ELMo-NB at SemEval-2020 Task 7: Assessing Sense of Humor in Edited News Headlines Using ELMo and NB

Enas Khwaileh
Dept. of Computer Science
Jordan University of Science
and Technology-Irbid, Jordan
*e*khwaileh18@cit.just.edu.jo

Muntaha Al-as'ad
Dept. of Computer Science
Jordan University of Science
and Technology-Irbid, Jordan
*m*aalasaad17@cit.just.edu.jo

Abstract

In this paper, we present our submission for SemEval-2020 competition subtask 1 in Task 7 (Hossain et al., 2020a): Assessing Humor in Edited News Headlines. The task consists of estimating the hilariousness of news headlines that have been modified manually by humans using micro-edit changes to make them funny. Our approach is constructed to improve on a couple of aspects; preprocessing with an emphasis on humor sense detection, using embeddings from state-of-the-art language model (ELMo), and ensembling the results came up with using machine learning model Naïve Bayes (NB) with a deep learning pretrained models. ELMo-NB participation has scored (0.5642) on the competition leader board, where results were measured by Root Mean Squared Error (RMSE).

1 Introduction

Checking the degree of sentence sense of Humor through understanding and analyzing humans natural language and by connecting the text to an intelligent system is considered a critical task (Rastogi et al., 2020). Hence Expressing the readers and writers opinions can increase several emotions, we still need to expand the positive texts and establish a way for analyzing it (Salminen et al., 2020). Humor is considered a great way to the reader, it is resembling therapy (Ziabari and Treur, 2020).

Producing machines that can determine whether the sentence contains some degree of sense of humor or not is gaining a great attention recently (Abdullah and Shaikh, 2018). Since social media is taking over most of people's daily life routines , the culture and environment affect the content greatly (Downey et al., 2006)(Zhao et al., 2020). Numerous factors have brought increasing attention to real-life tasks such as text classification (Howard and Ruder, 2018) (Conneau et al., 2016) (Al-Omari et al., 2020) and other text analysis like pun classification (Diao et al., 2020). Moreover, (Miller et al., 2020) came up with an idea to detect the tweets humorousness using Gaussian Process.

Several Natural Language Processing (NLP) applications and tools are proliferating recently (Kumar and Garg, 2020), (Hirschberg and Manning, 2015), especially with the rise of Deep Learning (DL) and Machine Learning (ML) enhancements (Duerr and Ramdeen, 2017) (Young et al., 2018) (Gardner et al., 2018). One of NLP state of the art approaches is ELMo language preprocessing model as a pretrained model on general NLP tasks of language modeling (Reimers and Gurevych, 2019). ELMo can be fine-tuned on specific tasks like next word prediction (Siddiqui and Hassan, 2019), translation (Li and Chen, 2019) or question answering (McCann et al., 2018) and semantic text (Al-Asa'd et al., 2019). Another NLP state of the art is BERT (Peters et al., 2018).

SemEval-2020 competition in Task 7 [1] has 313 as a total number of participants. The goal of this task is to assess humor in news headlines that have been modified using short edits to make them funny. There are two subtasks as follows: Sub-task 1 (Funniness Estimation): regression problem, the goal is to assign a funniness grade to an edited headline between [0-3], where the systems will be ranked by Root Mean Squared Error (RMSE). Sub-task 2 (Funnier of the Two): a classification problem, given two different

[1] https://competitions.codalab.org/competitions/20970

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 1001–1007
Barcelona, Spain (Online), December 12, 2020.

edited versions of the same headline. The goal of this sub task is to predict which is the funnier of the two. Systems will be ranked by prediction accuracy. Participant "oyx" scored the first place in both subtasks, scoring RMSE (0.50157) for subtask 1, and accuracy (0.67237) for subtask 2.

In this paper, we have experimented ELMo with NB regression to predict the mean funniness of the edited headline in subtask 1. The main target of this task is discovering the atomic change and the tipping between the original and the humorous sentence. As a result, we need to rank the sentences between [0-3]: scores (0) it is not funny, Slightly Funny score (1), moderately funny score (2), and very funny score (3). The micro-editing made to the sentence to make them funny is defined as replacing a noun by a different noun phrase, an entity with different nouns and a verb with a different verb as in (Kanakaraj and Guddeti, 2015). ElMo with NB is showing an amazing performance in predicting the value of humor in the news headline.

The rest of this paper is presented as follows: Sections 2 overviews related work. Section 3 describes the methodology proposed in this paper. Sections 4 overviews results and discusses the most important findings of some experiments and models evaluation. Finally, Section 5 concludes this research and provides possibilities for future work.

2 Literature Reviews

Creating a model that is able to judge the sensitivity to be humorist or not is still a critical task. Several Machine Learning (ML) and Deep Learning (DL) approaches are recommended strongly for working on detecting humorist sentences. For example, using BERT DL pretarined model has a significant role in detecting sentiments and emotions in text (Al-Omari et al., 2019). A team (Mao and Liu, 2019) participated in the HAHA 2019 task used BERT as a bi-directional representation and Fine-tuned pretrain dataset. They obtained the output layer after training the model with the Mean Squad Error (MSE). Other researchers (Potash et al., 2017) added a new task called shared task between the first and second approaches to explore humour. They focused on experimentally comparing hashtag wars from TV show @midnight. The neural network-based system recorded the higher rank.

Researchers (Joshi et al., 2016) proposed a new approach to detect sarcasm. In their experiment, they created a dataset based on quotes GoodReads website, which is one of the largest sites for reading book recommendations. They used word embedding with four types LSA, GloVe, Dependency-based, and Word2Vec. A similar task is what researchers (Hossain et al., 2020b) did recently, where they created a competitive game called Funlines. The users can edit the news headlines. The new sentence has some degree for the sense of humour; they set a method to define the funlines and organizing the sentences to categorize (fun, interactive, collaborative, rewarding, and educational). The classification improvements used to check the performance with and without this dataset augmentation. They were showing that using BERT gave much better results than using LSTM with GloVe word vectors as a benchmark results. The application provides useful feedback to users, to improve their ability to learn and upgrade the level of humourist sentence. In this way the newly generated dataset is performing better.

3 Methodology

3.1 Dataset Preparing and Cleansing

The dataset in this paper is obtained from subtask 1 in Task7-SemEval-2020 competition [2] (Hossain et al., 2019). The researchers collected dataset from the Reddit website related to news headlines. The number of headlines on the train (9652), dev (2419), and test (3025). The teams are asked to predict the mean funniest of each edited headline. In our proposed model, we replace the target word instead of the word between the tags $< \ / >$ in the original headline in both train and dev datasets in terms of predicting the mean funniest value. Then, we replace some of the abbreviations in the data, such as "he's " to "he is" applied in on all dataset as shown in **Table 1**. To make this dataset more understandable, useful and ready fit in any models, we have applied a set of preprocessing techniques like converting the data to

[2] https://competitions.codalab.org/competitions/20970

lower case, remove stemming, stop words, tokenization, punctuation marks, common & rare words, and lemmatization.

Original Headline	New Headline	Grade
appar first iran israel **engag** militarili	appar first iran israel **slap** militarili	0.4
told week ago flynn misl **vice** presid	told week ago flynn misl **school** presid	0
franc hunt citizen join **isi** without trial iraq	franc hunt citizen join **twin** without trial iraq	0.2
john kerri get presidenti fever might **challeng** 2020	john kerri get presidenti fever might **snuggl** 2020	2.6

Table 1: Sample data from the training set

3.2 Word Embeddings

We have used different pre-trained word embeddings to convert each word in the input into a vector representation of 300-Dimensional word vectors. The most popular NLP pretrained models are ELMo and BERT systems. We have used ELMo as the main pretrained system for our submission.

ELMo is a pre-trained model developed by Matthew Peters in 2017 and available on TensorFlow hub. This model is a contextualized deep model, which means it looks at the whole sentence before putting the embedding for the words. The ELMo is a novel technique that assigns each word vectors or embedding based on the context and used Bidirectional LSTM idea. In other words, it applies the forward and backward on each word and concatenates the two values at each layer as shown in the below figure. ELMo can deal with different NLP tasks like question answering, named entity extraction and sentiment analysis as shown **Figure 4**.

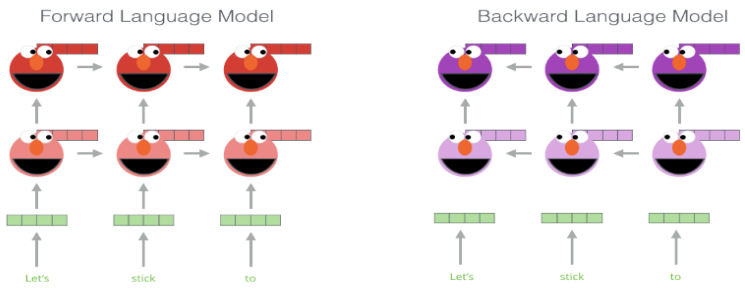

Figure 1: Forward and Backward Techniques [3]

BERT is a language model developed by Google in 2018 and trained on large datasets like Wikipedia. This model is performed on NLP tasks like sentiment and emotion analysis (Sun et al., 2019), and question answering (Yang et al., 2019). BERT converts the words into vectors or embeddings based on the context and uses the transformer method. It is a deeply bidirectional way, which means from right to left and left to right. The transformers contains encoder (read the dataset) and decoder (produce the prediction task). Through examples training, it uses two strategies which are Masked Language Model (MLM), and Next Sentence Prediction (NSP). The MLM, works by replacing 15% of words by masking each word, and try to predict these words based on the non-mask words. While the NSP, works by learning the relationship between the two sentences and produce a label in terms of the second sentence is the next sentence on not based on the meaning between them. As shown in **Figure 2**, BERT applies some of the operations on the dataset before reading it: 1) add [CLS] at the beginning of the sentence and [Sep] at the end of each sentence, 2) Apply Token Embeddings, 3) Sentence Embeddings, 4) Transformer positional Embeddings.

3.3 Model Evaluation Metrics

We have used the Root Mean Square Error (RMSE) value to measure the performance. To calculate RMSE, we need first to calculate Mean Square Error (MSE). So, we take the difference for each Observed (O_i) and Predicted value (P_i) and take the difference squared. Then, we divide the sum of all the values

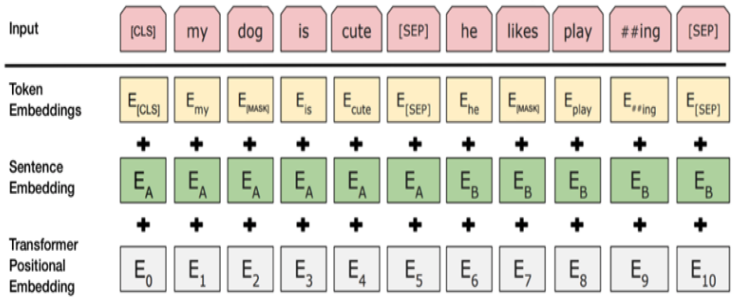

Figure 2: Architecture of the BERT model [4]

by the number of observations to get the MSE value. Finally, we take the root of MSE to get RMSE value.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(P_i - O_i)^2}{(n)}} \tag{1}$$

4 Experimentation and setup

Traditional ML algorithms are being widely used to make predictions based on data. In this paper, ELMo with NB (Chen et al., 2019) is implemented as the proposed model. In our experiments, we compared the performance of the proposed model verses ELMo with Bagging NB and BERT performances. We starts with replacing one word for each original sentence by the requested new word. Next, we find the level of humor in the sentence between [0-3] scale. Different perspectives are experimented to modify and identify the humorist sentences.

According to the performance measures, the results showed that proposed model overrides both BNB and BERT in solving the problem of humor evaluation. The proposed model achieved an RMSE of 0.5642, BERT achieved an RMSE of 0.5747, while BNB an RMSE of 0.5682 as you can see in **Figure 3**

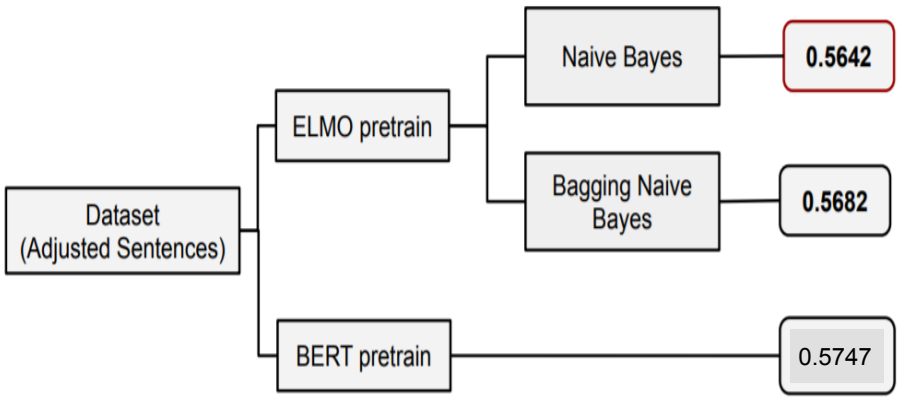

Figure 3: The performance of ELMo-NB, ELMo-BNB, and BERT models

In ML, the NB belongs to ”Probabilistic Classifiers” family based on the Bayes theorem. The main idea of NB is finding a relationship between features using **Equation 2**, which represents the relationship given class label (Y) and dependent feature vector (X)

$$p(Y|X) = \frac{P(X|Y) * P(Y)}{P(X)} \tag{2}$$

Figure. 4 illustrates the general framework of the ELMo-NB model with the dataset. The NB regression is unique of its kind, it is known as the best according to the running time, high accuracy and features handling since it deals with the features as an independent member, so the decision taken is not affected by an absence of some features. Although we have a large set of data and a large number of records, NB is still giving the best RMSE over all experimented regressions.

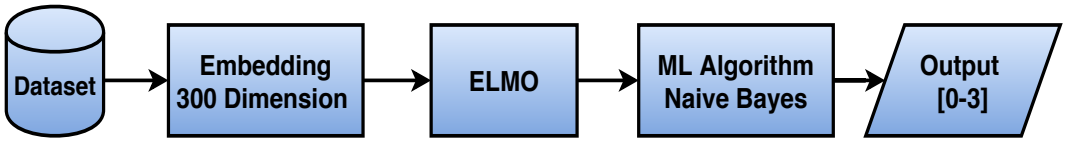

Figure 4: Work-Flow of ELMo-NB Model

Regarding Bagging Naïve Bayes (BNB), the basic concept of bagging is to build new models using the same regression and dataset variance. The concept of bagging is based on taking the dataset to be chosen more than one time and for each time it is running the NB model. Therefore, it is allowed for records to appear in several runs. As we discussed earlier, the NB prediction gives great performances with unbalanced and independent feature variables. We tried to run this kind of NB using the bagging algorithm to measure the regression behavior and make it conductible for comparison.

The third experiments is with DL, we have found that Recurrent Neural Network (RNN) is a well known architecture for NLP. It is proper to handle inputs of different lengths in order to its structure, so RNN serves us well in finding assessing humour in edited news headlines. In our experiment, we uses BERT as a standalone model.

5 Conclusion and Future Work

Through this challenge (SemEval-2020 competition subtask 1 in Task 7), we uses Embeddings from Language Models (ELMo) with the Naïve Bayes (NB) model as the primary baseline. The main focus is using the best text manipulation algorithm, where we recommend using ELMo for it is usefulness and its ability to generate "contextualized" word embeddings. Our trial is to use the ELMo pretrained and then let the ML models make the prediction. Where the NB with ELMo recorded lowest RMSE. For future work, we want to use different ML and DL models with the dataset. Also, we plan to use different-dimensional and pre-trained embedding. Using the XLNET, with well-spotted parameters could be useful too.

Acknowledgement

We would like to extend our sincere thanks to Dr. Malak Abdullah for her efforts and support. In order to finish this work, we had a lot of straight directions and advice from her, during the fall semester, 2019.

References

Malak Abdullah and Samira Shaikh. 2018. Teamuncc at semeval-2018 task 1: Emotion detection in english and arabic tweets using deep learning. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 350–357.

Muntaha Al-Asa'd, Nour Al-Khdour, Mutaz Bni Younes, Enas Khwaileh, Mahmoud Hammad, and AL-Smadi Mohammad. 2019. Question to question similarity analysis using morphological, syntactic, semantic, and lexical features. In *2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)*, pages 1–6. IEEE.

Hani Al-Omari, Malak Abdullah, and Nabeel Bassam. 2019. Emodet at semeval-2019 task 3: Emotion detection in text using deep learning. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 200–204.

Hani Al-Omari, Malak A Abdullah, and Samira Shaikh. 2020. Emodet2: Emotion detection in english textual dialogue using bert and bilstm models. In *2020 11th International Conference on Information and Communication Systems (ICICS)*, pages 226–232. IEEE.

Wei Chen, Xusheng Yan, Zhou Zhao, Haoyuan Hong, Dieu Tien Bui, and Biswajeet Pradhan. 2019. Spatial prediction of landslide susceptibility using data mining-based kernel logistic regression, naive bayes and rbfnetwork models for the long county area (china). *Bulletin of Engineering Geology and the Environment*, 78(1):247–266.

Alexis Conneau, Holger Schwenk, Loïc Barrault, and Yann Lecun. 2016. Very deep convolutional networks for text classification. *arXiv preprint arXiv:1606.01781*.

Yufeng Diao, Hongfei Lin, Liang Yang, Xiaochao Fan, Di Wu, and Kan Xu. 2020. Homographic pun location using multi-dimensional semantic relationships. *Soft Computing*, pages 1–11.

Mark O Downey, Nick K Dokoozlian, and Mark P Krstic. 2006. Cultural practice and environmental impacts on the flavonoid composition of grapes and wine: a review of recent research. *American Journal of Enology and Viticulture*, 57(3):257–268.

Ruth Duerr and Sarah Ramdeen. 2017. Natural language processing (nlp), machine learning (ml), and semantics in polar science. *AGUFM*, 2017:IN13D–03.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. Allennlp: A deep semantic natural language processing platform. *arXiv preprint arXiv:1803.07640*.

Julia Hirschberg and Christopher D Manning. 2015. Advances in natural language processing. *Science*, 349(6245):261–266.

Nabil Hossain, John Krumm, and Michael Gamon. 2019. " president vows to cut¡ taxes¿ hair": Dataset and analysis of creative text editing for humorous headlines. *arXiv preprint arXiv:1906.00274*.

Nabil Hossain, John Krumm, Michael Gamon, and Henry Kautz. 2020a. Semeval-2020 Task 7: Assessing humor in edited news headlines. In *Proceedings of International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain.

Nabil Hossain, John Krumm, Tanvir Sajed, and Henry Kautz. 2020b. Stimulating creativity with funlines: A case study of humor generation in headlines. *arXiv preprint arXiv:2002.02031*.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Aditya Joshi, Vaibhav Tripathi, Kevin Patel, Pushpak Bhattacharyya, and Mark Carman. 2016. Are word embedding-based features useful for sarcasm detection? *arXiv preprint arXiv:1610.00883*.

Monisha Kanakaraj and Ram Mohana Reddy Guddeti. 2015. Performance analysis of ensemble methods on twitter sentiment analysis using nlp techniques. In *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015)*, pages 169–170. IEEE.

Akshi Kumar and Geetanjali Garg. 2020. The multifaceted concept of context in sentiment analysis. In *Cognitive Informatics and Soft Computing*, pages 413–421. Springer.

Hanji Li and Haiqing Chen. 2019. Human vs. ai: An assessment of the translation quality between translators and machine translation. *International Journal of Translation, Interpretation, and Applied Linguistics (IJTIAL)*, 1(1):43–54.

Jihang Mao and Wanli Liu. 2019. A bert-based approach for automatic humor detection and scoring. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019). CEUR Workshop Proceedings, CEUR-WS, Bilbao, Spain (9 2019)*.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*.

Tristan Miller, Erik-Lân Do Dinh, Edwin Simpson, and Iryna Gurevych. 2020. Predicting the humorousness of tweets using gaussian process preference learning. *Procesamiento del Lenguaje Natural*, 64:37–44.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Peter Potash, Alexey Romanov, and Anna Rumshisky. 2017. Semeval-2017 task 6:# hashtagwars: Learning a sense of humor. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 49–57.

Rohit Rastogi, DK Chaturvedi, Santosh Satya, Navneet Arora, Piyush Trivedi, Akshay Kr Singh, Amit Kr Sharma, and Ambuj Singh. 2020. Intelligent analysis for personality detection on various indicators by clinical reliable psychological tth and stress surveys. In *Computational Intelligence in Pattern Recognition*, pages 127–143. Springer.

Nils Reimers and Iryna Gurevych. 2019. Alternative weighting schemes for elmo embeddings. *arXiv preprint arXiv:1904.02954*.

Joni Salminen, Maximilian Hopf, Shammur A Chowdhury, Soon-gyo Jung, Hind Almerekhi, and Bernard J Jansen. 2020. Developing an online hate classifier for multiple social media platforms. *Human-centric Computing and Information Sciences*, 10(1):1.

M Farhan Siddiqui and M Hassan. 2019. Effective word prediction in urdu language using stochastic model. *Sukkur IBA Journal of Computing and Mathematical Sciences*, 2(2):38–46.

Chi Sun, Luyao Huang, and Xipeng Qiu. 2019. Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence. *arXiv preprint arXiv:1903.09588*.

Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019. End-to-end open-domain question answering with bertserini. *arXiv preprint arXiv:1902.01718*.

Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. 2018. Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine*, 13(3):55–75.

Yunpeng Zhao, Mattia Prosperi, Tianchen Lyu, Yi Guo, and Jing Bian. 2020. Integrating crowdsourcing and active learning for classification of work-life events from tweets. *arXiv preprint arXiv:2003.12139*.

S Sahand Mohammadi Ziabari and Jan Treur. 2020. An adaptive cognitive temporal-causal network model of a mindfulness therapy based on humor. In *Information Systems and Neuroscience*, pages 189–201. Springer.