Identifying Intent-Sentiment Co-reference from Legal Utterances

Pinaki Karkun
Department of Computer Sc. & Engineering
Jadavpur University
Pinaki.Karkun@gmail.com

Dipankar Das
Department of Computer Sc. & Engineering
Jadavpur University
dipankar.dipnil2005@gmail.com

Abstract

Co-reference is always treated as one of challenging tasks under natural language processing and has been explored only in the domain of anaphora resolution to an extent. However, the benefit of it to identify the relations between multiple entities in a single context can be explored better while we aim to identify intent and sentiment from the utterances of a dialogue or conversation. The utilization of co-reference becomes more elegant while tracking users' intents with respect to their corresponding sentiments explored in a specialized domain like judiciary. Thus, in the present attempt, we have identified not only intent and sentiment expressions at token level in an individual manner, we also classified the utterances and identified the co-reference between intent and sentiment entities in utterance level context. Last but not the least, the deep learning algorithms have shown improvements over traditional machine learning in all cases.

1 Introduction

Discourse, a one-way dialogue where both parties participate and one of the main purposes of a cooperative, two-way conversation is to transfer knowledge from the speaker/writer to the listeners/readers. Discourse study [1] in general aims to answer two types of questions: (1) *What information is provided in the extended sequences of utterances that are not contained in the meaning of the individual utterances?* (2) *What effect does the context in which an utterance is used have on the meaning of individual utterances or parts of them?*

One of the reasons for considering discourse and dialogue rather than just the sentences is because information is sometimes presented or requested over multiple sentences, and we want to recognise various phrases or relationships among them that identify the *who, what, when, where*, and *why* of the event. We could take material from articles in newspapers and magazines, as well as chapters from books, and save it in tables that are more easily searchable for dialogue. We might want to extract information from dialogue to complete activities like booking travel plan or making a restaurant reservation, or teaching a kid etc. Alternatively, we might wish to be able to detect explicit or implicit demands in purely social exchanges that have no clear purpose. Intent or sentiment identification in an individual manner may not always contribute whereas their relationship in terms of co-reference actually helps in tracking these entities during the discourse trail, from start to the end.

Intents are the tags that we can assign to words or phrases in a dialogue or discourse dataset. It differs from topic identification in general as topic identification itself is a separate research area (Styoanov and Cardie, 2008) (Passonneau, 2004). However, it identifies what fine-grained topics/aspects are being discussed in utterances or in dialogues. Intents are subjective to the conversation topic; there can be unique sets of intents for different conversations. Let's take two sentences from a chat as follows,

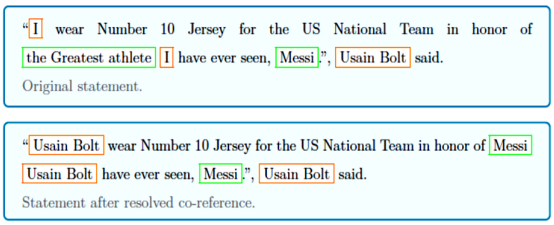
1. I am hungry.
2. I have not eaten anything since morning.

When addressing traditional text classification issues, these two statements have different syntactic meanings, which are treated as facts. But, if we try to see things from a different

[1] Barbara Grosz, Discourse and Dialogue, chapter 6, Harvard University, Cambridge, Massachusetts, USA

perspective (intent of the sentences), sentence 1 clearly indicates that the individual is 'hungry'. Despite the fact that sentence 2 does not contain the word 'hungry', we may grasp the person's 'intent' that he or she is hungry at that point because it is considered as "not eating anything".

Thus, one of the prime objectives of co-reference resolution (CR) is to locate all linguistic expressions (called mentions) that relate to the same real-world thing in a given text. We can fix these mentions by replacing pronouns with noun phrases.



In the present context, it is the analysis of polarity towards different types of emotions. Sentiment analysis is quickly becoming a crucial tool for monitoring and understanding sentiment in all forms of data, as humans communicate their thoughts and feelings more openly than ever before. Brands can learn what makes customers happy or frustrated by automatically evaluating customer feedback, such as comments in survey replies and social media dialogues. This allows them to customise products and services to match their customers' demands.

Using sentiment analysis to evaluate the responses of users using any chat-bot is important to provide necessary responses and to control the flow of conversation by handling the emotions of users such a way a fruit-full result comes out of the conversation.

Co-reference technique is a terrific method to get unambiguous statements that computers can understand much more readily. Text interpretation, information extraction, machine translation, sentiment analysis, and document summarising are just a few of the NLP activities that can benefit from co-reference resolution. However, all such topics have not been yet implemented on judicial dataset. One of the reasons may be the judicial dataset has its unique issues to address, mainly- it contains very low amount of positive statements, and most of the utterances are facts. Therefore, the number of neutral statements tends to be more than others. It is also very important for a chat-bot or an agent to track the sentiments of the users while having a conversation with them. When a conversation screws to some extreme emotion, the conversation becomes very inefficient. So it is an important challenge for the chat-bot to have clear connection to which intent can lead to those polarising emotions.

Apart from sentiment or emotion, a chat-bot is mainly built for judicial assistance and possesses some inherited problems like the intents of the users which are convoluted and overlapped. So it is hard to identify intents based on the words spoken by the user. The context, phrases changes the intents entirely.

Our objective in this present work is to develop models based on judicial conversations; one model identifies the intents whereas another one captures the sentiment from utterances. The models also extract tokens which determine the corresponding intent or sentiment of the utterances at word / phrase level. Finally, we build a co-reference model to identify the relation between intent and sentiment if any.

We have employed several machine learning and deep learning techniques to classify and identify possible intents that are appeared in the dialogues of users and bot. We also developed similar models to identify and classify the sentiments too. Finally, we then developed deep learning base model to identify and classify if any utterance has intent(s) that is/are potentially co-referred to the sentiment(s) of the utterance. The deep learning models have shown satisfactory results in terms of F1-score.

The rest of the paper is organized as follows. Related work on this particular topic is discussed in Section 2 whereas Section 3 briefly shows the insights of the datasets. Section 4 describes the method we used to detect the intent of the utterance whereas sentiment identification models are discussed in Section 5. The models of intent-sentiment co-reference along with their experiments and results are dictated in Section 6. Finally, in Section 7, we present the conclusions and briefly discuss about future work.

2 Related Work

Intent Identification: The authors (Xu and Sarikaya, 2013) employed CNN followed by triangular CRF to forecast intent and identify an entity using extracted characteristics from CNN. On the other hand, Kundu and Choudhury (2017) proposed Elman-type, Jordan-type, and

bi-directional Jordan-type RNN followed by basic CRF and later Peng *et al.* (2014) employed a modified deep LSTM, followed by CRF to explicitly represent the dependencies between semantic labels for greater understanding. The intent identification was further explored by using bi-directional RNN, bidirectional RNN with attention mechanism, and Encoder-Decoder LSTM by several researchers (Louvan and Magnini, 2020) (Su *et al.*, 2016).

In (Qin *et al.*, 2019), a self-attentive shared encoder has been used to produce better context-aware representations which applied as the extracted and summarized features for IC at sentence and the token level. Recently in (Chen *et al.*, 2019), authors have used a pre-trained BERT model and a fine-tuned BERT model for IC and joint IC - slot filling tasks, respectively.

Sentiment Identification: As per traditional developments, the activities have shown that in hotel reviews, Kasper and Vela, (2011) presented a "Web Based Opinion Mining method". The research described a system for evaluating online user reviews and comments in order to aid quality control in hotel management systems.

Examining sentiments and opinions can be done in several ways. Some of the researchers investigated public opinion and blogs in (Das and Bandyopadhyay, 2009). In context of their individual task, it divides prior labour into two groups (sentiment analysis for news and blogs).

Furthermore, other important research divides similar work into two categories (Wiseman *et al.*, 2016): the first focuses on detecting term direction, while the second focuses on detecting term subjectivity. These distinctions only apply to term/word level classification research studies, not document level categorization.

The prior research on sentiment-based categorization of input documents implicated either by the use of polarity-popularity models (Das *et al.*, 2020) or the manual or semi-manual creation of discriminant-word lexicons proposed by (Mondal and Das, 2021).

Co-reference Identification: Till date, several ways have been adopted in order to deal with the problem of co-reference in texts. However, in contrast to the earlier approaches, the present approach not only identifies the relations among intent and sentiment entities but also tries to provide the future steps to track such important components in the discourse of a discussion at the pragmatic level.

Rule Based: The task of co-reference resolution in NLP has long been regarded as the crucial task that inevitably relies on certain hand-crafted rules. These principles are based on the syntactic and semantic characteristics of the text at hand. A constant point of contention has been which elements aid resolution and which do not. There have also been studies particularly designed to address this issue (Bengtson and Roth, 2008). While most early Anaphora Resolution (AR) and Co-reference Resolution (CR) algorithms relied on a complex set of hand-crafted rules (often knowledge expensive) and so were knowledge intensive, others attempted to reduce this dependency. One of the first algorithms to deal with AR was Hobb's naive algorithm (Hobb, 1978). To find an antecedent, the algorithm used a rule-based, left-to-right breadth-first traversal of the syntactic parse tree of a sentence. Hobb's approach also used selectional constraints based on world knowledge for antecedent removal. Despite the fact that the majority of rule-based algorithms were knowledge-rich, some of the researchers (Haghighi and Klein, 2009) tried to reduce the level of dependency of rules on external knowledge. The "knowledge-poor algorithms" labelled as such CogNIAC was a resource-constrained high-precision co-reference resolver. This early strategy got us closer to understand how humans resolve references.

When there is no access to extensive training data in the desired target scheme, the change in CR research from rule-based systems to deep learning systems has resulted in a loss of the ability of CR systems to adapt the varied co-reference phenomena and boundary definitions. Dependency syntax was also used as input in a recent rule-based algorithm. It attempted to target co-reference kinds such as cataphora, compound modifier, i-within-i, and others that were not annotated by the CoNLL 2012 shared task. This method, known as Xrenner, was tested on two quite distinct corpora, the GUM corpus and the WSJ corpus.

Statistical and ML Based: During the late 1990s, the area of co-reference resolution shifted from heuristic and rule based methods to learning-based methods. The availability of labelled co-reference corpora like as MUC and ACE was a major factor in this transition. From linguists to machine learning, research community of CR has grown. The mention-pair,

entity-mention, and ranking models are three types of learning-based co-reference models. Soon et al., heuristic's mention creation approach that was the most widely used algorithm for creating mention instances (Soon *et al.*, 2001).

Despite its immense popularity, the mention rankers were unable to effectively use earlier decisions to make current decisions. This prompted the development of "cluster ranking" algorithms. The greatest aspects of entity-mention and ranking models were combined in cluster ranking approaches. In recent deep learning models for CR, a mix of mention ranker and cluster ranker has been used (Clark and Manning, 2016a). In addition, the mention-ranking model did not distinguish between anaphoric and non-anaphoric NPs. The difficulty is addressed by recent deep learning-based mention ranking systems (Clark and Manning, 2016b) which teach anaphoric senses alongside mention ranking.

Deep Learning Based: Despite depending on few characteristics of machine learning (Ng and Cardie, 2002), the current state-of-the-art model is an end-to-end CR system that outperforms prior techniques. This end-to-end neural model (Lee *et al.*, 2017) mentions detection and CR in the same sentence.

3 Dataset Preparation

3.1 Data Crawling

Initially, the raw data in the legal forum [2] was present in the form of a series of user-posted legal concerns and their possible legal advice from Indian legal professionals. It was then turned into a conversational format for usage by a user and a trained conversational agent. A total of 350 distinct legal cases were collected. Each of the raw examples is then processed further to get it into a conversational format. Table 1 shows a sample of the scrapped corpus in its raw form.

3.1 Annotation Guidelines

The crawled data contains information in a set of question and answers from a user and multiple lawyers. Each of the unprocessed legal cases was turned into a series of legal talks between a client and an automated agent. We noticed that the majority of the raw cases lacked information, which could lead to incorrect solutions. As a result, one of the automated agent's key responsibilities would be to ask relevant questions depending on the given scenario in order to extract missing critical pieces of information via question answering. Thus, in order to accomplish the goal, diverse legal advice from different specialists was studied to identify the events' direction and chronology. Each incident is thereafter represented as a pair of issues with related legal advice. Finally, the informative ones are picked and various conversation flows based on the raw text were prepared.

Speaker	Statement
User	Q: My husband has been physically harassing me for years. Need help.
legal expert 1	1. You can apply for divorce
legal expert 2	1. File for divorce 2. Apply for maintenance.
legal expert 3	1. Make a police complaint 2. Shift to your maternal home 3. Send him legal notices 4. If you want a divorce then file for it

Table 1: Sample of scrapped corpus

We focus on understanding the user's intention as well as the gravity of the situation. For the very reason, the utterance of each user has been annotated with two attributes, intent name and sentiment score within the range of -5 to +5. In case of the utterances of a bot, the annotation has been limited with intent name only. Figure 1 represents the hierarchical structure of the intent tree. The intents are categorized as a hierarchical tree to show how they are linked and how different intents determine each other. The hierarchy of intents looks like this:

Statistical analysis on Table 2 signifies that the sentiment of the legal corpus is negatively biased in general which in turn validates the psychological state of worried clients. A total number of 1440 user utterances have been taken into consideration. The data can be visualized by a normal distribution of N (-0.5211, 1.3479). Figure 2 represents the details of sentiment distribution.

[2] https://indiankanoon.org/search/?formInput=forums

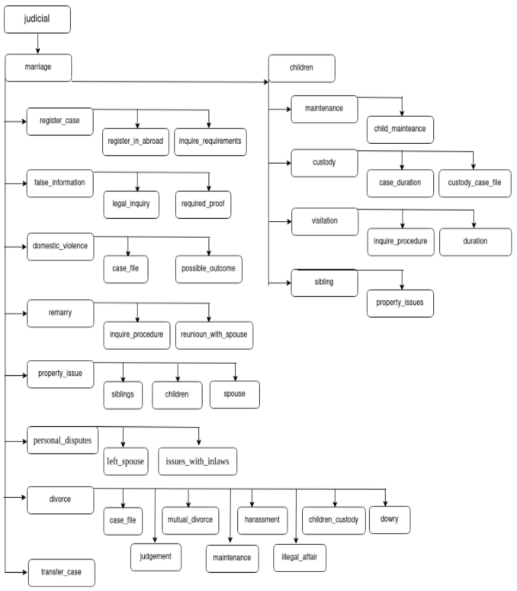


Figure 1: The hierarchical structure of the intent tree

Feature name	Value
Maximum length of utterance by user (in character)	242
Minimum length of utterance by user (in character)	1
Average length of utterance by user (in character)	24.94
Total number of user utterances	1440
Maximum length of utterance by bot (in character)	220
Minimum length of utterance by bot (in character)	1
Average length of utterance by bot (in character)	19.58
Total number of bot utterances	1414
Total number of dialogues	430
Maximum length of a dialogue	15
Minimum length of a dialogue	3
Average length of a dialogue	7.81
Total number of positive utterance by a user	242
Total number of negative utterance by a user	622
Total neutral utterance by user	728
Maximum different intents appeared in a dialogue	13
Minimum different intents appeared in a dialogue	1
Average Number of different intents appeared in a dialogue	2.683795712
Number of time sentiment changes to +ve to -ve	20
Number of time sentiment changes to -ve to +ve	5

Table 2: Statistical details of the dataset

4 Intent Extraction

It has been observed that the identification of intent words is more crucial than classifying intents. Moreover, the task to identify the words or tokens that are responsible for that particular utterance related to that intent has been performed first. There can be one or more than one word that can be marked as intent-word. Moreover, the utterances can contain one or more than one word that can be potentially intent word so that we later on can deduct the relationship between intent and sentiment by using these words.

The main motivation of our task is to identify the potential feature words that better represent the intent classes. Therefore, we first use Tf-Idf technique to extract feature words that potentially can be used for identifying the intent-word. We prepared a dataset and marked the feature words that are present in the utterances as potential-intent words.

Secondly, we prepared a dataset to mark tokens in the utterances if they are intent words but we did not feed them into an entity recognition model. The utterances are broken into tokens; so each row of the dataset contains the tokens, their corresponding sentence number, pos tag and entity recognition tag.

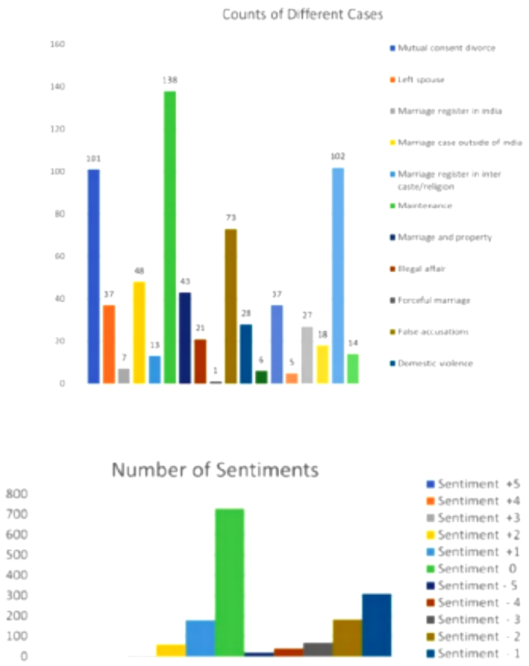


Figure 2: Intent and Sentiment Distributions

4.1 Experimental Setup

The Tf-Idf model has been developed on the primary dataset to identify potential intent words. The primary dataset contains 3178 utterances whereas the intent words were found in 2841 utterances. On the other hand, 337 utterances contain no intent word. The dataset is then fed into the entity recognition model based on BERT (a neural network) to identify and classify the

tokens if they are intent or not. The description and the hyper-parameters of the entity recognition model are as follows:

• Model Name : bert-base-uncased, Training data size : 64%, Validation data size : 16%, Test data size : 20%, Number of epochs : 20, Learning rate : 1e-5, Train batch size : 16, Evaluation batch size : 16, Number of output classes : 39. The precision, recall and F1-score of the model are 0.64, 0.59 and 0.61, respectively.

Intent classification is the automatic classification of text in dialogues or in conversations. The categorization depends on the domain words and the intents are used to tag keywords in conversation so users can easily identify different topics discussed in a conversation. The present dataset contains conversational utterances and each utterance is consists of question and answer on different topics. There are two speakers of each utterance, namely the user and the bot. The dataset has been annotated to have one or more than one intent tagged to each utterance irrespective of the speaker. Our task is to classify those intents based on the given hierarchy of intent classes.

Understanding the context of any natural language text is one of the challenges in dealing with text data. As a result, we need to contextualise texts in vector format. While constructing machine learning models, we tried to deploy embedding using the Tf-Idf vectorizer first. For text feature extraction, we utilised the Tf-idf vectorizer module from the scikit-learn[3] package. In order to feed machine learning models, the dataset has been developed in the format as shown in Figure 3.

4.2 Results

The following are the machine learning classifiers that we have employed in the present task to assess the performance of our intent categorization. The BERT-based architecture, on the other hand, beats the rest of the machine learning and deep learning models substantially. The results of different models developed for intent classification are shown in Table 3.

Multinomial Naive Bayes (MNB): This classifier is suitable for classification with discrete features (e.g., word counts captured from Tf-Idf vectorizer). The multinomial distribution normally requires integer feature counts. We have used the Multinomial NB module with default parameters.

Bot	The decree granted by a Foreign Court is considered to be legal, validity and binding in the Indian Courts by the virtue of Section 14 of the Civil procedure Code.	marriage@divorce @legal_enquiry
Bot	Kindly share the details.	marriage@divorce @mu-tual@legal_enquiry
Bot	Where did you get married?	marriage@divorce @validity
Bot	There is no such law which rescribes to use husband only.Depending upon the con-venience she may choose her surname.	marriage@modify_docs
User	He recorded her voice/call conver-sation with that person in mobile. He has sufficient audio recordings of both which clearly shows that they are in physical relation.	proofs
Bot	welcome, have a nice day	thank

Figure 3: Annotation Snapshot of a Conversation

Stochastic Gradient Descent (SGD): This estimator implements regularized linear models with stochastic gradient descent learning. The gradient of the loss is estimated each sample at a time and the model is updated along the way with a decreasing strength schedule (aka learning rate). SGD allows mini-batch (online/out of core) learning via the partial fit method. For the best results using the default learning rate schedule, the data should have zero mean and unit variance. We have used the SGD Classifier module with hinge loss, l2 penalty and learning rate of 1e-5.

Support Vector Machine (SVM): is a supervised machine learning algorithm used for both classification and regression. We have used the svm module with linear kernel.

Logistic Regression (LR): is a process of modelling the probability of a discrete outcome given an input variable. We have used a logistic regression module with learning rate of 1e5, maximum iteration of 100.

BERT: it makes use of transformer, an attention mechanism that learns contextual relations between words (or sub-words) in a text. In its vanilla form, transformer includes two separate mechanisms — an encoder that reads the

[3] https://scikit-learn.org/stable/

text input and a decoder that produces a prediction for the task. We have fine-tuned the 'bert-base-uncase' model to perform the multi-label classification. The hyper-parameters of the intent recognition model are as follows:

• Model Name : bert-base-uncased, Training data size : 64%, Validation data size : 16%, Test data size : 20%, Number of epochs : 20, Learning rate : 1e-5, Train batch size : 24, Eval batch size : 24, Threshold : 0.2

Metrics	MNB	SGD	SVM	LR	m-BERT
Accuracy	0.27	0.41	0.41	0.40	0.95
Precision	0.15	0.38	0.36	0.38	0.59
Recall	0.26	0.38	0.42	0.42	0.49
F1-score	0.17	0.37	0.37	0.39	0.53

Table 3: Results of the Intent Classification

4.3 Error Analysis

In the case of intent data, it has been found that a single utterance can include several intents. For instance, the utterance "*Can I visit my daughter after the divorce*" combines the intents of '*visitation*' and "*case file*" of divorce. In these circumstances, intent detection by intent classification models is only partially successful.

A single utterance may occasionally have many intentions. For instance, when I brought up divorce, she said, "*I can't endure his abuse." I want to file a case*". The user does not specify whether she wants to launch a lawsuit for divorce, harassment, or both. Some of the sample cases are mentioned in Table 4.

Utterance	Actual Intent	Predicted Intent
So, we were separated for 3 yrs. Now she has an affair with another person. I need divorce.	left spouse, illegal affair, case file (divorce)	left spouse, child visitation
My wife uses bad and abusive language to me and my old father and mother. She has asked about divorce and 20L for maintenance	harassment, case file(divorce) maintenance	child custody, transfer case 38
My mother-in-law is forcing me to sell our ancestral unglow. But I have four brothers, they are not ready for it.	in-law harassment, property issues	property issues, remarry

Table 4: Sample Results of Intent

5 Sentiment Extraction

The second challenge is to determine which words or tokens are responsible for that specific sentiment in an utterance. There can be one or more words that are identified as sentiment-words. Our goal is to find words that better convey different sentiment classes so that we can later use these words to deduce a correlation between intent and sentiment.

We must first identify the feature words that determine the sentiment classes. In order to achieve this task, we first utilise text-blob[4] to extract feature words that could be used to identify sentiment of the whole utterance. We prepared a dataset and label feature words that appear in utterances as potential sentiment-words. Second, we developed a dataset to label tokens in utterances as sentiment-words or not, which we used to feed into a BERT-based entity identification model to discover potential sentiment classes in the dialogue. Tokens have been employed from the utterances, and each row of the dataset contains the tokens, their related sentence number, pos and entity recognition tag.

5.1 Classification

The technique of identifying positive or negative sentiment in text is referred to as sentiment classification. The present dataset has been annotated with values 0, 1, 2 to indicate neutral, positive and negative sentiments, respectively. The snippet of the dataset is shown in Table 5. Already, we discussed on the details of multinomial Naive Bayes, Support Vector Machine and Logistic Regression in the previous sections. In addition to these three, we employed other three classifiers as follows:

Random Forests: is an ensemble learning method for text classification. In the RF classifier a bunch of independent trees is built. Every document is classified by the trees independently. The class of the document is defined by the largest number of votes of all trees.

CNN: In the case of CNN, 1D convolution layers have been used. Each of the convolution layers is followed by a Max-pooling layer. A layer called Max-pooling comes after each convolution layer. The first and second convolution layers, respectively, have 256 and 128 filters set up,

[4] https://textblob.readthedocs.io/en/dev/

while the size of the filter for both layers is set to 5. The training approach is conducted with a 64-person batch size and a 0.0001 learning rate. In both convolution and fully connected layers, the exponential linear unit (ELU) was applied. In the first of the three dense layers, there are 128 hidden neurons with a dropout value of 0.7, and there are 64 hidden neurons with a dropout value of 0.5. For the first dataset, six *softmax* units and eleven *softmax* units were utilised to categorise each user input.

RNN: We chose an embedding layer of size 256 as the first layer in the RNN model based classification framework, and then a 256-layer RNN layer. There were two further LSTM layers placed after these first two. Both of the two LSTM layers have 256 hidden neurons, with the first having a dropout value of 0.3 and the second having a dropout value of 0.3 and recurrent dropout 0.2. A dense layer with *softmax* activation function is present at the end. Using a learning rate of 0.001 and *softmax* units, each user input for the dataset was classified.

BERT: As we know, it is a transformer based model which can be used for various types of classification problems. It uses stacked encoders to train it's language model so well that if it fine-tuned with even a small training dataset it produces wonderful results. The parameters are as

Model Name : bert-base-uncased, Training data size : 64%, Validation data size : 16%, Test data size : 20%, Number of epochs : 20, Learning rate : 1e-5, Train batch size : 32, Eval batch size : 32, Number of output classes : 1

5.2 Results and Observations

The models are evaluated with respect to three evaluation metrics. Results show that deep learning algorithms outperform machine learning techniques and specially, the performance of BERT is significantly better than others. The later part of Table 5 shows that the BERT fails only in case of identifying positive (1: 0 /-1) sentiments.

Models	Precision	Recall	F-Score
Sentiment Class : -1			
M1	0.69	0.42	0.52
M2	0.59	0.89	0.71
M3	0.64	0.79	0.69
M4	0.65	0.71	0.68
M5	0.64	0.85	0.73
M6	0.69	0.81	0.74
M7	0.91	0.90	0.90
Sentiment Class : 0			
M1	0.54	0.63	0.58
M2	0.78	0.72	0.75
M3	0.76	0.78	0.77
M4	0.79	0.75	0.77
M5	0.76	0.78	0.78
M6	0.72	0.86	0.78
M7	0.89	0.91	0.89
Sentiment Class : +1			
M1	0.14	0.21	0.17
M2	0.25	0.31	0.27
M3	0.39	0.21	0.28
M4	0.35	0.33	0.34
M5	0.62	0.10	0.17
M6	0.50	0.45	0.47
M7	0.57	0.63	0.59

Actual Class : Predicted Class							
	M1	**M2**	**M3**	**M4**	**M5**	**M6**	**M7**
-1:0	282	6	16	16	20	17	0
-1:1	134	8	5	6	2	11	0
0:-1	66	34	22	20	18	13	0
0:1	208	19	19	21	6	10	0
1:0	418	18	11	10	7	3	52
1:-1	166	11	2	3	5	6	43

M1: Naive Bayes, **M2:** Support Vector Machine **M3:** Logistic Regression, **M4:** Random Forest, **M5:** CNN, **M6:** RNN, **M7:** BERT

Table 5: Results of the Sentiment Classification

6 Intent-Sentiment Co-reference

One of the primary objectives of co-reference resolution is to locate all phrases in a text that refer to the same thing. It's a crucial stage in a variety of higher-level NLP activities involving natural language comprehension, such document summarising, question answering, and information extraction. Although co-reference resolution has been used in noun-noun or noun-pronoun based resolution, we in this present work are trying to detect co-reference between the intent of user in a dialogue or conversation to the sentiment of the user. The main purpose of detecting intent-sentiment co-reference is to better understand how sentiment of a user depends on the topics they are discussing.

Better scoring means those topics are more sensitive to the users, hence the conversation needs better monitoring. This significantly

improves the performance of a chat-bot to mimic human interactions. We divided the detection of co-reference into two sections. First, we will discuss how to detect if utterances have any sentiment-word co-relates to intent-word. We are going to achieve that we prepared lexicons containing intent and sentiment words and we parse the word dependency using Stanford NLP parser (Stanza) to detect if any relation between them exists.

Pre-processing - We have used the dependency parser developed by Stanford namely "Stanza". The dependency parser takes an utterance as input and generates all the dependencies possible between each and every token. From the input phrase, the dependency parsing module creates a tree structure of words that depicts the relationships between words' syntactic dependencies.

6.1 Experiments

Now, we developed a classification model based on BERT to classify if the utterances have any sentiment-intent co-reference or not thus achieving our thesis objective. The experiment has been carried out in three stages and results are shown in Table 6.

In the first experiment (EXP1), we annotated the dataset to have co-reference label attached to all the utterances. We assigned 1 to utterances where pairs of co-references are there and 0 where there isn't. Then the utterances and the co-reference labels are fed into BERT.

In second experiment (EXP2), we take the dataset already prepared in the previous experiment and add the intent words and the sentiments words extracted from the dataset produced by the dependency parser. Then the concatenated utterances with the intent word and sentiment words are used as input vector and the co-reference as labels to feed into BERT.

In third experiment (EXP3), we take the dataset already prepared in the previous experiment but instead of concatenating utterances with the intent and sentiment word directly we concatenate them by special character used in BERT for separation of input token '[SEP]'. Then the concatenated utterances with the intent word and sentiment words are used as input vector and the co-reference as labels to feed into BERT. The model descriptions and hyper-parameters are given follows:

Model Name : bert-base-uncased, Training data size : 64%, Validation data size : 16%, Test data size : 20%, Number of epochs : 20, Learning rate : 1e-5, Train batch size : 32, Eval batch size : 32, Number of output classes : 1, Number of utterances : 4914, Number of co-referenced utterances : 2378

Experiments	Precision	Recall	F1-Score
EXP1	0.91	0.91	0.92
EXP2	0.93	0.93	0.93
EXP3	0.98	0.98	0.98

Table 6: Results of the Co-reference Identification

7 Conclusion

The goal of co-reference resolution is to find all phrases that refer to the same entity in a text. It's an important step in a number of higher-level NLP tasks requiring natural language understanding, such as document summarization, question answering, and information extraction. Despite the fact that co-reference resolution has been employed in noun-noun or noun-pronoun based resolution, we are attempting to identify co-reference between the user's purpose in a dialogue or discussion and the user's sentiment in this thesis study. The basic goal of intent-sentiment co-reference detection is to better comprehend how a user's sentiment varies depending on the subjects they're talking. Better ranking indicates that certain issues are more sensitive to users, necessitating more careful monitoring of the discourse. This boosts the ability of chat-bot AI to simulate human interactions dramatically.

The detection of co-reference was split into two sections. To begin, we identified all of the utterances in which a sentiment-word is related to an intent-word. We were able to accomplish this by using a database that already contained intent-word and sentiment word information. We then used the Stanford NLP parser (Stanza) to analyze the word dependence and discover co-reference between sentiment and intent.

Acknowledgments

This work supported by the Defence Research Development Organization (DRDO), India, under the sanction code: DFTM/02/3111/M/01/SC-17/P9/JCBCAT-23. Center for Artificial Intelligence and Robotics (CAIR) is acting as reviewing lab for the work concerned.

References

Bengtson, E and D. Roth. 2008. Understanding the value of features for Coreference resolution. In *proceedings of Emnlp'08*, pages 294–303.

Chen, Q Z. Zhuo, and W. Wang. 2019. Bert for joint intent classification and slot filling, arXiv preprint arXiv:1902.10909.

Clark, K. and C. D. Manning. 2016a. Deep reinforcement learning for mention-ranking co-reference models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2256–2262, Association for Computational Linguistics.

Clark, K. and C. D. Manning. 2016b. Improving co-reference resolution by learning entity level distributed representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistic*s (Volume 1: Long Papers), pp. 643–653.

Das, D. and S. Bandyopadhyay. 2009. Word to Sentence Level Emotion Tagging for Bengali Blogs. In *Proceedings of ACL-IJCNLP*, pp.149-152.

Das, S., A. Kolya, D. Das. 2020. Sentiment classification with GST tweet data on LSTM based on polarity-popularity model, *SADHNA* vol. (2020) 45:140

Haghighi A and D. Klein. 2009. Simple co-reference resolution with rich syntactic and semantic features. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pp. 1152–1161, Association for Computational Linguistics.

Hobbs, J. R. 1978. Resolving pronoun references. *Lingua*, 44(4):311–338.

Jansen, J., D. Booth, and A. Spink. 2008. Determining the informational, navigational, and transactional intent of web queries. In *Journal of Information Processing Management*, 44:1251– 1266.

Kasper, Walter & Vela, Mihaela. 2011. Sentiment Analysis for Hotel Reviews. In *journal of Speech Technologies*, Vol. 2, pp. 96-109

Kundu, B and S. Choudhury. Demystifying topology of autopilot thoughts: A computational analysis of linguistic patterns of psychological aspects in mental health. 2017. In *Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017)*, pp. 435–446, Kolkata, India, Dec. 2017. NLP Association of India.

Lee, K., L. He, M. Lewis, and L. Zettlemoyer. End-to-end neural co-reference resolution. 2017. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 188–197, Association for Computational Linguistics.

Lee, U., Z. Liu, and J. Cho. 2005. Automatic identification of user goals in web search. In *Proceedings of the 14th International Conference on World Wide Web, WWW '05*, pp. 391–400, Association for Computing Machinery.

Louvan, S. and B. Magnini. 2020. Recent neural methods on slot filling and intent classification for task-oriented dialogue systems: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 480–496, Barcelona, Spain (Online).

Mondal, A. and D. Das. 2021. Ensemble Approach for Identifying Medical Concepts with Special Attention to Lexical Scope, *SADHNA*, vol. (2021) 46:77

Ng, V. and C. Cardie. 2002. Improving machine learning approaches to co-reference resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 104–111.

Passonneau, R.2004. Computing reliability for co-reference annotation. In *Proceedings of International Conference on Language Resources and Evaluation*, Lisbon, 2004.

Qin, L., W. Che, Y. Li, H. Wen, and T. Liu. 2019. A stack-propagation framework with token-level intent detection for spoken language understanding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP*) Association for Computational Linguistics, pp. 2078–2087.

Soon, W. M., H. T. Ng, and D. C. Y. Lim. 2001. A machine learning approach to Coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544.

Stoyanov, V. and Cardie, C. 2008. Topic Identification for Fine-Grained Opinion Analysis. In *Proceedings of COLING*, pp. 817-824.

Su, J., K. Duh, and X. Carreras. 2016. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, Texas, Nov. 2016. Association for Computational Linguistics.

Wiseman, S., A. M. Rush, and S. M. Shieber. 2016. Learning global features for Coreference resolution. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, *Association for Computational Linguistics* pp. 994–1004, San Diego, California, June 2016.

Xu, P., and R. Sarikaya. 2013. Convolutional neural network based triangular crf for joint intent detection and slot filling. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 78–83.