A Chatbot Framework for the Children's Legal Centre

Jay Morgan, Adeline Paiement, Monika Seisenberger, Jane Williams, Adam Wyner

A Chatbot Framework for the Children's Legal Centre

Jay MORGAN [a], Adeline PAIEMENT [a], Monika SEISENBERGER [a], Jane WILLIAMS [b], Adam WYNER [a,b],
[a] *Swansea University, Department of Computer Science*
[b] *Swansea University, School of Law*

Abstract. This paper presents a novel method to address legal rights for children through a chatbot framework by integrating machine learning, a dialogue graph, and information extraction. The method addresses a significant problem: we cannot presume that children have common knowledge about their rights or express themselves as an adult might. In our framework, a chatbot user begins a conversation, where based on the circumstance described, a neural network predicts both speech acts, relating to a dialogue graph, and legal types. Based on the legal types, relevant legal rights are returned to the user. Information is extracted throughout the conversation in order to create a case for a legal advisor. In collaboration with the Children's Legal Centre Wales, who advocate for the improvement and dissemination of legal rights in Wales, a corpus has been constructed and a prototype chatbot developed. The framework has been evaluated with classification measures and a user study. The framework can be extended and adapted to other domains.

Keywords. Children's Legal Rights, Chatbot, Natural Language Processing, Machine Learning, Recurrent Neural Networks

1. Introduction

Chatbots or conversational agents are computer programs that allow for interaction with systems through natural language inputs and outputs [9,15]. What differentiates them from other systems that use natural language, e.g. search engines or controlled natural languages, is the dialogic interaction [7]. Chatbots can automate mundane or difficult tasks by asking the user for some information and responding so as to lead them through the process one step at a time.

Chatbots can be used in legal consultations to make legal processes more accessible to the public by reducing the burden of legal knowledge that may be applicable in a situation. However, such legal chatbots are designed with adult users in mind and presume "everyday" legal and commonsense knowledge.

In this paper, we present a novel chatbot framework, which aims to improve children's and young people's access to information about their legal rights, where we cannot presume that they have any prior knowledge of the law. We use a combination of machine learning (topic modelling with recurrent neural networks) and natural language processing (parsing and textual information extraction).

For a demonstration of this framework, we have developed a prototype in collaboration with the Childrens Legal Centre Wales (CLC)[1]. The goals of the chatbot are: 1) identify the legal circumstances of the users; 2) identify some of the parties in the circumstances; and 3) given the information in 1 and 2, create a case that can be followed up by an advisor. To achieve these goals, we create a new corpus to train a neural network to recognise the relevant legal circumstances.

In Section 2, we provide background on chatbots, the implications of GDPR, and the work of the CLC. In Section 3, we discuss the corpus and its features. Section 4 presents the methodology, and Section 5 evaluates the framework through classification measures and user studies. This paper concludes with a discussion of limitations and future work in Section 6.

2. Background

The first chatbot was *Eliza* in 1966 [14], which analysed the input text for keywords and linguistic features that were then used by the rules to trigger and fill slots in its response. More recently, corpus-based chatbots have been developed [12] using machine learning to learn the appropriate response given the input message [15]. We are interested in corpus-based legal chatbots.

2.1. Legal Chatbots

There have been successful chatbots that provide legal services. Most notable is DONOTPAY bot[2]; for example, the tool asks motorists questions about their situation, then generates a letter to appeal a parking ticket.

Visabot is another domain-specific chatbot that helps users with immigration issues; the tool has a range of functionalities - asking questions, gathering data, providing forms, analysing and extracting textual information, and drafting documents for the user to provide to a lawyer. DONOTPAY uses fixed options, rather than free writing. Both tools assume an adult's level of comprehension. The dialogue interactions are limited to asking specific questions in order. No ongoing legal support is provided. As proprietary tools, they are unavailable for academic development.

Existing general purpose or domain specific tools do not address our requirements. That is, we do not know of any chatbots designed for children's rights, where we must relate the language of children to legal concepts, as a child may describe a problem in everyday rather than legal terms. We must also classify input speech acts for interaction whilst extracting information to populate a case.

We have considered the implications of the General Data Protection Regulation (GDPR), which stipulates consent for the usage of personal data with respect to children under 16. In situations where automated decisions are made using personal data, consent has to be granted [8]. However, in situations where consent can place the child in further harm, parental consent should not be necessary.

[1] `https://childrenslegalcentre.wales/`
[2] `https://www.donotpay.com/`

	NS		LR		NT		WT		AL (Std)	
Speech Act	**A**	**R**	**A**	**R**	**A**	**R**	**A**	**R**	**A**	**R**
Greeting	153	36	9.89	4.00	633	172	64	43	17.46 (13.31)	19.14 (9.42)
Statement	467	148	12.22	6.30	4642	2174	380	345	45.35 (16.63)	65.53 (42.68)
Positive	151	24	8.23	1.82	255	60	31	33	6.97 (4.15)	9.52 (11.81)
Negative	143	17	7.26	2.56	385	92	53	36	11.11 (7.63)	20.59 (15.94)
Legal Type	**A**	**R**	**A**	**R**	**A**	**R**	**A**	**R**	**A**	**R**
Abuse	187	46	9.43	3.57	1706	550	181	154	40.72 (15.88)	56.09 (38.84)
Hate-crime	78	19	7.08	2.80	913	227	128	81	55.23 (14.79)	53.00 (41.48)
Cyber-Crime	105	15	5.84	2.80	998	319	171	114	45.33 (16.90)	102.20 (35.50)
Underage Sex	97	54	7.59	6.32	1025	973	135	154	46.35 (15.48)	76.91 (41.33)

NS Number of Statements - Total number of statements for the classification type.
LR Lexical Richness - Number of unique tokens that occur throughout the classification type.
NT Number of Tokens - Number of non-unique tokens that occur.
WT Word Types - Number of unique tokens that occur.
AL Average Sentence Length - Average sentence length by token (standard deviation of average).

Table 1. Analysis of artificial and real statements for both speech acts and legal types.

2.2. The Children's Legal Centre Wales

The centre provides consultations and information about laws that affect children and young people in Wales. In 2016, they received a grant to develop a Virtual Legal Practice (VLP) that would allow law school students and practising lawyers from across Wales to help with the consultations. Part of VLP allows parents and young people to get in touch with a representative of the practice, so as to develop and manage legal cases on the client's behalf. The chatbot framework fits within the scope of VLP activities; in addition, the CLC helped with user evaluation.

3. Corpus Dataset

For supervised machine learning, a corpus is required. While there are legal corpora, e.g. the British Law Report Corpus (BLaRC) [13], some of which bears on children's rights and family law [4], the terminology, style, and content is not such as we might expect children to use. We are unaware of an available "naturally occurring" corpus of chats by children or young adults about legal matters

Therefore, we created, based on the expertise of the CLC, a novel corpus of chat messages. Each message is in (English) language that experts deemed to be similar to that of children in Wales. The corpus (Table 1) is comprised of "artificial" statements (labelled A), which were generated by the CLC in order to approximate how a child would describe the situation. Additional "real" statements (labelled R) were taken from how participants in a user study would describe the same situation. The messages are classified in terms of *speech act* and *legal type*.

As artificial and real statements appear in the corpus and are used to train the model, we compare them linguistically so as to identify any impacts on the model. In Table 1, we find the lexical richness (LR) to be higher by a factor of two for artificial statements than the real statements for almost all types. However, the number of unique tokens (WT) are similar, indicating that even though the

number of artificial statements is greater than the real statements, the sentences are constructed from the same words but with varying sentence styles. Variety is helpful for training a model that can generalise to the styles of different people.

We additionally measure the statistical difference between the statements types through the Linguistic Inquiry and Word Count (LIWC) method. This provides a quantitative measure to the conveyance of people's circumstance and emotional state by looking at the wording and language style [16]. Table 2 shows that both types of statements are equally expressive of positive emotions, with the real statements being more negative. There is also clear divergence with the usage of pronouns and cognitive processes, where the artificial statements use a lot more pronouns (such as I or me). The real statements focus more on describing the circumstance that results in a larger cognitive processes score.

We discuss in Section 5.2 the impact of training a neural network on the artificial and real statements.

LIWC Measure	**Artificial Samples**	**Real Samples**	**Difference**
Pronouns	25.2	12.0	13.2
Social Words	15.2	15.1	0.1
Positive Emotions	1.0	0.9	0.1
Negative Emotions	3.4	4.5	1.1
Cognitive Processes	7.9	13.2	5.3

Table 2. LIWC statistical difference between the artificial and real statements.

4. Methodology

When the user accesses the chatbot's interface, a new instance of a dialogue graph is created that tracks the interactions between the chatbot and user (Section 4.1). To respond naturally to the user, the chatbot must perform two tasks (detailed below). First, it reasons as to the role of each input within the dialogue graph as well as to the legal type being discussed. These two classification tasks are performed by a neural network (Section 4.2). The classifications are considered by the internal program logic in relation to the current position in the dialogue graph to determine the next move in the conversation. The internal program logic will select a predefined response from a database. Any unclassifiable statements are deemed to be outside the intended dialogue graph, and a default response is returned to keep the conversation going. Second, named entities are extracted from the user's messages (Section 4.3), and saved in the database to be used in the conversation.

Given the legal type as determined by the neural network, the chatbot attempts to help the user learn about their legal rights. It does this by retrieving the most relevant legal content from its host website (the CLC) to the user.

At the end of the conversation, the chatbot's internal memory is used in order to generate a formatted, readable PDF document that can be sent to an advisor.

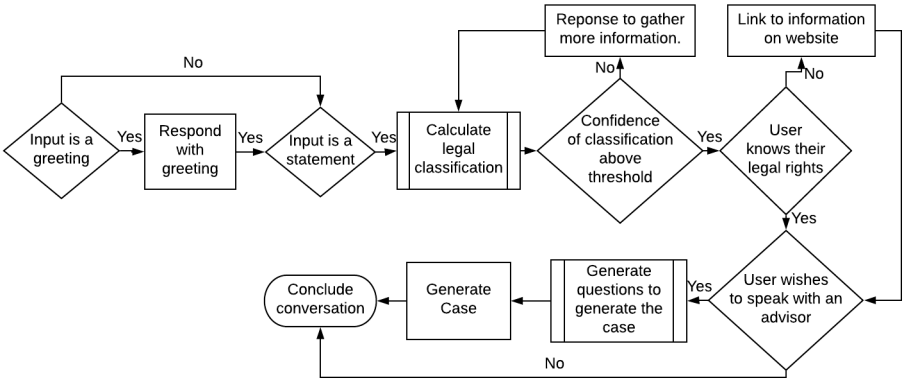

Figure 1. Dialogue graph used by the chatbot.

Legal Type	**Required Information**
General Information	Contact name, time of event, contact information, and contact time.
Abuse	Event location. Who is the abuser and abused.
Cyber-Crime	Which platform the event occurred. The reason behind the case.
Hate-Crime	Who committed the act. What act was committed.
Underage-Sex	The age of parties involved. The reason behind the request.

Table 3. Information that the chatbot must acquire for the advisor's case.

4.1. Dialogue Model

We design the dialogue graph (Fig. 1) according to its three goals, i.e. identifying legal circumstances, accessing one's rights, and engaging with an advisor (Section 1). The initial stage of the conversation consists of turns between the user and the chatbot in order to discover the nature of the legal type. This implicitly follows an information-seeking dialogue model as described by Walton and Krabbe [2]. Once the legal type is discovered, the dialogue shifts to an action-deliberation type [2], wherein the chatbot and user attempt to decide upon a plan of action [5], in particular, whether the user wishes to talk with a legal advisor. If the user does wish to contact an advisor, the conversation may return to that of an information-seeking type to gather the necessary information that the advisor will need (see Table 3). Some information may have already been acquired and stored in the database, allowing the chatbot to focus only on the missing data.

4.2. Classification of the Message's Functions and Contents

Given a user's input statement, a neural network is used to classify the speech act and legal type. The resulting classifications are used to progress through the dialogue graph. The classification process is illustrated in Fig. 2.

Words are tokenized and converted to word vectors of 200 dimensions. These word vectors aim to capture the semantic similarities between words [3], allowing for the usage of synonyms and improving the model's generalisability [17].

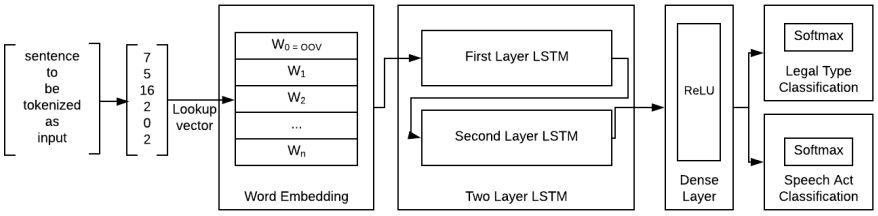

Figure 2. Neural Network Architecture

Network Type	**Speech Act F_1 Score**	**Legal Type F_1 Score**	**Avg F_1 score**
Dense Neural Network	97.36% (+/- 0.57%)	93.93% (+/- 1.21%)	95.65%
2 Layer RNN	95.16% (+/- 1.73%)	93.23 (+/- 1.24%)	94.20%
Pre-trained Embedding	98.41% (+/- 0.90%)	98.06% (+/- 0.35%)	98.24%

Table 4. Comparison between Dense and Recurrent Neural Network classification scores. Tests consist of a dense network with an untrained embedding layer; replacing the dense layers for a 2 recurrent layers; using a pre-trained embedding.

We utilise a pre-trained word embedding from GloVE, that was trained on Twitter data. This word embedding provides the best textual similarity between the intended audience of the chatbot platform and the pre-trained embedding.

These word vectors then go through two LSTM layers that assist with the classification tasks by encoding the impact of word (vector) order in the sentence's meaning. Each sentence is encoded individually by these LSTM layers. We use the final internal state of the second LSTM layer as the sentence encoding [18].

It is further transformed by a dense layer with a ReLU activation function, and a dropout rate of 20% to reduce overfitting. The sentence is classified as a speech act and legal type by two parallel dense layers with a softmax activation.

4.3. Named Entity Recognition Class

The system identifies and extracts named entities from within the user's statement to be used later in the creation of the case without explicitly asking for the information. For well-formatted input, e.g. email addresses, we use regular expressions (REGEX). In other cases, we use a neural net to recognise named entities as well as provide syntactic dependency relations to identify the syntactic roles of arguments [6,11,1]. The chatbot uses a set of rules about how to interpret these syntactic relations in order to extract entities relative to syntactic roles.

5. Results and Evaluation

5.1. Evaluation of the Classification

We perform a 5-fold cross validation to evaluate how well the model classifies both legal types and speech acts, describing the accuracy through the F_1 score.

Training Corpus	**Speech Act F_1 Score**	**Legal Type F_1 Score**	**Avg F_1 Score**
Artificial Data Only	92.00%	92.88%	92.44%
Real Data Only	88.00%	83.99%	86.00%
Artificial & Real Data	98.41%	98.06%	98.24%

Table 5. Impact of training with artificial data on the classification scores.

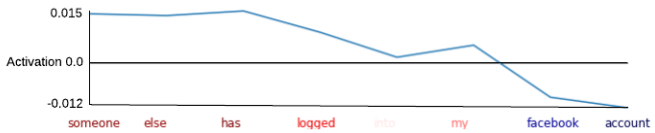

Figure 3. Salient words for a sample sentence of the cyber-crime legal type.

We compare the effect of using a LSTMs and a pre-trained embedding (see Table 4), beginning with a simple dense neural network with an untrained embedding layer. As we replace the dense layers with LSTM layers (Fig. 2), the classification scores for both speech act and legal type drop. This could be due to the small size of our corpus which may not allow the LSTM layers to learn accurate sentence representations. Using a pre-trained embedding layer, the subsequent layers are provided with more descriptive word vectors, thus the LSTMs can form accurate sentence representations quickly and with less training data. This results in the best score for this experiment, at 98.24%.

Additional outputs could be included to predict the sentiment of the statement and the degree of harm the child could be in. This would then be included into a triage system where more serious cases are ranked at a higher priority. Due to the implications of the network creating an incorrect prediction, we see this as future work that benefits organisations with many incoming cases.

5.2. Classification Scores on the Artificial & Real Statements

As the corpus is constructed from both artificial and real statements, there is a benefit to measuring the classification performance by adding real statements, assessing any limitations in online learning from future conversations.

We perform 3 experiments that consist of: 1) training only on artificial data; 2) training only on real data; 3) training using a both artificial and real data (Table 5). In all experiments, a sample of the real statements are reserved for testing. The resulting F_1 scores show that training on a combination of artificial and real statements has a positive impact on the final test accuracies. This is perhaps due to the real statements having a higher cognitive processes score from a longer explanation of the circumstance (Section 3). Using only the real statements for training does not provide a diverse enough training set to generalise well. Therefore, we argue that creating a corpus from both artificial and real statements yields the best result for this corpus size.

5.3. Explainable Neural Network Classifications through Word Salience

As neural networks can be viewed as composite functions that transform an input x into an output y, the hidden proprieties of this transformation are often

Legal Type	Excerpt
Abuse	You are worried about how angry your mum and dad get with each other.
Cyber-Crime	Another pupil took photos and posted them on Instagram making fun of you.
Underage Sex	You are 20 and she is 16. You want to know if it is statutory rape.
Hate Crime	You even a heard a teacher saying you came from a Gypsy family.
Underage Sex	You want to know if it is against the law have sex with your boyfriend who is 15.

Table 6. Excerpts from the 5 circumstances that are given to participants in the user study.

overlooked. Understanding the reasoning behind why user's sentence has been classified as a certain type can lead to a qualitative performance. We investigate whether the network is learning keywords that justify the classification.

We transfer the trained weights into a duplicated network that includes an additional output: the LSTM activation of cell states at each timestep (Fig. 3). A strong activation may indicate a salient word that results in the prediction.

For all legal types, we remove the 5 most salient words from the testing set, perform a prediction for all sentences, and measure the rate of correct statement classification. With no salient words removed, the F_1 score is 98.41%, with 5 words removed the score drops to 91.25%. Removal of these salient words from the sentences therefore has a dramatic impact on the legal type classifications.

This method of identifying possible salient words provides two methods for further variation within our corpus: 1) restructuring of the original statements whilst keeping the salient word the same, presenting the salient word in different contexts; 2) replacing a salient word with a synonym, making the model more robust when no salient words appear.

5.4. User Studies

The goal of the study was: 1) to understand how participants would phrase the circumstances; 2) to perform a usability survey based on the experiences.

We invited 14 participants to select 3 situations from a total of 5 given situations (Table 6), then to begin conversing with the chatbot about each of the selected circumstances and complete a questionnaire describing their experience. As the corpus was formed from an adult interpretation of how a child may express a situation, the participants played the role of a child to best match how the corpus was conceived to how it should be tested.

The questionnaire consisted of 11 questions. Responses were ranked 1 to 5, with 1 being the lowest score (strongly disagree) and 5 being the highest (strongly agree). Questions that required a no, maybe, yes response, were translated to 1, 3, 5 respectively. Two questions for each measure were accumulated, and the average score for all participants were taken (Table 7).

1. How easy was the chatbot to use?
2. How easy was it to create a case with an advisor?
3. How well do you feel that the chatbot understood you?
4. The pace of the interactions were suitable.
5. If the chatbot did not understand, was it easy to reformulate the response?
6. How friendly was the chatbot?
7. Were the questions that the chatbot was asking you clear?

User Study Measure	Minimum	Maximum	Average
Ease of Use (Q1, Q2)	6	10	7.42
Interaction Performance (Q4, Q8)	6	9	6.71
Politeness & Responses (Q6, Q7)	7	10	7.57
Perceived Understanding (Q3, Q5)	3	10	5.00
Future Use (Q9, Q10)	4	10	8.29

Table 7. User study questionnaire responses.

8. The conversation felt natural.
9. Would you use the system again?
10. Overall, do you find yourself satisfied with the experience?
11. Free form feedback.

Ease of use shows the degree of difficulty in using the chatbot to create a case. As this system is designed to be an alternative method of creating a case, the ease of use should not be hampered. *Interaction performance & politeness* is determined by the dialogue graph and templating of responses. With this prototype, participants found the pace of the conversation to be suitable.

Perceived understanding evaluates the participant's belief that they have been understood by the chatbot. Despite the neural network's classification score, we see this measure drop. From the free form feedback in the questionnaire, we find this due to the templated responses. The chatbot's response did not include information to indicate that it had understood, rather, it would move on to the next question without acknowledgement. To present a prosocial chatbot, it would be beneficial to include more slots within the response that would rephrase the user's input to create a psychological echo effect [10].

6. Conclusion and Future Work

We have presented a chatbot framework to improve children's access to their legal rights, automatically detecting the legal type and creating a case. Our method uses machine learning to perform joint predictions of the speech act and the legal type being described by the user. The framework relates the predicted speech act with the current position in the dialogue graph and dynamically constructs a set of questions for the predicted legal type in order to best assist the child in accessing their legal rights. We incorporate named entity extraction and syntactic relationships to extract knowledge for the case creation.

We see several steps that can be addressed to further improve this framework: 1) improving the chatbot's response through echoing techniques to demonstrate understanding and provide a more prosocial experience; 2) performing sentiment analysis within the circumstance that can provide additional insight for a triage basis on the case; 3) scaling of the framework to a greater array of legal types and corpora size; 4) enhancing the existing corpus through child participation; 5) increasing compatibility with GDPR, from data privacy to automated decision making; 6) finally, further refining and evaluating the scope of parsing through dependency graphs to extract further detailed information from the conversation.

Based on the results presented from the user evaluations and classification abilities of the neural network, this framework has shown the ability to naturally detect and create a case based on the legal type. As this framework takes care of identification of the legal type from the circumstance, a child may more easily learn the legal rights applicable to their situation and get access to a legal professional without the need for common knowledge about the law.

References

[1] J. D. Choi and M. Palmer. Guidelines for the CLEAR Style Constituent to Dependency Conversions. *Center for Computational Language and EducAtion Research University of Colorado Boulder Institute of Cognitive Science Technical Report 01-12*, 2012.

[2] F. H. V. Eemeren, R. Grootendorst, J. A. Blair, A. Charles, and D. N. Walton. Types of Dialogue, Dialectical Shifts and Fallacies. *Argumentation Illuminated*, (1984):133–147, 1992.

[3] M. Faruqui and C. Dyer. Improving Vector Space Word Representations Using Multilingual Correlation. *Proceedings of the 14th Conference of the EACL*, pages 462–471, 2014.

[4] Flax. British Law Report Corpus (BLaRC).

[5] D. Hitchcock, P. McBurney, and S. Parsons. A framework for deliberation dialogues. *Proceedings of the Fourth Biennial Conference of the Ontario Society for the Study of Argumentation*, 2(73):275, 2001.

[6] M. Honnibal and M. Johnson. An Improved Non-monotonic Transition System for Dependency Parsing. *Emnlp 2015*, (September):1373–1378, 2015.

[7] D. Jurafsky and J. Martin. Speech and Language Processing. In *Speech and Language Processing.*, volume 3, pages 441–458. Pearson London, 2014.

[8] D. Kamarinou, C. Millard, and J. Singh. Machine Learning with Personal Data: Profiling, Decisions and the EU General Data Protection Regulation. *Queen Mary School of Law Legal Studies Research Paper No. 247/2016*, pages 1–7, 2016.

[9] A. Kerly, P. Hall, and S. Bull. Bringing chatbots into education: Towards natural language negotiation of open learner models. *Knowledge-Based Systems*, 20(2):177–185, 2007.

[10] W. Kulesza, D. Dolinski, A. Huisman, and R. Majewski. The Echo Effect: The Power of Verbal Mimicry to Influence Prosocial Behavior. *Journal of Language and Social Psychology*, 33(2):183–201, 2014.

[11] M. C. D. Marneffe and C. D. Manning. Stanford typed dependencies manual. *Technical report, Stanford University*, pages 1–10, 2008.

[12] M. F. McTear. The rise of the conversational interface: A new kid on the block? In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 10341 LNAI, pages 38–49, 2017.

[13] M. J. M. Pérez and C. R. Rizzo. Design and compilation of a legal English corpus based on UK law reports: the process of making decisions. *Las Tic: Presente y Futuro en el analisis de corpus*, pages 101–110, 1996.

[14] B. A. Shawar, E. Atwell, and A. Roberts. FAQchat as an Information Retrieval System. In *Proceedings of the 2nd Language and Technology Conference*, pages 274–278, 2005.

[15] B. A. Shawar and E. S. Atwell. Using corpora in machine-learning chatbot systems. *International Journal of Corpus Linguistics*, 10(4):489–516, 2005.

[16] Y. R. Tausczik and J. W. Pennebaker. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54, 2010.

[17] J. Turian, L. Ratinov, Y. Bengio, and J. Turian. Word Representations: A Simple and General Method for Semi-supervised Learning. *Proceedings of the 48th Annual Meeting of the EACL*, pages 384–394, 2010.

[18] P. Zhou, Z. Qi, S. Zheng, J. Xu, H. Bao, and B. Xu. Text Classification Improved by Integrating Bidirectional LSTM with Two-dimensional Max Pooling. 2(1):3485–3495, 2016.