International Journal of Engineering Applied Sciences and Technology, 2021
Vol. 6, Issue 4, ISSN No. 2455-2143, Pages 354-360
Published Online August 2021 in IJEAST (http://www.ijeast.com)

A SURVEY ON DEEP LEARNING FOR THE DETECTION OF THE INAPPROPRIATE CONTENT PRESENT IN TEXT

Shivakumar H Teli
Department of Electronics and Communication Engineering, RVCE Bangalore

Dr Kiran V
Associate Professor
Department of Electronics and Communication Engineering, RVCE Bangalore

***Abstract*—certain piece of textual information produced by any user or agent is said to be inappropriate if the expressed intent can cause hate, annoyance to other users or exhibits lack of respect, rudeness, which is disrespectful towards certain individuals or communities who may cause harm to oneself or others. In the present day scenario the different classification techniques are used to filter this kind of annoying text or messages. And browsers this days should be able to filter such kind of searches done in the searching engines which will be done every day. Providing such classification technique to filter such messages or searches which are not appropriate using some of the deep learning algorithms and considering the web search conversations such kind of searches which is found as abusive or which might cause hatred can be eliminated.**

Keywords— Query classification, Deep learning, Query autosuggest, Web search, Supervised learning.

I. INTRODUCTION

Rapid growth of chat bots and Interactive Gaming Systems expect large, real time, clean human conversation data to train their models. High usage of social networking sites provide a forum for users to communicate and express opinions publicly. Inappropriate interactions and posts in these forums can also lead to loss of business and damaging company's reputation.

Any natural language text phrase is described as inappropriate if one of the following is intended-

(a) Rude, discourteous or lack of respect for other individuals or groups of individuals
(b) Capable of inflicting harm on oneself or other persons
c) in connection with an activity which is illegal under the laws of The country, or d) Is experiencing extreme violence.
Below are some of the unacceptable examples from the document.

1. You feel sort of like a zombie.

2. Blacks are swindlers.

3. Clear ways of committing suicide.

4. I 'd love to break you apart.

5. How to sow marijuana?

6. Govt. should just kill all the people who do business in leather.

7. Kill every Jew.

In the examples above, first and second show rudeness towards a individual and disrespect / racism towards an individual Individuals section. Third and fourth are examples of harm to oneself and harm to others. Fifth Speech About Marijuana, A drug that is illegal in many countries and the last two cases are the most serious. These are unacceptable to use and insulting for business people and Jews in clothing.

With over a billion searches every day, the search which are done by us have become a representation of the society Humans attitudes and prejudices. Therefore, besides clean-intentioned requests, search query logs also Includes questions that convey abuse, hate speech, discrimination, pornography, profanity and illegality. Accordingly, While the search engines can sometimes unconsciously offer alternative completions from search logs demand finishes which are unacceptable for users. Although users still have the right to search for anything they need, there is a browser that offers such Inadvertently-

a) Unprofessional suggestions may be considered as promoting those views And the brand name is tarnished.

(b) Harm the image of other persons or communities leading to legal problems or

International Journal of Engineering Applied Sciences and Technology, 2021
Vol. 6, Issue 4, ISSN No. 2455-2143, Pages 354-360
Published Online August 2021 in IJEAST (http://www.ijeast.com)

(c) Helping someone who is attempting to injure themselves or others. In the history, there have been cases in which search engines have been dragged into legal tussles over such unexpected ideas.

Importance of Identifying Inappropriate text in Conversations

Chat bots are computer programs that mimic interaction with artificial intelligence-users. these programs are also designed to simulate convincingly how a person will act as a conversation partner. Chat bots are usually used for various functional purposes in dialog systems including customer service or the gathering of information. Most of the popular chat bots use sophisticated natural language processing systems and machine learning techniques.

Chat bots acquire the knowledge by using the pre-trained computer to respond to user messages learning models equipped with large sets of human conversations in real time data. The probability is high that human chat data can be filled in with inappropriate real-time conversations that will lead the chat bots to produce some inadequate answers to clean conversations.

The use of chat bots has grown rapidly, especially in services such as sales, marketing, and customer service. Company Insider's survey showed that nearly 44 per cent of the US consumers would like to use customer service chat bots and more than 37% of Americans would like to be prepared to make a purchase via chat bot interfaces.

Since the demand for chat bot is strong in many fields, training the chat bots to communicate politely to customers otherwise it will vitiate both the business and the image of the respective business enterprise. In addition, interactive gaming systems and social networking sites provide users with a platform to share their views.

Challenges involved in Identifying Inappropriate Text

Due to lack of sufficient context, the web browsers which are used for the searching of queries or questions is quite challenging and syntactic structure in web requests, the occurrence of spelling errors and the ambiguity of the natural language for the machine to identify which are the messages or text spread any hatred or rudeness. For example, a question such as "defecate on my face video" may sound extremely offensive and therefore inappropriate but this is the name of a famous song. Likewise, "what to do when you tweak by yourself" is a query where the term "tweaking" refers to the use of meth-a drug that is illegal and, therefore, the drug recommendation is wrong. A query such as "hore in bible" has a spelling error which refers to Whore which makes an offensive demand.

Similarly, the detection of improper conversations is often difficult for the following reasons :

1. We consider conversations usually of highly variable duration. The normal talk between two users of any online applications may differ from a simple chat to a long comment, like hi. We consider the maximum duration in our data of 250 words of speech.

2. Many misspelled words are usually seen in the conversations. We found in our data that 70 per cent at least one term of the conversations was misspelled.

3. There are also icons and smileys in conversations.

II. LITERATURE SURVEY

The identification of improper messages is especially difficult. Non-graphmatic structure of query, misspellings and query length variable make this classification problem difficult. Investigators worked on detecting abusive question flames on a given domain. Subsections after give details on the research work involved.

Vandersmissen et al. [1] applied Naive Bayes and Support Vector Machines(SVM) techniques are used for the automatically detect messages which may contain the text or messages which one might find it out as offensive language on Dutch social networking site Netlog. The developed multi class classifier is designed to detect ”Sexist”, ”Racist” and ”Irrelevant” messages.

This classifier has 3 steps where the 3rd step of the classifier also detectes ”Outrage” class of offensive messages using pre-defined wordlists. They report, Naive bayes classifier with Precision 9% and Recall 86% performs very badly compare to SVM for detecting offensive messages which gives a precision of 69% and recall of 62%. They have combined a wordlist based semantic model to the SVM and this hybrid model achieves a precision of 93% and Recall of 46%. “In this context the paper which used the a combination of SVM and word list based on the classifier which are used in the paper to find the results which performs better and observed their other models and also checks for the part which performs badly with less context which is usually the case with the search questions as well”.

International Journal of Engineering Applied Sciences and Technology, 2021
Vol. 6, Issue 4, ISSN No. 2455-2143, Pages 354-360
Published Online August 2021 in IJEAST (http://www.ijeast.com)

Xiang et al. [2] created a dataset for offensive message classification on twitter corpus using Bootstrapping technique. They used a method called 'statistical topic modeling (LDA)' [5] and lexicon based features for the detection of the offensive tweets that are made. These try various models (J48 'decision tree learning', 'SVM's, 'Logistic Regression (LR) and Random Forests (RF)' with these features and check for which of the method that are used will perform better and found that LR performs best with F1 Value of 0.849 using 50 topic features. They also report other metric using true positive rate. The report says that "their approach achieves a true positive rate (TP) of 75.1% over 4029 testing tweets using Logistic Regression which is significantly outperforming the keyword matching baseline, which has a TP of 69.7%, while keeping the false positive rate (FP) at the same level as the baseline at about 3.77%. It is quite difficult to find the topical features from the search queries made which will be as usual small and have less context".

Razavi et al. [4] detect flames (offensive/abusive rants) from text messages using a multi-level classification approach. In the first level they use Complement Nave Bayes classifier for selecting the most discriminative features from a tokenized raw features. In the next step Multinomial Updatable Naive Bayes classifier is used to update the model. They use Insulting or Abusive Language Dictionary (IALD) to generate aggregated features form the output of second level classifier. In the last level, they run a rule-based classifier named DTNB (Decision Table/Naive Bayes hybrid classifier) to make choice on the final label of the message.

Xu et al. [3] made use of the grammatical relations and a curated offensive word list which are used to identify and filter the offensive language messages in online social forums. They have created a dataset by manually filtering over 11,000 text comments from the YouTube website. They designed a procedure to construct a Relation Tree (RelTree) using Parts-of-Speech(POS) and Typed Dependency Relations among words in a sentence. The heuristic rules created by them finds the offensive chunks in the sentences and also replaces them with clean words without losing the readability of the sentence i.e. the grammar of the sentence won't change much. Their performance metric is reported as, "with 2063 sentences containing offensive words, the number of insufficient filtering is 58 (i.e. 2.81%), and the number of excessive filtering is 129 (i.e. 6.25%). To sum up, the overall ratios of accuracy of their implemented semantic filter is 90.94%." They have also implemented an extension for Mozilla Firefox browser to filter the text which might contain some of the unwanted contexts which are found to be offensive in Online Social Networking websites.

Chuklin et al. [6] the method used will automatically classify the search queries made in the browsers or any other online social network platforms with adult intent into three categories namely – "black (adult intent), grey, white (clean intent). They use gradient boosted decision trees for classification using Adultness of the query and manually created black and white list as the features".

Researchers who are doing works on this contexts from past years they tried combining CNN and LSTM architectures in the context of other text mining problems such as Named Entity Recognition (NER) and Sentiment Analysis. Zhou et al. [12] said they tried "for the combination of the CNN and LSTM architectures to create a hybrid model C-LSTM and apply it for sentiment analysis of movie reviews and question type classification. On both these tasks, they show that the C-LSTM model not only outperforms hand-crafted feature based baseline models", but also individual CNN and LSTM models.

Sainath et al. [13] said that "we used the method to sequentially combine convolutional, LSTM and fully connected layers into a single architecture named CLDNN for the problem of speech recognition. They report that CLDNN performs better than other baselines and achieves 4-6% reduction in Word Error Rate(WER)". They also prove that CLDNN is better than individual CNN and LSTM models.

III. C-BILSTM USED FOR INAPPROPRIATE QUERY DETECTION

C-BiLSTM turns to The input search query and the likelihood of an inappropriate class query is output. The input search query is fed into the model in a word embedding matrix format. Form C-BiLSTM Consists of three layers- a) Convolution (CONV) Layer

b) Bi-directional LSTM (BLSTM) Layer and

c) Fully Connected (FC) layer

Provided the embedding matrix for the input application, the CONV layer learns which is a new lower-dimensional representation for the input query function, which is then fed into the BLSTM stratum. The BLSTM layer takes representation of the CONV layer query as input and in turn outputs a function representation that encodes the sequential

International Journal of Engineering Applied Sciences and Technology, 2021
Vol. 6, Issue 4, ISSN No. 2455-2143, Pages 354-360
Published Online August 2021 in IJEAST (http://www.ijeast.com)

patterns in the forward query from both and the directions reverse. This kind of a feature represented as above which will then goes through the FC layer, which models the various interactions between these features and the probability of a question which is used in the search queries belonging to them is finally output to the class inappropriate.

The output of the BLSTM layer (32 dimensional feature vector) is given as input to a Fully Connected (FC) layer that models the interactions between those characteristics. In the FC layer the final softmax node outputs the probability of inappropriate class query.

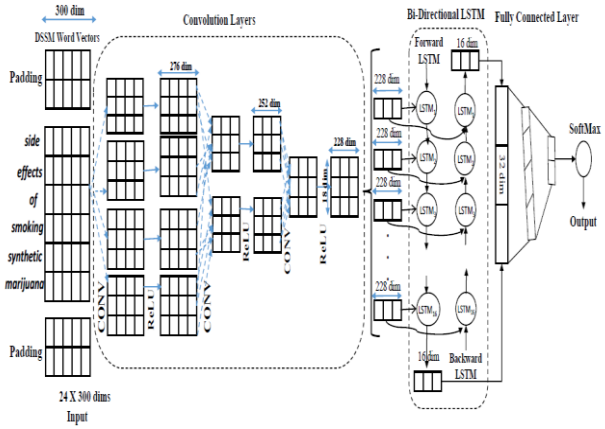

Fig. 1 Architecture of the Convolutional Bi-Directional LSTM (C-BiLSTM) Model

IV. INAPPROPRIATE CONVERSATION DETECTION IN CHAT DATA

Web platforms have grown suddenly, allowing users to share their opinion or comment publicly on specific individual / incident / community. They also offer the possibility to communicate with other users in natural language or with system. Some Gaming systems also allow users to communicate with each other automated machine that imitates human conduct in conversations.

These automated systems incorporate lot of human conversation training info, which may have been inappropriate along with clean chat interviews. The improper feeding of data into the systems helps them learn to generate improper responses to users that may reduce product confidence and decrease customer base. And it is the filtering of training data on automated machines and publicly available social media is mandatory networking sites to safeguard client faith and brand.

Bi-directional LSTM for Inappropriate Conversation Detection

BLSTM performs an input conversation and outputs the interaction likelihood that belongs to the class unacceptable. The series of characters trigrams which are taken from the chat between two users feedback is fed in the model with its one-hot portrayal.

The BLSTM model which is having the layer of three sequential layers namely "- a) Embedding Layer b) Bi-directional LSTM (BLSTM) Layer and c) Fully Connected (FC) layer". The given message which will be used as the input, the Embedding Layer which will be forming the new lower-dimensional feature representation for every character trigram in the input text message which are then sequentially given to the BLSTM layer. Then the BLSTM model will do the encoding of the sequential which will be having patterns in the input message in the both forward and reverse directions.

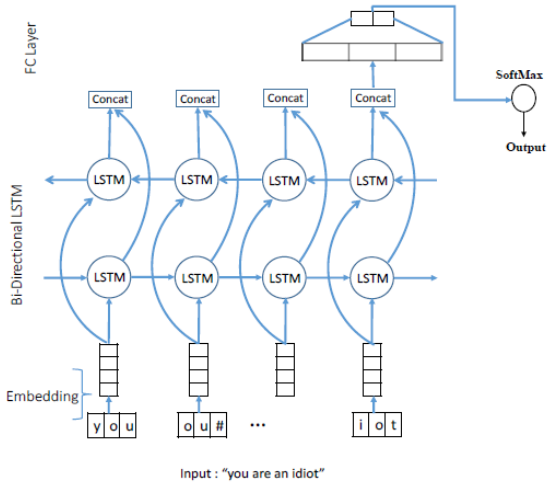

Fig.2 Architecture of the Bi-Directional LSTM (BLSTM) Model

Though LSTMs which is used in the last contexts only, Bi-Directional LSTMs (BLSTMs) tackle this constraint by analyzing this kind of inputs from both going forward and backward, and then combining the details from both ends to obtain a single results. This will help in the enabling of the to capture of much good patterns from both directions and this also keeps on helping in the learning of a much better feature representation for the input question.

The LSTM cells which are present in the inside of the forward and backward LSTM networks of BLSTM will read

International Journal of Engineering Applied Sciences and Technology, 2021
Vol. 6, Issue 4, ISSN No. 2455-2143, Pages 354-360
Published Online August 2021 in IJEAST (http://www.ijeast.com)

the character trigram representations which will be in the forward and reverse orders and each of these will output a 25 dimensional representation which is then combined to give a 50 dimensional feature representation which will be used in the encoding of the various semantic patterns in the given input.

V. SIAMESE FOR SIMILAR QUESTION RETRIEVAL

Siamese Convolutional Neural Network for cQA(SCQA) which will be containing the deep convolutional neural networks which is termed as the twin networks with a contrastive energy function at the top. These twin networks share the weights with each other (parameter sharing). SCQA discovers the parameters and similarities of a common model metric by reducing the binding power function of twin networks. Sharing parameters guarantees that question and its relevant answers in the semantic space are closer to one another while the question is and any trivial solutions to that are far from one another.

The representations, for example 'President of the United States' and 'Barack Obama' would be closer to one another than 'President' of the United States and "Tom Cruise lives in the United States." The taught similitude metric is used for semantic recovery related archive questions about a newly posted article.

Similar question pairs are needed for the training of the SCQA which is essentially the hardest part to get in the large numbers. Hence, SCQA achieves to overcome this constraint by doing the question-answer pairs from the cQA archives. Using question-answer training data which will lead to the richer concept or term associations in SCQA learning.

As mentioned earlier, Siamese Neural Networks are very common in Computer Vision verification of hand written signatures, face verification, visual embedding learning, etc. It is composed of A pair of weight sharing sub-networks that are connected at their outputs. The networks will map your input vectors to common semiconductor space. The joining node measures the distance in the semantic space which acts as the similarity metric.

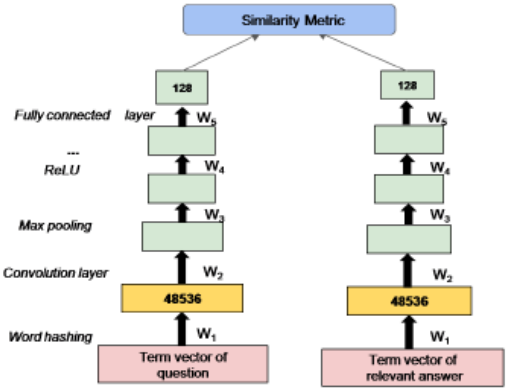

Fig.3 Architecture of SCQA.

Let, F(X) be a function group with parameter set W. F(X) is considered differentiable as for the W. Siamese network follows a parameter W value such that the symmetric value is similarity metrics are small if X1 and X2will be belonging to the same category and high if they belong to that of the different categories classes. In SCQA, question and relevant answer pairs are given as the input to this kind of models to train the network. The loss function is minimized so that S (Q; A) will be smaller if the answer A is relevant to the question Q or if it is not the case then it will bw in the larger otherwise. The entries to the SCQA twin networks are named as the hashed term vectors of a question-answer pair and their mark. And the result which is a label tries to indicate that the checked input which is a sample should be kept in the nearer or the farther place based on the semantic of the context. Twin networks for positive samples (which are expected to be closer in semantic space) are fed with word hashed question vectors and appropriate responses that are called "best-answer" or in Yahoo's cQA list, "Most Voted Answers" (question-relevant answer pairs). In testing, similar questions need to be retrieved for a given query. We do pairs during tests for all the query questions, and feed them to SCQA. Question Pair word vectors are term hashed and nourished twin sub-networks.

The SCQA project the educated shared weights vector problem in the checking domain. The sample results obtained by this will be made as a pair in order to check the similarities in the in the pairs if present. SCQA thus gives out a distance measure (score) which will be a value for all the

International Journal of Engineering Applied Sciences and Technology, 2021
Vol. 6, Issue 4, ISSN No. 2455-2143, Pages 354-360
Published Online August 2021 in IJEAST (http://www.ijeast.com)

pairs taken by the questions. And for each pair of the output values the threshold for questions is dynamically set to the average similarity score and it only issue those issues that are more similar than the average similarity ranking. After experimenting with many techniques of deep learning, we realized the question of seeking semantic

Relationship between the question-answer pairs can be summarized to the joint learning problem question-answer representations combined with the goal of minimizing the semantic gap between the two they do. Therefore we suggest a solution to the problem based on the Siamese network. The Architecture of SCQA contains strongly convolutionary neural networks as twin networks with an energy contrasting feature at the superior. These twin networks share the weights amongst themselves. SCQA knows the parameters of a shared model and metric similarity by reducing the power function linking twin networks. SCQA correlates the additional topics "to the question topic which the original askers may not even be aware of. For example, for the question "How to code deep neural networks using Python", an expert may give an answer containing the concepts "keras", "Theano", etc. Due to this, the concept deep neural networks using Python gets associated with these extended concepts as well". Hence, SCQA learns richer term associations.

TABLE I. SUMMARY OF PUBLISHED PAPERS

Reference	**Dataset Description**	**Algorithm used**	**Performance metric**
Vandersmissen et al. [1]	messages containing offensive language	Naive Bayes and Support Vector Machines(SVM) techniques	Naive bayes classifier with Precision 9% and Recall 86% performs very badly compare to SVM
Xiang et al. [2]	dataset for offensive message classification on twitter corpus	statistical topic modeling (LDA)	true positive rate (TP) of 75.1% over 4029 testing tweets
Xu et al. [3]	dataset by manually filtering over 11,000 text comments from the YouTube website	construct a Relation Tree(RelTree) using Parts-of-Speech(POS) and Typed Dependency Relations among words in a sentence.	ratios of accuracy of their implemented semantic filter is 90.94%.
Razavi et al. [4]	Used in the detection of the flames (offensive/abusive rants) from text messages	Complement Nave Bayes classifier for selecting the most discriminative features from a tokenized raw features.	keyword matching baseline, which has a TP of 69.7%,
Chuklin et al. [6]	Adultness of the query and manually created black and white list as the features	gradient boosted decision trees for classificat	-

VI. CONCLUSIONS

Nowadays it's important to recognize and filter inappropriate text from queries and conversations. This keeps the medium safer and therefore maintains the credibility of the company by not dragging it into legal position questions. Things such as chat bot and gaming systems aim to automate human actions and mimic it. Conversations that need a lot of data from the study. Many users intentionally supply unacceptable data to the systems and attempt to tarnish its name. And it's still useful to have an improper text filtering tackling these events.

The conversation data has few distinct features that distinguish them from the search queries. The length of the singleton conversation and use of smileys etc. didn't allow us to use the similar technique to treat data concerning conversations. Bi-directional LSTM is considered to be working better in classification conversations between Pattern Matching, Boosted Decision Tree and LSTM improper and clean modules.

We would like to build a standardized model for inappropriate text detection as part of future research this works with all data forms including questions, conversations etc. We want to expand the function as well constructing a multi-class classifier that determines the particular type of improper class to which the text belongs.

VII. REFERENCES

[1] F. D. T. Vandersmissen, Baptist and T. Wauters, "Automated Detection of Offensive Language Behavior on Social Networking Sites," pp. xiv, 81 p.: ill., 2012.

[2] G. Xiang, B. Fan, L.Wang, J. Hong, and C. Rose, "Detecting Offensive Tweets via Topical Feature Discovery over a Large Scale Twitter Corpus," pp. 1980–1984, 2012.

[3] Z. Xu and S. Zhu., "Filtering Offensive Language in Online Communities using Grammatical Relations," In Proceedings of the Seventh Annual CEAS 2010.

[4] A. H. Razavi, D. Inkpen, S. Uritsky, and S. Matwin, "Offensive Language Detection Using Multilevel Classification," in AI'10, pp. 16–27, Springer-Verlag, 2010.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," JMLR, 2003.

[6] A. Chuklin and A. Lavrentyeva, "Adult Query Classification for Web Search and Recommendation," in Search and Exploration of X-rated Information (WSDM'13).

[7] Z. Bar-Yossef and N. Kraus, "Context-Sensitive Query Auto-Completion," in WWW'11, (New York, NY, USA), pp. 107–116, ACM, 2011.

[8] M. Shokouhi and K. Radinsky, "Time-Sensitive Query Auto-Completion," in SIGIR '12, (New York, NY, USA), pp. 601–610, ACM, 2012.

[9] S. Whiting and J. M. Jose, "Recent and Robust Query Auto-completion," in WWW'14.

[10] F. Cai and M. de Rijke, "A Survey of Query Auto Completion in Information Retrieval," Foundations and Trends in Information Retrieval, vol. 10, no. 4, pp. 273–363, 2016.

[11] G. Di Santo, R. McCreadie, C. Macdonald, and I. Ounis, "Comparing Approaches for Query Autocompletion," in SIGIR '15, (New York, NY, USA), ACM, 2015.

[12] J. P. C. Chiu and E. Nichols, "Named Entity Recognition with Bidirectional LTM-CNNs," CoRR, vol. abs/1511.08308, 2015. 10

[13] P. Huang, X. He, J. Gao, L. Deng, A. Acero, and L. P. Heck, "Learning deep structured semantic models for web search using clickthrough data," CIKM, 2013.

[14] V. Nair and G. E. Hinton, "Rectified Linear Units Improve Restricted Boltzmann Machines," in ICML-10, June 21-24, 2010, Haifa, Israel, pp. 807–814, 2010. 12

[15] J. Friedman, T. Hastie, and R. Tibshirani, The Elements of Statistical Learning, vol. 1. Springer series in statistics Springer, Berlin, 2001.

[16] F. Pedregosa, G. Varoquaux, A. Gramfort, B. Michel, V.and Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[17] C. D. Manning, P. Raghavan, and H. Sch¨utze, Introduction to Information Retrieval. New York, NY, USA: Cambridge University Press, 2008.

[18] T. Zhang, "Solving large scale linear prediction problems using stochastic gradient descent algorithms," in Proceedings of the Twenty-first International Conference on Machine Learning, ICML '04, (New York, NY, USA), pp. 116–, ACM, 2004.

[19] Y. LeCun and F. J. Huang, "Loss functions for discriminative training of energy-based models," AISTATS, 2005.

[20] G. Zhou, L. Cai, J. Zhao, and K. Liu, "Phrase-based translation model for question retrieval in community question answer archives," ACL:HLT, 2011.