Effective Deep Neural Network Method based Sentimental Analysis for Social Media Health Care Information

D. Sasikala[1], Dr.S. Sukumaran[2]

[1]Research Scholar, Department of Computer Science, Erode Arts and Science College (Autonomous), Erode, India.
[2]Associate Professor, Department of Computer Science, Erode Arts and Science College (Autonomous), Erode, India.
[1]sasikalad@gmail.com

Article History Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 28 April 2021

Abstract: With the advent of natural language processing and machine learning techniques, Sentimental Analysis (SA) is receiving attention among various communities due to interpreting and classifying emotions from subjective data. The way of categorizing the emotions as positive, negative, and neutral enables the brand providers to know about their intention and reaction to their products. Existing researches have been focused on text summarization, feature reduction, and sentiment prediction separately. In this paper, all the approaches are integrated to provide a novel sentimental analysis framework for classifying the patients' emotions towards the medical facility through Twitter, Facebook, and other social media. The main aim of the work is manifold. The proposed paper performs preprocessing at the outset, which includes word segmentation, tokenization, TF-IDF, and stemming. First, the text summarization is performed by applying query-based summarization, which adopts a Deep Neural Network (DNN) for encoding the input data encompassing text and icon to convert into a fixed size at the final state, followed by decoding the encoded text and icon using the unidirectional DNN for creating the summary of the input document. Second, to improve scalability, feature reduction is performed using the effective Genetic Algorithm. The proposed system is experimented with using the real-time datasets from social media platforms to analyze the proposed healthcare sentimental analysis model. The parameters such as accuracy, F-score, recall rate, and precision are chosen to analyze the performance against existing sentimental analysis models.

Keywords: Sentiment Analysis, Deep Neural Network, Feature Reduction.

1. Introduction

Public opinion plays a vital role in providing valuable information based on the sentiments observed in social media platform. Challenge arises when its efficiency and accuracy of the sentimental analysis has been hindered by the natural Language Processing (NLP). Deep learning has not failed in proving promising solutions to the challenges faced by NLP. This paper highlights the recent studies that have been employed for solving the problems. The proposed work involves the usage of frequency-inverse document frequency (TF-IDF) and word embedding to the dataset used. This paper develop a scalable healthcare sentiment analysis framework, which is a query based summarization using Deep Neural Network (DNN). It summarizes the text and icons that are used in social media platforms. Sentiment analysis here analyses the patient’s opinion in social media platform based on their various emotions. Along with machine learning techniques similarity of results are collected [1]. The data source for sentiment analysis are online social media where users provide wide range of opinions and those serve as an input for sentimental analysis. The data has to be grouped into big data approach where it requires efficient data storage, reliable access and processing that ensures results are efficient [2]. Subjective or objective classification has to be performed. If it found to be subjective it will determine whether comment is positive or negative. It’s observed that there is no impacting difference between both document and sentence level classification since they are just short documents.

Recent studies have proposed sentiment analyses based on deep-learning that have differing features and performance. Sentiment polarity and deep learning models with TF-IDF along with word embedding to social media datasets has been analyzed.

Deep Learning

Deep Learning adapts a multilayered approach for the layers that are hidden in nature for neural networks. A conventional approach of machine learning defines and extracts the features manually or adapts some feature selection methods. Deep neural networks possess complex mathematical models for processing the data in required ways. It’s an adjustable model that comprises output in the form of functional inputs with several layers that include input data, hidden layers and neurons which are termed as processing nodes. The output layer also involves several neurons which are the network outputs.

Convolutional Neural Network (CNN)

Convolutionl Neural Network comes under the category of feed-forward neural network that is employed in various applications like computer vision, recommender systems and natural language processing. The network architecture is composed of layers that are subsampled for providing inputs to a fully-connected classification layer. While processing the convolutional layers filter the inputs for extracting the features as a result of which multiple filters in the output are combined. Pooling process can also be termed as subsampling layers where the resolution of the features are reduced increasing the robustness against noise and distortion. Classification task is performed by the layers that are fully connected.

This paper focuses on two objectives where a scalable Healthcare sentiment analysis framework is developed using Deep neural network and proposing Effective Genetic Algorithm for optimal feature selection. Thereby Deep learning model is combined with Effective Genetic Algorithm for developing an Improved Patient Sentimental Analysis Framework for predictive health care.

2. Related Works

The process of extracting information from an entity and identifying their subjectivity is termed as sentiment analysis. It captures the emotions from the text generated by the users whether they create a positive, negative or neutral impact on various applications. The levels of extraction mentioned in [3] are (i) lexicon- based (ii) machine learning based and (iii) hybrid approach. Initially lexicon based was used in sentiment analysis which is further divided into dictionary based and corpus based [4]. Former one is based on the terms used in dictionary and the later relies on statistical analysis of the contents available in the documents using k-nearest neighbor algorithm and hidden Markov models (HMM) [5]. The proposed machine learning based techniques consists of two models traditional and deep learning models [6]. The classifiers used are Naive Bayes and Maximum entropy classifier [7-8].

Query based summarization was proposed on early 90's where the work evaluates methods for extractive query-based summarization. Parse trees and Sentence compression was performed on query based summarization but it was not purely extractive based summarization [9]. Neural network model that dealt with query based summarization [10]. Dynamic Memory network was introduced [11] where state of the heart performance was depicted in the task of Natural Language Processing. Machine translation model was built which generated abstractive summaries on various datasets through Convolutional Neural network [12,13].

User's opinion on social behavior helps in analyzing the current and uprising prediction. Comprehensive overview of proposed algorithms with enhancements and applications was presented by Medhal [14]. For determining the polarity the score can be calculated by SentiWordNet. The method has outperformed machine learning methods with accuracy of 76.8% for feedback level, at sentence level accuracy is 86.6% methods. Genetic algorithms based optimization for feature selection is implemented.

3. Proposed Methodology

The deep learning models involve features that are learned and extracted automatically and maintaining efficiency and accuracy. The Fig1 below depicts the outset of preprocessing where word segmentation, tokenization, TF-IDF is involved. Text summarization is query based feature that adopts Deep Neural Network (DNN). It encodes the input data comprising of text and icon which adopts a Deep Neural Network (DNN) for encoding the input data encompassing text and icon to convert into a fixed size at the final state, followed by decoding the encoded text and icon using the unidirectional DNN for creating the summary of the input document.

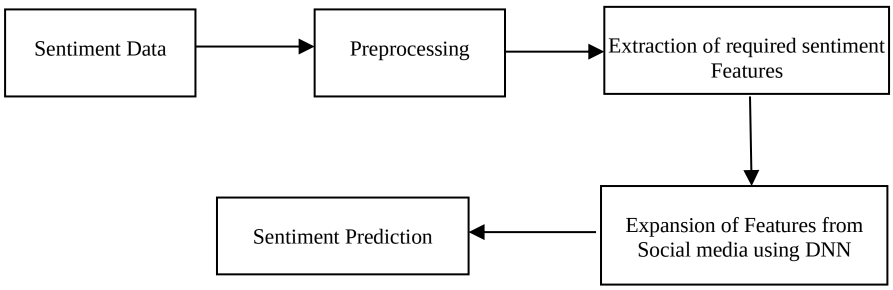


Fig. 1. Sentiment analysis through machine learning techniques

Important term that is used in Query based text and icon summarization is **Recognizing Named Entity and Gated Recurrent units (GRU)**. Extracting information is one among the task where structures information is extracted from various documents. The task is termed as named entity, it is classified into various categories based on which it may belong to either a person, location or no category. For example a name involved can consist of both a general name accompanied with a country name. It has to be divided and categorized into which it is related.

GRU belongs to neural network technique for exploiting the gradient problem that is a great hindrance from capturing dependencies on long term. It belongs to LSTM model and without any complicated computations [15,16]. The formulas described below depict the GRU architecture. X_t is defined as input at the time t, h_t is considered as an output. Both r_t and z_t represent the scaling vectors for generating the information in the form of gates as [0,1]. H_t is carry data with elements [-1,1].

$$r_t = \sigma(W^r[x_t, h_{t-1}] + b^r \quad (4.1)$$

$$z_t = \sigma(W^z[x_t, h_{t-1}] + b^z \quad (4.2)$$

$$h_t = GRU[h_{t-1}] + x_t \quad (4.3)$$

Embedding the Words

Consider a vocabulary where each word is encoded in a unique manner. This generates a vector of a particular length and every word is mapped to certain dimension available in the vector dimension. If word is mapped, the value is 1 and otherwise it is said to be 0. The vector has been transformed to embedding of the word and multiplying with embedded metric with particular dimensionality. It is the important parameter in adapted Deep neural Network. As a result of which two related words are expected to be in same vector space. The methods performed in word embedding are mentioned in [17,18].

Query based Text and Icon Summarization

A sequence model with pointer mechanism is introduced where the Input is a text through a query. The text is passed to the text encoder and a query encoder respectively. The output produced is transferred to decoder which generates the summary they make use of Deep neural network combined with GRU. Once GRU is occurred with a subscript with separate weights and biases.

Encoding the Text Query

The function of the encoder is to process the input query and generating a state for every input word. For obtaining a representation for the query convolutional neural network is used. At a particular time frame the intermediate states from both forward and backward reader is computed where a word from the query belongs to vocabulary and its reversed input is also taken into consideration.

Decoder

The decoder is CNN based and unidirectional for constructing the summary of input query. It depends on the final state of encoder in the input, it uses attention accompanied with pointer mechanism. The embedding query is considered as an input for the time process of decoder.

3.1 Query based Sentiment Analysis Framework

In query based summarization, input query and the related text documents are taken into consideration where the output is a summarized text. The process of summarization is to collect the sentences that are semantically related to the sentences given in the query. A score is obtained between a query sentence and the input sentence and that is considered for computing the relativity with the query[19]. Content word plays a vital role in delivering information in a sentence. While sentences are picked from a summary there is a high possibility that it can be conveyed in multiple meaning [20]. Eliminating the redundancy is an important task in text summarization. This reduces the size of the summary obtained as an output.

Extract the important sentence from the text document which is given as an input in prior to finding semantic relatedness score between the query and input text. Important sentences carry more vital information for the summary.

Query based Summarization using DNN Algorithm

Step 1: Selecting Important Sentence and content words are identified based on a prescribed criteria.
Step 2 : Word phrases are identified from the query along with their extended terms.
Step 3: Identifying word sense using Sentiword Net and obtain the score.
Step 4 : Identifying Sentence relativity score based on HSO
Step 5: Identify sentences that are redundant free selection for summary generation.
Step 6 : Generate the summary.

3.2 Proposed Effective Genetic Algorithm for Feature Extraction and Reduction

The proposed framework consists of all the features required for sentiment analysis. Genetic algorithm involves effective optimizations resulting in proper governing of the system. Entire framework is automated with various modules that are mentioned earlier along with which GA is applied in the part of feature reduction and extraction. Initially the process goes through query based summarization using deep neural network where redundant free data through query analysis has been calculated. Polarity score has to be obtained before the algorithm begins. It is determined by ontology of Sentiword Net. The score always ranges from 0 to 1(extremely negative to extremely positive). The stages involved are cleaning of data, pre- processing and analysis part. The tagger used is Maxent tagger from StansfordCoreNLP[16].

Feature Optimization

From entire set of words features are extracted through data structures which hold large feature size of the vector. The keywords used are associated with sentiment value included in feature vector. Drawback is it occur scalability issues in large dataset. To overcome this issue feature vector is optimized and size is reduced for maintaining the accuracy.

Calculating Fitness Value

When GA is applied binary string is chosen as chromosome. The fitness function which is a simple form of objective function is an essential part of feature selection using GA. The fitness function defined is attempted to provide optimal solution from first generation. The function depends on the distance from labeled sentiment value based on the distance with respect to polarity score from lexical database. Minimal the polarity distance between the label of the class and score the solution is more feasible for surviving in upcoming generation.

Fitness value determines the quality of the solution. The function used to evaluate the solution is the accuracy obtained for performing machine learning techniques. The steps involved in evaluating the result with Boolean vector is shown below:

(i) Generating index file with the words present in the solution.
(ii) Dataset is partitioned into two where major portion of it undergoes learning and minor version performs tests.
(iii) The model is studied by Multinomial Naïve Bayes algorithm.
(iv) The learned model is tested in dataset based on the words present in the solution.
(v) When fitness function is maximized accuracy is obtained by the model.
(vi) The accuracy calculated is computed below:

$$Accuracy = \frac{Total\ number\ of\ classified\ documents}{Total\ number\ of\ documents} \qquad (6.1)$$

Feature Selection Algorithm and Analysis

GA based feature selection algorithm is simulated for N number of generations as a result of which entire population converges for single optimal solution. For every generation GA works the way shown below:

Algorithm for Feature Selection

Step 1: *A finite list of tokens is considered where each consists of labelled sentiment value T. Optimal features are obtained as an output.*

Step 2 : *Initially randomly a seeded population is generated which has k number of generations and count value is assumed to be zero.*

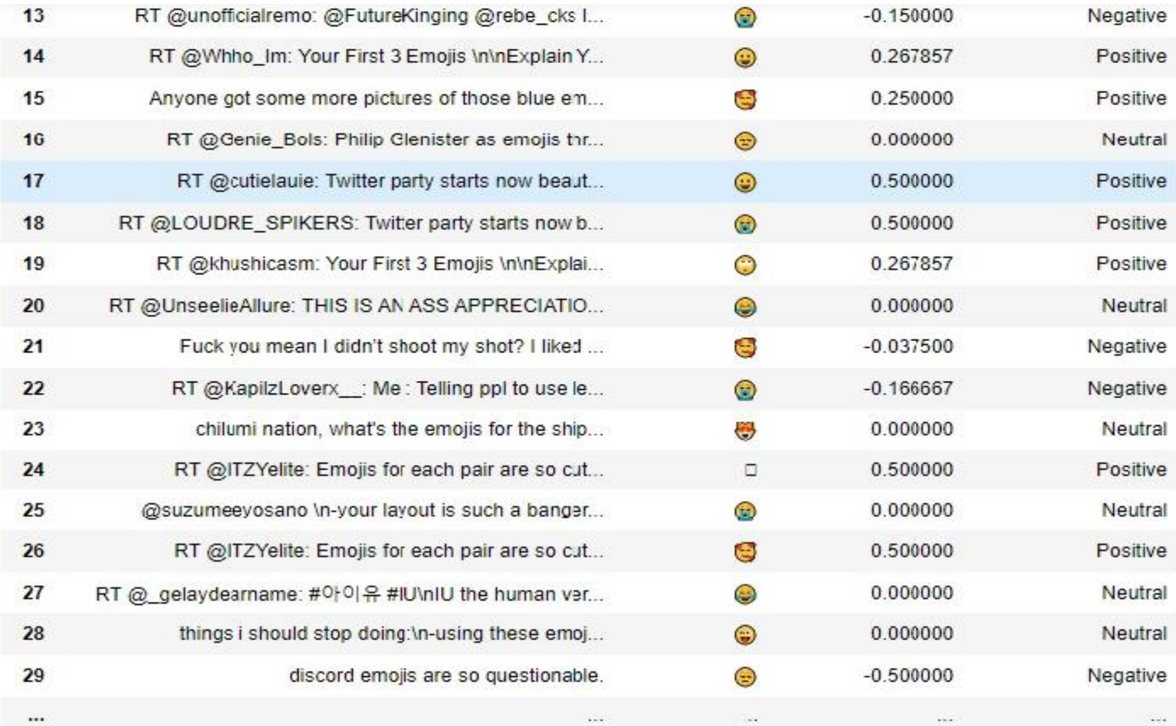
Step 3 : *While count<k, next generation is produced.*

Step 4 : *Get optimal features as result.*

This also includes crossover, mutation, generating offspring and evaluating fitness level. The time complexity of the applied GA is based on the fitness function. The ways adopted in general are generating new population one after the other until an optimal solution is obtained or fixing up or predefining the total number of generations in prior and converge to that particular level.

4. Experiments and Results

For evaluating various approaches on this framework use UCI ML dataset that consists of three different social media platform twitter, instagram and Facebook where the user share their opinions on various hospitals and their treatment procedures through comments as well as emojis. The framework was built in Python language. From the experimentations results using deep neural networks in sentiment classification on the dataset with normalization and eliminating diacritics the feature space has reduced. Decision tree algorithm, a best classifier in deep neural network has been adopted. Below figure depicts how the classifier has classified the comments and emojis into positive, negative and neutral.



13	RT @unofficialremo: @FutureKinging @rebe_cks I...		-0.150000	Negative
14	RT @Whho_Im: Your First 3 Emojis \n\nExplain Y...		0.267857	Positive
15	Anyone got some more pictures of those blue em...		0.250000	Positive
16	RT @Genie_Bols: Philip Glenister as emojis thr...		0.000000	Neutral
17	RT @cutielauie: Twitter party starts now beaut...		0.500000	Positive
18	RT @LOUDRE_SPIKERS: Twitter party starts now b...		0.500000	Positive
19	RT @khushicasm: Your First 3 Emojis \n\nExplai...		0.267857	Positive
20	RT @UnseelieAllure: THIS IS AN ASS APPRECIATIO...		0.000000	Neutral
21	Fuck you mean I didn't shoot my shot? I liked ...		-0.037500	Negative
22	RT @KapilzLoverx__: Me : Telling ppl to use le...		-0.166667	Negative
23	chilumi nation, what's the emojis for the ship...		0.000000	Neutral
24	RT @ITZYelite: Emojis for each pair are so cut...	□	0.500000	Positive
25	@suzumeeyosano \n-your layout is such a banger...		0.000000	Neutral
26	RT @ITZYelite: Emojis for each pair are so cut...		0.500000	Positive
27	RT @_gelaydearname: #아이유 #IU\nIU the human ver...		0.000000	Neutral
28	things i should stop doing:\n-using these emoj...		0.000000	Neutral
29	discord emojis are so questionable.		-0.500000	Negative
...	...	...	...	...

Fig. 2. Analysis of Comments Observed in Social Media Platform

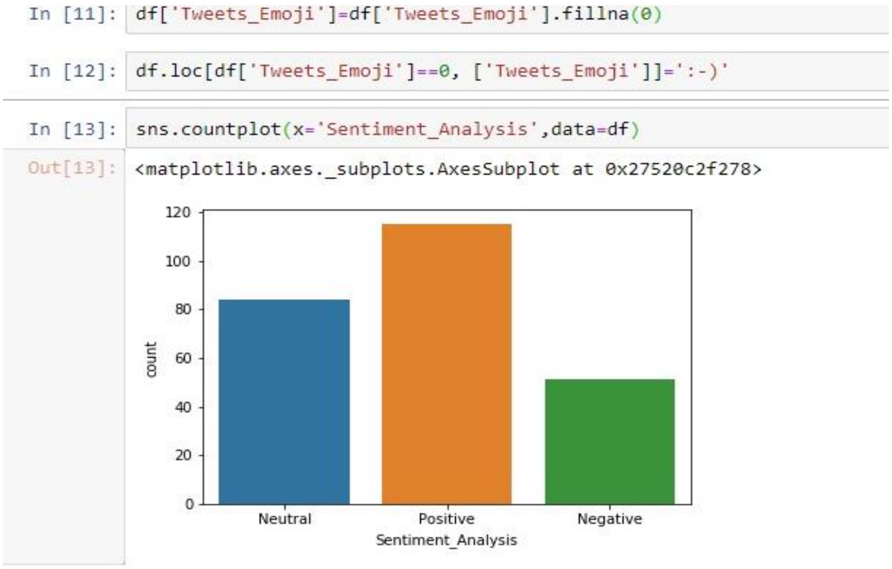


Fig. 3. Classification Result

Table 1. Comparison Table for Accuracy, F-score, Recall Rate and Precision

Methods	**Accuracy**	**F-score**	**Recall**	**Precision**
MNB	0.83	0.79	0.71	0.73
Decision tree	0.75	0.72	0.69	0.76
SVM	0.81	0.81	0.78	0.63
Deep Neural with GA	**0.96**	**0.93**	**0.92**	**0.95**

5. Conclusion

Experimentation show that feature selection with Effective GA overcomes the result of the classifier. Query based summarization with GA has achieved 96% accuracy in feature reduction compared to other machine learning algorithms. The paper has described query based summarization using Deep Neural Network and the outcome of it undergoes Effective Genetic algorithm supported feature extraction and classification achieving more accuracy is proved through experimentation. The result proves that accuracy, f-score, recall rate and precision of Deep neural network with GA is more better than other deep learning techniques that include MNB, Decision tree and SVM.

References

1. Pouli, V., Kafetzoglou S., Tsiropoulou E.E., Dimitriou A., & Papavassiliou, S. “Personalized multimedia content retrieval through relevance feedback techniques for enhanced user experience” International Conference on Telecommunications (ConTEL), Graz, Austria, 1–8, 2015.
2. Thai T., Weili Wu., & Xiong H., “Big Data in Complex and Social Networks”, CRC Press: Boca Raton, FL, USA, 2016.
3. Bhavitha B., Rodrigues A.P. & Chiplunkar N.N., “Comparative study of machine learning techniques in sentimental analysi”, in International Conference on Inventive Communication and Computational Technologies (ICICCT), 216–221, 2017.
4. Medina-Moreira J., Lagos-Ortiz K., Luna-Aveiga H., Rodriguez-Garcia M.A., & Valencia-García R., “Sentiment analysis on tweets about diabetes: An aspect-level approach”, Journal of *Computational and Mathematical Methods in Medicine*, 2017**.**
5. Huq M.R., Ali A. & Rahman A., “Sentiment analysis on Twitter data using KNN and SVM”, International Journal of Advanced Computer Science and Applications, 8(6), 19– 25, 2017.
6. Soni S. & Shara A., “Sentiment analysis of customer reviews based on hidden markov model” In Proceedings of International Conference on Advanced Research in Computer Science Engineering & Technology, 1-5, 2015.
7. Malik V., Kumar Amit, “Sentiment Analysis of Twitter Data Using Naive Bayes Algorithm”, International Journal on Recent and Innovation Trends in Computing and Communication, 6(4), 120–125, 2018.
8. Wu H., Li J., & Xie J., “Maximum entropy-based sentiment analysis of online product reviews” in Chinese. In Automotive, Mechanical and Electrical Engineering, CRC Press, 559-562, 2017.

9. Lu Wang, Hema Raghavan, Vittorio Castelli, Radu Florian and Claire Cardie, “A sentence compression based framework to query-focused multi document summarization” In ACL, 2013.
10. Preksha Nema, Mitesh Khapra, Anirban Laha, and Balaraman Ravindran, “Diversity driven Attention Model for Query-based Abstractive Summarizatio”, ACL, 2017.
11. Ankit Kumar, Ozan Irsoy, Peter Ondruska, Mohit Iyyer, James Bradbury, Ishaan Gulrajani, Victor Zhong, Romain Paulus, and Richard Socher, “Ask me anything: dynamic memory networks for natural language processing”, In International Conference on Machine Learning. PMLR, 48, 1378-1387, 2016.
12. Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, “Neural machine translation by jointly learning to align and translate”, ICLR, arXiv:1409.0473, 2015.
13. Caglar Gulcehre, Sungjin Ahn, Ramesh Nallapati, Bowen Zhou and Yoshua Bengio, “Pointing the unknown words”, In Proceedings of the Association for Computational Linguistics, 140–149, 2016.
14. W. Medhat, A. Hassan, and H. Korashy, “Sentiment analysis algorithms and applications: A survey,” Ain Shams Engineering Journal, 5(4) 4,1093–1113, 2014.
15. Junyoung Chung, Caglar Gulcehre, Kyung Hyun Cho, and Yoshua Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling”, arXiv:1412.3555, 2014.
16. M. McCandless, E. Hatcher, and O. Gospodnetic, “Lucene in Action: Covers Apache Lucene 3.0”, Manning Publications Co., 2010.
17. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean, “Distributed representations of words and phrases and their compositionality”, In Proceeding of NIPS, 3111–3119, 2013.
18. Jeffrey Pennington, Richard Socher, and Christopher D. Manning, “Glove: Global vectors for word representation”, In Empirical Methods in Natural Language Processing, 1532–1543. 2014.
19. Nazreena Rahman, and Bhogeswar Borah, “Improvement of query-based text summarization using word sense disambiguation”, Complex & Intelligent Systems, 75–85, 2020.
20. Niwattanakul S, Singthongchai J, and Wanapu S, “Using of jaccard coefficient for keywords similarity”, In Proceedings of the International Multi Conference of Engineers and Computer Scientists, 380–384, 2013.
21. C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, “The Stanford Core NLP natural language processing toolkit,” in Proceedings of the Association for Computational Linguistics: System Demonstrations, 55–60, 2014.