Factuality Classification Using BERT Embeddings and Support Vector Machines

Biswarup Ray, Avishek Garain

Department of Computer Science and Engineering,
Jadavpur University,
Kolkata-700032, West Bengal, India

Abstract

For any topic, its factuality can be defined as the category that determines the status of events with certainty of presentation of them. The first edition of the FACT task mainly focused on determination of the factuality of verb based events. The present edition is aimed at identifying noun based events and determine the factuality of all events be it verbs or nouns. We have participated in Subtask-1 of FACT 2020 task which is to automatically propose a factual tag for each event in the text. In this paper we have presented a method which extracts various features like BERT embeddings, Word2Vec embeddings and TF-IDF (Term Frequency-Inverse Document Frequency) scores of commonly recurring words, along with other manually extracted features as input features and passes them through a SVM (Support Vector Machine) classifier for classification purposes. Our system has achieved a f1-score of 36.6% and accuracy of 59.9% which is quite satisfactory relative to performance of other systems.

Keywords

BERT, embeddings, Factuality, SVM, Word2Vec

1. Introduction

With the advent of ever-growing craze for Internet and increasing popularlity of social media platforms, there has been a noticeable exponential growth of user-generated content, rumors in these platforms. Every second a new news, a new writeup, a new post is either created or shared. Now it might be a real fact or some fake content created and shared as a hate speech propaganda or some other selfish motives. Classification of factuality of such content is highly important to decrease negativity over the platforms as well as preventing any danger at personal level that might be immediate result of defaming caused by fake news. Recently a tech giant like Facebook lost 7.2 billion USD at one go by losing trust of giant companies like Verizon Communications Inc., Hershey Co. and Coca Cola. The companies Verizon Communications Inc. and Hershey Co. have stopped social media ads after critics said that Facebook has failed to sufficiently police hate speech and disinformation on the platform. Coca-Cola Co. said it would pause all paid advertising on all social media platforms for at least 30 days. Such is the importance of Factuality Analysis these days. However, identification of the factual status of events early is a complex task with unavailabilty of enough evidence such as responses and fact checking sites. Making

Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)
EMAIL: raybiswarup9@gmail.com (B. Ray); avishekgarain@gmail.com (A. Garain)
ORCID: 0000-0001-6225-3343 (A. Garain)

the fact-verification pipeline is quite challenging, despite the recent progress in natural language processing, databases and information retrieval [1]. Many prior studies had begun by manual inspection of tweet messages in the training datasets to come up with an initial curated list of word features. It was found that these words could be categorized into meaningful groups which have been useful in identifying an author's certainty in journalism, detecting disagreement in online dialogue and determining veracity of rumors [2, 3, 4] .
The verification of facts in this task, that is, the commitment of the source to the truth-value of the event is not in context of real word problems but is just assessed on basis of the way these facts are presented by the annotator. In view of this perspective, the task can be conceived as a core methodology for other tasks ranging from fake-news to fact-checking, paving way for future tasks like comparison of what is narrated in the text (fact tagging) to what is happening in the world (fact-checking and fake-news). Here, we have designed a system which extracts various word based features making use of BERT pretained vectors along with other manually extracted features as input features and classifies them into corresponding classes using a SVM (Support Vector Machine) classifier.
The rest of the paper has been organized as follows. Previous works related to this task have been briefly described in Section 2. Section 3 describes the data on which the task was performed. The methodology followed is described in Section 4. This is followed by the results and concluding remarks in Section 5 and 6 respectively.

2. Related Works

In 2019, the FACT shared task (Factuality Annotation and Classification Task), was hosted by the First Iberian Languages Evaluation Forum (IberLEF) [5]. The corpus contained Spanish texts with more than 5,000 events classified as F (Fact), CF (Counterfact), U (Undefined). The corpus was divided in two subcorpora: the training corpus (80%), and the testing corpus (20%). Many teams participated and proposed different system designs for the task.

Premjith et al.[6] proposed a system based on word embeddings using a Random Forest classifier. Since the data was unbalanced in nature, the implementation assigned a higher weight to the (CF) label to improve the prediction of the less frequent categories. The model obtained an accuracy of 72.1% and a Macro F1-score of 0.561 when tested.

Giudice [7] proposed a system with a character-level convolutional recurrent neural network. It took advantage of tokenization to classify individual words within the text. Each word was represented as a fixed-size list of vectors. An event flag was added to indicate whether the word represented an event or not. In the final step a dense layer was applied to get, classification for each word among one of the three classes.The system obtained an accuracy of 63.5% and a Macro F1-score of 0.554 when tested.

For this task, Mao [8] chose BERT-Base, multilingual cased model. The corpus was divided into two parts, Uruguayan texts and Spanish texts for the training task. The training and the prediction for the categories for each subcorpus was made separately for the two models. Finally, both outputs were combined in order to create the final result. The model obtained an accuracy of 62.2% and a Macro F1-score of 0.489 when tested.

Another team macro128 (Pastorini) utilized SentencePiece tokenizer in its pre-processing phase. Pre-trained BERT language model was used having one end layer classifying each token among the three possible categories. Since each and every words were not initially classified, each token was randomly assigned to a category.The model was initially trained without the classification layer, until convergence, for a maximum of 100 epochs. The entire model was then trained to converge for a maximum of 100 epochs using F1 measure for early stopping. The system obtained an accuracy of 57.9% and a Macro F1-score of 0.362 when tested.

3. Data

The corpus that has been used for this task contains Spanish texts with approximately 6,300 events classified into respective categories of factuality. The texts belong to the journalistic register and most of them are from the political sections from Spanish and Uruguayan newspaper. The three possible categories established in the dataset are:

- Fact (F): current and past situations in the world that are presented as real.
- Counterfact (CF): current and past situations that the writer presents as not having happened.
- Undefined (U): Possibilities, future situations, predictions, hypothesis and other options.

Words have been tagged with these factuality tags. All these words are related to some event and proper placement in the sentences converts the sentences into events. We have divided the dataset in the ratio 80:20 for training and validation purposes respectively. The distribution of data instances is given in Table 1.

Label	Train	Validation
F	2777	694
CF	223	55
U	1155	289
All	4155	1038

Table 1: Distribution of the labels in dataset

4. Methodology

For FACT Task, our method is based on a SVM classifier with BERT embeddings, Word2Vec embeddings and TF-IDF (Term Frequency-Inverse Document Frequency) of commonly recurring words as input features. The system architecture for the proposed method is represented in Fig.1

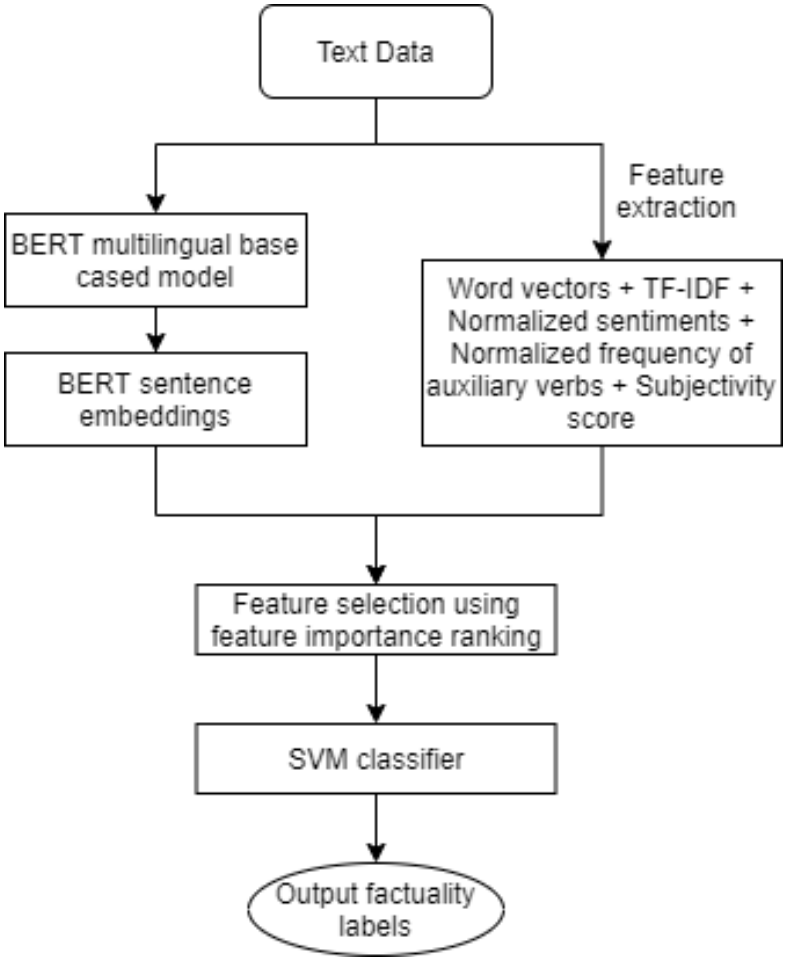

Figure 1: The architecture for the proposed method for the factuality classification task.

4.1. Feature Extraction

For the BERT [9] embeddings firstly, for a given sentence, first token representation is obtained from the pre-trained BERT model using the help of a WordPiece model (cased). The pre-trained BERT models are pre-trained on large corpus (Wikipedia + BookCorpus). In FACT, we utilised a BERT-Base, multilingual cased model as: First, multilingual model are much better than English-only models for Spanish documents in FACT as the strictly-English models split tokens unavailable in its vocabulary into sub-tokens, which affects the accuracy of the classification task. Also BERT-Large generally outperforms BERT-Base in NLP tasks in English language, BERT-Large versions of multilingual models haven't been published yet.The tokens are indexed and alongwith their segment ids are fed to a BERT model as torch tensors which are termed as the input embeddings. The output given by the final hidden layer has four dimensions, in the order of:

- The number of the layer (13 layers). The first element is the input embeddings, the rest are the outputs of each of BERT's 12 layers.

- Batch number (number of sentences)
- Word / token number (number of tokens in the sentence)
- Feature number (768 features)

The batch dimension is not needed hence it is removed. The output is then permuted into the desired dimension of $[tokens, layers, features]$. To get a single vector for an entire sentence a simple approach is used by averaging the second to last hidden layer of each token producing a single vector of length 768. The value of these vectors are contextually dependent.
An example of input embeddings for a particular sentence to find the BERT embeddings from the BERT model is shown in Fig.2

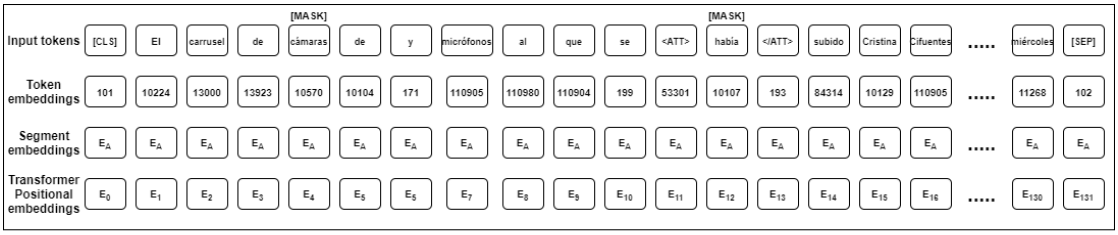

Figure 2: An example containing for a set of input tokens with positional, segment, transformer embedding.

The vector representation of each of the word is produced using the Gensim module. The vectors of each word are the context in which the words appear (Word2Vec). Each of the texts is also converted into numerical vectors using the word vectors produced. Thus, we train the Doc2Vec model by feeding in our text data. By applying this model on the texts, we get the representation vectors.
The manually extracted features thus added to the model for classification apart from the BERT embeddings are:

1. The vector representations obtained using Word2Vec.
2. The TF-IDF for the words that frequently occur in the text are also added to the feature list. The TF computes the number of times a word recurs in the dataset, and IDF computes the relative importance of the word which depends on how many times the word can be found, and are added as features to filter and reduce the size of the final output.
3. Normalized counts of words with positive sentiment, negative sentiment and neutral sentiment in Spanish by dividing with word count of the corresponding sentence. [10]
4. Normalized Frequency of auxiliary verbs like "es", "estaba", "son", "fueron" by dividing by word count of the sentence.
5. Subjectivity score of the text (calculated using predefined libraries). [11]

4.2. Classification

For the classification task the features extracted through the various above mentioned processes are selected by using the feature importance rankings for each feature. To train the SVM (Support Vector Machine) with a linear kernel [12] model the categorical labels present for the data were label encoded into numerical values. The numerical labels along with the selected features were used to train the SVM model. As done in the work [13], class weights were assigned to the SVM model as the dataset is unbalanced and contains a very low quantity of texts having the Counterfact (CF) label.

5. Results

The FACT corpus contains Spanish texts with approximately 6,300 events classified into respective categories of F (Fact), CF (Counterfact), and U (Undefined).
In FACT, the performance is measured against the evaluation corpus using the Macro-F1(%), Macro-Precision(%), Macro-Recall(%), Accuracy(%) and Global accuracy metrics. Macro-F1 is the most important measure for this task. We have presented the results for the unknown test set in Table.2. As shown in Table.2 our proposed method outperformed the FACT baseline in all of the metrics: Macro-F1(%), Macro-Precision(%), Macro-Recall(%), Accuracy(%) and Global accuracy. Using class weights is one of the main reason for the better performance as it ensures lesser misclassification of the label CF present in much lesser quantity than other labels. Also using various features and implementing feature selection for the classification task ensures better performance.

Model	Macro-F1(%)	Macro-Precision(%)	Macro-Recall(%)	Accuracy(%)
Our model	36.6	35.7	39.4	59.9
FACT_baseline	24.6	25.4	25.1	52.4

Table 2: Result Metrics of our system for Subtask-1

However the lower performance of the proposed method may be due to the poor performance of the SVM model as word vectors may be highly non-linearly separable. Also even with class weights the lower number of instances in the CF class in the training data is a reason for misclassification.

6. Conclusion

We have presented our system that we have used for participating in the FACT: Factuality Analysis and Classification Task in IberLEF 2020. Considering previous approaches, our approach is a comparatively different approach in terms of architecture as well as methodology of feature extraction. It is a generalized and versatile framework and has showed satisfactory performance among all participating systems during the FACT 2020 evaluations. In future works, we will use

an ensemble of different classification models to increase the performance and we will explore some practical applications such as fake news detection, etc.

References

[1] A. Vlachos, S. Riedel, Fact checking: Task definition and dataset construction, in: Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science, 2014, pp. 18–22.

[2] A. Misra, M. Walker, Topic independent identification of agreement and disagreement in social media dialogue, arXiv preprint arXiv:1709.00661 (2017).

[3] U. D. Reichel, P. Lendvai, Veracity computing from lexical cues and perceived certainty trends, arXiv preprint arXiv:1611.02590 (2016).

[4] S. Soni, T. Mitra, E. Gilbert, J. Eisenstein, Modeling factuality judgments in social media text, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2014, pp. 415–420.

[5] A. Rosa, I. Castellón, L. Chiruzzo, H. Curell, M. Etcheverry, A. Fernández, G. Vázquez, D. Wonsever, Overview of fact at iberlef 2019: Factuality analysis and classification task, 2019.

[6] B. Premjith, K. P. Soman, P. Poornachandran, Amrita_cen@fact: Factuality identification in spanish text, in: IberLEF@SEPLN, 2019.

[7] V. Giudice, Aspie96 at fact (iberlef 2019): Factuality classification in spanish texts with character-level convolutional rnn and tokenization, in: IberLEF@SEPLN, 2019.

[8] J. Mao, W. Liu, Factuality classification using the pre-trained language representation model bert, in: IberLEF@SEPLN, 2019.

[9] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018. arXiv:1810.04805.

[10] A. Garain, S. K. Mahata, Sentiment analysis at sepln (tass)-2019: Sentiment analysis at tweet level using deep learning (2019).

[11] A. Garain, Humor analysis based on human annotation (haha)-2019: Humor analysis at tweet level using deep learning (2019).

[12] C. Cortes, V. Vapnik, Support-vector networks, Machine Learning 20 (1995) 273–297. URL: https://doi.org/10.1007/BF00994018. doi:10.1007/BF00994018.

[13] A. Garain, A. Basu, The titans at semeval-2019 task 5: Detection of hate speech against immigrants and women in twitter, in: Proceedings of the 13th International Workshop on Semantic Evaluation, 2019, pp. 494–497.

[14] A. Rosá, L. Alonso, I. Castellón, L. Chiruzzo, H. Curell, A. Fernández, S. Góngora, M. Malcuori, G. Vázquez, D. Wonsever, Overview of fact at iberlef 2020: Events detection and classification (2020).

[15] J. Mao, W. Liu, Factuality classification using the pre-trained language representation model bert., in: IberLEF@ SEPLN, 2019, pp. 126–131.

[16] A. Garain, A. Basu, The titans at semeval-2019 task 6: Offensive language identification, categorization and target identification, in: Proceedings of the 13th International Workshop on Semantic Evaluation, 2019, pp. 759–762.

[17] A. Garain, S. K. Mahata, S. Dutta, Normalization of numeronyms using nlp techniques, in: 2020 IEEE Calcutta Conference (CALCON), 2020, pp. 7–9.