Applying Distilled BERT for Question Answering on ASRS Reports

Samuel Kierszbaum
Ecole Nationale de l'Aviation Civile
Toulouse, France
samuel.kierszbaum@enac.fr

Laurent Lapasset
Ecole Nationale de l'Aviation Civile
Toulouse, France
laurent.lapasset@enac.fr

***Abstract*—This paper employs the Bidirectional Encoder Representations from Transformers (BERT), a language model, fine-tuned on the question answering task, on the Aviation Safety Reporting System (ASRS) dataset's free text reports, that describe incident occurrences in an International aviation safety context. A four-step method is used to evaluate the produced results. This paper outlines what are the limitations of this approach, as well as its usefulness in trying to extract information from thirty randomly selected free text reports when asking the following question: "When did the incident happen?". We aim to try to integrate one of the algorithms resulting of the recent advances in Natural Language Processing (NLP) to leverage information in natural language narratives, as opposed to working directly with the structured part of the ASRS dataset. We find that our approach yields interesting results, with roughly seventy percent correct answers, including answers that have information that is not overlapping with the ASRS dataset's metadata.**

***Index Terms*—NLP, aviation, safety, BERT, ASRS**

I. Introduction

The Aviation Safety Reporting System (ASRS) contains incident reports published since 1976, which consists in a total report number of 1 625 738 (07/2019) [1]. It is used by safety experts to study incidents and trends in aviation safety. Many questions arise around this data.

One of the main directions of research in the field of NLP applied to safety in aviation aims at augmenting the data automatically to help experts query the database efficiently [2]. For instance, the ability to query the database is needed to find similar occurrences when investigating an incident, or when trying to find rare occurrences. Augmenting the data is often done by using text classification techniques [3]. In this article, we present an attempt to help safety experts in the study of the reported incidents where we use a different approach.

Recently, the field of Natural Language Processing (NLP) has developed dramatically, with the rise of models such as BERT [4] from Google or Open AI's GPT2 [5]. In the context of these advances, our intention is to explore one of the ways to incorporate these algorithms to try producing hindsight on the aviation-safety related NASA ASRS database [6].

We work with the information filled in by the reporters [7], which consists in reports written in natural language as well as metadata.

Information from metadata and natural language reports are both challenging to use, as shown by Tulechki Nikola [2]. Information in natural language can hardly be used as such to produce statistics or search occurrences. Typically, it is converted in metadata or taxonomies, following the coding scheme of the database [6], that can help querying the ever expanding incident database, via SQL queries. These SQL queries are the only means currently used by experts to extract data.

Producing this data augmentation is highly time consuming and requires expert knowledge. Additionally, metadata is not as concise as natural language to express complex events. It may require thousands of fields to accurately depict factual information described in the reports.

Often, natural language processing (NLP) is used as a tool to automatize this process through information retrieval or text classification techniques [3] [8] [9]. However, the complexity problem is not overcome.

Our working assumption is that for a low granularity research, taxonomy/metadata is a good tool to look for documents, but that for finer granularity, preserving natural language might be a more reasonable choice. However, we understand that reading complex documents returned by a query to discriminate them or gathering them might be too time expensive. That is why we propose to use algorithms that show only extracts of the text pertaining to topics of interest for an analyst.

II. Method

A. Data

We have seen in the introduction that the data we are working on is from the ASRS database. Time-consuming work from experts is needed to make useful augmentations the data with an averaging 396 new reports per working day [1] (07/2019). In the context of this article, we have only worked with natural language and metadata that was provided by reporters (e.g. flight phase, altitude, narrative...) as opposed to using the augmented data, such as the synopses or the various metadata present in the database, which are not seen in the reporting forms. This is in line with the idea that NLP should be used to reduce the time spent by experts on pre-processing the input data to render it usable.

The incident reports we worked on are from after 2011 excluded. This is because writing style before and after this date are too different for our algorithm to be equally efficient on both, with strong use of abbreviations such as TFC for traffic, WX for weather or THRU for through, as well as the exclusive use of capital letters before 2012 [10]. We have chosen to work on the later reports because they are more readable and give a more modern vision of what are the current aviation safety issues.

B. Algorithm used

In this section, for clarity purposes, a more in-depth presentation of the NLP recent advances in transfer learning is given, and in particular, the algorithm that we worked with.

NLP is the branch of data science that focuses on natural language data. The work is done on textual data. However, all algorithms work on numerical data. Thus, a preliminary necessary step towards completing any kind of task in NLP is the act of converting the textual data into numerical data. This process is called embedding.

We distinguish between contextual and non-contextual embedding. In the non-contextual versions of embedding, one word matches a single vector. In the contextual embedding, the value of the vector that represents a word depends on the other words around it. One can understand intuitively the advantage of this kind of embedding, with these two sentences: ”He lives near the river bank.” and ”He works in a bank.” Here the word ”bank” has different meanings, that are determined by the context.

Contextual embedding is a key element in transfer learning in NLP. The main idea behind transfer learning is that an algorithm can use what he learned on a first task, the so-called pre-training task, to perform better on another task (downstream task). An example from computer vision that can help understand this intuitively is the following: an algorithm that was trained to recognize dogs in a picture will be faster at learning to recognize cats than an untrained algorithm (it will need fewer cat pictures to train on).

It has been found that teaching an algorithm to contextually embed language is a powerful pre-training task [11]. The advantage of this task is that it does not require labelled data. It is typically done by masking a word in a sentence and making a prediction on what the masked word is, using the words around it. By comparing the result with the real word, one can create a loss function, that allows for the training of our algorithm. We refer to this task as the language modelling task. We refer to algorithms that have been pre-trained on this task as language models.

As a general rule, pre-trained language models are used to boost performance on downstream NLP tasks in two ways. The first one is to use the embedding they produce as the input for the algorithm trained on the downstream task.

The other approach is to retrain the pre-trained model on the downstream tasks. This generally involves making task-specific changes to the model architecture and other techniques. This process is called fine-tuning.

The considerations above help understand the model we have chosen to use in the case of this article, which is a distilled version of BERT, fine-tuned on the Question Answering task.

BERT is a language model introduced by Google in 2018, that obtained state-of-the-art results on several tasks designed to gauge a language model’s ability to understand language, the Natural Language Understanding tasks (NLU).

Question Answering (QA) is one of those tasks where BERT achieved state-of-the-art performance when it was published. Typically, QA is done by providing a piece of text to our model, the so-called context, as well as a question. The algorithm’s goal is to provide an answer to the question by using information from the context. This can be done in various ways. In our case, the answer is directly extracted from the context. That is to say, the output of the algorithm is a span of the text.

The version of BERT we are working with is distilled, meaning it has fewer parameters than the original version (40% less). It runs 60% faster while preserving 95% of the original model performance, as measured on GLUE [12], which is a widely used Natural Language Understanding benchmark, that includes a QA data set: the Stanford Question Answering Dataset (SQuAD). All these characteristics make the model convenient in a time-sensitive research context while giving us a good idea of what are the language models’ abilities.

The model used in this paper has been trained on the SQuAD v1 data set.

Before presenting the method of evaluation of the algorithm, the following subsection introduces safety specific vocabulary; where some of the terms that were used in this paper when describing an incident occurrence are defined precisely.

C. Safety vocabulary

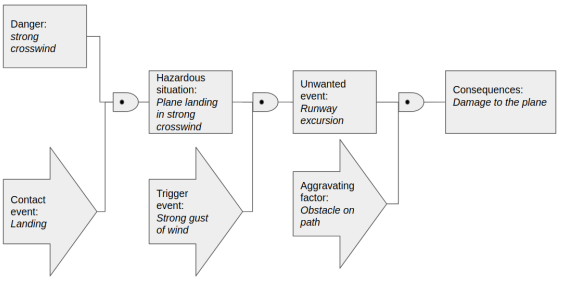

Fig. 1. Example of an accident scenario, adapted from Desroches et al. (2016, p.14)

In this subsection, we intend to give a brief overview of the Safety-related vocabulary that we will use when describing the occurrence of an incident [13]. It must be well understood when trying to answer any question on the topic.

Within an accident, we distinguish between three events: a contact event, a trigger event and an unwanted event.

The contact event is what causes an otherwise ’normal’ situation to become hazardous. It puts the system under study into contact with the hazard. As an example, entering a bad

weather zone puts the aircraft operated (system under study) into contact with a hazard (bad weather), creating a hazardous situation: flying into bad weather. Likewise, a strong crosswind (hazard), during a landing (contact event: aircraft ready to land), creates a hazardous situation: aircraft landing in a strong crosswind.

So far, nothing significant has happened. That stands in contrast with a trigger event, of which occurrence in a hazardous situation leads to an unwanted event. For instance, a strong gust upon landing (trigger event) which leads to a lateral runway excursion (the so-called unwanted event).

Finally, the final piece in this sequential view is the consequences. In our last example, that would be the damage to the plane or any kind of injuries sustained by the people flying on the plane. An aggravating factor could have been the presence of an obstacle on the path of the plane.

The considerations above are valuable when trying to understand what composes the often multifaceted chain of events described in a report. They give a linear, simplified, yet powerful working frame, that encompasses most of the steps that can happen.

The ASRS only has voluntarily submitted incident reports as opposed to accident reports [1]. It will be seen that for the reporter, an incident is not always reported through such a complete scenario or combination of events. Depending on what is mentioned in the report, the expert may consider that the 'incident' is the trigger event, or contact event or unwanted event or a series of them.

D. Evaluation method

The extracting kind of Question-Answering (Q&A) algorithm is used on thirty randomly selected reports to answer the following question:

"When did the incident happen?"

However, it has been seen in the subsection above that when one wants to make sense of an incident report, the term "incident" can mean various things, depending on what is written in the report. Also, one of the characteristics of natural language report is that the information in it is not codified. As such, one can characterize the time of happening of an event in the way that one sees fit. Effectively, many reporters choose to use the flight phase as their main indicators of when an incident happens. However, this is not always the case. The manifestations of the information of interest are of multiple nature, such as spatial ("10 000 ft") or contextual ("third touchdown", "training flight") for instance. This is why we chose to use a question with a large scope, that makes as few assumptions as possible regarding the content of the report. In this way, we hoped to find information that escapes the grasp of fixed taxonomies.

Because the scope of interpretation of the question can not be rigidly defined, we chose to rely on the following heuristic when trying to validate or invalidate the extracted answer.

According to Clark et al. (2013) [14], "NLP is concerned with producing artifacts that accomplish tasks. **The operative question in evaluating NLP is therefore the extent to which it produces the results for which it was designed.**"

We claim that our tool is useful if it does one or both of the following:

- Reporter's reading time to get the information of interest from the context is reduced.
- The extracted information is not captured by the metadata.

To evaluate our approach, we propose the sequential frame of work below:

1) Extract an answer from the context without reading the context first.
2) Write down information that can be deduced from the extracted answer, regarding the time of the incident. For instance, if the extracted answer is "touchdown", the following hypothesis can be made: "The incident happened during landing, upon touchdown.".
3) Verify if the deduced information from the extracted answer is correct, by reading the entire context, and/or using the metadata provided by the reporter.
4) Distinguish between an answer that is only correct, from an answer that provides information that is not covered by the metadata.

Because the question is factual, we expect it will not be hard to verify the veracity of the proposed answers. It does not mean that the task is easy, as we are working with domain-specific language, but no particular domain-specific training. Hence, one of our underlying intentions when doing this evaluation is to gauge the ability of the algorithm to provide reliable answers in this context.

III. Results

Among the thirty randomly selected reports, we get twenty-two 'good' answers.

Among those twenty-two answers, only eight give information that is not properly captured by the metadata. As an example, one answer gives the context of when the incident happened: 'unusually high volume of IFR traffic into and out of the airport.'. Another one of the extracted answers gives a range of altitude where the incident happened, as opposed to a single value, as imposed by the format of the reporting form (group A).

For two answers, the answer is arguably correct, but other extracts could have been selected as an answer as well. This is an inherent limit to our model that outputs a span of text as opposed to multiple ones (group B).

The rest of the answers overlap with the metadata provided by the reporter (group C).

Among the eight answers that can not be used to deduce anything regarding the incident, we notice the following.

One of the bad answers comes from the fact that there is no information regarding when the incident has happened in the text. A good answer, in that case, would have been to return an empty string (Group D).

Two of the bad answers occurred on reports describing a hazardous situation that lasted during multiple stages of the

flight. Here, the metadata also gives a fake sense that the incident, which is here a hazardous situation, happened during a single flight phase (group E).

Finally, for the five others, it seems the algorithm follows a heuristic by answering with extract of the text that describe time lengths with units such as “minutes” or “hours” (group F).

The extracted answers, the deduced information, as well as the ’item id’ for all the reports studied above, are provided in the first author’s GitHub page [15].

In the annexe, a sample for each of the groups above is provided, with the report and the extracted answer.

IV. Discussion

The proposed analysis is qualitative. We do not claim that these results can be generalized to the whole dataset, as the size of our sample is too small. However, it still gives a sense of what can be achieved with the kind of algorithm that was used and allows to learn more on developing a methodology adapted to this kind of evaluation.

Evaluating an NLP tool in a highly specialized context is not an easy task. We distinguish between two kinds of evaluations, the first being intrinsic, as opposed to the second which is extrinsic (Jones and Galliers, 1995) [16]. One can treat the ability to extract an answer that is correct as of the intrinsic criteria. This is because no matter what the context of use is, respecting the criteria above will always be part of the task given to the algorithm. It stands in contrast with the expectations that one has when working from an extrinsic approach, where one evaluates how well the algorithm fills its purpose in the specific context of the task. Here, we assume that what makes our task-specific are the following: the use of aviation safety language, the need for information that can not be captured by the metadata.

In this case, authors believe that the chosen evaluation method presents both qualities to a certain extent. From an intrinsic point of view, twenty-two answers out of the thirty answers are correct. The skills involved in producing these kinds of answers are:

- distinguishing written text related to the moment of an incident from the rest.
- when an answer exists in the text, extracting enough words so that information regarding the moment of the incident can be deduced upon reading the extract.

From an extrinsic point of view, only eight answers out of the thirty answers are considered correct. Producing these kinds of answers require a finer understanding of what is interesting from the user side. Our claim was that finding information produced through the extraction and deduction process that can not be captured by the metadata, constituted a fine criterion to adopt in that endeavour.

We feel that in the context of this article, we have shown promising results regarding the intrinsic abilities of the algorithm, even when used in a domain-specific language setting. We have also proposed a novel way to evaluate this kind of algorithm, that relies on a four-step method, that encompassed both intrinsic and extrinsic qualities.

V. Conclusion

In this article, our general intention was to gauge the usefulness of one of the fine-tuned algorithms, from the recent generation of NLP attention-based algorithms [17], in the context of the aviation safety industry. We have found promising results, however, there is still work to do on different levels. We believe the three following points would be worth exploring further:

First is the transparency of the model. Having the ability to explain why an algorithm chose an answer instead of another [18] [19].

Second is the choice of the question. Our question was quite factual, as such, a strong overlap with the metadata should have been expected. For instance, a question such as “What contributed to the incident?” should provide answers that may be extrinsically more interesting.

Thirdly, we have not studied how differently the model behaves once it receives specific training, for instance through the use of safety expert feedback under the form of a supplementary training data set.

Finally, on a broader level, we have not yet explored all the possibilities offered by the most recent technologies in the NLP field, such as the T5 algorithm from Google [20]. The more we know about their strong and weak points in our specified context of use, the more we will be able to produce useful hindsight, in particular, if we cleverly combine them.

References

[1] ASRS. Asrs program briefing. [Online]. Available: https://asrs.arc.nasa.gov/docs/ASRS_ProgramBriefing.pdf

[2] N. Tulechki, “Natural language processing of incident and accident reports : application to risk management in civil aviation,” phdthesis, Université Toulouse le Mirail - Toulouse II, Sep. 2015. [Online]. Available: https://tel.archives-ouvertes.fr/tel-01230079

[3] M. S. Ahmed, L. Khan, N. C. Oza, and M. Rajeswari, “Multi-label ASRS Dataset Classification Using Semi Supervised Subspace Clustering,” in *CIDU*, 2010.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *arXiv:1810.04805 [cs]*, May 2019, arXiv: 1810.04805. [Online]. Available: http://arxiv.org/abs/1810.04805

[5] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language Models are Unsupervised Multitask Learners,” 2019.

[6] ASRS. Asrs database online. [Online]. Available: https://asrs.arc.nasa.gov/search/database.html

[7] ——. Electronic report submission. [Online]. Available: https://asrs.arc.nasa.gov/report/electronic.html

[8] B. Xu and S. Kumar, “A Text Mining Classification Framework and its Experiments Using Aviation Datasets,” Jan. 2015.

[9] M. A. U. Abedin, V. Ng, and L. Khan, “Cause Identification from Aviation Safety Incident Reports via Weakly Supervised Semantic Lexicon Construction,” *Journal of Artificial Intelligence Research*, vol. 38, pp. 569–631, Aug. 2010, arXiv: 1401.4436. [Online]. Available: http://arxiv.org/abs/1401.4436

[10] L. Tanguy, N. Tulechki, A. Urieli, E. Hermann, and C. Raynal, “Natural language processing for aviation safety reports: From classification to interactive analysis,” *Computers in Industry*, vol. 78, pp. 80–95, May 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0166361515300464

[11] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," *arXiv:1802.05365 [cs]*, Mar. 2018, arXiv: 1802.05365. [Online]. Available: http://arxiv.org/abs/1802.05365
[12] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding," *arXiv:1804.07461 [cs]*, Feb. 2019, arXiv: 1804.07461. [Online]. Available: http://arxiv.org/abs/1804.07461
[13] A. N. D. M. D. S. Desroches, A., *Analyse globale des risques: principes et pratiques*. Lavoisier-Hermes, 2016.
[14] F. C. Clark, A. and S. Lappin, *The Handbook of computational linguistics and natural language processing.* Willey-Blackwell, 2013.
[15] Bert_qa_asrs. [Online]. Available: https://github.com/Sharing-Sam-Work/BERT_QA_ASRS/
[16] K. S. Jones and J. R. Galliers, *Evaluating Natural Language Processing Systems: An Analysis and Review.* Springer Science, 1995.
[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," *arXiv:1706.03762 [cs]*, Dec. 2017, arXiv: 1706.03762. [Online]. Available: http://arxiv.org/abs/1706.03762
[18] J. Vig, "A Multiscale Visualization of Attention in the Transformer Model," *arXiv:1906.05714 [cs]*, Jun. 2019, arXiv: 1906.05714. [Online]. Available: http://arxiv.org/abs/1906.05714
[19] D. Yogatama, C. d. M. d'Autume, J. Connor, T. Kocisky, M. Chrzanowski, L. Kong, A. Lazaridou, W. Ling, L. Yu, C. Dyer, and P. Blunsom, "Learning and Evaluating General Linguistic Intelligence," *arXiv:1901.11373 [cs, stat]*, Jan. 2019, arXiv: 1901.11373. [Online]. Available: http://arxiv.org/abs/1901.11373
[20] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *arXiv:1910.10683 [cs, stat]*, Oct. 2019, arXiv: 1910.10683. [Online]. Available: http://arxiv.org/abs/1910.10683

ANNEX:

- Group A: 'Before boarding, I walked to main cabin to deliver catering items. While there I noticed that one of the oxygen walk-around masks did not appear to be correctly connected to the bottle. As I have been finding these regularly, I took it upon myself to check more closely and found that both oxygen bottles in the aft cabin did not have the mask tubing properly connected to the bottle. I told the crew of my discovery and shared my previous experience with them. I suggested that they look more carefully during preflight checks because sometimes it is not immediately evident that the tubing is not connected. I reported the discrepancy to the Captain and asked that the issue be entered in the aircraft logbook. He did and called Maintenance who came and rectified the issue during boarding. Flight attendant training should include emphasis on checking tubing connection on oxygen bottles during every pre-flight check. As mentioned, it is sometimes difficult to see if the tubing is correctly attached because of the confined locations of emergency equipment and other factors.'
  Extracted answer:Before boarding
- Group B: 'A buzz-squeal picked up in VHF1 from transmissions from VHF3 covered transmissions from ATC, from approach at 2000 AGL until we were at the gate. My FO and I could not hear the taxi instructions during rollout from tower and were forced to ask for re-transmit. The buzz-squeal interference was picked up from approach, during runway exit, during taxi, and during parking. This aircraft, seems worse than others in the past, but this is a problem that affects a majority of the fleet. We experienced this problem in DEN on 133.3, 121.85, and 131.97. Honeywell radios have a long standing problem that needs to be fixed '
  Extracted answer:during runway exit
- Group C: 'We were on a heading diverting north of ZZZ at FL340 when the First Officer side pitot froze up. Right at this time we both saw lightning flashes and smelled something burning (may have been ozone). EICAS showed an SPS [Stall Protection System] Advanced and MFD had a CAS message. Airspeed on First Officer side started dropping and eventually went to zero and displayed no data (X's). Autopilot failed and we requested a descent to get down to an altitude where the temperatures would be above freezing. Our alternate was ZZZ1 and we declared an emergency and told Center that we wanted to divert there. In the descent the pressurization failed. We notified Dispatch and Maintenance that we were diverting. We told the Flight Attendant that we were diverting and made an announcement to the passengers. Once we got around 10,000 the pitot unfroze and the aircraft was back in a desirable aircraft state. We continued with the diversion to ZZZ1 and landed uneventfully. Taxi to gate and called Maintenance Control to have Contract Maintenance sent to the aircraft. Pitot system frozen (ADC Failure) with convective weather nearby. Other than exiting freezing conditions I don't think there was much more we could do. Our destination airport had gone into holding for the same line of weather we were trying to avoid and we were not going to head into holding with an ADC failure & possible lightning strike.'
  Extracted answer: FL340
- Group D: 'AFT FA said she need[ed] to make an ice bag for a passenger. I ask[ed] her did the passenger injure herself on our equipment or was it a preexisting condition. She replied that the passenger hit her head on the overhead bin. The AFT FA gave the passenger the ice bag to the passenger. I explained to the AFT FA the she needs to get the details of what happened and the name of the passenger and complete a [report]. I personally did not see what happen[ed], but the passenger sat with the ice bag to the right side of her head during the entire flight. The passenger asked for additional ice during flight. Passenger deplaned the AC without a complaint related to the injury, returned the bag of melted ice and said Thank You for the assistance. Passenger did not use caution while moving around the seating area.'
  Extracted answer: the passenger sat with the ice bag
- Group E: 'During our pre-flight the Captain advised us that it may be bumpy going in to our destination and that he may have us sit down and then just ding us at 10,000. I explained that although the company and its pilots have been using this as standard practice to give customers more time to use the onboard WiFi that it was in violation of both company (manual) policy as well as a violation

of two FARs. I thought that we had the situation under control until we hit the bumps and the Captain told us that he was sitting us down but the customers could continue to use all electronic devices. We only have a secure cabin or an unsecure cabin. Our flight landed unsecure. The customers had tray tables back down and all electronics on. Had an evacuation happened, customers would have wasted great amounts of time trying to get around all the laptops and other electronic devices.'
Extracted answer: pre-flight

- Group F: 'We departed Runway 24 for right traffic pattern while practicing landings with a student. Traffic that departed after us was also closed traffic for runway 24 right traffic, they were advised to follow us. As my student turned downwind, I noticed a shadow of an airplane over the ground flying towards our direction. As that caught my attention, I started looking for traffic. After scanning for about 3-4 seconds, I noticed a low wing aircraft flying right at us. I immediately took the controls and initiated a climb to take evasive action. The airplane continued their downwind turn as we continued to climb and we were exactly above the airplane. I advised the controller that we started an immediate climb due to traffic. The traffic below us was asked to make a left 360 degree turn on the downwind and then we continue the traffic pattern in front of them. We landed normally.'
Extracted answer: 3-4 seconds