

# French Tweet Corpus for Automatic Stance Detection

Marc Evrard, Rémi Uro, Nicolas Hervé, and Béatrice Mazoyer

Institut national de l’audiovisuel (INA)  
18, avenue des Frères-Lumière  
94366 Bry-sur-Marne Cedex  
{mevrard, ruo, nherve, bmazoyer}@ina.fr

## Abstract

The automatic stance detection task consists in determining the attitude expressed in a text toward a target (text, claim, or entity). This is a typical intermediate task for the fake news detection or analysis, which is a considerably widespread and a particularly difficult issue to overcome. This work aims at the creation of a human-annotated corpus for the automatic stance detection of tweets written in French. It exploits a corpus of tweets collected during July and August 2018. To the best of our knowledge, this is the first freely available stance annotated tweet corpus in the French language. The four classes broadly adopted by the community were chosen for the annotation: *support*, *deny*, *query*, and *comment* with the addition of the *ignore* class. This paper presents the corpus along with the tools used to build it, its construction, an analysis of the inter-rater reliability, as well as the challenges and questions that were raised during the building process.

**Keywords:** Stance Detection, French, Dataset, Fake News, Twitter, Inter-Rater Reliability, Natural Language Processing

## 1. Introduction

Fake news proliferation is a conspicuous issue, which seems particularly difficult to overcome. Their detection is a complex task, requiring both a large and precise knowledge of the world. Indeed, most domains seem to be affected by fake news, while some rumors refer to very specific topics that are known by only small groups of initiated people, which implies that extensive knowledge of all existing concepts would be necessary to tackle this problem. Since a direct approach leading to a high accuracy does not yet seem practical, we are instead currently focusing on the gathering of information to describe and analyze fake news. One of the parameters we wish to investigate is the determination of the stance of a text toward another text it relates to. This is a typical intermediate task of the fake news analysis or detection, as the discussions concerning them are often controversial and debated, especially on social networks (Zubiaga et al., 2018). Discussion threads about rumors are thus expected to exhibit contrasting stances. Moreover, it has been shown that the veracity of a rumor is correlated to the stance of discussion toward it (Ferreira and Vlachos, 2016; Enayet and El-Beltagy, 2017).

We decided to resort to social networks to gather the text corpus since it is a major source of fake news. According to the Reuters Institute (Nic et al., 2018) in their Digital News Report of 2018, 36% of French adults use social media as a source of news. Although this share has declined since 2017, it is mostly due to a decrease in the use of Facebook, while other networks like Twitter are stable or growing rapidly. Besides being a prominent source of public event information, Twitter also provides API endpoints allowing the realtime fetching of data.

At the *Institut national de l’audiovisuel* (France), through the OTMedia+ project (Hervé, 2019), a large corpus of tweets in the French language was gathered in 2018 (Mazoyer et al., 2018). It serves as a basis for this annotation work. To the best of our knowledge, no corpus of French tweets annotated for stance detection is readily available.

This paper presents the construction of such a corpus, as well as the challenges and questions that were raised during this process.

### 1.1. Automatic stance detection

Three SemEval tasks have been proposed for automatic stance detection (Mohammad et al., 2016; Derczynski et al., 2017; Gorrell et al., 2019). The first task is applied to a Twitter dataset targeting six selected topics. The second uses a Twitter rumor dataset, while the third extends it by providing a dataset from Reddit—another social media network with a similar structure to Twitter when it comes to discussion threads. Different approaches have been proposed to solve these tasks. For the first task, the best performances were drawn using standard text classification features (i.e., n-grams, word vectors, and sentiment lexicons). The highest macro-averaged  $F_1$  for the second task was achieved by (Kochkina et al., 2017) with a sequential approach using a Branch-Long short-term memory recurrent neural network (Branch-LSTM) architecture. It also resorts to several computed features such as a lexicon, the similarity to other tweets, or the punctuation. The highest score on the third task was achieved by (Yang et al., 2019) using a chain-based generative pre-trained transformer (GPT) model. Other approaches such as BUT-FIT (Fajcik et al., 2019) based on BERT—the pre-trained deep bidirectional transformers used for language understanding (Devlin et al., 2019)—also achieve good scores.

For a more extensive review of stance and rumor detection approaches in social media, the reader is invited to refer to the work of Zubiaga et al. (2018).

### 1.2. Corpus

The tweets annotated in this work were sampled from a 38 million tweets (retweets excluded) corpus in the French language, collected during a three weeks period in July and August 2018. An initial manual annotation process was applied to 130,000 tweets by three annotators in the context

of another project (Mazoyer et al., 2020). It consisted of choosing whether or not each tweet was related to an event out of a list of 243 events, selected from both the press and Twitter trending subjects during the annotation period. Out of the 130,000 tweets, more than 95,000 were annotated as related to one of these events.

We chose to base our work on this dataset since the event labels bring more context and thus open the perspective of application to other complex tasks. In our case, the focus is aimed at the stance of a tweet toward another within a given discussion thread. Two annotation levels were chosen for each tweet: one for the stance toward its direct predecessor—the *previous* tweet—the other toward the first tweet of the thread—referred here as the *root* tweet. To select the most interesting tweets, only the threads displaying a minimum number of replies were retained in the corpus. However, on these remaining threads, only the first and second level of replies were presented to the annotator to ease the annotation process, since tweet topics tend to quickly drift away from the initial topic as the depth of the thread increases. After this filtering process, more than 15,000 tweets remained for annotation.

To prevent too much bias, the selection was limited to a few parameters: number of retweets from the original tweet and the length of threads. For instance, the number of words in the tweets were not taken into account, although the initial gathering process was based on keyword searches, which statistically tends to increase the length of tweets. Keeping longer tweets would have eased the annotation process by bringing more contextual information and avoiding very short potentially unintelligible tweets, but this would have biased the corpus toward a particular writing style. We also chose not to filter according to the subject of a tweet. Keeping subjects as diverse as possible allows a wider range of application tasks for this corpus.

## 2. Annotation process

The four classes usually adopted by the community (Procter et al., 2013) were chosen as a basis for this annotation task: *support*, *deny*, *query*, *comment* to which an *ignore* class was added to make sure all cases were taken into consideration. Although we believed this classification scheme was not optimal, in particular, because of the presence of intersections between the classes, we decided to use them to keep our corpus compatible with other frameworks.

The tweets were presented to the annotator as a *pack* containing the *root* and its children belonging to all threads initiated from that *root*. We initially thought of presenting each tweet pair individually, since the automatic task we foresaw was to predict the stance of a tweet source toward its target without any other form of context. The global approach simplifies the annotation process, by allowing the annotators to gather more context from the whole thread and thus helping them stay focused on a consistent topic for the whole annotation *pack*, instead of constantly switching topics at each annotation pair. Moreover, this corpus is meant to be used for various task configurations, thus we did not want to be too specific to the task at hand and to let the range of applications of this corpus open to other task schemes.

The *packs* were then grouped in *sets*. The number of *packs* included in a *set* depended on the number of tweets contained in each *pack*. The aim was to obtain a roughly equivalent annotation workload for each *set*. The *packs* were allocated so that the number of tweets per *set* stayed constantly close to 100, as shown in Figure 1. The *sets* were then sorted in descending order of the average number of tweets per *pack*, so that the first *sets* held few larger *packs*, while the last held several smaller *packs*. This order allowed prioritizing the annotation of *packs* containing a large number of tweets, which means that either the *root* tweet received several responses, or that each response generated long discussion threads. Longer threads were expected to correlate with more controversial or more polarized discussions, which would then lead to stronger stance expression.

Although longer threads provide more context, which is expected to ease the annotation process, too long threads can also be fastidious to handle. Indeed, very long threads (91 tweets for the longest) require staying focused for an extended period on the same subject. We thus chose to randomize the first 52 *sets* to alternate between long threads and smaller ones.

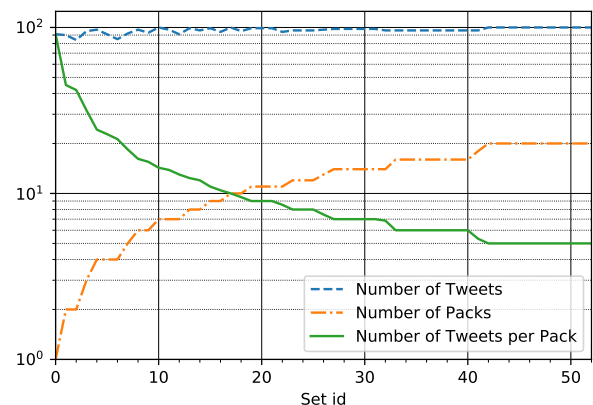


Figure 1: Organization of the tweets for annotation, for the 52 first *sets*. The plot shows the number of tweets per *set*, the number of *packs* per *set*, and the number of tweets per *Pack*.

Each tweet target was annotated toward its predecessor (*previous*) and toward the *root*—if the latter was not its predecessor. Thus they are two types of annotation pairs: *target-previous* and *target-root*. We chose to call *target-previous* the pairs whose link is direct, even if the targets in question respond to the *root*.

A graphical interface was designed to facilitate the annotation task. It shows the tweets present in the *pack* successively, in chronological order, as the annotator is scrolling down. Their order of appearance is similar to the one of the depth-first search algorithm. One or two selection menus—depending on the distance of the target tweet to the *root*—present the five classes to use for annotation, as illustrated on the screenshot of the annotation graphical interface in Figure 2. The menu on the right refers to the *root* tweet and the menu on the left to the preceding tweet. To further

ease the annotation work, the color of the menu matches the color of the tweet it refers to. At the end of the *pack*, two buttons let the annotator tell whether the *pack* was interesting or not. In either case, the annotation is saved, and a corresponding tag is recorded in the database. The next *pack* in the *set* is then presented to the annotator until the last *pack* of the *set* is finished. Then a page proposes to start the annotation of the next *set*.

Tweets quoting another tweet are considered like responses and are annotated in the same way. However, to let the annotator know, the quoting tweets contain an extract of the quoted text, as shown in the second tweet in Figure 2.

### 3. Results analysis

Eight annotators (five males and three females) who are regular Twitter users, participated in the labeling of the corpus. At the current state of the annotation process, a total of 5803 tweet pairs were labeled, 4291 pairs were annotated by at least two annotators, and 3117 were by at least three. Among the tweet pairs, 72.35% consist of directly linked tweets (*target-previous*), and 27.65% of indirectly linked tweets (*target-root*). Figure 3 shows the distribution of annotated pairs, according to a minimal number of annotators. Labels need to be extracted from the annotation results to be used as a training or testing corpus for a classifier. A typical strategy is to resort to majority votes. To avoid undetermined votes, the number of annotators theoretically required to annotate each tweet pair is the number of classes plus one, which would mean that at least six annotators would be required. However, we observed that the number of undetermined cases was acceptable for a lower number of annotators. First, as shown in Table 1, when the classes *ignore* and *comment* are merged, the four resulting classes are obviously fully determined with five annotators. The number of indeterminations is negligible with four annotators. It increases significantly with three annotators but seems still acceptable, with 3.85% for the *target-previous* pairs and 2.01% for the *target-root* pairs. These undetermined annotation votes will need to be discarded to use this corpus to train a classifier, but the concerned tweet pairs are likely to be the most prone to ambiguity, thus this may be seen as a way of cleaning the corpus.

An analysis of the inter-rater reliability was performed on tweet pairs annotated by two groups of three raters and by a group of five. They were chosen from those who share the largest number of annotated tweets. Table 2 shows the results of the first group of three annotators. On 1757 *target-previous* pairs, they achieve an inter-rater reliability score of  $\kappa = 0.42$  (using the Randolph (2005)’s score) on the four classes. This moderate agreement seems coherent with the complexity of the task. Merging *ignore* and *comment* proves useful, since, with the five classes, the score goes down to  $\kappa = 0.34$ —a fair agreement according to Cohen (1960)’s scale. To improve the score, the undetermined annotated votes were removed—which is realistic since they cannot be used for training or testing purposes—the score gets up to  $\kappa = 0.47$ . The second group of three annotators goes up to  $\kappa = 0.53$  in the same condition, as shown in Table 3. Globally the two groups seem to present rather similar agreement scores. However, it can be noted that

the first group tends to do slightly better for the *target-root* pairs, while the second shows more coherence on the *target-previous* pairs.

The agreement of the group of five annotators is also in the moderate range if four classes are kept, as shown in Table 4. With a score of  $\kappa = 0.42$ , it is on par with the first group of three, when the undetermined cases are still present in the *target-previous* pairs. However, the agreement reaches  $\kappa = 0.48$  for the *target-root* pairs. These results should be taken with caution, because of the strong class imbalance, in particular for the *target-root* pairs, presenting a very high ratio of the *comment-ignore* merged class. Indeed, as shown in Figure 4, the *comment* and *ignore* classes account together for more than half (63%) of the annotations.

Nb. of annot.	Previous		Root	
	Und. ratio	Total	Und. ratio	Total
2	47.60%	834	42.35%	340
3	3.85%	1064	2.01%	398
4	0.35%	289	0.00%	112
5	0.00%	846	0.00%	296

Table 1: Ratio of undetermined labels according to the number of annotators, in the four classes scenario (*ignore* and *comment* merged). The third and fifth columns indicate the corresponding total number of pairs.

Nb. of classes	Nb. of tweet pairs		Stance toward	Inter-rater score ( $\kappa$ )	
4	1757	1641	Previous	0.42	0.47
5	1757	1508	Previous	0.34	0.43
4	639	603	Root	0.45	0.49
5	639	527	Root	0.26	0.37

Table 2: The inter-rater reliability score based on the Randolph (2005)’s score for the group of 3 annotators having the largest number of common tweet pairs annotated. The right-hand side values are the results when the undetermined annotation votes are removed. ( $\leq 0$ : no agreement; 0.01–0.20: none to slight; 0.21–0.40: fair; 0.41–0.60: moderate; 0.61–0.80: substantial; 0.81–1.00: almost perfect agreement (Cohen, 1960)).

Nb. of classes	Nb. of tweet pairs		Stance toward	Inter-rater score ( $\kappa$ )	
4	1015	957	Previous	0.48	0.53
5	1015	907	Previous	0.42	0.50
4	366	346	Root	0.41	0.46
5	366	286	Root	0.21	0.35

Table 3: The inter-rater reliability score based on the Randolph (2005)’s score for 3 annotators having the second largest number of common tweet pairs annotated. The right-hand side values are the results when the undetermined annotation votes are removed.

**Root tweet (purple):**  
 NB contrairement à ce que dit @gerardcollomb ce n'est pas nécessairement à l'autorité hiérarchique de faire un signalement article 40 CPP mais à "toute autorité constituée", tout fonctionnaire dans l'exo de ses fonctions qui "a connaissance" d'un crime ou délit <https://t.co/QHtLyTfW60>

**Reply 1 (cyan):**  
 Encore une fois, #Collomb ment comme s'il ignorait que tout le monde peut avoir accès à l'article en question en quelques clics. Ah, les vieux réflexes...  
 "NB contrairement à [...] ou délit <https://t.co/QHtLyTfW60>"

**Reply 2 (cyan):**  
 En réponse à : @combatsdh @LarrereMathilde @gerardcollomb  
 Quelqu'un aurait une photo ou vidéo de Benalla aux côtés de G. Collomb?

**Reply 3 (yellow):**  
 Obnoxious,Jul  
 En réponse à : @david\_dewaele @combatsdh @LarrereMathilde @gerardcollomb  
 Non mais on peut trouver les traces du conseiller de #collomb (Jean Marie Girier) disant à #benalla : non on n'achète pas d'armes pour la campagne Le conseiller de Collomb savait depuis la campagne, il a oublié d'avertir son boss, on dirait <https://t.co/5RNkaq1tsi>  
<https://t.co/G5180Od9NH>

**Reply 4 (cyan):**  
 Donc mensonge sous serment :) #CollombDemission #Collomb #AffaireMacronBenalla #AffaireBellanaMacron #CommissionEnquete  
 "NB contrairement à [...] ou délit <https://t.co/QHtLyTfW60>"

**Annotation buttons:**  
 Ignorer, Soutient (checked), Contredit, Questionne, Commente

**Bottom buttons:**  
 Intéressant, Pas intéressant

Figure 2: Screenshot of the graphical interface built for the annotation process. The page shown corresponds to the *pack* 7 of the *set* 43 (*jeu*)—which includes 20 *packs* in total. The *root* tweet is in purple, the first level answers are in cyan, and the second level in yellow. The menu holding the 5 classes appear on two columns next to the tweet to annotate. The one on the left side corresponds to the stance toward the *previous* tweet and those on the right refer to the stance toward the *root*.

Nb. of classes	Nb. of tweet pairs	Stance toward the:	Inter-rater score ( $\kappa$ )
4	838	Previous	0.42
5	838	Previous	0.35
4	310	Root	0.48
5	310	Root	0.29

Table 4: The inter-rater reliability score based on the Randolph (2005)’s score for the group of 5 annotators having the largest number of common tweet pairs annotated. No undetermined annotation vote occurs in this subset.

## 4. Discussion

This annotation scheme does not seem optimal for several reasons. First, as such, the initial four classes are not exhaustive of all the possible cases. For example, a response to a tweet that is completely unrelated to the original content, or a tweet whose meaning is utterly not understood by the annotator would not fit any of the provided options. Assigning it to the *comment* class would seem the most logical choice since this class represents the least informative general case from the stance point of view. Second, all classes are not perfectly dissociated and could be used simultaneously, which makes the forced-choice task ambiguous to the annotator. For instance, a response to a tweet could support the general stance of the source, while adding more

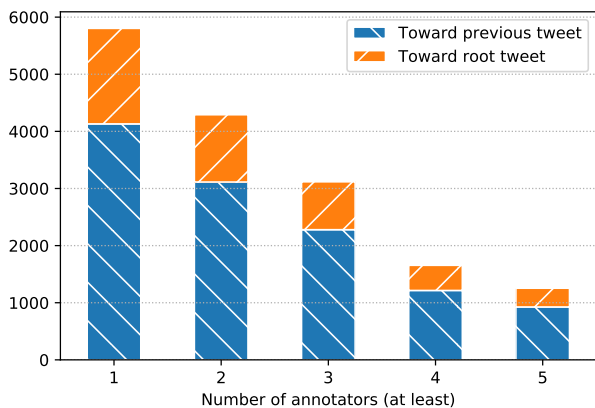


Figure 3: Number of annotated pairs for both the target tweet stance toward the *previous* and toward the *root* tweet. They are presented according to a minimal number of annotators per tweet.

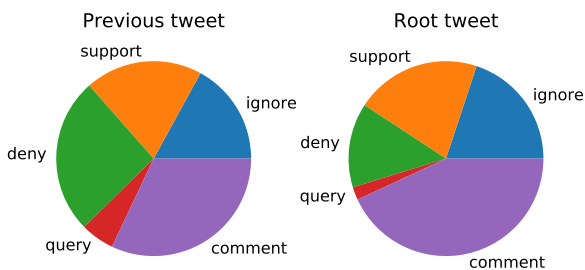


Figure 4: Ratio of annotated classes on the whole corpus for the *target-previous* pairs and for the *target-root* pairs.

information, thus commenting or even questioning part of the content. The annotator would then have to judge if the target tweet mostly agrees or mostly adds information to the source tweet, to decide between *support* and *comment*, respectively.

Figures 5 and 6 show the confusion matrices of two annotators (*rater 1* and *rater 2*) on the 5 classes case, for the *target-previous* and the *target-root* tweet pairs, respectively. On both matrices, *support* and *deny* are very rarely confused, which is reassuring since these two classes are of most importance for the stance detection task. *Support* is mostly confused with *comment*. Interestingly, the *rater 1* is much more prone to the use of the *comment* class, while the *rater 2* seems to prefer the *ignore* class. This is another illustration of the need to merge the two classes. We could argue that the addition of the *ignore* class is thus not very useful since increasing the number of classes is likely to decrease the inter-rater agreement. However, certain annotators reported less frustration when having the possibility to use the *ignore* class.

More complex annotation approaches exist such as the one proposed by Joseph et al. (2017) with the ConStance reasoning model. The authors argue that the lack—or the overabundance—of context information prevents annotators from making a reliable assessment, which leads to



Figure 5: Normalized confusion matrix of two annotators (raters) for 1757 *target-previous* pairs. The normalization is performed according to *rater 1*.



Figure 6: Normalized confusion matrix of two annotators (raters) for 655 *target-root* pairs. The normalization is performed according to *rater 1*.

noisy and uncertain annotations. They propose to present different context information to different raters and learn the optimum combination of annotation conditions. Although this approach could improve the moderate inter-rater agreement of our annotations, the production of gold standard labels needed to train the model would prove very challenging for this task of broad tweet pair stance assignment. Moreover, we currently lack the resources in terms of the number of annotators required to implement the multiplicity of context information conditions.

## 5. Conclusion

In this paper, we presented a new tweet corpus for stance detection. The annotation scheme was presented in detail, as well as the annotation process. The corpus annotations

were described in-depth, using inter-rater reliability analysis, confusions matrices, and other statistical data. We discussed the fact that the moderate agreement obtained in most configurations seems coherent with the complexity of the task.

The tools used to construct the corpus (graphical interface, web server, analysis tools, etc.) and corpus annotations (including the tweet ids) are all open-sourced and freely available. The tools are hosted on GitHub<sup>1</sup> and the annotated Twitter dataset is available upon registration on the INA scientific dataset website<sup>2</sup>, after acceptance of the terms of use agreement. The corpus will be further developed by increasing the number of annotators per tweet pairs. The inter-rater reliability score is expected to improve and the resulting labels to be more consistent and reliable.

## 6. Acknowledgments

The authors would like to warmly thank Laurent Vinet, Valérie Gauffreteau, Élisabeth Chapalain, Florent Lioret, and Thomas Petit for their remarkable annotation work.

## 7. Bibliographical References

- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.
- Derczynski, L., Bontcheva, K., Liakata, M., Procter, R., Wong Sak Hoi, G., and Zubiaga, A. (2017). SemEval-2017 Task 8: RumourEval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 69–76, Vancouver, Canada, August. Association for Computational Linguistics.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.
- Enayet, O. and El-Beltagy, S. R. (2017). NileTMRG at SemEval-2017 Task 8: Determining Rumour and Veracity Support for Rumours on Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 470–474, Vancouver, Canada, August. Association for Computational Linguistics.
- Fajcik, M., Smrz, P., and Burget, L. (2019). BUT-FIT at SemEval-2019 Task 7: Determining the Rumour Stance with Pre-Trained Deep Bidirectional Transformers. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1097–1104, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.
- Ferreira, W. and Vlachos, A. (2016). Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1163–1168, San Diego, California, June. Association for Computational Linguistics.
- Gorrell, G., Kochkina, E., Liakata, M., Aker, A., Zubiaga, A., Bontcheva, K., and Derczynski, L. (2019). SemEval-2019 Task 7: RumourEval, Determining Rumour Veracity and Support for Rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 845–854, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.
- Hervé, N. (2019). OTMedia, the TransMedia News Observatory. In *FIAT/IFTA Media Management Seminar 2019*, Stockholm, Sweden, May. [http://www.herve.name/research/nherve\\_ina\\_fiat\\_stockholm\\_otmedia\\_preprint.pdf](http://www.herve.name/research/nherve_ina_fiat_stockholm_otmedia_preprint.pdf).
- Joseph, K., Friedland, L., Hobbs, W., Lazer, D., and Tsur, O. (2017). ConStance: Modeling Annotation Contexts to Improve Stance Classification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1115–1124, Copenhagen, Denmark, September. Association for Computational Linguistics.
- Kochkina, E., Liakata, M., and Augenstein, I. (2017). Turing at SemEval-2017 Task 8: Sequential Approach to Rumour Stance Classification with Branch-LSTM. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 475–480, Vancouver, Canada, August. Association for Computational Linguistics.
- Mazoyer, B., Cagé, J., Hudelot, C., and Viaud, M.-L. (2018). Real-Time Collection of Reliable and Representative Tweets Datasets Related to News Events. In *Bro-Dyn@ ECIR*, pages 23–34.
- Mazoyer, B., Cagé, J., Hervé, N., and Hudelot, C. (2020). A french corpus for event detection on twitter. In *LREC 2020*.
- Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., and Cherry, C. (2016). Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41.
- Nic, N., Fletcher, R., Kalogeropoulos, A., Levy, D. A., and Nielsen, R. K. (2018). *Reuters Institute Digital News Report 2018*. Reuters Institute for the Study of Journalism.
- Procter, R., Vis, F., and Voss, A. (2013). Reading the riots on Twitter: methodological innovation for the analysis of big data. *International journal of social research methodology*, 16(3):197–214.
- Randolph, J. J. (2005). Free-Marginal Multirater Kappa (multirater K [free]): An Alternative to Fleiss’ Fixed-Marginal Multirater Kappa. *Online submission*.
- Yang, R., Xie, W., Liu, C., and Yu, D. (2019). BLCU\_nlp at SemEval-2019 Task 7: An Inference Chain-based GPT Model for Rumour Evaluation. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1090–1096.
- Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M., and Procter, R. (2018). Detection and Resolution of Rumours in Social Media: A Survey. *ACM Computing Surveys*, 51(2):1–36, February.

<sup>1</sup><https://github.com/ina-foss/tweet-stance-annotation>

<sup>2</sup><https://dataset.ina.fr/corpus>