A Peer-to-Peer Corpus for Conversational Agents for Long-Distance Relationships

Naryn Samuel[1], Nicholas Caporusso[1], Devyn Ferman[1]

[1] Department of Computer Science, Northern Kentucky University,
Louie B Nunn Dr, 41099 Highland Heights, United States

ABSTRACT

Recent advances in machine learning, including the development of more effective natural language processing (NLP) models, have enabled the use of text classification and generation algorithms, sentiment and emotion detection models, and intelligent conversational agents, in different domains, from business to healthcare. Specifically, intelligent and conversational agents (e.g., chatbots) are currently incorporated in many applications (e.g., customer care and decision support systems) to automate tasks while simultaneously providing users with a more credible and natural human-like interaction. The availability of NLP corpora is crucial for training conversational agents and increasing their quality and performance. Nevertheless, the availability of domain-specific NLP corpora is crucial for training conversational agents, especially in applications that focus on mental health counseling and support. In this paper, we introduce a corpus especially designed for NLP tasks that focus on providing bi-national couples in a long-term relationship with mental health support. Our dataset contains over 4000 posts and users' reactions published on social media groups dealing with COVID-19 travel restrictions. We detail the content of the dataset, its format, and its use in the development of NLP applications.

Intelligent Human Systems Integration (IHSI 2022): Integrating People and Intelligent Systems
https://openaccess.cms-conferences.org/#/publications/book/978-1-7923-8988-7

Keywords: Human-Computer Interaction, Natural Language Processing, Machine Learning, conversational agents

INTRODUCTION

In addition to automated Decision Support Systems, intelligent agents are being increasingly utilized in many different domains to support humans in accomplishing their tasks. Research demonstrated the applicability of virtual agents, with specific regard to conversational agents and chatbots, as an effective system for providing users with support in the form of answers to frequent questions, simplified and more natural interfaces for searching for information, and human-like interaction. Several studies introduced the use of chatbots in the healthcare domain for a variety of tasks: in addition to analyzing their performance in accomplishing the expected goals, research confirmed the effectiveness of conversational agents in creating engaging and positive user experiences. Recently, dedicated chatbots have been introduced for addressing mental health conditions. Several systems have been developed for supporting individuals suffering from mental disorders, in situations of stress and anxiety, or who are reluctant to seek mental health advice. The availability of corpora is crucial for developing and training conversational agents. To this end, dedicated groups on social networks, forums, and online peer support communities are especially useful for acquiring datasets for addressing specific circumstances and issues, improving the design of Natural Language Processing (NLP) systems, and training Machine Learning algorithms. The scientific literature documents repositories and corpora generated by collecting information publicly available on Twitter, Reddit, Facebook, and other websites.

In this paper, we introduce a corpus especially dedicated to supporting couples who are in long-distance relationships. Our work leverages the experience of thousands of individuals who have been separated during the COVID-19 pandemic when the governments of many countries enacted travel restrictions that prevented couples from reuniting and forced them into long-distance relationships: travel bans, which were introduced in March 2020 and maintained by several countries (e.g., the United States) until most of 2021, especially affected thousands of bi-national couples subject to VISA-related restrictions, who experienced a prolonged situation of stress, anxiety, and depression, that impacted their mental health. In our work, we collected publicly available data published on websites and groups dedicated to offering peer-to-peer support to individuals who have been separated from their partners due to COVID-19-related restrictions. Our corpus contains over 16 months of data and more than 4000 posts and their reactions, which provide a rich representation of conversations and interactions that happened in the group and offers insight into peer-support dynamics. Although our data refer to situations caused by COVID-19-related travel restrictions, the content of the corpus is applicable to long-distance relationships in general, and it is particularly suitable for realizing research on peer-support as well as for developing new applications, conversational agents, and intelligent systems that can offer psychological and sentimental help to individuals and couples.

RELATED WORK

The availability of NLP corpora is fundamental for developing, training, and fine-tuning conversational agents and other systems based on machine learning. Most corpora for training conversational agents consist of information retrieved from large collections of text documents such as books, newspapers, and popular websites (Pháp, H.C., 2016), which provide researchers and developers with vast datasets that are suitable for creating extensive and robust language models. In recent years, several NLP corpora have been developed and made available through public and open access repositories, which, in turn, enabled to improve the accuracy of NLP models and build better applications. Nowadays, thanks to the effort of many research groups worldwide, there are models for most languages. Although many existing datasets are crucial for building language models that are suitable for general-purpose applications, due to their lack of specificity they are not applicable to tasks that require using domain knowledge or reproducing dynamics such as emotion and sentiment, which are more typical of human-like interaction. Several studies demonstrated that in addition to more general models, many NLP applications require topic-specific corpora that reduce ambiguity and enable fine-tuning the models and specializing them for the specific need of a particular application (Ezzini et al., 2021) (Gu et al., 2021). For instance, prior research demonstrated that smaller models trained with domain-specific corpora can achieve better performances compared to larger pre-trained language models, such as BERT (Edwards et al., 2020).

As a result, recently, the focus of researchers has shifted toward creating corpora that enable training NLP models that are focused on the task, that is, with datasets that contain domain-specific knowledge. In addition, many developed sentiment analysis and emotion classification models that improve the affective component of interaction by incorporating human dynamics. To this end, user-generated content in forums, online communities, and, more recently, social networking websites has been especially instrumental in providing researchers with an invaluable source of data. For instance, the authors of (Manohar & Kulkarni, 2017) trained their NLP model with content from social media to better detect emotion and classify tweets based on their sarcasm. This, in turn, has enabled the development of NLP toolkits especially designed to work with datasets acquired from social media (Pérez et al., 2021).

Consequently, collecting specialized corpora, making them available to the scientific community, and rendering them easily accessible has become even more crucial for fostering the development of new NLP applications that require taking into consideration the dynamics of human emotions. In addition to fostering the development of new systems, the availability of corpora enables automated analysis of human dynamics from text, which, in turn, promotes a better understanding of human dynamics and the development of practices to deal with them. For instance, in (Afshar et al., 2019) a corpus consisting of electronic health records was utilized to identify alcohol misuse in patients affected by trauma. As a result of the increasing demand, new datasets are being collected to train models based on the emotional response to specific events such as catastrophes, political events, global commemorations, and global strikes (Plaza del Arcoet al., 2020). As an example, the authors of (Glasgow et al., 2016) present a corpus that enables analyzing individuals' sense of gratitude for support in the aftermath of a disaster such as a mass shooting.

During the COVID-19 pandemic, most individuals have been required to comply with health-safety measures such as the use of masks and social distancing practices. Moreover, especially during the acute phases of the pandemic, shelter in place and lockdown orders have been enacted to prevent new outbreaks and to contain the burden on the health system. In addition, the governments of many countries introduced restrictions to international travel as soon as the emergency related to the SARS-CoV-2 virus was declared a pandemic, in March 2020. Such restrictions applied to most individuals during the acute stage of the pandemic, that is, from March to July 2020. Many research studies investigated different non-clinical aspects of the COVID-19 pandemic such as individuals' perception of health-safety measures and novel systems for dealing with the global health crisis (Christen et al., 2021) (Farber et al., 2021) (Clark & Caporusso, 2021) (Niehaus & Caporusso, 2021). In addition, as the consequences of the COVID-19 emergency had major impacts on individuals' mental health, several research groups focused on collecting data that could be utilized to further analyze different aspects of the pandemic using NLP tools. Specifically, posts from social media websites have been utilized to study public opinion and the impact of COVID-19 related restrictions on individuals' mental health (Praveen et al., 2021) (Valdez et al., 2020).

THE CORPUS

The objective of our work is to analyze the consequences of COVID-19 restrictions on bi-national couples that were separated by travel bans and to evaluate the effectiveness of online groups in providing users with peer-to-peer support. To this end, we collected secondary data publicly available on two of the largest Facebook groups dedicated to individuals facing the situation of being apart from their partner due to COVID-19-related restrictions. The first group (i.e., Group A), was created at the end of June 2020 and has over 48000 total members at the date of this study. In the first months of its activity, the group grew at a very fast rate (e.g., with peaks of 2000 new members per day, in June 2020). However, as travel resumed and restrictions were progressively lifted in most countries, after June 2021, the group experienced a lower growth rate (about 100 new members every month) and overall activity (about 50 new posts every month). Group B was also created in June 2020 and currently has over 17500 members. It experienced a similar trend as Group A in terms of membership and activity. Currently, its growth and engagement statistics report 45 new members/week and about 300 posts/month. Group B primarily focuses on the European (EU/UK) and American travel bans, whereas Group A advocates for couples from all nationalities. Both groups have the main goal of tackling issues related to couples separated by travel restrictions, though they also support people separated from children, parents, and other loved ones. The aim of this paper is to share our data in the form of a corpus that can be utilized to realize further studies on long-distance relationships, develop NLP models, and incorporate conversational agents in counseling and support tools.

Methodology

In our data collection, we focused on post content and engagement, only. We did not collect any statistics about users' interaction with the group such as membership demographics or engagement over time, because they are beyond the purpose of our work. Also, this type of information is accessible to group administrators only. Although Facebook's Application Programming Interfaces (API) enable researchers and developers to conveniently extract data from pages and groups, it supports querying the history up to 90 days only, which was not suitable for our research. Also, as Facebook APIs retrieve partial information (e.g., they do not incorporate reactions and comments), we developed a software tool that enabled extracting other publicly available properties for each post, such as the comments and reactions, which could provide the research community with additional insight into user interactions. The tool was designed to retrieve the following information about the post: identifier of the group (i.e., A or B), creation date and time, unique numeric user identifier of the author, text content, permalink (i.e., URL), top comments, and total count of comments, shares, and reactions. As per the latter type of information, we collected the count and type of reactions as specified on Facebook (i.e., like, care, love, anger, sadness, surprise, and laughter). We decided to skip posts that did not have any text. Also, we did not acquire any images or videos, because they could potentially expose the identity of the author or other individuals; furthermore, this type of information was not relevant for our research. The data collection software was designed to avoid acquiring any identifiable information (e.g., the first and last name of the author, as well as their profile picture). Also, we implemented systems for protecting the privacy of the users who authored posts and comments. For instance, we utilized a one-way encryption algorithm (i.e., MD5) to convert the numeric user identifier into a unique hash. Furthermore, we processed the content of posts and comments with algorithms that anonymized the data by removing any personally identifiable information (e.g., names, phone numbers, and emails), and we double-checked the dataset with a human agent because some of the posts incorporated content that could identify the authors or their friends (e.g., emails, phone numbers, website URLs, the first and last name of users tagged within the post, or text mentioning individuals' and groups' names).

Content of the Corpus

A total of 4047 posts were acquired during this research and are reported in this paper. However, data collection is ongoing because COVID-19-related travel restrictions are still in place in some countries, the groups are still active, and users are continuing to publish and comment on posts to share their experience and receive support in dealing with the consequences of long-distance relationships. As per the content acquired during the reported data collection window, Group A and B account for 1721 (42.52%) and 2326 posts (57.47%), respectively. The posts were published by a total of 2802 users, that is, 1414 in group A (53.78%) and 1295 in group B (49.53%). 93 users (3.31%) were members of both groups. On average, each user who generated content published 1.44 posts. In this regard, users in the two groups had a similar behavior (1.14 posts/user in group A and 1.67 in group B).

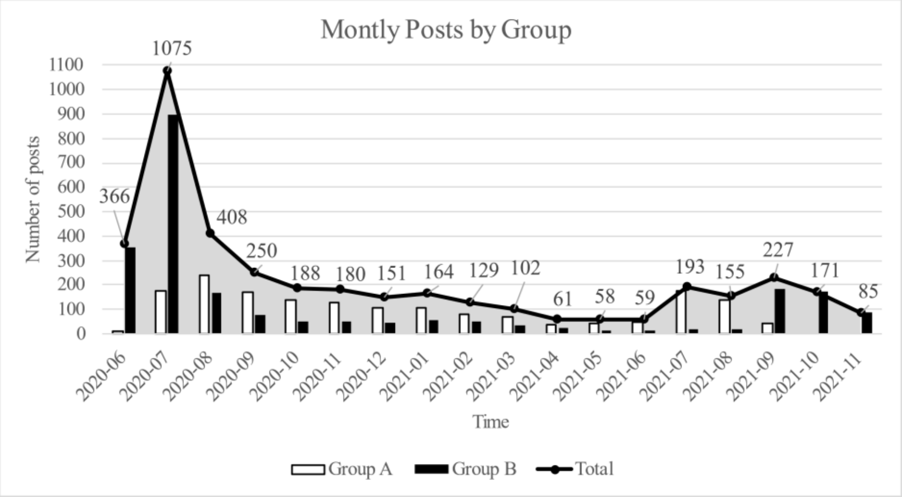

Figure 1. Posts published monthly in Group A and Group B since the creation of the groups and until November 11th, 2021. After an initial peak, the activity of the groups has been steady throughout the pandemic. As a result, our corpus also provides researchers with a dataset that is suitable for longitudinal studies.)

Figure 1 represents the number of posts published in each group from the date on which the group was created until November 11th, 2021. A total of 2618 posts (64.69%) were published in 2020, and 1404 (34.70%) in 2021. Specifically, 977 posts (56.77%) were created in Group A in 2020 and 740 (43%) in 2021. Nevertheless, as shown in Figure 1, Group A experienced a decline in the last three months, with no new posts being created in October and November 2021. As per Group 2, 1641 posts (70.55%) were published in 2020 and 664 (28.54%) in 2021. The trend in the total number of posts shows that users were very active during the first three months, though the overall activity of the groups has been constant throughout the pandemic and the different phases of the health emergency.

The corpus contains the entire text published in each of the posts, which consists of a total of 496736 words divided into 9769 paragraphs (i.e., approximately 123 words and 2.41 paragraphs in each post, on average). Figure 2 depicts a word cloud showing the 100 most frequent words in the content published by the users, which offers a thorough representation of the main themes discussed in the groups. They reference relationships (e.g., love, boyfriend, together), travel restrictions (e.g., visa, passport, documents, and travel) and country-specific travel bans (e.g., us for the United States and uk for the United Kingdom), COVID-19-related topics (e.g., covid, test), and more personal themes related to individuals' feelings (e.g., hope, finally, help, and feel) and to the duration of the separation (e.g., days, months, since, and time). Although barely captured in the word cloud, many posts document successes such as reunions, and marriages, as well as stories of break-ups, divorces, and separation. Indeed, most posts specifically deal with the issues of bi-national couples forced into long-distance relationships by travel restrictions enacted in response to the COVID-19 pandemic. Nevertheless, the content of the corpus provides researchers with material that can be suitable for realizing studies on peer-to-peer support. Furthermore, in addition to the text of the posts, the corpus contains data about the interaction of the members of the groups with the content, which includes 104208 comments and 367685 reactions classified into likes and key emotion labels as defined by Facebook (i.e., love, support, surprise, anger, sadness, and excitement). However, currently, our corpus contains text from the top comments, only. User interaction data, comments, and reactions can be utilized to model the behavior of conversational agents and train intelligent affective systems to show empathy in response to stimuli having a very strong emotional component.

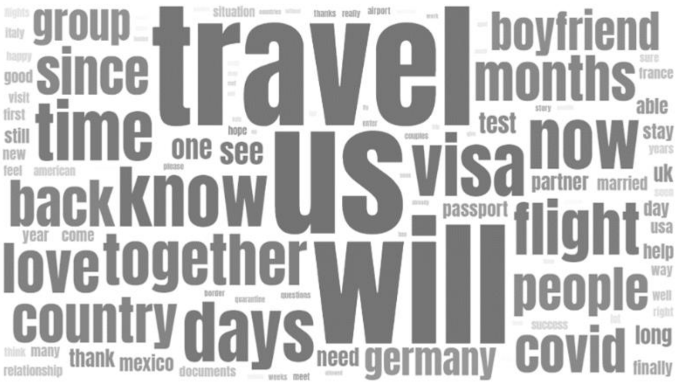

Figure 2. Word cloud summarizing the key topics discussed in the groups.

CONCLUSIONS AND FUTURE WORK

Advancing research and development in the context of conversational agents relies on the availability of NLP corpora that enable fine-tuning general pre-trained models with respect to the requirements of each specific task. This is especially true for applications that deal with domain-specific knowledge or human dynamics such as affection. As a result, there is an increasing demand for specialized corpora containing information that capture emotional aspects. To this end, in this paper, we presented a dataset especially focused on conversations of people forced into long-distance relationships by COVID-19-related travel restrictions. Our corpus includes publicly-available user-generated content posted on Facebook groups dealing with the topic from the creation of the groups (i.e., June 2020) until November 11th 2021, which covers most of the duration of the COVID-19 pandemic. As a result, the dataset offers a representation of the conversations, the interaction dynamics between the members, and the different reactions of couples and partners to being forced into long-distance relationships. Our methodology and the amount of data we collected are comparable with other studies published in the literature (e.g., (Pérez et al., 2021)). The corpus described in this paper and any new data associated with our work and collected after the publication date of this paper are publicly available as a JSON-formatted file at the URL of the GitHub repository of the project (Caporusso, 2021). Also, the repository contains a set of JavaScript utilities developed for NodeJS for analyzing the data and converting them into other formats such as comma-separated values (i.e., csv). The data published in the repository do not include the post permalink to protect the privacy of the authors.

In our future work, we will extend our corpus by incorporating more posts from the groups featured in this paper and updating the GitHub repository. Furthermore, we will expand the data about user interaction by acquiring additional data, such as comments and comment interaction: in addition to resulting in a richer and more detailed corpus, additional user interaction data may be useful to other applications such as conversational agents and text generation tasks (e.g., question and answer). Also, we will expand our work to other relevant platforms and groups, such as Discord channels and forums that discuss the topics of interest for our research. Initially, we did not include them in our work because of the nature of the interaction on these channels, that is, either short messages having a conversational style or more informative and articulated content.

REFERENCES

Booher, Harold, ed. (2003). Handbook of human systems integration. New Jersey: Wiley.

Booher, H.R., Minninger, J. (2003) “Human systems integration in army systems acquisition”, in: Handbook of human systems integration, Booher, Harold (Ed.). pp. 663-698

Pháp, H.C., 2016. Solutions of creating large data resources in natural language processing. In Recent Developments in Intelligent Information and Database Systems. Springer, pp. 243–253.

Ezzini, S. et al., 2021. Using domain-specific corpora for improved handling of ambiguity in requirements. In 2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE). IEEE, pp. 1485–1497.

Gu, Y. et al., 2021. Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH), 3(1), pp.1–23.

Edwards, A. et al., 2020. Go simple and pre-train on domain-specific corpora: On the role of training data for text classification. In Proceedings of the 28th International Conference on Computational Linguistics. pp. 5522–5529.

Manohar, M.Y. & Kulkarni, P., 2017. Improvement sarcasm analysis using NLP and corpus based approach. In 2017 International Conference on Intelligent Computing and Control Systems (ICICCS). IEEE, pp. 618–622.

Pérez, J.M., Giudici, J.C. & Luque, F., 2021. pysentimiento: A Python Toolkit for Sentiment Analysis and SocialNLP tasks,

Afshar, M. et al., 2019. Natural language processing and machine learning to identify alcohol misuse from the electronic health record in trauma patients: development and internal validation. Journal of the American Medical Informatics Association, 26(3), pp.254–261.

Plaza del Arco, F.M. et al., 2020. EmoEvent: A Multilingual Emotion Corpus based on different Events. In Proceedings of the 12th Language Resources and Evaluation Conference. Marseille, France: European Language Resources Association, pp. 1492–1498. Available at: https://aclanthology.org/2020.lrec-1.186.

Glasgow, K. et al., 2016. Grieving in the 21st Century: Social Media's Role in Facilitating Supportive Exchanges Following Community-Level Traumatic Events. In Proceedings of the 7th 2016 International Conference on Social Media & Society. pp. 1–10.

Christen, L., Farber, T. & Caporusso, N., 2021. Face masks as awareness and engagement platforms. In International Conference on Applied Human Factors and Ergonomics. Springer, pp. 617–624.

Farber, T., Christen, L. & Caporusso, N., 2021. Evaluating Users' Perception of Health-Safety Measures Against Pandemics. In International Conference on Applied Human Factors and Ergonomics. Springer, pp. 633–639.

Clark, J. & Caporusso, N., 2021. A dedicated platform for health-safety reviews. In International Conference on Applied Human Factors and Ergonomics. Springer, pp. 625–632.

Niehaus, J. & Caporusso, N., 2021. An Infrastructure for Integrated Temperature Monitoring and Social Tracking. In 2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO). IEEE.

Praveen, S., Ittamalla, R. & Deepak, G., 2021. Analyzing Indian general public's perspective on anxiety, stress and trauma during Covid-19-a machine learning study of 840,000 tweets. Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 15(3), pp.667–671.

Valdez, D. et al., 2020. Social media insights into US mental health during the COVID-19 pandemic: longitudinal analysis of twitter data. Journal of medical Internet research, 22(12), p.e21418.

Caporusso, N., 2021. Long-Distance Relationship NLP Corpus, GitHub. Available at: https://github.com/NicholasCaporusso/COVID-19-Long-Distance-Relationships-NLP-Corpus.git.