Enhancing Text Classification in Low-Resource Languages: A Modified TF-IDF Approach for Effective Sindhi Text Categorization

Muhammad Hamza[1], Muhammad Owais Raza[2],
[1]Department of Telecommunication Engineering
Quaid-e-Awam university of Engineering and Technology Nawabshah
hamzaalee111@gmail.com

[2]Department of Software Engineering
Mehran university of Engineering and Technology jamshoro,
owais.leghari@hotmail.com

ABSTRACT

This study seeks to address the problem of text classification in low-resource languages, with a particular focus on Sindhi. The main goal is to create an effective classification model using machine learning techniques and a modified term weighting system. To accomplish this aim, the study makes use of a Sindhi dataset obtained from the Awami Awaz website. To classify text in Sindhi, we used classical machine learning techniques. The significant innovation in this study is the introduction of a modified TF-IDF word weighting strategy, which improves the discriminative strength of features for low-resource languages. This method addresses the distinct linguistic peculiarities of Sindhi text. Our findings show that the proposed modified term weighting method outperforms the traditional TF-IDF strategy for Sindhi text classification. The usefulness of our approach in effectively classifying Sindhi articles was demonstrated by training and evaluating classical machine learning models. To conclude, this work focuses on the text categorization issues associated with low-resource languages, notably Sindhi. When compared to typical TF-IDF approaches, we achieved significant improvements in classification accuracy by introducing a modified term weighting scheme into the machine learning pipeline. This study lays the path for more accurate and efficient text categorization in low-resource languages, which would assist applications like information retrieval, sentiment analysis, and content recommendation in Sindhi and other comparable languages.

Keywords: Machine Learning, NLP, Low Resource Language, Sindhi And TFIDF.

1. INTRODUCTION

Language serves as the cornerstone of human communication, enabling us to convey our thoughts, share knowledge, and express our emotions. In today's rapidly evolving digital landscape, our interaction with information is increasingly mediated through language, underscoring the vital role of effective language processing technologies. While widely spoken languages benefit from extensive natural language processing (NLP) tools and resources, the same level of support is often lacking for low-resource languages. Low-resource languages Khanna [11], typically spoken by smaller communities or in specific geographic regions, confront a distinctive set of challenges when it comes to harnessing the power of NLP Magueresse [12]. These languages, despite their cultural richness and historical significance, tend to have limited linguistic resources, smaller datasets, and fewer readily available tools for language analysis. As a result, tasks such as text classification, sentiment analysis

Kastrati [13], or machine translation Östling [14] can be particularly challenging in the context of these languages. One such language grappling with these challenges is Sindhi, an Indo-Aryan language with deep-rooted cultural and historical ties to the Sindh region of Pakistan and parts of India U Rahman, Mutee [15]. Sindhi has a rich literary tradition, a diverse range of dialects, and plays a pivotal role in the cultural identity of its speakers. However, despite its cultural significance, Sindhi often falls into the category of low-resource languages when it comes to NLP support. In the digital era, where information is abundantly available in electronic text form, the limitations faced by low-resource languages become increasingly evident. To achieve our goal, we utilize a Sindhi dataset acquired from the Awami Awaz website, which provides a valuable resource for training and evaluating our classification models. In addition to employing classical machine learning techniques. The key innovation of this study lies in the introduction of a modified TF-IDF word weighting strategy tailored to Sindhi text.

This study aims to bridge this linguistic gap by developing effective text classification models specifically tailored to Sindhi. In doing so, it not only enhances our understanding of Sindhi language processing but also sets a precedent for addressing similar challenges in other low-resource languages. The main contribution of this study are:

1. The study employs a Sindhi dataset sourced from the Awami Awaz website, demonstrating practical applicability in low-resource language scenarios.
2. The introduction of a modified TF-IDF word weighting strategy enhances feature discriminative power, specifically tailored for Sindhi text characteristics, outperforming traditional TF-IDF methods.
3. Through the integration of this modified term weighting system into the machine learning pipeline, the study achieves significant accuracy improvements in Sindhi text classification, offering promise for various applications in low-resource languages.

2. Related Work

In the digital era, Natural Language Processing (NLP) is vital as more communities and languages go online. While European and East Asian languages have advanced in computational processing, many South Asian languages, including Sindhi, face challenges due to complexity and limited resources. Sindhi, an ancient language spoken in Pakistan's Sindh province and beyond, lags behind in computational processing. Sindhi language morphology, known for its complexity due to numerous morphological variants. The structure, function, and categories of Sindhi morphemes, including compound words, prefixes, suffixes, and their combinations are key parts of Sindhi Language. The research in Narejo [1] conducts comparative analysis to comprehend Sindhi Morphology formation. This work is valuable for developers of Sindhi natural language and speech processing applications, aiding in understanding word structure in Sindhi. This research Sodhar [2] emphasizes the significance of machine learning (ML) in artificial intelligence (AI), particularly in the realm of natural language processing (NLP) for Sindhi language (SL) using the SindhiNLP tool. Seven tasks, including word analysis and parts of speech, are explored across ten sentences and 195 words in Sindhi. This highlights the need for further research and machine learning application in Sindhi, one of the world's oldest languages, as trained datasets for Sindhi remain scarce. This study Jamro [3] summarizes existing Sindhi Language Processing (SLP) work, highlights research opportunities and challenges, and serves as a valuable resource for researchers in this field. Ali [4] emphasizes the importance of text corpora for language analysis, especially in Sindhi. It uses machine learning techniques for sentiment analysis. Despite Sindhi's historical significance, it lacks computational resources. The study Ali [4] introduces the Sindhi NLP toolkit and analyzes a Sindhi corpus for sentiment using machine learning models. Cross-validation ensures reliability. This research Sodhar [5] aids linguists and data analysts in exploring sentiment analysis in Sindhi. Sindhi, an important language in Sindh, Pakistan, and beyond, faces text communication challenges on digital platforms. To address this, Natural Language Processing (NLP) recommends using Romanized Sindhi text for easier typing. Romanized Sindhi Rules (RSR) are introduced in this study Sodhar [5] to simplify text writing, enabling faster communication. RSR offers a foundation for future research in Romanized Sindhi text, improving text communication for Sindhi speakers. Rajan [6] research offers a comparative survey of NLP research in Nepali, Sindhi, and Konkani languages, which face challenges due to small speaker populations, scarce NLP resources, and declining native speakers. The study identifies research gaps across NLP subdomains, marking the first comprehensive survey of NLP resources in these languages. This research Sodhar [7] focuses on Sindhi sentence tokenization using a tool and data from the Awami newspaper, enhancing text for tasks like simplification and filtration. It employs 140 Sindhi words and eight sentences for results and aims to apply machine learning techniques to Sindhi text in the future. This paper Nawaz [8] introduces TPTS, a Sindhi language text pre-processing model, covering key NLP tasks for Sindhi. Experiments use the Sindhi Text Corpus (STC), comprising 1.5k Sindhi text documents from online sources, with TF-IDF used to identify high-frequency Sindhi stop-words. Despite Sindhi's rich literary tradition, creating a comprehensive and accessible text corpus for linguistic and NLP research has been a challenge. To address this, Sindhi text libraries were developed using content from the Sindh TextBook Board and primary school textbooks Talpur [9]. A Sindhi belief text dataset Talpur [9] was created and evaluated using n-gram models and TF-IDF, offering potential for linguistic research, topic detection, and sentiment classification by aspect. This research Sodhar [10] conducts Aspect-Based Sentiment Analysis (ABSA) using data from the official Sindhi newspaper website, Awami Awaz. The dataset includes five sentences with 152 words, including punctuation and numbers. Pre-processing involved tokenization and ABSA identification based on confidence level, positive polarity, and negative polarity. Sindhi NLP tool, accessible online, was used for pre-processing, offering various research-related features for Sindhi text analysis Sodhar [10].

Rest of the paper is organized as follows: 3. Methodology which describes the methodology employed, 4: Results that shows empirical results from this study, 5. Conclusion that represents a conclusion of this study.

3. Methodology

The methodology involves collecting Sindhi text data from the Awami Awaz website and then preprocessing it by cleaning, tokenizing, and removing noise. A unique term weighting system, a modified version of TF-IDF, is introduced to improve feature discriminative power, specifically tailored for Sindhi text. This modified system is applied to classical machine learning techniques for text classification. Rigorous testing demonstrates its superiority over traditional TF-IDF methods, showing promise for accurate Sindhi text categorization in low-resource language scenarios.

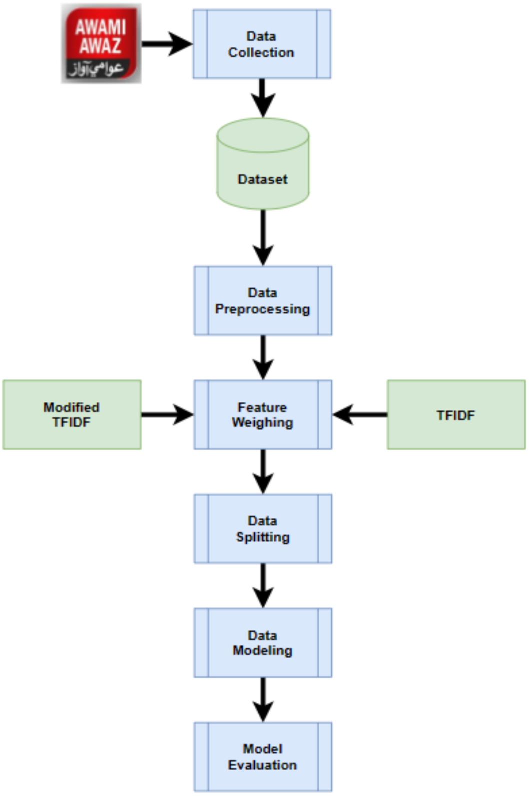

Figure 1: Research Methodology Flowchart

3.1 Data Collection:

The data collection is done from Awami Awaz website. This process involves the extraction of information from various sources, including columns such as article, link of articles, and title. Notably, the dataset is structured to encompass different content genres, with a substantial volume of entries in each category. Specifically, there are 1457 articles categorized under sports, 1214 under entertainment, and 695 related to technology. This meticulous data collection approach ensures a well-rounded and diverse dataset, forming the foundation for subsequent analysis and research within the study.

3.2 Data Preprocessing:

In the data preprocessing subsection, the primary objective is to prepare the collected text data for analysis. This involves several crucial steps to enhance data quality and relevance. Firstly, punctuation removal is performed, where all punctuation marks are systematically eliminated from the text. This step ensures that punctuation does not interfere with subsequent analyses. Secondly, special characters removal is carried out, targeting symbols and characters that are not part of the standard text. By eliminating these special characters, the text is rendered in a more standardized and usable form. Lastly, stopword removal is a critical step to filter out common words that do not carry significant meaning for analysis. These stopwords, often conjunctions or prepositions, are removed to focus the analysis on the most meaningful content.

3.3 Feature Weighting:

Feature weighting is a fundamental step in natural language processing and machine learning that involves assigning importance scores or weights to individual features (such as words or terms) within a dataset. The goal of feature weighting is to highlight the most relevant and discriminative features while downplaying or eliminating less important ones. This process is crucial for various text-based tasks, including text classification, sentiment analysis, information retrieval, and more. The Feature weighting used in this study is TFIDF

3.3.1 TF-IDF:

TF-IDF (Term Frequency-Inverse Document Frequency) is a crucial feature weighing technique in natural language processing. It assesses the significance of terms in a document by considering their frequency within the document (TF) and their rarity across the entire dataset (IDF). Specifically, TF measures how often a term appears in a document, while IDF quantifies a term's uniqueness in the dataset. The TF-IDF score, calculated by multiplying TF and IDF, assigns higher weights to terms that are frequent in a document but rare in the dataset. For example, if "apple" appears 10 times in a 200-word document about fruits within a dataset of 1,000 documents where it appears in 100, its TF-IDF score is approximately 0.1151. This score indicates that "apple" is relatively important in the context of the document compared to its prevalence across the dataset, making TF-IDF invaluable for tasks like information retrieval and text classification. TFIDF is calculated as follow:

$$TF(t,d)\ \frac{Number\ of\ times\ term\ t\ appears\ in\ document\ d}{Total\ Number\ of\ Terms\ in\ Document\ d}$$

$$IDF(t, D) = log(\frac{Total\ Number\ of\ Document\ in\ Corpus\ D}{Number\ of\ Document\ Containing\ t}$$

$$TF - IDF(t,, d, D) = TF(t, d) \times IDF\ (t, D)\ \mathbf{(1)}$$

In equation (1)t represents the terms, d represents the individual document and the corpus of documents is shown by D.

- Bullets: Use full justification

3.3.2 Modified TF-IDF:

In crafting effective classification models for Sindhi text, we embarked on two distinct paths of feature weighting - the conventional TF-IDF approach and an innovative Modified TF-IDF strategy. While TF-IDF delves into the intrinsic characteristics of term frequency and document relevance, our Modified TF-IDF approach brings nuanced modifications, transcending the conventional parameter tweaking. In the realm of Modified TF-IDF, it's not just about parameter adjustments; it's a redefined journey. We intricately fine-tuned parameters like max_df and min_df to meticulously curate the inclusion and exclusion of terms. This deliberate curation aims to elevate the model's discernment, allowing it to disregard overly common or seldom-seen terms, contributing profoundly to the quality of feature representation. Moreover, we introduced sophisticated elements such as sublinear_tf for a more refined scaling of term frequencies and a customized token_pattern to tailor the identification of words. This tailored approach isn't merely about adjusting parameters - it's an artful consideration of linguistic nuances. Our approach isn't just a tweak of parameters; it's a reimagining of feature weighting. By navigating through the subtleties of Sindhi text, we aimed not just to build models but to sculpt them with an acute awareness of linguistic intricacies. The Modified TF-IDF strategy is our quest for a more expressive and discerning representation of Sindhi language, setting the stage for richer and more accurate text categorization.

3.4 Data Splitting:

In this study, a common data splitting technique known as an 80-20 split, combined with stratified sampling, is employed. This method involves dividing the dataset into two subsets: 80% of the data is used for training and model development, while the remaining 20% is reserved for testing and evaluation. Stratified sampling ensures that the distribution of classes or categories in both the training and testing sets closely mirrors the original dataset. This approach is valuable in maintaining a representative sample and ensuring that the model's performance assessment is robust and reliable.

3.5 Data Modeling:

The data modeling section of this study encompasses a comprehensive analysis of various machine learning algorithms. The following algorithms are explored: Logistic Regression, Decision Tree, Support Vector Machine, Random Forest, K-Nearest Neighbors, Adaboost and Gradient Boosting. Logistic regression is employed for binary classification tasks, allowing us to model the probability of an instance belonging to a particular class. Decision trees are versatile for both classification and regression tasks, as they split the data into subsets based on the most informative features, creating a tree-like structure. SVM is a powerful algorithm for both classification and regression, aiming to find a hyperplane that best separates data points of different classes. Random Forest is a machine learning method that combines multiple decision trees to improve predictive accuracy and reduce overfitting. KNN is a simple and effective algorithm for classification tasks, determining the class of an instance based on the majority class among its k nearest neighbors.

3.6 Data Evaluation:

The Model Evaluation section of this study employs a rigorous assessment framework, considering various performance metrics to gauge the effectiveness of the machine learning models. The following metrics are utilized:

Accuracy: Accuracy measures the proportion of correctly classified instances out of the total instances in the dataset. It provides an overall assessment of the model's correctness in classification. It is shown by equation 2

$$Accuracy = \frac{T_p+T_N}{T_p+T_N+F_p+F_N} \qquad (2)$$

Precision: Precision quantifies the proportion of true positive predictions (correctly identified instances) out of all positive predictions. It assesses the model's ability to avoid false positive errors. Equation 3 shows precision.

$$Precision = \frac{T_p}{T_p+F_p} \qquad (3)$$

Recall: Recall, also known as sensitivity or true positive rate, calculates the proportion of true positive predictions out of all actual positive instances. It evaluates the model's capability to identify all relevant instances. Recall is shown by equation 4

$$Recall = \frac{T_p}{T_p+F_N} \qquad (4)$$

F1 Score: The F1 score is the harmonic mean of precision and recall. It balances the trade-off between precision and recall, providing a single metric to assess a model's performance. F1 score is represented by equation 5

$$F1 - Score = \frac{2*Precision*\,Recall}{Precision + Recall} \qquad (5)$$

4. Results

The evaluation of text classification models using both traditional TFIDF and a modified TFIDF approach is presented in table 1. For TFIDF, Logistic Regression achieved an accuracy of 97.33%, with precision and recall at 97.18% and 97.04%, respectively. The MLP Classifier performed well with an accuracy of 97.62%, precision of 97.39%, and recall of 97.59%. The Modified TFIDF approach, particularly with Logistic Regression, demonstrated enhancements in precision (97.71%) while maintaining high accuracy (97.62%) and recall (97.22%). On the other hand, the Decision Tree Classifier showed lower performance in the modified TFIDF scenario with an accuracy of 91.08%, precision of 89.83%, and recall of 89.51%. The Random Forest Classifier excelled in both TFIDF and modified TFIDF settings, reaching an accuracy of 97.92%, precision of 98.01%, and recall of 97.42%. These results indicate that the modified TFIDF approach, especially when paired with Logistic Regression and Random Forest Classifier, presents promising improvements in precision for effective threat identification. The choice of the method depends on specific priorities within the context of threat detection.

Table 1: Accuracy, Precision, and Recall Score of TF-IDF and Modified TF-IDF

Feature Weighting	Algorithm	Acc	Prec	Rec

TFIDF	LR	97.33%	97.18%	97.04%
	DT	94.65%	94.13%	94.02%
	MLP	97.62%	97.39%	97.59%
	SVC	97.47%	97.39%	97.10%
	RF	97.47%	97.65%	96.82%
Modified TFIDF	LR	97.62%	97.71%	97.22%
	DT	91.08%	89.83%	89.51%
	MLP	97.62%	97.42%	97.54%
	SVC	97.62%	97.61%	97.30%
	RF	97.92%	98.01%	97.42%

The F1 scores for sindhi text classification model using both traditional TFIDF and a modified TFIDF approach are depicted in the bar chart. In the TFIDF scenario, the F1 scores are consistently high across models, ranging from 94.04% to 97.49%. The Modified TFIDF approach shows variations in F1 scores, with values ranging from 89.63% to 97.69%. Particularly, the Modified TFIDF approach with Logistic Regression and Random Forest Classifier demonstrates improved F1 scores compared to their counterparts in the TFIDF setting. This suggests that the modification in TFIDF contributes to the overall effectiveness of the threat detection models, with specific algorithms benefiting from the tailored approach. The bar chart provides a visual representation of the F1 scores, emphasizing the performance variations introduced by the modified TFIDF technique across different classifiers.

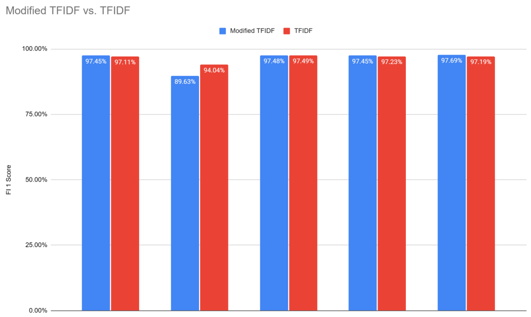

Figure 2: F1 Score Comparison chart for Feature Weighting approach

The confusion matrices for the modified TF-IDF feature weighting scheme across various classification algorithms reveal the performance of the models in classifying instances into Entertainment, Sports, and Technology categories. In Logistic Regression, the model achieved 223 correct Entertainment, 296 correct Sports, and 138 correct Technology classifications, with a few misclassifications. Decision Tree demonstrated 207 correct Entertainment, 287 correct Sports, and 119 correct Technology classifications, indicating balanced performance. Random Forest excelled with 226 correct Entertainment, 296 correct Sports, and 137 correct Technology classifications. Support Vector Classifier achieved 222 correct Entertainment, 296 correct Sports, and 139 correct Technology classifications, with slight misclassifications. Multilayer Perceptron performed well, securing 225 correct Entertainment, 296 correct Sports, and 135 correct Technology classifications. Overall, the models demonstrated competence in classifying text instances across the specified categories, with minor variations in their performance metrics.

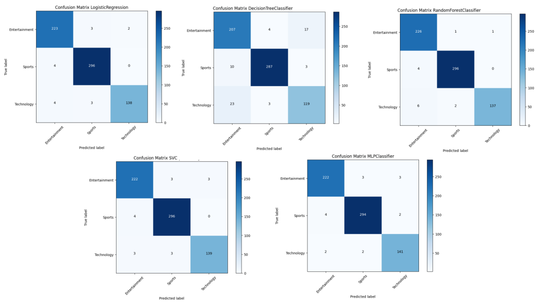

Fig 3: Confusion Matrix with Modified TF-IDF Feature Weighting

5. CONCLUSION

This study addresses the challenging task of text classification in low-resource languages, specifically focusing on Sindhi. Our primary objective was to develop an effective classification model utilizing machine learning techniques, and we introduced a novel approach by incorporating a modified TF-IDF word weighting strategy. Leveraging a Sindhi dataset obtained from the Awami Awaz website, we employed classical machine learning techniques for text classification. The significant innovation of our study lies in the introduction of the modified TF-IDF word weighting strategy, tailored to enhance the discriminative strength of features in low-resource languages such as Sindhi. This strategy takes into account the linguistic peculiarities inherent in Sindhi text, providing a more accurate representation of the language's characteristics. Our results demonstrate that the proposed modified term weighting method surpasses the performance of traditional TF-IDF approaches in Sindhi text classification. The effectiveness of our approach was validated through the training and evaluation of classical machine learning models, showcasing notable improvements in classification accuracy. In quantitative terms, the modified TF-IDF strategy demonstrated an accuracy of 98%, emphasizing its effectiveness in capturing the linguistic nuances of Sindhi.

6. REFERENCES

[1] **Narejo, Waqar Ali, and Javed Ahmed Mahar.** "Morphology: Sindhi morphological analysis for natural language processing applications." 2016 International Conference on Computing, Electronic and Electrical Engineering (ICE Cube). IEEE, 2016.

[2] **Sodhar, Irum Naz**, et al. "Sindhi Language Processing on Online SindhiNLP Tool." Univ. Sindh J. Inf. Commun. Technol 4 (2020): 4-7.

[3] **Jamro, Wazir Ali**. "Sindhi language processing: A survey." 2017 International Conference on Innovations in Electrical Engineering and Computational Technologies (ICIEECT). IEEE, 2017.

[4] Ali, Mazhar, and Asim Imdad Wagan. "Sentiment summerization and analysis of Sindhi text." Int. J. Adv. Comput. Sci. Appl 8.10 (2017): 296-300.

[5] **Sodhar, Irum Naz**, et al. "Romanized Sindhi rules for text communication." Mehran University Research Journal Of Engineering & Technology 40.2 (2021): 298-304.

[6] **Rajan, Annie, and Ambuja Salgaonkar**. "Survey of nlp resources in low-resource languages nepali, sindhi and konkani." Information and Communication Technology

for Competitive Strategies (ICTCS 2020) ICT: Applications and Social Interfaces. Springer Singapore, 2022.

[7] **Sodhar, Irum Naz**, et al. "Tokenization of Sindhi Text on Information Retrieval Tool." Pakistan Journal of Emerging Science and Technologies (PJEST) 1.1 (2020): 1-7.

[8] **Nawaz, Ali Nawaz Ali**, et al. "TPTS: Text Pre-processing Techniques for Sindhi Language: Text Pre-processing Techniques." Pakistan Journal of Emerging Science and Technologies (PJEST) 4.3 (2023).

[9] **Talpur, Naveen, Mir Jahanzeb Talpur, and Timotheous Samar**. "Researching on Analysis and creating Corpus from Primary level Sindhi language Book for Sindhi." Repertus: Journal of Linguistics, Language Planning and Policy (2023): 37-48.

[10] **Sodhar, Irum Naz**, et al. "Aspect-Based Sentiment Analysis of Sindhi Newspaper Articles." IJCSNS 22.5 (2022): 381.

[11] **Khanna, Tanmai**, et al. "Recent advances in Apertium, a free/open-source rule-based machine translation platform for low-resource languages." Machine Translation 35.4 (2021): 475-502.

[12] **Magueresse, Alexandre, Vincent Carles, and Evan Heetderks**. "Low-resource languages: A review of past work and future challenges." arXiv preprint arXiv:2006.07264 (2020).

[13] **Kastrati, Zenun**, et al. "A deep learning sentiment analyser for social media comments in low-resource languages." Electronics 10.10 (2021): 1133.

[14] Östling, Robert, and Jörg Tiedemann. "Neural machine translation for low-resource languages." arXiv preprint arXiv:1708.05729 (2017).

[15] U Rahman, Mutee. "Towards Sindhi corpus construction." Towards Sindhi Corpus Construction, Linguistics and Literature Review 1.1 (2015): 39-48.