IDAT at FIRE2019: Overview of the Track on Irony Detection in Arabic Tweets

Bilal Ghanem
PRHLT Research Center,
Univ. Politècnica de València
Spain
bigha@doctor.upv.es

Jihen Karoui
AUSY R&D
Paris
France
jkaroui@ausy.fr

Farah Benamara
IRIT-CNRS
Université de Toulouse
France
farah.benamara@irit.fr

Véronique Moriceau
IRIT-CNRS
Université de Toulouse
France
veronique.moriceau@irit.fr

Paolo Rosso
PRHLT Research Center,
Univ. Politècnica de València
Spain
prosso@dsic.upv.es

ABSTRACT

This overview paper describes the first shared task on irony detection for the Arabic language. The task consists of a binary classification of tweets as ironic or not using a dataset composed of 5,030 Arabic tweets about different political issues and events related to the Middle East and the Maghreb. Tweets in our dataset are written in Modern Standard Arabic but also in different Arabic language varieties including Egypt, Gulf, Levantine and Maghrebi dialects. Eighteen teams registered to the task among which ten submitted their runs. The methods of participants ranged from feature-based to neural networks using either classical machine learning techniques or ensemble methods. The best performing system achieved F-score value of 0.844, showing that classical feature-based models outperform the neural ones.

CCS Concepts

•**Artificial Intelligence** → **Natural Language Processing;**

Keywords

Irony detection ; Arabic language ; Social media

1. AIMS AND MOTIVATIONS

Irony is a complex linguistic phenomenon widely studied in philosophy and linguistics. In the standard pragmatic model [10], irony is viewed as an apparent violation of the maxim of quality, stating that the speaker does not say what he believes to be false. In this model, when one ironically utters P, one conversationally implicates its opposite, that is $Not(P)$. For example, if one says to his colleague "Congratulation for your great presentation" after a disappointing talk. This vision has been criticized by several authors who pointed out that logical opposition between what is said and what is intended captures only one type of irony. To overcome this deficiency, different theories have been proposed to deal with the multi-dimensional nature of opposition. Among them, we cite [29, 6, 12, 2, 31] that respectively describe irony in terms of echoic mention, allusional pretense, predicate and propositional negations, relevant inappropriateness, and implicit display. Irony is used here as an umbrella term that covers a variety of other figurative devices such as satire, parody, and sarcasm [6, 9].

Irony detection has gained relevance recently, due to its importance in various NLP applications such as sentiment analysis, hate speech detection, author profiling, fake news detection, and crisis management (e.g., terrorist attacks, public disorder). For example, recent studies on irony show that the performances of sentiment analysis systems drastically decrease when applied to ironic texts [3, 7, 14, 32]. This is mainly due to the complexity of ironic contents that make use of figures of speech to convey non-literal meaning.

Most state of the art approaches to irony detection consider social media data and tweets in particular, as specific hashtags (#irony, #sarcasm) are often employed by users to help readers understand their ironic contents. These hashtags are used as gold labels to detect irony in a supervised learning setting. Most related work concern English [13] with some efforts in French [17], Portuguese [4], Italian [8], Dutch [21], Hindi [30] and Arabic [16]. Also, many shared tasks on irony have been proposed, such as SemEval 2018 task 3 for English [13], DEFT 2017 for French [3], IronITA@Evalita 2018 for Italian [5], and IroSvA@IberLEF-2019 for Spanish variants [26] (from Spain, Cuba and Mexico). As far as we know, this is the first shared task on irony for the Arabic language and will be a good opportunity to compare the performances of Arabic irony detection to those reported in recent shared tasks in other languages.

2. PROCESSING ARABIC TWEETS

Computational processing of the Arabic language has received a great attention in the literature for over a twenty

FIRE '19 December 12–15, 2019, Kolkata, India

DOI: 10.1145/3368567.3368585

years[1]. Several resources and tools have been built to deal with Arabic nonconcatenative morphology and Arabic syntax [22]. There is also a wide range of Arabic NLP (ANLP) applications including question answering [24], automatic translation [28] and sentiment analysis [20]. However, the field of ANLP is still very vacant at the layer of pragmatics. As far as we know, the sole effort towards Arabic irony detection was done by Karoui et al. [16] who proposed a supervised approach to detecting ironic tweets. The performance of several groups of features (like surface, sentiment, shifter and contextual features) have been assessed achieving an accuracy of 72.36% on a dataset composed of 3,466 tweets among which 50% were ironic.

3. DATA AND ANNOTATION

The collected dataset is composed of tweets posted on Twitter during the years 2011 to 2018 about different political issues and events related to the Middle East and the Maghreb. A set of predefined keywords is used to collect tweets, which targeted specific political figures which were the subject of the Arab spring and the presidential elections of Egypt and US. From these retrieved tweets, we selected those containing or not the Arabic ironic hashtags #سخرية, #مسخرة, #تهكم, #استهزاء[2].

The collection process resulted in a set of 22,318 tweets (6,809 ironic tweets and 15,509 are not). These tweets are written using standard (formal) and different Arabic language varieties: Egypt, Gulf, Levantine and Maghrebi dialects.

We took a sample of 6,000 tweets to investigate the validity of using the original tweets labels. This sample consists of 3,000 tweets as ironic and 3,000 as not. It has been manually annotated by two Arabic native speakers. We measured the inter-annotator agreement using Cohen's Kappa and obtained a score of 76%, which referred to a strong agreement. This score is inline with agreements reported in annotating irony in tweets from other languages such as English (e.g., $Kappa = 0.72$ in the SemEval-2018 task 3) [13]), French ($Kappa = 0.69$ in the Deft 2017 shared task [3]), and Spanish ($Kappa = 0.67$ in IroSvA@IberLEF [26]). The disagreement between the annotators is due to two main factors: (1) the misinterpretation or comprehension of some dialectal words; and (2) the lack of context knowledge to understand the ironic sense of the tweet.

The distribution of tweets in the final IDAT dataset is given in Table 1. The class distribution (ironic vs. non ironic) is quite similar, with a proportion of ironic tweets of about 52% in both train and test.

Table 1: Tweet distribution the IDAT dataset.

	#Ironic	# Not-Ironic	Total
Train	2,091	1,933	4,024
Test	523	483	1,006
Total	2,614	2,416	5,030

[1]For a detailed description of Modern Standard Arabic and an overview of Arabic NLP, see [11].

[2]All of these words are synonyms meaning "Irony".

4. TASK DESCRIPTION AND EVALUATION MEASURES

The task consists of classifying a tweet as ironic or not ironic. The IDAT training set has been released on May 31^{th} and participants had one month and a half to train their systems. The test was then released on July 15^{th} and each participant was allowed to submit a maximum of 3 runs within 10 days.

Participating systems were evaluated using standard evaluation metrics, namely accuracy and F-score as follows. Official rankings is given according to F-score.

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total number of instances}} \quad (1)$$

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{True Negatives}} \quad (2)$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (3)$$

$$\text{F-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

5. METHODS OF PARTICIPANTS AND RESULTS

Eighteen teams have registered to the shared task among which **ten** submitted their runs. Participants were from 7 different countries: Algeria, Canada, Egypt, India, Jordan, Pakistan and UK. All team members were from public entities (Universities, Research Centers).

Participants used either traditional machine learning approaches (SVM, Multimodel Naive Bayes, Logistic Regression, Ensemble models) and/or deep learning methods (CNN, RNN, LSTM, Gated Recurrent Unit, Transformers). The tweet contents are represented by traditional bag of words (*YOLO* [19], *SSN-NLP* [18]), n-grams (*BENHA* [25]) eventually weighted with TF-IDF (*BENHA, YOLO, PITS* [15]), emotion features (*Kinmokusu* [23], *PITS, YOLO*) and word embeddings (*Kinmokusu, Ali_Allaith* [1], RGCL [27], *Amrita_CEN, Tha3aroon*). Embeddings were obtained using different models such as Word2Vec, FastText and BERT.

Table 2 presents participants' results for each submitted run. The results are ranked according to the F-score. For each system, best run is given in bold font. We also compare the results with those of two baselines: SVM with unigrams term frequency (BOW) and a random baseline.

6. CONCLUSION

This paper overviews the first shared task on irony detection in Arabic social media that aims at classifying a tweet as ironic or not. 18 teams participated in the task and a total of 10 teams submitted their runs. Systems have been trained on a nearly balanced dataset composed of ironic and non ironic tweets about political issues that raised between 2011 and 2018 in the Middle East and Maghreb. The dataset has been manually annotated and inter-annotator agreement was good ($Kappa = 0.76$). The methods proposed by participants ranged from traditional features-based approaches relying on bag of words features to neural methods using pretrained word embeddings. Several neural architectures were tested such as CNN, LSTM and Transformers. Ensemble methods have also been used. The best system achieved an

Table 2: Participants results ranked in terms of F-score. Baselines are in italic font.

Team	F-score			
	1	2	3	Rank
YOLO	**0.844**	0.833	0.823	**1**
Chiyu_Zhang_UBC	0.819	**0.824**	0.811	**2**
BENHA	0.816	0.811	**0.821**	**3**
RGCL	**0.818**	0.804	0.816	4
Ali_Allaith	**0.817**	0.794	–	5
SSN_NLP	**0.816**	0.793	0.709	6
PITS	**0.807**	–	–	7
Tha3aroon	**0.794**	0.75	–	8
BOW Baseline	*0.793*			
Kinmokusu	**0.695**	0.687	0.689	9
Amitra_CEN	**0.687**	0.534	0.434	10
Random Baseline	*0.496*			

F-score of 0.844 showing that classical features-based models outperform deep learning methods when applied to the IDAT dataset.

Acknowledgments

This publication was made possible by NPRP grant 9-175-1-033 from the Qatar National Research Fund (a member of Qatar Foundation). The findings achieved herein are solely the responsibility of the last author.

7. REFERENCES

[1] A. Allaith, M. Shahbaz, and M. Alkoli. Neural Network Approach for Irony Detection from Arabic Text on Social Media. In *Proceedings of the IDAT@FIRE2019. In Metha P., Rosso P., Majumder P., Mitra M. (Eds.) Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2019). CEUR Workshop Proceedings.*, 2019.

[2] S. Attardo. Irony as Relevant Inappropriateness. *Journal of Pragmatics*, 32(6):793–826, 2000.

[3] F. Benamara, C. Grouin, J. Karoui, V. Moriceau, and I. Robba. Analyse d'opinion et langage figuratif dans des tweets présentation et résultats du Défi Fouille de Textes DEFT2017. In *Actes de DEFT@TALN2017*, 2017.

[4] P. Carvalho, L. Sarmento, M. J. Silva, and E. D. Oliveira. Clues for Detecting Irony in User-Generated Contents: Oh...!! It's "so easy";-). In *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, pages 53–56. ACM, 2009.

[5] A. T. Cignarella, S. Frenda, V. Basile, C. Bosco, V. Patti, and P. Rosso. Overview of the EVALITA 2018 task on irony detection in italian tweets (ironita). In *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, December 12-13, 2018.*, 2018.

[6] H. H. Clark and R. J. Gerrig. On the Pretense Theory of Irony. *Journal of Experimental Psychology: General*, 113(1):121–126, 1984.

[7] A. Ghosh and D. T. Veale. Fracking Sarcasm using Neural Network. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 161–169, 2016.

[8] A. Gianti, C. Bosco, V. Patti, A. Bolioli, and L. D. Caro. Annotating Irony in a Novel Italian Corpus for Sentiment Analysis. In *Proceedings of the 4th Workshop on Corpora for Research on Emotion Sentiment and Social Signals, Istanbul, Turkey*, pages 1–7, 2012.

[9] R. W. Gibbs. Irony in Talk Among Friends. *Metaphor and symbol*, 15(1-2):5–27, 2000.

[10] H. P. Grice. Logic and Conversation. In P. Cole and J. L. Morgan, editors, *Speech Acts. Syntax and Semantics, Volume 3*, pages 41–58. Academic Press, New York, 1975.

[11] N. Habash. *Introduction to Arabic Natural Language Processing*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2010.

[12] H. Haverkate. A speech act analysis of irony. *Journal of Pragmatics*, 14(1):77 – 109, 1990.

[13] C. V. Hee, E. Lefever, and V. Hoste. SemEval-2018 Task 3: Irony Detection in English Tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT, New Orleans, Louisiana, June 5-6, 2018*, pages 39–50, 2018.

[14] D. I. Hernańdez Farías, V. Patti, and P. Rosso. Irony Detection in Twitter: The Role of Affective Content. *ACM Transactions on Internet Technology (TOIT)*, 16(3):1–24, 2016.

[15] N. Kanwar, R. Kumar, Mundotiya, M. Agarwal, and C. Singh. Emotion based voted classifier for Arabic irony tweet identification. In *Proceedings of the IDAT@FIRE2019. In Metha P., Rosso P., Majumder P., Mitra M. (Eds.) Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2019). CEUR Workshop Proceedings.*, 2019.

[16] J. Karoui, F. Benamara, and V. Moriceau. SOUKHRIA: Towards an Irony Detection System for Arabic in Social Media. In *Third International Conference On Arabic Computational Linguistics, ACLING 2017, November 5-6, 2017, Dubai, United Arab Emirates*, pages 161–168, 2017.

[17] J. Karoui, F. Benamara, V. Moriceau, N. Aussenac-Gilles, and L. H. Belguith. Towards a Contextual Pragmatic Model to Detect Irony in Tweets. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing : Volume 2 short papers*, ACL-IJCNLP'15, pages 644–650, 2015.

[18] S. Kayalvizhi, D. Thenmozhi, B. S. aKumar, and C. Aravindan. SSN_NLP@IDAT-FIRE-2019 : Irony Detection in Arabic Tweets using Deep Learning and Features-based Approaches. In *Proceedings of the IDAT@FIRE2019. In Metha P., Rosso P., Majumder P., Mitra M. (Eds.) Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2019). CEUR Workshop Proceedings.*, 2019.

[19] M. Khalifa and N. Hussein. Ensemble Learning for

Irony Detection in Arabic Tweets. In *Proceedings of the IDAT@FIRE2019. In Metha P., Rosso P., Majumder P., Mitra M. (Eds.) Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2019). CEUR Workshop Proceedings.*, 2019.

[20] S. Kiritchenko, S. M. Mohammad, and M. Salameh. Semeval-2016 task 7: Determining sentiment intensity of english and arabic phrases. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '16, 2016.

[21] C. Liebrecht, F. Kunneman, and B. A. van den. The Perfect Solution for Detecting Sarcasm in Tweets# Not. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 29–37, 2013.

[22] Y. Marton, N. Habash, and O. Rambow. Dependency Parsing of Modern Standard Arabic with Lexical and Inflectional Features. *Computational Linguistics*, 39(1):161–194, 2013.

[23] L. Moudjari1 and K. Akli-Astouati. An embedding-based approach for irony detection in Arabic tweets. In *Proceedings of the IDAT@FIRE2019. In Metha P., Rosso P., Majumder P., Mitra M. (Eds.) Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2019). CEUR Workshop Proceedings.*, 2019.

[24] H. Mozannar, K. E. Hajal, E. Maamary, and H. M. Hajj. Neural arabic question answering. *CoRR*, abs/1906.05394, 2019.

[25] H. A. Nayel, W. Medhat, and M. Rashad. BENHA@IDAT: Improving Irony Detection in Arabic Tweets using Ensemble Approach. In *Proceedings of the IDAT@FIRE2019. In Mehta P., Rosso P., Majumder P., Mitra M. (Eds.) Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2019). CEUR Workshop Proceedings.*, 2019.

[26] R. Ortega-Bueno, F. Rangel, D. Hernández Farıas, P. Rosso, M. Montes-y Gómez, and J. E. Medina Pagola. Overview of the Task on Irony Detection in Spanish Variants. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019), co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019). CEUR-WS. org*, 2019.

[27] T. Ranasinghe, H. Saadany, A. Plum, S. Mandhari, E. Mohamed, C. Orasan, and R. Mitkov. RGCL at IDAT: Deep Learning models for Irony Detection in Arabic Language. In *Proceedings of the IDAT@FIRE2019. In Metha P., Rosso P., Majumder P., Mitra M. (Eds.) Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2019). CEUR Workshop Proceedings.*, 2019.

[28] F. Sadat and E. Mohamed. Improved Arabic-French machine translation through preprocessing schemes and language analysis. In *Proceedings of the Canadian Conference on Artificial Intelligence*, AI, pages 308–314, 2013.

[29] D. Sperber and D. Wilson. Irony and the use-mention distinction. *Radical pragmatics*, 49:295–318, 1981.

[30] S. Swami, A. Khandelwal, V. Singh, S. S. Akhtar, and M. Shrivastava. A Corpus of English-Hindi Code-Mixed Tweets for Sarcasm Detection. 19th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing), 2018.

[31] A. Utsumi. Stylistic and Contextual Effects in Irony Processing. In *Proceedings of the 26th Annual Meeting of the Cognitive Science Society*, pages 1369–1374, 2004.

[32] S. Zhang, X. Zhang, J. Chan, and P. Rosso. Irony detection via sentiment-based transfer learning. *Information Processing Management*, 56(5):1633 – 1644, 2019.