow$ super.

6. **Handling of Imbalanced Datasets:** Data imbalance usually reflects the number of observations per class which is not equally distributed within the dataset. If we look at our dataset, we observe the same kind of problem. To handle this problem we randomly duplicate samples in the minority class called **Random Oversampling** with the help of sklearn[3].

[3]https://github.com/scikit-learn-contrib/imbalanced-learn

4 Our Approach

In this section, we present three different types of deep neural network architectures. We model each task in Memotion Analysis as a multi-class classification problem where given an Internet meme and text description, the model outputs probabilities of it belonging to output classes. The number of output classes will vary for each individual task. The full code of system architecture can be found on GitHub[4] .

4.1 Bidirectional Long Short-Term Memory (BiLSTM) Network with Glove Embeddings

BiLSTM is a recurrent neural architecture (Schuster and Paliwal, 1997) consists of forward and backward LSTM's. The forward LSTM reads the input text in forward order and uses the contextual information from the past. The backward LSTM reads the text description in the reverse order and preserves the contextual information from future. These two LSTM's generate two independent sequence output vectors. We obtain an output vector for each word by concatenating these forward and backward vectors. Figure 2 shows the architecture of BiLSTM with Glove embeddings.

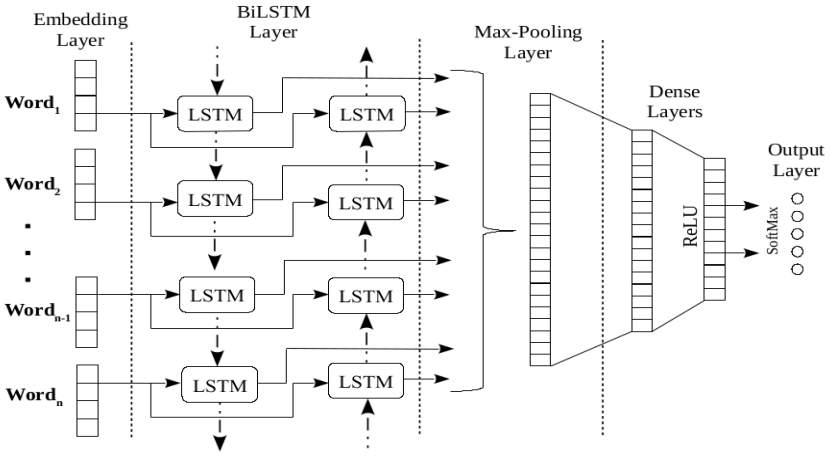

Figure 1: BiLSTM network with Glove embeddings

The input text description is tokenized and fed to the embedding layer. The embedding layer maps the input sequence to a matrix of shape $\mathbf{n} \times \mathbf{d}$, $\mathbf{n}$ is the number of words in the text description, and $\mathbf{d}$ is the dimension of the Glove vector. Output matrix of an embedding layer used to feed the BiLSTM layer. A dropout of 0.2 was applied to the input of the BiLSTM layer and a dropout of 0.1 was used for the output of BiLSTM layer. After BiLSTM layer placed a global max-pooling layer, resulting in an output shape of $\mathbf{d} \times \mathbf{1}$. This output is forwarded to two Dense layers with the activation of Rectifier Linear Unit (ReLU). Dense layer output was passed through a SoftMax layer having $\mathbf{m}$ units. The dimension of $\mathbf{m}$ depends on the number of classes in each task.

4.2 Multimodal Neural Network (MNN) - I

In this section, we demonstrate a multimodal neural network architecture for memotion analysis. In the previous architecture we use the only textual features but in this multimodal architecture join two different data modalities (text and image) for better results.

4.2.1 Image Embeddings

Training a CNN model from scratch can be challenging as a huge amount of data is required and many different models have to be tried before achieving satisfying performances. To avoid this issue, we are using 42-layer deep learning pre-trained network called Inception-v3 that trained to recognize images through the ILSVRC-2012-CLS[5] image classification dataset. The third edition of Google Inception

[4]https://github.com/SunilGundapu/Memotion-Analysis
[5]http://image-net.org/challenges/LSVRC/2015/

network stacks 11 inception modules where each module consists of convolutional filters with rectified linear units and pooling layers.

4.2.2 System Architecture

Figure 3 presents a detailed description of the architecture. On the one side of architecture, the internet meme image, resized to (224,224) is run through the Inception-v3 network, the output is a vector of shape **2048** $\times$ **1** called image embedding, that captures the content of the image. The image embedding is forward to a fully connected layer that converts the vector of shape **2048** $\times$ **1** into the vector of shape **128** $\times$ **1**.

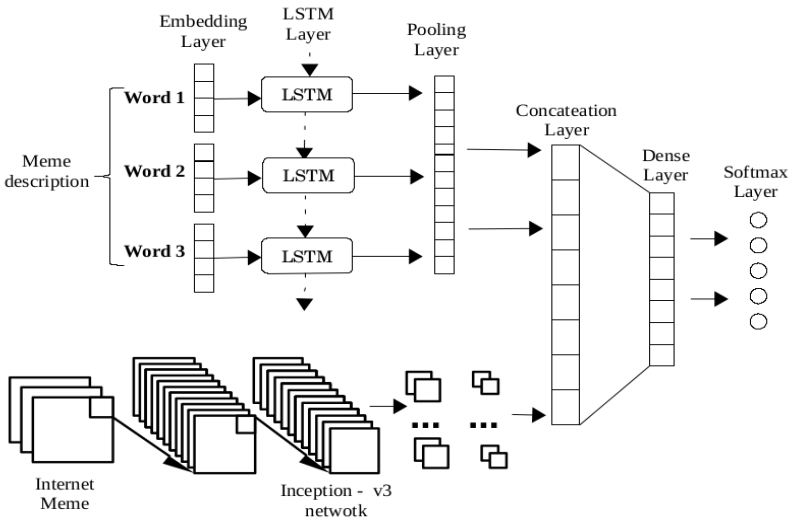

Figure 2: Architecture of Multimodal Neural Network - I

On the other side, the preprocessed meme description is tokenized into a sequence of words and pad with zeros so that each sequence has length **n** (=75). The word embedding layer maps the input sequence to a matrix of shape **n** $\times$ **d**, here **d** (=200) is the dimension of the Glove vector. The output matrix is feed to a LSTM layer, outputs a vector of shape **128** $\times$ **1**.

Further, concatenation of both the textual and image modalities results in a vector of shape **256** $\times$ **1**. In the end, we have a dense layer with the activation of ReLU followed by a softmax layer. The output from the softmax layer is a vector of shape **m** $\times$ **1**. That refers to class probabilities for the **m** classes in a particular task. We used both PyTorch[6] and Keras libraries to build this model.

4.3 Multimodal Neural Network - II

We experiment this model with two word embedding layers and one image embedding layer. In two word embedding layers, one is Sentiment Specific Word Embeddings (SSWE), and the other is Glove embeddings. Gupta et al. (2017) proved that GloVe embeddings capture semantic information and SSWE (Tang et al.,2014) embeddings capture sentiment information in the continuous sequence of words.

An overview of the system architecture can be found in Figure 4. As already said like our approach is built on two components (Word embeddings, Image embeddings). Initially, the input meme image is forwarded to the image embedding layer that generates image feature representations by using the Inception-v3 network. The output dimension of image embeddings is **2048** $\times$ **1** is projected in a vector shape of **256** $\times$ **1** by using a Dense layer. Then we passed the input as a meme description to two LSTM layers using two word embedding matrices with the shape of **n** $\times$ **d**. One layer uses a sentiment specific word embedding, whereas the other layer uses a Glove word embedding. From meme description, these two word embedding layers comprehend sentiment and semantic feature representation and observe

[6]https://pytorch.org/

sequential patterns. These two feature representations are concatenated and passed to a Dense layer with ReLU as an activation function. The output is a vector of dimension $\mathbf{256} \times \mathbf{1}$.

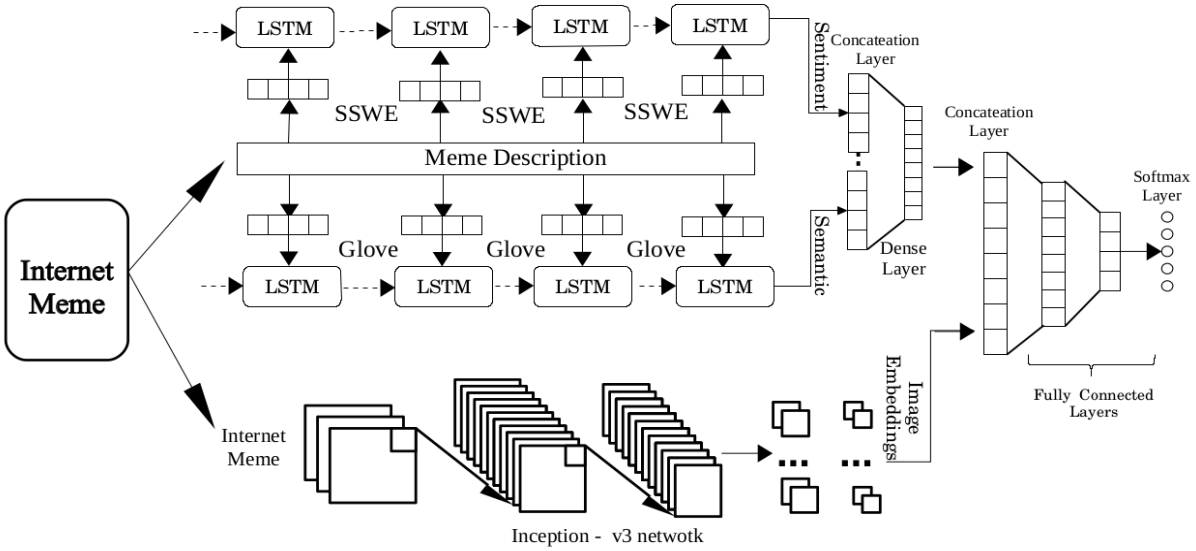

Figure 3: Architecture of Multimodal Neural Network - II

The two modalities (image and text) outputted vectors are then concatenated and the output is a vector of shape $\mathbf{512} \times \mathbf{1}$. The concatenated vector is passed through a fully connected network and then a softmax output layer to give the probability distribution over the different classes in a task.

5 Results

In Table 3, we show the performance and comparison of all the systems presented in Section 4. We observed that Multimodal Neural Network approaches often performed better than the single textual or visual models for all the three tasks. These MNN approaches capture the sentiment along with semantic information in the meme description and extracting the emotion in internet memes.

By comparing the results of the three models, MNN-I outperforms for task A and MNN-II outperforms for tasks B and C in the official macro F1 score. Bi-directional LSTM with Glove vectors did not yield very good results. This system failed to capture the meaning of polysemous word in different contexts and failed to handle short sentences. The MNN systems performed very well on the noisy and imbalanced dataset and failed in some cases such as identifying sarcasm, extracting the emotion (when text covered the expression in memes), and identifying visual cue in the Internet memes.

	Task A		**Task B**		**Task C**	
Model	F1-Score (Macro)	F1-Score (Micro)	F1-Score (Macro)	F1-Score (Micro)	F1-Score (Macro)	F1-Score (Micro)
Baseline	0.2176	0.3077	0.5002	0.5686	0.2483	0.3328
BiLSTM	0.2984	0.3848	0.4713	0.5364	0.2991	0.3236
MNN - I	**0.3391**	**0.4627**	0.4944	0.5846	0.3074	0.3420
MNN - II	0.3261	0.3972	**0.5014**	**0.5896**	**0.3123**	**0.3489**

Table 3: Results of different systems.

To find the right set of hyper-parameters, we used the grid search and development dataset. By considering the dropout (Srivastava et al., 2014), we found the hyper-parameters like the number of LSTM layers, number of epochs, and learning rate. We used the GPU for training all our models.

6 Conclusion

In this paper, we developed a novel multimodal neural network method using deep learning techniques. Our model is constructed by concatenating the textual and image feature representations. We discovered that a combination of text and vision modalities give better predictions than single modalities (text or vision) for memotion analysis. Till now we handled challenges like the unstructured/elongated words, phrase and word contractions, multiple sentences, noisy data, imbalanced datasets, etc. In future work, we will concentrate on problems like short meme descriptions, free word ordering in sentences, more features to identify expressions in memes, sarcasm in descriptions, etc. We would like to explore more deep neural network architectures that are able to capture humor and sarcasm in Internet memes.

References

Chhavi Sharma, Deepesh Bhageria, William Paka, Scott, Srinivas P Y K L, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari and Björn Gambäck. 2020. SemEval-2020 Task 8: Memotion Analysis-The Visuo-Lingual Metaphor! In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain,Sep. Association for Computational Linguistics.

Truong, Quoc-Tuan and Hady W. Lauw. 2017. Visual Sentiment Analysis for Review Images with Item-Oriented and User-Oriented CNN. *Proceedings of the 25th ACM international conference on Multimedia.*

Hu, Anthony and Flaxman, Seth. 2018. *Multimodal Sentiment Analysis To Explore the Structure of Emotions*. 10.1145/3219819.3219853.

Beskow, David and Kumar, Sumeet and Carley, Kathleen. 2019. The Evolution of Political Memes: Detecting and Characterizing Internet Memes with Multi-modal Deep Learning. In *Information Processing & Management*, 57. 102172. 10.1016/j.ipm.2019.102170.

Peirson, V and Tolunay, E. 2018. *Dank Learning: Generating Memes Using Deep Neural Networks*, CoRR, abs/1806.04510.

Johan Bollen, Huina Mao, and Alberto Pepe. 2011. Modeling Public Mood and Emotion: Twitter Sentiment and Socio-Economic Phenomena. In *Proceedings of the Fifth International Conference on Weblogs and Social Media*, Barcelona, Catalonia, Spain, July 17-21, 2011. The AAAI Press.

Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. 2013. Large-scale Visual Sentiment Ontology and Detectors Using Adjective Noun Pairs. In *ACM International Conference on Multimedia (ACM MM)*.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Enhanced Sentiment Learning Using Twitter Hashtags and Smileys. In COLING 2010, *23rd International Conference on Computational Linguistics*, Posters Volume, 23-27 August 2010, Beijing, China. Chinese Information Processing Society of China, 241–249.

Xia Hu, Jiliang Tang, Huiji Gao, and Huan Liu. 2013. Unsupervised sentiment analysis with emotional signals. In *22nd International World Wide Web Conference*, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013. International World Wide Web Conferences Steering Committee / ACM, 607–618.

Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2015. Robust Image Sentiment Analysis Using Progressively Trained and Domain Transferred Deep Networks. In *Association for the Advancement of Artificial Intelligence.*

Sabat, Benet and Ferrer, Cristian and Giró-i-Nieto, Xavier. 2019. *Hate Speech in Pixels: Detection of Offensive Memes towards Automatic Moderation.*

M. Schuster and K. K. Paliwal. 1997. Bidirectional recurrent neural networks. In *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673-2681.

Jeffrey Pennington and Richard Socher and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing*.

C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. 2015. Rethinking the Inception architecture for computer vision, arXiv preprint, 1512.00567, 2015

P. Blandfort, D. U. Patton, W. R. Frey, S. Karaman, S. Bhargava, F.-T. Lee, S. Varia, C. Kedzie, M. B. Gaskell, R. Schifanella, et al. 2019. Multimodal social media analysis for gang violence prevention. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 114-124.

William Yang Wang and Zhenhao Hua. 2014. A semi-parametric gaussian copula regression model for predicting financial risks from earnings calls. In *Proceedings of Association for Computational Linguistics (ACL)*.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky,Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networksfrom overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Umang Gupta, Ankush Chatterjee, Radhakrishnan Srikanth, and Puneet Agrawal. 2017. *A sentiment-and-semantics-based approach for emotion detection in textual conversations*, CoRR, abs/1707.06996.

Chen, Tao and Salaheldeen, Hany and He, Xiangnan and Kan, Min-Yen and Lu, Dongyuan. 2015. *VELDA: Relating an Image Tweet's Text and Images*.

Tang, Duyu and Wei, Furu and Yang, Nan and Zhou, Ming and Liu, Ting and Qin, Bing. 2014. Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification. *52nd Annual Meeting of the Association for Computational Linguistics, ACL*, Proceedings of the Conference. 1. 1555-1565. 10.3115/v1/P14-1146.