

Digging Communicative Intentions: The Case of Crises Events

Farah Benamara

FARAH.BENAMARA@IRIT.FR

*IRIT, Université de Toulouse, CNRS, Toulouse INP, UT3. France
IPAL, CNRS-NUS-ASTAR. Singapore*

Alda Mari

ALDA.MARI@ENS.FR

IJN CNRS/ENS/EHESS/PSL. France

Romain Meunier

ROMAIN.MEUNIER@IRIT.FR

IRIT, Université de Toulouse, CNRS, Toulouse INP, UT3. France

Véronique Moriceau

VERONIQUE.MORICEAU@IRIT.FR

IRIT, Université de Toulouse, CNRS, Toulouse INP, UT3. France

Leila Moudjari

LEILA.MOUDJARI@IRIT.FR

IRIT, Université de Toulouse, CNRS, Toulouse INP, UT3. France

Valentin Tinarrage

V.TINARRAGE@GMAIL.COM

IJN CNRS/ENS/EHESS/PSL. France

Editor: Junyi Jessy Li

Submitted 07/2023; Accepted 05/2024; Published online 05/2024

Abstract

In emergency situations users of social networks convey all sorts of what have been called communicative intentions, well-known since the work of [Austin \(1962\)](#) and [Searle \(1969\)](#) as speech acts (SA). While speech acts have been the focus of close scrutiny in the philosophical and linguistic literature (see [\(Portner, 2018\)](#) for extended discussion), their role has been only rarely understood and exploited in processing social media content about crisis events, our focus here. Current work on communicative intentions in social media are *topic-oriented*, focusing on the correlation between SA and specific topics such as crisis (e.g., earthquakes) but also politics, celebrities, cooking, travel, etc. It has been observed that people globally tend to react to natural disasters with SA distinct from those used in other contexts (e.g., celebrities, which are essentially made up of comments). Here, we explore the further hypothesis of a correlation between different SA types and urgency and propose an in depth linguistic and computational analysis of communicative intentions in tweets from an *urgency-oriented* perspective. Indeed, SA are mostly relevant to identify intentions, desires, plans and preferences towards action and to ultimately produce a system intended to help rescue teams. Our contribution is four-fold and consists of: (1) A two-layer annotation scheme of speech acts both at the tweet and sub-tweet levels, (2) A new French dataset of about 13K tweets annotated for both urgency and SA, targeting both expected (e.g., storms) and unexpected or sudden (e.g., building collapse, explosion) events, (3) A thorough analysis of the annotations studying in particular the correlation between SA and the urgency of the message, SA and intentions to act categories (e.g., human damages), and SA and crisis types, finally, (4) A set of deep learning experiments to detect SA in crises related corpora. Our results show a strong correlation between SA and urgency annotations at both the tweet and sub-tweet levels with a particular salient correlation in the latter case, which constitutes a first important step towards SA-aware NLP-based crisis management on social media.

Keywords: Speech acts, Crisis events, Social Media

1. Introduction

1.1 Motivation

In ordinary interaction as well as in social networks, speakers unveil a variety of communicative intentions among which, make content known, express their own views and opinions or enhance action. Since [Austin \(1962\)](#) and later and more prominently [Searle \(1975\)](#), these communicative intentions are known under the term *speech acts*.

Before percolating into the computational literature, speech acts (henceforth SA) have been the object of extensive discussion in the philosophical and the linguistic communities (([Hamblin, 1970](#); [Brandom, 1994](#); [Sadock, 2004](#); [Asher and Lascarides, 2008](#); [Portner, 2018](#); [Bach and Harnish, 1979](#)) to mention just a few). According to the Austinian initial view, SA are to achieve action rather than conveying information. When uttering *I now baptize you*, the priest accomplishes the action of baptizing rather than just stating a proposition. Beyond these prototypical cases, the literature has quickly broadened the understanding of the notion of SA as a special type of linguistic object that encompasses questions, orders and assertions and transcends propositional content revealing communicative intentions on the part of the speaker ([Bach and Harnish, 1979](#); [Gunlogson, 2008](#); [Asher and Lascarides, 2008](#); [Giannakidou and Mari, 2021c](#)): With an *assertion*, the speaker intends to present the propositional content and to add it to the common ground ([Portner, 2018](#)); with a *question*, the speaker asks the addressee to provide new information; with an order the speaker asks that the content be realized and with *exclamatives*, a subjective evaluation towards propositional content is conveyed.

Our study investigates the communicative intentions that SA conveys in **urgency situations** and more importantly, how intentions vary according to the degree of urgency of the information (urgent vs. not urgent vs. not useful – cf. examples below) when posted in social networks. We focus on messages posted on Twitter as tweets are widely used to generate valuable information in crisis situations ([Reuter et al., 2018](#)). For example, the Notre Dame fire that occurred in France has been the most used in Twitter in 2019¹ and in the recent earthquake in Turkey and Syria, some victims trapped in the rubble have been saved thanks to the messages they posted ([Toraman et al., 2023](#)).

SA are particularly helpful in identifying urgent messages. These are messages that raise situational awareness over a crisis situation and some specific aspects that include human/infrastructure damages, security instructions, etc. They provide actionable information that will help human teams to set priorities and decide appropriate actions ([Vieweg et al., 2014](#); [Castillo, 2016](#); [Reuter and Kaufhold, 2018](#)). Therefore, speaking subjects perform qualitatively very different language acts depending on the situation they find themselves in. They mostly aim to make interlocutors react (i.e., *perlocutionary level*) by different linguistic means (*illocutionary level*, this is the level at which the speech acts are encoded), in view of achieving a purpose.²

1.2 When Communicative Intentions Reveal Urgency

By revealing speakers communicative intentions and aiming at triggering the addressee reaction, speech acts become essential in emergency situations where action is to be enhanced. We have thus used two different independent classifications: (i) a new, two level classification of speech acts, and (ii) an independent classification for urgency and actionability elaborated in [Kozłowski et al. \(2020\)](#).

1. https://blog.twitter.com/en_us/topics/insights/2019/ThisHappened-in-2019

2. On perlocutionary / illocutionary, see ([Austin, 1962](#); [Searle, 1975](#)).

The following are two examples³ of how these two classifications proceed. We use the \rightarrow notation, with, at its left, the tweet-level categories, and, at its right, the sub-tweet level categories. A precise definition of the labels will be provided later in the paper (see Section 3).

- (1) a. [The fire situation in the Landiras area is getting worse.]¹ [Please follow the instructions of the fire brigade and the police.]²
 (SA annotation) JUSSIVE \rightarrow 1. PROPER ASSERTIVE; 2. OPEN OPTION
 (Urgency annotation) URGENT: WARNING/ADVICE
- b. [5th day of fire fighting, about 6000 hectares of our forest charred here. Still the same means at the disposal of our firemen: 2 air-crafts and 1 dash.]¹ [What are you waiting for to give them the means to stop this fire? @EmmanuelMacron @GDarmanin #landiras]²
 (SA annotation) SUBJECTIVE \rightarrow 1: PROPER ASSERTIVE; 2: EVALUATION
 (Urgency annotation) URGENT: MATERIAL DAMAGE

As shown in these examples, a tweet is composed of several parts that contribute to the construction of the communicative intention of the whole message. These parts may convey (and they indeed often do) very different speech acts types. Therefore tweets need *also* to be analyzed at the sub-tweet level, in order to search for more precise and specific content that provides useful actionable information.

For (1-a), the writer publicly expresses an explicit demand (hence a JUSSIVE⁴ speech act at the tweet level) for the population to follow the authorities' instructions as the wildfires in the Landiras region keep spreading. At the sub-tweet level, (1-a) first presents a description of the situation (cf. segment 1 that triggers a speech act of PROPER ASSERTIVE) and then provides an advice on how to behave (see segment 2 which qualifies as an OPEN OPTION in our classification, cf. *infra*). The latter is the most useful piece of content as it provides new and actionable information triggering action expectation. For emergency and actionability, (1-a) qualifies as URGENT at the tweet level, specifically providing content that falls in the actionability category ADVICE.

As a further example, insofar as the speech act annotation is concerned, (1-b) expresses an intention to complain about the current means at the disposal of the fire brigades. The overall tweet is considered as expressing a subjective stance of the speaker (hence the overall label SUBJECTIVE) in virtue of the question, which reveals a complaint (the part containing the question is labeled as EVALUATION, cf. *infra* for details). The first segment is a PROPER ASSERTIVE. As for the emergency annotation of the same tweet, (1-b) qualifies as URGENT at the tweet level, providing content that is labeled MATERIAL DAMAGES at the actionability level.⁵

1.3 Previous Approaches and Research Questions

Since the introduction of dialogue acts (see, a.o., the DAMSL framework (Allen and Core, 1997; Core et al., 1998)), SA have been dedicated an extensive body of work in the computational linguistics literature where various approaches have been proposed to detect them in both synchronous (e.g., meeting, phone) (Stolcke et al., 2000; Keizer et al., 2002; Carvalho and Cohen, 2005; Joty and Mohiuddin, 2018) as well as asynchronous dialogues (e.g., emails, live chats, tweet threads)

3. These are examples taken from our French corpus translated into English.

4. We borrow the Latin word for order as standard practice in linguistics, see Portner (2018)

5. Kozłowski et al. (2020) classification also comprises a prior level of relatedness as we explain later in the paper.

(Carvalho and Cohen, 2005; Joty and Mohiuddin, 2018; Bracewell et al., 2012). SA have shown to be an important step in many downstream NLP applications such as strategic actions prediction (Cadilhac et al., 2013), dialogues summarization (Goo and Chen, 2018) and conversational systems (Higashinaka et al., 2014). However, SA for emergency detection has received less attention in the literature and most of related work on communicative intentions in social media are *topic-oriented*, focusing on the correlation between SA and specific topics such as crisis (e.g., earthquakes, bombing, attacks) but also politics, celebrities, cooking, travel, etc. (Zhang et al., 2011; Vosoughi, 2015; Elmadany et al., 2018a; Saha et al., 2020b). These corpus-based studies show that there is a greater similarity of distribution between topics of the same type than between topics of different types. In particular, it has been observed that people globally tend to react to natural disasters with SA distinct from those used in other contexts (e.g., celebrities, which are essentially made up of *comments*).

Here, we explore the further hypothesis of a correlation between different SA types and urgency. We thus investigate whether SA can be used to sort urgent from not urgent messages. As far as we know, this is the first study that proposes an in depth linguistic and computational analysis of communicative intentions in tweets from an *urgency-oriented* perspective: *What are the most frequent intentions in urgent vs. not urgent message? Are these intentions different from those found in non useful messages? And more importantly, are they particularly correlated with fine-grained urgency categories (such as human/infrastructure damages, donations, security instructions etc.)? Finally, are the observed SA stable across different types of crisis (flood, hurricane, fire, attack, etc.)?* To answer these questions and before moving to real scenarios that rely on SA-aware automatic detection of urgency (this is left for future work), we propose to (1) measure the impact of SA in detecting urgency during crisis events in manually annotated data, and (2) explore the feasibility of SA automatic detection in crisis corpora.

1.4 Overview of the Main Contributions

We build on Laurenti et al. (2022a) where we performed a preliminary analysis of the role of SA on urgency detection in about 6,6K tweets with of a focus on natural disasters (flood, hurricane, storm, etc.). In Laurenti et al. (2022a), we relied on a new annotation scheme of SA that takes into account the variety of linguistic means whereby SA are expressed (including lexical items, punctuation, etc), both at the message and sub-message level. We further extend this initial work by proposing:

- The first largest French dataset of about 13,300 tweets annotated for both urgency and SA following the same annotation scheme. In addition, we expend the annotations to 6 new sudden crisis making the dataset spans over 20 crises.⁶
- A qualitative and quantitative analysis of the annotation campaign intersecting the two-level classification of speech acts with a classification of urgency. In particular, we explore the correlations between SA vs. urgency, SA vs. intention to act categories as well as SA vs. the types of crises for both levels of SA annotations. Our results show a strong correlation between SA and urgency annotations at both the tweet and sub-tweet levels with a particular salient correlation in the latter case which constitutes a first important step towards SA-aware NLP-based crisis management on social media.

6. The annotated dataset will be available for research purposes upon request.

- A set of deep learning experiments to detect speech acts relying on deep learning architectures coupled with relevant linguistic features about how SA are linguistically expressed. We consider several experimental settings ranging from monotask to multitask learning including multi-label classification. Our results show that SA detection achieve very encouraging results proposing to the community a novel state of the art of SA detection in French social media.
- An error analysis of the automatic detection at both SA levels, highlighting main cases of mis-classification.

This paper is organized as follows. Section 2 presents related work in SA detection in social media as well as main existing crisis datasets. Section 3 provides the classification of SA we propose and the annotation guidelines to annotate them. Sections 4 and 5 respectively detail the dataset we relied on and the results of the annotation campaign. Section 6 focuses on the experiments we carried out to detect SA automatically. We end by some perspectives for future work.

2. Related Work

Speech acts have been extensively studied in the computational linguistics literature since early 2000's. Most studies focus on SA in human-human dialog conversations where several datasets have been annotated relying on various taxonomies of SA (also known as *dialogue acts*), such as QUESTION, ACKNOWLEDGMENT and FOLLOW-UP QUESTIONS (see Serban et al. (2018); Gonçalo et al. (2022) for recent surveys in the field). Dialogues being out of the scope of this paper, we focus in this section on SA for social media content, a relatively under-explored area of research compared to dialogue. We first provide an overview of SA used to annotate tweets about various events including crises as well as other domains (politics, offensive language, etc.). We then review main approaches for SA automatic detection. As our dataset for the first time combines SA and urgency annotations, we end this section by presenting existing crisis-related datasets highlighting the novelty of this study.

2.1 Speech Acts in Social Media

2.1.1 SPEECH ACTS IN THE CRISIS DOMAIN

The main line of analysis of the role of SA in tweets consists in unveiling how speech acts (as used on Twitter) vary qualitatively according to the *topic* discussed. In this line of questioning, SA have been studied as filters for new topics. Zhang et al. (2011) in particular, resorts to a Searlian typology of SA that distinguishes between assertive STATEMENTS (description of the world) and expressive COMMENTS (expression of a mental state of the speaker). Zhang et al. (2011) also distinguish between interrogative QUESTIONS and imperative SUGGESTIONS. Finally, a category MISCELLANEOUS brings together the Searlian DECLARATIVES and the COMMISSIVES, used to make promises. Concerning the question of emergency, Zhang et al. (2011) showed that the SA's distribution on Twitter in the context of a natural disaster (e.g., earthquake in Japan) is distinctive: it is essentially composed by statements, associated to comments and suggestions / orders. In this context new information or ideas on how to (re)act are indeed expected and assertions are the most suitable to this aim. By contrast, discussion over a celebrity will mostly generate comments and

almost no order or suggestion. Indeed, in this context, subjectivity matters more than immediate action.

Also inspired by Searle's typology, [Vosoughi \(2015\)](#); [Vosoughi and Roy \(2016\)](#) distinguish six categories: ASSERTIONS, RECOMMENDATIONS, EXPRESSIONS, QUESTION REQUESTS and MISCELLANEOUS. The authors use the definitions of [Zhang et al. \(2011\)](#), by distinguishing the *topic* discussed in the tweets, from the *type* of topic (*Entity-oriented*, *Event-oriented topics*, or *Long-standing topics* which are topics about subjects that are commonly discussed). Six topics were then selected (2 of each type): for *entity-oriented*, they are interested in Ashton Kuser and the Red Sox; for *event-oriented*, they studied the Boston bombings in 2013 and the Ferguson demonstrations in 2014; for *Longstanding topics*, they considered cooking and travel. The distribution of speech acts shows a greater similarity of distribution between topics of the same type than between topics of different types. On the other hand, the *entity-oriented* and *event-oriented* types are closer to each other, with a majority of assertions and expressions, whereas for the *long-standing* types, assertions are less abundant and recommendations well represented.

In this same perspective of topic identification and relying on the same topic characterization as above, [Elmadany et al. \(2018b\)](#) manually annotate 21,000 tweets in Arabic according to their topic type and distinguish events like Sinai bombings, Gulf crisis, Arab spring and world cup qualifications, entities (especially people) and various issues such as travel or cooking. Each tweet is associated to a pair of speech act/sentiment according to the following classification: ASSERTIONS, RECOMMENDATIONS, EXPRESSIONS AND REQUESTS, and among sentiments, the standard Positive, Negative, Mixed and Neutral categories. Their study reveals a salient association between assertions and people/events and neutrality on the one hand and an association between expressivity long-standing topics and negativity on the other.

2.1.2 SA IN OTHER DOMAINS

In a recent and extensive study of SA in social media, [Bell \(2020\)](#) takes on a different approach than other studies in the literature on Speech Act Theory and conducts an empirical investigation into the identity of illocutionary force indicating devices, which are the elements responsible for encoding a speaker's intentions. A corpus of 1,000 twitter threads is collected, manually segmented by an expert and annotated at the sub-tweet level, allowing multiple speech acts per tweet, as opposed to most other studies. They consider the following SA: ASSERTIVE, DIRECTIVE, INTERROGATIVE, EXPRESSIVE, COMMISSIVE, EXERCITIVE (with a commissive, the speaker commits themselves, with an exercitive, the speaker requires someone else's commitment). This study distinguishes direct and indirect illocutionary acts (i.e. acts performed by way of performing another). Regarding the direct force, the majority of segments (64.5%) were annotated as assertive. The second most frequent category was expressive, with 16.3%. On the other end, the least frequent category was exercitive, with 0.25%. Regarding the indirect force, 83.9% of tweets were determined to perform no indirect act, and those annotated as performing one were about 80% expressives.

In [Plakidis and Rehm \(2022\)](#), an annotation of SA is done using a subset of 600 tweets taken from a German corpora of offensive and non-offensive tweets. Mainly inspired by [Searle \(1975\)](#), and building upon [Compagno et al. \(2018\)](#) and [Weisser \(2018\)](#), the tweets are segmented in sentences, which are then annotated on two main levels : the syntactical level (eg. declarative, exclamative, imperative, etc.), which describes the type of sentence, and the speech act level, consisting of a

coarse-grained and a fine-grained level, which describes the type of speech act. The categories used for the first speech act level are as follows: ASSERTIVE, EXPRESSIVE, DIRECTIVE, COMMISSIVE, OTHER and UNSURE. They are subsequently detailed into 23 sub-classes at the sub-tweet level. For example, the category ASSERTIVE is further detailed into the following 6 sub-categories: ASSERT ("It costs 200\$"), SUSTAIN ("I'm going to buy it because it's very convenient"), GUESS ("I'm unsure he's right for her"), PREDICT ("It will be a few hundreds at most"), AGREE ("You are right"), DISAGREE ("I don't think so"). The results suggest that offensive language contains more expressives and less assertive than non-offensive language. Tweets with implicit offensive language have a lower frequency of expressives and a higher frequency of assertives than tweets with explicit offensive language.

In view of the topic – offensive language – the distinction between assertives and expressives is reported as a prominent issue, which does not arise in the context of urgency detection, where the description of facts (assertives) and the evaluation of said facts (expressives) are more clearly distinct.

For completeness, we note that SA have also been studied in the context of political campaigns, notably by [Subramanian et al. \(2019\)](#), with a corpus of 258 official documents related to the 2016 Australian "federal election cycle": official statements, tweets, press clippings, etc. from which 7641 utterances are extracted. Each utterance is annotated with a SA and a target party (liberal or conservative). The categorization of SA articulates: ASSERTIVES, COMMISSIVES-ACTION-SPECIFIC, COMMISSIVE-ACTION-VAGUE, COMMISSIVES-OUTCOME (about a future reality state), DIRECTIVES, EXPRESSIVES, PAST-ACTIONS and VERDICTIVES (an assessment on prospective or retrospective actions). They observe an over-representation of assertives (40%), followed by verdictives (25%) and specific action (12%). The other categories represent less than 10% of the annotations.

It is interesting to note that commissives make up almost a quarter of the assigned speech acts, whereas they are almost absent from our corpus, which is related to emergency.

2.1.3 SA AUTOMATIC DETECTION

SA prediction has been tackled either as a primary task (i.e., multi-class classification problem) or auxiliary task where SA information are used to boost the performances of classification tasks such as sentiment analysis, emotion detection or hate speech detection. Some works consider speech acts at the message level while others consider dialogue acts when uttered in conversations.

At the message level, most state of the art approaches make use of feature-based machine learning algorithms (SVM, Naive Baise, Decision Tree) relying on various surface, lexicon and syntactic features such as unigrams, punctuations, POS, emoticons and sentiment words ([Zhang et al., 2011](#); [Rojas-Barahona et al., 2012](#); [Franovic and Šnajder, 2012](#); [Vosoughi and Roy, 2016](#); [Sherkawi et al., 2018](#); [Algotiml et al., 2019](#)). Deep learning architectures have also been explored. [Saha et al. \(2021\)](#) propose a multi-modal approach for detecting SA in Arabic tweets relying on a multi-tasking framework based on dyadic attention mechanism ([Vaswani et al., 2017](#)) and adversarial loss to predict simultaneously sentiment, emotion and speech acts. It employs intra-modal and inter-modal attention to fuse multiple modalities and learn generalized features across all the tasks. [Subramanian et al. \(2019\)](#) propose a target based speech act classification on a dataset of political discourse using a semi-supervised learning approach (biGRU) by incorporating contextualized word representations

(ELMo) and a cross-view training framework to augment the initial dataset with in-domain unlabeled text. Finally, [Saha et al. \(2020a\)](#) combine BERT and capsule networks ([Sabour et al., 2017](#)) to assess the intent of tweets (expression, statement, suggestion, threat, request, question).

Another line of research focuses on predicting SA in social media conversational thread casting it into a sequence labeling problem. For example, [Cerisara et al. \(2018\)](#) use a two-level hierarchical recurrent network (Bi-LSTM and RNN) to predict dialog acts and sentiments. [Joty and Mohiuddin \(2018\)](#) experiment with an LSTM-RNN architecture to represent sentences of a conversation then CRF models to extract the inter-sentence dependencies. The approach has been evaluated on many synchronous and asynchronous corpora, including forum conversations from TripAdvisor. Other works propose to model SA in dialogues as a multi-label classification problem. For example, [Xu et al. \(2017\)](#) rely on a CNN model on top of pre-trained word vectors by utilizing a threshold learning mechanism. The model has been evaluated on the task of dialog state tracking.

2.2 Crisis Datasets

The literature on emergencies detection in social media has been growing fast in the recent years and several datasets (mainly tweets) have been proposed to account for crisis related phenomena such as flood, hurricane, storm and attacks.⁷ Messages are annotated according to relevant categories that are deemed to fit the information needs of various stakeholders like humanitarian organizations, local police and firefighters. Annotations are usually done at the text level relying either on crowd-sourced workers, humanitarian volunteers or domain experts.⁸ Relevance criteria found in the literature can be grouped into the following dimensions:

- *Relatedness* (also known as usefulness or informativeness) to identify whether the message content is useful provides valuable information that might be relevant to rescue teams. This is generally cast into a binary classification problem: is the message useful vs. non useful. This dimension is used in almost all state of the art annotation guidelines ([Imran et al., 2016](#); [Kaufhold et al., 2020](#)).
- *Urgency* (also known as criticality or priority) to filter out *on-topic relevant* information that can aid people in making decisions, advise others or offer immediate post-impact help, and *on-topic irrelevant* including offers, supports and solicitations for donations to charities ([Imran et al., 2013](#); [McCreadie et al., 2019a](#); [Sarioglu Kayi et al., 2020](#); [Kozlowski et al., 2020](#); [Kejriwal and Zhou, 2020](#)).
- *Intention to act*, also known as humanitarian information type ([Alam et al., 2021](#)). Urgency is often associated with a taxonomy of intention to act categories such as: caution or advice, donations, people missing, found, or seen and damage infrastructure ([Imran et al., 2016](#); [Olteanu et al., 2015](#)).
- *Eyewitnesses types*. It is used to identify direct (first-hand knowledge and experience of an event), indirect (messages sharing valuable information from direct witnesses) and vulnerable direct eyewitness (users reporting warnings and alerts) ([Zahra et al., 2020](#)). Annotations in most existing datasets are usually carried out at the message level.

7. See <https://crisisnlp.qcri.org/> for an overview.

8. Some studies propose to additionally annotate images within the tweets (see for example [Alam et al. \(2018\)](#))

Existing datasets are either annotated according to one of the dimensions above or using several dimensions in cascade like relatedness or urgency first, then information type for messages that have been identified as relevant. Most annotated datasets are in English. Well known datasets include TREC-IS ⁹ (McCreadie et al., 2019b, 2020), a shared task that aims to develop real-time monitoring systems capable of monitoring the development of incidents such as natural disasters, terrorist incidents or public health crises from online text data feeds. We also cite the CrisisFACTS2022 dataset ¹⁰ which aims at generating a summary of crisis. Few crisis datasets exist in other languages such as Spanish (Cobo et al., 2015), Arabic (Alharbi and Lee, 2019), Italian (Cresci et al., 2015). For French, the only publicly available dataset is the one developed by Kozłowski et al. (2020) who propose a three-level classification of tweets : Relatedness, Urgency and Intention to act categories to deal with missing people, human/infrastructure damage, etc. This dataset focuses on several natural disasters (hurricanes, flood, storms, etc.) going beyond the French portion of CrisisNLP ¹¹ that only focuses on one type of crisis (landslide).

2.3 Contributions

As far as we are aware, communicative intentions have been explored in connection with urgency detection in two previous works. First, Laurenti et al. (2022a) propose a SA classification for French tweet in the crisis domain. They focus on ecological crises and propose a two-layer annotation scheme to manually annotate a dataset of 6,669 tweets both for urgency (URGENT, NOT URGENT and NOT USEFUL) and SA (tweet level: ASSERTIVE, SUBJECTIVE, INTERROGATIVE and JUS-SIVE). Quantitative analysis of the annotations showed a correlation between tweet-level SA and urgency categories. This dataset has been used for supervised SA classification where a set of deep learning experiments have been carried out based on the CamemBert transformer architecture to classify each tweet into four SA categories at the tweet level. Laurenti et al. (2022b) built on this pre-trained classifier and propose SA-aware urgency detection models, showing that injecting SA as external semantic feature is a promising direction to improve urgency detection in social media.

In the present paper, we rely on the annotation scheme initially proposed in Laurenti et al. (2022a) and advance these previous studies, making six new contributions:

1. We double the dataset in Laurenti et al. (2022a) and create the largest French dataset annotated for SA with a total of about 13K at both the tweet and sub-tweet levels.
2. We extend to 6 new sudden crises, making the dataset cover both sudden and expected crises for a total of 20 events. This new dataset is the first that combines both SA and urgency annotations, to the best of our knowledge.
3. We correct the initial dataset addressing shortcomings related to the annotations at the sub-tweet level. We propose an automatic procedure to check for annotation inconsistencies which yields to a significantly improved version of the annotations.
4. In addition to the quantitative analysis we made on the initial portion of the dataset (6,6k) (i.e., SA vs. (urgent vs. non urgent vs. non useful)), we newly explore: (1) The correlation between SA and 6 intention to act categories, among human/infrastructure damages, warning advice,

9. <https://www.dcs.gla.ac.uk/~richardm/TREC-IS/>

10. <https://crisisfacts.github.io/>

11. <https://crisisnlp.qcri.org/>

critics, supports, etc. (2) The distribution of SA across crisis types (sudden vs. expected events), and (3) A study of the SA evolution across time.

5. In addition to SA detection at the tweet-level relying on baseline architectures, we newly:
(1) Address sub-tweet SA detection as well as joint tweet/sub-tweet predictions relying on monotask and multitask learning approaches while evaluating models performances to classify each message into a single class vs. multi-label. **As far as we know, handling SA in social media content** as a multi-label problem has not been explored before, (2) Experiment model adaptability across crisis types and layers. To the best of our knowledge, this is the first attempt to the automatic SA detection in a French social media dataset.
6. Finally, we provide a detailed error analysis of our results at both tweet and sub-tweet levels.

Overall this paper proposes an in-depth study of speech act in view of their contribution to enhance emergency detection. Before moving to real scenarios that rely on SA-aware automatic detection of urgency – which we leave for future work – our aim here is (a) unveil the contribution of speech act to emergency detection on a distributive basis, and (b) explore SA detection in French social media across various crisis types. This is, as far as we know, the first work that addresses the issue in such exhaustive manner. The second step that will consider injecting SA to improving urgency detection is out of the scope of this paper.

3. A Two-level Annotation of Speech Acts for Urgency Detection

We deployed two layers of annotation for speech acts:

- **SA1**: at the first level, we use a classification including 5 distinct categories, which we apply to the tweet as an atomic unit.
- **SA2**: at the second level, 8 categories are used to annotate tweets at the sub-tweet level as opposed to the tweet as a whole.

The goal of this two-layers annotation is to allow us to dig fine-grained information about speaker’s posture towards the event, to ultimately identify the main communicative intention of the tweet as a whole. In this section, all examples are taken from our corpus and provided in French together with their English translations. URLs and private user mentions have been replaced by <URL> and <USER> respectively. Each example comes with its SA1 (cf. Sections 3.1 and 3.2) and SA2 (cf. Section 3.2). Notation-wise, recall from Section 1.2 that we use arrows (→) to signify the relation between first-level and second-level SA categories, at the left and right of the arrow respectively. In addition, in order to show the interplay between SA and urgency annotations, all examples come with urgency annotations (URGENT vs. NOT URGENT vs. NOT RELEVANT) as well as six intention to act categories as follows: (1) URGENT applies to messages mentioning HUMAN, INFRASTRUCTURE DAMAGES as well as SECURITY INSTRUCTIONS to limit these damages during crisis events, (2) NOT URGENT groups SUPPORT messages to the victims, CRITICS or any OTHER MESSAGES that do not have an immediate impact on actionability but contribute in raising situational awareness, and finally (3) NOT URGENT for messages that are not related to the targeted crisis. Please note that all urgency annotations have been removed during the SA annotation campaign (cf. Section 4.1).

3.1 Tweet level

Our classification of SA elaborates on the foundational Austinian and later Searlian distinction by (i) relying on propositional content and lexical clues such as modals (*should, must, can, ...*), evaluative adjectives, attitude verbs (*think, believe, want, hope ...*); (ii) introducing the category SUBJECTIVE, which reshuffles some of the earlier classifications ('wishes', for instance are SUBJECTIVE rather than JUSSIVE in our classification (e.g., [Condoravdi and Lauer \(2012\)](#))); (iii) considering presuppositional content as well (see [Mari \(2016\)](#) on French).

We distinguish four first-level categories which are mutually exclusive and define tweets as wholes, at a holistic level, as shown in Figure 1.

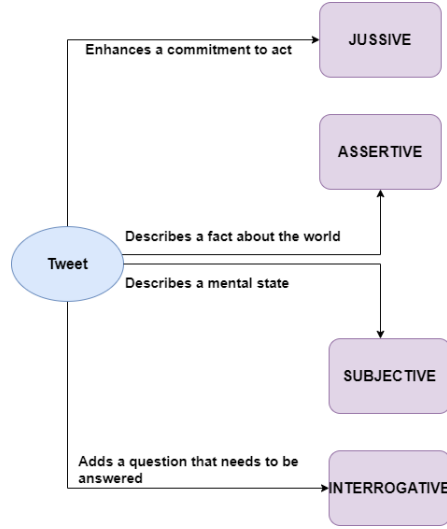


Figure 1: A classification for tweets that makes use of four illocutionary categories.

(1) **JUSSIVE**, as defined by [Zanuttini et al. \(2012\)](#), enhance commitment to take action, as in (2). Importantly, there is no strict correlation between the imperative form and JUSSIVE. As the example shows, the imperative form is not needed to enhance action. In this respect, our classification aligns with accounts that do not ground speech acts in sentence types (see [Portner \(2018\)](#) for extended discussion).

- (2) Incendies #Feuxdeforêt #Gironde 1. Ne pas se fier uniquement aux prévisions de Météo France 2. Si fumée lire le communiqué 3. Laisser les #SapeursPompiers effectuer leur rotation de 12 heures au feu 4. 96% des sinistres sont d'origine humaine (source SDIS 33) Merci <URL>
 (Wildfires #Forestfires #Gironde 1. Do not rely solely on Météo France forecasts 2. If smoke read the press release 3. Let the #FireBrigade carry out their 12-hour rotation at the fire 4. 96% of fires are of human origin (source SDIS 33) Thank you <URL>)
SA1: JUSSIVE
Urgency: URGENT → SECURITY INSTRUCTIONS

(2) **ASSERTIVE**. Assertions, like in (3), are considered to convey objective truth (as opposed to subjective truth ([Giannakidou and Mari, 2021c](#))). With ASSERTIVE, the speaker is committed

toward the truthfulness of the proposition that is being uttered ((Portner, 2018) a.o.) and require their interlocutor to update the common ground (Ginzburg, 2012).

- (3) DIRECT. Deux immeubles s’effondrent à Lille: les secours cherchent une victime dans les décombres <URL> via @lavoixdunord
(DIRECT. Two buildings collapse in Lille: rescue workers search for a victim in the rubble <URL> via @lavoixdunord)
SA1: ASSERTIVE
Urgency: URGENT → HUMAN DAMAGE

At this level of the classification, this is a simplification of what assertions are. When asserting, speakers can lie, or they can use a partial knowledge that undermines the likelihood of the assertion to express true. To nuance this simplification, we elaborate on the notion of assertion at the second level of the annotation, where we introduce some evidentiality-based distinctions.

(3) INTERROGATIVE. This category is dedicated to those questions that require an informative answer, like in (4) The questions that, besides triggering an answer, reveal bias and expectations on the part of the speaker (see Ladd (1981)) are classified as SUBJECTIVE (see below).

- (4) @EmmanuelMacron Où sont les renforts censés arrivés à Saint-Martin et que comptez-vous faire. #sxmirma #SaintMartin #Irma #sxmstrong #SXM
(@EmmanuelMacron Where are the reinforcements supposed to arrive in St. Martin and what are you planning on doing. #sxmirma #SaintMartin #Irma #sxmstrong #SXM)
SA1: INTERROGATIVE
Urgency: NOT URGENT → CRITICS

(4) SUBJECTIVE. Finally, with SUBJECTIVE, as in (5) the speaker shares a mental state that can be either a personal evaluation or preference (see among many others (Lasersohn, 2005)) or an expressive state (an emotion or a feeling, (Giannakidou and Mari, 2015)). The interlocutor is asked to update the common ground not just with the content of the evaluation but with the evaluation itself (see Simons (2007), and for recent discussion on French: Mari and Portner (2021)). In our classification, ‘wishes’, for instance, are SUBJECTIVE rather than JUSSIVE as they do not trigger any commitment to act so to make the content of the wish true (this is the emotive content of the wish (Giannakidou and Mari, 2021a)).

- (5) #incendie L’abbaye de Frigolet..la catastrophe... Un désastre..
(#fire at Frigolet Abbey..the catastrophe... A disaster...)
SA1: SUBJECTIVE
Urgency: URGENT → INFRASTRUCTURE DAMAGE

(5) OTHER. Additionally, OTHER is added to the classification, for undecidable cases.

- (6) Feu d’artifices du 14 Juillet @villedeputeaux <URL>
(14th of July fireworks @cityofputeaux)
SA1: OTHER
Urgency: NOT USEFUL

One important feature of our classification is that it does not rely on sentence type, but on sentence interpretation. For instance, an imperative is not necessarily classified as a JUSSIVE. Imperatives that convey wishes, as we noted, are considered to be SUBJECTIVE. Likewise, the interrogative

form, does not necessarily correlate with the interrogative category. An interrogative can express a point of view, or even knowledge, as in the case of rhetorical questions. In (7), the speaker is not really asking a question, but rather wants to express their opinion that the authorities are not doing the right thing, hence expressing a subjective point of view.

- (7) @Prefet974 #Berguitta .. Alerte orange pour rien hier qui a penalisé l'économie et pas d'alerte rouge pour ne pas pénaliser l'économie quand le danger est réel.. on marche sur la tête?

(@Prefet974 #Berguitta . Orange alert for nothing yesterday that penalized the economy and no red alert to not penalize the economy when the danger is real ... are we walking on our heads ?)

SA1: SUBJECTIVE

Urgency: URGENT → SECURITY INSTRUCTIONS

3.2 Sub-tweet Level

We consider each tweet as a discourse unit, composed of one or more statements or sub-segments, so that it can not only be classified at the holistic level but also at the level of its segments (identified in the following examples between '[...]'). In order to achieve this, we have elaborated on each of the four categories at the tweet level to annotate the tweets at the segment level relying on eight categories (see Figure 2).

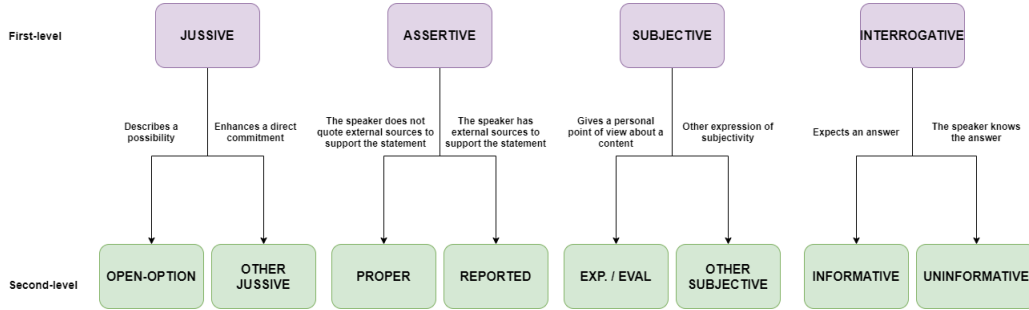


Figure 2: Two-layers annotation for tweets and inner segments.

For JUSTIVE, the annotation distinguishes between (a) **OPEN-OPTION** – the speaker puts forward a possibility and leaves the addressee free to realize it or not (cf. (8)) – , and (b) utterances that enhance a direct commitment on the part of a discourse participant, *ie.* **COMMISSIVES, EXHORTATIVES, ORDERS AND PROHIBITIONS**, that are called **OTHER-JUSTIVE** (cf. (9)).

- (8) [Cyclone #Irma : Qu'est ce que le CIC?]¹ [Présentation de cet outil de #Gescrise ci-dessous! <USER> <URL>]²

([Cyclone #Irma : What is the CIC?]¹ [Presentation of this #Gescrise tool below! <USER> <URL>]²)

SA2: 1. INTERROGATIVE→ UNINFORMATIVE, 2. JUSTIVE→ OPEN OPTION

Urgency: NOT URGENT → OTHER MESSAGES

- (9) [Un été caniculaire de tous les dangers avec des incendies dans plusieurs régions.]¹ [Alors redoublons de prudence et de vigilance. Pas de barbecues en forêt, de cigarettes allumées...]²
 ([A dangerous hot summer with fires in several regions.]¹ [So let's be extra cautious. No

barbecues in the forest, no lit cigarettes...]²)

SA2: 1. ASSERTIVE→ PROPER, 2. JUSSIVE→ OTHER JUSSIVE

Urgency: URGENT → SECURITY INSTRUCTIONS

For ASSERTIVE, both second-level categories are determined by the source of knowledge that the speaker relies upon, i.e. the evidentiality condition as defined by [Saurí and Pustejovsky \(2009\)](#). If the speaker grounds their utterance on a third-party source, the assertive utterance is (a) a **REPORTED ASSERTIVE**, whereas if there is no such explicit source, it is a (b) **PROPER ASSERTIVE**, see (10) and (11) respectively.

- (10) [Inondations dans l'Aude. Macron promet 80 millions d'euros:]¹ [est ce suffisant ???]²

([*Floods in the Aude. Macron promises 80 million euros:*]¹ [*is this enough???*]²)

SA2: 1. ASSERTIVE→ REPORTED, 2. INTERROGATIVE→ INFORMATIVE

Urgency: NOT URGENT → OTHER MESSAGES

- (11) [Le feu de Landiras (au départ à 40km) s'approche de chez moi. Encore 2 villages et c'est à nous d'évacuer. Ce soir on sent vachement le brulé.]¹ [On ne panique pas mais le stress monte. Vais mal dormir.]²

([*The Landiras fire (initially 40km away) is approaching my home. Two more villages and it's up to us to evacuate. Tonight it really smells like burnt.*]¹ [*No panic but the stress is mounting. I'm not going to sleep well.*]²)

SA2: 1. ASSERTIVE→ PROPER, 2. SUBJECTIVE→ EXPRESSIVE

Urgency: URGENT → HUMAN DAMAGE

It is important to note that the distinction between reported and proper assertive is meant to reveal a difference in degrees of commitment on the part of the speaker. On the assumption that a proper assertive reveals total commitment to the truthfulness of the content of the assertion, by signaling that the content of the assertion is reported, the speaker is considered as willing to distance themselves from the truth of that content (see discussion in [Aikhenvald \(2004\)](#); [Giannakidou and Mari \(2021a\)](#) and subsequent literature).

While we are aware that a certain amount of simplification remains (assertions can be lies, for instance (see extended discussion in [Giannakidou and Mari \(2021c\)](#)), this distinction allows us to introduce a certain degree of complexity in our treatment of the attitudinal domain.

For SUBJECTIVE, a distinction is made between (a) **EXPRESSIVES/EVALUATIVES** whereby the speaker describes a personal evaluation or an expressive state that it is not deemed to become common ground or truth (cf. (12))([Lasersohn, 2005](#); [Giannakidou and Mari, 2021c](#); [Mari and Portner, 2021](#)) and (b) **OTHER SUBJECTIVE** for utterances that do not explicitly fall in the previous category (eg: puns, greetings...), see (13).

- (12) [@Prefet29 Enfin !]¹ [Un grand bravo aux pompiers et aux agriculteurs qui sont venus aider à maîtriser cet incendie]²

([*@Prefet29 Finally!*]¹ [*Congratulations to the firefighters and farmers who came to help control the fire*]²)

SA2: 1. SUBJECTIVE→ EXPRESSIVE, 2. SUBJECTIVE→ EVALUATIVE

Urgency: NOT URGENT → SUPPORT

- (13) [Le paradis en feu.]¹ [Grosses pensées aux pompiers, policiers, bénévoles qui se battent s'en relâche depuis 1 semaine maintenant, nos cœurs sont serrés et nous prenons notre mal en patience.. #bassindarcachon #feuxdeforet #sud #DuneduPilat #IncendiesGironde

#incendies <URL>]²
 ([*Heaven on fire.*]¹[*My thoughts go out to the firemen, policemen, volunteers who have been fighting relentlessly for a week now, our hearts are heavy but we grin and bear it..* #bassindarcachon #forestfires #south #DuneduPilat #FiresGironde #fires <URL>]²)
SA2: 1. SUBJECTIVE→ OTHER SUBJECTIVE, 2. SUBJECTIVE→ EXPRESSIVE
Urgency: NOT URGENT → SUPPORT

The expressive/evaluative category is a complex one which can be enhanced by a variety of linguistic means, such as evaluative adjectives (including moral adjectives, *good*, *right*, epistemic adjective such as *clear*, *evident*) modality (*must*, *might*, *should*, *would* etc), adverbs (*obviously*, *regretfully*, etc.), particles (in French, *bien* (*ok*), *bon* (*good*), ...) (see Giannakidou and Mari (2021c)).

For INTERROGATIVE, a distinction is made between (a) **INFORMATIVE** questions to which the speaker cannot answer and which require an answer triggering new information and the ones that are (b) **UNINFORMATIVE** indicating that the speaker is biased towards an answer, as in (14) and (15) respectively.¹²

- (14) [Une semaine après le drame, on continue d’éclairer les zones d’ombre.]¹ [Pourquoi le médecin qui a trouvé la mort dans l’ #effondrement n’a pas été évacué ? #lille <URL> <URL>]²
 ([*One week after the tragedy, we continue to shed light on the grey areas.*]¹[*Why was the doctor who died in the #collapse not evacuated? #lille <URL> <URL>*]²)
SA2: 1. ASSERTIVE→ PROPER, 2. INTERROGATIVE→ INFORMATIVE
Urgency: NOT URGENT → OTHER MESSAGES
- (15) [Etes-vous en zone inondable ?]¹ [Retrouvez l’actualisation de la carte de prévention du #risque #inondation à #paris sur <URL>]²
 ([*Are you in a flood zone?*]¹[*Find the update of the #flood risk prevention map in #paris on <URL>*]²)
SA2: 1. INTERROGATIVE→ UNINFORMATIVE, 2. JUSSIVE→ OTHER JUSSIVE
Urgency: NOT URGENT → OTHER MESSAGES

As for SA1, we also add **OTHER** to the SA2 classification, for undecidable cases.

4. Data and Annotation

In this section, we provide details on the dataset used, the annotation procedure, and the results of the annotation campaign.

4.1 Dataset

Since our focus is on crises that occur in metropolitan France and its overseas departments, we rely on the only available corpus of French tweets by Kozłowski et al. (2020)¹³ and augmented later on by Bourgon et al. (2022b) with sudden crises (attacks, explosion, fires, etc.). The collection is composed of 19,595 tweets collected using dedicated keywords about ecological crises that occurred in France from 2016 to 2022 and posted 24h before, during (48h) and up to 72h after the crisis: 2

12. See Larrivé and Mari (2022) for French and Ginzburg (2012); Giannakidou and Mari (2021b) for a more general discussion and cross-linguistic observations.

13. https://github.com/DiegoKoz/french_ecological_crisis

floods that occurred in Aude and Corsica regions, 8 storms (Béryl, Berguitta, Fionn, Eleanor, Bruno, Egon, Ulrika, Susanna), 2 hurricanes (Irma and Harvey), 2 building collapses (Marseille, Lille), 2 chemical plants explosions (Lubrizol, Sanary), 2 fires (Notre-Dame fire, Gironde and Landes wildfires) and 1 terrorist attack (Trèbes).¹⁴ The data comes with additional metadata including: number of likes, retweets, followers and followings of the user.

In this dataset, each tweet is annotated following an urgency classification composed of three urgency categories as well as 6 intentions to act categories: (1) URGENT that applies to messages mentioning HUMAN/INFRASTRUCTURE DAMAGES as well as SECURITY INSTRUCTIONS to limit these damages during crisis events, (2) NOT URGENT that groups SUPPORT messages to the victims, CRITICS or any OTHER MESSAGES that do not have an immediate impact on actionability but contribute in raising situational awareness, and finally (3) NOT USEFUL for messages that are not related to the targeted crisis or information pertaining to events occurring outside the French territories. This scheme has been used to annotate the dataset by two annotators who achieved a Kappa inter-annotator agreement of 0.67 and 0.65 for urgency and intention to act classification respectively (Kozłowski et al., 2020).

	Not Useful	Urgent			Not urgent			Total
		Security instruc.	Human damage	Infra. damage	Support	Other messages	Critics	
Flood Aude	1,065	150	34	157	157	184	26	1,773
Flood Other	993	292	35	111	231	16	19	1,697
Flood Corse	468	51	58	12	52	66	13	720
Storm Beryl	612	91	0	2	3	10	2	720
Storm Bruno	586	107	5	11	2	9	0	720
Storm Susanna	484	129	11	38	4	54	0	720
Storm Ulrika	650	47	2	18	0	4	0	721
Storm Berguitta	587	56	5	9	12	46	5	720
Storm Fionn	552	138	6	10	0	8	6	720
Storm Egon	609	66	1	35	0	10	0	721
Storm Eleanor	590	82	22	19	1	6	0	720
Hurricane Harvey	628	78	10	2	1	1	0	720
Hurricane Irma	790	121	47	55	199	199	29	1,440
Collapse Marseille	627	9	24	11	11	19	19	720
Collapse Lille	320	2	39	27	12	117	32	549
Wildfire Gironde Landes	1,394	51	23	93	317	380	165	2,423
Wildfire Notre-Dame	86	224			209			519
Plant Explosion Lubrizol	137	583			627			1,347
Plant Explosion Sanary	6	363			164			533
Attack Trèbes	174	398			810			1,382
Total	11,358	3,970			4,257			19,595

Table 1: Urgency distribution in our dataset per crisis.

Table 1 presents the distribution by class for all available crises. Some crises (Plant Explosion Lubrizol, Plant Explosion Sanary, Notre-Dame Wildfire, Attack Trèbes) are only annotated for urgency. The ecological crisis (flood, storm, hurricane) are the most represented with 12,112 messages against 7,483 messages for sudden crisis (collapse, wildfire, plant explosion, attack). We also

14. For long-term crises, the end of the crisis has been fixed to the date of resolution of the crisis, e.g. extinction of the first fires in the case of fires land.

notice that for sudden crises, there are fewer SECURITY INSTRUCTION messages than ecological crisis, explained by the fact that these latter crisis are predictable.

The collection is extremely imbalanced with 57.96% NOT USEFUL and 20.26% for URGENT. This is largely due to how tweets are collected. Indeed, since tweets posted 24 hours before the crisis have been collected, a large amount of them are NOT USEFUL. The corpus is also imbalanced regarding the sub-level of urgency categories: 1.93% of the tweet are annotated as HUMAN DAMAGE with 306 messages while SECURITY INSTRUCTION represents 9.88% of the corpus with 1,470 tweets. These proportions are in line with the ones reported in other crisis corpora (see Section 2.1).

4.2 Annotation Procedure

A subset of this dataset composed of 13,378 tweets has been selected for SA1 annotations, among them 11,229 have been annotated for both SA1 and SA2. Regarding SA1 dataset, it comprises almost all URGENT (3,857) and NOT URGENT (4,222) messages. Only 5,299 NOT USEFUL tweets have been selected, in order to reduce the size of that category, but keep it as the majority class. Similar urgency annotations split holds for SA2 dataset. Note that, during the annotation process, pre-existing urgency tags and metadata information are removed, as to not bias the annotators.

The annotators were native French speakers, both master’s degree students in Linguistics. The procedure was as follows. First, each segment in a given tweet is annotated at the sub-tweet level (i.e., SA2), then the tweet level annotation (i.e., SA1) is deduced accordingly:

- If the tweet is composed of one or several SA2 annotations that subsume the same SA1 category, the final annotation is SA1. For example, for a tweet composed of two segments annotated with SA2=[INFORMATIVE, UNINFORMATIVE], then SA1=INTERROGATIVE.
- In case of several segments annotated with SA2 that do not belong to the same SA1 category, annotators are asked to determine the main communicative purpose of the tweet, and what segment signifies the main communicative intention of the speaker ((Simons, 2007; Mari and Portner, 2021) a.o.). The main criterion to identify the main intention relies on the determination of the background (known) - foreground (new) information. For example in (16), a tweet is composed of two segments: a PROPER ASSERTIVE, followed by an UNINFORMATIVE question that conveys an evaluation. The annotators have considered the second segment to be dominant, as the first half is a description of a fact that occurred in the past and that is already part of the common ground. The main point of the tweet is the uninformative question about the present situation, as an expression of a *criticism*.¹⁵ The tweet is thus labeled at the first level as SUBJECTIVE.

The SA2 annotation and the background-foreground distinction provides a solid heuristic to identify the main point of the tweet. Furthermore, as we shall see in Section 5.1, the first segment is mostly responsible for determining the overall categorization, this providing a reliable criterion to settle undecided cases. Finally, as we show in Section 5.3, specific sub-segments correlate with urgency, thus enhancing emergency detection.

- (16) [#Marseille - La mairie de Marseille a touché des millions pour la rénovation des immeubles urbains.]¹ [Qu’en ont-ils fait ?! @joelle_dago #GGRMC <URL>]²
 ([The Marseille city council has received millions for the renovation of urban buildings.

15. Recall that questions can convey a subjective stance rather than a request of information.

]¹[*What have they done with it?!]²)*
SA2: 1. ASSERTIVE→ PROPER, 2. SUBJECTIVE→ EVALUATIVE
SA1: SUBJECTIVE
Urgency: NOT URGENT→ CRITICS

The annotation has been performed using the BRAT annotation tool. (Stenetorp et al., 2012)¹⁶ To ensure consistency between annotations at the SA1 and SA2 levels (i.e., a tweet composed of one segment and annotated with SA1=INTERROGATIVE and SA2=PROPER ASSERTIVE), automatic checks have been conducted and annotators are asked to solve their errors before moving to the next tweet. Figure 3 shows an example of the tweet "A fire is currently in progress in #SaintDizier in the city center. Avoid the area" annotated in BRAT, highlighting both the tweet level (in red) and the sub-tweet levels SA annotations (in white).

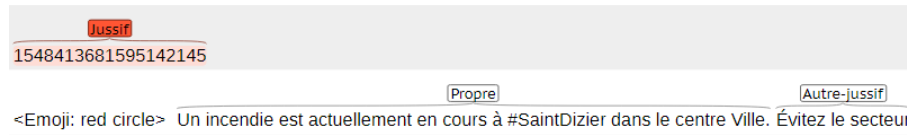


Figure 3: Example of a tweet annotated in BRAT. *Jussif* stands for JUSSIVE, while *Propre* and *Autre-jussif* for PROPER ASSERTIVE and OTHER JUSSIVE respectively.

The annotators performed a two-step annotation with an intermediate analysis of agreement and disagreement between the annotators. 448 tweets have been annotated in the first step by both annotators to compute the inter-annotator agreement (Cohen’s Kappa=0.62 for SA1 and 0.48 for SA2¹⁷). This agreement exhibits a comparatively lower score than what is typically encountered in similar studies involving SA annotations in tweets, with for example 0.78 in Vosoughi and Roy (2016) and between 0.72 and 0.92 depending on task in Subramanian et al. (2019). We found that it is mostly caused by the level of subjectivity involved in this task, in particular, the choice of the dominant segment, as mentioned earlier, has been the source of a lot of discrepancies. To address this issue, we encouraged regular feedback sessions and discussions between the two annotators to address discrepancies, clarify guidelines, and ultimately improve their agreement levels.

Another cause of disagreement were due to the difficulty of disentangling SUBJECTIVE from ASSERTIVE, in particular when attitudes and modal expressions are used such as *believe*, *think that*, etc. Indeed, both the subjective expressions (*think*, *believe*, or even more complex modal-tense-aspect combinations as *fallait* (which translates as ‘should have been’ with an additional implicature of preference in (17))) or its content can be targeted, according to their contextual relevance.

- (17) <USER> Et maintenant il n’y a presque plus de fumée... Il fallait arrêter le trafic ce matin et pas au milieu de la journée.
 (<USER> *And now there’s hardly any smoke... Should have stopped the traffic this morning, not in the middle of the day.*)
SA1: SUBJECTIVE
Urgency: NOT URGENT

16. <http://brat.nlplab.org>

17. We computed SA2 inter-annotation agreements on the basis of the dominant segment.

5. Results of the Annotation Campaign

We provide in this section a detailed analysis of the annotation campaign. We focus in particular on: (a) quantitative results of the SA annotations at both the tweet (SA1) and sub-tweet levels (SA2), (b) an analysis of how SA are expressed across different types of crisis, (c) the correlation between SA and urgency annotations, and finally (d) the evolution of SA over time since the crisis occurs. We end this section highlighting the main findings of this corpus-based study.

5.1 SA Annotations: Quantitative Results

Table 2 shows the distribution of categories of SA1 annotations (i.e., tweet level). We observe that a majority of the tweets are classified as ASSERTIVE, with 53.42%. The second-most frequent class is SUBJECTIVE, with 28.18% followed by JUSSIVE with 11.72%. INTERROGATIVE and OTHER are the less frequent with 3.36% and 3.32% respectively. These distributions indicate that in crisis situations, users predominantly tweet to assert their thoughts and views, to express their personal opinions and feelings, and to share information and updates on the given situation. Conversely, the low percentage of JUSSIVE and INTERROGATIVE suggests that they are less likely to give advice or ask questions in these circumstances (see also (Zhang and Liu, 2014)).

ASSERTIVE	SUBJECTIVE	JUSSIVE	INTERROGATIVE	OTHER	Total
7,147 (53.42%)	3,770 (28.18%)	1,568 (11.72%)	449 (3.36%)	444 (3.32%)	13,378

Table 2: Frequency of tweet level (SA1) annotations.

Figure 4 provides the distribution of the SA2 dominant labels (i.e., the ones that drive the SA1 annotations). We observe that PROPER ASSERTIVE is the most frequent with 37.19% while the other ASSERTIVE sub-class, namely REPORTED ASSERTIVE, was dominant in 14.36% of the tweets. Regarding NON ASSERTIVE content, EVALUATIVE and EXPRESSIVE SA2 annotations obtained similar frequencies of about 14.94% and 13.19% respectively.

Figure 5 combines the previous two tables illustrating the distribution of each SA2 sub-categories with their corresponding SA1 annotations. We observe that the pattern (SA1 = ASSERTIVE, SA2 = PROPER ASSERTIVE) is the most frequent with 72.13%. For INTERROGATIVE, 72.58% of the segments are INFORMATIVE vs. 27.42% for UNINFORMATIVE while for JUSSIVE, 63.17% are OPEN OPTION vs. 36.83% OTHER JUSSIVE. Similar observations hold for the two remaining SA1 categories. Finally, the very low percentage of OTHER (i.e., 0.43%) suggests that annotators were able to easily associate a SA2 category to a given segment. This is not the case for SA1 annotations where this frequency increases to 3.32% showing that sub-level SA annotations are important to better capture users’ communicative intentions. The number of OTHER SA2 annotations being relatively low (48 instances), we discard them for the further analysis below.

When analyzing tweet segmentation for SA2 annotations (recall that SA2 annotations consist of a sequence of segments $[s_1, s_2, \dots, s_n]$, each with its associated SA2 category), we observe (see Table 3) that, among the 11,229 tweets annotated for SA2, only about 23% are made up of more than one segment. Furthermore, 18.01% and 4.12% of tweets contain two and three segments respectively. While all SA2 classes display over 50% of presence in the first position, an interesting observation regarding the distribution of SA2 tags among possible positions is that it differs from class to class. Notably, while PROPER ASSERTIVE and REPORTED ASSERTIVE segments are over-



Figure 4: Distribution of SA2 annotations.

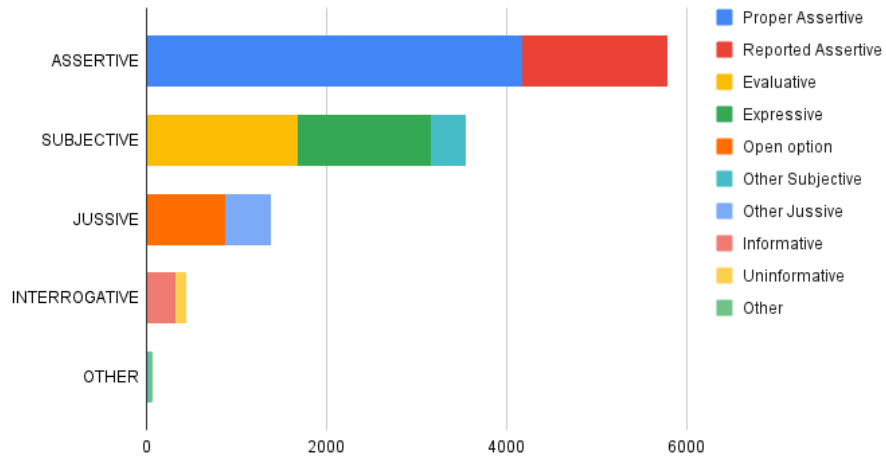


Figure 5: Distribution of SA1 and SA2 annotations in our dataset.

whelmingly found in the first position (over 93%), all other classes display a much higher rate of non-first position in the sequence, ranging from 24,27% for INFORMATIVE to 44,60% for OTHER JUSSIVE.

Table 3 together with Table 4 that shows the distribution of the most frequent sequences within a tweet, suggest that relying only on the first label in the case of multi-label sequences might be a viable approach. However, this approach should consider two potential difficulties. First, it could introduce a bias in favor of the dominant class PROPER ASSERTIVE, which tends to appear as the first element in multi-label sequences a lot more than the other classes in our data (about 98% of the time). Second, a specific pattern is identified, where a PROPER ASSERTIVE is followed by a different type of SA2 that is considered dominant, with the latter being in relation or in reaction to that initial assertion. In these cases, the reaction is the main, new, informative content that the rescue teams might be interested in, whereas the assertive content provides background information.

When analyzing the data further, we indeed observe that, for tweets composed of two sequences, the forms [PROPER ASSERTIVE, EVALUATIVE] and [PROPER ASSERTIVE, EXPRESSIVE] are a majority with 414 and 388 tweets respectively, followed by [PROPER ASSERTIVE, OTHER JUSSIVE] and [PROPER ASSERTIVE, OPEN-OPTION] with 189 and 169 tweets respectively. For tweets composed of three sequences, the patterns [PROPER ASSERTIVE, EVALUATIVE, EXPRESSIVE] and [PROPER ASSERTIVE, OTHER JUSSIVE, EVALUATIVE] have been observed in 88 and 44 tweets respectively.

Examples (18) and (19) illustrate of the observed patterns. In (18), the INTERROGATIVE is a direct follow-up to the assertion while in (19), the JUSSIVE is a reminder/directives given directly in reaction to the assertion. A final interesting observation concerns the OTHER class, where PROPER ASSERTIVE is not over-represented. This is likely due to the fact that this category is used to classify tweets that do not fit any of the other classes, and it should therefore be expected that such tweets follow a different pattern than the other classes.

SA2 / Position	1st	2nd	3rd	4th	5th	Total
PROPER ASSERTIVE	5,195	89	1	0	0	5,285
REPORTED ASSERTIVE	1,656	85	27	3	2	1,773
EVALUATIVE	1,092	774	183	18	4	2,071
EXPRESSIVE	1,272	548	204	42	0	2,066
OTHER SUBJECTIVE	304	208	26	1	0	539
OTHER JUSSIVE	422	372	30	8	2	834
OPEN OPTION	766	299	26	3	1	1,095
INFORMATIVE	252	91	26	3	3	375
UNINFORMATIVE	213	89	19	3	0	324
OTHER	57	11	2	0	0	70
Total	11,229	2,566	544	81	12	14,432

Table 3: Distribution of SA2 labels based on their position in the sequence.

- (18) [Antony inondations du 11 juin arrêté de catastrophe naturelle enfin sorti ! Les 700 habitations d’Antony sinistrées, doivent s’attendre à d’autres désordres vue la situation climatique. Le réservoir de Fresnes serait sous dimensionné, un autre doit être construit,]¹ [mais quand ?!]²

Sub-tweet SA sequences	%	
PROPER ASSERTIVE	30.86	46.13
PROPER ASSERTIVE + other(s) SA2	15.27	
EVAL./EXPR.	19.10	21.04
EVAL./EXPR. + other(s) SA2	1.94	
REPORTED ASSERTIVE	13.37	14.22
REPORTED ASSERTIVE + other(s) SA2	0.85	
OPEN-OPTION	6.15	6.84
OPEN-OPTION + OTHER(S) SA2	0.69	

Table 4: Distribution of the most frequent sequences in SA2 annotations.

([*Antony floods on June 11: natural disaster decree finally issued! The 700 Antony homes affected by the floods can expect further disruption, given the weather situation. The Fresnes reservoir is said to be undersized, and a new one is due to be built*]¹ - [*but when?!*]²)

SA2: 1. ASSERTIVE → PROPER, 2. INTERROGATIVE → UNINFORMATIVE

- (19) [Un été caniculaire de tous les dangers avec des incendies dans plusieurs régions.]¹ [Alors redoublons de prudence et de vigilance. Pas de barbecues en forêt, de cigarettes allumées...]²
 ([*A dangerous hot summer with fires in several regions.*]¹[*So let's be extra cautious. No barbecues in the forest, no lit cigarettes...*]²)

SA2: 1. ASSERTIVE → PROPER, 2. JUSSIVE → OTHER JUSSIVE

5.2 SA Annotations vs. Crisis Types

Our dataset is composed of 7 types of crisis, among them five 5 are unexpected or sudden events: Floods, Storms, Hurricanes, Building Collapses, Explosions and Fires/ Wildfires, and Terrorist Attacks. In this section, we analyze whether the type of crisis impacts the distribution of SA annotations. Table 5 shows the results.

Overall, the distribution is quite similar across all crises (some more fine-grained observation will be provided in section 5.4), and are inline of those observed in Tables 2 and 5. The only exception being the Trèbes Attack, with 39.36% ASSERTIVES (the lowest frequency of Assertives in the corpus) and 45.88% SUBJECTIVES (the highest frequency). Tweets posted during the Sanary Explosion displays the polar opposite distribution: 79.92% ASSERTIVES (the highest frequency of Assertives in the corpus) and 9.38% SUBJECTIVES (the lowest frequency of Subjectives in the corpus). Those two events, despite having both resulted in several deaths and injuries, have, according to the difference in SA distribution, elicited vastly different reactions on twitter. A possible interpretation is that, in the case of the incident in Sanary, users simply shared and discussed facts, as opposed to the terrorist attack in Trèbes, where users expressed their emotions and sentiments.

The types of crises seem to highlight certain tendencies related to SA1 annotations. For example, the distribution is quite similar between the 3 Floods: with 3 of the highest numbers of ASSERTIVES, averaging to 64.73%, and the 3 lowest numbers (besides Sanary) of SUBJECTIVES, averaging to 17.78%. Similarly, the distribution for the 2 Fires is such that both sub-corpus display, by quite a margin (besides Trèbes), the lowest numbers for ASSERTIVES and the highest for SUBJECTIVES with, respectively, 42.18% and 39.26%. Finally, the frequency of OTHERS is consis-

DIGGING COMMUNICATIVE INTENTIONS

		% ASS.	% SUB.	%JUS.	% INT.	%OTH.
Floods	Aude (1,002)	68.66	17.96	8.38	2.00	3.00
	Autre (1,001)	62.14	17.48	12.39	2.80	5.20
	Corse (404)	61.39	18.07	11.39	5.94	3.22
	Total (2,407)	64.73	17.78	10.55	2.99	3.95
Storms	Beryl (320)	55.00	26.88	7.19	3.44	7.50
	Bruno (350)	57.43	26.29	4.86	4.86	6.57
	Susanna (391)	59.08	23.53	11.51	1.53	4.34
	Ulrika (316)	53.80	18.99	13.61	2.22	11.38
	Berguitta (330)	56.97	22.12	10.61	4.55	5.76
	Fionn Corse (352)	67.33	19.60	7.95	1.70	3.41
	Egon (332)	55.72	28.31	7.23	3.31	5.42
	Eleanor (316)	66.14	21.84	8.23	1.90	1.90
	Total (2,707)	59.03	23.50	8.90	2.92	5.65
Hurricane	Harvey (304)	55.59	19.08	11.84	7.57	5.92
	Irma (893)	54.88	28.22	10.96	4.03	1.91
	Total (1,197)	55.02	25.88	11.18	4.92	2.91
Collapse	Marseille (304)	47.37	33.55	8.88	5.59	4.61
	Lille (549)	45.53	25.86	21.49	4.73	2.38
	Total (853)	46.14	28.59	16.98	5.04	3.25
Incidents	Lubrizol (1,357)	48.77	23.12	18.05	4.57	0.59
	Sanary (533)	79.96	9.39	10.51	0.00	0.19
	Total (1,890)	61.01	19.24	15.92	3.27	0.48
Fires	Landes (2,423)	42.51	39.99	12.55	3.76	1.20
	NotreDame (519)	40.66	35.84	3.08	2.89	17.53
	Total (2,942)	42.18	39.26	10.88	3.60	4.08
Attacks	Trèbes (1,382)	39.36	45.88	12.52	2.03	0.22
Total	Total (13,378)	53.35	28.14	11.65	3.34	3.52

Table 5: SA1 distribution per crisis type.

tently very low for the whole corpus (averages 3.32%), with two exceptions: Ulrika with 11.39% and NotreDame with 17.53%.

Finally, when looking into the distributions of SA2 annotations across crisis types, we observe that PROPER ASSERTIVES are the most frequent first segment of the sequence for all the crises except the building collapse and terrorist attack where REPORTED ASSERTIVES were a majority. We also observe a high proportion of INFORMATIVES, OPEN OPTIONS and EVALUATIVES. The distributions for the storms, hurricanes, terrorist attack but also fire crises are different with more EXPRESSIVES than EVALUATIVES.

5.3 SA vs. Urgency Annotations

Our dataset is annotated both for urgency and speech acts. All the tweets in our corpus (13,378) have been annotated for SA1 and urgency (i.e. URGENT, NOT URGENT and NOT USEFUL), whereas 11,229 have been annotated tweets for SA1, SA2 and intentions to act, namely SECURITY INSTRUCTION, HUMAN DAMAGE and INFRASTRUCTURE DAMAGE for Urgent messages, and SUPPORT, OTHER and CRITICS for Not Urgent messages.

SA1 vs. Urgency. Table 6 details the frequency of SA1 tags comparatively with the original urgency annotations. Regarding the two most frequent SA1 (ASSERTIVE and SUBJECTIVE), two observations emerge: (1) Among 3,857 URGENT messages (resp. 4,222 NOT URGENT), 86.13% (resp. 33.82%) are ASSERTIVE; and (2) only 5.81% of URGENT messages are SUBJECTIVE while 44.69% of NOT URGENT messages are. Similarly, we observe that 6.82% of JUSSIVE are URGENT vs. 15.99% NOT URGENT. Regarding NOT URGENT messages, ASSERTIVES mainly occur when messages contain information that is irrelevant to the crisis. It is interesting to note that the proportion of INTERROGATIVES are higher for NOT URGENT messages when compared to the URGENT ones (3.79% vs. 0.93%). Finally, among the 444 messages that have been annotated as OTHER messages, 81.08% are NOT USEFUL. These frequencies are statistically significant using the χ^2 test ($\chi^2 = 2, 831.84$, $df = 8$, $p < 0.01$). When measuring the dependency strength between urgency and SA1 categories using the Cramer’s V test, we get ($V = .32$, $df = 8$) which confirms the statistical correlation between these two classifications. These observations indicate a strong correlation between assertivity and urgency when removing the NOT USEFUL class ($V = .54$, $df = 4$).

%	URGENT	NOT URGENT	NOT USEFUL	Total
ASSERTIVE	86.13	33.82	45.23	53.42
SUBJECTIVE	5.81	44.69	31.31	28.18
JUSSIVE	6.82	15.99	11.89	11.72
INTERROGATIVE	0.93	3.79	4.77	3.36
OTHER	0.31	1.71	6.79	3.32
Total	100	100	100	100

Table 6: Urgency vs. SA1 annotation pairs statistics.

%	INF.	HUM.	SEC.	SUP.	CRI.	OTH.	NOT. USF	Total
ASSERTIVE	86.87	90.52	83.33	21.80	17.30	59.21	46.66	54.56
SUBJECTIVE	7.08	6.21	5.04	70.34	70.75	10.31	31.52	26.95
JUSSIVE	4.49	2.61	10.35	7.10	2.52	20.93	11.43	11.22
INTERROGATIVE	1.21	0.65	0.92	0.51	9.12	5.58	4.69	3.73
OTHER	0.35	0.00	0.35	0.25	0.31	3.97	5.70	3.55
Total	100	100	100	100	100	100	100	100

Table 7: Intention to act categories vs. SA1 annotations pairs statistics.

Table 7 provides the same analysis, this time with SA1 vs. intention to act annotations. For all URGENT subcategories, ASSERTIVE has the highest frequency with a total of 5,243 tweets, among them 90.52% are HUMAN DAMAGES, 86.87% INFRASTRUCTURE DAMAGES, and 83.33% SECURITY INSTRUCTIONS. Regarding the 2,416 NOT URGENT messages, SUBJECTIVE make up 70.75% of CRITICS and 70.34% of SUPPORTS vs. 17.30% and 21.80% for ASSERTIVES respectively. However, for the 1,309 OTHER MESSAGES, which are not urgent messages that do not fall in either of the previous two categories, only 10.31% of them are classified as SUBJECTIVE, while 20.93% are JUSSIVES, and 59.21% ASSERTIVES. These frequencies are statistically significant using the χ^2 test ($\chi^2 = 2, 502.17$, $df = 24$, $p < 0.01$). When measuring the dependency strength

between intention and SA1 categories using the Cramer’s V test, we get ($V = .25$, $df = 24$) which confirms the statistical correlation between these two classifications.

SA2 vs. Urgency. Table 8 presents the frequency of sub-tweet SA tags (excluding OTHER) when paired with urgency labels. In this table, the frequencies of SA2 are statistically significant ($\chi^2 = 2,378.84$, $df = 16$, $p < 0.01$, $V = .32$), showing that SA2 annotations are of particular importance for urgency detection.

%	URGENT	NOT URGENT	NOT USEFUL
ASSERTIVE			
REPORTED ASSERTIVE	52.73	24.60	40.21
PROPER ASSERTIVE	33.32	9.69	8.53
SUBJECTIVE			
EVALUATIVE/EXPRESSIVE	5.34	42.20	27.75
OTHER SUBJECTIVE	0.69	3.25	4.83
JUSSIVE			
OPEN OPTION	2.08	10.42	8.47
OTHER JUSSIVE	4.20	5.44	3.92
INTERROGATIVE			
INFORMATIVE	1.22	2.94	3.71
UNINFORMATIVE	0.24	0.92	1.68

Table 8: Urgency vs. SA2 annotations pairs (OTHER SA2 tags have been removed).

When looking into the distributions of SA2 tags against intentions to act categories (cf. Table 9), we again observe an over-representation of PROPER ASSERTIVE with a total of 3,133 instances, among them 153 are about INFRASTRUCTURE DAMAGES whereas 72 HUMAN DAMAGES. UNINFORMATIVE has the lowest frequency of 112 instances. Overall, the relationship between SA2 and the urgency categories suggests that the degree of urgency of a message is correlated to the type of speech act used ($\chi^2 = 2,928.24$, $df = 42$, $p < 0.01$, $V = .25$). The strength of the correlation increases to ($V = 0.40$, $p < 0.01$) when excluding the NOT USEFUL.

5.4 Evolution of Speech Acts Over Time

Recall that all the tweets in our dataset has been collected in three periods: 24h before, during (48h) and up to 72h after the crisis. Our aim here is to analyze the evolution of speech acts over time focusing on three periods: BEFORE, DURING and AFTER the event happened.¹⁸ Table 10 shows the distribution of SA1 categories per period in terms of percentage.

When looking at tweets over time since the crisis happens, we notice some interesting trends. Before a crisis, tweets are a mix of assertions and to a little extent subjective content. During the crisis, tweets become more focused and include a lot of strong statements and questions, showing people intend to provide informative content and express opinions and evaluations. After the crisis,

18. This three time periods have been determined to better meet the French Civil Security and Crisis Management Department’s specifications who perceives actionability in terms of emergency.

%	INF.	HUM.	SEC.	SUP.	CRI.	OTH.	NOT USF.
ASSERTIVE							
REPORTED	26.97	40.82	9.81	3.5	3.18	13.84	8.39
PROPER	57.3	48.98	70.98	18.03	14.33	47.95	41.2
SUBJECTIVE							
EVAL/EXPR.	9.74	7.48	2.71	67.18	63.7	10.38	27.4
OTHER-SUB	0.37	0	0.84	3.63	07.01	0.4	5.32
JUSSIVE							
OPEN OPTION	1.87	0.68	1.88	3.5	0.00	14.96	8.13
OTHER-JUSS	1.12	0.68	11.06	3.63	2.55	6.11	3.53
INTERROGATIVE							
INFOR.	2.62	1.36	1.67	0.13	7.64	3.94	3.5
UNINFOR.	0.00	0.00	01.04	0.39	1.59	1.93	1.73

Table 9: Intentions to act categories vs. SA2 annotation pairs (percentage of each SA2 category per intention category).

there is still a focus on sharing information, but fewer opinions are shared. After the crisis, assertive language remains substantial, suggesting a continued focus on conveying information. The proportion of subjective and interrogative tweets decreases post-crisis. This nuanced understanding highlights the shifting dynamics in communication styles across different phases of a crisis, with assertiveness and information-seeking becoming more pronounced during heightened situations. Jussives are observed as more prominent before the crises happens, which is in line with the interpretation of the jussive: the speakers intend to enhance action most notably when preventing casualties is still possible, that is to say, before the crisis happens.

%	BEFORE	DURING	AFTER
ASSERTIVE	10.51	26.05	15.89
SUBJECTIVE	7.68	15.01	6.57
JUSSIVE	6.24	2.64	2.8
INTERROGATIVE	0.71	1.73	1.02
OTHER	0.51	1.45	1.2

Table 10: SA1 annotation vs. crisis period, in percentage.

We further detail our analysis, this time by studying the distribution of SA1 per crisis type and period (see Table 11).¹⁹ In the case of storms, floods and hurricanes, there is a notable surge in assertive messages, particularly *after* the event, indicating a shift toward providing clear information and directives to address the aftermath. Concurrently, there is an increase in subjective expressions, possibly reflecting the emotional impact on individuals. Likewise, Collapse sees a notable increase in assertive messages post-crisis, suggesting a focus on clear statements and instructions once the immediate danger has passed. For Explosion/Attack and Fire-related communication, there is a

19. We removed the OTHER category from Table 11.

significant uptick in assertive messages *during* the event, possibly aimed at providing immediate guidance.

	Storm	Flood	Hurricane	Collapse	Explosion	Attack	Fire
ASSERTIVE							
BEFORE	2.32	0.87	2.28	0.28	0.00	0.00	0.28
DURING	4.04	1.33	0.99	0.60	9.47	4.47	9.62
AFTER	6.75	4.33	2.14	2.36	0.01	0.00	0.30
SUBJECTIVE							
BEFORE	0.86	0.21	0.81	0.30	0.00	0.00	0.29
DURING	1.83	0.54	0.49	0.44	2.99	5.21	8.71
AFTER	2.52	1.06	1.24	1.26	0.00	0.00	0.48
JUSSIVE							
BEFORE	0.37	0.11	0.47	0.09	0.00	0.00	0.18
DURING	0.74	0.27	0.15	0.25	2.47	1.42	2.36
AFTER	0.87	0.51	0.48	0.85	0.00	0.00	0.09
INTERROGATIVE							
BEFORE	0.11	0.07	0.21	0.07	0.00	0.00	0.02
DURING	0.23	0.09	0.03	0.09	0.51	0.23	0.78
AFTER	0.30	0.21	0.24	0.20	0.00	0.00	0.07

Table 11: SA1 annotation vs. crisis period vs. crisis type, in percentage. The two best scores for each period are in bold font.

5.5 Interim Conclusions

The corpus-based study of speech acts in tweets annotated for urgency allows for multiple statistically relevant observations:

- The vast majority of tweets are ASSERTIVES, seconded by SUBJECTIVES. More specifically, PROPER ASSERTIVES is the dominant class at the sub-tweet level. These results seem to indicate that, in a reaction to a crisis, French Twitter users mostly tweet to share information, generally in the form of a single utterance. This corroborates the findings in (Zhang and Liu, 2014), tending to show that in an emergency, factual information is more relevant than the expression of a personal view-point.
- PROPER ASSERTIVES are over-represented in the first position in every SA1 category of tweets. In particular, the high frequency of PROPER ASSERTIVES in the INTERROGATIVE, JUSSIVE and SUBJECTIVE tweets is explained by the fact that a significant part of those tweets follow a format comprising an assertion, followed by the speaker’s reaction to said assertion, which constitutes the dominant SA, as shown in (16). This reveals an interesting finding: In crisis situations, speakers tend to assert or re-assert a piece of (already known) information, followed by their personal comment in relation to it, thus sharing their perspective.

- The distribution of SA1 annotations highlights a general consistency in the data across the different crises, as well as similarities in the SA distribution of similar crises. Finally, we found a statistically significant relationship between ASSERTIVITY and URGENCY, and between SUBJECTIVITY and absence of URGENCY.

The picture that emerges, is one on which speakers favor (what they consider) truthful information over orders and commands to enhance action (on the part of the rescuing teams, for instance). Indeed, in our classification ASSERTIVES do not include subjective evaluations, and thus convey content informationally reliable and objectively veridical (i.e. conform to the outer reality and not a mental state) (Giannakidou and Mari, 2017, 2018, 2021c) and thus ready for uptake and endorsement (e.g. Ginzburg (2012), Krifka (2019)) on the part of those who will bring help. The fact that speakers favor PROPER ASSERTIVES to indicate urgency reveals that they are fully committed to the truthfulness of the message, of which they might present themselves as the primary informational source.

On the contrary, we observe that SUBJECTIVES correlate with absence of urgency. Among subjectives EVALUATIVES/EXPRESSIVES are largely used to convey truths that are relativized to a ‘judge’ or an individual (a.o. (Lasersohn, 2005; Stephenson, 2007)) and are not eligible to function as reliable information for the rescuing services. A minority of subjectives encompass attitudes, whereby truth is also relativized to a particular mental state and cannot (without further negotiation) immediately become common ground (e.g., (Gunlogson, 2008; Mari and Portner, 2021)) and be ready for uptake on the part of the helpers. Collapses, while qualifying as sudden crises, behave like non-sudden ones, probably in virtue of the long searches for casualties that make them similar to non-sudden ones.

Finally, we have discovered that assertives are more prominent after the crises when these are non-sudden, and during the crises when these are sudden. This points to the fact that speakers are active in providing information during the aftermaths of non-sudden crises, which, most of the times, require sustained efforts in view of the intensity of the damages. Speakers are keener in using assertives during the crisis with sudden crises, aiming at providing contentful information as the unexpected crisis unfolds and no knowledge had been made previously available by media or other sources.

6. Automatic Detection of SA

6.1 Experimental Settings

Now the dataset has been annotated, the next step is to automatically detect SA. We cast the problem into a classification task, leaving the complex task of discourse-based tweet segmentation into non overlapping units to future work (Morabia et al., 2019; Aljebreen et al., 2021). We propose the following experimental settings:

- **SA1 detection:** Classify each tweet into one of our five SA1 categories, namely ASSERTIVE, SUBJECTIVE, INTERROGATIVE, JUSSIVE, and OTHER.
- **SA2 detection:** Classify each tweet into one of our eight SA2 categories. Note that the OTHER instances (48 tweets) have been removed from the dataset for the experiments as they are very less frequent in urgent tweets and have no regular linguistic patterns. We propose two settings:

- Multi-class classification. Given a tweet $t = [s_1, \dots, s_n]$ and its associated SA annotations $SA1 = [SA2_1, \dots, SA2_j, \dots, SA2_n]$ where $SA2_j$ is the dominant segment, predicts its SA2 category $SA2_{pred}$. We evaluate the results considering (i) a strict match where $SA2_{pred} = SA2_j$ (this is similar to a binary classification), as well as (ii) a partial match such that $SA2_{pred} \in \{SA2_1, \dots, SA2_j, \dots, SA2_n\}$.
- Multi-label classification. Multi-class classification only focuses on the dominant segment ignoring the speech acts conveyed by the other tweet segments (we recall that a tweet can be composed up to 5 segments, see Table 3). We believe these informations can be of particular importance for urgency detection. Therefore, we aim at capturing label dependencies among the segments by assigning multiple labels for each instance simultaneously (Zhang and Zhou, 2013; Liu et al., 2021).
- **Detecting SA1 and SA2 simultaneously.** This is a multi-task learning framework considering there are two classification tasks (SA1 and SA2). The classifiers for both tasks share and update the same low layers except the final task-specific classification layer.

6.2 Models

We rely on **FlauBERT**_{base} (Le et al., 2019) the base cased French Transformers models (Martin et al., 2020) pre-trained on French texts from various sources from the general domain (e.g., Wikipedia and books), as implemented in HuggingFace. In addition, we use **FlauBERT**_{tuned}, a FlauBERT model that was pre-trained on 358,834 unannotated tweets from the crisis domain (Kozłowski et al., 2020) achieving better performances compared to FlauBERT for urgency detection. We also experimented with CamemBERT_{base} (Martin et al., 2019), the other French transformer architecture.²⁰

Following Laurenti et al. (2022b), we also experiment with two multi-input models that use extra-features added on top of pre-trained contextual word embeddings, among which²¹: the presence of URLs, punctuation (exclamation marks and question marks) and the presence of numbers, as they are often used in tweets to indicate phone numbers of emergency rescue services or weather forecast. We refer to these models as **FlauBERT**_{tuned+Feat} and **FlauBERT**_{base+Feat}.

In addition to the cross entropy loss (hereafter +C) and to fight class imbalance, we consider the focal loss (hereafter +focal) (Lin et al., 2017) or weighted cross entropy (hereafter +W). Our aim here is to compare with one of the most effective approach for handling imbalanced data (Cui et al., 2019). All our models were trained for four epochs with a learning rate of $2e - 5$, on top of which a linear layer for classification was added. For better convergence, we use the Adam optimizer during backpropagation. To avoid exploding gradients, we use a gradient clipping of 1.0.

For the multi-label task, we adapt the FlauBert architecture²² to account for multilabel outputs relying on a sequence classification head on top of the pool layer. The input sequence comprises characters, sub-words, and words, which are processed by the transformer layers. On top of the pooled output, a linear layer is added for the classification task. We then examine each label independently for every message and determine whether the label is predicted by the model. We rely on label-based metrics (F1 macro) following the general trend in multi-label classification (Zhang and

20. The majority baseline achieved an accuracy of 0.534 and 0.372 for SA1 and SA2 classifications, respectively.

21. We also tested several other features including tweet meta-features, sentiment and emoticons, number of imperatives verbs, etc., but the results were not conclusive.

22. We used FlauBert as the results achieved by CamemBert on this task were lower.

Zhou, 2013). This architecture was successfully employed in multilabel classification in other NLP tasks including judicial documents (Dai and Liu, 2020), sentiment analysis (Tang et al., 2020) and diagnoses of patients prediction (Hart, 2022).

6.3 Evaluation Protocol

To evaluate SA1 and SA2 models, we designed two evaluation protocols:

- *Random sampling.* We mixed the tweets for all the crisis and randomly select 80% for train and 20% for test. For SA1 classification (resp. SA2), the final dataset is composed of 11, 181 tweets (resp. 13, 378) split into 80-20 for train-test while keeping the same distribution as in the train set.
- *Out-of-event.* Following the general trends in crisis management (Kersten et al., 2019; Al-giriyage et al., 2021; Bourgon et al., 2022a), we designed an out-of-type evaluation protocol by training on a pool of events related to different types of crises (e.g., Hurricane, Storm) and testing on a particular different type (e.g., Earthquake). The aim is to evaluate if a model can deal with new types of crisis, which is crucial to ensure the portability of the models to unseen events. To this end, we consider the distinction between *expected* vs. *sudden* events and experiment whether the use of speech acts differ according to the type of crises. Indeed, compared to ecological disaster like hurricanes and floods, sudden events (like earthquakes, terror attacks, explosions, technological incidents) are difficult to predict (Björck, 2016). These events, over which organizations have virtually no control, influence social behavior and the ways the emergency services are organized (James and Wooten, 2005; Coombs, 2014; Quarantelli et al., 2017). We propose two evaluation settings: (a) Train on expected events and test on sudden, (b) Train on sudden and test on expected. We consider flood, storm and hurricane as expected events while collapse, wildfire, plant explosion and attack as sudden events (see Table 1), which corresponds to a total of 6,311 tweets for the former vs. 7,067 for the latter.

All the SA1 (resp. SA2) models have been run five times on a randomly selected instances from the test set with a standard deviation of results being 5.3×10^{-6} (resp. 7.8×10^{-4}). We therefore report the averaged scores (accuracy, precision, recall, and macro F1). Finally, due to the high number of experiments, we only provide those achieved by the best configurations. Below we present our results. We end by a qualitative analysis highlighting main causes of misclassification.

6.4 Results

6.4.1 RANDOM SAMPLING RESULTS

The experimental results are presented in Table 12, showcasing the accuracy (A), precision (P), recall (R), and macro-averaged F1-scores (F1). The results are grouped according to mono vs. multitask learning and whether these models use extra-features. Best scores are highlighted in bold font.

The results show that FlauBERT_{tuned} has consistently achieved the best scores across all settings and that mono-task learning models outperform its multitask counterpart. However, it is interesting to note that FlauBERT_{MultitaskTuned+C+Feat} resulted in the best accuracy of 81.81%. Injecting additional features was very helpful when coupling with the focal loss for FlauBERT_{tuned+focal+Feat} and

Models	A	P	R	F1
CamemBERT _{base+W}	79.04	69.20	65.67	66.93
FlauBERT _{base+focal}	78.29	69.53	61.88	64.96
FlauBERT _{tuned+C}	79.15	68.87	65.19	66.81
FlauBERT _{base+C+Feat}	78.66	70.95	62.79	65.58
FlauBERT _{tuned+focal+Feat}	78.59	71.06	65.02	67.37
FlauBERT _{MultitaskBase+C}	62.98	53.01	49.33	49.76
FlauBERT _{MultitaskTuned+C}	80.69	60.69	57.57	58.98
FlauBERT _{MultitaskBase+C+Feat}	79.17	59.43	56.23	57.65
FlauBERT _{MultitaskTuned+C+Feat}	81.81	61.21	58.28	59.60

Table 12: SA1 classification results.

cross entropy for FlauBERT_{MultitaskTuned+C+Feat}. When looking into the detailed results per class (cf. Table 13), we observe that the predictions are closely aligned with the distribution of each class in the dataset. In particular, ASSERTIVE and SUBJECTIVE were well-predicted with an F-score of 85.32 and 76.14 respectively, whereas JUSSIVE and INTERROGATIVE exhibit lower scores.

	Precision	Recall	F1-Score
ASSERTIVE	84.46	86.19	85.32
SUBJECTIVE	75.47	76.82	76.14
JUSSIVE	64.26	64.47	64.37
INTERROGATIVE	63.24	55.84	59.31
OTHER	67.86	41.76	51.70
Accuracy = 78.59			

Table 13: SA1 results per class as given by FlauBERT_{Tuned+focal+Feat} our best model.

Level	Models	A	P	R	F1
Monotask strict	FlauBERT _{base+C}	66.58	56.76	52.06	52.79
	FlauBERT _{tuned+W}	67.72	59.56	56.98	57.82
	FlauBERT _{base+C+Feat}	65.49	57.55	51.88	51.90
	FlauBERT _{tuned+W+Feat}	67.95	56.90	55.12	55.44
Monotask partial	FlauBERT _{base+C}	73.71	63.46	60.23	60.42
	FlauBERT _{tuned+W}	74.74	67.91	64.74	65.67
Multitask	FlauBERT _{MultitaskBase+C}	59.83	48.64	39.69	41.15
	FlauBERT _{MultitaskTuned+C}	67.59	59.76	53.09	55.17
	FlauBERT _{MultitaskBase+C+Feat}	65.71	52.78	50.53	50.05
	FlauBERT _{MultitaskTuned+C+Feat}	67.99	60.31	53.51	54.72
Multi-Label	FlauBERT _{base+C}	94.68	96.35	82.85	87.80
	FlauBERT _{tuned+C}	88.11	78.17	57.50	62.36

Table 14: SA2 classification results.

The results of the fine-grained SA experiments are presented in Table 14. The results indicate that partial evaluation leads to an improvement of approximately 8% in terms of accuracy with FlauBERT_{tuned+W} achieving the best performance, resulting in an F-score of 65.67. This shows that a strict evaluation is not suitable to determine the dominant segment which is predictable given the pragmatic nature of selecting these types of segment. Regarding multitask architectures, the results are inline with those observed with SA1, making them less productive for multi-level SA detection. More importantly, the multi-label classification was the best, with FlauBERT_{base+C} yielding the highest scores. Finally, although features have been very productive for SA1 classification, their injection into the FlauBERT architecture for SA2 achieved mitigated results, see for example the boost in the F1-scores achieved by FlauBERT_{MultitaskBase+C+Feat} vs. FlauBERT_{MultitaskBase+C} while we observe a drop when comparing FlauBERT_{MultitaskTuned+C} vs. FlauBERT_{MultitaskTuned+C+Feat}.

We end this section by detailed results per class, as given by the multi-label model (cf. Table 15). Overall, the model achieves very good results for all the classes except the less frequent (see for example OTHER SUBJECTIVE and UNINFORMATIVE).

	Precision	Recall	F1-Score
REPORTED ASSERTIVE	100	96.78	98.37
PROPER ASSERTIVE	91.27	99.81	95.35
EXPRESSIVE/EVALUATIVE	98.22	93.96	96.02
OTHER-SUBJECTIVE	100	43.14	60.27
OPEN-OPTION	100	92.36	96.03
OTHER-JUSSIVE	98.36	90.91	94.49
UNINFORMATIVE	100	64.29	78.26
INFORMATIVE	81.13	70.49	75.44
Accuracy = 94.68			

Table 15: SA2 results per class as given by the multi-label FlauBERT_{base+C}.

6.4.2 OUT-OF-EVENT RESULTS

Table 16 shows the results of our best SA1 (resp. SA2) models when tested following the out-of-event protocol, i.e., FlauBERT_{tuned+focal+Feat} (resp. FlauBERT_{base+C}), in terms of precision (P), recall (R) and the averaged F1-score (F1).

For SA1, and when compared to random sampling, we observe a small drop in the performances and this is more salient when the model is trained on sudden events. When we look into the results per class, we notice that three out of the five SA1 categories achieved similar results when trained vs. tested on expected events: 82.30 vs. 85.40 F1-score for ASSERTIVE, 51.58 vs. 54.16 for INTERROGATIVE, and 59.76 vs. 60.11 for JUSSIVE. Note that these scores are close to the one reported in Table 13 where the SA1 model has been evaluated in a random sampling scenario. OTHER and SUBJECTIVE however exhibits a different behavior: OTHER scores 44.44 vs. 31.58 resulting in an important decrease in performances up to 7.3% in terms of F1-score when compared to random sampling. For SUBJECTIVE, the drop depends on the test set: When trained on expected events and tested on sudden, this category achieved around -6% compared to a random test. On the other hand, when trained on sudden and tested on expected events, the scores were similar (76.85 vs. 76.14 in the random configuration).

This drop can be explained by the diverse linguistic means by which speech acts are expressed in sudden events (we recall the the distribution of each class in both settings are quite similar (see Table 5)). The SUBJECTIVE class encompasses a series of expressions that belong to different grammatical categories (verbs, adjectives, particles, interjections, ...) at different levels of the semantic and pragmatic interpretation (sentence, discourse, ...) and can either have an informational or just an expressive function. A finer grained typology of expressions is to be established to pin down the linguistic differences between the subjective expressions involved in sudden crises and those used in non-sudden crises situations.

Finally, regarding the SA2 results, we notice that testing on expected (resp. sudden) events does not significantly impact the results when compared to the random sampling. This shows that casting SA2 detection as a multi-labeling problem is quite effective.

	RANDOM	EXPEC. → SUDDEN			SUDDEN → EXPEC.		
	F1	P	R	F1	P	R	F1
SA1: FlauBERT _{tuned+focal+Feat}	67.37	68.18	59.81	62.98	65.86	60.88	60.32
SA2: FlauBERT _{base+C}	87.80	94.62	76.90	82.35	94.35	83.23	87.47

Table 16: SA1 (resp. SA2) best models results when tested in the out-of-event protocol.

6.5 Error Analysis

We end this paper by analyzing most causes of misclassifications as given by our best SA1 (resp. SA2) models, namely FlauBERT_{tuned+focal+Feat}, the fine-tuned FlauBERT with focal loss and feature injection (resp. the multi-label FlauBERT_{base+C} trained with a cross entropy loss). Figure 6 presents our results. We also provide the confusion matrix for SA1 classification (see Table 17).²³

It shows that most errors come from the difficult distinction between ASSERTIVE and SUBJECTIVE, as also observed during the annotation campaign (see Section 4.2). In practice, the objective-subjective distinction may not always be clear-cut, leading to a preference for ASSERTIVE classification, see the examples (9) and (10) in Table 18. We also observe other complex cases like in (2), (3), (7) and (8) that contain declarative statements and for which the model fails to distinguish between assertives and jussives. Some other examples lack context, and the gold label may not always be accurate (e.g., (1)), but they can still express assertive sentiments (e.g., (4)). Finally, the model struggles with some interrogative texts that are phrased declaratively (e.g., (5)) or affirmatively (e.g., (6)), leading to difficulties in identifying them as interrogative, therefore the model has trouble taking into consideration the interrogative mark present at the end of the text.

	ASSERTIVE	SUBJECTIVE	JUSSIVE	INTERROGATIVE	OTHER
ASSERTIVE	0	119	66	6	8
SUBJECTIVE	135	0	31	14	2
JUSSIVE	59	35	0	6	8
INTERROGATIVE	13	19	2	0	0
OTHER	18	23	10	2	0

Table 17: SA1 misclassification results.

23. We only report the confusion matrix for SA1 classification as the one for SA2 is too sparse.

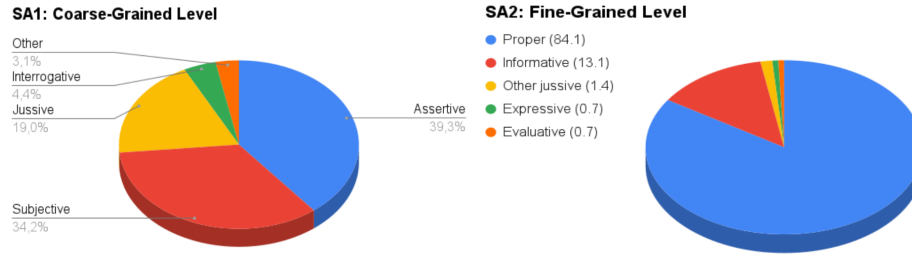


Figure 6: Distribution of misclassified examples.

	Tweet	Predicted	Gold
1	Tempête #nekfeu (Storm #nekfeu)	OTH.	ASSERT
2	Aurore Bergé sur les inondations: "Il faut donner des cours de natation aux Français" (Aurore Bergé on the floods: "We need to give swimming lessons to the French.")	JUSS	ASSERT
3	regarde le prix des billets pour la Guadeloupe (Look at the price of tickets to Guadeloupe)	JUSS	ASSERT
4	Ouragan #Irma : l'inquiétude des Antillais de métropole #BourdinDirect (Hurricane #Irma: Concerns of Antilleans in mainland France #BourdinDirect)	ASSERT	OTH.
5	J'suis le seul a quitter une reunion en plein milieu parce qu'elle ne m'intéresse pas? (Am I the only one who leaves a meeting in the middle because it doesn't interest me?)	ASSERT	INTER
6	- Toulouse prêt pour le déluge ? (Is Toulouse ready for the deluge?)	ASSERT	INTER
7	#Rouen sous la fumée noire de l'entreprise #lubrizol :: URL (#Rouen under the black smoke of the company #lubrizol :: URL)	ASSERT	JUSS
8	Les mégapoles face aux risques d'inondation, passé / présent (Megacities facing flood risks, past and present.)	ASSERT	JUSS
9	Journée noire à #Rouen après l'incendie de l'usine #Lubrizol #pollution URL (Black day in #Rouen after the fire at the #Lubrizol factory #pollution URL)	ASSERT	SUBJ
10	nan mais le temps ici y'a deux secondes c'était inondation et là y'a du soleil (But the weather here, two seconds ago, was flood and now it's sunny.)	ASSERT	SUBJ

Table 18: SA1 misclassified examples.

For SA2, more than 84% of the misclassified tweets were due to the PROPER ASSERTIVE class, followed by INFORMATIVE. Table 19 presents some representative examples of such errors. It appears that the model fails to distinguish between proper assertives and subjective (expressive or evaluative) content, as seen in examples (1) and (2). We observed that the classifier is misguided by the interrogation (e.g., (7)) and exclamation (e.g., (5) and (6)) marks. Other examples lack context and present some grammatical errors (e.g., (4)). Finally, the model is challenged with some of the interrogative sentences, one possible reason is that these often have a different syntactic structure and may resemble declarative sentences (e.g., (9) and (10)).

7. Conclusion

In this paper, we presented the first corpus-based study to measure the impact of speech acts in messages posted in social media during various types of crises. We first proposed a new annotation guideline to annotate speech acts both at the tweet and sub-tweet levels, then a new dataset annotated

	text	Predicted	Gold
1	Ce que l'on sait de l'incendie à Rouen qui ravage l'usine Lubrizol URL (Rouen's Lubrizol factory fire: what we know URL)	OTH-JUSS	PROPER
2	Regarder la nuit tomber c'est dans mon top 3 des activités (Watching the night fall is in my top 3 activities.)	PROPER	EXPR./EVAL.
3	Calmez vraiment la tempête (Really calm the storm.)	EXPR./EVAL.	OTH-SUBJ
4	T'es bon pour intéresser aussi au Papi (Programmes d'Actions de Prévention des Inondations). (You're good to also be interested in the Papi (Programs of Action for Flood Prevention).)	INFO	EXPR.
5	Wow! RT @PellepX3: Les premières images de l'inondation à Montréal le 29 mai #inondation (Wow! RT @PellepX3: The first images of the flooding in Montreal on May 29 #flooding)	INFO	EVAL
6	@dragonduclos Qui sème le vent récolte la tempête ! (@dragonduclos Who sows the wind reaps the storm!)	INFO	OTH-SUBJ
7	USER RT @PellepX3: Et toi tu t'effondres quand sous le poids de tes âneries ? (And you, do you collapse under the weight of your nonsense?)	INFO	OTH-SUBJ
8	Une sinistrée des inondations demande que Macron apporte "des solutions" (A flood victim asks Macron to bring "solutions".)	PROPER	RAPPORTED
9	Hello @BFMTV ça vous intéresse des vidéos de l'incendie de la forêt de Brocéliande faites avec un drone ??? (Hello @BFMTV, are you interested in videos of the fire in the Brocéliande forest made with a drone???)	OTH-JUSS	INFO
10	Le rétablissement de la continuité écologique des zones humides ne fait-il pas oublier le risque d'inondation ? (Does the restoration of ecological continuity in wetlands not forget the risk of flooding?)	PROPER	UNINFO

Table 19: SA2 misclassified examples.

for both speech acts and urgency categories in French. We conducted a deep corpus-based analysis of the correlation between SA and urgency, SA and intention to act categories, SA and crisis type, and finally, SA evolution over time since the event happens. Our results show a strong correlation (i) between Assertive messages (particularly those that rely on first hand knowledge, i.e. PROPER ASSERTIVES) and urgency, (ii) Subjective messages and absence of urgency, with a high frequency of expressives and evaluatives. In addition, we found a strong correlation between ASSERTIVES and HUMAN/INFRASTRUCTURE DAMAGES.

We finally conducted a set of experiments to detect SA relying on transformer architectures augmented with dedicated features. We propose a set of monotask and multitask learning settings to classify a given tweet in either tweet-level speech act or sub-tweet level speech act categories casting the problem into a multi-label and multi-class task. We also experiment models portability to unseen events to measure SA detection performances in real scenario. Our results are encouraging and constitute a new state of the art of speech act detection in French tweets.

The next step now is to inject SA information while detecting urgency. Our preliminary study on SA-aware urgency detection were very encouraging (Laurenti et al., 2022b). These experiments were however conducted on a small set of 6,6K tweets and only focusing on tweet-level speech acts. We plan to extend this work by injecting sub-tweet speech acts as well through new deep learning architectures in a multilingual setting.

Acknowledgment

This work has been supported by the INTACT project funded by a CNRS pre-maturation grant. Alda Mari gratefully acknowledges ANR-17-EURE-0017 FrontCog. The research of Farah Benamara is also partially supported by DesCartes: the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CRE-ATE) program.

Authors contributions

The order of the authors is alphabetical to reveal contributions of comparable importance. Farah Benamara and Alda Mari are the project leaders, and have furthermore coordinated the research and the writing.

References

- Alexandra Aikhenvald. *Evidentiality*. Oxford University Press, Kettering, Northamptonshire, UK, 2004.
- Firoj Alam, Ferda Ofli, and Muhammad Imran. CrisisMMD: Multimodal Twitter Datasets from Natural Disasters. *arXiv:1805.00713 [cs]*, may 2018. URL <http://arxiv.org/abs/1805.00713>.
- Firoj Alam, Hassan Sajjad, Muhammad Imran, and Ferda Ofli. Crisisbench: Benchmarking crisis-related social media datasets for humanitarian information processing. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15 of *ICWSM '21*, pages 923–932, May 2021. URL <https://ojs.aaai.org/index.php/ICWSM/article/view/18115>.
- Nilani Algiriyage, Rangana Sampath, Raj Prasanna, Emma EH Doyle, Kristin Stock, and David Johnston. Identifying disaster-related tweets: a large-scale detection model comparison. In *Social Media in Crises and Conflicts, Proceedings of the 18th ISCRAM Conference*, pages 731–743, 2021. URL http://idl.iscram.org/files/nilanialgiriyage/2021/2368_NilaniAlgiriyage_etal2021.pdf.
- Bushra Algotiml, AbdelRahim Elmadany, and Walid Magdy. Arabic tweet-act: Speech act recognition for Arabic asynchronous conversations. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 183–191, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4620. URL <https://aclanthology.org/W19-4620>.
- Alaa Alharbi and Mark Lee. Crisis Detection from Arabic Tweets. In *Proceedings of the 3rd Workshop on Arabic Corpus Linguistics*, pages 72–79, 2019. URL <https://www.aclweb.org/anthology/W19-5609>.
- Abdullah Aljebreen, Weiyi Meng, and Eduard Dragut. Segmentation of tweets with urls and its applications to sentiment analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12480–12488, May 2021. doi: 10.1609/aaai.v35i14.17480. URL <https://ojs.aaai.org/index.php/AAAI/article/view/17480>.
- James Allen and Mark Core. Draft of dams1: Dialog act markup in several layers. 1997. URL <http://www.fb10.uni-bremen.de/anglistik/ling/ss07/discourse-materials/DAMSL97.pdf>.
- Nicholas Asher and Alex Lascarides. Commitments, beliefs and intentions in dialogue. In *Proceedings of the 12th Workshop on the Semantics and Pragmatics of Dialogue*, pages 29–36. Citeseer, 2008.

- John Langshaw Austin. *How to do things with words*. Oxford University Press, Kettering, Northamptonshire, UK, 1962.
- Kent Bach and Robert M Harnish. *Linguistic communication and speech acts*. MIT Press, 1979.
- Laura Beth Bell. *Illocution on Twitter: The Construction and Analysis of a Social Media Speech Act Corpus*. Georgetown University, Washington, D.C., USA, 2020.
- Albena Björck. Crisis Typologies Revisited: An Interdisciplinary Approach. *Central European Business Review*, 2016(3):25–37, 2016. doi: 10.18267/j.cebr.156. URL <https://ideas.repec.org/a/prg/jnlcbr/v2016y2016i3id156p25-37.html>.
- Nils Bourgon, Farah Benamara, Alda Mari, Véronique Moriceau, Gaetan Chevalier, and Laurent Leygue. Are Sudden Crises Making me Collapse? Measuring Transfer Learning Performances on Urgency Detection. In *19th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2022)*, 2022a. URL <https://ut3-toulouseinp.hal.science/hal-03707241/document>.
- Nils Bourgon, Farah Benamara, Alda Mari, Véronique Moriceau, Gaetan Chevalier, Laurent Leygue, and Yasmine Djadda. Are sudden crises making me collapse? measuring transfer learning performances on urgency detection. In Rob Grace and Hossein Baharmand, editors, *19th International Conference on Information Systems for Crisis Response and Management, ISCRAM 2022, Tarbes, France, May 22-25, 2022*, pages 701–709. ISCRAM Digital Library, 2022b. URL <https://idl.iscram.org/show.php?record=2449>.
- David Bracewell, Marc Tomlinson, and Hui Wang. Identification of social acts in dialogue. In *Proceedings of COLING 2012*, pages 375–390, Mumbai, India, December 2012. The COLING 2012 Organizing Committee. URL <https://aclanthology.org/C12-1024>.
- Robert Brandom. *Making it explicit: Reasoning, representing, and discursive commitment*. Harvard university press, Cambridge, MA, USA, 1994.
- Anaïs Cadilhac, Nicholas Asher, Farah Benamara, and Alex Lascarides. Grounding strategic conversation: Using negotiation dialogues to predict trades in a win-lose game. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 357–368, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL <https://aclanthology.org/D13-1035>.
- Vitor R. Carvalho and William W. Cohen. On the collective classification of email ”speech acts”. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’05*, page 345–352, New York, NY, USA, 2005. Association for Computing Machinery. ISBN 1595930345. doi: 10.1145/1076034.1076094. URL <https://doi.org/10.1145/1076034.1076094>.
- Carlos Castillo. *Big crisis data: Social media in disasters and time-critical situations*. Cambridge University Press, Cambridge, England, 2016. ISBN 9781316476840. doi: 10.1017/9781316476840.

- Christophe Cerisara, Somayeh Jafaritazehjani, Adedayo Oluokun, and Hoa T. Le. Multi-task dialog act and sentiment recognition on mastodon. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 745–754, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL <https://aclanthology.org/C18-1063>.
- Alfredo Cobo, Denis Parra, and Jaime Navón. Identifying Relevant Messages in a Twitter-based Citizen Channel for Natural Disaster Situations. In *Proceedings of the 24th International Conference on World Wide Web, WWW’15*, pages 1189–1194, 2015.
- Dario Compagno, Elena V. Epure, Rebecca Deneckere, and Camille Salinesi. Exploring digital conversation corpora with process mining. *Corpus Pragmatics*, 2:193–215, 2018. doi: 10.1007/s41701-018-0030-6. URL <https://hal.univ-lorraine.fr/hal-01722928>.
- Cleo Condoravdi and Sven Lauer. Imperatives: Meaning and illocutionary force. *Empirical issues in syntax and semantics*, 9:37–58, 2012.
- W. T. Coombs. Ongoing crisis communication: Planning, managing, and responding. *Thousand Oaks, CA: Sage.*, 2014.
- Mark Core, Masato Ishizaki, Johanna Moore, Christine Nakatani, Norbert Reithinger, David Traum, and Syun Tutiya. The report of the third workshop of the discourse resource initiative, chiba university and kazusa academia hall. Technical report, 1998.
- Stefano Cresci, Maurizio Tesconi, Andrea Cimino, and Felice Dell’Orletta. A Linguistically-driven Approach to Cross-Event Damage Assessment of Natural Disasters from Social Media Messages. In *Proceedings of the 24th International Conference on World Wide Web, WWW’15*, pages 1195–1200, 2015.
- Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- Mian Dai and Chao-Lin Liu. Multi-label classification of chinese judicial documents based on bert. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 1866–1867. IEEE, 2020.
- AbdelRahim Elmadany, Hamdy Mubarak, and Walid Magdy. Arsas: An arabic speech-act and sentiment corpus of tweets. In Hend Al-Khalifa, King Saud University, KSA Walid Magdy, University of Edinburgh, UK Kareem Darwish, Qatar Computing Research Institute, Qatar Tamer Elsayed, Qatar University, and Qatar, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, may 2018a. European Language Resources Association (ELRA). ISBN 979-10-95546-25-2.
- AbdelRahim Elmadany, Hamdy Mubarak, and Walid Magdy. Arsas: An arabic speech-act and sentiment corpus of tweets. *OSACT*, 3:20, 2018b.
- Tin Franovic and Jan Šnajder. Speech act based classification of email messages in croatian language. 2012.
- Anastasia Giannakidou and Alda Mari. Mixed (non) veridicality and mood choice with emotive verbs. In *CLS 51*, 2015.

- Anastasia Giannakidou and Alda Mari. Epistemic future and epistemic must: nonveridicality, evidence, and partial knowledge. In *Mood, Aspect, Modality Revisited*, pages 75–118. University of Chicago Press, Chicago, IL, USA, 2017.
- Anastasia Giannakidou and Alda Mari. A unified analysis of the future as epistemic modality. *Natural Language & Linguistic Theory*, 36(1):85–129, 2018.
- Anastasia Giannakidou and Alda Mari. A linguistic framework for knowledge, belief, and veridicality judgment. *KNOW: A Journal on the Formation of Knowledge*, 5(2):255–293, 2021a.
- Anastasia Giannakidou and Alda Mari. Modalization and bias in questions. *University of Chicago and Institut Jean Nicod*, 2021b.
- Anastasia Giannakidou and Alda Mari. *Truth and veridicality in grammar and thought: Mood, modality, and propositional attitudes*. University of Chicago Press, Chicago, IL, USA, 2021c.
- Jonathan Ginzburg. *The interactive stance*. Oxford University Press, Kettering, Northamptonshire, UK, 2012.
- Oliveira Hugo Gonalo, Patr cia Ferreira, Daniel Martins, Catarina Silva, and Ana Alves. A brief survey of textual dialogue corpora. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1264–1274, Marseille, France, June 2022. European Language Resources Association. URL <https://aclanthology.org/2022.lrec-1.135>.
- Chih-Wen Goo and Yun-Nung Chen. Abstractive dialogue summarization with sentence-gated modeling optimized by dialogue acts. In *Proceedings of 7th IEEE Workshop on Spoken Language Technology*, 2018.
- Christine Gunlogson. A question of commitment. *Belgian Journal of Linguistics*, 22(1):101–136, 2008.
- Charles L Hamblin. Fallacies. *Tijdschrift Voor Filosofie*, 33(1), 1970.
- Hanna’t Hart. Predicting diagnoses of patients in the emergency room: a multi-label text classification approach. Master’s thesis, 2022.
- Ryuichiro Higashinaka, Kenji Imamura, Toyomi Meguro, Chiaki Miyazaki, Nozomi Kobayashi, Hiroaki Sugiyama, Toru Hirano, Toshiro Makino, and Yoshihiro Matsuo. Towards an open-domain conversational system fully based on natural language processing. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 928–939, Dublin, Ireland, August 2014. Dublin City University and Association for Computational Linguistics. URL <https://aclanthology.org/C14-1088>.
- Muhammad Imran, Shady Mamoon Elbassuoni, Carlos Castillo, Fernando Diaz, and Patrick Meier. Extracting information nuggets from disaster-related messages in social media. *Proc. of ISCRAM, Baden-Baden, Germany*, 2013.
- Muhammad Imran, Prasenjit Mitra, and Carlos Castillo. Twitter as a Lifeline: Human-annotated Twitter Corpora for NLP of Crisis-related Messages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, 2016. European Language Resources Association (ELRA). ISBN 978-2-9517408-9-1.

- Erika James and Lynn Wooten. Leadership as (un)usual: How to display competence in times of crisis. *Organizational Dynamics*, 34(2):141–152, 2005.
- Shafiq Joty and Tasnim Mohiuddin. Modeling speech acts in asynchronous conversations: A neural-CRF approach. *Computational Linguistics*, 44(4):859–894, December 2018. doi: 10.1162/colia.00339. URL <https://aclanthology.org/J18-4012>.
- Marc-André Kauffhold, Markus Bayer, and Christian Reuter. Rapid relevance classification of social media posts in disasters and emergencies: A system and evaluation featuring active, incremental and online learning. *Information Processing & Management*, 57(1):102–132, 2020.
- Simon Keizer, Rieks op den Akker, and Anton Nijholt. Dialogue act recognition with Bayesian networks for Dutch dialogues. In *Proceedings of the Third SIGdial Workshop on Discourse and Dialogue*, pages 88–94, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1118121.1118134. URL <https://aclanthology.org/W02-0213>.
- Mayank Kejriwal and Peilin Zhou. On detecting urgency in short crisis messages using minimal supervision and transfer learning. *Social Network Analysis and Mining*, 10(1):58, 2020.
- Jens Kersten, Anna Kruspe, Matti Wiegmann, and Friederike Klan. Robust Filtering of Crisis-related Tweets. In *ISCRAM 2019 conference proceedings-16th international conference on information systems for crisis response and management*, 2019. URL http://idl.iscram.org/files/jenskersten/2019/1909_JensKersten_etal2019.pdf.
- Diego Kozłowski, Elisa Lannelongue, Frédéric Saudemont, Farah Benamara, Alda Mari, Véronique Moriceau, and Abdelmoumene Boumadane. A three-level classification of french tweets in ecological crises. *Inf. Process. Manag.*, 57(5):102284, 2020. doi: 10.1016/j.ipm.2020.102284. URL <https://doi.org/10.1016/j.ipm.2020.102284>.
- Manfred Krifka. Commitments and beyond. *Theoretical Linguistics*, 45(1-2):73–91, 2019.
- D Robert Ladd. A first look at the semantics and pragmatics of negative questions and tag questions. In *Papers from the... Regional Meeting. Chicago Ling. Soc. Chicago, Ill*, number 17, pages 164–171, 1981.
- Pierre Larrivé and Alda Mari. Interpreting high negation in negative interrogatives: the role of the other. *Linguistics Vanguard*, 8(2):219–226, 2022.
- Peter Lasersohn. Context dependence, disagreement, and predicates of personal taste. *Linguistics and philosophy*, 28(6):643–686, 2005.
- Enzo Laurenti, Nils Bourgon, Farah Benamara, Alda Mari, Véronique Moriceau, and Camille Courgeon. Give me your intentions, i’ll predict our actions: A two-level classification of speech acts for crisis management in social media. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, pages 4333–4343. European Language Resources Association, 2022a. URL <https://aclanthology.org/2022.lrec-1.462>.

- Enzo Laurenti, Nils Bourgon, Benamara Farah, Alda Mari, Moriceau Véronique, and Camille Courgeon. Speech acts and communicative intentions for urgency detection. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 289–298, Seattle, Washington, July 2022b. Association for Computational Linguistics. doi: 10.18653/v1/2022.starsem-1.25. URL <https://aclanthology.org/2022.starsem-1.25>.
- Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. FlauBERT: Unsupervised Language Model Pre-training for French. *arXiv preprint arXiv:1912.05372*, 2019.
- Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.
- Weiwei Liu, Haobo Wang, Xiaobo Shen, and Ivor W Tsang. The emerging trends of multi-label learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(11):7955–7974, 2021.
- Alda Mari. Assertability conditions of epistemic (and fictional) attitudes and mood variation. In *Semantics and Linguistic Theory*, volume 26, pages 61–81, 2016.
- Alda Mari and Paul Portner. Mood variation with belief predicates: Modal comparison and the raisability of questions. *Glossa: a journal of general linguistics*, 40(1), 2021.
- Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. CamemBERT: a Tasty French Language Model. *arXiv e-prints*, art. arXiv:1911.03894, Nov 2019.
- Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de La Clergerie, Djamé Seddah, and Benoît Sagot. CamemBERT: a Tasty French Language Model. In *ACL 2020 - 58th Annual Meeting of the Association for Computational Linguistics*, Seattle / Virtual, United States, July 2020. doi: 10.18653/v1/2020.acl-main.645. URL <https://hal.inria.fr/hal-02889805>.
- Richard McCreadie, Cody Buntain, and Ian Soboroff. TREC incident streams: Finding actionable information on social media. In Zeno Franco, José J. González, and José H. Canós, editors, *Proceedings of the 16th International Conference on Information Systems for Crisis Response and Management, València, Spain, May 19-22, 2019*. ISCRAM Association, 2019a. URL http://idl.iscrum.org/files/richardmccreadie/2019/1867_RichardMcCreadie_et al2019.pdf.
- Richard McCreadie, Cody Buntain, and Ian Soboroff. TREC incident streams: Finding actionable information on social media. In *Proceedings of the 16th ISCRAM Conference*, 2019b.
- Richard McCreadie, Cody Buntain, and Ian Soboroff. Incident streams 2019: Actionable insights and how to find them. In *Proceedings of the 17th ISCRAM Conference*, 2020.
- Keval Morabia, Neti Lalita Bhanu Murthy, Aruna Malapati, and Surender Samant. SEDTWik: Segmentation-based event detection from tweets using Wikipedia. In *Proceedings of the 2019*

- Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 77–85, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-3011. URL <https://aclanthology.org/N19-3011>.
- Alexandra Olteanu, Sarah Vieweg, and Carlos Castillo. What to Expect When the Unexpected Happens: Social Media Communications Across Crises. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, CSCW '15*, pages 994–1009, 2015.
- Melina Plakidis and Georg Rehm. A dataset of offensive german language tweets annotated for speech acts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4799–4807, 2022.
- Paul Portner. *Mood*. Oxford University Press, Kettering, Northamptonshire, UK, 2018.
- EL Quarantelli, A Boin, and P Lagadec. Studying Future Disasters and Crises: A Heuristic Approach. *Handbook of Disaster Research*, pages 61–83, 2017.
- Christian Reuter and Marc-André Kaufhold. Fifteen years of social media in emergencies: A retrospective review and future directions for crisis informatics. *Journal of Contingencies and Crisis Management (JCCM)*, 26(1):41–57, 2018. ISSN 0966-0879. doi: <https://doi.org/10.1111/1468-5973.12196>. URL <http://tubiblio.ulb.tu-darmstadt.de/108144/>. Special Issue: Human-Computer-Interaction and Social Media in Safety-Critical Systems.
- Christian Reuter, Amanda Lee Hughes, and Marc-André Kaufhold. Social media in crisis management: An evaluation and analysis of crisis informatics research. *International Journal of Human-Computer Interaction*, 34(4):280–294, 2018.
- Lina M. Rojas-Barahona, Alejandra Lorenzo, and Claire Gardent. Building and exploiting a corpus of dialog interactions between French speaking virtual and human agents. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1428–1435, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/505_Paper.pdf.
- Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. *Advances in neural information processing systems*, 30, 2017.
- Jerrold Sadock. 3 speech acts. *The handbook of pragmatics*, page 53, 2004.
- Tulika Saha, Srivatsa Ramesh Jayashree, Sriparna Saha, and Pushpak Bhattacharyya. Bert-caps: A transformer-based capsule network for tweet act classification. *IEEE Transactions on Computational Social Systems*, 7(5):1168–1179, 2020a.
- Tulika Saha, Aditya Prakash Patra, Sriparna Saha, and Pushpak Bhattacharyya. A transformer based approach for identification of tweet acts. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020b.

- Tulika Saha, Apoorva Upadhyaya, Sriparna Saha, and Pushpak Bhattacharyya. Towards sentiment and emotion aided multi-modal speech act classification in Twitter. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5727–5737, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.456. URL <https://aclanthology.org/2021.naacl-main.456>.
- Efsun Sarioglu Kayi, Linyong Nan, Bohan Qu, Mona Diab, and Kathleen McKeown. Detecting urgency status of crisis tweets: A transfer learning approach for low resource languages. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4693–4703, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.414. URL <https://aclanthology.org/2020.coling-main.414>.
- Roser Saurí and James Pustejovsky. Factbank: a corpus annotated with event factuality. *Language resources and evaluation*, 43(3):227, 2009.
- John R Searle. Indirect speech acts. In *Speech acts*, pages 59–82. Brill, Leiden, Netherlands, 1975.
- John Rogers Searle. *Speech acts: An essay in the philosophy of language*. Cambridge university press, Cambridge, England, 1969.
- I. V. Serban, R. Lowe, P. Henderson, L. Charlin, and J. Pineau. A survey of available corpora for building data-driven dialogue systems: The journal version. *Dialogue & Discourse*, 9(1):1–49, 2018.
- Lina Sherkawi, Nada Ghneim, and Oumayma Al Dakkak. Arabic speech act recognition techniques. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 17(3), feb 2018. ISSN 2375-4699. doi: 10.1145/3170576. URL <https://doi.org/10.1145/3170576>.
- Mandy Simons. Observations on embedding verbs, evidentiality, and presupposition. *Lingua*, 117(6):1034–1056, 2007.
- Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun’ichi Tsujii. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations Session at EACL 2012*, Avignon, France, April 2012. Association for Computational Linguistics.
- Tamina Stephenson. Judge dependence, epistemic modals, and predicates of personal taste. *Linguistics and philosophy*, 30(4):487–525, 2007.
- Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–374, 2000. URL <https://aclanthology.org/J00-3003>.
- Shivashankar Subramanian, Trevor Cohn, and Timothy Baldwin. Target based speech act classification in political campaign text. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 273–282, Minneapolis, Minnesota,

- June 2019. Association for Computational Linguistics. doi: 10.18653/v1/S19-1030. URL <https://aclanthology.org/S19-1030>.
- Tiancheng Tang, Xinhui Tang, and Tianyi Yuan. Fine-tuning bert for multi-label sentiment analysis in unbalanced code-switching text. *IEEE Access*, 8:193248–193256, 2020.
- Cagri Toraman, Izzet Emre Kucukkaya, Oguzhan Ozcelik, and Umitcan Sahin. Tweets under the rubble: Detection of messages calling for help in earthquake disaster. *arXiv preprint arXiv:2302.13403*, 2023.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- Sarah Vieweg, Carlos Castillo, and Muhammad Imran. Integrating Social Media Communications into the Rapid Assessment of Sudden Onset Disasters. In *Proceedings of the 6th International Conference of Social Informatics*, SocInfo’14, pages 444–461, 2014.
- Soroush Vosoughi. *Automatic detection and verification of rumors on Twitter*. Thesis, Massachusetts Institute of Technology, 2015. URL <https://dspace.mit.edu/handle/1721.1/98553>.
- Soroush Vosoughi and Deb Roy. Tweet acts: A speech act classifier for twitter. *Proceedings of the Tenth International AAAI Conference on Web and Social Media (ICWSM 2016)*, 2016.
- M. Weisser. *How to Do Corpus Pragmatics on Pragmatically Annotated Data: Speech Acts and Beyond*. John Benjamins Publishing Company, 2018.
- Guanghao Xu, Hyunjung Lee, Myoung-Wan Koo, and Jungyun Seo. Convolutional neural network using a threshold predictor for multi-label speech act classification. In *2017 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 126–130, 2017. doi: 10.1109/BIGCOMP.2017.7881727.
- Kiran Zahra, Muhammad Imran, and Frank O Ostermann. Automatic identification of eyewitness messages on twitter during disasters. *Information Processing & Management*, 57(1):102–107, 2020. ISSN 0306-4573.
- Raffaella Zanuttini, Miok Pak, and Paul Portner. A syntactic analysis of interpretive restrictions on imperative, promissive, and exhortative subjects. *Natural Language & Linguistic Theory*, 30(4): 1231–1274, 2012.
- Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8):1819–1837, 2013.
- Renxian Zhang and Naishi Liu. Recognizing humor on twitter. In *Proceedings of the 23rd ACM international conference on conference on information and knowledge management*, pages 889–898, 2014.
- Renxian Zhang, Dehong Gao, and Wenjie Li. What are tweeters doing: Recognizing speech acts in twitter. In *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*. Citeseer, 2011.