Speech acts and Communicative Intentions for Urgency Detection
Enzo Laurenti, Nils Bourgon, Farah Benamara, Alda Mari, Véronique Moriceau, Camille Courgeon

To cite this version:
Enzo Laurenti, Nils Bourgon, Farah Benamara, Alda Mari, Véronique Moriceau, et al.. Speech acts and Communicative Intentions for Urgency Detection. 11th Joint Conference on Lexical and Computational Semantics (*SEM 2022), Jul 2022, Seatle, United States. 10.18653/v1/2022.starsem-1.25. hal-03730461

HAL Id: hal-03730461
https://hal.science/hal-03730461
Submitted on 11 Dec 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License
Speech acts and Communicative Intentions for Urgency Detection

Enzo Laurenti1 Nils Bourgon2 Farah Benamara2
1IJN, CNRS/EHESS, PSL University
firstname.lastname@ens.fr
Alda Mari1 Véronique Moriceau2 Camille Courgeon1
2 IRIT, Université de Toulouse, CNRS, Toulouse INP, UT3
firstname.lastname@irit.fr

Abstract
Recognizing speech acts (SA) is crucial for capturing meaning beyond what is said, making communicative intentions particularly relevant to identify urgent messages. This paper attempts to measure for the first time the impact of SA on urgency detection during crises, in tweets. We propose a new dataset annotated for both urgency and SA, and develop several deep learning architectures to inject SA into urgency detection while ensuring models generalisability. Our results show that taking speech acts into account in tweet analysis improves information type detection in an out-of-type configuration where models are evaluated in unseen event types during training. These results are encouraging and constitute a first step towards SA-aware disaster management in social media.

1 Introduction
Discovered by (Austin, 1962) and extensively promoted by (Searle, 1975), speech acts (henceforth SA) have been the object of extensive discussion in the philosophical and the linguistic literature (Sadock, 2004; Portner, 2018). According to the Austinian initial view, SA are to achieve action rather than conveying information. When uttering I now pronounce you man and wife, the priest accomplishes the action of marrying rather than just stating a proposition. Beyond these prototypical cases, the literature has quickly broaden 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; Portner, 2018; Giannakidou and Mari, 2021).

Because recognizing speakers’ intentions is crucial for capturing meaning beyond what is said (Noveck, 2018), SA have given rise to 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) as well as asynchronous dialogues (e.g., emails, tweet threads) (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 action prediction (Cadilhac et al., 2013), dialogue summarization (Goo and Chen, 2018) and conversational systems (Higashinaka et al., 2014). In this paper, we attempt to measure for the first time the role of SA on urgency detection in tweets, focusing on natural disasters (hurricanes, storms, floods, etc.).

SA are particularly relevant to identify urgent messages, i.e. those that raise situational awareness over a crisis (including human/material damages, security instructions, etc.), providing therefore actionable information that will help to set priorities for the human teams and decide appropriate rescue actions. By tweeting, speakers seek to achieve impact via enhancing a chain of reactions. They do not necessarily seek to merely express themselves. The greater the number of re-tweets and replies the greater the impact. Therefore, tweets are not only public, but they are also interactive. They mostly aim to make interlocutors react (perlocutionary level) by different linguistic means (illocutionary level), in view of achieving a purpose (on perlocutionary / illocutionary, see (Austin, 1962; Searle, 1975)). We illustrate this in the following examples1 where speaking subjects perform qualitatively very different language acts depending on the situation they find themselves in. In the tweet (1a), the writer publicly expresses an explicit commitment to provide help after the Irma hurricane tragedy, using an explicit action verb (“to help”) which is under the scope of an explicit attitude verb (“want”), thus aiming to obtain a reply on what to do to provide help. (1b) on the other hand ex-

---

1These are examples taken from our French corpus translated into English.
presses an intention to complain about the absence of assistance without using any explicit intent keywords and thus raise awareness and attention on the part of the people in charge of assistance.

(1) #Irma Hurricane: “I want to go there to help.”

(2) Irma hurricane: where is disaster assistance one month later?

When annotating tweets posted during a crisis (like earthquakes, bombing, attacks) according to different taxonomies of SA, state of the art corpus-based studies observe a majority of statements, essentially supplemented by suggestions and comments – in contrast, the topics dealing with e.g. celebrities are essentially made up of comments (Zhang et al., 2011; Vosoughi, 2015; Elmadany et al., 2018a; Saha et al., 2020). These results have however been obtained after manual annotations, the focus being rather on SA classification of topic oriented tweets. The next step now is to show to what extent these observations are still valid from a computational point of view. Our contribution is threefold and consists in:

1. A new dataset of 6,669 tweets in French annotated for both urgency and SA for disaster events of various types that occurred in France;

2. A set of deep learning experiments to inject SA information into urgency detection using monotask and multitask architectures. We investigate the role of communicative intentions in three classification settings: relatedness (i.e., useful vs. non useful for emergency responders), urgency detection (i.e., non useful vs. urgent vs. non urgent), and information type following a predefined taxonomy of six actionable categories;

3. An evaluation of the proposed classifiers while measuring their ability to generalize over new events. Our results show that SA are helpful for filtering out urgent from non urgent messages. This is particularly salient for information type detection in an out-of-type configuration where models are evaluated in unseen event types during training. These results are encouraging and constitute a first step towards SA-aware disaster management in social media.

beating several SA agnostic state of the art baselines.

This paper is organized as follows. We first provide related work on NLP-based approaches to crisis management as well as SA in social media. We then describe our data, the annotation procedure and the results of the annotation campaign. We detail the experiments we carried out on injecting SA in urgency detection in Section 4 and discuss our results in Section 5. We end the paper by some perspectives for future work.

2 Related Work

2.1 Crisis Datasets

The literature on emergencies detection has been growing fast in the recent years and several datasets (mainly tweets) have been proposed to account for crisis related phenomena. 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. Well-known dimensions include relatedness (also known as usefulness or informativeness) to identify whether the message content is useful (Jensen, 2012), situation awareness (also known as urgency, criticality or priority) to filter out on-topic relevant (e.g., immediate post-impact help) vs. on-topic irrelevant information (e.g. supports and solicitations for donations) (Imran et al., 2013; McCreadie et al., 2019; Sarioglu Kayi et al., 2020; Kozlowski et al., 2020), and eyewitness types to identify direct and indirect eyewitnesses (Zahra et al., 2020). For most of the existing datasets, annotations usually apply at the text level. Some studies propose to additionally annotate images within the tweets (see for example (Alam et al., 2018)).

The question of how speakers convey emergency at the sentence level has nonetheless been only tangentially addressed in a literature that has considered the correlation between specific speech acts and specific topics, without overtly addressing what the speech act shape of urgent messages is (see below).

--

2The dataset will be made available to the research community.

3See https://crisisnlp.qcri.org/ for an overview.
2.2 Speech Acts in Social Media

Some amount of attention has been indeed devoted to understanding 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, resort to a Searlian typology of SA that distinguishes between assertive statements (description of the world), expressive comments (expression of a mental state of the speaker), interrogative questions and imperative suggestions. 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. The same conclusions have been drawn by Vosoughi (2015); Vosoughi and Roy (2016) when distinguishing the topic discussed in the tweets, from the type of topic (Entity-oriented—celebrities, Event-oriented topics—bombing events, or Long-standing topics—cooking).

Their corpus study shows that there is a greater similarity of distribution of SA between entity-oriented and event-oriented, with a majority of assertions and expressions.

In this same perspective of topic identification, Elmadany et al. (2018b) classify 21,000 tweets in Arabic according to their topic type and distinguish events (for example, in our case, natural disasters), 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 makes emerge 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.

Our classification of speech acts relies on the fourfold distinction between asserting, ordering, asking and expressing a subjective view (cf. infra, section 3.2 for the definitions and specifications of these categories). The novelty of our work lies in exploring communicative intentions in the context of urgency detection, an enterprise which, to our best knowledge, has never been undertaken. This paper fills this gap by crossing the urgency classification and the SA classification in order to elucidate the interactions between speaker’s attitudes and urgency categories (and their associated actions).

3 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 (Kozlowski et al., 2020) composed of about 12k tweets collected using dedicated keywords about ecological crises that occurred in France from 2016 to 2019 and posted 24h before, during (48h) and 72h after the crisis: 2 floods that occurred in Aude and Corsica regions, 10 storms—Béryl, Berguitta, Fionn, Eleanor, Bruno, Egon, Ulrika, Susanna, Fakir and Ana, and 2 hurricanes—Irma and Harvey, and 1 sudden crisis (Marseille building collapse). It is important to note that in this dataset, some crises occurred in the same time period which implies that some messages that were scraped for some crises actually belonged to other (they were annotated as NOT USEFUL in this case, as they are not related to the targeted crisis, see below).

3.1 Urgency Annotation Layer

In this dataset, each tweet is annotated according to its relatedness, urgency and six information type categories, namely HUMAN DAMAGES and MATERIAL DAMAGES which concern missing, injured, displaced and dead people or any damaged infrastructure that was caused by a crisis, WARNING-ADVICE that gives security instructions, tips to limit the damage or weather reports, SUPPORT messages to the victims, CRITICS messages that denounce the lack of effectiveness of rescue services, and OTHER messages that do not have an immediate impact on actionability but contribute to raising situational awareness. The first three types are subcategories of urgent messages while the last three are subcategories of non urgent messages. The dataset comes with additional metadata including: number of likes and retweets of the tweet, and number of likes, followers, following of the user.

The collection is extremely imbalanced with 11.24% useful but NOT URGENT, 16.74% URGENT and 72.02% NOT USEFUL messages, which is in line with the proportions reported in other crisis corpora. A subset of this dataset composed of 6,669 tweets have been selected for SA annotations, so that almost all URGENT (2,080) and NON URGENT (1,401) messages have been annotated. Only 3,188 NOT USEFUL tweets have been selected in order to reduce the size of this class but keep it majoritary. Note that pre-existing urgency tags and metadata information have been removed to prevent annotators from getting biased by specific urgency-SA pairs.

3.2 Speech Act Annotation Layer

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 ‘subjectives’, which reshuffles some of the earlier classifications (‘wishes’, for instance are ‘subjectives’ rather than ‘jussives’ in our classification (e.g. (Condoravdi and Lauer, 2012)); (iii) considering presuppositional content as well (see (Mari, 2016) on French).

We distinguish four categories which are mutually exclusive and define tweets as wholes, at a holistic level, as follows:

(1) **JUSSIVES**, as defined by (Zanuttini et al., 2012), enhance commitment to take action, as in (3). In our classification we distinguish: commissives (i.e. the speaker commits himself or herself), exhortatives (i.e. the speaker commits some relevant individuals), orders (i.e. the speaker commits the addressee, in the case of authority relations), and open-options (i.e. the speaker describes the existence of a possibility).

(3) #Inondation Si vous êtes en zone inondable, découvrez comment préparer un kit de survie (#Flooding If you are in an area at risk of flooding, discover how to prepare a survival kit).

(2) **ASSERTIVES**. Assertions are considered to convey objective truth (as opposed to subjective truth (Giannakidou and Mari, 2021)). With assertives, 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).

(4) Inondations dans l’Aude : la région débloque 25M€, le président Macron sur place lundi (Flooding in Aude: the region unlocks 25M€, the president Macron on the spot on Monday).

(3) **INTERROGATIVES**. This category is dedicated to a variety of questions including both those that require an informative answer and those that, besides triggering an answer, reveal bias and expectations on the part of the speaker (see (Ladd, 1981)).

(5) Salut Chelsea, comment ça va, la tempête, par chez vous? (Hi Chelsea, how is the storm at your place?).

(4) **SUBJECTIVES**. Finally, with subjectives, 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). 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 ‘subjectives’ rather than ‘jussives’ as they do not trigger any commitment to act so to make the content of the wish true.

(6) Grosse pensée à ma Laure qui est en Martinique avec l’ouragan (My thoughts are with my Laure, who is in Martinique with the hurricane.)

Finally, **OTHERS** is added to the classification, for uncertain or unclassifiable cases, as in (7).

(7) Simulation #3D d’une #inondation à Issy-les-Moulineaux merci à @Ubick3D pour le prêt #ortho3D #InterAtlas (3D simulation of a flood in Issy-les-Moulineaux thanks to @Ubick3D for the loan #ortho3D #InterAtlas).

The final dataset is therefore composed of 6,669 tweets. Here is a representative example
of a tweet in our dataset, along with its corresponding annotation: Relatedness=USEFUL, Urgency=URGENT, Information type=HUMAN DAMAGE, SA=ASSERTIVE:

(8) #irma st martin: nouveau bilan provisoire avec 8 morts et 21 blessés à St. Martin (#irma st martin: new provisional death toll of 8 dead and 21 injured in St. Martin)

3.3 Results of the Annotation Campaign

We hired two native French speaking annotators, both master’s degree students in Linguistics in order to annotate tweets. We performed a two-step annotation where an intermediate analysis of agreement and disagreement between the annotators was carried out. 448 tweets have been annotated in the first step by both annotators so that the inter-annotator agreement could be computed (Cohen’s Kappa=0.62). Most cases of disagreement come from the difficulty of disentangling SUBJECTIVES from ASSERTIVES, 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 such as fallait, which translates as ‘should have been’ with an additional implication of preference in (9)) or their content can be targeted, according to their contextual relevance. This delicate distinction is often resolved in different manners by annotators.

(9) Et maintenant il n’y a presque plus de fumée... Il fallait arrêter le trafic ce matin et pas au milieu de la journée.
(And now there is almost no more smoke... Traffic should have been stopped this morning and not in the middle of the day).

Table 1 details the frequency of SA tags when paired with the original urgency annotations. The final distribution of annotated tweets is 59.8%, 22.3%, 10%, 4.5% and 3.3% for ASSERTIVE, SUBJECTIVE, Jussive, OTHER and INTERROGATIVE respectively. Concerning the two most frequent SA (ASSERTIVE and SUBJECTIVE), two observations emerge: (1) Among URGENCY messages (resp. NON URGENCY), 86.6% (resp. 48.7%) are ASSERTIVE; and (2) Only 5% of URGENCY messages are SUBJECTIVE while 29% of NON URGENCY messages are. Similarly, we observe that 7% of JUS-SIVE are URGENCY vs. 14% NON URGENCY. All these frequencies are statistically significant using the $\chi^2$ test ($\chi^2 = 1, 1011.62$, $df = 8$, $p < 0.01$). When measuring the dependency strength between urgency and SA categories using the Cramer’s $V$, we get ($V = .28$, $df = 2$) which confirms the statistical correlation between these two classifications.

Table 2 further details the SA distribution for each crisis. We can see that ASSERTIVE messages are the most frequent ones regardless of the crisis. Another interesting finding concerns the distribution of SA in sudden crisis. Indeed, SA frequencies are relatively similar in natural disaster crisis (flood, storms and hurricane) with about 60% of ASSERTIVE and 20% of SUBJECTIVE. However in the Marseille building collapse, we observe a higher proportion of SUBJECTIVE (35% vs. 49% for ASSERTIVE) showing that people tend to express fewer messages of warning-advice but many critics denouncing the lack of effectiveness of government social action.

4 Speech Acts for Urgency Detection

We propose several models to automatically classify a tweet according to its relatedness (binary classification–REL), urgency (three classes–URG) and information type categories (multiclass–INF) while injecting SA information into the learning process. Our models have been compared to SA-agnostic baselines while analyzing the impact of SA on generalization to new disaster events which is important for this application, since disasters can vary widely with respect to both their specific properties as well as their types. Although SA detection is an important preliminary step, this is however out of the scope of this paper. Note that a baseline CamemBERT model (Martin et al., 2019) fine-tuned to predict the five SA tags achieves a macro F-score of 0.686 with a precision of 0.690 and recall of 0.701. Improvement of these results is left for future work.
4.1 SA-agnostic Models

SA-agnostic models have been compared to Kozlowski et al. (2020), the only existing work in French that has shown to outperform state of the art on urgency detection. Kozlowski et al. (2020) models rely on a language adaptation version of FlauBERT base cased model (Le et al., 2020), initially trained on a general domain, and fine-tuned for the crisis domain using a set of French unlabeled dataset of 358,834 tweets. Our baselines are:

– FlauBERT\textsuperscript{tuned}. This is the original tuned version of FlauBERT trained on our dataset with a cross-entropy loss. We newly add FlauBERT\textsuperscript{tuned} \textsuperscript{wl}, a variant that uses the weighted loss instead to handle class imbalance.\textsuperscript{5} The results obtained with this variant model being more productive, the weighted variant that uses the weighted loss instead to handle class imbalance.

\textsuperscript{5}We also experimented with focal loss (Lin et al., 2017) but the results were lower.

4.2 SA-aware Models

SA are incorporated into FlauBERT models in two ways. First, rely on SA gold annotations as additional extra-features. We experimented with several ways to inject SA among which representing SA as numerical values (0 for ASSERTIVE, 1 for SUBJECTIVE, etc.), inserting SA tags at the end of the tweet using a specific marker (e.g., <Assertif > for ASSERTIVE tweets), representing SA as one hot vector, and finally consider each SA tag as a unique binary feature to model its presence or absence. The last option was the most productive and is used in four models: FlauBERT\textsuperscript{tuned} \textsuperscript{wl} +SA, FlauBERT\textsuperscript{tuned} \textsuperscript{wl} +SA+Meta, ML\textsuperscript{3}+SA, and ML\textsuperscript{3}+SA+Meta.

The previous configuration is an ideal case where urgency detection benefits from gold SA which may not be available for unseen/new disaster events. We therefore designed a more realistic scenario where SA detection is considered as an auxiliary task. This is a multitask learning approach that jointly learns urgency detection with SA classification as a secondary task. Two models are newly proposed:

– ML\textsuperscript{2}: It corresponds to FlauBERT\textsuperscript{tuned} \textsuperscript{wl} trained to perform SA together with one urgency task (i.e., two tasks among REL+SA, URG+SA or INF+SA). This configuration aims to investigate what are the tasks that may benefit the most from injecting SA information among relatedness, urgency and information type.

– ML\textsuperscript{4}: FlauBERT\textsuperscript{tuned} \textsuperscript{wl} learns SA together with the three urgency tasks. This is a four task configuration that corresponds to SA+REL+URG+INF.

\begin{table}[h]
\begin{tabular}{|m{3cm}|m{1cm}|m{1cm}|m{1cm}|m{1cm}|}
\hline
& ASSERTIVE & SUBJECTIVE & JUSSIVE & INTERROGATIVE \\
\hline
Flood & Aude & 718 (71.37%) & 184 (18.29%) & 84 (8.35%) & 20 (1.99%) \\
& Corse & 248 (63.75%) & 73 (18.77%) & 45 (11.57%) & 23 (5.91%) \\
& Other Flood & 631 (64.65%) & 180 (18.44%) & 137 (14.04%) & 28 (2.87%) \\
& Total & 1,397 (67.36%) & 437 (18.43%) & 266 (11.22%) & 71 (2.99%) \\
\hline
Storms & Beryl & 174 (59.18%) & 87 (29.59%) & 22 (7.48%) & 11 (3.74%) \\
& Bruno & 201 (61.47%) & 94 (28.75%) & 17 (5.20%) & 15 (4.95%) \\
& Susanna & 230 (61.66%) & 92 (24.66%) & 45 (12.06%) & 6 (1.61%) \\
& Ulrika & 170 (60.71%) & 60 (21.43%) & 43 (15.36%) & 7 (2.55%) \\
& Berguitta & 189 (60.77%) & 73 (23.47%) & 35 (11.25%) & 14 (4.54%) \\
& Fionn Corse & 238 (69.79%) & 69 (20.23%) & 28 (8.21%) & 6 (1.76%) \\
& Egon & 185 (58.92%) & 95 (30.25%) & 24 (7.64%) & 10 (3.18%) \\
& Eleanor & 208 (67.10%) & 69 (22.26%) & 26 (8.39%) & 7 (2.26%) \\
& Total & 1,595 (62.55%) & 639 (25.06%) & 240 (9.41%) & 76 (2.98%) \\
\hline
Hurricane & Harvey & 168 (58.74%) & 59 (20.63%) & 36 (12.59%) & 23 (8.04%) \\
& Irma & 487 (55.72%) & 251 (28.78%) & 100 (11.44%) & 36 (4.12%) \\
& Total & 655 (56.47%) & 310 (26.72%) & 136 (11.72%) & 39 (3.09%) \\
\hline
Collapse & Marseille & 143 (49.48%) & 102 (35.39%) & 27 (9.34%) & 17 (5.88%) \\
\hline
\end{tabular}
\caption{SA distribution for each crisis.}
\end{table}
These two models are further augmented with tweet meta features, resulting in two other models: ML$^2$+Meta and ML$^4$+Meta.

4.3 Experimental Settings

Following the general trends in evaluating urgency detection during disaster events, we designed two evaluation protocols:

- **[OE]** **out-of-event** by testing on unseen events for which no manually annotated data is available during training. To ensure a fair comparison with (Kozłowski et al., 2020), we used the same test sets composed of crises Eleanor and Bruno. This choice is also motivated by the fact that these two crises did not show the mentioned overlap with other crises and hence there was no information leak from one event to another (cf. Section 3);

- **[OT]** **out-of-type** by training on a pool of events related to different types of crises and testing on a particular different type. We used the building collapse as a test set. While the hurricanes and floods are known with anticipation, a building collapse is a sudden event with pretty different distributions in terms of urgency categories, making the [OT] configuration more challenging.

During the experiments, all the five SA tags have been taken into account for urgency detection. $^6$

5 Results

5.1 Out-of-event and Out-of-type Detection

The results of [OE] and [OT] configurations in terms of macro-F1 scores are given in Table 3. It shows that SA-enhanced models beat SA agnostic ones for urgency and information type detection in both the [OE] and [OT] evaluation settings. In [OE], ML$^5$+SA+Meta improves over the FlauBERT$_{tuned}^{vol}$ and FlauBERT$_{tuned}^{vol}$+Meta baselines and this is more salient for information type classification. The same observations hold for [OT] where SA boost the scores when injected both as extra-features and as an auxiliary task. Another interesting finding is that joint learning of SA and urgency tags (i.e., ML$^2$) achieves results comparable to those obtained in the ideal case, i.e. when incorporating gold SA annotations as extra-features. Also, when coupling SA with tweet meta features, the results improve in most experiments, confirming the importance of extra-linguistic information for urgency detection. On the other hand, when compared to the best baseline, SA injection into relatedness detection achieves similar scores in [OE] while they decrease in [OT]. This was however expected as the relatedness baseline classifiers perform relatively well (F-score=0.849 and F-score=0.856 for [OE] and [OT] respectively). This can be explained by the same proportions of SA we observed in each of the USEFUL and NOT USEFUL class where ASSERTIVE messages are a majority followed by SUBJECTIVE ones (see Table 2).

When looking into the scores per class for urgency detection in [OE] (see Table 4), we observe that SA are the most helpful for predicting URGENT messages with an important boost up to (+3%) for NON URGENT tweets. A boost is observed in [OT] where SA injection improves by +1.2% over the SA-agnostic best model. Regarding the ability of the models to filter-out irrelevant messages, we observe that the results with SA are stable in [OE] (with an F-score=0.887) while they increased in [OT]. It is interesting to note that the results obtained in real scenario via multitask learning models (i.e., $ML^2$ and $ML^4$) achieve good results compared to the models that rely on SA gold annotations. More importantly, multitask models outperform SA-agnostic baselines which show the importance of SA for fine-grained urgency detection in social media.

Concerning information type classification, Table 5$^7$ shows that the SA-aware model in the [OE] setting is able to predict MATERIAL DAMAGES, NOT USEFUL as well as OTHER non urgent messages (related to animals, messages that aim to provide additional information via external links via URLs, photos or videos, and prevention messages that provide general-purpose safety instructions upstream of crisis). When testing on a particular different event (i.e., a sudden event like the building collapse in Marseille), the [OT] configuration shows an improvement on MATERIAL DAMAGES and WARNING ADVICE. Finally, it is also interesting to note that major improvements concern the

---

$^6$We tried several groupings of SA tags among which ASSERTIVE vs. NOT ASSERTIVE, (ASSERTIVE+SUBJECTIVE) vs. (INTERROGATIVE+IMPERATIVE+OTHER) to measure what are the SA combinations that contribute the most to the task at hand. Our results show that all SA are relevant.

$^7$The two events used for testing do not have any CRITICS messages.
### Out-of-event

<table>
<thead>
<tr>
<th></th>
<th>OUT-OF-EVENT</th>
<th>OUT-OF-TYPE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>REL</td>
<td>URG</td>
</tr>
<tr>
<td><strong>SA-agnostic‡</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FlauBERT\textsuperscript{tuned}</td>
<td>0.846</td>
<td>0.681</td>
</tr>
<tr>
<td>FlauBERT\textsuperscript{tuned} \textsuperscript{w+k}</td>
<td>0.847</td>
<td>0.688</td>
</tr>
<tr>
<td>FlauBERT\textsuperscript{tuned} \textsuperscript{w}+\textsuperscript{Meta}</td>
<td>0.837</td>
<td>0.698</td>
</tr>
<tr>
<td>ML\textsuperscript{3}</td>
<td>0.842</td>
<td>0.654</td>
</tr>
<tr>
<td>ML\textsuperscript{3}+\textsuperscript{Meta}</td>
<td>0.849</td>
<td>0.679</td>
</tr>
<tr>
<td><strong>SA-aware as extra-features</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FlauBERT\textsuperscript{tuned} \textsuperscript{w}+\textsuperscript{SA}</td>
<td>0.849</td>
<td>0.680</td>
</tr>
<tr>
<td>ML\textsuperscript{3}+\textsuperscript{SA}</td>
<td>0.849</td>
<td>0.693</td>
</tr>
<tr>
<td>ML\textsuperscript{3}+\textsuperscript{SA}+\textsuperscript{Meta}</td>
<td>0.844</td>
<td>0.708</td>
</tr>
<tr>
<td><strong>SA-aware as an auxiliary task</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ML\textsuperscript{2}</td>
<td>0.841</td>
<td>0.703</td>
</tr>
<tr>
<td>ML\textsuperscript{2}+\textsuperscript{Meta}</td>
<td>0.841</td>
<td>0.693</td>
</tr>
<tr>
<td>ML\textsuperscript{4}</td>
<td>0.847</td>
<td>0.697</td>
</tr>
<tr>
<td>ML\textsuperscript{4}+\textsuperscript{Meta}</td>
<td>0.842</td>
<td>0.689</td>
</tr>
</tbody>
</table>

Table 3: Urgency detection results in terms of Macro F1-score. ‡: SA agnostic strong baselines. Bold font: Outperforming models over the baselines.

---

# 6 Conclusion

This paper newly addresses the role of speech acts in urgency detection in tweets. In particular, we propose a dataset of French tweets about urgent situations and create models that utilize speech acts to classify the tweets. We also analyze the generalization of the models over new urgent events. Our results are encouraging and demonstrate that SA improve urgency detection. This is more salient for out-of-type evaluation setting, where the SA-aware approach has shown to have a good generalisation power in fine-grained classification.

This work could be very useful to government workers who need to respond to natural disasters and to decide how to deploy possibly limited resources. As future work, we plan to explore a finer-grained SA taxonomy on urgency detection.

---

### Acknowledgment

This work has been supported by the INTACT project funded by the AAP CNRS - INHESJ 2022 and the FIESPI grant with the French Interior Min-
Table 5: Impact of SA injection for information type classification per class in terms of macro F1-scores.

<table>
<thead>
<tr>
<th>Best models</th>
<th>NOT USF</th>
<th>HUM. DAM.</th>
<th>MAT. DAM.</th>
<th>WAR. ADV.</th>
<th>SUP.</th>
<th>CRI.</th>
<th>OTH.</th>
</tr>
</thead>
<tbody>
<tr>
<td>FlauBERTwl</td>
<td>0.855</td>
<td>0.746</td>
<td>0.638</td>
<td>0.843</td>
<td>0.545</td>
<td>–</td>
<td>0.246</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ML¹</td>
<td>0.886</td>
<td>0.721</td>
<td>0.656</td>
<td>0.844</td>
<td>0.500</td>
<td>–</td>
<td>0.356</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FlauBERTwl+Meta</td>
<td>0.899</td>
<td>0.759</td>
<td>0.432</td>
<td>0.000</td>
<td>0.917</td>
<td>0.326</td>
<td>0.250</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ML²</td>
<td>0.895</td>
<td>0.737</td>
<td>0.500</td>
<td>0.308</td>
<td>0.783</td>
<td>0.323</td>
<td>0.188</td>
</tr>
</tbody>
</table>

References


Diego Kozlowski, Elisa Lannelongue, Frédéric Saudemont, Farah Benamara, Alda Mari, Véronique


