Can Social Media Inform Dietary Approaches for Health Management? A Dataset and Benchmark for Low-Carb Diet

Skyler Zou[1,4*] **Xiang Dai**[1] **Sarvnaz Karimi**[1] **Pennie Taylor**[2] **Grant Brinkworth**[3]
[1]CSIRO Data61 [4]University of New South Wales
[2]CSIRO Health and Biosecurity, AEHRC
[3]CSIRO Health and Biosecurity, Human Health
{firstname.lastname}@csiro.au

Abstract

Social media offers an accessible avenue for individuals of diverse backgrounds and circumstances to share their unique perspectives and experiences. Our study focuses on the experience of low carbohydrate diets, motivated by recent research and clinical trials that elucidates the diet's promising health benefits. Given the lack of any suitable annotated dataset in this domain, we first define an annotation schema that reflects the interests of healthcare professionals and then manually annotate data from the Reddit social network. Finally, we benchmark the effectiveness of several classification approaches that are based on statistical Support Vector Machines (SVM) classifier, pre-train-then-finetune RoBERTa classifier, and, off-the-shelf ChatGPT API, on our annotated dataset. Our annotations and scripts that are used to download the Reddit posts are publicly available at https://data.csiro.au/collection/csiro:59208.

1 Introduction

Current practice and general perception recommend carbohydrates as a major contributor to dietary energy intake—Acceptable Macronutrient Distribution Ranges specify carbohydrates $45\% - 65\%$ of energy intake (eatforhealth, 2002). However, recent research spanning clinical trials, systematic reviews, and meta-analyses substantiates the benefits of *low-carb diets*—$< 26\%$ of total energy intake from carbohydrate or less than 130g of carbohydrate per day—especially for individuals with Type II Diabetes, including weight loss, blood glucose control and reducing cardiovascular disease risk and remission of Type II Diabetes (Alzahrani et al., 2021; Goldenberg et al., 2021). A low-carb diet typically involves reducing the intake of carbohydrates and increasing the proportion of dietary protein and fats (Feinman et al., 2015). Whilst a large body of controlled clinical trials has been conducted (Pavlidou et al., 2023), there is limited research on the experience of customers following a low-carb diet. The term *experience* entails health effects observed in practice, the challenges or barriers that people may face, and the extent of social support or advice, etc. Public perception is another aspect to probe. By understanding commonly perceived health effects and misconceptions of low-carb diets, health professionals can tailor strategies to educate and increase the accessibility of low-carb diets.

We consider social media as a valuable potential source of real-world insight to support scientific research and inform health professionals on personal perspectives and experiences of low-carb diets. Our aim is to quantify the extent of insights that can be mined from social media firsthand which could indirectly inform dietitians in clinical health management. Given the lack of similar studies in the NLP community, the first step is to create a dataset that facilitates such research. We create a dataset that reflects the interests of health professionals and focus on identifying Reddit posts about health responses, barriers, and advice. We manually annotate comments and submissions from the low carbohydrate diet community on Reddit (r/lowcarb), ensuring reasonable inter-annotator agreement. Finally, we benchmark the effectiveness of several classifiers, including Support Vector Machine-based, pre-train-then-finetune RoBERTa-based, and off-the-shelf ChatGPT-based classifiers.

Related work Social media has been extensively investigated to inform health care professionals regarding epidemic intelligence (Joshi et al., 2019), pharmacovigilance (Nikfarjam et al., 2015; Karimi et al., 2015), or vaccination hesi-

*This work was partially done when Skyler was a summer intern at CSIRO Data61.

The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, pages 406–412

tancy (Dunn et al., 2015; Khoo et al., 2022). However, there are hardly any studies from the NLP community paying attention to dietary practices. Hansen and Hershcovich (2022) investigated mining arguments for the transition to sustainable dietary practices (plant-based diets) on Twitter with crowd-sourced annotations. They focused on identifying claims supported with sufficient evidence, including anecdotal, expert, study, fact, or normative. They also annotated their dataset for stance. Their study highlighted the need for sustainability aspects to be considered for design of diet programs. They also mention a restriction in their dataset being lack of context given the nature of the tweets.

To our knowledge, however, there is no study that focuses on identifying perspectives and experiences of the low-carb diet and that is the focus of our study.

2 Dataset

We explored two data sources—Twitter and Reddit—in the early stages of our work. We found that different from tweets that were retrieved through keyword searching and thus are often off-topic, the low-carb subreddit (r/lowcarb) exhibits longer posts with guided discussion and richer information. Note that the subreddit fosters a community for people interested in low-carb diets, and may have a selection bias toward more positive user experiences. However, there is still ample conversation on the challenges and obstacles of a low-carb diet. Thus, Reddit was selected as the main data source as we believe its pros outweigh the cons. We leave the investigation of other data sources for future work.

The Reddit data follows a tree-like structure. Users can submit *submissions* with a title and an optional body paragraph, which initiates discussion on a defined topic relevant to low-carb diets. Redditors may reply to submissions in the form of *comments*, which can also be directly replied to. We use the Pushshift API (https://github.com/pushshift/api), which can be used to search all publicly available comments and submissions on Reddit, to collect data.[1] When performing the search, we specify only the name of the subreddit (i.e., *lowcarb*), and the API will search for the most recent comments and submissions within the low-carb subreddit. The searches were conducted several times in 2022-December and 2023-January, and finally, we collect 1570 unique comments and 1210 unique submissions, respectively.

Annotation schema Initial exploration of the Reddit data provided a glimpse into its nature and characteristics. We note that many posts are advisory, responding with recommendations to a query or experience, with low-carb recipes frequently posted. Nutritionists are often interested in understanding what challenges impede the progression of the low-carb diet and what health response users experience. Therefore, we defined three categories of interest: (1) health response, (2) barrier, and (3) advice. Any remaining post is categorised as "Other". Table 1 provides a short description of these categories.[2]

Annotation process The annotations were conducted using the Label Studio interface (https://labelstud.io/). There were five annotators: two are dietary experts and the remaining three with computer science backgrounds. We conducted a total of four annotation rounds: for the first three rounds, the same set of examples/posts were annotated by all annotators, after which the annotators met to discuss the disagreement at the end of each round. In the main final round, annotators were assigned different sets of data. We frame the task as a multi-label classification problem. That is, annotators were allowed to assign more than one category to the post. However, *Other* category can not be annotated together with other categories. In other words, only if an example does not belong to any of the three categories of interest, it is annotated as *Other*.

Inter-annotator agreement There are in total 200 posts annotated by all five annotators, and we measure inter-annotator agreement on these multi-annotated examples. For each category of interest, we compute pair-wise inter-annotator Cohen's kappa coefficient (Cohen, 1960; McHugh, 2012). We observe that the *Health Response* category is relatively straightforward to annotate and annotators reach a moderate agreement level (averaged Cohen's kappa across annotator pairs: 0.59). In contrast, *Barrier* is the most challenging category,

[1] We note that this API has since been discontinued. We therefore only share the annotations with post IDs. The original posts can be directly obtained from Reddit.

[2] Detailed annotation guidelines are available at https://data.csiro.au/collection/csiro:59208.

Category	Definition	Example
Health Response	Physiological or psychological response, perceived or experienced. This includes changes in weight, changes in body composition and body fat levels, changes in blood sugar metrics (HbA1c, A1c, glucose spikes and variability, post-prandial glucose, hypoglycemia, etc.), changes in Blood lipid (fat) levels, changes in Blood pressure, changes in medication requirements, kidney health or function, risks or side effects (e.g., constipation, insomnia, hair loss, diseases), emotional responses (e.g., happy, sad, frustrated), mental health (e.g., experiencing anxiety or depression).	Since starting this diet last year lost 40 pounds. I was about 280 in the first picture and weigh about 240 now. Goal weight is 190.
Barrier	Circumstances or challenges impeding initiation or progression of a low-carb diet. This includes lack of support (e.g., professional or family), inaccessibility of resources, lack of availability of desirable food, lack of knowledge or understanding, craving sugar or carbs, and mental health.	Cannot keep my diet if my partner buys fast food for dinner every day!
Advice	Recommended actions to take on the topic of low-carb diets. This includes the suggestion of action (i.e., how to, try this, or go to), seeking advice, implicit or passive advice (e.g., providing a recipe or mentioning a resource), and intention to assist.	Is there any tips and tricks you have found that helped you with starting low carb for someone with diabetes?/ My biggest tip is don't try to find low carb replacements for carby foods like low carb pasta or low carb bread.
Other	Posts that do not satisfy the criteria of the Health Response, Barriers, and Advice categories fall under "Other". This embodies the texts containing information that is not of interest to the research or in a format suitable for effective NLP analysis.	Interesting. Thanks for sharing.

Table 1: Defined categories, their definitions, and examples.

where averaged Cohen's kappa of 0.37 indicates a fair level of agreement. The other observation is that dietary experts tend to disagree with each other more often than layman annotators. One reason behind this scenario is that dietary experts may use their domain knowledge to interpret users' experience (e.g., pregnancy is considered a barrier, as it may impede the progression of low-carb diets), in contrast, layman annotators rely primarily on language patterns without inferring any information that is not explicitly stated.

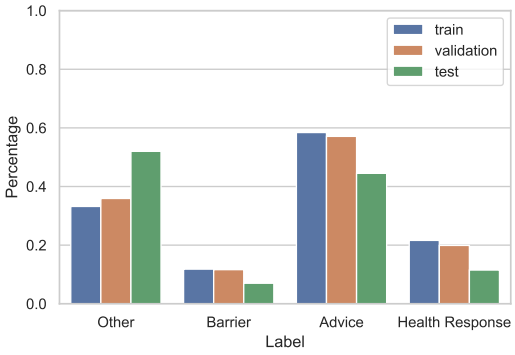

Figure 1: Label distribution of our annotated dataset. Note that one example can be assigned to multiple categories, the summed percentage over all categories in each set can thus be larger than one.

3 Benchmark Results

To evaluate the viability of our annotated dataset, we build several representative classifiers and test their effectiveness on the dataset.

Train-validation-test split We split our dataset into training, validation, and testing sets. The testing set contains all examples that are annotated by multiple annotators (200), and the remaining examples are randomly split into training (85%, 2193) and validation (15%, 387) examples. The label distribution is shown in Figure 1.

Evaluation metrics We first create the gold label for each test example via a simple majority rule. That is, each category (*Health response*, *Barrier*, *Advice*) is added to the gold labels if more than half annotators choose that category. If no category has more than half votes, the *Other* category is assigned. We then compute the Micro F_1 and Macro F_1 scores of the model predictions against these created gold labels. We call this evaluation approach *Hard* evaluation.

We also employ the *Soft* evaluation (Uma et al.,

2021) via comparing model predictions against each annotator's annotations. That is, if one predicted label for the test example could be found in one annotator's annotations, it is counted as $1/5$ true positive,[3] otherwise, $1/5$ false positive. Similarly, if the annotated label by one annotator is not found in the model predictions, it is counted as $1/5$ false negative. Finally, the Micro F_1 and Macro F_1 scores are calculated.

Classifiers We test the effectiveness of three representative classification approaches that are based on statistical Support Vector Machines (SVM), pre-train-then-finetune RoBERTa classifier, and, off-the-shelf ChatGPT API.

- NB-SVM (Wang and Manning, 2012) is a strong and efficient SVM variant that uses Naive Bayes (NB) log-count ratios as feature values. We train three NB-SVM binary classifiers for each category of interest on the training set. During inference, these three classifiers are applied separately, and, if no positive label is predicted by any of these three classifiers, 'Other' is assigned to the example.

- RoBERTa (Liu et al., 2019) is a transformer-based encoder that is pre-trained using a masked language modeling objective. To adapt RoBERTa for text classification, we built a multi-label classification head on top of the RoBERTa encoder. The additional head takes the contextualized representations generated by the encoder and maps them to the target labels. During the fine-tuning process, the parameters of the RoBERTa encoder and the classification head are jointly optimised using a binary cross-entropy loss on the training set. We test both the base and large versions of RoBERTa in our experiments.

- ChatGPT (Ouyang et al., 2022) is a large language model based on the GPT (Generative Pre-trained Transformer) architecture, where the original GPT-3 model (Brown et al., 2020) is further fine-tuned using supervised learning on a dataset of demonstrations of the desired model behavior, and reinforcement learning from human feedback. To use ChatGPT for text classification, we construct a prompt for each test example (Figure 2) which is taken as the input of the model, and the model returns a free-text response, from which the predicted label could be inferred. Similar to NB-SVM, we build three binary classifiers that use zero-shot or 5-shot learning—five randomly selected exemplars from the training set—with OpenAI API (gpt-3.5-turbo, https://platform.openai.com/docs/models/gpt-3-5).

[3]One test example is annotated by 5 annotators.

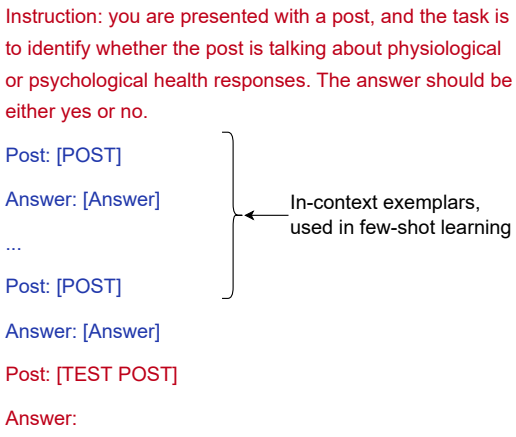

Figure 2: An illustration of the prompt, consisting of the instruction, a test example, and optionally in-context exemplars.

Method	Category			Overall	
	H.R.	B.	A.	Micro	Macro
	Hard evaluation				
NB-SVM	60.0	48.0	80.7	73.5	62.9
RoBERTa-Base	**86.4**	66.7	76.4	77.0	76.5
RoBERTa-Large	82.4	**68.8**	**81.0**	**79.9**	**77.4**
ChatGPT (0-shot)	68.1	43.2	54.5	55.5	55.3
ChatGPT (5-shot)	47.6	33.8	66.3	55.3	49.2
	Soft evaluation				
NB-SVM	58.1	43.6	71.8	66.0	57.8
RoBERTa-Base	72.6	47.8	72.7	69.7	64.4
RoBERTa-Large	**73.2**	**57.5**	**76.5**	**73.6**	**69.1**
ChatGPT (0-shot)	59.8	44.7	50.2	51.2	51.6
ChatGPT (5-shot)	42.4	38.4	61.1	52.2	47.3
Majority rule	78.5	67.1	82.6	80.0	76.1

Table 2: Comparison of different classifiers in terms of F_1 scores. H.R. stands for Health Response; B. for Barrier; and, A. for advice.

Classification results Table 2 shows the evaluation results in terms of F_1 scores of each category as well as micro-averaged and macro-averaged

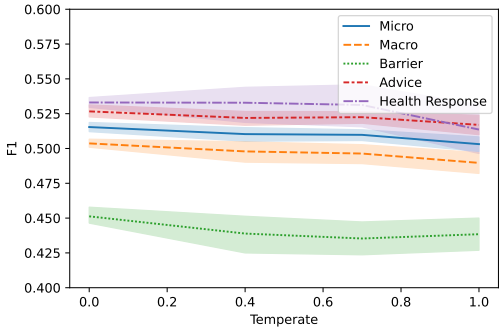

Figure 3: The impact of temperate using ChatGPT (0-shot). We repeat all experiments 10 times. The plot shows an estimate of the central tendency and the confidence interval.

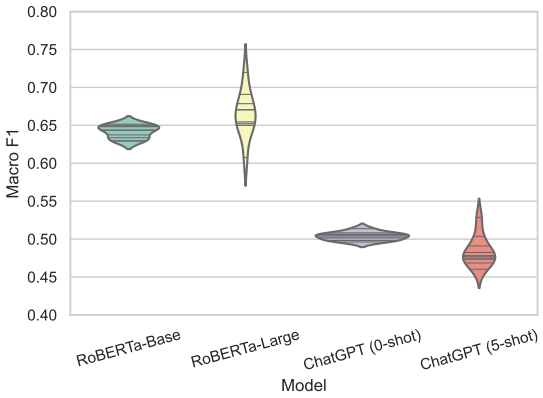

Figure 4: The impact of randomness—random weights initialization in RoBERTa-based classifiers and random sampling and in-context exemplars in ChatGPT-based classifiers—on Macro F_1 scores. We repeat all experiments 10 times. The stick shows each experimental result.

scores. First, we observe that supervised fine-tuned models (NB-SVM, RoBERTa) outperform ChatGPT-based classifiers by a large margin in both hard evaluation and soft evaluation results. Secondly, we find that hard evaluation results tend to overestimate the effectiveness of these classifiers. We conjecture that creating a single set of gold labels may reduce the task difficulty via tolerating more 'Other' predictions. For example, if one example is annotated by two annotators as 'Barrier' and the other three as 'Other', the merged gold label is 'Other' via simple majority rule, because there are no more than half votes on any category of interest. A model prediction of 'Other' does not count as an error under hard evaluation, but contributes 0.4 false negatives under soft evaluation.

Lastly, we observe that there is no clear benefit of 5-shot over zero-shot with ChatGPT (except for 'Advice' category), showing that the classifier relies primarily on semantic priors from pretraining rather than in-context exemplars.

How stable are ChatGPT classifiers? One important feature of the ChatGPT-based classifier is that it outputs a sequence of text rather than discrete labels. Although we specify in the prompt that "The answer should be either yes or no", the responses are still very diverse (e.g., "...not enough information...", "...unclear...", "...cannot be determined..."). To control the randomness of sampling, we choose different temperature values from a list of numbers: {0, 0.4, 0.7, 1} and observe its impact on zero-shot F_1 scores. Figure 3 shows that lower values like 0 make the classifier more effective and stable. Therefore, we choose a temperature of 0 for the following experiments.

Another factor that may cause instability in ChatGPT-based few-shot classifiers is the selection of in-context exemplars (Zhang et al., 2022). We test the impact of in-context exemplars by randomly selecting different examples. From Figure 4, we find that although ChatGPT 5-shot is indeed more unstable than zero-shot where only instructions are provided in the prompt without in-context exemplars, it is still more stable than supervised training with RoBERTa-large. The latter has been observed by Mosbach et al. (2021) where training with multiple random seeds results in a large variance of effectiveness.

4 Summary

We create a dataset that consists of Reddit posts talking about low-carb diets. These posts contain rich information about health responses relating to low-carb diet consumption, barriers impeding the progression, and suggested actions to take. Our benchmark results on the annotated dataset show that although ChatGPT is promising at classifying Reddit posts into defined categories of interest, it still underperforms supervised trained models by a large margin.

Acknowledgements This study has ethics approval from CSIRO's ethics committee (2020_050_LR) for CSIRO Low-Carb Diet Branded Convenient Meals - Impact Evaluation Study project. Skyler Zou worked on this project during his internship at CSIRO's Data61.

References

Ahmad H Alzahrani, Mads J Skytte, Amirsalar Samkani, Mads N Thomsen, Arne Astrup, Christian Ritz, Elizaveta Chabanova, Jan Frystyk, Jens J Holst, Henrik S Thomsen, et al. 2021. Body weight and metabolic risk factors in patients with type 2 diabetes on a self-selected high-protein low-carbohydrate diet. *European Journal of Nutrition*, 60(8):4473–4482.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, and Amanda Askell. 2020. Language models are few-shot learners. In *Conference on Neural Information Processing Systems*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Adam G Dunn, Julie Leask, Xujuan Zhou, Kenneth D Mandl, and Enrico Coiera. 2015. Associations between exposure to and expression of negative opinions about human papillomavirus vaccines on social media: an observational study. *Journal of medical Internet research*, 17(6):e4343.

eatforhealth. 2002. Nutrient Reference Values for Australia and New Zealand. *Webpage (last accessed 24 April 2023)*.

Richard D. Feinman, Wendy K. Pogozelski, Arne Astrup, Richard K. Bernstein, Eugene J. Fine, Eric C. Westman, Anthony Accurso, Lynda Frassetto, Barbara A. Gower, Samy I. McFarlane, Jörgen Vesti Nielsen, Thure Krarup, Laura Saslow, Karl S. Roth, Mary C. Vernon, Jeff S. Volek, Gilbert B. Wilshire, Annika Dahlqvist, Ralf Sundberg, Ann Childers, Katharine Morrison, Anssi H. Manninen, Hussain M. Dashti, Richard J. Wood, Jay Wortman, and Nicolai Worm. 2015. Dietary carbohydrate restriction as the first approach in diabetes management: Critical review and evidence base. *Nutrition*, 31(1):1–13.

Joshua Z Goldenberg, Andrew Day, Grant D Brinkworth, Junko Sato, Satoru Yamada, Tommy Jönsson, Jennifer Beardsley, Jeffrey A Johnson, Lehana Thabane, and Bradley C Johnston. 2021. Efficacy and safety of low and very low carbohydrate diets for type 2 diabetes remission: systematic review and meta-analysis of published and unpublished randomized trial data. *bmj*, 372.

Marcus Hansen and Daniel Hershcovich. 2022. A dataset of sustainable diet arguments on Twitter. In *Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*, pages 40–58, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Aditya Joshi, Sarvnaz Karimi, Ross Sparks, Cécile Paris, and C Raina Macintyre. 2019. Survey of text-based epidemic intelligence: A computational linguistics perspective. *ACM Computing Surveys (CSUR)*, 52(6):1–19.

Sarvnaz Karimi, Chen Wang, Alejandro Metke-Jimenez, Raj Gaire, and Cecile Paris. 2015. Text and data mining techniques in adverse drug reaction detection. *ACM Computing Surveys*, 47(4).

Aleney Khoo, Maciej Rybinski, Sarvnaz Karimi, and Adam Dunn. 2022. The role of context in vaccine stance prediction for twitter users. In *Proceedings of the The 20th Annual Workshop of the Australasian Language Technology Association*, pages 16–21.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv*, 1907.11692.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2021. On the Stability of Fine-tuning BERT: Misconceptions, Explanations, and Strong Baselines. In *International Conference on Learning Representations*.

Azadeh Nikfarjam, Abeed Sarker, Karen O'connor, Rachel Ginn, and Graciela Gonzalez. 2015. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association*, 22(3):671–681.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. 2022. Training language models to follow instructions with human feedback. In *Conference on Neural Information Processing Systems*.

E Pavlidou, SK Papadopoulou, A Fasoulas, M Mantzorou, and C Giaginis. 2023. Clinical evidence of low-carbohydrate diets against obesity and diabetes mellitus. *Metabolites*, 13(2).

Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from Disagreement: A Survey. *Journal of Artificial Intelligence Research*, 72.

Sida Wang and Christopher Manning. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 90–94, Jeju Island, Korea. Association for Computational Linguistics.

Yiming Zhang, Shi Feng, and Chenhao Tan. 2022. Active example selection for in-context learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9134–9148, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.