Amrita_CEN@FACT: Factuality Identification in Spanish Text

Bhavukam Premjith[1][0000−0003−1188−1838], Kutti Padannayil Soman[1], and Prabaharan Poornachandran[2]

[1] Center for Computational Engineering and Networking (CEN)
Amrita School of Engineering, Coimbatore
Amrita Vishwa Vidyapeetham, India
prem.jb@gmail.com
[2] Center for Cybersecurity Systems and Networks
Amrita School of Engineering, Amritapuri
Amrita Vishwa Vidyapeetham, India

Abstract. This paper presents the description of the system used by the team Amrita_CEN for the shared task on FACT (Factuality Analysis and Classification Task) at IberLEF2019 (Iberian Languages Evaluation Forum) workshop. The goal of the task was to automatically annotate an event with its factuality status. Factuality status is categorized into three as Fact, Counter Fact and Undefined. Our proposed system predicts the factuality of an event with a prediction accuracy of 72.1%. The classification model for this task was trained using Random Forest classifier which uses word embedding of the events as input features. The word embedding of an event was generated by using Word2vec algorithm. Random Forest was implemented by giving higher weights to minority classes and lesser weights to majority classes so that more number of elements in the minority class will be predicted precisely.

Keywords: Factuality classification · Spanish text · Word2vec · Weighted Random Forest.

1 Introduction

In Natural Language Understanding (NLU), identification of the characteristics of an event has greater significance. Factuality is one of the principal characteristics of an event [1]. The factuality of an event shows the happening of an event in the past or present. It also helps to know whether an event has not yet happened or it is just an illusion of a writer. However, in day-to-day conversations, factuality of an event often expressed in a vague manner and thereby leaves some degree of ambiguity in the context of occurrence. This uncertainty is ubiquitous in all sorts of situations [2] and hence makes the automatic prediction a tough task. The accurate prediction of the factuality of an event is vital in deducing various knowledge related to that event. The understanding of an event when it is identified as a fact is different from the reasoning about that event when it is recognized as a counter fact or an undefined event [3]. Therefore, the proper categorization of events into its actual factuality is very important and is widely used in many applications such as temporal organization of events, sentiment analysis and opinion detection and question answering [3]. Despite its considerable importance in NLU, this task is underexplored especially in Spanish. Wonsever et al. [4] and Wonsever et.al [5] put significant effort in developing an annotated corpora as well as automatic models for the analysis and classification of event factuality in Spanish texts. But, still this research is in its fledgling stages.

Factuality Analysis and Classification Task (FACT) is a shared task organized as part of IberLEF2019 for recognizing the factuality of an event in a Spanish text. In this task, events are tagged with three labels - Fact, Counter Fact and Undefined. The goal of the task was to encourage the research in this field through the development of computational models for the automatic prediction of the factuality of an event. Our team, Amrita_CEN developed a machine learning model which used Word2vec [6], [7] for extracting features from the event words and Random Forest algorithm [9] [10] for classification. We used a weighted Random Forest algorithm [11] for classifying events because the number of instances in Counter Fact class was very less compared to other two classes (Fact and Undefined). The performance of the model was evaluated using F1-score (macro averaging) and accuracy score and our model achieved the scores of 0.561 and 72.1% respectively.

2 Description of the task

The objective of the shared task "FACT: Factuality Analysis and Classification Task" was to classify the events expressed in Spanish texts as Fact, Counter Fact or Undefined by considering their factuality status into account. The events which belong to the category of "Facts" are those events which are expressed as real in either past or current circumstances. The "Counter Facts" events are those which never happened so far whereas the "Undefined" events are neither Fact nor Counter fact because the author was uncertain about the existence of such events.

The training data contains 56 Spanish texts of which 4,343 events were labelled as Fact(F) or Counter Fact(CF) or Undefined(U). Among these labelled events, the number of distinct event names was 2,053. 1,428 words in the vocabulary occurred only once and the word with highest frequency of occurrence was "es" with 171 occurrences. The word "ha" also appeared more than 100 times in the list with 162 appearances and is visible in the Figure 1 which shows the frequency of occurrences of top 50 words and their counts in the training data.

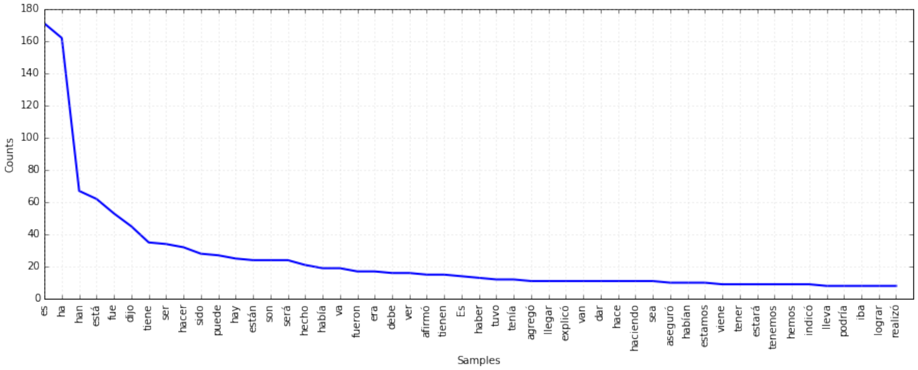

Fig. 1. Plot of the most frequently occurred 50 events and their counts in the training data

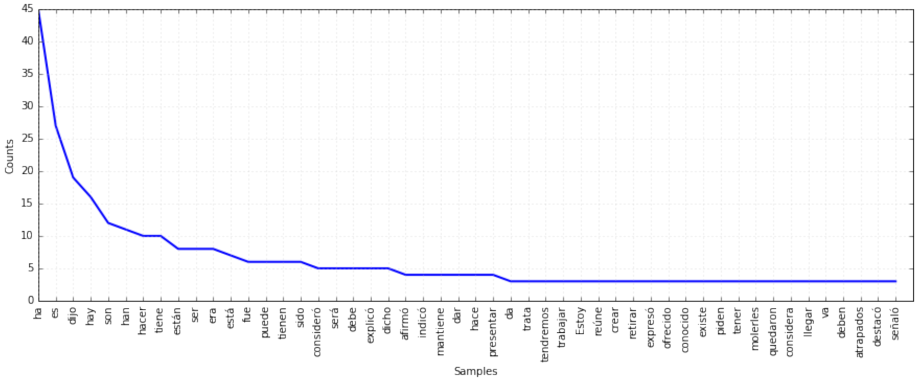

Fig. 2. Plot of the most frequently occurred 50 events and their counts in the test data

In the test data, there were 15 Spanish texts with 1,075 unidentified events. Out of these 1,075 events, 715 were unique words and 580 of these unique words were found only once in the dataset. Among these words, only 8 words existed more than 10 times. Another interesting trend we found in both dataset was that, out of 20 most frequently occurred events in the training set, 16 events were present in the list of top 20 events in the test data with highest frequency of occurrence. This trend can be observed in Figures 1 and 2.

Table 1. Statistics of the dataset.

Class label	Number of instances
Counter Fact	255 (5.87%)
Fact	2917 (67.16%)
Undefined	1171 (26.96%)

3 System description

The training as well as testing dataset for the task were given as an XML file. The first task was to extract features for the events from the data and represent them in terms of vectors. Word embedding algorithms were used for this representation. We tried both Word2vec[3] and FastText[4] [8] algorithms with varying embedding dimensions and observed that Word2vec performs better than FastText in the classification. Various parameters used for building the Word2vec model is given in the Table 2. We also observed that the embedding dimension beyond 300 didn't produce a significant change in the performance of the classifiers.

Table 2. Word2vec parameters.

Parameter	Parameter value
Algorithm	Word2vec
Embedding size	300
Window size	1
Minimum count	1
Aplha (Initial learning rate)	0.025
Workers	4
Sg	0 (Continuos Bag-of-Words)

We used various classification algorithms defined in the Scikit-learn (version = 0.20.1) python library [12] for modeling the data. Random Forest (RF), Decision Tree (DT) and Naïve Bayes (NB) classifiers achieved reasonable accuracy.

[3] https://radimrehurek.com/gensim/models/word2vec.html

[4] https://radimrehurek.com/gensim/models/fasttext.html

The performance of Support Vector Machine (SVM) was poor and hence we concluded that, the word vectors were highly non-linearly separable. Among all the classifiers, Random Forest achieved the best training accuracy. When the model was trained with the word vectors as features, it was found that most of the data points in Class "CF" were classified as "F". The less number of instances in the CF class in the training data was the reason for this misclassification. Confusion matrix obtained for this modeling is shown in Figure 3.

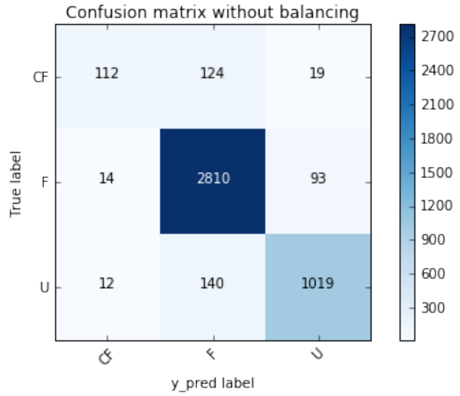

Fig. 3. Confusion matrix for Random Forest classifier without using any balancing scheme to deal with minority classes

Even though the model gave a good training accuracy of 90.74%, we decided to use a weighted Random Forest classifier for training with the motivation to increase the classification accuracy of minority class "CF". From, the confusion matrix in Figure 3, it is clear that only 43.92% of "CF" class was correctly classified as "CF". This may affect the performance of the system when tested with unknown samples. Therefore, we applied a weighted Random Forest classifier. It attained an overall accuracy of 88.46% which is relatively less than the unweighted Random Forest accuracy. However, when the class-wise classification was analysed, most of the instances (71.76%) in "CF" were classified as "CF" itself. The confusion matrix for the weighted Random Forest is shown in Figure 4. The weights used for "CF", "F" and "U" were 5.68, 0.5 and 1.24 respectively which was computed using the Equation 1.

$$weights = \frac{number\_of\_instances}{number\_of\_classes \times bin\_count\,(y)} \tag{1}$$

Where $bin\_count\,(y)$ is the number of instances in each classes.

The training performance of both unweighted and weighted Random Forest is described in Table 3. We used accuracy score, macro-Precision, macro-Recall and macro-F1-score evaluating the training model.

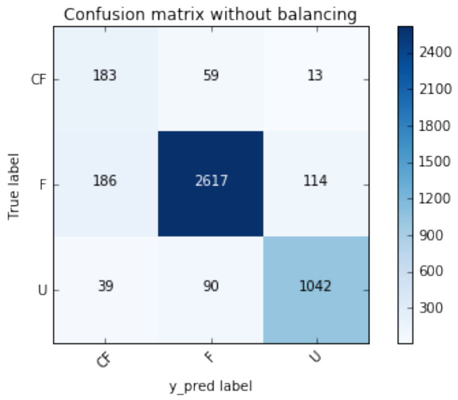

Fig. 4. Confusion matrix for the weighted Random Forest algorithm

Table 3. Training scores of the classification.

Metrics	Unweighted RF	Weighted RF
Accuracy score	90.74	88.46
Precision	0.8756	0.7620
Recall	0.7576	0.8349
F1-score	0.7978	0.7879

The subset of parameters which were used to build the Random Forest classifier is presented in Table 4.

Table 4. Parameters of Random Forest classifier.

Parameter	Parameter value
Splitting criteria	Gini
Class weight	Balanced
Number of trees in the forest	50
Minimum number of data points at the leaf	1

The shared task organizers used macro-F1-score and accuracy for evaluating the predictions of class labels for the test data. Six teams participated in the contest including the baseline system, of which our system scored the highest both in terms of macro-F1 and accuracy. The results are shown in Table 5.

4 Conclusion

The identification of the factuality of an event is an important task in Natural Language Understanding (NLU). The factuality of an event acts as an additional feature for many Natural Language Processing (NLP) applications like question

Table 5. Results of the FACT shared task.

Participant	macro-F1	Accuracy
premjithb	0.561	72.1
Aspie96	0.554	63.5
jimblair	0.489	62.2
macro128	0.362	57.9
fact (baseline)	0.340	52.4
garain	0.301	51.2

answering and opinion detection. Automatic identification of an event as Fact or Counter Fact or Undefined is a multi-class classification problem. In this paper, we used weighted Random Forest classifier for learning the patterns in the data which was represented using Word2vec algorithm. The model obtained an accuracy of 72.1 and an F1-score (macro) of 0.561 when tested with a set of unknown events.

References

1. Rudinger, Rachel, Aaron Steven White, and Benjamin Van Durme, Neural models of factuality, arXiv preprint arXiv:1804.02472 (2018)
2. Saur, Roser, and James Pustejovsky, Are you sure that this happened? assessing the factuality degree of events in text, Computational Linguistics, 38(2), 261-299 (2012)
3. Saur, Roser, A factuality profiler for eventualities in text, Unverffentlichte Dissertation, Brandeis University. Zugriff auf http://www.cs.brandeis.edu/ roser/pubs/sauriDiss (2008)
4. Wonsever, Dina, Marisa Malcuori, and Aiala Ros Furman, Factividad de los eventos referidos en textos, Reportes Tcnicos 09-12 (2009)
5. Wonsever, Dina, Aiala Ros, and Marisa Malcuori, Factuality Annotation and Learning in Spanish Texts, LREC (2016)
6. Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781 (2013)
7. Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean, Distributed representations of words and phrases and their compositionality, Advances in neural information processing systems, 3111–3119 (2013)
8. Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov, Enriching word vectors with subword information, Transactions of the Association for Computational Linguistics, 5, 135-146 (2017)
9. Liaw, Andy, and Matthew Wiener, Classification and regression by randomForest, R news, 3;2(3), 18-22 (2002)
10. Premjith, B., Soman, K.P., Kumar, M.A. and Ratnam, D.J, Embedding Linguistic Features in Word Embedding for Preposition Sense Disambiguation in English-Malayalam Machine Translation Context, Recent Advances in Computational Intelligence, Springer, Cham, 341-370 (2019)
11. Xie, Yaya, Xiu Li, E. W. T. Ngai, and Weiyun Ying, Customer churn prediction using improved balanced random forests, Expert Systems with Applications, Elsevier, 36(3), 5445–5449 (2009)

12. Pedregosa, Fabian and Varoquaux, Gaël and Gramfort, Alexandre and Michel, Vincent and Thirion, Bertrand and Grisel, Olivier and Blondel, Mathieu and Prettenhofer, Peter and Weiss, Ron and Dubourg, Vincent and others, Scikit-learn: Machine learning in Python, Journal of machine learning research, 12, 2825–2830 (2011)