Sarcasm and sentiment detection in Arabic language: A hybrid approach combining embeddings and rule-based features

Kamel Gaanoun[1]
[1]SI2M Lab
National Institute of Statistics
and Applied Economics
Rabat, Morocco
kamel.gaanoun@gmail.com

Imade Benelallam[1,2]
[1]SI2M Lab
National Institute of Statistics
and Applied Economics
[2]AIOX Labs
Rabat, Morocco
i.benelallam@insea.ac.ma

Abstract

This paper presents the ArabicProcessors team's system designed for sarcasm (subtask 1) and sentiment (subtask 2) detection shared task. We created a hybrid system by combining rule-based features and both static and dynamic embeddings using transformers and deep learning. The system's architecture is an ensemble of Gaussian Naive Bayes, MarBERT and Mazajak embedding. This process scored an F1-sarcastic score of 51% on sarcasm and an F1-PN of 71% for sentiment detection.

1 Introduction

Automatic sarcasm detection is a sub-discipline of sentiment analysis. This process, however, presents specific challenges (Nigam and Hurst, 2006; Pang and Lee, 2008) related to the fact that sarcasm can present opposite polarities between surface and intended sentiment. Macmillan English dictionary show the complexity of this phenomenon by defining it as *"the activity of saying or writing the opposite of what you mean, or of speaking in a way intended to make someone else feel stupid or show them that you are angry"*. This imposes additional layers to go beyond the methods used so far for sentiment analysis; since in addition to the sentimental polarity of the text, it will also be necessary to consider, among others, pragmatic factors (Kreuz and Caucci, 2007), based on context and background knowledge (Hazarika et al., 2018), behavior modelling (Rajadesingan et al., 2015; Agrawal et al., 2020) and the linguistic theory of incongruity (Joshi et al., 2015).

This observation is even more accentuated when it comes to the treatment of Arabic sources. Undoubtedly, the scarcity of data available in Arabic relative to English and the different Arabic dialects, for example, makes the detection of sarcasm even more complicated. Should sarcasm in the Arabic language be treated the same way it is treated in the English language? (Karoui et al., 2017) is sarcasm also present in Arabic the same way in its different dialects? This complexity is reflected in the scarcity of studies on automatic Arabic sarcasm detection, such that first works Karoui et al. (2017) and Ghanem et al. (2019) dealing with this problem, were recently published in 2017 and 2019.

Shared task on sarcasm and sentiment detection in Arabic (Abu Farha et al., 2021) is a new step in exploring this problem, aimed at both sarcasm detection (Subtask 1) and the analysis of sentimental polarity (Subtask 2). In this paper, we present our contribution to this challenge, by illustrating a novel approach based on a hybrid system that combines rule-based features and both static and dynamic embeddings using transformers and deep learning.

In the next sections, we describe used data in Section 2, describe our system in Section 3, present and discuss our results in Section 4, and finally summarize our work in Section 5.

2 Data

To train our models, we were provided with the same dataset for both subtasks; ArSarcasm-v2 (Abu Farha et al., 2021). It comprises of four variables, namely; the text of the tweet, a boolean label for sarcasm, sentimental polarity label (Positive, neutral and negative) as well as a dialect label.

In the cause of this shared task, a total of 12,548 tweets were made available to us, of which 2168 are labeled as sarcastic tweets, representing 17.3% of all tweets (Figure 1). Therefore, it is an unbalanced dataset, as is the case for the majority of sarcasm and irony datasets (Joshi et al., 2017).

With regards to sentiment distribution, approximately half were neutral, while 37% were negative and 17% positive (Figure 1). It is noteworthy, how-

Proceedings of the Sixth Arabic Natural Language Processing Workshop, pages 351–356
Kyiv, Ukraine (Virtual), April 19, 2021.

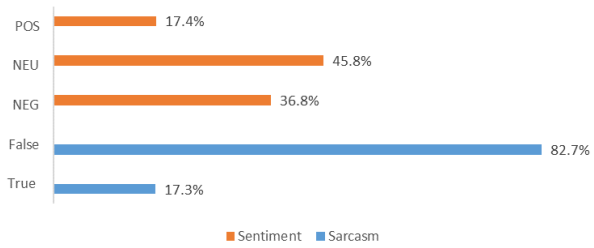

Figure 1: Sarcasm and Sentiment distribution.

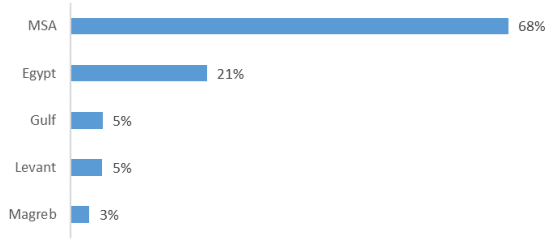

Figure 2: MSA and dialects distribution.

ever, that the vast majority of tweets were written in MSA (68%) and dominantly in Egyptian dialect (Figure2).

With respect to the sentimental polarity label, we established overriding amount of negative sentiments among sarcastic tweets (Figure 3). Moreover, it was noted that the Egyptian dialect, which has the same provided dataset general distribution with non-sarcastic tweets, it represents the same share as MSA among sarcastic tweets.

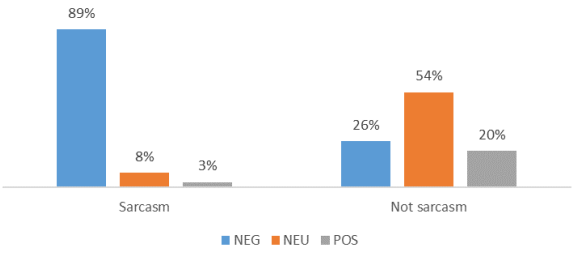

Figure 3: Sarcasm per sentiment polarity.

In analysing the most common words among sarcastic tweets, we noted two main groups, political personalities, or entities or events (Sissi, Trump, Morsi,«انقلاب»-coup d'etat-, Hilary, etc.) and punctuation dominated by quotation marks, colons, replies (@) and exclamation mark.

We further observed that averagely, there are more words in sarcastic tweets (18 words) than in the non-sarcastic ones (15 words). Sarcastic tweets are, however, on an average, shorter with 5.6 characters as against 6.6 characters for other tweets. The same applies to stopwords, which are more frequent among sarcastic tweets, reaching an average of 2.1 stopwords as against 1.7 for the rest of the tweets.

The most common emoticons in sarcastic and non-sarcastic tweets include the "face with tears of joy", the "new moon face", and the "pensive face". The emoticon showing a "face with tears of joy" sums up to 32% of sarcastic tweets, far ahead of the rest of the emoticons, while it sums up to only 7% of non-sarcastic tweets.

Regarding the interjection «هههه» with n repetition of the letter «هـ», equivalent to «haha», we found out that it is more present in sarcastic tweets with 4% compared to other tweets not exceeding 0.5%.

These findings are a strong indication of the potential efficiency of the use of features extracted from tweets. Indeed, this would be a factor for sarcasm detection. In that regard, we analyzed the features presented in the following section.

2.1 Features set

In addition to features extraction, the only preprocessing step was the removal of urls. We took inspiration from, and extended the features used by Karoui et al. (2017) to create features that have been classified into three groups:

(1) Surface features:

- Punctuation marks
- Quotation mark
- Sequence of exclamation or question marks
- Combination of question and exclamation marks
- Hashtags : presence and count
- Emoticons: total count and count of "face with tears of joy"
- Replies (@)
- Number of repeated words (stopwords excluded)
- Word count, char count, char per word, number of stopwords
- Diacritics percentage and boolean diacritics more than 45% to detect Quranic texts
- Url existence
- Interjections : «هههه ،هههههه»
- Sarcasm words, see the list in Appendix A
- Opposition words, see the list in Appendix A

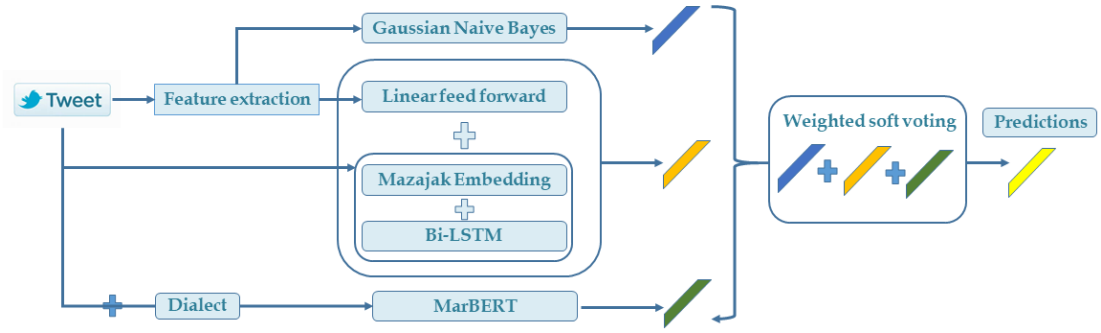

Figure 4: System architecture

Illustration inspired from Kang and Gwak (2020).

- Politic entities, see the list in Appendix A

(2) Sentiment features:

In order to get word sentiment scores, we combined two publicly available Arabic sentiment lexicons, namely *The Arabic Emoticon Lexicon* and *The Arabic Hashtag Lexicon (dialectal)*[1]. We also used *Emoji Sentiment Ranking 1.0* (Kralj Novak et al., 2015) to extract emoticons sentiment

- Total sentiment score and average sentiment score
- Positive/negative sentiment words count and percent
- Reversed sentiment score between the beginning and end of the tweet (Boolean), i.e. is the sentiment of the first part of the tweet the opposite of the last part's sentiment?
- Emoticons sentiment: total sentiment score, average sentiment score, positive/negative (Boolean)
- Emoticons count

(3) Intensifiers: As «إطلاقا، دهشة، مذهل», complete list is available in Appendix A

3 System

We adopted an ensemble hybrid system (Figure 4) combining extracted features and embeddings. The final prediction is the result of a weighted soft voting of the following models:

- Gaussian Naïve Bayes, with scaled features.
- MarBERT *"a large-scale pre-trained masked language model focused on both Dialectal Arabic (DA) and MSA"* (Abdul-Mageed et al., 2020). We first concatenated the tweets' text with their dialect label before feeding them to the model.
  - MarBERT, AraBERT (Antoun et al., 2020),and Arabic-BERT (Safaya et al., 2020) were compared. The former presented better results.
- Bidirectional long short term memory (Bi-LSTM) using Mazajak embeddings (Abu Farha and Magdy, 2019), concatenated with a fully connected layer with extracted features as input.

A 5-fold stratified cross-validation was used to evaluate and compare models. Also, the weights for the soft voting step were obtained by Bayesian optimization.

We used the hyper-paremeters as shown in Table 1.

GPU	Tesla P100-PCIE-16GB
Language	Python 3.6.9
Main librairies	Hugging Face Transformers 4.4.2, Torch 1.7.0, Sklearn 0.22.2, bayesian-optimization-1.2.0
Bert Hyperparameters	Learning rate: 2e-5, Epochs: 4, Batch size: 64, Embedding maximum length: 125
Bi-LSTM - Mazajak-	Hidden units: 64, Dropout: 0.2, Activation: ReLU, Spatial dropout: 0.3, Epochs: 20, Batch size: 64

Table 1: Used infrastructure and hyperparameters

[1] http://saifmohammad.com/WebPages/ArabicSA.html

4 Results and discussion

We oriented the development of our system towards the use of an ensemble model including MarBERT, Bi-LSTM with Mazajak embedding and Gaussian Naive Bayes. Gaussian Naive Bayes was chosen after comparison with Logistic regression, SVM, XGBoost, LightGBM, and Random forest was done.

Results obtained for individual models on the 5-fold cross validation, and ensembled one are reported in Table 2.

Model	Sarcasm	Sentiment
Gaussian Naive Bayes	43	46
MarBERT	57	71
Bi-LSTM -Mazajak-	46	58
Weighted ensemble	59	73

Table 2: 5-fold cross validation scores (%)
Sarcasm: F1-score for sarcastic class, Sentiment: Macro average of the F-score of the positive and negative classes

We draw attention to the fact that among the ensembled models, MarBERT accounts for the largest share. This confirms the effectiveness of dynamic embeddings and BERT based models.

Results obtained on the official evaluation set are reported in the next subsection.

4.1 Official results

The organizers provided us with an unlabeled TEST dataset, intended for the final evaluation of the system, composed of 3000 tweets as well as their respective dialects. The scores obtained on this dataset was used for system ranking (see Table 3).

Model	Sarcasm	Sentiment
Gaussian Naive Bayes	39.34	32.33
MarBERT	54.17	73.32
Bi-LSTM -Mazajak-	23.15	51.32
Final official results	**50.86**	**71.45**

Table 3: Scores obtained on final evaluation set (%)
Sarcasm: F1-score for sarcastic class, Sentiment: Macro average of the F-score of the positive and negative classes

These scores reflect the complexity of the sarcasm detection when compared to sentiment analysis. Furthermore, our model shows better generalization capacity on sentiments, by loosing only 1.5 points between cross-validation and the final evaluation results. Exploring the evaluation set distribution, we note that 27% of the tweets are labeled as sarcastic compared to only 17% among Train set tweets. Covariate shift is then one possible reason for the generalization issue raised regarding sarcastic tweets.

We also note that the Bi-LSTM is the model suffering the most from this generalization issue by loosing 34 points. This leads us to conclude that this model could be enhanced adopting a better regularization strategy, or even not considering it in the final ensemble model.

Finally the results are confirming the higher accuracy of Bert based model upon other models, and thus going deeper in this direction can lead to better results, by adopting some eventual improvement methods as mentioned in the next subsection.

4.2 Discussion

This shared task presents three main challenges: the few amount of data, the unbalanced labels, and the complexity of the sarcasm detection even for humans. Indeed, with no information about the text context nor the writer profile, this indicates a lack of primordial clues for sarcasm detection. Adding this information could enhance the system's performance. Furthermore, balancing the data is a possible way of improvement by upsampling the minority label.

We could also use a hierarchical system by taking advantage of the efficiency of the sentiment prediction model. Indeed, a first step would be to predict the tweets' sentiment, then use this prediction as a feature for sarcasm prediction, since we have seen that sarcasm is strongly linked to the polarity of the tweet.

Finally, we can still explore other ways to improve BERT-based models by concatenating more information in addition to the dialect.

5 Conclusion

In this paper, we have described our contribution to the WANLP'21 sarcasm and sentiment detection in Arabic subtasks. The system's architecture is an ensemble of Gaussian Naive Bayes, MarBERT and Mazajak embedding. By recording an F1-score of 51% at subtask 1, and 71% at subtask 2, this approach demonstrates the efficiency of our hybrid system in the challenging field of sarcasm and sentiment detection for Arabic texts.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2020. Arbert & marbert: Deep bidirectional transformers for arabic. *arXiv preprint arXiv:2101.01785*.

Ibrahim Abu Farha and Walid Magdy. 2019. Mazajak: An online Arabic sentiment analyser. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 192–198, Florence, Italy. Association for Computational Linguistics.

Ibrahim Abu Farha and Walid Magdy. 2020. From Arabic sentiment analysis to sarcasm detection: The ArSarcasm dataset. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 32–39, Marseille, France. European Language Resource Association.

Ibrahim Abu Farha, Wajdi Zaghouani, and Walid Magdy. 2021. Overview of the wanlp 2021 shared task on sarcasm and sentiment detection in arabic. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*.

Ameeta Agrawal, Aijun An, and Manos Papagelis. 2020. Leveraging transitions of emotions for sarcasm detection. pages 1505–1508.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 9.

Bilal Ghanem, Jihen Karoui, Farah Benamara, Véronique Moriceau, and Paolo Rosso. 2019. Idat at fire2019: Overview of the track on irony detection in arabic tweets. In *Proceedings of the 11th Forum for Information Retrieval Evaluation*, FIRE '19, page 10–13, New York, NY, USA. Association for Computing Machinery.

Devamanyu Hazarika, Soujanya Poria, Sruthi Gorantla, Erik Cambria, Roger Zimmermann, and Rada Mihalcea. 2018. CASCADE: Contextual sarcasm detection in online discussion forums. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1837–1848, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Aditya Joshi, Pushpak Bhattacharyya, and Mark J Carman. 2017. Automatic sarcasm detection: A survey. *ACM Computing Surveys (CSUR)*, 50(5):1–22.

Aditya Joshi, Vinita Sharma, and Pushpak Bhattacharyya. 2015. Harnessing context incongruity for sarcasm detection. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 757–762, Beijing, China. Association for Computational Linguistics.

Jaeyong Kang and Jeonghwan Gwak. 2020. Ensemble learning of lightweight deep learning models using knowledge distillation for image classification. *Mathematics*, 8(10).

Jihen Karoui, Farah Banamara Zitoune, and Veronique Moriceau. 2017. Soukhria: Towards an irony detection system for arabic in social media. *Procedia Computer Science*, 117:161–168.

Petra Kralj Novak, Jasmina Smailović, Borut Sluban, and Igor Mozetič. 2015. Sentiment of emojis. *PLoS ONE*, 10(12):e0144296.

Roger Kreuz and Gina Caucci. 2007. Lexical influences on the perception of sarcasm. In *Proceedings of the Workshop on Computational Approaches to Figurative Language*, pages 1–4, Rochester, New York. Association for Computational Linguistics.

Kamal Nigam and Matthew Hurst. 2006. Towards a robust metric of polarity. In *Computing Attitude and Affect in Text: Theory and Applications*, pages 265–279. Springer.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2:1–135.

Ashwin Rajadesingan, Reza Zafarani, and Huan Liu. 2015. Sarcasm detection on twitter: A behavioral modeling approach. *WSDM 2015 - Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, pages 97–106.

Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. 2020. Kuisail at semeval-2020 task 12: Bert-cnn for offensive speech identification in social media. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

A Appendices

- Sarcasm words: استهبال ، إستهبال، إستهبل، استهبل، مسخرة، سخر، إستهزاء، اِستهزاء ، تهكم، طنز، طنز

- Opposition words: لكن ،مع أنه ، مع ان ، مع انه ، مع ذلك ،في حين أن،في حين ان، من ناحية ثانية،من ناحية تانية،من ناحيه ثانية،من ناحيه تانية،من ناحية ثانيه،من ناحية تانيه، مع إنو،بالرغم،غير أن، غير أنه،فيما أن،فيما ان،غير ان، غير انه،بينما،بيد ان،بيد أن،إلا أن،الا ان،إلا ان،الا أن،بخلاف،

- Politic entities:سوريا، سيسي، مرسي، ترامب، ، مصر، الانقلاب، انقلاب، داعش، الاخوان ، هيلاري، الشعب، حزب، كلنتون، السعودية ، بوتن، انتخاب، انتخب، حرب، بشار، ثورة،سياسة

- Intensifiers list: إطلاقا، دهشة، مذهل، اطلاق، مرارة، دموي، تماما، جنون، ميت، مخيف هائل، خصوصا، استثنائي، إفراط، ابعد حد، غير عادي، خيالي، للغاية مخيف، لا يصدق، حرفيا، غاضب،بقوة، شائن، ثمين، كبير، جذريا، حقيقة،لافت للنظر، وبالتالي، قليلا، ممتاز، بدرجة عليا، رهيب، رائع، جدا، لا يصدق،حقيقي