Online Dating Dialogue Annotation Task for Sentiment Analysis

Troy Kelley, Hana La, Nidhi Sinha
New York University
New York City, New York
tk2208@nyu.edu, hl3063@nyu.edu, ns4143@nyu.edu

1. Abstract

Although online dating platforms are increasing in impact on the current dating scene, there is no existing corpus which focuses on this type of online communication. This paper will describe how we created an annotated dataset for online dating conversations sourced from popular Reddit page r/Tinder, and our peers in order to allow for efficient sentiment analysis on the conversation. Each line in the conversation is tagged with dialogue acts from Stolcke et. al., as well as our own tags specialized for the nature of the corpus. We then tagged each line based on positivity and optional humour, sexual, and other tags in order to accurately describe sentiment and the overall outcome of each exchange.

2. Introduction

Compared to other online communication platforms, applications and websites designed for conversing specifically for the purpose of dating are relatively new, but have a huge cultural and social impact on the present-day dating scene, where dating app usage has greatly increased in the younger generation, and their popularity continues to grow (Anderson 2020). Despite online dating's growing prevalence, there is a lack of publicly available datasets focusing on dialogue from dating apps. We have therefore created a small but high quality annotated dataset for online dating conversations. This dataset can be used for further research on dating app conversations and their conclusions through the use of sentiment analysis. Researchers can investigate the cause and effect of pickup lines, explicit content, and related messages to see how these various types of dialogue impact the overall result. What is important about having a corpus specific to online dating conversations is that there are features unlike regular social messaging services, with a much greater importance placed on the first exchange. Depending on the opening line, the whole tone and success of the conversation is affected, where the recipient has no obligation to continue the conversation if their interest is not held. We will also analyze how our annotation tagging changed depending on the levels of inter-annotator agreement for each tag and the issues that occurred.

Using standardized annotation guidelines and tags, we have created a dataset that can be trusted for use in sentiment analysis, chat bots and other analyses. Approximately 100 conversations were sourced from the popular Reddit page "reddit.com/r/Tinder", as well as from our peers who use dating apps. A subset of conversations were cross-annotated by two members of our team to compare results and ensure we all used consistent tagging, with good inter-annotator agreement for each tag. All identifiable information has been redacted from the conversations in order to maintain privacy. We will also discuss the possible ramifications of a publicly available dating conversation corpus, and how we would go about publishing a corpus with sensitive information like the data with which we are dealing.

3. Method and Approach

The first step in the annotation process was to decide on which dating application to obtain conversations from. For the sake of consistency, we decided to get all exchanges from the app Tinder which had the greatest number of registered users in the United States according to Statista. A majority of the exchanges were obtained from various public online forums and websites, such as Reddit, a social forum website, or Imgur, an image-sharing social platform. A portion was also obtained from an annotator's own exchanges as well as from peers with permission.

3.1 Dialogue Formatting

In order to ensure that our transcriptions were consistent in format as well as maintaining a certain level of privacy, all lines are labelled A, the initiator of the conversation, or B, the recipient. Changes done to the conversations in the process of transcription are rewritten in all capital letters. Any sensitive information, such as phone numbers or social media usernames, are replaced with "REDACTED" in conversation. Considering the prevalence of emojis in online conversations, we have also decided to transcribe emojis according to their official name from the unicode emoji list after "EMOJI" (i.e. EMOJI_grinning_face).

For this annotation task we also decided to focus on the first and last exchanges, since a main feature of online dating conversations is the importance of the opening line and its effect on the overall success of the conversation, where success is an implied connection between the two participants. For this purpose, during transcription of longer conversations, the A/B exchanges after the first and before the last were removed. In the cases where the initiator A starts an exchange and the recipient B does not complete their turn in the conversation, B's line is replaced with "NO_RESPONSE" and the sentiment is negative as a result.

3.2 Dialogue Labels

Each line of each conversation in the corpus is tagged with one of the following dialogue acts from the dialogue act tagging system described in Stolcke et. al., or one of our custom dialogue tags. These custom tags were created after gathering the first thirty conversations, where prevalent occurrences of unique features in the conversation were not represented in the original set of tags. This list includes creative openings, conventional closings, pick-up lines, and text corrections (i.e. "A: there*"). In the corpus, their respective tags are represented as UNCONVENTIONAL-OPENING, CONVENTIONAL-CLOSING, PICKUP-LINE, and CORRECTION.

3.2 Sentiment Labelling

After each line in every conversation is labelled with a dialogue act, the annotators then tag the line's overall sentiment. After analyzing the first conversations we obtained, sentiment tags were created specific to the online dating corpus. The tags each fall under one of five categories: humor/serious, positivity, sexual intent, unsolicited information, and speaker or recipient focus. The speaker or recipient focus represents if the line is about the recipient or the speaker, or relating to both. Other than the dialogue act and positivity, all other tags are optional,

where a line can have one, many, all, or no extra sentiments. If a sentiment is not annotated explicitly, the line is considered to not have that sentiment and is null during evaluation. Each tag is also mutually exclusive within their categories. The following is a list of sentiment tags in each category that appear in the corpus (Figure 3.2.1.)

Positivity	**Humorous/Serious**	**Speaker or Recipient Focus**	**Sexual-Intent-Related**	**Unsolicited-Information**
positive negative neutral	humorous serious	self other both	sexual-intent sexual-decline sexual-response	unsolicited-information unsolicited-information-response

Figure 3.2.1

As an example tagged sentence, see Figure 3.2.1, where each tag is tab separated for easy parsing. These tags, rather than just generic positive, neutral, and negative tags provide additional information specific to online dating.

```
A: If you were a potato, you'd be a damn fine potato
PICKUP-LINE    humorous  positive  other
B: Ooo that's a good one. I have one too     STATEMENT
positive  self
B: If you were my girlfriend, my parents would finally be
proud of me    PICKUP-LINE    humorous  positive  both
CONV neutral
```

Figure 3.2.1

3.3 Conversation Result

As seen in the example annotation, the end of each conversation as an additional tag, where the annotator grades the overall success of the conversation, where explicit or implied connection and future contact between A and B is expressed is a success, outright or implied rejection, including no responses, are considered failures, and ambiguity is neutral.

3.4 Evaluation

By the nature of the task being done - building a corpus of transcribed and tagged conversations, precision, recall, accuracy and similar metrics cannot be computed since there is no ground-truth answer key to compare our results against. This task aims to build a corpus that is consistent and reliable in its construction.

To measure our consistency, we used the Cohen Kappa score for inter-annotator agreement. This is a binary scoring system which ranges from [1.0, -1.0]. A higher score corresponds with higher agreement and a lower score corresponds to less agreement. A negative score is almost rarely observed but it typically means the annotators agreement is worse than random (McHugh, 2012) Since our tags consist of multiple options within each category (e.g.

positivity, subject), we scored our annotations using the presence of a tag across the same annotated sentences/conversations. If a given tag category has the same value across annotations, then they are said to be in agreement. In this way a kappa score is calculated for every tag value and conversation result.

4. Results and Discussion

4.1 Results

At the time of writing, the team has created a corpus of approximately 100 conversations. Once 30 conversations were tagged, the annotations were scored by using the Cohen kappa score for inter-annotator agreement. By using Cohen's guidelines seen in Figure 4.1.1 (McHugh, 2012), we received a moderate score of 0.58. The team then discussed various discrepancies between their annotations and how tags could be more uniformly used in future annotations. This same dataset was re-tagged by both annotators on the team and the resulting average kappa score was 0.70, which is deemed as substantial agreement. This led the team to believe that we reached an acceptable amount of agreement that could be compiled into one corpus without needing sentence to be tagged twice - once by each annotator. The final corpus is an accurately annotated dataset, despite having two contributors.

Score Range	Agreement
-1.0 - 0.0	Less than chance
0.01 - 0.20	Slight
0.21 - 0.40	Fair
0.41 - 0.60	Moderate
0.61 - 0.80	Substantial
0.81 - 1.0	Almost Perfect

Figure 4.1.1

In this second scoring, the average score for conversation result tags was 0.24. While the score for the conversation "results" category is relatively low, it still represents fair agreement. This also shows the ambiguity which lies in a conversation's meaning - there is never just one way to interpret a message.

The average score for sentiment tags was 0.54. This shows that it is easier to glean meaning and categorize the sentiment sentence-by-sentence rather than viewing the dialogue as a whole. Most notably the positivity category performed the best within sentiment tags, which received scores of 0.75, 0.88, and 0.64 for "positive", "negative", and "neutral" respectively.

Overall, these results confirm our initial hypothesis that enough agreement had been reached between annotators. See Figure 5.1.2 for the scores of all tags and conversation results.

Positivity Tag	Kappa Score	Number of Occurrences
positive	0.75	124
negative	0.88	27
neutral	0.64	75
average	0.75	N/A

Humor Tag	Kappa Score	Number of Occurrences
humorous	0.68	46
serious	0.34	11
average	0.51	N/A

Subject Tag	Kappa Score	Number of Occurrences
self	0.65	40
other	0.56	93
both	0.64	15
average	0.79	N/A

Sexual Tag	Kappa Score	Number of Occurrences
sexual-intent	0.76	28
sexual-response	0.60	16
sexual-decline	0.00	1
average	0.45	N/A

Information Tag	Kappa Score	Number of Occurrences
unsolicited-information	0.38	10
unsolicited-information-response	0.00	1
average	0.45	N/A

Conversation Result Tag	Kappa Score	Number of Occurrences
success	0.16	14
failure	0.39	26
neutral	0.18	13
average	0.24	N/A

Sentence Action Tag	Kappa Score	Sentence Action Tag	Kappa Score
STATEMENT	0.79	OPEN-QUESTION	0.55
BACKCHANNEL/ACKNOWLEDGE	0.00	RHETORICAL-QUESTION	0.00
OPINION	0.85	HOLD-BEFORE-ANSWER/AGREEMENT*	1.00
ABANDONED/UNINTERPRETABLE*	1.00	REJECT	0.79
AGREEMENT/ACCEPT	0.66	NEGATIVE-NON-NO-ANSWERS	0.66
APPRECIATION	0.00	SIGNAL-NON-UNDERSTANDING*	1.00
YES-NO-QUESTION	0.85	OTHER-ANSWERS*	1.00
YES-ANSWERS*	1.0	CONVENTIONAL-OPENING	0.80
CONVENTIONAL-CLOSING	1.0	OR-CLAUSE	1.0
WH-QUESTION	0.32	DISPREFERRED-ANSWERS*	1.0
NO-ANSWERS	0.00	3RD-PARTY-TALK*	1.0
RESPONSE-ACKNOWLEDGMENT	0.00	OFFERS-OPTIONS-COMMITS*	1.0
HEDGE*	1.00	SELF-TALK*	1.0
DECLARATIVE-YES-NO-QUESTION*	1.00	DOWNPLAYER	1.0

OTHER	0.76	MAYBE/ACCEPT-PART*	1.0
BACKCHANNEL-QUESTION	0.00	TAG-QUESTION*	1.0
QUOTATION*	1.00	DECLARATIVE-WH-QUESTION*	1.0
SUMMARIZE/REFORMULATE	1.00	APOLOGY*	1.0
AFFIRMATIVE-NON-YES-ANSWERS	1.00	THANKING*	1.0
ACTION-DIRECTIVE	0.49	UNCONVENTIONAL-OPENING	0.55
COLLABORATIVE-COMPLETION*	1.00	PICKUP-LINE	0.72
REPEAT-PHRASE	1.00	CORRECTION	1.00
Average		0.72	

Key: * denotes a tag not occurring in the dataset

Figure 4.1.2

Action Tag	**Number of Occurrences**
STATEMENT	46
CONVENTIONAL-OPENING	34
PICKUP-LINE	19
OPINION	14
OPEN-QUESTION	14
UNCONVENTIONAL-OPENING	14

Most Frequent Sentence Action Tags

Figure 4.1.3

4.2 Discussion

Some of the most varying results occur in the sentence action tags, where 31 out of 44 tags had a score of either 1.0 or 0.0. For the 25 tags receiving a score of 1.0, this is largely because the actions they are describing simply do not occur very often in instant messaging. For example tags like “COLLABORATIVE-COMPLETION”, which describes a sentence started by

one party and completed by the other, would simply never occur digitally since all messages are typed out before they are sent. If these types of tags were to be omitted, then the average kappa score drops to 0.63 which is still in the "substantial" agreement category, but is a 0.9 difference from the previous average. This shows that within the realm of more frequent sentence actions, it is harder for annotators to pick between various possible tags. For the tags with a score of 0, these typically consist of actions which could easily be categorized in a different way. "RHETORICAL-QUESTION" is a good example of this, since an annotator could easily misinterpret cues to label a sentence one way or another. Figure 4.2.1 shows an example of a conversation where the last sentence could be ambiguous as a rhetorical question or a serious yes/no question.

```
A: REDACTED_IDENTITY you are tooooo damn yummy baby
A: EMOJI_PEACH EMOJI_sweat_droplets EMOJI_tongue
B: Lmao
B: Is that how we're gonna start, REDACTED_IDENTITY?
```

Figure 4.2.1

Notably the tags "PICKUP-LINE", and "CONVENTIONAL-OPENING", both received high agreement scores of 0.72 & 0.80, and were in the top 3 most occuring action tags. Conventional openings belong to a predefined set of colloquial sayings agreed upon by society as normal ways to start a conversation. Inversely, a pickup line is a rather novel and possibly egregious greeting to another person. However they still typically follow some sort of pattern with a "play on words", or clever use of language in an alluring way. These types of actions are very easy for annotators to spot and identify as such. With high scores and frequency, our methods show high promise for tagging the common and arguably important utterances in a dialogue.

Some roadblocks that the team encountered were relating to the manual entry of all the annotations. This caused tags to be occasionally misspelled, and the format of the annotations to be broken. In turn this made more work for us to go back and correct these errors before results could be scored or made into one corpus. In the future, the use of some widely available tagging programs would be of great assistance to both improve the accuracy of our transcriptions and speed of input. A barrier to this method is the rather specialized transcription and annotation format that would need modifications to fit a predefined annotation standard.

Alternatively, a simple command line script could be developed by the authors to easily provide an annotation tool to annotators. This would wholly eliminate any human error in the actual creation of the corpus. Furthermore, creating or organizing an efficient data pipeline could greatly increase the ability to create a larger dataset. The most time consuming part of creating annotations is the finding of conversations, and the transcription from images to typed text. Lastly, the actual tagging of action, sentiment, and result tags consumes the least time. A theoretical workflow would be to assign different people to these separate tasks where each member completes their task using the output of the previous member's task. Even further improvements would include using a text-extraction tool for images which would eliminate the need to manually transcribe sentences with a quick review of the output to ensure correctness.

An important point of note is the potential bias of using conversations primarily sourced from Reddit, and similar online forums. Because there is no way to verify the validity of the

content in a post, some conversations may be fake or not authentic to what a user would actually say in a real conversation. The utterances in conversations may be said strictly for the purpose of comedic effect. Similarly, users on online forums are less likely to post a conversation that is not distinct in some way. Therefore many conversations may lean to be more humorous, sexual, or unique in nature. Some conversations were sourced directly from acquaintances of the team members and in this data there were much fewer sentences categorized as humorous, pickup-lines, or unsolicited-information.

The team ultimately envisions this dataset to be a resource for tasks focused on research within the field of online-dating conversations, since currently there is no widely available corpus for this type of data. This would enable tasks like sentiment analysis to be more accurately completed. It would also allow for easier modelling and prediction of opening and closing remarks used in these types of conversations.

In order to prevent this dataset being used for possible malicious intent, the team took two measures of redacting identity and eliminating dialogue between the first and last exchanges. While this helps in that regard, it is also a tradeoff between the amount of data in the corpus and user anonymity. This lack of entire conversation data also allows for quicker annotation, allowing more of them to be completed, yet the data is less robust. The shortcomings of this method is that it could not be used to model entire conversations and what exactly caused the greeting state to lead to the ending state.

For further research, this data could be used in tandem with another dataset and integrated into a larger system to better inform a task. Even though the corpus is small, it provides relevant data that can be used in systems in the future. This project is a stepping stone for more work to be done using it and analysis done therein. These annotations provide an important insight into how humans communicate romantically and sexually through online messaging,

4.3 Publishing

After researching into the possibility of publishing this corpus, we discovered that there would be possible ramifications for dealing with a corpus that includes sensitive information. Although all of the data has been transcribed in such a way as to anonymize the participant's identity, there still might be methods of malicious attacks using this corpus. A portion of our data is from peers and an annotator's own exchanges on the Tinder application, which, according to IRB guidelines (U.S Dept. of Health and Human Services, 2021), we would need to have provided a consent form informing the contributors of such risks. We could hypothetically publish our corpus without review if all data was sourced from completely public online forums, but we did not separate the conversations by how they were obtained. If we had initially identified and separated all conversations from offline sources, we could then publish only the conversations which were readily available online. To publish the full dataset, we could have gone through proper IRB procedures at the beginning of data gathering and transcription. Unfortunately, we did not have the resources to perform these steps and the corpus remains private. However, we have proceeded with this annotation task as if we had done so.

5 Conclusion

The transcription and annotation work completed is certainly only the beginning of a potentially much larger task. This task outlines the initial steps needed to gather adequate data , create an annotation standard & guidelines, and finally build a usable corpus for future work. Based on a satisfactory inter-annotator agreement score, our data from multiple annotators is deemed to be acceptable to be used together as one corpus. The tags used in conjunction with the transcription of the dialogues includes both standard conversational action tags used in similar tasks in the field, as well new tags which describe overall meaning in better detail. We believe the additional tagging of sentence actions and sentiment to be important relevant information which will allow for deeper analysis and understanding, rather than simply creating a dataset of transcribed dating conversations. We look forward to how other members of the community build upon and use this data in the future.

References

Anderson, Monica, Emily A. Vogels, and Erica Turner. 2020. The Virtues and Downsides of Online Dating. Pew Research Center.

Ivanovic, Edward. 2005. Dialogue Act Tagging for Instant Messaging Chat Sessions. *Association for Computational Linguistics.*

McHugh, Mary L. Interrater reliability: the kappa statistic. 2012. *Biochemia Medica*.

Napoles, Courtney, Joel Tetreault, Enrica Rosato, Brian Provenzale & Aasish Pappu. 2017. Finding Good Conversations Online: The Yahoo News Annotated Comments Corpus. *Proceedings of the 11th Linguistic Annotation Workshop*, pages 13–23. *Association for Computational Linguistics.*

Olshefski, Emily Grace. 2015. Game-Changing Event Definition and Detection in an eSports Corpus. *Proceedings of the 3rd Workshop on EVENTS at the NAACL-HLT* 2015, pages 77–81. *Association for Computational Linguistics.*

Rosenberg, Andrew and Ed Binkowski. 2004. Augmenting the kappa statistic to determine Interannotator reliability for multiply labeled data points. *Association for Computational Linguistics.*

Stolcke, Andreas, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *Association for Computational Linguistics.*

UCI Office of Research. Guidance for Reviewing Protocols that Include Online Sources or Mobile Devices. Web. 14 April 2021.

U.S. Department of Health and Human Services, U.S Food and Drug Administration. Institutional Review Board (IRB) Written Procedures: Guidance for Institutions and IRBs. Web. 14 April 2021.