# I2C at SemEval-2022 Task 6: Intended Sarcasm in English using Deep Learning Techniques

**Adrián Moreno Monterde**
Escuela Técnica Superior de Ingeniería
Universidad de Huelva (Spain)
adrian.moreno521@alu.uhu.es

**Laura Vázquez Ramos**
Escuela Técnica Superior de Ingeniería
Universidad de Huelva (Spain)
laura.vazquez005@alu.uhu.es

**Victoria Pachón Álvarez**
Escuela Técnica Superior de Ingeniería
Universidad de Huelva (Spain)
vpachon@dti.uhu.es

**Jacinto Mata Vázquez**
Escuela Técnica Superior de Ingeniería
Universidad de Huelva (Spain)
mata@uhu.es

## Abstract

Sarcasm is often expressed through several verbal and non-verbal cues, e.g., a change of tone, overemphasis in a word, a drawn-out syllable, or a straight looking face. Most of the recent work in sarcasm detection has been carried out on textual data. This paper describes how the problem proposed in *Task 6: Intended Sarcasm Detection in English* (Abu Arfa et al. 2022) has been solved. Specifically, we participated in *Subtask B: a binary multi-label classification task*, where it is necessary to determine whether a tweet belongs to an ironic speech category, if any. Several approaches (classic machine learning and deep learning algorithms) were developed. The final submission consisted of a BERT based model and a macro-F1 score of 0.0699 was obtained.

## 1 Introduction

Existing social media analysis systems are limited by their inability to accurately detect and interpret figurative language. Sarcasm is often used by individuals to express opinions on complex matters and regarding specific targets (Carvalho et al. 2009).

Early computational models for verbal and irony and sarcasm detection have relied on shallow methods exploiting conditional token count regularities. But lexical clues alone are insufficient to discern sarcasm intent. Appreciating the context of expression is critical for this; even for humans (Wallace et al. 2014). Indeed, the exact same sentence can be interpreted as literal or sarcastic, depending on the speaker. Consider the sarcastic tweet in Figure 1 (ignoring for the moment the attached *#sarcasm* hashtag). Without knowing the author's political inclination, it would be difficult to conclude with certainty whether the tweet was intended as sarcastic or not.


Figure 1: An illustrative tweet.

This task is about the binary classification of tweets in English based on the category of ironic speech. As a Multilabel-Classification, a tweet can belong to multiple categories or none. This research is important because social networking sites have changed our lives in recent years. For companies, this social reality has turned into an obligation to set up communication and marketing channels through social networks. Besides, companies require feedback from consumers, through comments or messages that mark the acceptance or rejection of each of the proposals, products or services.

For this competition 2 models/strategies were used:

- The first model is a binary multilabel classifier using Bayesian networks.
- The second model and final submit on the competition consisted of six different binary classifiers using BERT. In other words, one for each evaluable label: sarcasm, irony, satire,



*Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 856 - 861

understatement, overstatement and rhetorical question.

In this task we have analysed the sarcastic behaviour of people on social networks, in this case twitter, and the ways in which people express it. The dataset proposed by the organization was very unbalanced, so we had to apply several data balancing techniques in order to achieve good results.

The rest of this paper is organized as follows: in section 2 we explain the dataset, the meaning of the labels and data distribution. Therefore, we refer to other research that has helped us in our approach to this one. In section 3, several techniques and methods applied to our models to improve their performance and metrics are described. In section 4 we explain the libraries used and their usefulness. Finally, in section 5, the scores obtained with our proposed approaches are presented.

## 2 Background

As mentioned above, this paper is focused on Subtask B: binary multilabel classification. The original dataset has 3467 tweets in English with a maximum size of 280 characters (tweet limit). As a matter of fact, only the first 867 tweets will be useful for our task because the rest of the tweets do not have the labels that we are going to work with. Furthermore, the columns "Unnamed:0", "rephrase" and "sarcastic" have been removed because they are useless for our task perfomance.

For a better analysis and understanding of the multiple labels that we have in our dataset it is important to mention the *ironic speech* exposed in (Leggitt and Gibbs 2000):

1. *Sarcasm*: tweets that contradict the state of affairs and are critical towards an addressee
2. *Irony*: tweets that contradict the state of affairs. but are not obviously critical towards an addressee.
3. *Satire*: tweets that appear to support an addressee but contain underlying disagreement and mocking.
4. *Understatement*: tweets that undermine the importance of the state of affairs they refer to
5. *Overstatement*: tweets that describe the state of affairs in terms that are obviously exaggerated.
6. *Rhetorical question*: tweets that include a question whose invited inference (implication) is obviously contradicting the state of affairs.

In Figure 2, two samples of the dataset with the information used in our approaches can be seen.

**Tweet:** The only thing I got from college is a caffeine addiction
**Sarcasm:** [0]
**Irony:** [1]
**Satire:** [0]
**Understatement:** [0]
**Overstatement:** [0]
**Rhetorical question:** [0]

**Tweet:** do i just blast maneskin to get hyped for my osce or??
**Sarcasm:** [1]
**Irony:** [0]
**Satire:** [0]
**Understatement:** [0]
**Overstatement:** [0]
**Rhetorical question:** [1]

Figure 2: Example of two rows

When it comes to the multiple labels (categories) of sarcasm, Table 1 shows the distribution of the tweets into these categories. As can be seen, the dataset is imbalanced so, in the next section, we explain how the dataset was balanced for a better performance.

| Category | Number of tweets |
|---|---|
| Sarcasm | 713 |
| Irony | 155 |
| Satire | 25 |
| Understatement | 10 |
| Overstatement | 40 |
| Rhetorical question | 101 |

Table 1: Distribution of tweets in each category

This challenge has been approached by different researchers. In (Davidov et al., 2010), experiments with semi-supervised sarcasm identification on a Twitter dataset (5.9 million tweets) were carried out using 50 Twitter tags and 15 emojis as sentiment labels. They used a 5-fold cross validation on their classifier getting a F1-score of 0.55.

In addition, in (Tsur et al.,2010), they propose a semi supervised system for sarcasm recognition over 66,000 products reviews from Amazon. They used the same strategy as in the previous mention and obtained and F-score of 0.83 on the product reviews dataset.

More recently, other approaches have been developed to solve the task of sarcasm detection. In (Ashwita et al. 2021), the authors experimented by varying the amount of context used along with the response (text to be classified) and found that including the last utterance in the dialogue along with the response improved the performance of their system.

In (Khatri, P, Pranav, y M, Dr. Anand Kumar 2020), a model using machine learning techniques with BERT and GloVe embeddings to detect sarcasm in tweets was proposed.

## 3 System Overview

This section describes the two types of models that were submitted and the techniques and methods applied to each model to improve their performance and metrics.

### 3.1 Data augmentation

One of the main problems with the dataset is the small number of tweets to train our models (only 867 tweets). To solve this, a data augmentation technique was applied. In particular, a synonym augmenter (Wordnet, English) (McCrae et al. 2019) was used to create a new tweet but only swapping one random word by its synonym and keeping their labels. An example of this technique can be seen in Table 2.

| | |
|---|---|
| **Original tweet** | The quick brown fox jumps over the lazy dog |
| **New tweet** | The quick gray fox jumps over the lazy dog |

Table 2: Example of data augmentation

We suggest applying this technique only once because our model could overfitting the data and could yield overrated results of the metrics.

### 3.2 External databases

Another technique applied in the proposed models for the data augmentation was the manual insertion of tweets and labels (Oprea and Magdy 2019). Most of the tweets inserted belong to minority labels (we can see the minority classes on Table 1) such as satire, overstatement or understatement. Finally, once the new tweets were manually added, the dataset consisted of 904 tweets.

### 3.3 Text Processing

We have applied three versions of text processing to clean and simplify the text based on the work described in (Alzahrani and Jolonian 2021).

**Text processing v1.0:** this is the most basic pre-process. For this version, the following guidelines were applied:

- Conversion of all characters to lowercase.
- Extent of all possible contractions in English (e.g., what's → what is).
- Removal of emojis.
- Removal of special characters.
- Removal of multiple spaces between characters.

**Text processing v2.0:** this is the intermediate version. In addition to the features described at v1.0, the following features were added:

- Removal of emojis made from keyboard characters
- Removal of mentions
- Removal of links

**Text processing v3.0:** this is the full version. In addition to the features presented in v2.0, a removal of stopwords in English was added.

### 3.4 First model: Bayesian networks

The first model that was developed involves the classic algorithm of Bayesian networks (Heckerman and Wellman 1995) to study the pattern of behavior that the categories of sarcasm may present in our dataset.

We used a naive Bayes classifier (NBC) which assumes that the attributes are independent of each other. That is to say, the probability can be obtained by calculating the product of the individual conditional probabilities of each attribute given the class node as it can be seen on Figure 3.

In this model, an input (a single tweet) is provided, and it returns a vector of size six (one for each tag) with the predicted label.

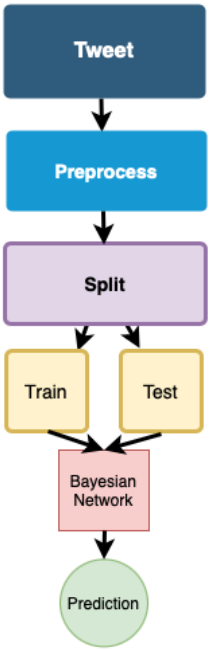


Figure 3: Steps followed on the first model

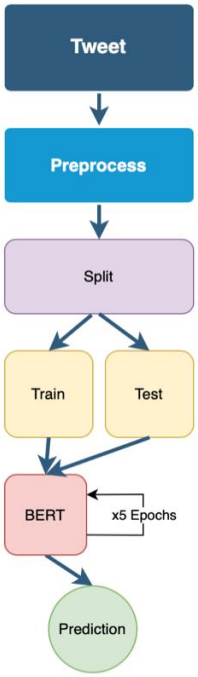


Figure 4: Steps followed on the second model

### 3.5 Second model: BERT

Bidirectional Encoder Representations from Transformers (BERT) is a transformer-based machine learning technique for natural language processing (NLP) pretraining developed by (Devlin et al. 2018). *BERT-base-uncased* model is pretrained from unlabeled data extracted from BooksCorpus (Bandy and Vincent 2021) which have 800M of words and from English Wikipedia with 2,500M of words.

BERT uses Transformers (Wolf et al. 2019) as an attention mechanism that learns contextual relations between words (or sub-words) within a text. Transformers includes two separate mechanisms: an encoder that reads the text input and a decoder that produces a prediction of the label.

For this model, the binary relevance (BR) strategy (Luaces et al. 2012) was used, which splits the learning process of the dataset into a sets of binary classification tasks, in other words, one per label. The main disadvantage of this strategy is that BR ignores any label dependency and could fail in predicting some combination of labels that presents any dependency.

We have trained our second model with a batch of 32 instances and 5 epochs. Figure 4 represents the strategy of this model.

## 4 Experimental Setup

To obtain the above models, some libraries were used:

- For the data augmentation, the *nltk* library (Wang and Hu 2021, 1041-1049).
- For padding the sequences of the inputs id's, the *keras* library.
- *Sklearn* library for the metrics and splitting the dataset (Hao and Ho 2019).
- *Pandas* library for working with dataframes (Stepanek 2020).
- *Transformers* library for everything related with BERT.

For all the experiments, tweets were preprocessed and then they were randomly split using a stratified method (80% training and 20% testing). That means that the proportion of values in the samples produced is the same as the proportion of values provided for the parameter to stratify.

During the training phase of the competition, we have only focused on the macro-F1 score as the key metric of the Subtask B.

## 5 Results

Two submissions were sent using Bayesian Networks: the first one with the text processing v1.0 and the second one with the text processing v3.0. Table 3 shows the results obtained during the training phase.

| Metrics | v1.0. Text processing | v3.0. Text processing |
|---|---|---|
| F1 Sarcasm | 0.89 | 0.84 |
| F1 Irony | 0.26 | 0.20 |
| F1 Satire | 0.38 | 0.51 |
| F1 Overstatement | 0.17 | 0.48 |
| F1 Understatement | 0.47 | 0.47 |
| F1 Rhetorical Question | 0.12 | 0.20 |
| Macro F1-Score | 0.38 | 0.45 |

Table 3: Results obtained using Bayesian Networks

| Metrics | v1.0 | v2.0 | v3.0 | v4.0 | v5.0 |
|---|---|---|---|---|---|
| F1 Sarcasm | 0.45 | 0.78 | 0.81 | 0.80 | 0.70 |
| F1 Irony | 0.45 | 0.72 | 0.77 | 0.77 | 0.64 |
| F1 Satire | 0.49 | 0.63 | 0.66 | 0.64 | 0.57 |
| F1 Understatement | 0.50 | 0.60 | 0.63 | 0.62 | 0.52 |
| F1 Overstatement | 0.49 | 0.65 | 0.69 | 0.71 | 0.60 |
| F1 Rhetorical Question | 0.84 | 0.90 | 0.91 | 0.89 | 0.87 |
| Macro F1-Score | 0.54 | 0.71 | **0.75** | 0.74 | 0.65 |

Table 4: Results obtained using BERT

Regarding the final submission using BERT-base-uncased, different approaches were used. Table 4 shows the results obtained during the training phase. The approaches were:

- **v1.0.** Nothing extra applied
- **v2.0.** Previous versions + Data Augmentation in minority class only.
- **v3.0.** Previous versions + insert data of an external database.
- **v4.0.** Previous versions + v2.0 of text processing.
- **v5.0.** Previous version + v3.0 of text processing.

Taking a look at Table 3 and Table 4, can be seen that in v3.0 of Table 4, the best macro F1-score is obtained. So that was our final submission.

According to the official metrics, as was mentioned before, we achieved a macro F1-score of 0.0699 and we were ranked 10$^{th}$ among 22 teams that participated on this subtask.

Analyzing our systems, we can state that the main problem found in the subtask was the lack of data towards unbalanced data at the dataset, which is why we have been constantly applying data augmentation on the minority classes and even inserting data from an external database. Applying these two techniques, a big improvement in the performance of our systems can be seen.

Furthermore, our research shows that any kind of preprocessing technique is mostly useless because any character, capital letter, overextended word or symbol, could be the determining factor in recognizing ironic speech.

## 6 Conclusion

In this paper our approach to solve *Task 6 (iSarcasmEval) – Subtask B: Given a text, determine which ironic speech category it belongs to, if any; in English*, has been described.

Our best result was reached with a deep learning algorithm (BERT) model, with which we achieved a macro F1-score of 0.0699. We obtained the 10th position in the ranking.

For future works, an improved version of our BERT model could be developed by training with a bigger dataset. It is also possible to look for new preprocessing techniques that enable the removal of information that is useless to the meaning of the tweet but still maintain the ironic speech patterns (if any).

## References

Abu Farha, Ibrahim, Silviu Oprea, Steven Wilson, Walid Magdy. "SemEval-2022 Task 6: iSarcasmEval, Intended Sarcasm Detection in English and Arabic". In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics, 2022.

Carvalho, Paula, Luís Sarmento, Mário J. Silva, and Eugénio de Oliveira. 2009. "Clues for Detecting Irony in User-Generated Contents: Oh...!! It's 'so Easy' ;-)." In *Proceeding of the 1st International*

*CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion - TSA '09*. New York, New York, USA: ACM Press.

Wallace, Byron C., Do Kook Choe, Laura Kertz, and Eugene Charniak. 2014. "Humans Require Context to Infer Ironic Intent (so Computers Probably Do, Too)." In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 512–16. Stroudsburg, PA, USA: Association for Computational Linguistics.

Leggitt, John S., and Raymond W. Gibbs. 2000. "Emotional Reactions to Verbal Irony." *Discourse Processes* 29 (1): 1–24. https://doi.org/10.1207/s15326950dp2901_1.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Semi-supervised recognition of sarcastic sentences in twit- ter and amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, CoNLL '10, pages 107–116, Stroudsburg, PA, USA. Association for Computational Linguistics.

Oren Tsur, Dmitry Davidov, and Ari Rappoport. 2010. ICWSM - a great catchy name: Semi-supervised recog- nition of sarcastic sentences in online product reviews. In William W. Cohen and Samuel Gosling, editors, *Proceedings of the Fourth International Conference on Weblogs and Social Media, ICWSM 2010, Washington, DC, USA, May 23-26, 2010*. The AAAI Press.

Oprea, Silviu, and Walid Magdy. 2019. "ISarcasm: A Dataset of Intended Sarcasm." *ArXiv [Cs.CL]*. https://paperswithcode.com/paper/isarcasm-a-dataset-of-intended-sarcasm.

Alzahrani, Esam, and Leon Jololian. 2021. "How Different Text-Preprocessing Techniques Using the BERT Model Affect the Gender Profiling of Authors." *ArXiv [Cs.CL]*. http://arxiv.org/abs/2109.13890.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." CoRR abs/1810.04805.

Bandy, Jack, and Nicholas Vincent. 2021. "Addressing 'Documentation Debt' in Machine Learning Research: A Retrospective Datasheet for BookCorpus." *ArXiv [Cs.CL]*. http://arxiv.org/abs/2105.05241.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, et al. 2019. "HuggingFace's Transformers: State-of-the-Art Natural Language Processing." ArXiv [Cs.CL]. http://arxiv.org/abs/1910.03771.

Luaces, Oscar, Jorge Díez, José Barranquero, Juan José del Coz, and Antonio Bahamonde. 2012. "Binary Relevance Efficacy for Multilabel Classification." *Progress in Artificial Intelligence* 1 (4): 303–13. https://doi.org/10.1007/s13748-012-0030-x.

Wang, Meng and Fanghui Hu. 2021. "The Application of NLTK Library for Python Natural Language Processing in Corpus Research." Theory and Practice in Language Studies 11 (9): 1041-1049. doi:10.17507/tpls.1109.09.

Hao, Jiangang and Tin Kam Ho. 2019. Machine Learning made Easy: A Review of Scikit-Learn Package in Python Programming Language. Vol. 44. Los Angeles, CA: SAGE Publications. doi:10.3102/1076998619832248. https://journals.sagepub.com/doi/full/10.3102/1076998619832248

Stepanek, Hannah. 2020. Thinking in Pandas : How to use the Python Data Analysis Library the Right Way. Berkeley, CA: Apress. doi:10.1007/978-1-4842-5839-2. https://library.biblioboard.com/viewer/087e187a-b660-11ea-a44d-0a7fc7c4e64f.

McCrae, John Philip, Alexandre Rademaker, Francis Bond, Ewa Rudnicka, and Christiane Fellbaum. 2019. "English WordNet 2019 – An Open-Source WordNet for English." In *Proceedings of the 10th Global Wordnet Conference*, 245–52.

Ashwitha, Shruthi, Shruthi, Makarand Upadhyaya, Abhra Pratip Ray, y Manjunath. 2021. "Sarcasm Detection in Natural Language Processing". Materials Today: Proceedings 37: 3324–31. https://doi.org/10.1016/j.matpr.2020.09.124.

Khatri, Akshay, P, Pranav, y M, Dr. Anand Kumar. 2020. "Sarcasm detection in tweets with BERT and GloVe embeddings". https://doi.org/10.48550/ARXIV.2006.11512.

Heckerman, David, and Michael P. Wellman. "Bayesian networks." *Communications of the ACM*, 1995, 38(3), 27-31