IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 3, March 2022

DOI: 10.17148/IJARCCE.2022.11348

Comparative Study on Sentiment Analysis on IMDB Dataset

Debarghya Banerjee[1], Sreya Mazumder[2], Samik Datta[3]

[1]Undergraduate, Department of Computer Science and Engineering, Techno Engineering College Banipur, West Bengal, India

[2]Undergraduate, Department of Computer Science and Engineering, Techno Engineering College, Banipur, West Bengal, India

[3]Assistant Professor, Department of Computer Science and Engineering, Techno Engineering College, Banipur, West Bengal, India

Abstract: The main part of information gathering is to find out what other people think. In case of movies, the movie reviews can provide an in-depth and detailed understanding of the movie and can help decide whether it is worth watching or not. However, with the growing amount of data in reviews, it is quite prudent to automate the process, saving a lot of time. Sentiment analysis is an important field of study in machine learning that practically deals with extracting useful information of subjects from the textual reviews. The sentiment analysis is closely related to Natural Language Processing (NLP) and text mining. It is used to determine the sentiments of the reviewer in regard to various topics or the overall polarity of the review. In case of movie reviews, along with giving a rating in numeric to a movie, they can give us information on the approval or the disapproval of a movie quantitatively. A collection of those information then gives us a comprehensive qualitative understanding on different facets of the movie. Since human language is complex, we face many kinds of challenges during opinion mining from movie reviews which might leads us to situations where a positive word has a negative connotation and vice versa.

Keywords: Machine Learning, Natural Language Processing, Text Mining, Opinion Mining, Analysis of Sentiments, Extracting Information.

1. INTRODUCTION

Sentiment analysis, which is also known as opinion mining, deals with the analysis of people's opinions, sentiments, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, and their attributes. "The views, opinion, choices of other people" has always played an important role during the decision-making process of recommendation systems, like recommendation of goods in Amazon or Walmart is based on the sentiment analysis of the user reviews.

For the area of analysis of movie review, sentiment analysis means finding the mood of the public about how do they judge a specific movie. For details, the documents, from the user reviews, are classified based on the mood they are expressing, such as positive, somehow positive, neutral, somehow negative and negative. Sentiment analysis uses text mining, Natural Language Processing (NLP), and other computational techniques.

The sentiment scores and the sentiment polarities are the two major aspects for sentiment analysis. Sentiment polarities can be represented as a binary value (0,1) which can be positive or negative. A model known as Parts of Speech (POS) tagging which tags language grammar, particularly (noun, adverbs, verbs, and adjectives). The main purpose of sentiment analysis is to evaluate all opinions to find out the cumulative polarity of reviews for concerned topics based on their levels of classification, for instance negative or positive. Existing concurrences of available reviews could be divided into three levels:
Sentence level, Document level, and Entity/Aspect level.

Sentence level analysis the sentiment on each sentence. Document level classifies the entire document as binary class or multi-class. Entity/Aspect level is a more complicated level, which is identifying the different aspects of a corpus first, and then classifying each document with respect to the observed aspects of each document.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 3, March 2022

DOI: 10.17148/IJARCCE.2022.11348

2. LITERATURE SURVEY: -

Related Works:

Earlier works on sentiment classification using machine learning approaches were carried by Pang et al. in 2002 [18]. It was performed on IMDb reviews using n-gram approaches and Bag of Words (BOW) as features. The model was trained using different classifiers like Naïve Bayes (NB), Maximum Entropy (ME), and Support Vector Machines (SVM). The unigram features outperformed when compared to other features. Another similar kind of work was done by Tripathy et al. [19], where TF, TF-IDF was used as feature. Experimentation was done with n-gram approach and its combination were tried to get the best results.

Apart from the word features which are considered for the classification task, special symbols known as emoticons can also be used as features. Neetu et al [20] used these special features along with the word features. The use of an ensemble classifier which classifies based on the results obtained by different classifiers like NB, ME and SVM is the major highlight of the work. Many researchers have worked on extracting features based on the parts of speech tagger. Geetika et al [21], used a unigram model to extract adjective as a feature to describe the positivity or negativity of the sentence.

The challenging task for a machine learning algorithm is to identify the semantics or the meaning of the text. Lexicon features were used in this regard to extract the opinions expressed in the text. The major advantage of choosing lexicon features was sarcasm detection. Anukarsh [22] et al. focused on the slangs and emojis which were present in the text to detect sarcasm. The efficiency of sarcasm detection was increased by using slang and emoji dictionaries during pre-processing. Taboada et al. [23], used dictionaries to calculate the Semantic Orientation (SO) and termed it as Semantic Orientation Calculator (SO-CAL). Various factors such as Parts of Speech (Adjectives, Nouns, Verbs and Adverbs), Intensifiers (Somewhat, Very, Extraordinary etc.,), Negations, etc., were considered to calculate sentiment orientation. Results showed that the Lexicon based sentiment analysis gives better results and can be applied to wide domains.

Problem Statement:

To analyze the movie review sentiments on the IMDb dataset and to conduct a comparative analysis of well-known classifiers where the classification accuracy can be improved by extracting the significant features and aspects from the contexts and to propose the best classification model for sentiment analysis on this particular dataset.

3. METHODOLOGY: -

3.1. Pre-processing: -

The gathered movie reviews which need to be analysed consist of words, numbers, and special symbols as its constituents. This stage is performed by techniques like lower case conversion, HTML tags removal, emoji removal, expanding the contraction, punctuation removal, number removal, URL removal, stop words removal, and lemmatization.

3.1.1. Lower case conversion:

The “lower case letters” are more often utilized by users. The “lower case letters” are mostly viewed when compared with the uppercase letters due to the common utilization. Thus, the conversion of the uppercase letters to the lowercase letters is necessary and provide a simpler classification of the sentiments in this proposed system.

3.1.2. Html tags removal:

Another common pre-processing technique that comes handy in multiple places is removal of html tags. This is especially useful when we scrap the data from different websites.

3.1.3. Emoji removal:

With more and more usage of social media platforms, there is an explosion in the usage of emojis in our day-to-day life as well. The gathered movie reviews include emojis and thus, it has to be removed for enhancing the efficiency of the model.

3.1.4. Expanding the contraction:

Contractions are words or combinations of words that are shortened by dropping letters and replacing them by an apostrophe. Expanding contractions contributes to text standardization and is useful when we are working on movie reviews as the words play an important role in sentiment analysis.

3.1.5. Punctuation removal:

The gathered movie reviews include more symbols like “.” & “,” and thus, it has to be reduced for enhancing the efficiency of the model.

3.1.6. Number removal:

The gathered movie reviews include numbers and thus, it has to be removed for enhancing the efficiency of the model.

3.1.7. URL removal:

Next pre-processing step is to remove any URLs present in the data. The movie reviews contain URLs of other websites or social handles of the reviewer. We need to remove them for our further analysis.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 3, March 2022

DOI: 10.17148/IJARCCE.2022.11348

3.1.8. Stop word removal:

It is employed for removing the most often used words like adverbs, conjunction, prepositions and article, which results in reduction of the dimensionality of the datasets. Some of the words are 'they', 'she', 'but', 'he', 'if' and 'we', so on. Thus, the stop words must be eradicated to enhance the quality of sentiment analysis.

3.1.9. Lemmatization

Lemmatization is the process of switching any kind of a word to its base or root form. In other words, Lemmatization basically converts different inflected forms of words, having same meaning into their root form. It gives the stripped word that has some meaning in the dictionary. It clearly identifies the base form of 'troubled' to 'trouble'' denoting some meaning.

3.2. Dataset Description: -

For the sentiment analysis on IMDB movie reviews, we have collected the input data from "https://www.kaggle.com/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews". This dataset contains 50k movie reviews. It includes 25,000 positive movie reviews and 25000 negative movie reviews. The review column is having 49582 unique values and the sentiment column consist of two unique values which are positive and negative respectively. In our analysis, we have a large amount of data for one class, for that reason the whole dataset has been divided into two parts, where 80% of data have been used for training and 20% of data have been used for testing.

3.3. Feature Extraction: -

Feature Extraction [10] identifies the features that have a positive effect towards classification. In this work, feature extraction is carried in two different parallel stages namely- Machine learning based feature extraction and Lexicon based feature extraction.

Machine learning based feature extraction method used to extract the features using popularly known technique Bag of Words, wherein the column corresponds to words and row corresponds to value of weighing measures such as Team Frequency (TF) and Team Frequency-Inverse Document Frequency (TF-IDF).

3.3.1. Bag of Words:

The Bag of Words is a Natural Language Processing approach to text modelling. The Bag of Words model is one of the most basic yet effective methods for extracting characteristics from text texts. A bag of words is a text representation that describes the frequency with which words appear in a document.

3.3.2. TF-IDF:

The TF-IDF (term frequency-inverse document frequency) statistic examines the relevance of a word to a document in a collection of documents. Because the TF-IDF weights words according to their significance, it may be used to discover which words are the most essential. This may be used to more quickly summarize articles or just identify keywords (or even tags) for a document.

3.4. Classification: -

Classification is the process of assigning labels to the reviews whose label is unknown. In the proposed work, K-Nearest Neighbour Classifier (KNN) [14], Decision tree classifier [10], Random Forest classifier [12], Logistic Regression [11] and SVM [14] are used.

3.4.1. K-Nearest Neighbour Classifier (KNN):

KNN [14] can be utilized in both regression and classification tasks, which is a simple and easy technique. It employs the principle of local approximation, which is also explored its neighbours and also analysed them. Therefore, it predicts the value through lazy learner. KNN performs by measuring the distance among the neighbour and selected the neighbour with lesser distance during the classification tasks. However, KNN is considered as the slowest classifier due to the direction relation of classification time is with the amount of data. It performs prediction by measuring the distance among the query point and the context from the samples, which is carried out through Euclidean distance. It classifies the texts through training samples and attributes. At last, the KNN predicts the classes into negative or positive sentiments.

3.4.2. Decision Tree Classifier:

A hierarchical structure is followed in DT [10], which can be applied for handling the classification and regression problems. It also works by considering the tree structure, in which the labelling of each node is done with a certain condition. Before giving the node as negative, the condition of every node is tested. This procedure is continued till it attains the leaf node and then, it predicts the output. It also works by considering the independent variables, where the classification is performed by using Gini-index. Finally, for the next iteration, the attributes with maximum Gini index value are chosen. Further, the DT has predicted the negative or positive sentiments.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 3, March 2022

DOI: 10.17148/IJARCCE.2022.11348

3.4.3. Random Forest Classifier:

The RF [12] classifier is a supervised learning algorithm which you can use for regression and classification problems. The Random Forest classifier randomly selects a subset of the training set and then creates a set of decision trees based on it. It then collects the votes from different decision trees to decide the final prediction.

3.4.4. Logistic Regression:

LR [11] is generally a supervised classification algorithm. In a classification problem, the target variable (or output), y, can take only discrete values for a given set of features (or inputs), X. It is a regression model. The model predicts the probability by building a regression model. It predicts whether a given data falls under the category numbered as "1". Just like Linear regression assumes that the data follows a linear function, Logistic regression models the data using the sigmoid function- $\mathbf{g(z) = 1/ (1 + e^{-z}).}$

3.4.5. Support Vector Machine (SVM):

SVMs [14] are one of the machine learning classification approaches that employs kernel function for mapping a space of data points. The SVM gives better performance when compared to the other classification approaches because of its features like better handling of huge scale features. However, it is robust while there is a sparse set of contents as it is not linearly separable one. It offers high accuracy, ability of dealing with non-linear data, handles over fitting. Moreover, SVM is more suited for both regression and classification. It has also several tuning constraints like regularization, gamma, margin and kernel.

The support vector machine algorithm used in scikit-learn is implemented using stochastic descent to converge on a solution. We can also access the SGD algorithm directly through a separate implementation in scikit-learn, SGD Classifier. SGD Classifier provides a handful of different loss functions, but of these, loss="hinge", the default, is the one which causes SGD Classifier to perform equivalently to a linear SVM.

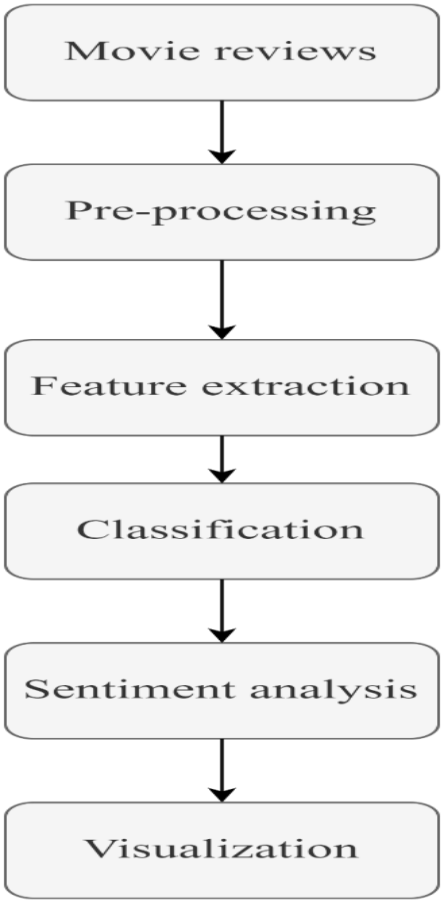


Fig. - Flowchart

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 3, March 2022

DOI: 10.17148/IJARCCE.2022.11348

.4. RESULT AND ANALYSIS:

- **Analysis on classifiers: -**

❖ Comparative Analysis of different Classification Algorithm: -

Algorithm	Performance	Pros	Cons
KNN	It is very popular in NLP analysis. It can parse sentence from large corpora.	1. Very Easy to implement for multi Class problem 2. Easy to implement	1.Performance degrades if the dataset size is increased 2. Outlier sensitivity
DT	Perform well for small datasets. Depends on Splitting Criteria, Splits, stopping criteria	1. It is good enough to represent any discrete value classifier 2. Its Self-explanatory	1.It examines single field at a time 2.Since it follows divide and Conquer strategy. Performance degrades if the number of features are high
RF	The Random Forest classifier randomly selects a subset of the training set and then creates a set of decision trees based on it.	1. Works well on large datasets. 2. Can be used to extract variable importance.	1. Overfitting in case of noisy data. 2. Hyperparameters need good tuning for high accuracy.
LR	The model predicts the probability by building a regression model. It predicts whether a given data falls under the category numbered as “1”.	1. One of the simplest machine learning algorithms 2. Updated easily to reflect new data	1. Difficult to capture complex relationships 2. Repetition of information could lead to wrong training of parameters
SVM	SVM creates the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future.	1. It is effective in high dimensional spaces. 2. It uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.	1. It doesn’t perform well when we have large data set because the required training time is high 2. It also doesn’t perform very well, when the data set has more noise i.e. target classes are overlapping

The performance analysis on suggested sentiment analysis model is evaluated with different classifiers. From the performance evaluation, the LR classifier gets superior results while comparing with other classifiers for all the performance measures. The accuracy of LR is 74.13% superior to KNN, DT, RF and SVM respectively. Similarly, the LR classifier on sentiment classification gets higher performance while comparing with other approaches in terms of all the performance metrics.

CLASSIFIER	ACCURACY
KNN	50.29999%
Decision Tree	49.55999%

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 3, March 2022

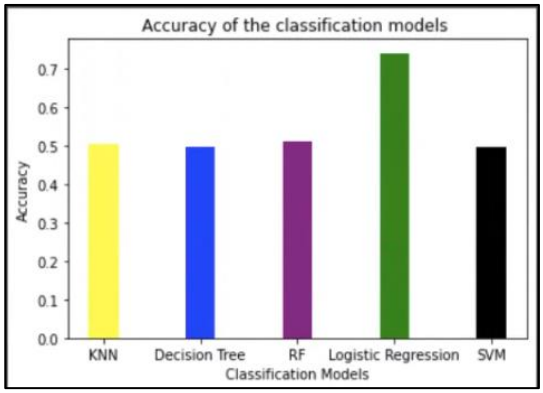
DOI: 10.17148/IJARCCE.2022.11348

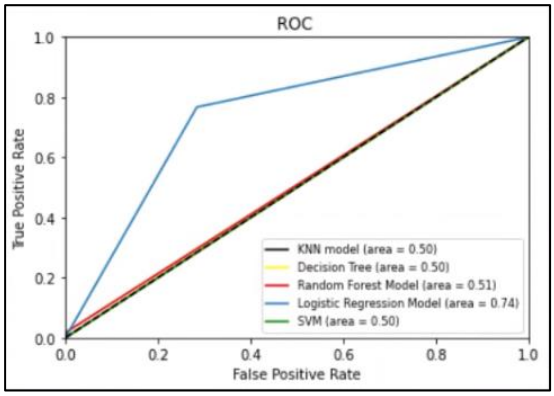
Random Forest	51.25999%
Logistic Regression	74.13333%
SVM	49.57777%

CLASSIFIER	SCORE
KNN	0.50
Decision Tree	0.50
Random Forest	0.51
Logistic Regression	0.74
SVM	0.50

- **Analysis on ROC: -**

The efficiency of the designed sentiment analysis model with different classifiers is evaluated by taking the false positive rates with true positive rates. The ROC [17] also helps us in finding out the accuracy of the model and sometimes it gives a better accuracy rate.





5. CONCLUSION: -

From the collected dataset, the pre-processing was performed through “lower case conversion, HTML tags removal, emoji removal, expanding the contraction, punctuation removal, number removal, URL removal, stop words removal, and lemmatization”. Then, the opinion words were gathered and the features were extracted by computing the polarity score. The machine learning algorithms like “KNN, DT, RF, LR and SVM” were used for predicting the sentiments. The performance of the suggested model was analysed with standard performance measures by evaluating all the classifiers. Through the performance analysis, we can conclude that by using TF-IDF as features, the accuracy of the designed sentiment analysis model using LR was having an accuracy of 74.13%, sensitivity of 0.8330689 and specificity of 0.7671509 which is better than KNN, DT, RF and SVM respectively. Thus, the designed model with LR has attained better superior performance while comparing with other classification methods. In our proposed method, the comparative analysis has been done only on machine learning algorithms. In future, a comparison of the performance analysis of deep learning algorithms will be conducted and that might result in the improvement of the accuracy of the models.

6. REFERENCES: -

1. B. Liu, "Sentiment analysis and opinion mining", Synthesis Lectures on Human Language Technologies, vol. 5, no. 1, pp. 1-167, May 2012.

DOI: 10.17148/IJARCCE.2022.11348

2. S. Poria, E. Cambria, G. Winterstein and G.-B. Huang, "Sentic patterns: Dependency-based rules for concept-level sentiment analysis", Knowledge-Based Systems, vol. 69 pp. 45–63, 2014.
3. M.Z. Asghar, A. Khan, S. Ahmad, F.M. Kundi, "A Review of Feature Extraction in Sentiment Analysis", Journal of Basic and Applied Scientific Research, vol. 4 pp. 181–186, 2014.
4. Cambria, E, "Affective computing and sentiment analysis", IEEE Intelligent Systems, vol. 31, pp. 102–107, 2016.
5. Chen, P., Sun, Z., Bing, L., and Yang, W., "Recurrent attention network on memory for aspect sentiment analysis", in Proceedings of the 2017 conference on empirical methods in natural language processing[2], pp. 452–461, 2017.
6. Asghar MZ, Khan A, Ahmad S, Kundi FM 2014 A Review of Feature Extraction in Sentiment Analysis, J Basic and Applied Sci Res, 4: 181–186.
7. Fu X, Liu W, Xu Y and Cui L 2017 Combine How Net lexicon to train phrase recursive autoencoder for sentence-level sentiment analysis, Neurocomputing, 241: 18-27.
8. Cambria E 2016. Affective computing and sentiment analysis, IEEE Intel Syst, 31: 102–107.
9. S.Shayaa et al., "Sentiment Analysis of Big Data: Methods, Applications, and Open Challenges," IEEE Access, vol. 6, pp.37807-37827, 2018.
10. Samik Datta, Satyajit Chakrabarti, 'Evaluation of Deep Learning Approaches for Aspect Based Sentiment Analysis on Movie Dataset', Proceedings of the Fourth International Conference on Smart Systems and Inventive Technology (ICSSIT-2022) January 2022, Pages 970 - 979.
11. Prajval Sudhir, Varun Deshakulkarni Suresh "Comparative study of various approaches, applications, and classifiers for sentiment analysis, "Global Transitions Proceedings, Volume 2, Issue 2, Pages 205-211, November 2021.
12. Annamalai Suresh and C R Bharathi "Sentiment Classification using Decision Tree Based Feature Selection, " International Journal of Control Theory and Applications, vol. 9, issue. 36, pp. 419-425, January 2016.
13. Ashwin Sanjay Neogi, Kirti Anilkumar Garg, Ram Krishn Mishra, Yogesh K Dwivedi, "Sentiment analysis and classification of Indian farmers' protest using twitter data," International Journal of Information Management Data Insights, Vol. 1, Issue. 2, November 2021.
14. Mohammad Rezwanul Huq, Ahmad Ali and Anika Rahman, "Sentiment Analysis on Twitter Data using KNN and SVM," International Journal of Advanced Computer Science and Applications, Vol. 8, No. 6, 2017.
15. Alqaryouti, N. Siyam, K. Shaalan, "A Sentiment Analysis Lexical Resource and Dataset for Government Smart Apps Domain", Proceedings of the International Conference on Advanced Intelligent Systems and Informatics on Advanced Intelligent Systems and Informatics 2018 (AISI2018), Springer, Cham, 2019.
16. Pranali Borele and Dili Kumar A. Borikar, "An Approach to Sentiment Analysis using Artificial Neural Network with Comparative Analysis of Different Techniques,"IOSR Journal of Computer Engineering (IOSR-JCE), Vol. 18, Issue. 2, pp. 64-
17. Vandewiele G, Dehaene I, Kovács G, Sterckx L, Janssens O, Ongenae F, VanHoecke S (2020) Overly optimistic prediction results on imbalanced data: flaws and benefits of applying over-sampling. Preprint at https://arxiv.org/abs/quant-ph/2001.06296.
18. B. Pang, L. Lee, and S. Vaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques." In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, Association for Computational Linguistics, pp. 79-86. 2002.
19. A. Tripathy, A. Agrawal, and S.K. Rath. "Classification of sentiment reviews using n-gram machine learning approach." Expert Systems with Applications, Vol. 57, pp. 117-126. 2016.
20. M. S. Mubarok, Adiwijaya, and M. D. Aldhi. "Aspect-based sentiment analysis to review products using Naïve Bayes." In AIP Conference Proceedings, vol. 1867, AIP Publishing, no. 1, pp 1-8.2017.
21. G. Gautam, and D. Yadav. "Sentiment analysis of twitter data using machine learning approaches and semantic analysis." In Contemporary computing (IC3), 2014 seventh international conference on, pp. 437-442. IEEE, 2014.
22. A. G. Prasad, S. Sanjana, S. M. Bhat, and B. S. Harish. "Sentiment analysis for sarcasm detection on streaming short text data." In Knowledge Engineering and Applications (ICKEA), 2017, 2nd International Conference on, pp. 1-5. IEEE, 2017.
23. M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede. "Lexicon-based methods for sentiment analysis." Computational linguistics, Vol. 37, no. 2, pp.267-307. 2011.