C O V E N T R Y
U N I V E R S I T Y
Faculty of Engineering, Environment and Computing
School of Computing, Electronics and Mathematics

Data Science and Computational Intelligence

7151CEM – Computing Individual Research Project

Citation Analysis Using Ensemble Techniques

Author: Jyothi Yendamuri

SID: 11467683

Supervisor: Dr. Xiaorui Jiang

Submitted in partial fulfilment of the requirements for the Degree of Master of Science in Data Science and Computational Intelligence

Academic Year: 2021/22

Declaration of Originality

I declare that this project is all my own work and has not been copied in part or in whole from any other source except where duly acknowledged. As such, all use of previously published work (from books, journals, magazines, internet etc.) has been acknowledged by citation within the main report to an item in the References or Bibliography lists. I also agree that an electronic copy of this project may be stored and used for the purposes of plagiarism prevention and detection.

Statement of copyright

I acknowledge that the copyright of this project report, and any product developed as part of the project, belong to Coventry University. Support, including funding, is available to commercialise products and services developed by staff and students. Any revenue that is generated is split with the inventor/s of the product or service. For further information please see www.coventry.ac.uk/ipr or contact ipr@coventry.ac.uk

Statement of ethical engagement

I declare that a proposal for this project has been submitted to the Coventry University ethics monitoring website (https://ethics.coventry.ac.uk/) and that the application number is listed below (Note: Projects without an ethical application number will be rejected for marking)

Signed: Jyothi Yendamuri Date: 19-04-2022

Please complete all fields.

First Name:	Jyothi
Last Name:	Yendamuri
Student ID number	11467683
Ethics Application Number	P133961
1st Supervisor Name	Dr. Xiaorui Jiang
2nd Supervisor Name	Dr. Alireza Daneshkhah

Table of Contents

Acknowledgements

I would like to express my sincere gratitude to Coventry University for allowing me fulfil my dream of being a student here. I am also extremely grateful for my supervisor Dr. Xiaorui Jiang, for his enthusiasm, patience, suggestions, encouragement and for pushing me further throughout my research project. I would want to express my gratitude to every one of my friends and family for their encouragement and support.

Abstract

Citation analysis is a major topic in academics for a variety of reasons. Over the last few years, there has been tremendous development in the use of citation to enhance the methods of assessing the originality of publications. The citation work reveals the author's motivation for referencing a particular book. Furthermore, several machine learning and deep learning techniques have been suggested for citation semantics analysis such as citation function classification (CFC) or citation significance identification (CSI). There are several noticeable problems. Due to the difficulties of annotation, dataset sizes were restricted, particularly for minority classes. Existing research were difficult to map and compare due to various annotation schemes. State-of-the-art deep neural networks for CFC created a feature vector for the entire citation context (or phrase), rather than modelling individual in-text citations, which is theoretically problematic. CSI, on the other hand, was commonly performed regardless of the citation context. Traditional approaches for determining the quality of articles, such as the weighting rely solely on the number of citations.

Earlier, researchers merged and re-annotated 4784 citations from six existing datasets of computational linguistics publications to build a huge citation context dataset. They have trained a series of Deep learning models for citation context analysis by experimenting with different ways of encoding features; but different models performed differently using the same dataset, and no single model achieved high performance. We used the same dataset to investigate the differences between various Deep learning models. The dataset was provided by Professor Xiaorui Jiang to investigate the research further.

The manual proposed technique takes a long time and is prone to errors. Despite the fact that numerous research have been done on citation function categorization, the performance of existing studies does not meet the criteria of analysis on large-scale data. Thus, for this research, we believe that using an ensemble approach is a good idea to boost the performance. By introducing various outputs of the dataset, the final performance has assessed.

First, we investigated the similarities and differences between the variant models' performances. The base classifiers in an ensemble must be distinct in order to enhance performance over a single model. So, we have checked variety of classifiers. To study the diversity, we employed metrics such as pair-wise and non-pair-wise diversities. We have performed ensemble majority voting for all epochs and best epochs to test the performance. We found that the suggested strategy outperforms existing strategies in terms of accuracy.

We achieved best 76.19% macro for best epochs with 6-class scheme. After performing the disagreement measure, 79.53% in class-6 scheme depicted as highest, and also we observed 72.89% in class-10 scheme and 72.54% in class 11 scheme respectively. We expect that the provided dataset results will serve as a valuable benchmark for researchers.

Keywords: Citation function classification, Citation context analysis, Ensemble Approach, Disagreement Measure, Pair-wise and Non pair-wise diversities.

1. Introduction

Overview

Over the last fifty years, a large number of research papers have focused on citation activity. Citations are main component of a scientific field's current condition, indicating how researchers organize their work and affecting future researchers' uptake. Recently a study have been provided by researchers and found out that citation classification is a helpful tool for research evaluation. The incentive or purpose to why authors reference particular papers in their work is classified as the citation function classification. Acknowledge sources and mention the original author as the easiest way to prevent plagiarising other people's work. The impact factor of scientific publications in the journal is calculated using citations. Evaluating scientific journals, particularly the flow of information transmission throughout the innovation process, necessitates classification of citation functions. Studying why writers cite specific papers is important for many scientific applications. Authors have a variety of reasons for acknowledging other writers' work. For example, researcher can use earlier work to get some kind of guidance in the way of prior knowledge, tools, and ideas. This can also be used to assess procedures and evaluate prior work.

1.1 Background to the Project

The citation work reveals the author's motivation for referencing a particular article. Furthermore, past research relied heavily on several machine learning approaches. To answer a major question, many citation function systems have been developed. Why does the author mention a research study with so many functions and granularity levels? However, quantitative measurements alone may be deceiving when it comes to a research's beneficial effect. A publication on a severely criticized work, for example, may receive a large number of citations. Systematic qualitative approach of articles is difficult since it necessitates the processing of all citing articles' textual content.

However, many works on citation function classification have been done; the performance of existing studies does not satisfy the requirement of analysis on large data. So, we choose an ensemble technique is a suitable technique to improve the performance.

The categorization study of the citation function can help publishers to discover the relationships between publications or articles. It aids to improve the design citation indexers, and the long-term examination of a scientific work's overall argumentation framework. The performance results of this study will provide some information about the quality of models which will be used to improve the citation performance. This will also be helpful to future researchers to find the models for better performance.

1.2 Project Objectives

Previously, a series of Deep learning models are trained for citation context analysis by experimenting with different ways of encoding features; observed different models performed differently using the same dataset, and no single model achieved high performance. We believed that we could investigate the differences between various Deep learning models and create ensemble techniques to out of proper analysis.

1.2.1 Analyse and model the requirements

The major goal is to create an ensemble model majority voting classifier to combine the all model predictions to find out the performance of citation function classification. Ensemble models mainly used for improve the performance using previous model predictions. A Python framework is used to create the models. Thus, the coding was carried out in the Google Colab (a web-based Google research platform that allows developers to create and execute python code) GPU or Anaconda jupyter to train and evaluate models.

1.2.2 Investigation of possible solutions

The following are the stepwise goals will deal with the research problem:

- A data frame should be used to store the required information.
- To build an ensemble model we need to analyze classification results of variety of models.
- Ensemble model will be created to combine the multiple model predictions to improve the overall performance.
- To analyze the diversities between the classifiers we use some measuring techniques like pair wise and non-pair wise diversities
- The disagreement measure between each pair of classifiers can be calculated to combine all the classifiers and based on this we can find a classifier which is most different from others.
- Different classes are used to test and increase the citation function performance
- Performance will be evaluated adding different thresholds range in disagreement measure
- After implementing the strategies, models will be evaluated using the F1 score, Precision, Recall, and confusion matrix, as well as other citation function ontologies.

1.3 Overview of This Report

Numerous citation strategies have been proposed in order to determine the intent of a specific citation and improve the performance. Citations are major anchors in scientific content that assist to organize the large publication area. Scientific knowledge is still maintained and transferred in quite unorganized pieces of information, vulgo research articles or other related, textual representations, even in the current age of Semantic Web, structured repositories and massive databases. The citation data was interpreting with suitable citation strategies depending on their interests, and the citation function performance can be calculated using multiple approaches. Citations are supportive to understand and reproduce the outcomes. Previously many machine learning and deep learning techniques are used to improve the performance. But performance metrics were incomparable as they worked different datasets with different classification schemes (citation function annotation schemes). So, in this report we introduce ensemble techniques and different diversity measures to increase the performance.

2. Literature Review

Citations are helpful for new researchers in a scientific field since they eventually refer to seminal, original work and information that is not public accessible or replicated in every publication. Citations are also the major communication links in scientific debates that can last over years or even decades. Moreover citations are assisting to understand and reproduce the outcomes. As a result, they form a significant text characteristic for all readers.

Analyzing citation context is a significant task in scientific texts identification, particularly for understanding the flow of knowledge dissemination in the innovation process. The categorization of citation functions is defined as a method of determining the reasons for citing prior publications. For many scientific applications, such as ranking, recommendation of related articles, revealing the structure of the science domain, research assessment, and summarizing a scientific issue, it is difficult for us to grasp the intents of authors producing citations to specific publications. Several citation semantics are considered to be helpful for clarifying the evolutionary nature of scientific publishing connections (Ghosal et al., 2022) and thus have wide range of downstream applications (Ding et al., 2014).

Over the last two decades numerous datasets with various identification techniques have been used. Deep learning has lately greatly improved the state-of-the-art (SOTA) in citation study field in terms of classification accuracy (Cohan et al., 2019; Beltagy et al., 2019).

Teufel et al. (2006) published the biggest dataset of 2,829 assessed citations. These citations were taken from the ACL anthology reference corpus of data science articles. The authors of the article generated the 12 different sorts of annotations that were employed in their citation categorization scheme.

Jurgens et al. (2018) enhanced the accuracy of overall CFC from 54.9 percent macro F1 on a dataset using an application-friendly six-class annotation system, i.e. "Background", "Future" (Work), "Comparison Or Contrast", "Motivation", "Uses", and "Extends". Cohan (Cohan et al., (2019) got 69 percentage and Beltagy (Beltagy et al., 2019) got 70.98 percentage macro f1 using a state of the art feature engineering technique.

Teufel and her colleagues used Instance based learning (IBK) to achieve an accuracy of.77 percent in their study using annotated data from their citation scheme, and they have also employed a k-NN classifier with cue phrases, self-citation and the position of the citing sentence as features (Teufel et al., 2006).

Zhu and Goldberg (2009), used an idea of self training to utilize unsupervised data, but choose fresh labelled data or new data in a more reliable way during the self-training process by utilising the ensemble learning method. Aljohani et al. (2021) combined Tuefel and jurgens Datasets. However it is very difficult to mapping over various datasets identification systems.

Cue sentences and rule-based techniques are used in the beginning, which demands a lot of manual effort to construct the rules and expressions, which is time-consuming and hard. Wan et al. (2009) provide a research on user demands for exploring scientific articles, demonstrating that citations are important. Schafer et al. (2010) proposed a unique user interface for browsing written citation graphs.

Munkhdalai et al. (2016) used a simple citation system and a deep learning architecture to develop a solution for end-to-end citation function categorization. Long short term memory (LSTM), LSTM + Global Attention, Bi-LSTM, and Bi-LSTM + Global Attention were the models utilized. Regarding citation context, unidirectional LSTMs with global attention had the highest F1 score of 75.86 percent. And for the citation phrase, bidirectional LSTMs with CAN (Compositional attention network) received the greatest F1 score of 68.61 percent.

Anne Lauscher et al. (2017) used collective word embeddings to build either of those Support vector Machines (SVM) and Convolutional Neural Networks (CNN) and contrasted their outcomes. The researchers utilized a dataset with 3271 citation contexts and six forms of citations: comparison, usage, substantiation, criticism, basis, and neutral. When utilizing the CNN model, best accuracy was achieved. Only CNN models better basic linear SVM utilizing lexical features by a little percentage, with F1 scores of 74.3 and 73.1, respectively.

In ensemble learning, the final prediction for a specific instance is usually made using two basic combination rules. One is probability-based, in which the fundamental classifier decision that generates the highest prediction confidence for the instance is chosen. Another option is majority voting (Dong and Ulrich Schafer, 2011).

The past two decades have observed many datasets with variety of annotation techniques. Deep learning has recently greatly improved the state-of-the-art (SOTA) of this study area in terms of classification accuracy (Cohan et al., 2019). Citation function classification performance on certain academic techniques (Turado et al., 2021) or (Zhao et al, 2019; Zheng et al, 2021) may be better but these are not the focus of this work.

Xiaorui et al., (2022) proposed a Strong SciBERT-based CFC (Citation Function Classification) models. The top models outperformed the state-of-the-art with 73.99 percent best (macro) F1 and 71.70 percent average F1 using a standard 6-class scheme. The CSI (Citation Significance Identification) model that relied just on citation context earned an F1 of 87.03 percent, which is comparable to the current state-of-the-art.

2.1 Ensemble: Background Study:

The Machine Learning Research has expressed an interest in the technique of integrating several categorization models. This approach is described as combination of specialists, ensemble techniques or ensemble of predictors. An ensemble of classifiers is made up of a group of classifiers known as base learners, each of their unique choices are integrated in some way to categorise future samples.

Many years ago, the very first research publications on ensembles of predictors were published. Major ensemble approaches including Bagging, Boosting and Stacking were introduced in the 1990s. These studies paved the way for a highly promising machine learning technique, ensemble classifier.

Many scientific frameworks have shown that the precision and variety of the individuals of any ensemble of classifiers are associated to the ensemble's accomplishment. According to Sagi et al., (2018) the objectives why ensemble approaches generally increase forecasting accuracy, are to minimise prediction error, reduce the chance of getting a local minimum, and enhance the hypotheses research scope.

Gabriele et al., (2001) proposes an approach for constructing classifier ensembles that prioritises variability between ensemble members. The extensive experiments from that study show that ensembles based on diversity are more accurate than ensembles relying on error rate. As a result, if an ensemble of predictors has a minimum error and their flaws are not synchronous, it can increase the precision of any of its individual members.

We've gone through the many models percentages, methodology, citation schemes, and datasets used by researchers for citation categorization analysis, as well as the advantages and disadvantages of each model. Although some of the research projects' accuracy is not greater, the results are not comparable since each author's citation method and dataset are distinct.

3. System Requirements

To begin working on this citation project, we used certain software. The python programming language is used to create the code. Because it is one of the most widely used languages in the industry and has also been successful in a number of programming languages. Several libraries are available in python, including Pandas, NumPy, Tensorflow, Seaborn, and Sklearn.

4. Examining Previous Datasets

Several citation contextual datasets have been presented for automatic citation semantics analysis in the last two decades. Detailed examinations of the datasets, challenges, and techniques could be found in Kunnath et al. (2021), Lyu et al. (2021), and Hernández-Alvarez and Gómez (2016).

Earlier research on citation categorization systems relied on datasets documented by specialists and professional annotators, which made the performance appraisal system slow and expensive. As a consequence, existing datasets in the domain are constrained to a certain domain, primarily information science and medical disciplines, because this is where the authors can identify the occurrences. The majority of datasets have reference contexts of varying durations. These references were evaluated in their complete contexts by Teufel et al. 18 (2006a, 2006b) and Hernández-Alvarez et al. (2017).

Below table provides an overview of available citation semantics datasets; however it is far from exhaustive. Kunnath et al. (2021) provide a more comprehensive overview.

Table 1: Survey of existing citation semantics datasets

Dataset	Fields*	Size	Annotation Scheme	Fulltext	Context	OA	Authoritative
Teufel et al. (2006a, 2010)	CL	4022	Neut, Weak, CoCoXY, CoCoGM, CoCoR0, CoCo-, PSim, PSup, PMot, PUse, PModi, PBas	•	variable	•	
Agarwal et al. (2010)	BM	3491	Background/Perfunctory, Contemporary, Contrast/Conflict, Evaluation, Explanation, Method, Modality, Similarity/Consistency	•	[-1, +1]	•	
Dong and Schäfer (2011)	CL	1728	Level 1: Background, Compare, Fundamental Level 2: Background_GRelated, Background_SRelated, Background_MRelated, Compare, Fundamental_Idea, Technical_Basis	•		•	
Jochim and Schütze (2012)	CL	2008	Aspect 1: conceptual vs operational; Aspect 2: evolutionary vs juxtapositional; Aspect 3: organic vs perfunctory; Aspect 4: confirmative vs negational.	•	variable	•	
Li et al. (2013)	BM	6335	Based_on$^{\top}$, Corroboration$^{\top}$, Discover$^{\top}$, Positive$^{\top}$, Practical$^{\top}$, Significant$^{\top}$, Standard$^{\top}$, Supply$^{\top}$, Contrast$^{=}$, Co-citation$^{=}$, Neutral$^{=}$, Negative^{-} (+/=/-: positive/neutral/negative)	•	[-?, +?]		
Abu-Jbara et al. (2013) also Jha et al. (2016)	CL	2098	Neutral, Criticizing, Comparison, Substantiating, Basis, Use				
Hernández-Alvarez et al. (2017)	CL	3013	acknowledge, corroborate, weakness, hedge, useful, based		[-1, +2]	•	
Jurgens et al. (2018)	CL	1954	Background, Compare or Contrasts, Motivation, Uses, Continuation (=> Extends), Future	•	variable	•	
Cohan et al. (2019)	CS, BM	11020	Background introduction, Method, Result comparison				
Su et al. (2019)	CL	1402	Neut, Weak, CoCo, Pos	partial**	•***	•	
Kunnath, Pride, et al. (2020, 2021)	CS, BM	3000	Background, Compares_Contrasts, Motivation, Uses, Extension, Future			•	
Pride and Knoth (2020)	various	11233	Background, Compare_Contrast (subclasses: similarities, differences, disagreement), Motivation, Uses, Extension, Future		[-1, +1]	•	
						•	•
Ferrod et al. (2021)	various	1380	Proposes, Analyzes (subclass: critiques), Compares (subclass: contrasts), Uses (subclass: dataset), Extends *Additional aspect: role – subj v.s. obj*			•***	•
Lauscher et al. (2021) *Multi-label annotation*	CL	12653	Background, Differences, Similarities, Motivation, Uses, Extends, Future Work			•	
Zhang et al. (2021)	CL	9594	Relationship – Motivation, Comparison, Extension, Application; Content – Background, Method, Data, Result; Sentiment – Positive v.s. Negative		variable	•	
				•	??	•***	
Meyers (2013)	BM	291	Corroborate v.s. Contrast	??	[-?, +?]		
Zhao et al. (2019) *Resource citation*	CL, ML, BM	3088	Use, Produce, Introduce, Compare, Extend, Other *Role: Material – Data; Method – Tool, Code, Algorithm; Supplement – Website, Document, Paper, Media, License*		[-2, +2]	•	
Tuarob et al. (2020)	CS	8796	Level 1: UTILIZE v.s. NONUTILIZE Level 2: USE, EXTEND v.s. MENTION, NOTALGO	•	****	•	
Wan et al. (2014)	CL	~800	5-grade	--	--		
Zhu et al. (2015)	various	140+	2-grade: influential v.s. non-influential	--	--	•	•
Valenzuela et al. (2015)	CL	465	4-grade; 2-grade (important v.s. incidental)	--	--	•	
Qayyum and Afzal (2019)	CS	488	2-grade	--	--		•
This study *CITSEG- and citation-level annotation*	CL	4784/ 3854	11-/10-class: Neutral, Weak, CoCoXY, CoCoGM, CoCoRes, Similar, Motivation, Usage, Extends, Future, (Support)*****	•	[-2, +3] or variable	•	

** Not all citations and not all citation contexts were annotated.
*** The authors of the datasets claimed to share their datasets, but they were still unavailable at the time of writing this paper.
**** A context window of a certain number of characters around the citation were extracted by ParsCit. Context size is not in measured in sentence count.
***** 11- or 10-class depending on whether including a Support class or re-annotating Support into other categories

4.1 Annotation Scheme For citation

The annotation scheme of complete dataset is given by Professor Xiaorui Jiang and most of the complete dataset is done by using Teufel's citation scheme.

According to Jurgens et al. (2018), the citation function scheme for science and technology like computer science has been "fixed" to a 6-class scheme that is the lines from Kunnath et al. (2020) to Lauscher et al. (2021). However, Teufel et al. (2006a) and Teufel et al. (2010) 12-class method continues to become the most conceptually reasonable; Researchers who aren't experts in citations function categorization will find the 6-class approach easier to comprehend and evaluate (Section 4.2).

Whereas most biomedical publications are freely accessible via PubMed Central, the volume of datasets available is insufficient. On the other hand, Computer science journals are sometimes unavailable for free, with the exception of a few disciplines such as AI or ML (machine learning). Because the Association for Computational Linguistics (ACL) keeps an open access database of published papers in ACL-sponsored venues, most extant datasets were annotated on CL articles. Four evaluation functions were defined by Teufel2010. "CoCoGM", "CoCoR0", and "CoCo-" are three of them that have been integrated into a unified comparative function.

According to Jiang 2021, the citation function scheme mapping and CITSEG –level statistics after re-annotating psup and psupport, the complete classes of annotation data are shown below in below table.

Table 2: Annotation Schemes, Statistics, and (Partial) Conceptual Mappings between Six Citation Datasets

Teufel2010			Dong2011[3]				Jha2016				Alvarez2017			Jurgens2018			Su2019		
Type	#	%	Type		#	%	Type		#	%	Type	#	%	Type	#	%	Type	#	%
PSup[2]	46	1.14	**Background**[4]	Neu	953	55.15	**Substantiate??** *Unmappable!!*		126	6.01	Background			**ComOrCon**	-	-	**Neut** (Neutral)	993	70.83
Neut Neutral description of cited work, or not enough textual evidence for other categories	2398	59.6	**- GRelated** General description of related work								**cor**roborate	0	0	**Future**	69	3.53			
			- SRelated Specifically related aspect of cited work: method, parameter, …				**Neutral**		1283	61.15	**ack**nowledge	982	32.59	**Background**	999	51.13			
CoCoXY Contrast between 2 cited methods	125	3.11		Pos *PMot?*	149	8.62					**deb**ate (0)	0	0						
							Criticising	Pos *PMot?*	71	3.38	Use (**use**ful) *Partly PMot?*	857	28.44						
Weak Weakness of cited approach	127	3.16	**- MRelated** Method mentions in related work that may be usable in current or future work*	Neg	46	2.66		Neg	150	7.15	Critique - **wea**kness	141	4.68	**ComOrCon** (Compare or Contrast)	353	18.07	**Weak**	30	2.14
											- **hed**ge	40	1.33						
CoCo- Unfavourable contrast/comparison (against cited work)	62	1.54	**Compare**		70	4.05	**Comparison**		122	5.82	Contrast (con)	136	4.51				**CoCo**	90	6.42
CoCoGM[1] (Neutral) contrast/comparison in Goals or Methods	187	4.65																	
CoCoR0 (Neutral) comparison in Results	51	1.27																	
PSim Citing work similar to cited work	133	3.31	**Fundamental** **- (Fundamental) Idea** *+PSim*		127	7.35					Use (**bas**ed) = PSim + PUse + PModi + PBas + PMot	491	16.30						
PMot Citation is positive about approach or problem, to motivate this paper	131	3.26					**Basis**		74	3.53				**Motivation**	89	4.55	**Pos** Positive (usage) to citing work	289	20.61
PBas Cited work as starting point	60	1.49												**Extends** *mixing PBas*	78	3.99			
PModi Adapt or modify algorithms, tools, data and etc. of cited work	60	1.49	**- (Technical) Basis** *+PSim*		420	24.31													
PUse Use algorithms, tools, data and etc.	642	15.96					**Use**		272	12.96				**Uses**	366	18.73			
Total	4022				1728				2098			3013			1954			1402	

Table 3: Citation function scheme mapping and CITSEG-level statistics.

Teufel2010+ (12+1 class)				Jiang2021 (11-class)			Jiang2021 (9-class)			Jiang2021 (7-class)			Jurgens2018 (6-class)		
label	Size		ratio	label	size	ratio	label	size	ratio	label	size	ratio	label	size	ratio
	citstr	citseg	citseg												
Future	97	85	2.21%	Future	85	2.21%	Future	85	2.21%	Future	85	2.21%	Future	85	2.21%
CoCoXY	200	152	3.94%	CoCoXY	152	3.94%	Neutral	1615	41.90%	Background	1773	46.00%	Background	1615	41.90%
Neut	1924	1463	37.96%	Neutral	1463	37.96%*									
Weak	223	158	4.10%	Weakness	158	4.10%	Weakness	158	4.10%						
CoCoGM	390	299	7.76%	CoCoGM	328	8.51%	ComOrCon	479	12.43%	ComOrCon	479	12.43%	ComOrCon	944	24.49%
CoCo-	108	80	2.08%												
CoCoR0	107	100	2.59%	CoCoRes	151	3.92%									
PSup	123	100	2.59%	Support	100	2.59%	Support	100	2.59%	Similar**	307	7.97%			
PSim	247	207	5.37%	Similar	207	5.37%	Similar	207	5.37%						
PMot	365	288	7.47%	Motivation	288	7.47%	Motivation	288	7.47%	Motivation	288	7.47%	Motivation	288	7.47%
PUse	794	755	19.59%	Usage	755	19.59%	Usage	755	19.59%	Uses	755	19.59%	Uses	755	19.59%
PModi	72	65	1.69%	Basis	167	4.33%	Basis	167	4.33%	Extends	167	4.33%	Extends	167	4.33%
PBas	134	102	2.65%												

Table 4: Citation function scheme mapping and CITSEG-level statistics after Re-annotating "PSup"/"Support".

Teufel2010+ (12+1 class)				Jiang2021 (11-class)			Jiang2021 (10-class)			Jiang2021 (8-class)			Jurgens2018 (6-class)		
label	size		ratio	label	size	ratio	label	size	ratio	label	size	ratio	label	size	ratio
	citstr	citseg	citseg												
Future	97	85	2.21%	Future	85	2.21%	Future	89*	2.31%	Future	89	2.31%	Future	89	2.31%
CoCoXY	200	152	3.94%	CoCoXY	152	3.94%	CoCoXY	153	3.97%	Background	1673	43.41%	Background	1670	43.38%
Neut	1924	1463	37.96%	Neutral	1463	37.96%	Neutral	1520	39.44%						
PSup	123	100	2.59%	Support	100	2.59%									
PSim	247	207	5.37%	Similar	207	5.37%	Similar	235	6.10%	Similar	235	6.10%			
Weak	223	158	4.10%	Weakness	158	4.10%	Weakness	158	4.10%	Weakness	158	4.10%			
CoCoGM	390	299	7.76%	CoCoGM	328	8.51%	CoCoGM	328	8.51%	ComOrCon	485	12.58%	ComOrCon	877	22.78%
CoCo-	108	80	2.08%												
CoCoR0	107	100	2.59%	CoCoRes	151	3.92%	CoCoRes	157	4.07%						
PMot	365	288	7.47%	Motivation	288	7.47%	Motivation	289	7.50%	Motivation	289	7.50%	Motivation	289	7.52%
PUse	794	755	19.59%	Usage	755	19.59%	Usage	758	19.67%	Usage	758	19.67%	Usage	755	19.59%
PModi	72	65	1.69%	Basis	167	4.33%	Basis	167	4.33%	Basis	167	4.33%	Basis	167	4.33%
PBas	134	102	2.65%												

As per the above citation function scheme each category or labels are described below.

Teufel's citation scheme	Jurgens functions	Description
Weak	Weakness	The cited approach flaws
CoCoGM	Comparison Or Contrast	Goals or approaches that contrast or compare (neutral).
CoCoRO	Comparison Or Contrast	In the Results, there is a contrast/comparison (neutral).
CoCo-	Comparison Or Contrast	Unfavorable Contrast/Comparison (modern work is superior than that referenced).
CoCoXY	Background	Compare and contrast the two cited strategies.
PBas	Basis	As a preliminary step, the author refers to the mentioned work.
PUse	Usage	The author employs techniques, methods, information, and ideas.
PModi	Extends	The effort of the author and the work mentioned are compatible/ supportive of one another.
PMot	Motivation	Used to encourage people to work on the present paper
Psim	Similar	The work of the authors and the work mentioned are comparable.
Psup	Support	The author's work and the quoted work are compatible and complement each other.
Neut	Neutral	Delivers information related to this domain
Future	Future	Source of future work

Table 5: Citation function scheme description

The following are the categories: If the researchers address a mistake in prior research, one category 'Weak' is allocated for it. The distinction among categories is to understand whether the contrast is between techniques or (CoCoGM), or outcomes. Furthermore, in the event of achievements, the referenced results are worse than the present comparable results (CoCoRo). Thereby, the annotators are pivotal to designate citation functions which provide literary proof for each function they assign.

The original dataset used in this research was provided by Professor Dr. Xiaorui Jiang, and annotation of the entire dataset is developed by utilizing Teufel's citation method (Refer Table 3 & 5). In this research we introduced previous intermediate outputs which are trained by deep learning models for citation function classification (Section 4.2). Requisite studies are to be analysed the logs of predictions made on each validation and test set after each epoch for the purpose of building ensembles. An overview of complete dataset has been represented in the Figure 1.

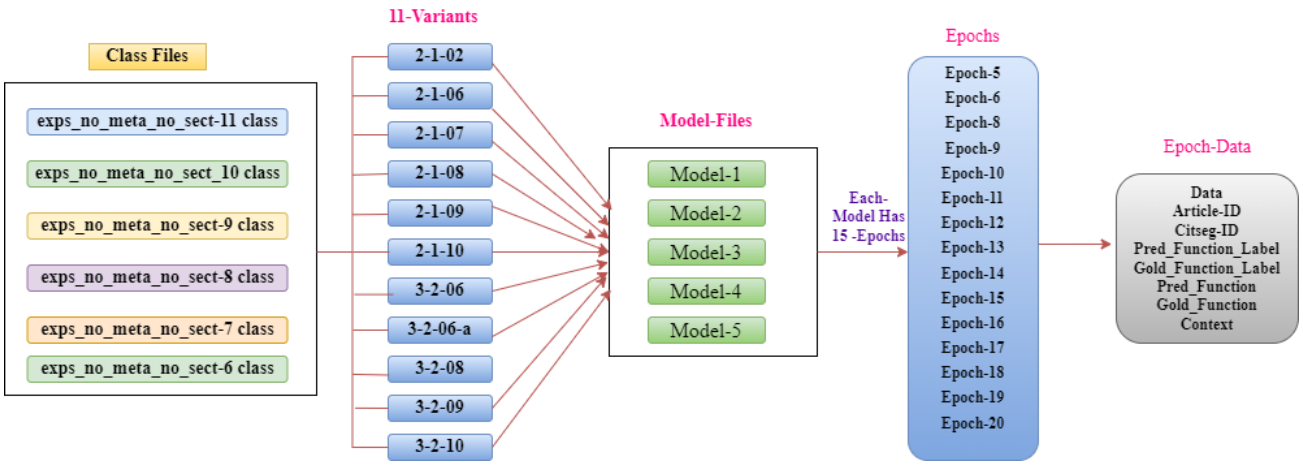


Figure 1: Complete Overview of the Dataset

There are six class files, as shown in the flow diagram above (Class 11, 10, 9, 8, 7, 6). Each class file has a set of eleven variants having five model files for each individual. Furthermore, 15 epochs are integrated in each model file. Each epoch consistently includes citation data such as Article ID, Citseg ID, Prediction function label, Gold function label, Prediction function, Gold function, and Context. A detailed representation of Dataset files containing Epoch Data is shown in Figure 2.

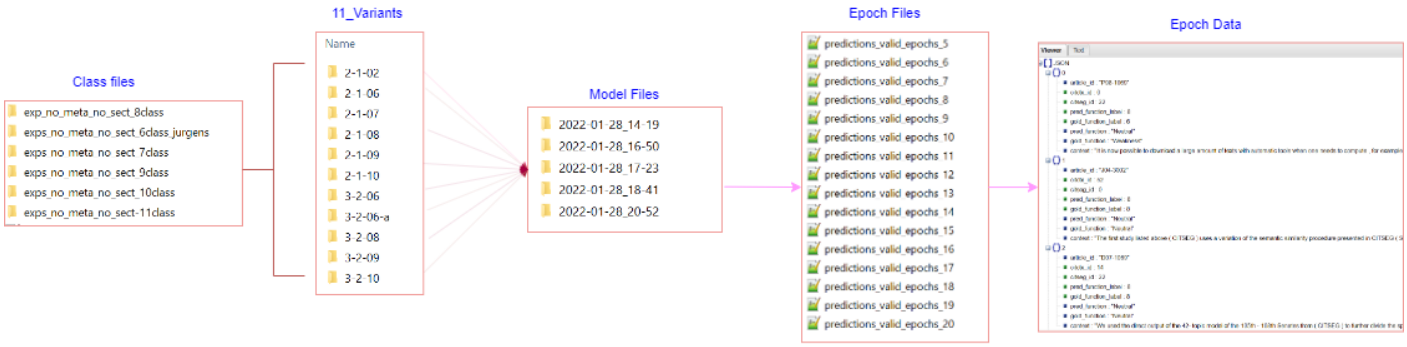


Figure 2: Original Dataset Files Overview

The epoch data is stored in a json format and figure 3 illustrates the structure of the data in json format.

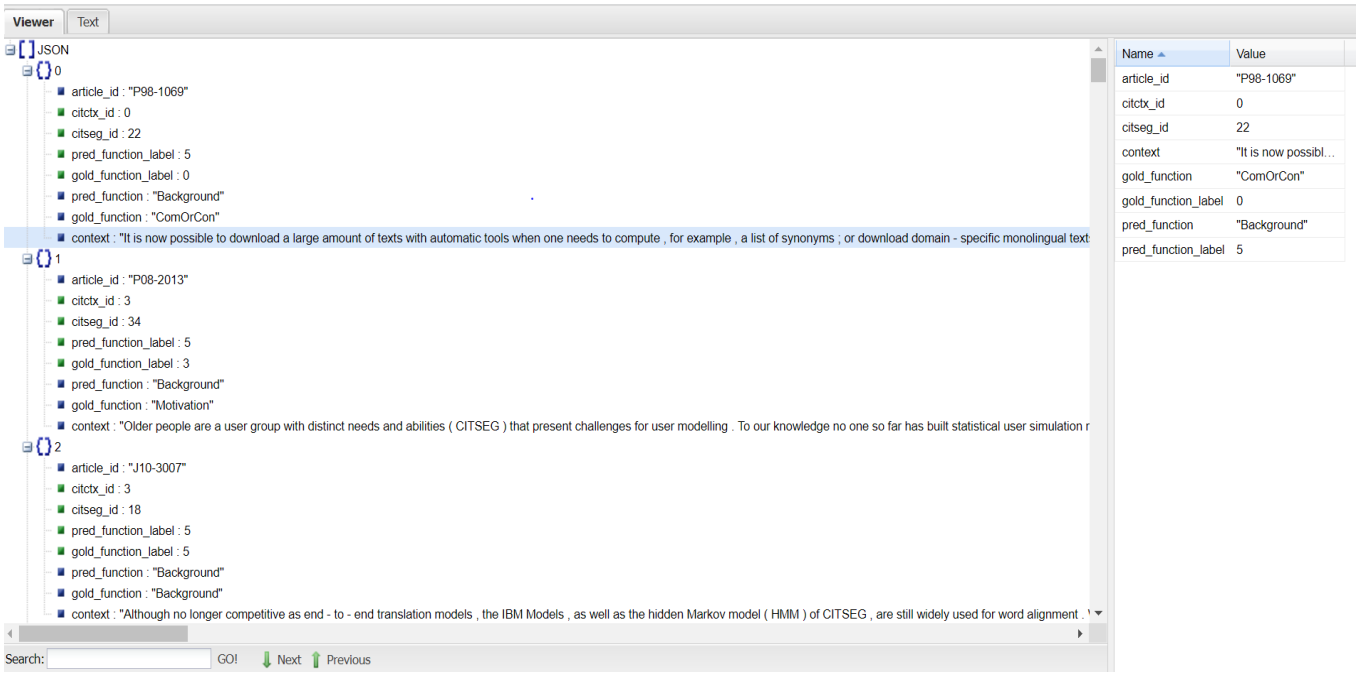


Figure 3: Citation Data shown in a json format

After importing data from a Json file it is stored in a Panda data frame as demonstrated in Figure 4.

	article_id	citctx_id	citseg_id	pred_function_label	gold_function_label	pred_function	gold_function	
0	P98-1069	0	22	5	0	Background	ComOrCon	It is now possible to download a large amount of texts with automatic tool models for bilingual lexicon compilation and machine translation (CITSE
1	P08-2013	3	34	5	3	Background	Motivation	Older people are a user group with distinct needs and abilities (CITSEG) th
2	J10-3007	3	18	5	5	Background	Background	Although no longer competitive as end - to - end translation models , the IBM instance , transferring annotations between languages (CITSEG__TARG
3	N12-1009	30	9	1	1	Uses	Uses	
4	E95-1016	2	33	1	1	Uses	Uses	When considering the prior probability , the more independent of the context
5	W97-1507	3	49	3	3	Motivation	Motivation	Grammar extraction algorithm Systemic Functional Grammar (SFG) (CITS successfully employed in some of the largest and most in

Figure 4: Json data stored in data frame

4.2. Base Classifiers of Citation Semantics Analysis:

A set of SciBERT-based deep learning models are developed for citation semantics analysis. The figure 5 shows the overview of SciBERT-based Citation Semantics Analysis Model. To execute component CFC, the pseudo word "CITSEG" was included to the vocabulary. The CITSEG Encoder employed the CITSEG sequences as the citation representation 'h'. The Context Pooler created the contextual representation 'c' to deal citations that required multi-sentence contexts. And the context window has been fixed to [-2,+3], that means two left and three right phrases. However, Lauscher et al. (2021) found that just a small percentage of citations require contexts longer than six words. The resulting feature vector, f = [h;s;c], was the combination of these three sections. For citation function categorization an MLP (Multiple-Layer Perceptron) was utilised. The citation format is required to differentiate between distinct citations within the same citance, while the citance and context representations are optional. f = [h; c] if just citation context is used. To demonstrate the importance of citation contexts the citance was tested and utilized exclusively i.e. f = [h; s].

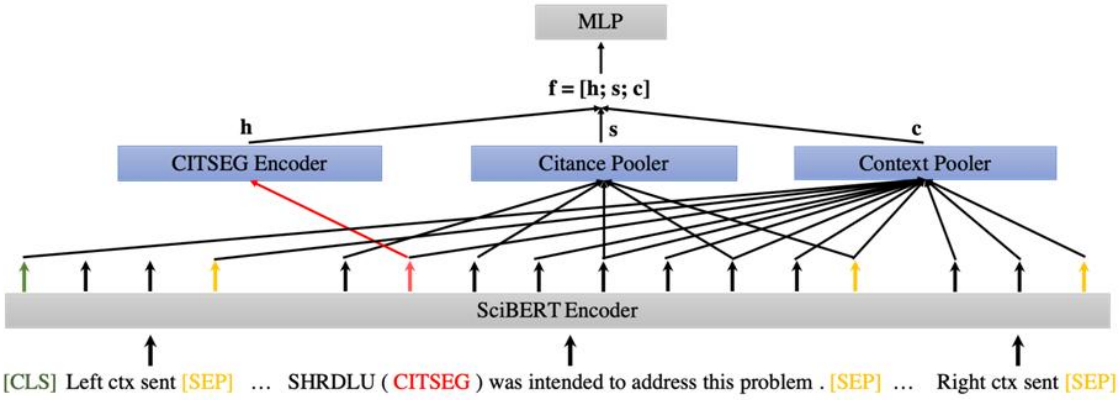


Figure 5: SciBERT-based Citation Semantics Analysis Mode

In conclusion, various factors influenced the construction of the citation semantics analysis model as shown in table 6. The ctx level specified whether the context was hierarchical (ctx level = sentence) or sequential (ctx level = word). The citance pooler and the context pooler are defined by citance and context, respectively. "max pool," "self attend," and "X" were all valid options. Because 'sent_enc' did not apply ("N/A"), Citance and context poolers created feature representations from tokens in a sequential context. In the scenario of a hierarchical context, Sent enc specifies the sentence encoder. max pool," "self attend," and "SEP" were all valid options. Finally, whether CITSEG encoder was utilised (i.e., citseg = O) or not (i.e., citseg= X) was stated by ref. The former required segment-by-segment CFC. The latter was created with the goal of emulating existing deep learning systems that conducted CFC at the citance or context level.

Modelling and Encoding Options					
ID	Ref	Citance	Context	Ctx_level	Sent_enc
1	o	max_pool	max_pool	word	N/A
2	o	x	max_pool	word	N/A
3	o	x	self_attend	word	N/A
4	o	max_pool	x	N/A	N/A
5	o	self_attend	x	N/A	N/A
6	o	x	x	N/A	N/A
2x	x	x	max_pool	word	N/A
3x	x	x	self_attend	word	N/A
4x	x	max_pool	x	word	N/A
5x	x	self_attend	x	word	N/A
6x*	x	x	[CLS]	word	N/A
7	o	max_pool	max_pool	sentence	[SEP]
8	o	x	max_pool	sentence	[SEP]
9	o	x	max_pool	sentence	max_pool
10	o	x	self_attend	sentence	max_pool
11	o	x	self_attend	sentence	[SEP]
12	o	max_pool**	x	N/A	N/A
13	o	self_attend	x	N/A	N/A
14	o	x	x	N/A	N/A

Table 6: Citation Function Classification Performances with Different Annotation Schemes

5. Challenges:

Categorizing citations according to their type is not an easy activity. To begin with, the citing language may not always include the required semantics clues that allow us to detect the citation type. Furthermore, researchers commonly utilize named items, such as the labels of the employed methods, tools, or data repositories, to refer to a previously referenced article later in their paper without directly specifying the citation (Kaplan, Tokunaga & Teufel, 2016). When describing citations, ignoring such indirect citations leads in a loss of information. Authors may employ excessive praise to conceal criticism and prevent adverse citations, as well as a refusal to recognize the usage of a certain approach from past study (Teufel, Siddharthan & Tidhar, 2006a).

6. Methodology

In this report ensemble technique has been chosen to improve the performance. Ensemble techniques are believed to be the best advanced approach for several machine learning problems. This method assists in the reduction of manual annotating time. Thereby, instructing many models and integrating their outputs of such strategies increase the forecasting performance of an individual model.

As part of this study, three procedures are performed.

1. Ensemble technique applied for all epochs of each model seed (Section 6.2).
2. Ensemble technique applied for best epochs of each model seed (Section 6.3).
3. Disagreement measure performed for respective model seeds to check the final accuracy(6.4).

The concept of Ensemble learning's is to merge several models to analyze the performance. The spread or dispersion of the forecasts and reliability of the model is reduced by using an ensemble.

A majority voting ensemble technique is applied to all epochs in each model seed where an each model obtains an ensemble result. Finally, using the majority results of five model seeds once again ensemble executed to get the final output.

Subsequently, we performed ensemble majority voting technique for best epochs of each model seed to test the performance.

Ensemble process in disagreement measure is different when compared to previous steps. Initially, disagreement measures between each pair of classifiers are calculated to combine the all classifiers. Followed by, disagreement matrix has been applied to display the pair of classifier results. Eventually, sum of rows are calculated in disagreement matrix. Therefore, using the sum of subset pairs we choose highest subset pairs to do ensemble.

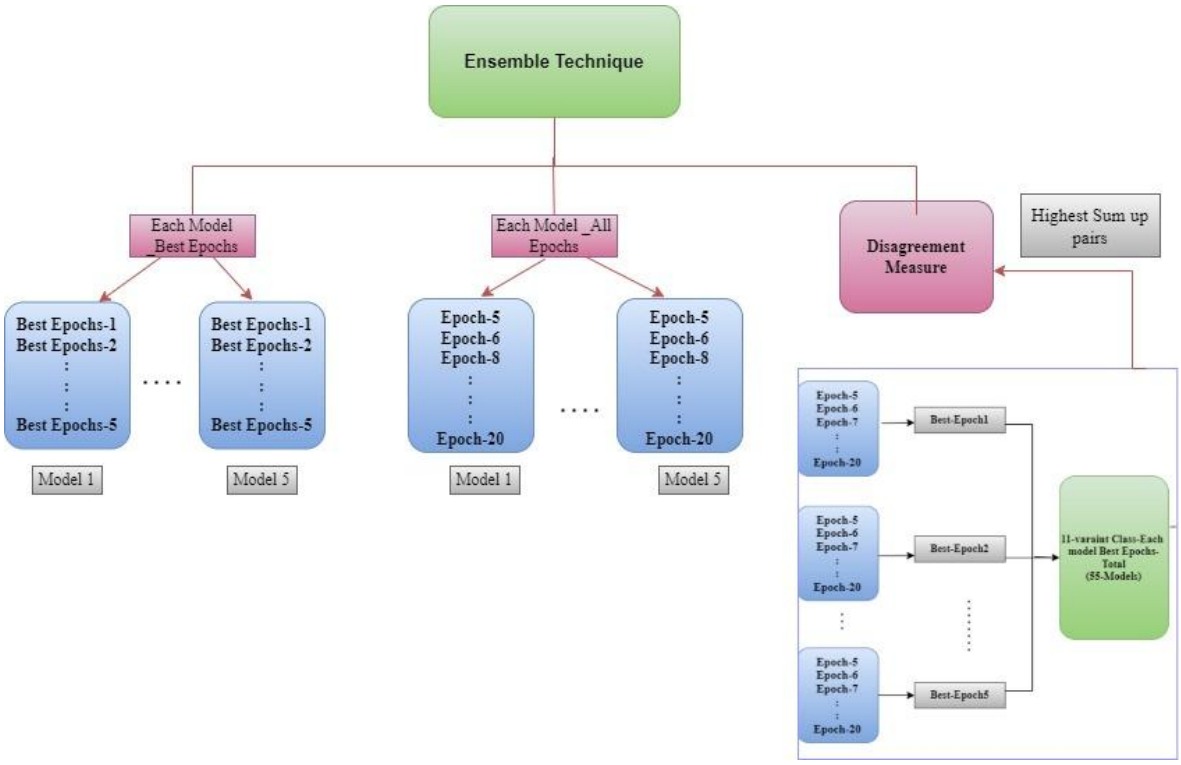


Figure 6: Methodology Overview

The figure 6 represents the complete overview of three procedures in this research. In the first part ensemble applied for best epochs of each model seed to check the performance. In the second part ensemble has been applied for all epochs. Finally disagreement measure applied to all best epochs models highest sum up pairs to check the final performance.

6.1 Ensemble Learning

Ensemble techniques are recognized the advanced approach for several machine learning issues. These strategies increase the predictive performance of a single model by having trained distinct models and integrating their outputs. In this research paper we are applying ensemble techniques to different model outputs to get the high predictive performance.

Ensemble learning refers to strategies that integrate numerous inducers to create a decision, which are commonly used in supervised machine learning problems. Furthermore, most of the ensemble learning studies is focused on homogeneous ensembles, despite the fact that heterogeneous ensembles may be highly effective when integrating pre-trained models, which are typically easily available.

In 1997, Dietterich described how machine-learning studies were progressing in four main ways. Training ensemble of classifiers to increase prediction performance; apply strategies for increasing supervised learning techniques, reinforced learning, and sophisticated unpredictable model learning

Ensemble learning study makes a significant contribution dramatically during the last 30 years. Since 1990 to 2019, the number of articles published in the core set of Science citation Index on the concept of 'ensemble learning' has consistently increased each year, as shown in below figure.

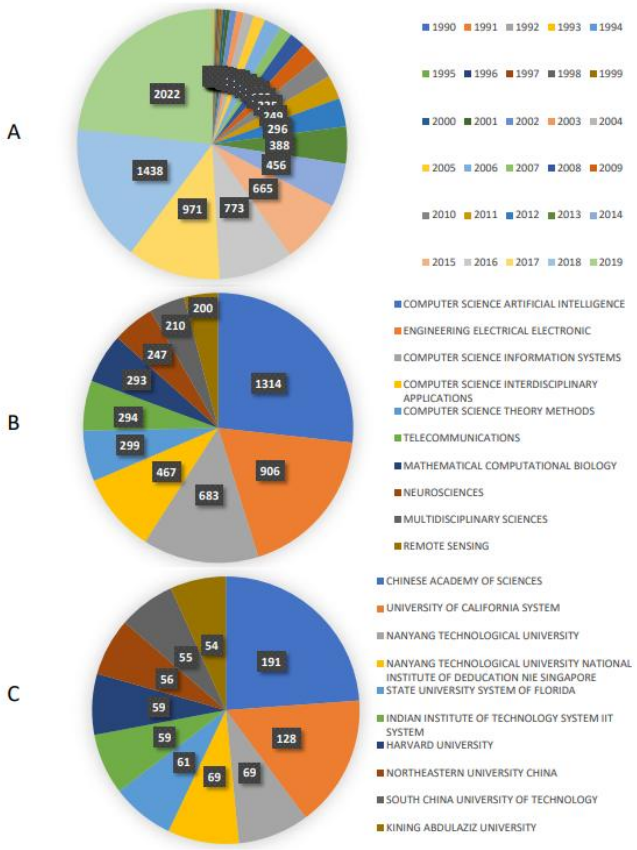


Figure 7: Articles published in the set of Science citation index using ensemble learning

6.1.2 Why Ensemble Learning?

There are two primary reasons to utilise an ensemble over a single model, and both are interrelated.

1. Performance: When compared to single model an ensemble can generate higher predictions

2. Robustness: The spread of forecasts and prediction performance are reduced by using an ensemble.

The primary strategy of ensemble learning is that by merging many models, the inaccuracies of an one inducer will most likely be corrected by other inducers, resulting in the ensemble's total performance evaluation being greater than that of a single inducer. Ensemble learning is divided into two categories based on the learning segment: parallel ensemble and sequential ensemble.

To take use of the qualities of autonomy between the base classifiers, parallel ensemble instruct them in parallel. One of the benefits of parallel ensemble is the process of performing training and prediction on various CPU cores or computers at the same time. Distinct and single learning techniques can be used as base forecasters. These are raising the chance of distinct error kinds due to their varied mathematical foundations. As a result, the total precision of prediction can be enhanced. As demonstrated, a flowchart for multimodal ensemble majority voting.

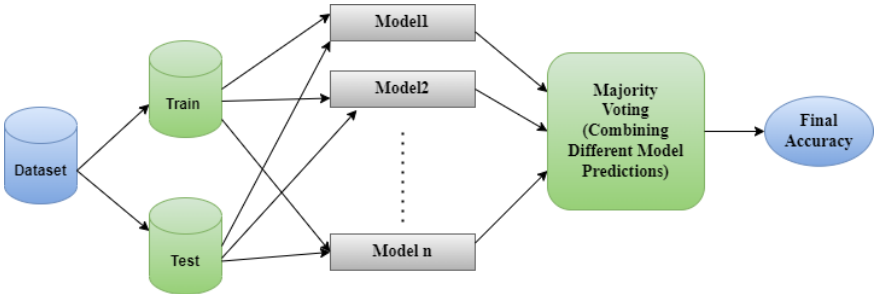


Figure 8: Ensemble process for different models

Uses:

- Ensemble techniques can be used to reduce the issues of class imbalance.
- The allocation of features and labels in several true machine learning systems tends to alter over time. This phenomena commonly has an impact on the model's prediction overall performance. Ensemble-based techniques are frequently used to address this issue.

Three concepts guide the development of ensemble models:

- **Diversity:** The outputs of all trained models are referred to as diversity.
- **Training ensemble members:** Members of a classifier are selected using a procedure for developing classifier members.
- **Combine:** Combing ensemble members to make a decision.

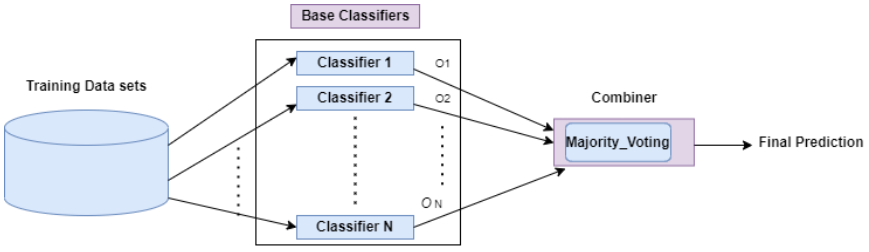


Figure 9: Simple Ensemble Learning System

6.2 Building Ensemble Model For All Epochs:

Initially studies to understand the logs of predictions are made on each validation and test set after each epoch to build an ensemble model. The intermediate outputs of all trained models used in this research were provided by Professor Xiaorui Jiang.

Figure 10, clearly portrays the ensemble overview of current project. Each class file has 11 model variants and every model variant contain five log files with epochs. Each epoch data is stored in a json format and comprises elements like 'Context', Citation Context id, citation segment Id, prediction function label, gold function label, Prediction function, gold function. We have applied majority voting for all epochs of model seed. Eventually, using the five model seed majority results, ensemble process has been implemented to check the final performance.

Rule:

To create an ensemble, we must examine the similarities and differences between the variant model's performances.

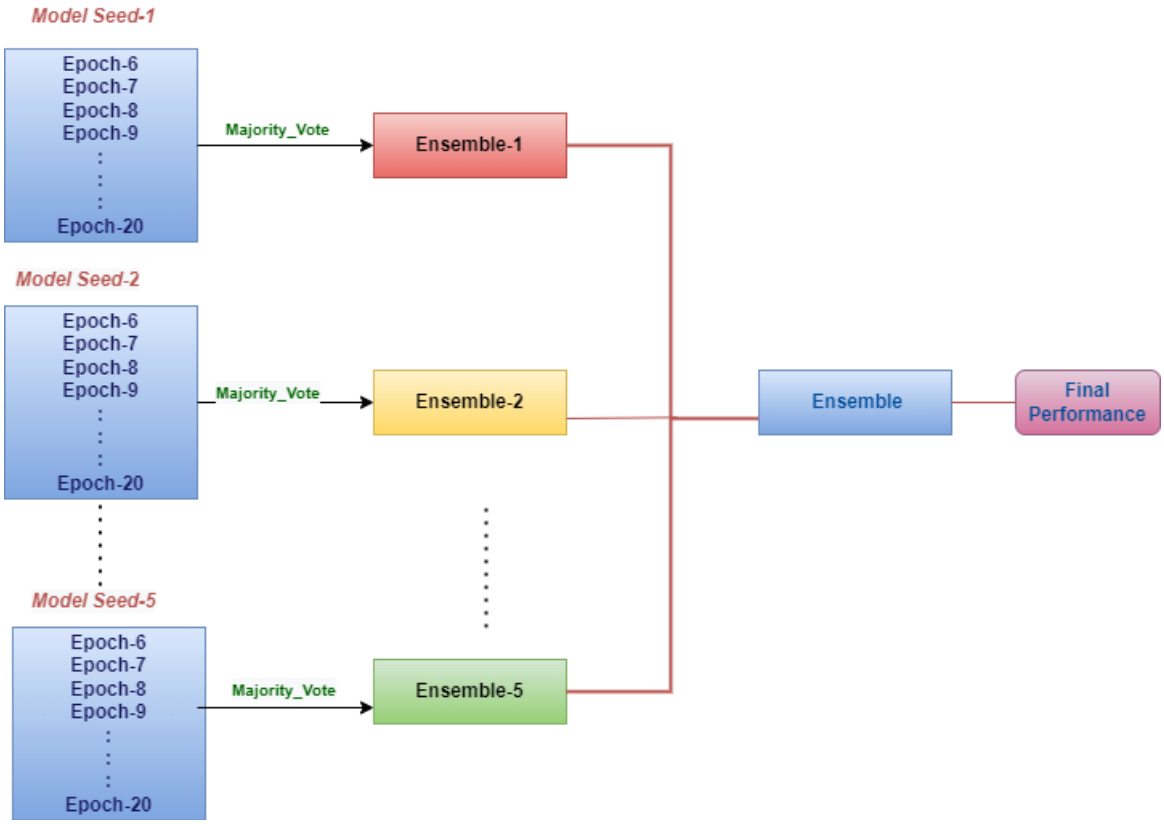
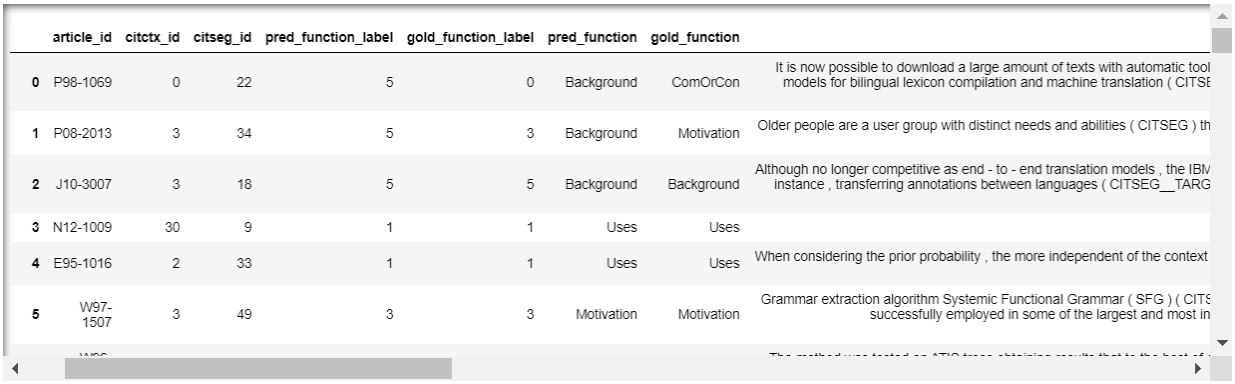


Figure 10: Ensemble process overview for all epochs

As observed in the figure10, Ensemble majority voting has been applied to all epochs of each model seeds. Finally using the five model seed results again ensemble has been applied to check the final performance.

6.2.1 Data stored in a data frame: Initially, reading the data from json file and storing it in a Panda data frame. As illustrated in the image below, the data frame looks like this.

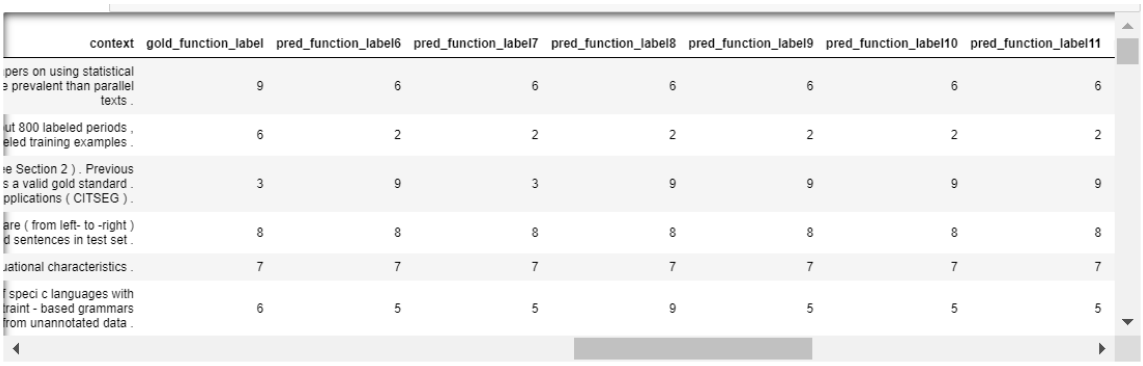


	article_id	citctx_id	citseg_id	pred_function_label	gold_function_label	pred_function	gold_function	
0	P98-1069	0	22	5	0	Background	ComOrCon	It is now possible to download a large amount of texts with automatic tool models for bilingual lexicon compilation and machine translation (CITSE
1	P08-2013	3	34	5	3	Background	Motivation	Older people are a user group with distinct needs and abilities (CITSEG) th
2	J10-3007	3	18	5	5	Background	Background	Although no longer competitive as end - to - end translation models , the IBM instance , transferring annotations between languages (CITSEG__TARG
3	N12-1009	30	9	1	1	Uses	Uses	
4	E95-1016	2	33	1	1	Uses	Uses	When considering the prior probability , the more independent of the context
5	W97-1507	3	49	3	3	Motivation	Motivation	Grammar extraction algorithm Systemic Functional Grammar (SFG) (CITS successfully employed in some of the largest and most in

Figure 11: Json data stored in data frame

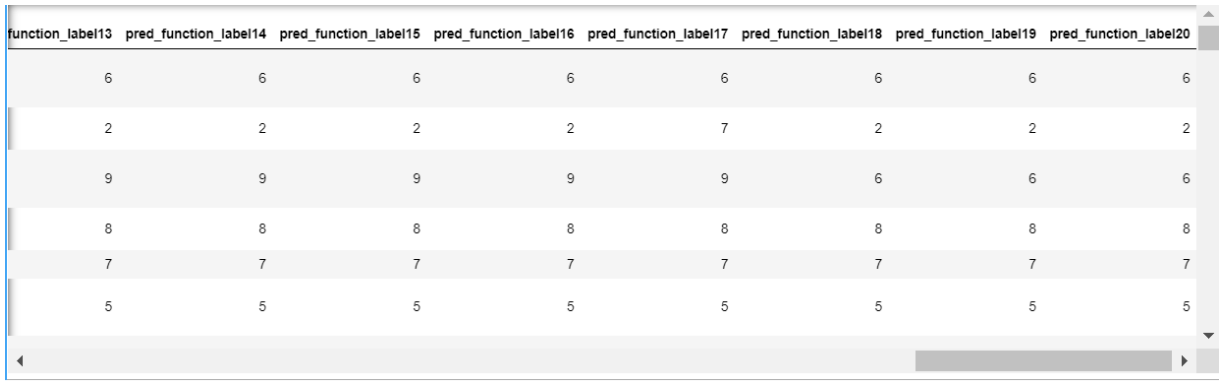
As we observe in the figure 11, all citation data such as article id, citation context id, context, gold function_ label, prediction function, and prediction function labels.

6.2.2 Data Selection: We have different model outcomes in a data frame like as shown in Figure 12 and 13. Here, we have selected context, gold function label and prediction function label (6 to 20) from data.



context	gold_function_label	pred_function_label6	pred_function_label7	pred_function_label8	pred_function_label9	pred_function_label10	pred_function_label11
pers on using statistical e prevalent than parallel texts .	9	6	6	6	6	6	6
ut 800 labeled periods , eled training examples .	6	2	2	2	2	2	2
e Section 2) . Previous s a valid gold standard . pplications (CITSEG) .	3	9	3	9	9	9	9
are (from left- to -right) d sentences in test set .	8	8	8	8	8	8	8
uational characteristics .	7	7	7	7	7	7	7
f speci c languages with traint - based grammars from unannotated data .	6	5	5	9	5	5	5

Figure 12: Prediction labels in data frame 1



function_label13	pred_function_label14	pred_function_label15	pred_function_label16	pred_function_label17	pred_function_label18	pred_function_label19	pred_function_label20
6	6	6	6	6	6	6	6
2	2	2	2	7	2	2	2
9	9	9	9	9	6	6	6
8	8	8	8	8	8	8	8
7	7	7	7	7	7	7	7
5	5	5	5	5	5	5	5

Figure 13: Prediction function label data frame 2

6.2.3 Apply Ensemble Majority Voting:

Majority Voting:

A majority voting is an ensemble machine learning technique that integrates forecasts from several different models. By voting, the majority voting strategy can be utilized to increase predictive accuracy, with the goal of outperforming any one model in the ensemble. It can be used for regression and classification problems. In the regression problems, this effectively takes the average of the models' outputs. When it comes to categorization, the forecasts for each label are added together, and the label with the most votes is chosen. Hard voting and soft voting are two effective ways to predicting the majority vote for categorization.

Hard Voting: Hard voting entails adding up all of the forecasts for each class label and calculating which one will receive the most votes.

Soft Voting: Soft voting entails adding up the anticipated probabilities for every classifier and forecasting the one with the highest likelihood.

When we have multiple models that score well with a predictive analysis, we should use a voting ensemble. The ensemble models must agree on the majority of their forecasts. When compared to single models, a voting ensemble can provide lesser variation in results. This is demonstrated by a smaller variation in regression prediction error. We can also be seen in decreased variance in categorization task accuracy. Considering the model's enhanced stability and confidence, this smaller variation can result in a significant mean performance of the ensemble, which may be desired.

3. Next, using gold function and prediction function label data ‘majority voting’ has been applied to integrate the multiple model outcomes.

```
In [845]: final
```

pred_function_label14	pred_function_label15	pred_function_label16	pred_function_label17	pred_function_label18	pred_function_label19	pred_function_label20	majority_2102
6	6	6	6	6	6	6	6.0
2	2	2	7	2	2	2	2.0
9	9	9	9	6	6	6	9.0
8	8	8	8	8	8	8	8.0
7	7	7	7	7	7	7	7.0
5	5	5	5	5	5	5	5.0

Figure 14: Majority voting result

From figure 14, it is evident that majority voting has been applied successfully for all prediction function labels and stored the majority result.

6.2.4 Performance Check:

Using gold function and Majority voting results performance evaluation matrix has been calculated.

The classification performance of the majority voting result is depicted in the figure 15.

```
: from sklearn.metrics import classification_report
  target_names = ["Basis","CoCoGM","CoCoRes","CoCoxy","Future","Motivation","Neutral","similar","usage","weakness"]
  print(classification_report(final_2102['gold_function_label'], final_2102['majority_2102'],target_names=target_names))
```

```
              precision    recall  f1-score   support

       Basis       0.50      0.40      0.44        25
      CoCoGM       0.69      0.76      0.72        50
     CoCoRes       0.67      0.64      0.65        25
      CoCoxy       0.67      0.35      0.46        23
      Future       0.69      0.69      0.69        13
  Motivation       0.57      0.70      0.63        44
     Neutral       0.78      0.75      0.77       228
     similar       0.71      0.75      0.73        36
       usage       0.70      0.78      0.74       114
    weakness       0.79      0.62      0.70        24

    accuracy                           0.71       582
   macro avg       0.68      0.65      0.65       582
weighted avg       0.71      0.71      0.71       582
```

Figure 15: Classification performance of one model

Similarly, same process is applied for five model seeds epoch data and then using the gold and majority voting results performance matrix has been calculated.

6.2.5 Applying the Ensemble method once again

The results of ensemble majority voting for five model seeds are shown in the image below.

```
Combine
```

	gold_function_label	majority_2102	majority2	majority3	majority4	majority5
0	9	6.0	6.0	6.0	6.0	6.0
1	6	2.0	2.0	2.0	5.0	5.0
2	3	9.0	9.0	9.0	6.0	6.0
3	8	8.0	8.0	8.0	8.0	8.0
4	7	7.0	7.0	7.0	7.0	7.0
5	6	5.0	6.0	6.0	9.0	6.0
6	6	6.0	6.0	6.0	6.0	6.0
7	6	6.0	6.0	6.0	6.0	6.0
8	6	6.0	6.0	6.0	6.0	6.0
9	6	6.0	6.0	6.0	6.0	6.0
10	8	8.0	8.0	8.0	8.0	8.0

Figure 16: Five models ensemble results

Finally, again ensemble technique is applied for all previous majority voting results to get the final result.

```
for i, j in Combine.iterrows():
    lst=[j[1],j[2],j[3],j[4],j[5]]
    Combine.at[Combine.index[i], 'majority_combine'] = int(max(lst,key=lst.count))
```

```
Combine
```

	gold_function_label	majority_2102	majority2	majority3	majority4	majority5	majority_combine
0	9	6.0	6.0	6.0	6.0	6.0	6.0
1	6	2.0	2.0	2.0	5.0	5.0	2.0
2	3	9.0	9.0	9.0	6.0	6.0	9.0
3	8	8.0	8.0	8.0	8.0	8.0	8.0
4	7	7.0	7.0	7.0	7.0	7.0	7.0
5	6	5.0	6.0	6.0	9.0	6.0	6.0
6	6	6.0	6.0	6.0	6.0	6.0	6.0
7	6	6.0	6.0	6.0	6.0	6.0	6.0
8	6	6.0	6.0	6.0	6.0	6.0	6.0
9	6	6.0	6.0	6.0	6.0	6.0	6.0
10	8	8.0	8.0	8.0	8.0	8.0	8.0

Figure 17: Combining all five ensemble results

Classification report has been performed to gold function label and final majority result.

```
from sklearn.metrics import classification_report
target_names = ["Basis","CoCoGM","CoCoRes","CoCoxy","Future","Motivation","Neutral","similar","usage","weakness"]
print(classification_report(Combine['gold_function_label'], Combine['majority_combine'],digits=4,target_names=target_names))
```

```
              precision    recall  f1-score   support

       Basis     0.5333    0.3200    0.4000        25
     CoCoGM      0.7660    0.7200    0.7423        50
     CoCoRes     0.6923    0.7200    0.7059        25
      CoCoxy     0.6875    0.4783    0.5641        23
      Future     0.9000    0.6923    0.7826        13
  Motivation     0.5882    0.6818    0.6316        44
     Neutral     0.7773    0.8114    0.7940       228
     similar     0.7714    0.7500    0.7606        36
       usage     0.7280    0.7982    0.7615       114
    weakness     0.7895    0.6250    0.6977        24

    accuracy                         0.7388       582
   macro avg     0.7234    0.6597    0.6840       582
weighted avg     0.7376    0.7388    0.7348       582
```

Figure 18: Classification report for final ensemble result

Same ensemble process is applied for all 11 model variants five model seed epochs to check the performance and final results are shown in below table.

Table 7: Class 6 all epochs 11 variants final ensemble performance

Class-6 All 11 Variants Ensemble Result		Ensemble 2_1_02	Ensemble 2_1_06	Ensemble 2_1_07	Ensemble 2_1_08	Ensemble 2_1_09	Ensemble 2_1_10	Ensemble 3_2_06	Ensemble 3_2_06_a	Ensemble 3_2_08	Ensemble 3_2_09	Ensemble 3_2_10
ComorCon	Precision	79.56	76	75.84	78.47	79.58	77.93	76.19	77.55	79.45	77.08	76.16
	Recall	76.22	79.72	79.02	79.02	79.02	79.02	78.32	79.72	81.12	77.62	80.42
	F1-Score	77.86	77.82	77.4	78.75	79.3	78.47	77.24	78.62	80.28	77.35	78.23
Uses	Precision	71.09	72.95	77.78	75.63	78.26	78.76	77.31	77.31	78.99	79.17	79.13
	Recall	79.82	78.07	79.82	78.95	78.95	78.07	80.7	80.7	82.46	83.33	79.82
	F1-Score	75.21	75.42	78.79	77.25	78.6	78.41	78.97	78.97	80.69	81.2	79.48
Extends	Precision	50	53.85	81.82	75	72.22	71.43	78.57	64.71	60	92.31	70.59
	Recall	28	28	36	36	52	40	44	44	48	48	48
	F1-Score	35.9	36.84	50	48.65	60.47	51.28	56.41	52.38	53.33	63.16	57.14
Motivation	Precision	64.29	70	67.5	64.29	65.12	66.67	65.85	63.04	71.79	65.12	70.73
	Recall	61.36	63.64	61.36	61.36	63.64	63.64	61.36	65.91	63.64	63.64	65.91
	F1-Score	62.79	66.67	64.29	62.79	64.37	65.12	63.53	64.44	67.47	64.37	68.24
Future	Precision	75	83.33	75	90	90	81.82	69.23	76.92	83.33	90.91	84.62
	Recall	69.23	76.92	69.23	69.23	69.23	69.23	69.23	76.92	76.92	76.92	84.62
	F1-Score	72	80	72	78.26	78.26	75	69.23	76.92	80	83.33	84.62
Background	Precision	77.91	77.96	78.26	78.43	77.95	78.6	78.23	80	80.49	78.09	80.41
	Recall	79.84	78.6	81.48	82.3	81.48	83.13	79.84	79.01	81.48	80.66	81.07
	F1-Score	78.86	78.28	79.84	80.32	79.68	80.8	79.02	79.5	80.98	79.35	80.74
Macro_Avg		67.1	69.17	70.38	71	73.45	71.51	70.73	71.81	73.79	74.79	74.74
Best		69	69	72	70	69	71	70	72	71	72	73

Class-7		Ensemble 2_1_02	Ensemble 2_1_06	Ensemble 2_1_07	Ensemble 2_1_08	Ensemble 2_1_09	Ensemble 2_1_10	Ensemble 3_2_06	Ensemble 3_2_06_a	Ensemble 3_2_08	Ensemble 3_2_09	Ensemble 3_2_10
Comparision	Precision	78.12	75	80	79.41	75.71	78.57	71.62	76.81	80	75.34	77.78
	Recall	68.49	73.97	76.71	73.97	72.6	75.34	72.6	72.6	76.71	75.34	76.71
	F1-Score	72.99	74.48	78.32	76.6	74.13	76.92	72.11	74.65	78.32	75.34	77.24
Similar	Precision	60.38	63.64	71.74	68	68.09	69.57	72.73	68.89	74.47	68.89	71.74
	Recall	69.57	60.87	71.74	73.91	69.57	69.57	69.57	67.39	76.09	67.39	71.74
	F1-Score	64.65	62.22	71.74	70.83	68.82	69.57	71.11	68.13	75.27	68.13	71.74
Uses	Precision	76.47	80.17	77.31	80	80	78.76	73.39	77.19	76.27	78.51	76.47
	Recall	79.82	81.58	80.7	80.7	80.7	78.07	79.82	77.19	78.95	83.33	79.82
	F1-Score	78.11	80.87	78.97	80.35	80.35	78.41	76.47	77.19	77.59	80.85	78.11
Extends	Precision	64.29	60	61.11	61.9	73.68	70	69.23	60	57.89	85.71	72.22
	Recall	36	36	44	52	56	56	36	60	44	48	52
	F1-Score	46.15	45	51.16	56.52	63.64	62.22	47.37	60	50	61.54	60.47
Motivation	Precision	68.42	71.05	68.42	60.47	70.73	65.85	66.67	75	75	71.79	73.68
	Recall	59.09	61.36	59.09	59.09	65.91	61.36	59.09	54.55	61.36	63.64	63.64
	F1-Score	63.41	65.85	63.41	59.77	68.24	63.53	62.65	63.16	67.5	67.47	68.29
Future	Precision	90	90	90	90.91	90	83.33	83.33	90.91	83.33	90.91	90.91
	Recall	69.23	69.23	69.23	76.92	69.23	76.92	76.92	76.92	76.92	76.92	76.92
	F1-Score	78.26	78.26	78.26	83.33	78.26	80	80	83.33	80	83.33	83.33
Background	Precision	81.34	79.09	82.21	82.12	83.21	81.79	79.71	80.42	81.43	82.08	81.65
	Recall	86.52	85.02	86.52	84.27	87.27	85.77	82.4	86.14	85.39	85.77	85.02
	F1-Score	83.85	81.95	84.31	83.18	85.19	83.73	81.03	83.18	83.36	83.88	83.3
Macro_Avg		69.63	69.81	72.31	72.94	74.09	73.48	70.11	72.81	73.15	74.36	74.64
Best		69	69	71	70	71	71	70	70	71	72	75

Table 8: Class 7 all epochs 11 variants final ensemble performance

Table 9: Class-8 all epochs 11 variants final ensemble performance

Class-8		Ensemble 2_1_02	Ensemble 2_1_06	Ensemble 2_1_07	Ensemble 2_1_08	Ensemble 2_1_09	Ensemble 2_1_10	Ensemble 3_2_06	Ensemble 3_2_06_a	Ensemble 3_2_08	Ensemble 3_2_09	Ensemble 3_2_10
ComorCon	Precision	79.41	73.68	82.09	76.39	81.16	81.16	78.26	79.17	81.43	78.38	79.17
	Recall	72	74.67	73.33	73.33	74.67	74.67	72	76	76	77.33	76
	F1-Score	75.52	74.17	77.46	74.83	77.78	77.78	75	77.55	78.62	77.85	77.55
Similar	Precision	74.29	68.42	66.67	72.22	75	72.22	63.16	71.43	72.22	73.53	71.05
	Recall	72.22	72.22	72.22	72.22	75	72.22	66.67	69.44	72.22	69.44	75
	F1-Score	73.24	70.27	69.33	72.22	75	72.22	64.86	70.42	72.22	71.43	72.97
Uses	Precision	76.27	76.27	80.34	79.13	76.03	79.17	77.5	77.31	76.07	76.03	79.65
	Recall	78.95	78.95	82.46	79.82	80.7	83.33	81.58	80.7	78.07	80.7	78.95
	F1-Score	77.59	77.59	81.39	79.48	78.3	81.2	79.49	78.97	77.06	78.3	79.3
Basis	Precision	66.67	70.59	81.82	64.29	62.5	61.11	71.43	70	66.67	71.43	81.25
	Recall	40	48	36	36	40	44	40	28	48	40	52
	F1-Score	50	57.14	50	46.15	48.78	51.16	51.28	40	55.81	51.28	63.41
Motivation	Precision	60.47	73.17	69.44	66.67	64.29	73.17	65	65.91	67.5	70	78.38
	Recall	59.09	68.18	56.82	63.64	61.36	68.18	59.09	65.91	61.36	63.64	65.91
	F1-Score	59.77	70.59	62.5	65.12	62.79	70.59	61.9	65.91	64.29	66.67	71.6
Future	Precision	75	78.57	81.82	81.82	81.82	71.43	91.67	81.82	81.82	76.92	71.43
	Recall	69.23	84.62	69.23	69.23	69.23	76.92	84.62	69.23	69.23	76.92	76.92
	F1-Score	72	81.48	75	75	75	74.07	88	75	75	76.92	74.07
Weak	Precision	76.47	92.31	81.25	76.47	81.25	72.22	61.9	72.22	73.68	72.22	72.22
	Recall	54.17	50	54.17	54.17	54.17	54.17	54.17	54.17	58.33	54.17	54.17
	F1-Score	63.41	64.86	65	63.41	65	61.9	57.78	61.9	65.12	61.9	61.9
Background	Precision	77.37	79.62	77.19	78.55	77.86	81.2	77.99	78.75	77.12	80.6	79.56
	Recall	84.46	84.06	87.65	86.06	84.06	86.06	83.27	85.66	83.27	86.06	86.85
	F1-Score	80.76	81.78	82.09	82.13	80.84	83.56	80.54	82.06	80.08	83.24	83.05
Macro_Avg		69.04	72.24	70.35	69.79	70.44	71.56	69.86	68.98	71.02	70.95	72.98
Best		68	71	72	70	70	73	70	70	70	69	71

Class-9		Ensemble 2_1_02	Ensemble 2_1_06	Ensemble 2_1_07	Ensemble 2_1_08	Ensemble 2_1_09	Ensemble 2_1_10	Ensemble 3_2_06	Ensemble 3_2_06_a	Ensemble 3_2_08	Ensemble 3_2_09	Ensemble 3_2_10
Comparison	Precision	78.87	74.29	75	84.13	73.33	79.41	77.78	79.71	76.71	79.45	79.71
	Recall	76.71	71.23	78.08	72.6	75.34	73.97	76.71	75.34	76.71	79.45	75.34
	F1-Score	77.78	72.73	76.51	77.94	74.32	76.6	77.24	77.46	76.71	79.45	77.46
Similar	Precision	70.97	63.64	69.7	74.19	75.76	72.73	74.19	70	70.97	68.97	68.75
	Recall	70.97	67.74	74.19	74.19	80.65	77.42	74.19	67.74	70.97	64.52	70.97
	F1-Score	70.97	65.62	71.88	74.19	78.12	75	74.19	68.85	70.97	66.67	69.84
Usage	Precision	79.28	79.46	74.8	78.63	79.46	80.7	77.39	79.31	78.95	76.47	80
	Recall	77.19	78.07	80.7	80.7	78.07	80.7	78.07	80.7	78.95	79.82	80.7
	F1-Score	78.22	78.76	77.64	79.65	78.76	80.7	77.73	80	78.95	78.11	80.35
Basis	Precision	55.56	66.67	73.33	70.59	63.16	64.71	66.67	73.68	75	64.71	54.55
	Recall	40	48	44	48	48	44	48	56	48	44	48
	F1-Score	46.51	55.81	55	57.14	54.55	52.38	55.81	63.64	58.54	52.38	51.06
Motivation	Precision	61.36	63.64	71.79	69.23	63.64	71.43	55.77	63.27	63.64	70.73	69.23
	Recall	61.36	63.64	63.64	61.36	63.64	68.18	65.91	70.45	63.64	65.91	61.36
	F1-Score	61.36	63.64	67.47	65.06	63.64	69.77	60.42	66.67	63.64	68.24	65.06
Support	Precision	50	53.85	77.78	43.75	50	56.25	57.14	40	66.67	61.54	60
	Recall	46.67	46.67	46.47	46.67	53.33	60	53.33	53.33	40	53.33	40
	F1-Score	48.28	50	58.33	45.16	51.61	58.06	55.17	45.71	50	57.14	48
Weakness	Precision	60	73.68	76.47	86.67	77.78	63.16	73.33	77.78	72.22	76.47	61.9
	Recall	50	58.33	54.17	54.17	58.33	50	45.83	58.33	54.17	54.17	54.17
	F1-Score	54.55	65.12	63.41	66.67	66.67	55.81	56.41	66.67	61.9	63.41	57.78
Future	Precision	90	90	81.82	90	90	69.23	90	90	90	83.33	81.82
	Recall	69.23	69.23	69.23	69.23	69.23	69.23	69.23	69.23	69.23	76.92	69.23
	F1-Score	78.26	78.26	75	78.26	78.26	69.23	78.26	78.26	78.26	80	75
Natural	Precision	76.43	76.81	78.38	76.28	79.61	78.46	78.82	80.08	77.15	77.78	76.81
	Recall	82.72	83.13	83.54	86.01	83.54	83.95	82.72	82.72	84.77	83.54	83.13
	F1-Score	79.45	79.84	80.88	80.85	81.53	81.11	80.72	81.38	80.78	80.56	79.84
Macro_Avg		66.15	67.75	69.57	69.44	69.72	68.74	68.44	69.85	68.86	69.55	67.16
Best		69	69	68	67	68	68	69	68	68	68	67

Table 10: Class-9 all epochs 11 variants final ensemble

Table 11: Class-10 all epochs 11 variants final ensemble performance

Class-10		Ensemble 2_1_02	Ensemble 2_1_06	Ensemble 2_1_07	Ensemble 2_1_08	Ensemble 2_1_09	Ensemble 2_1_10	Ensemble 3_2_06	Ensemble 3_2_06_a	Ensemble 3_2_08	Ensemble 3_2_09	Ensemble 3_2_10
Basis	Precision	53.33	64.29	66.67	60	61.9	68.75	68.75	71.43	66.67	71.43	70.59
	Recall	32	36	40	36	52	44	44	40	48	60	48
	F1-Score	40	46.15	50	45	56.52	53.66	53.66	51.28	55.81	65.22	57.14
CoCoGM	Precision	76.6	79.17	81.63	79.07	77.78	76.6	77.08	67.24	75	75	70.91
	Recall	72	76	80	68	70	72	74	78	78	72	78
	F1-Score	74.23	77.55	8081	73.12	73.68	74.23	75.51	72.22	76.47	73.47	74.29
CoCoRes	Precision	69.23	73.08	64.29	75	70.83	72	65.52	70.83	61.29	60.71	73.91
	Recall	72	76	73	72	68	72	76	68	76	68	68
	F1-Score	70.59	74.51	67.92	73.47	69.39	72	70.37	69.39	67.86	64.15	70.83
CoCoXY	Precision	68.75	73.33	71.43	61.9	66.67	70.59	62.5	69.23	69.23	71.43	64.29
	Recall	47.83	47.83	43.48	56.52	43.48	52.17	43.48	39.13	39.13	43.48	39.13
	F1-Score	56.41	57.89	54.05	59.09	52.63	60	51.28	50	50	54.05	48.65
Future	Precision	90	69.23	90	90	75	81.82	90	81.82	90	81.82	81.82
	Recall	69.23	69.23	69.23	69.23	69.23	69.23	69.23	69.23	69.23	69.23	69.23
	F1-Score	78.26	69.23	78.26	78.26	72	75	78.26	75	78.26	75	75
Motivation	Precision	58.82	65.12	63.83	59.18	68.29	59.57	58.33	64.44	70	57.14	65.91
	Recall	68.18	63.64	68.18	65.91	63.64	63.64	63.64	65.91	63.64	63.64	65.91
	F1-Score	63.16	64.37	65.93	62.37	65.88	61.54	60.87	65.17	66.67	60.22	65.91
Neutral	Precision	77.73	76.4	79.25	76.98	77.96	76.45	78.15	79.49	78.99	78.63	79.32
	Recall	81.14	83.77	83.77	84.65	83.77	81.14	81.58	81.58	82.46	80.7	82.46
	F1-Score	79.4	79.92	81.45	80.58	80.76	78.72	79.83	80.52	80.69	79.65	80.86
Similar	Precision	77.14	64.86	72.97	69.23	68.42	70	69.44	74.29	76.47	66.67	68.42
	Recall	75	66.67	75	75	72.22	77.78	69.44	72.22	72.22	66.67	72.22
	F1-Score	76.06	65.75	73.97	72	70.27	73.68	69.44	73.24	74.29	66.67	70.27
Usage	Precision	72.8	72.88	76.03	78.95	75.83	77.69	75.61	77.69	77.87	75.21	77.87
	Recall	79.82	75.44	80.7	78.95	79.82	82.46	81.58	82.46	83.33	79.82	83.33
	F1-Score	76.15	74.14	78.3	78.95	77.78	80	78.48	80	80.51	77.45	80.51
Weakness	Precision	78.95	72.22	75	81.25	66.67	75	77.78	59.26	58.33	70	71.43
	Recall	62.5	54.17	62.5	54.17	58.33	50	58.33	66.67	58.33	58.33	62.5
	F1-Score	69.77	61.9	68.18	65	62.22	60	66.67	62.75	58.33	63.64	66.67
Macro_AVG		68.4	67.14	69.89	68.78	68.11	68.88	68.44	67.96	68.89	67.95	69.01
Best		67	66	71	68	68	68	69	67	69	69	68

Class-11		Ensemble 2_1_02	Ensemble 2_1_06	Ensemble 2_1_07	Ensemble 2_1_08	Ensemble 2_1_09	Ensemble 2_1_10	Ensemble 3_2_06	Ensemble 3_2_06_a	Ensemble 3_2_08	Ensemble 3_2_09	Ensemble 3_2_10
Basis	Precision	64.71	66.67	66.67	63.16	57.14	70.59	68.42	75	60	61.11	68.18
	Recall	44	48	40	48	48	48	52	48	48	44	60
	F1-Score	52.38	55.81	50	54.55	52.17	57.14	59.09	58.54	53.33	51.16	63.83
CoCoGM	Precision	75	76	75	70.59	73.47	73.58	74.47	72.55	75.47	77.08	72.73
	Recall	66	76	78	72	72	78	70	74	80	74	80
	F1-Score	70.21	76	76.47	71.29	72.73	75.73	72.16	73.27	77.67	75.51	76.19
CoCoRes	Precision	72	69.23	60.71	68	65.38	60.71	62.96	65.38	66.67	60.71	65.38
	Recall	78.26	78.26	73.91	73.91	73.91	73.91	73.91	73.91	78.26	73.91	73.91
	F1-Score	75	73.47	66.67	70.83	69.39	66.67	68	69.39	72	66.67	69.39
CoCoXY	Precision	58.33	65	69.23	55.56	66.67	62.5	75	64.71	68.75	52.94	61.54
	Recall	60.87	56.52	39.13	43.48	34.78	43.48	52.17	47.83	47.83	39.13	34.78
	F1-Score	59.57	60.47	50	48.78	45.71	51.28	61.54	55	56.41	45	44.44
Future	Precision	81.82	90	90	90	90	90	81.82	90	81.82	81.82	90.91
	Recall	69.23	69.23	69.23	69.23	69.23	69.23	69.23	69.23	69.23	69.23	76.92
	F1-Score	75	78.26	78.26	78.26	78.26	78.26	75	78.26	75	75	83.33
Motivation	Precision	59.57	59.62	58.49	65.22	62.5	65.96	63.04	65.85	69.05	68.89	60.42
	Recall	63.64	70.45	70.45	68.18	68.18	70.45	65.91	61.36	65.91	70.45	65.91
	F1-Score	61.54	64.58	63.92	66.67	65.22	68.13	64.44	63.53	67.44	69.66	63.04
Neutral	Precision	76.19	77.83	77.49	75.86	76.07	77.53	77.49	74.15	77.02	77.73	78.73
	Recall	80	81.36	81.36	80	80.91	80	81.36	79.55	82.27	80.91	79.09
	F1-Score	78.05	79.56	79.38	77.88	78.41	78.75	79.38	76.75	79.56	79.29	78.91
Similar	Precision	75.86	70.97	71.88	70.59	70	68.57	70.97	73.33	74.29	66.67	70.97
	Recall	70.97	70.97	74.19	77.42	67.74	77.42	70.97	70.97	83.87	70.97	70.97
	F1-Score	73.33	70.97	73.02	73.85	68.85	72.73	70.97	72.13	78.79	68.75	70.97
Support	Precision	57.14	57.14	70	58.33	58.33	54.55	53.33	61.54	77.78	53.33	60
	Recall	53.33	53.33	46.67	46.67	46.67	40	53.33	53.33	46.67	53.33	40
	F1-Score	55.17	55.17	56	51.85	51.85	46.15	53.33	57.14	58.33	53.33	48
Usage	Precision	74.8	82.14	75.41	78.07	74.38	75.21	77.97	75	78.76	75	74.8
	Recall	80.7	80.7	80.7	78.07	78.95	79.82	80.7	81.58	78.07	78.95	80.7
	F1-Score	77.64	81.42	77.97	78.07	76.6	77.45	79.31	78.15	78.41	76.92	77.64
Weakness	Precision	70.59	78.95	87.5	66.67	73.68	82.35	66.67	72.22	71.43	77.78	72.73
	Recall	50	62.5	58.33	58.33	58.33	58.33	58.33	54.17	62.5	58.33	66.67
	F1-Score	58.54	69.77	70	62.22	65.12	68.29	62.22	61.9	66.67	66.67	69.57
Macro_Avg		66.95	69.59	67.43	66.75	65.85	67.33	67.77	67.64	69.42	66.18	67.76
Best		66	69	67	65	67	65	66	66	68	64	69

Table 12: Class-11 all epochs 11 variants final ensemble performance

The Table 7 to Table 12 shows the ensemble results of all 11 model variants five model seed epochs.

As observed in the all results tables (Table 7 to 12) Class 6 and 7 got high performance (Table 6,7) when compared to other classes. The below confusion matrix shows the performance of these classes and also class-11

Class -6 best performance confusion matrix

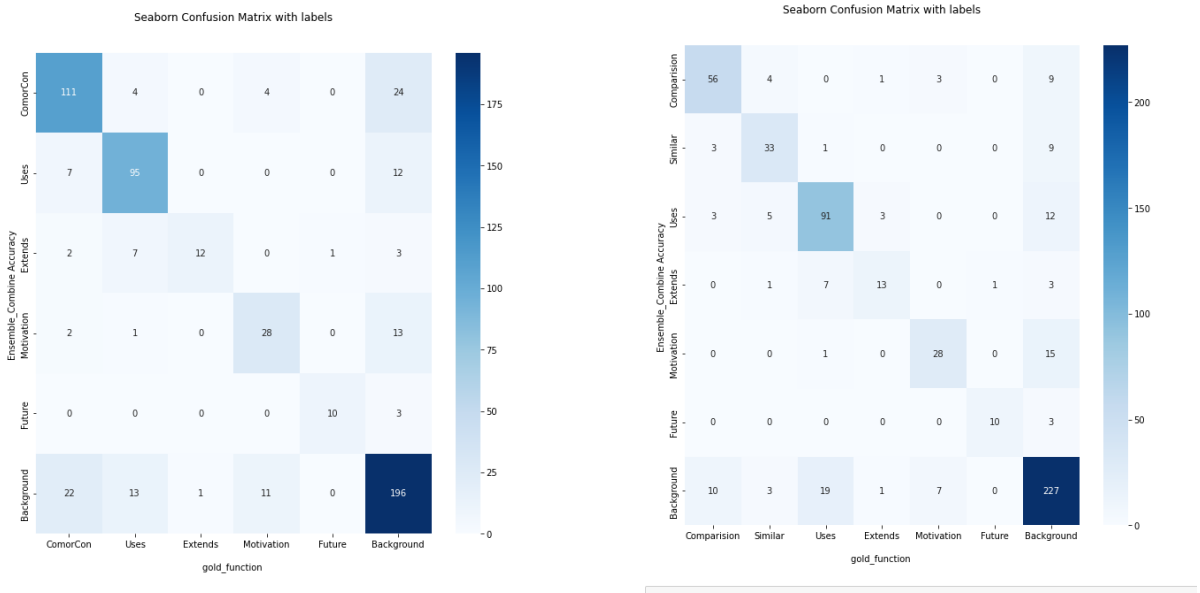


Figure19: Ensemble class-6 and 7 best performance confusion matrix

Class-11 best performance confusion matrix

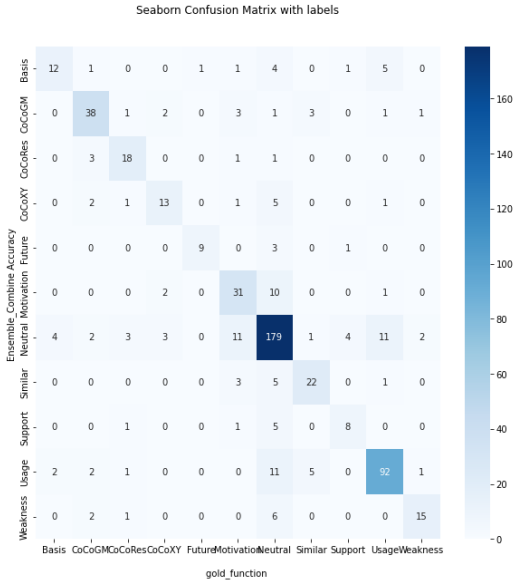


Figure 20: Class-11 ensemble al epoch performance

6.3 Building Ensemble Technique for Best Epochs

Each class file 11 model variants have five model seeds with 15 epochs. Here we have chosen best epochs according to seeds of respective model files, then applied enemble majority voting to all five best epochs to check the final performance.

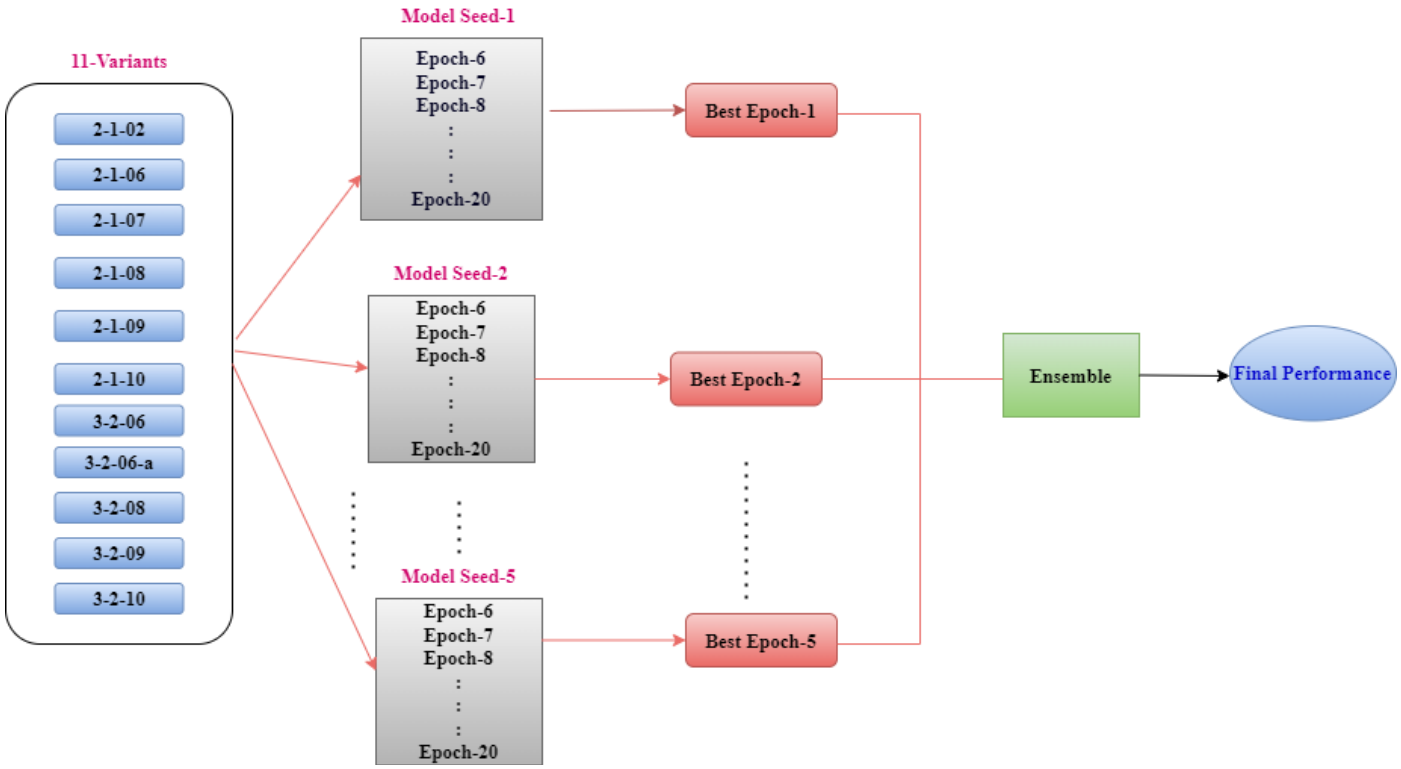


Figure 21: Best Epochs Ensemble Process

As observed in the figure 21, each model has 20 epochs. Ensemble majority voting has been applied to each model best epochs to evaluate the final performance.

The following steps are done to build an ensemble for best epochs.

6.3.1: Initially one beast epoch is chosen form model seed1 and read the epoch data and stored it in one data frame. The following Figure 22 shows the first model seed best epoch data.

```
d1['predictions_valid_epochs_13.json']
```

	pred_function_label	gold_function_label	
0	5	0	It is now possible to download a large amount of texts with automatic tools when one needs to compute , for example , a list models for bilingual lexicon compilation and machine translation (CITSEG__TARGET) , large amount of human effort and
1	5	3	Older people are a user group with distinct needs and abilities (CITSEG) that present challenges for user modelling . To our kr for both younger and older
2	5	5	Although no longer competitive as end - to - end translation models , the IBM Models , as well as the hidden Markov model (for instance , transferring annotations between languages (CITSEG__TARGET) ; discovery of paraphrases (CITSEG) ; ar the tendency of each source word to
3	1	1	We performed a
4	1	1	When considering the prior probability , the more independent of the context it is the better to measure actual associations .
5	3	3	Grammar extraction algorithm Systemic Functional Grammar (SFG) (CITSEG) is based on the assumption that the differenti successfully employed in some of the largest and most influential text generation projects -- such as , for exa

Figure 22: Chosen Model1 best epoch

Next, we have taken 'predictions_valid_epochs_13.json' best epoch data form model seed/file one to do ensemble. Before performing ensemble, similar process should be completed for remaining four models best epochs.

```
d2['predictions_valid_epochs_8.json']
```

	pred_function_label	gold_function_label	context
0	5	0	It is now possible to download a large amount of texts with automatic tools when one needs to compute , for example , a list of synonyms ; or download domain - specific monolingual texts by specifying a keyword to the search engine , and then use this text to extract domain - specific terms . It remains to be seen how we can also make use of the multilingual texts as NLP resources . In the years since the appearance of the first papers on using statistical models for bilingual lexicon compilation and machine translation (CITSEG__TARGET) , large amount of human effort and time has been invested in collecting parallel corpora of translated texts . Our goal is to alleviate this effort and enlarge the scope of corpus resources by looking into monolingual , comparable texts . This type of texts are known as nonparallel corpora . Such nonparallel , monolingual texts should be much more prevalent than parallel texts .
1	5	3	Older people are a user group with distinct needs and abilities (CITSEG) that present challenges for user modelling . To our knowledge no one so far has built statistical user simulation models for older people . The only statistical spoken dialogue system for older people we are aware of is Nursebot , an early application of statistical methods (POMDPs) within the context of a medication reminder system (CITSEG__TARGET) . In this study , we build SUs for both younger and older adults using n-grams . Our data comes from a fully annotated corpus of 447 interactions of older and younger users with a Wizard - of - Oz (WoZ) appointment scheduling system (CITSEG) . We then evaluate these models using standard metrics (CITSEG) and compare our findings with the results of statistical corpus analysis .
			Although no longer competitive as end - to - end translation models , the IBM Models , as well as the hidden Markov model (HMM) of CITSEG , are still widely used for word alignment . Word alignments are used primarily for extracting minimal translation units for machine translation (MT) (e.g. , phrases [CITSEG] and rules [CITSEG]) as well as for MT system combination (CITSEG) . But their importance has grown far beyond machine translation : for instance , transferring

Figure 23: Chosen Model 2 best epoch

The figure 23 is the best epoch data from model 2, stored it in a data frame. Here we have gold function label, prediction function label and context.

After choosing the best epochs from all five model files, adequate performing to ensemble majority voting to these five model best epochs to get the final performance.

	context	gold_function_label1	pred_function_label1	pred_function_label2	pred_function_label3	pred_function_label4	pred_function_label5	majority
0	It is now possible to download a large amount of texts with automatic tools when one needs to compute , for example , a list of synonyms ; or download domain - specific monolingual texts by specifying a keyword to the search engine , and then use this text to extract domain - specific terms . It remains to be seen how we can also make use of the multilingual texts as NLP resources . In the years since the appearance of the first papers on using statistical models for bilingual lexicon compilation and machine translation (CITSEG__TARGET) , large amount of human effort and time has been invested in collecting parallel corpora of translated texts . Our goal	0	5	5	5	5	5	5.0

Figure 24: Five best epoch's majority voting result

Next, using gold function label and all five model best epoch majority results we are calculating classification performance.

	precision	recall	f1-score	support
ComorCon	0.7534	0.7692	0.7612	143
Uses	0.7739	0.7807	0.7773	114
Extends	0.6000	0.4800	0.5333	25
Motivation	0.6667	0.5909	0.6265	44
Future	0.7143	0.7692	0.7407	13
Background	0.7823	0.7984	0.7902	243
accuracy			0.7577	582
macro avg	0.7151	0.6981	0.7049	582
weighted avg	0.7555	0.7577	0.7561	582

Figure 25: One model variant best epochs classification performance

Similarly same ensemble process has been applied for all 11 model variants and then final classification performance is calculated.

Class-6 Best Epoch Performance		Ensemble 2_1_02	Ensemble 2_1_06	Ensemble 2_1_07	Ensemble 2_1_08	Ensemble 2_1_09	Ensemble 2_1_10	Ensemble 3_2_06	Ensemble 3_2_06_a	Ensemble 3_2_08	Ensemble 3_2_09	Ensemble 3_2_10
ComorCon	Precision	75.34	77.03	75.51	78.01	79.72	78.01	75.33	75.8	80.56	78.57	78.87
	Recall	76.92	79.72	77.62	76.92	79.72	76.92	79.02	83.22	81.12	76.92	78.32
	F1-Score	76.12	78.35	76.55	77.46	79.72	77.46	77.13	79.33	80.84	77.74	78.6
Uses	Precision	77.39	74.38	79.28	79.83	78.15	78.45	78.99	79.49	80.33	81.08	79.46
	Recall	78.07	78.95	77.19	83.33	81.58	79.82	82.46	81.58	85.96	78.95	78.07
	F1-Score	77.73	76.6	78.22	81.55	79.83	79.13	80.69	80.52	83.05	80	78.76
Extends	Precision	60	75	81.25	80	76.47	80	85.71	78.57	70.59	68.75	82.35
	Recall	48	48	52	48	52	48	48	44	48	44	56
	F1-Score	53.33	58.54	63.41	60	61.9	60	61.54	56.41	57.14	53.66	66.67
Motivation	Precision	66.67	75	72.97	62.79	67.5	73.68	72.22	70	74.36	71.79	72.5
	Recall	59.09	61.36	61.36	61.36	61.36	63.64	59.09	63.64	65.91	63.64	65.91
	F1-Score	62.65	67.5	66.67	62.07	64.29	68.29	65	66.67	69.88	67.47	69.05
Future	Precision	71.43	90.91	83.33	90.91	100	83.33	83.33	83.33	83.33	90.91	90.91
	Recall	76.92	76.92	76.92	76.92	69.23	76.92	76.92	76.92	76.92	76.92	76.92
	F1-Score	74.07	83.33	80	83.33	81.82	80	80	80	80	83.33	83.33
Background	Precision	78.23	78.4	77.99	78.66	78.35	78.85	78.09	80.99	80.24	77.36	78.08
	Recall	79.84	80.66	83.13	81.89	81.89	84.36	80.66	80.66	81.89	84.36	83.54
	F1-Score	79.02	79.51	80.48	80.24	80.08	81.51	79.35	80.82	81.06	80.71	80.72
Macro_Avg		**70.49**	**73.97**	**74.22**	**74.11**	**74.61**	**74.4**	**73.95**	**73.96**	**75.33**	**73.82**	**76.19**
Best		**71**	**70**	**72**	**73**	**72**	**73**	**71**	**73**	**72**	**73**	**74**

Table 13: Class-6 Best epoch ensemble performance

Class-7 Best Epoch-Performance		Ensemble 2_1_02	Ensemble 2_1_06	Ensemble 2_1_07	Ensemble 2_1_08	Ensemble 2_1_09	Ensemble 2_1_10	Ensemble 3_2_06	Ensemble 3_2_06_a	Ensemble 3_2_08	Ensemble 3_2_09	Ensemble 3_2_10
Comparison	Precision	78.12	74.32	83.33	80.6	75	70.89	73.91	72.6	81.43	72.6	81.16
	Recall	68.49	75.34	75.34	73.97	73.97	76.71	69.86	72.6	78.08	72.6	76.71
	F1-Score	72.99	74.83	79.14	77.14	74.48	73.68	71.83	72.6	79.72	72.6	78.87
Similar	Precision	61.82	67.5	73.91	69.81	74.42	70.45	68.63	68.89	75	69.57	69.39
	Recall	73.91	58.7	73.91	80.43	69.57	67.39	76.09	67.39	71.74	69.57	73.91
	F1-Score	67.33	62.79	73.91	74.75	71.91	68.89	72.16	68.13	73.33	69.57	71.58
Uses	Precision	76.47	76.19	76.03	78.99	80	81.74	72.22	78.7	76.32	79.49	80.91
	Recall	79.82	84.21	80.7	82.46	80.7	82.46	79.82	74.56	76.32	81.58	78.07
	F1-Score	78.11	80	78.3	80.69	80.35	82.1	75.83	76.58	76.32	80.52	79.46
Extends	Precision	61.54	53.33	63.16	60.87	72.22	65.22	78.57	61.54	66.67	82.35	68.18
	Recall	32	32	48	56	52	60	44	64	48	56	60
	F1-Score	42.11	40	54.55	58.33	60.47	62.5	56.41	62.75	55.81	66.67	63.83
Motivation	Precision	66.67	71.05	68.42	66.67	72.5	69.23	67.5	73.53	70.27	72.5	77.78
	Recall	50.09	61.36	59.09	59.09	65.91	61.36	61.36	56.82	50.09	65.91	63.64
	F1-Score	62.65	65.85	63.41	62.65	69.05	65.06	64.29	64.1	64.2	69.05	70
Future	Precision	81.82	90	90	76.92	90	100	83.33	90.91	76.92	90.91	90.91
	Recall	69.23	69.23	69.23	76.92	69.23	76.92	76.92	76.92	76.92	76.92	76.92
	F1-Score	75	78.26	78.26	76.92	78.26	86.96	80	83.33	76.92	83.33	83.33
Background	Precision	80.78	79.93	82.62	83.58	83.1	82.35	80.74	80.7	80.42	82.37	82.11
	Recall	85.02	83.52	87.27	83.9	88.39	83.9	81.65	86.14	86.14	85.77	87.64
	F1-Score	82.85	81.68	84.88	83.74	85.66	83.12	81.19	83.33	83.18	84.04	84.78
Macro_Avg		**68.72**	**69.06**	**73.21**	**73.46**	**74.31**	**74.61**	**71.67**	**72.98**	**72.78**	**75.11**	**75.98**
Best		**69**	**71**	**72**	**70**	**72**	**71**	**70**	**71**	**72**	**72**	**75**

Table 14: Class-7 Best epoch ensemble performance

Class-8		Ensemble 2_1_02	Ensemble 2_1_06	Ensemble 2_1_07	Ensemble 2_1_08	Ensemble 2_1_09	Ensemble 2_1_10	Ensemble 3_2_06	Ensemble 3_2_06_a	Ensemble 3_2_08	Ensemble 3_2_09	Ensemble 3_2_10
ComorCon	Precision	81.67	74.67	83.58	78.57	81.43	80	80	81.16	86.36	78.38	79.45
	Recall	77.33	74.67	74.67	73.33	76	74.67	74.67	74.67	76	77.33	77.33
	F1-Score	79.45	74.67	78.87	75.86	78.62	77.24	77.24	77.78	80.85	77.85	78.38
Similar	Precision	75.68	70	66.67	73.68	76.47	68.29	68.42	71.43	71.05	70.59	72.5
	Recall	77.78	77.78	66.67	77.78	72.22	77.78	72.22	69.44	75	66.67	80.56
	F1-Score	76.71	73.68	66.67	75.68	74.29	72.73	70.27	70.42	72.97	68.57	76.32
Uses	Precision	79.28	78.95	79.83	82.14	76.07	80.17	78.15	81.74	75.65	78.45	80.91
	Recall	77.19	78.95	83.33	80.7	78.07	81.58	81.58	82.46	76.32	79.82	78.07
	F1-Score	78.22	78.95	81.55	81.42	77.06	80.87	79.83	82.1	75.98	79.13	79.46
Basis	Precision	56.52	72.22	81.82	72.22	63.16	71.43	64.29	82.35	64.71	73.33	70
	Recall	52	52	36	52	48	60	36	56	44	44	56
	F1-Score	54.17	60.47	50	60.47	54.55	65.22	46.15	66.67	52.38	55	62.22
Motivation	Precision	57.14	72.97	73.53	70.73	63.41	73.68	61.36	69.05	70.27	70.73	82.35
	Recall	54.55	61.36	56.82	65.91	59.09	63.64	61.36	65.91	59.09	65.91	63.64
	F1-Score	55.81	66.67	64.1	68.24	61.18	68.29	61.36	67.44	64.2	68.24	71.79
Future	Precision	83.33	83.33	81.82	81.82	81.82	76.92	91.67	81.82	90	83.33	73.33
	Recall	76.92	76.92	69.23	69.23	69.23	76.92	84.62	69.23	69.23	76.92	84.62
	F1-Score	80	80	75	75	75	76.92	88	75	78.26	80	78.57
Weak	Precision	87.5	85.71	76.47	81.25	81.25	75	70.59	77.78	72.22	82.35	68.42
	Recall	58.33	50	54.17	54.17	54.17	62.5	50	58.33	54.17	58.33	54.17
	F1-Score	70	63.16	63.41	65	65	68.18	58.54	66.67	61.9	68.29	60.47
Background	Precision	79.26	78.31	75.96	79.71	77.37	82.13	79.1	78.55	77.22	79.85	80.07
	Recall	85.26	84.86	86.85	87.65	84.46	86.06	84.46	86.06	86.45	86.85	86.45
	F1-Score	82.15	81.45	81.04	83.49	80.76	84.05	81.7	82.13	81.58	83.21	83.14
Macro_Avg		**72.06**	**72.38**	**70.08**	**73.14**	**70.81**	**74.19**	**70.39**	**73.53**	**71.02**	**72.54**	**73.79**
Best		**68**	**71**	**70**	**71**	**71**	**74**	**71**	**72**	**70**	**71**	**71**

Table 15: Class-8 Best epoch ensemble performance

Table 16: Class-9 Best epoch ensemble performance

Class-9		Ensemble 2_1_02	Ensemble 2_1_06	Ensemble 2_1_07	Ensemble 2_1_08	Ensemble 2_1_09	Ensemble 2_1_10	Ensemble 3_2_06	Ensemble 3_2_06_a	Ensemble 3_2_08	Ensemble 3_2_09	Ensemble 3_2_10
Comparison	Precision	74.08	73.24	78.08	83.08	77.78	80.88	76.71	74.67	78.87	82.35	82.09
	Recall	78.08	71.23	78.08	73.97	76.71	75.34	76.71	76.71	76.71	76.71	75.34
	F1-Score	76	72.22	78.08	78.26	77.24	78.01	76.71	75.68	77.78	79.43	78.57
Similar	Precision	71.43	67.65	68.57	71.43	71.43	73.53	74.19	73.53	72.41	72.41	69.7
	Recall	80.65	74.19	77.42	80.65	80.65	80.65	74.19	80.65	67.74	67.74	74.19
	F1-Score	75.76	70.77	72.73	75.76	75.76	76.92	74.19	76.92	70	70	71.88
Usage	Precision	78.45	78.63	75	78.15	80.91	78.81	78.45	76.86	75.41	80.36	81.74
	Recall	79.82	80.7	81.58	81.58	78.07	81.58	79.82	81.58	80.7	78.95	82.46
	F1-Score	79.13	79.65	78.15	79.83	79.46	80.17	79.13	79.15	77.97	79.65	82.1
Basis	Precision	66.67	68.75	80	66.67	64.71	68.42	75	68.42	81.25	64.71	73.68
	Recall	56	44	48	48	44	52	48	52	52	44	56
	F1-Score	60.87	53.66	60	55.81	52.38	59.09	58.54	59.09	63.41	52.38	63.64
Motivation	Precision	62.22	67.44	66.67	64.44	65.12	75	57.69	65.96	65.85	69.05	79.41
	Recall	63.64	65.91	63.64	65.91	63.64	68.18	68.18	70.45	61.36	65.91	61.36
	F1-Score	62.92	66.67	65.12	65.17	64.37	71.43	62.5	68.13	63.53	67.44	69.23
Support	Precision	58.82	72.73	75	43.75	58.33	66.67	50	47.06	77.78	70	66.67
	Recall	66.67	53.33	40	46.67	46.67	66.67	53.33	53.33	46.67	46.67	53.33
	F1-Score	62.5	61.54	52.17	45.16	51.85	66.67	51.61	50	58.33	56	59.26
Weakness	Precision	70.59	70	80	78.95	77.78	66.67	70.59	70	81.25	81.25	73.68
	Recall	50	58.33	50	62.5	58.33	58.33	50	58.33	54.17	54.17	58.33
	F1-Score	58.54	63.64	61.54	69.77	66.67	62.22	58.54	63.64	65	65	65.12
Future	Precision	90.91	81.82	81.82	81.82	81.82	81.82	90.91	90	83.33	90.91	81.82
	Recall	76.92	69.23	69.23	69.23	69.23	69.23	76.92	69.23	76.92	76.92	69.23
	F1-Score	83.33	75	75	75	75	75	83.33	78.26	80	83.33	75
Natural	Precision	79.42	78.38	78.76	78.74	78.41	79.3	79.6	80.75	77.07	76.17	77.94
	Recall	79.42	83.54	83.95	82.3	85.19	83.54	81.89	79.42	84.36	86.83	87.24
	F1-Score	79.42	80.88	81.27	80.48	81.66	81.36	80.73	80.08	80.55	81.15	82.33
Macro_Avg		**70.94**	**69.34**	**69.34**	**69.47**	**69.38**	**72.32**	**69.48**	**70.11**	**70.73**	**70.49**	**71.9**
Best		**68**	**68**	**67**	**68**	**67**	**68**	**71**	**68**	**69**	**69**	**69**

Class-10		Ensemble 2_1_02	Ensemble 2_1_06	Ensemble 2_1_07	Ensemble 2_1_08	Ensemble 2_1_09	Ensemble 2_1_10	Ensemble 3_2_06	Ensemble 3_2_06_a	Ensemble 3_2_08	Ensemble 3_2_09	Ensemble 3_2_10
Basis	Precision	58.82	61.9	70.59	59.09	68.75	63.64	66.67	70.59	66.67	71.43	73.33
	Recall	40	52	48	52	44	56	56	48	56	60	44
	F1-Score	47.62	56.52	57.14	55.32	53.66	59.57	60.87	57.14	60.87	65.22	55
CoCoGM	Precision	75	77.78	80	76.6	80.43	84.09	80	69.64	79.17	78.26	68.63
	Recall	72	70	80	72	74	74	72	78	76	72	70
	F1-Score	73.47	73.68	80	74.23	77.08	78.72	75.79	73.58	77.55	75	69.31
CoCoRes	Precision	69.23	73.08	60.71	66.67	72	70.37	67.86	70.83	65.52	60.71	73.91
	Recall	72	76	68	72	72	76	76	68	76	68	68
	F1-Score	70.59	74.51	64.15	69.23	72	73.08	71.7	69.39	70.37	64.15	70.83
CoCoXY	Precision	63.16	78.57	80	68.75	78.57	73.68	61.11	75	69.23	71.43	69.23
	Recall	52.17	47.83	52.17	47.83	47.83	60.87	47.83	39.13	39.13	43.48	39.13
	F1-Score	57.14	59.46	63.16	56.41	59.46	66.67	53.66	51.43	50	54.05	50
Future	Precision	90	83.33	100	81.82	81.82	90	90	90	90.91	90	100
	Recall	69.23	76.92	76.92	69.23	69.23	69.23	69.23	69.23	76.92	69.23	69.23
	F1-Score	78.26	80	86.96	75	75	78.26	78.26	78.26	83.33	78.26	81.82
Motivation	Precision	58	63.04	64.44	56.25	70.73	65.22	70	65.91	63.64	62.22	68.18
	Recall	65.91	65.91	65.91	61.36	65.91	68.18	63.64	65.91	63.64	63.64	68.18
	F1-Score	61.7	64.44	65.17	58.7	68.24	66.67	66.67	65.91	63.64	62.92	68.18
Neutral	Precision	76.11	75.4	78.93	77.27	79.92	78.66	78.69	78.46	80.7	78.1	78.95
	Recall	82.46	83.33	83.77	82.02	83.77	82.46	84.21	84.65	80.7	82.89	85.53
	F1-Score	79.16	79.17	81.28	79.57	81.8	80.51	81.36	81.43	80.7	80.43	82.11
Similar	Precision	75	71.43	74.29	71.05	70	71.79	68.29	76.47	75	69.44	72.97
	Recall	75	69.44	72.22	75	77.78	77.78	77.78	72.22	75	69.44	75
	F1-Score	75	70.42	73.24	72.97	73.68	74.67	72.73	74.29	75	69.44	73.97
Usage	Precision	78.38	76.99	77.12	79.31	73.08	77.05	78.81	79.31	74.6	75.21	76.8
	Recall	76.32	76.32	79.82	80.7	83.33	82.46	81.58	80.7	82.46	79.82	84.21
	F1-Score	77.33	76.65	78.45	80	77.87	79.66	80.17	80	78.33	77.45	80.33
Weakness	Precision	83.33	72.22	72.73	86.67	70	92.86	82.35	56.52	61.54	73.68	83.33
	Recall	62.5	54.17	66.67	54.17	58.33	54.17	58.33	54.17	66.67	58.33	62.5
	F1-Score	71.43	61.9	69.57	66.67	63.64	68.42	68.29	55.32	64	65.12	71.43
Macro_AVG		**69.17**	**69.68**	**71.91**	**68.81**	**70.24**	**72.62**	**70.95**	**68.68**	**70.38**	**69.2**	**70.3**
Best		**68**	**66**	**72**	**68**	**69**	**69**	**68**	**68**	**69**	**69**	**69**

Table 17: Class-10 Best epoch ensemble performance

Class-11		Ensemble 2_1_02	Ensemble 2_1_06	Ensemble 2_1_07	Ensemble 2_1_08	Ensemble 2_1_09	Ensemble 2_1_10	Ensemble 3_2_06	Ensemble 3_2_06_a	Ensemble 3_2_08	Ensemble 3_2_09	Ensemble 3_2_10
Basis	Precision	61.11	57.89	60	61.11	60	64.71	75	76.47	65	61.11	68.18
	Recall	44	44	36	44	40	44	60	52	52	44	60
	F1-Score	51.16	50	45	51.16	53.33	52.38	66.67	61.9	57.78	51.16	63.83
CoCoGM	Precision	71.11	73.47	74.51	69.23	68.52	66.07	75.56	72.55	78.43	78.26	78.43
	Recall	64	72	76	72	74	74	68	74	80	72	80
	F1-Score	67.37	72.73	75.25	70.59	71.15	69.81	71.58	73.27	79.21	75	79.21
CoCoRes	Precision	66.67	73.91	64	69.23	68	64.29	68.18	65.38	69.23	62.96	72.73
	Recall	78.26	73.91	69.57	78.26	73.91	78.26	65.22	73.91	78.26	73.91	69.57
	F1-Score	72	73.91	66.67	73.47	70.83	70.59	66.67	60.39	73.47	68	71.11
CoCoXY	Precision	57.89	52.17	83.33	58.82	61.11	70.59	64.71	61.11	63.16	60	64.29
	Recall	47.83	52.17	43.48	43.48	47.83	52.17	47.83	47.83	52.17	52.17	39.13
	F1-Score	52.38	52.17	57.14	50	53.66	60	55	53.66	57.14	55.81	48.65
Future	Precision	81.82	90	81.82	90	90.91	100	81.82	100	81.82	81.82	100
	Recall	69.23	69.23	69.23	69.23	76.92	69.23	69.23	69.23	69.23	69.23	76.92
	F1-Score	75	78.26	75	78.26	83.33	81.82	75	81.82	75	75	86.96
Motivation	Precision	58	61.22	63.04	63.83	61.7	64.44	68.18	66.67	70	67.39	69.77
	Recall	65.91	68.18	65.91	68.18	65.91	65.91	68.18	63.64	63.64	70.45	68.18
	F1-Score	61.7	64.52	64.44	65.93	63.74	65.17	68.18	65.12	66.67	68.89	68.97
Neutral	Precision	75.77	75.86	76.79	76.86	78.03	79.3	77.22	75.11	77.12	78.26	78.95
	Recall	78.18	80	82.73	80	79.09	81.82	83.18	78.18	82.73	81.82	81.82
	F1-Score	76.96	77.88	79.65	78.4	78.56	80.54	80.09	76.61	79.82	80	80.36
Similar	Precision	75	75	74.19	71.43	72.41	68.57	72.73	70.97	81.25	65.62	70.97
	Recall	67.74	67.74	74.19	80.65	67.74	77.42	77.42	70.97	83.87	67.74	70.97
	F1-Score	71.19	71.19	74.19	75.76	70	72.73	75	70.97	82.54	66.67	70.97
Support	Precision	53.33	47.06	63.64	63.64	63.64	66.67	66.67	56.25	70	50	53.33
	Recall	53.33	53.33	46.67	46.67	46.67	53.33	53.33	60	46.67	46.67	53.33
	F1-Score	53.33	50	53.85	53.85	53.85	59.26	59.26	58.06	56	48.28	53.33
Usage	Precision	75.61	78.63	74.4	78.95	73.39	77.97	76.23	74.4	77.59	74.79	75.59
	Recall	81.58	80.7	81.58	78.95	79.82	80.7	81.58	81.58	78.95	78.07	84.21
	F1-Score	78.48	79.65	77.82	78.95	76.47	79.31	78.81	77.82	78.26	76.39	79.67
Weakness	Precision	68.42	80	83.33	69.57	70	77.78	73.68	72.22	71.43	78.95	84.21
	Recall	54.17	50	62.5	66.67	58.33	58.33	58.33	54.17	62.5	62.5	66.67
	F1-Score	60.47	61.54	71.43	68.09	63.64	66.67	65.12	61.9	66.67	69.77	74.42
Macro_Avg		65.46	66.53	67.31	67.68	67.14	68.93	69.22	68.23	70.23	66.82	70.68
Best		65	64	65	66	67	67	67	67	70	64	71

Table 18: Class-11 Best epoch ensemble performance

As shown in all performance results(Table 13 to 18) of class-6, 7, 8, 9, 10, 11 it is evident that higher ensemble performance 76.19 on class 6 in variant 3_2_10 has been achieved.

Variant	3_2_10	3_2_10	2_1_10	2_1_10	2_1_10	3_1_10
Class	Class-6	Class-7	Class-8	Class-9	Class-10	Class-11
Percentage	**76.19**	**75.98**	**74.19**	**72.32**	**72.62**	**70.68**

Table 19: Higher Best epoch performance of each class

Confusion matrix of best performance classes

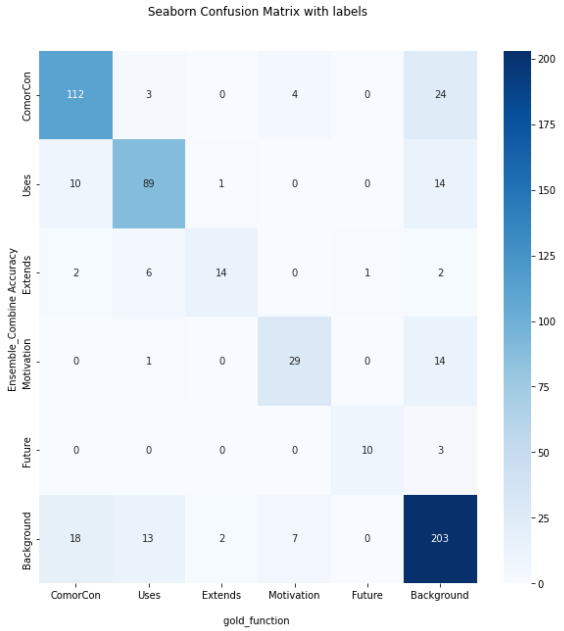


Figure 26: Class-6 best epoch confusion performance

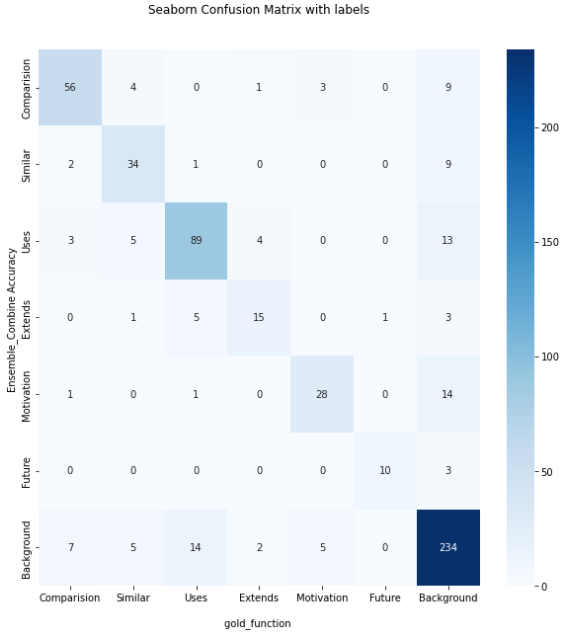


Figure 27: Class-7 best epoch confusion performance

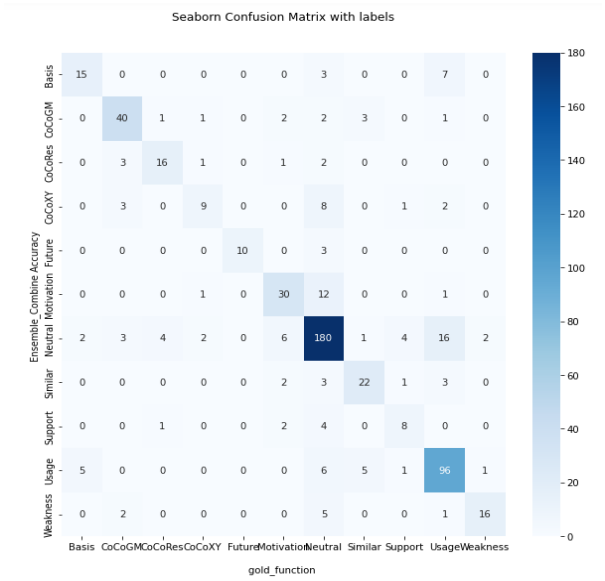


Figure 28: Class-11 best epoch confusion performance

After performing the ensemble we need to analyze the similarities and differences between the performances of the variant models. The base classifiers in an ensemble must be distinct in order to enhance performance over a single model. So we need to analyze the classifier diversity. We use some measures like pair wise and non-pair-wise diversities to analyze the diversity.

6.4 Diversity Measures:

A fundamental challenge in classifier aggregation is diversity among the members of the group of classifiers. Furthermore, assessing diversity is difficult due to the lack of a widely accepted mathematical formalism. Many strategies have been developed to build a decent classifier ensemble by enhancing both the efficiency and diversity of the basis classifiers. Moreover, there is no single universally accepted definition of diversity, and evaluating diversity is extremely challenging. Despite the fact that scientists organised multiple experimental trials to examine different diversity metrics, the outcomes were often contradictory. Investigation has been done the Q statistic, correlation, disagreement, and double fault are four aggregate pair wise measurements and six non-pair-wise metrics are voting entropy, coincident failure, difficulty index, Kohavi-Wolpert variance, interpreter agreement and generalised diversity.

In this research pair-wise 'Disagreement measure' has been used to calculate the similarities and dissimilarities between the two learners.

6.4.1 Pair-wise Measures:

A traditional technique to measuring ensemble diversity is to evaluate paired similarity/dissimilarity among two learners, and then averaged all of the paired measures to get the total diversity. As a result, when the dataset contains three or more base classifiers, the overall diversity is determined by the mean of all classifier pair measurements.

The paired Disagreement measure has employed in this research are described below.

6.4.2 Disagreement Measure

The disagreement measure was utilized by Skalak (Skalak, 1996, Ho, 1998) to evaluate the variation of two classifiers. It's expressed as a measure of the number of measurements divided by the total number of samples for which the classifiers differed. Means the measurement of the sum of instances on which one classifier is accurate and the other is incorrect to the total number of occurrences.

$$\mathrm{D2}_{ij} = \frac{N^{01} + N^{10}}{N},$$
$$\overline{\mathrm{D2}} = \frac{2}{M(M-1)} \sum_{\substack{i,j=1,\ldots,M \\ i \neq j}} \mathrm{D2}_{ij}.$$

Here D2ij state the number of cases or instances, where both learners i and j are accurate. D2 represents the number of cases in which both classifiers are in accurate.

The disagreement measure between each pair of classifiers has been calculated to combine the all classifiers. And based on this we can find a classifier which is most different from others. Finally majority voting classifier is used to combine the several base models' predictions in order to improve the performance.

The disagreement measure is depicted in detail shown in the Figure 29.

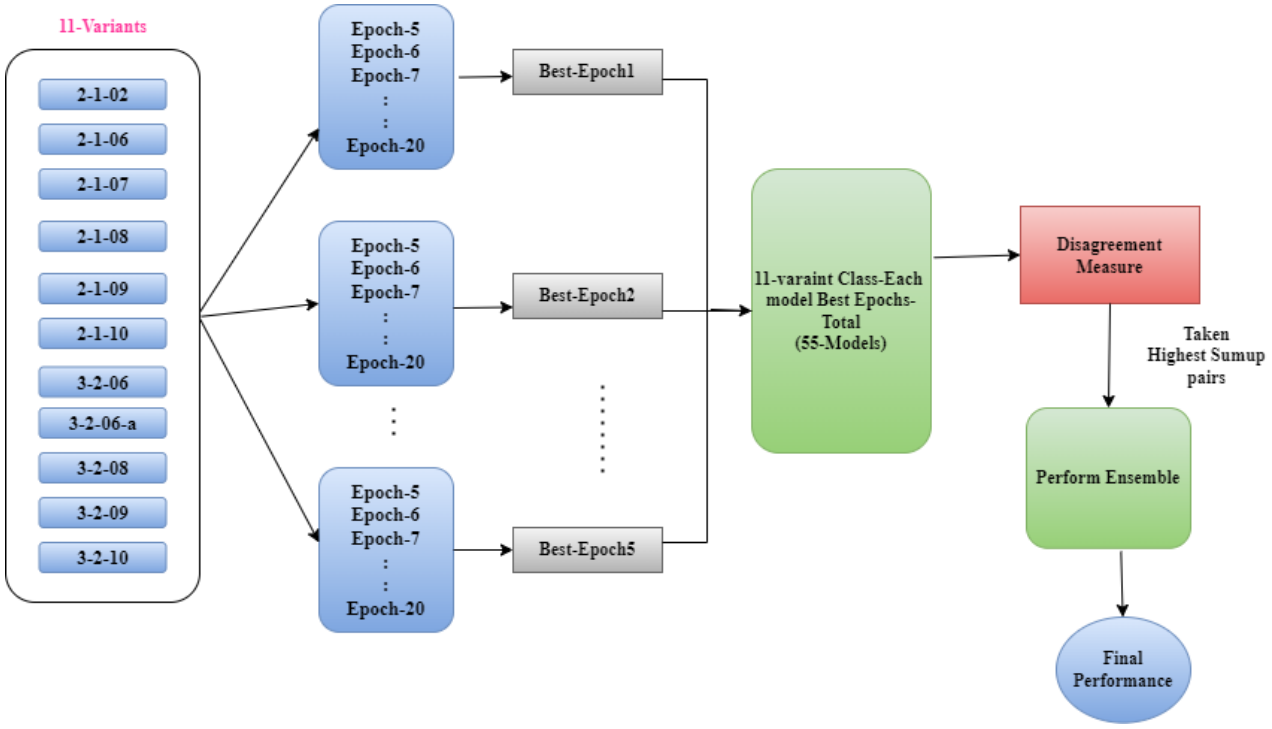


Figure 29: Complete Overview of Disagreement measure

The figure 29 clearly describes the overview of the disagreement process in this research. Initially, all model best epochs are selected to find the disagreement measure. Using these results sum of disagreement has been calculated. Finally highest sum up pairs has been taken form disagreement sumup results to evaluate the final performance.

The disagreement measure can be evaluated in different ways:

- We can choose the all epochs for each model variant to build an ensemble for one model variant plus one seed.
- We can use the all epochs in five model seed to build an ensemble for each model variant
- For best epochs we can take 11 best epochs for 11 model variants * 5 seeds per model variant to build an ensemble for CFC
- We can use 11 best models for the 11 model variants (the best epoch using the best seed) to build an ensemble

Here, chosen all best epochs of 5 seeds per model in 11 model variants to calculate the disagreement measure (total 55 models.)

First, the disagreement measure between each pair of classifiers has been calculated to combine the all classifiers.

```
i pred_function_label1
j pred_function_label1
Disagreement1 0.0
0 0
i pred_function_label1
j pred_function_label2
Disagreement1 0.14948453608247422
0 1
i pred_function_label1
j pred_function_label3
Disagreement1 0.12886597938144329
0 2
i pred_function_label1
j pred_function_label4
Disagreement1 0.14948453608247422
0 3
i pred_function_label1
j pred_function_label5
Disagreement1 0.14948453608247422
```

Figure 30: Disagreement measure of each pair of classifier

Figure 30, shows the disagreement measure results of each pair of classifiers. Using these results we performed disagreement matrix to display the pair of classifier results.

```
array2=numpy.copy(array)
array2.shape
```

```
(55, 55)
```

```
array2
```

```
array([[0.  , 0.15, 0.13, 0.15, 0.15, 0.63, 0.62, 0.63, 0.62, 0.61, 0.63,
        0.63, 0.62, 0.62, 0.62, 0.14, 0.11, 0.15, 0.14, 0.15, 0.62, 0.63,
        0.62, 0.62, 0.61, 0.58, 0.57, 0.58, 0.59, 0.57, 0.57, 0.56, 0.56,
        0.57, 0.57, 0.57, 0.55, 0.56, 0.57, 0.56, 0.13, 0.15, 0.15, 0.13,
        0.14, 0.14, 0.15, 0.16, 0.15, 0.16, 0.17, 0.14, 0.13, 0.15, 0.14],
       [0.15, 0.  , 0.15, 0.17, 0.17, 0.64, 0.63, 0.63, 0.63, 0.62, 0.63,
        0.64, 0.63, 0.63, 0.62, 0.16, 0.15, 0.15, 0.15, 0.14, 0.63, 0.63,
        0.64, 0.62, 0.62, 0.59, 0.58, 0.59, 0.6 , 0.57, 0.58, 0.56, 0.57,
        0.57, 0.58, 0.57, 0.55, 0.57, 0.58, 0.58, 0.17, 0.17, 0.15, 0.18,
        0.16, 0.17, 0.18, 0.17, 0.16, 0.16, 0.16, 0.14, 0.16, 0.16, 0.14],
       [0.13, 0.15, 0.  , 0.16, 0.15, 0.63, 0.62, 0.62, 0.61, 0.61, 0.62,
        0.62, 0.62, 0.61, 0.62, 0.14, 0.11, 0.15, 0.12, 0.17, 0.62, 0.63,
        0.62, 0.61, 0.61, 0.59, 0.58, 0.58, 0.6 , 0.58, 0.58, 0.56, 0.57,
        0.57, 0.57, 0.57, 0.55, 0.57, 0.57, 0.56, 0.12, 0.16, 0.14, 0.15,
        0.13, 0.16, 0.15, 0.13, 0.16, 0.15, 0.14, 0.14, 0.12, 0.14, 0.14],
       [0.15, 0.17, 0.16, 0.  , 0.16, 0.64, 0.64, 0.64, 0.63, 0.64, 0.64,
        0.65, 0.64, 0.64, 0.63, 0.16, 0.14, 0.17, 0.16, 0.16, 0.63, 0.65,
        0.64, 0.62, 0.63, 0.61, 0.59, 0.61, 0.61, 0.58, 0.59, 0.6 , 0.58,
        0.58, 0.59, 0.59, 0.57, 0.58, 0.59, 0.59, 0.14, 0.17, 0.16, 0.15,
```

Figure 31: Disagreement matrix

Figure 31 demonstrates the all disagreement measure results of each pair of classifiers. Using these results the sum of disagreement curve has been plotted. Same process applied to all classes.

Sum of disagreement curve: The sum of rows in Disagreement matrix of 55 models has been calculated. We obtained a list of pairs' results after computing the sum.

Using these results we plotted a curve of sum of disagreement measure of all classifiers.

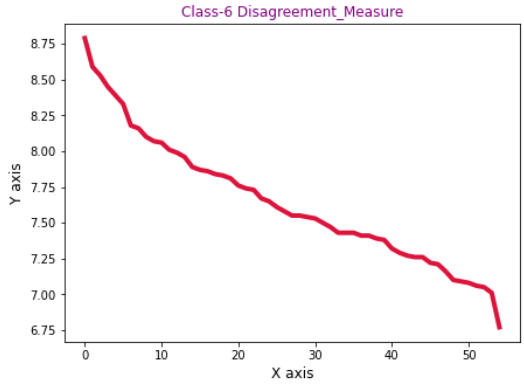


Figure 32: Class-6 disagreement curve

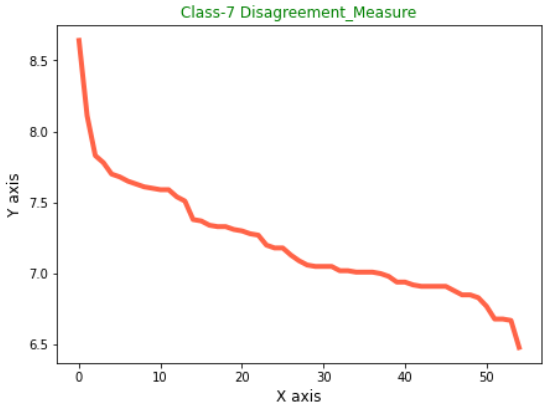


Figure 33: Class-7 disagreement curve

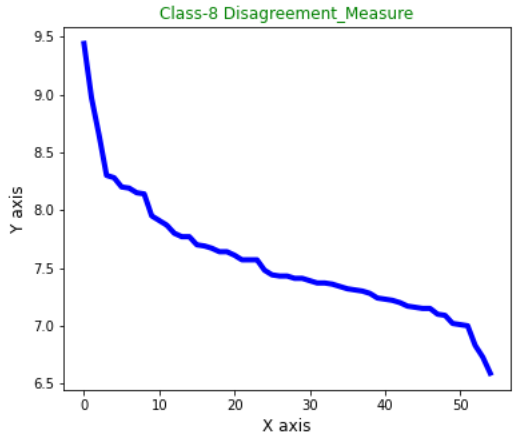


Figure 34: Class-8 disagreement curve

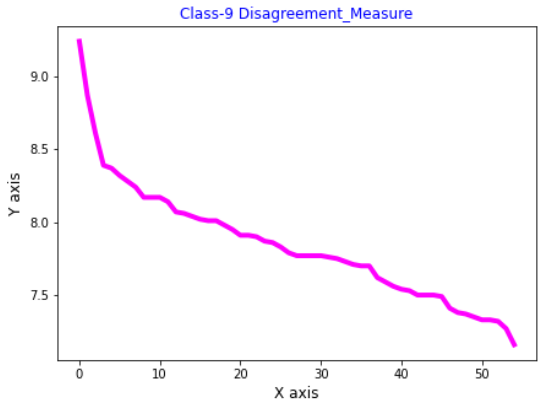


Figure 35: Class-9 disagreement curve

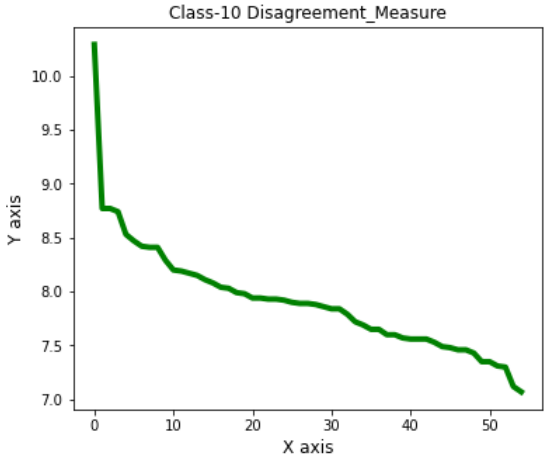


Figure 36: Class-10 disagreement curve

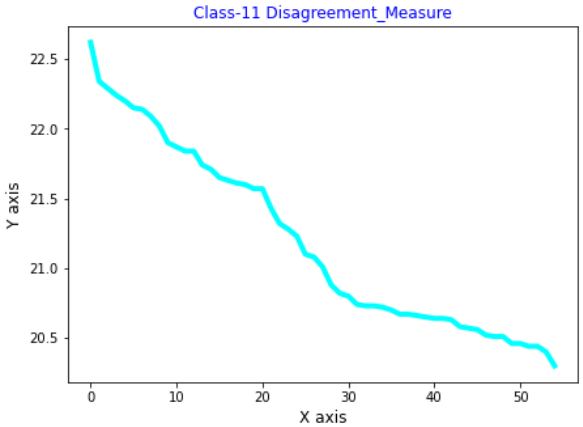


Figure 37: Class-11 disagreement curve

After sum up the pairs we have chosen some odd threshold range 'K' to select the subset of them. We have taken highest sum up pairs and set the threshold range 1, 3,5,7,9,11,13,15, 17 up to 19. Then the final performance was calculated using the ensemble majority voting procedure. We have applied same process to all classes' best epochs and results are shown in below tables. Table 20 to 25 demonstrates the disagreement results of all classes.

Class-6	Threshold	Thr-1	Thr-3	Thr-5	Thr-7	Thr-9	Thr-11	Thr-13	Thr-15	Thr-17	Thr-19
ComorCon	Precision	66.11	69.89	71.01	75.64	76.77	77.7	77.85	77.55	78.91	81.38
	Recall	83.22	86.01	83.92	82.52	83.22	80.42	81.12	79.72	81.12	82.52
	F1-Score	73.68	77.12	76.92	78.93	79.87	79.04	79.45	78.62	80	81.94
Uses	Precision	72.66	76.03	78.76	79.63	78.38	79.09	77.78	77.97	79.49	81.03
	Recall	81.58	80.7	78.07	75.44	76.32	76.32	79.82	80.7	81.58	82.36
	F1-Score	76.86	78.3	78.41	77.48	77.33	77.68	78.79	79.31	80.52	81.74
Extends	Precision	72.22	60.87	60.87	63.64	66.67	63.64	68.42	87.5	87.5	88.89
	Recall	52	56	56	56	56	56	52	56	56	64
	F1-Score	60.47	58.33	58.33	59.57	60.87	59.57	59.09	68.29	68.29	74.42
Motivation	Precision	70	70	72.97	78.79	75.68	71.79	73.68	77.78	75.68	75.68
	Recall	63.64	63.64	61.36	59.09	63.64	63.64	63.64	63.64	63.64	63.64
	F1-Score	66.67	66.67	66.67	67.53	69.14	67.47	68.29	70	69.41	69.14
Future	Precision	78.57	76.92	83.33	83.33	83.33	90	90.91	90.91	90.91	100
	Recall	84.62	76.92	76.92	76.92	76.92	69.23	76.92	76.92	76.92	76.92
	F1-Score	81.48	76.92	80	80	80	78.26	83.33	83.33	83.33	86.96
Background	Precision	85.15	84.69	82.89	80.08	80.08	78.26	79.44	79.13	79.53	80.86
	Recall	70.78	72.84	77.78	82.72	81.07	81.48	81.07	82.72	83.13	85.19
	F1-Score	77.3	78.32	80.25	81.38	80.57	79.84	80.24	80.89	81.29	82.97
Macro_Avg		**72.74**	**72.61**	**73.43**	**74.15**	**74.63**	**73.64**	**74.87**	**76.74**	**77.09**	**79.53**

Table 20: Class-6 Disagreement measure performance

Class-7	Thr-Range	Thr-3	Thr-5	Thr-7	Thr-9	Thr-11	Thr-13	Thr-15	Thr-17	Thr-19
Comparison	Precision	76.12	81.54	79.71	80.88	81.16	79.17	80.56	79.17	78.57
	Recall	69.86	72.6	75.34	75.34	76.71	78.08	79.45	78.08	75.34
	F1-Score	72.86	76.81	77.46	78.01	78.87	78.62	80	78.62	76.92
Similar	Precision	63.64	66.04	66.67	72	76.6	74.47	77.78	72.92	71.43
	Recall	76.09	76.09	78.26	78.26	78.26	76.09	76.09	76.09	76.09
	F1-Score	69.31	70.71	72	75	77.42	75.27	76.92	74.47	73.68
Uses	Precision	73.23	73.02	75.81	79.19	78.51	79.34	79.34	80	80.51
	Recall	81.58	80.7	82.46	83.33	83.33	84.21	84.21	84.21	83.33
	F1-Score	77.18	76.67	78.99	81.2	80.85	81.7	81.7	82.05	81.9
Extends	Precision	64.29	58.33	61.11	77.78	68.42	77.78	82.35	87.5	86.67
	Recall	36	28	44	56	52	56	56	56	52
	F1-Score	46.15	37.84	51.16	65.12	69.09	65.12	66.67	68.29	65
Motivation	Precision	65.22	66.67	70.73	78.38	77.78	77.78	77.14	77.14	75
	Recall	68.18	63.64	65.91	65.91	63.64	63.64	61.36	61.36	61.36
	F1-Score	66.67	65.12	68.24	71.6	70	70	68.35	68.35	67.5
Future	Precision	90	90.91	90.91	90.91	90.91	90.91	90.91	90.91	90.91
	Recall	69.23	76.92	76.92	76.92	76.92	76.92	76.92	76.92	76.92
	F1-Score	78.26	83.33	83.33	83.33	83.33	83.33	83.33	83.33	83.33
Background	Precision	82.51	81.32	82.64	83.45	82.8	83.39	83.63	82.5	82.69
	Recall	81.27	83.15	82.02	86.89	86.52	86.52	88.01	86.52	87.64
	F1-Score	81.89	82.22	82.33	85.14	84.62	84.93	85.77	84.46	85.09
Macro_Avg		**70.33**	**70.39**	**73.36**	**77.06**	**76.31**	**77**	**77.54**	**77.08**	**76.2**

Table 21: Class-7 Disagreement measure performance

Class-8	Thr-Range	Thrsh-1	Thrsh-3	Thrsh-5	Thrsh-7	Thrsh-9	Thrsh-11	Thrsh-13	Thrsh-15	Thrsh-17	Thrsh-19
ComorCon	Precision	**72.86**	**77.78**	**78.87**	**78.87**	**79.17**	**81.69**	**80.56**	**82.86**	**80.56**	**80.56**
	Recall	**68**	**74.67**	**74.67**	**74.67**	**76**	**77.33**	**77.33**	**77.33**	**77.33**	**77.33**
	F1-Score	**70.34**	**76.19**	**76.71**	**76.71**	**77.55**	**79.45**	**78.91**	**80**	**78.91**	**78.91**
Similar	Precision	**50.85**	**58**	**60.42**	**70**	**71.79**	**71.79**	**74.36**	**73.68**	**72.97**	**72.97**
	Recall	**83.33**	**80.56**	**80.56**	**77.78**	**77.78**	**77.78**	**80.56**	**77.78**	**75**	**75**
	F1-Score	**63.16**	**67.44**	**69.05**	**73.68**	**74.67**	**74.67**	**77.33**	**75.68**	**73.97**	**73.97**
Uses	Precision	**75.96**	**76.99**	**76.99**	**79.28**	**78.63**	**77.31**	**77.31**	**78.81**	**78.45**	**77.97**
	Recall	**69.3**	**76.32**	**76.32**	**77.19**	**80.7**	**80.7**	**80.7**	**81.58**	**79.82**	**80.7**
	F1-Score	**72.48**	**76.65**	**76.65**	**78.22**	**79.45**	**78.97**	**78.97**	**80.17**	**79.13**	**79.31**
Basis	Precision	**48**	**64**	**66.67**	**56**	**63.16**	**66.67**	**73.68**	**63.64**	**61.9**	**68.42**
	Recall	**48**	**64**	**64**	**56**	**48**	**56**	**56**	**56**	**52**	**52**
	F1-Score	**48**	**64**	**45.31**	**56**	**54.55**	**60.87**	**63.64**	**59.57**	**56.52**	**59.09**
Motivation	Precision	**60.87**	**65.22**	**56.36**	**60**	**66.67**	**68.29**	**69.05**	**68.29**	**69.05**	**66.67**
	Recall	**63.64**	**68.18**	**70.45**	**68.18**	**63.64**	**63.64**	**65.91**	**63.64**	**65.91**	**63.64**
	F1-Score	**62.22**	**66.67**	**62.63**	**63.83**	**65.12**	**65.88**	**67.44**	**65.88**	**67.44**	**65.12**
Future	Precision	**78.57**	**85.71**	**80**	**78.57**	**76.92**	**75**	**81.82**	**81.82**	**81.82**	**81.82**
	Recall	**84.62**	**92.31**	**92.31**	**84.62**	**76.92**	**69.23**	**69.23**	**69.23**	**69.23**	**69.23**
	F1-Score	**81.48**	**88.89**	**85.71**	**81.48**	**76.92**	**72**	**75**	**75**	**75**	**75**
Weak	Precision	**72.22**	**60.87**	**66.67**	**70**	**82.35**	**82.35**	**82.35**	**82.35**	**87.5**	**87.5**
	Recall	**54.17**	**58.33**	**58.33**	**58.33**	**58.33**	**58.33**	**58.33**	**58.33**	**58.33**	**58.33**
	F1-Score	**61.9**	**59.57**	**62.22**	**63.64**	**68.29**	**68.29**	**68.29**	**68.29**	**70**	**70**
Background	Precision	**77.24**	**82.01**	**82.13**	**81.67**	**79.85**	**80.53**	**80.23**	**80.38**	**79.78**	**79.78**
	Recall	**75.7**	**78.09**	**76.89**	**81.67**	**83.67**	**84.06**	**84.06**	**84.86**	**84.86**	**84.86**
	F1-Score	**76.46**	**80**	**79.42**	**81.67**	**81.71**	**82.26**	**82.1**	**82.56**	**82.24**	**82.24**
Macro_Avg		**67.01**	**72.43**	**72.21**	**71.9**	**72.31**	**72.8**	**73.96**	**73.39**	**72.9**	**72.96**

Table 22: Class-8 Disagreement measure performance

Class-9	Thr-Range	Thrsh-1	Thrsh-3	Thrsh-5	Thrsh-7	Thrsh-9	Thrsh-11	Thrsh-13	Thrsh-15	Thrsh-17	Thrsh-19
Comparison	Precision	52.63	63.22	68.67	70.89	72.5	79.17	80.28	79.45	79.17	81.43
	Recall	82.19	75.34	78.08	76.71	79.45	78.08	78.08	79.45	78.08	78.08
	F1-Score	64.17	68.75	73.08	73.68	75.82	78.62	79.17	79.45	78.62	79.72
Similar	Precision	67.74	68.57	72.73	74.19	74.19	73.53	72.73	71.88	68.75	68.75
	Recall	67.74	77.42	77.42	74.19	74.19	80.65	77.42	74.19	70.97	70.97
	F1-Score	67.74	72.73	75	74.19	74.19	76.92	75	73.02	69.84	69.84
Usage	Precision	74.56	77.68	78.95	78.81	78.15	78.15	77.97	78.15	77.87	78.33
	Recall	74.56	76.32	78.95	81.58	81.58	81.58	80.7	81.58	83.33	82.46
	F1-Score	74.56	76.99	78.95	80.17	79.83	79.83	79.31	79.83	80.51	80.34
Basis	Precision	58.33	56.52	60.87	72.22	72.22	65	72.22	75	86.67	86.67
	Recall	56	52	56	52	52	52	52	48	52	52
	F1-Score	57.14	54.17	58.33	60.47	60.47	57.78	60.47	58.54	65	65
Motivation	Precision	56	64.44	63.64	66.67	72.5	70.73	74.36	70.73	67.44	69.05
	Recall	63.64	65.91	63.64	63.64	65.91	65.91	65.91	65.91	65.91	65.91
	F1-Score	59.57	65.17	63.64	65.12	69.05	68.24	69.88	68.24	66.67	67.44
Support	Precision	50	58.33	61.54	63.64	61.54	72.73	66.67	61.54	66.67	72.73
	Recall	53.33	46.67	53.33	46.67	53.33	53.33	53.33	53.33	53.33	53.33
	F1-Score	51.61	52.85	57.14	53.85	57.14	61.54	59.26	57.14	59.26	61.54
Weakness	Precision	65.22	73.68	65	81.25	76.47	75	75	75	75	70.59
	Recall	62.5	58.33	54.17	54.17	54.17	50	50	50	50	50
	F1-Score	63.83	65.12	59.09	65	63.41	60	60	60	60	58.54
Future	Precision	83.33	83.33	83.33	83.33	83.33	83.33	83.33	81.82	81.82	83.33
	Recall	76.92	76.92	76.92	76.92	76.42	76.92	76.92	69.23	69.23	76.92
	F1-Score	80	80	80	80	80	80	80	75	75	80
Natural	Precision	85.86	80.59	80.42	78.82	80.56	79.38	79.09	79.31	79.15	78.33
	Recall	69.96	78.6	79.42	82.72	83.54	83.95	85.6	85.19	84.36	84.77
	F1-Score	77.1	79.58	79.92	80.72	82.02	81.6	82.21	82.14	81.67	81.42
Macro_Avg		**66.19**	**68.26**	**69.46**	**70.36**	**71.33**	**71.61**	**71.7**	**70.37**	**70.73**	**71.54**

Table 23: Class-9 Disagreement measure performance

Class-10	Thr-range	Thrsh-1	Thrsh-3	Thrsh-5	Thrsh-7	Thrsh-9	Thrsh-11	Thrsh-13	Thrsh-15	Thrsh-17	Thrsh-19
Basis	Precision	46.43	54.55	74	68.18	70	70	75	75	70	73.68
	Recall	52	48	56	60	56	56	60	60	56	56
	F1-Score	49.06	51.06	62.22	63.83	62.22	62.22	66.67	66.67	62.22	63.64
CoCoGM	Precision	57.63	69.81	80.43	77.78	78.72	79.17	80.85	79.59	81.25	81.25
	Recall	68	74	74	70	74	76	76	78	78	78
	F1-Score	62.39	71.84	77.08	73.68	76.29	77.55	78.35	78.79	79.59	79.59
CoCoRes	Precision	58.62	64.29	63.33	63.33	67.86	63.33	65.52	65.52	65.52	65.52
	Recall	68	72	76	76	76	76	76	76	76	76
	F1-Score	62.96	67.92	69.09	69.09	71.7	69.09	70.37	70.37	70.37	70.37
CoCoXY	Precision	50	50	60	68.42	73.33	75	76.47	68.42	72.22	72.22
	Recall	60.87	56.52	52.17	56.52	47.83	52.17	56.52	56.52	56.52	56.52
	F1-Score	54.9	53.06	55.81	61.9	57.89	61.54	65	61.9	63.41	63.41
Future	Precision	83.33	91.67	90.91	81.82	83.33	81.82	90	90	81.82	81.82
	Recall	76.92	84.62	76.92	69.23	76.92	69.23	69.23	69.23	69.23	69.23
	F1-Score	80	88	83.33	75	80	75	78.26	78.26	75	75
Motivation	Precision	58.49	66.67	68.18	66.67	66.67	63.64	64.44	67.44	69.05	69.05
	Recall	70.45	68.18	68.18	63.64	63.64	63.64	65.91	65.91	65.91	65.91
	F1-Score	63.92	67.42	18.18	65.12	65.12	63.64	65.17	66.67	67.44	67.44
Neutral	Precision	81.28	83.17	79.74	78.24	79.17	79.15	79.24	79.17	78.93	79.67
	Recall	66.67	75.88	81.14	82.02	83.33	81.58	82.02	83.33	83.77	84.21
	F1-Score	73.25	79.36	80.43	80.09	81.2	80.35	80.6	81.2	81.28	81.88
Similar	Precision	48.15	61.36	69.05	70.73	70.73	71.79	70	73.68	74.36	74.36
	Recall	72.22	75	80.56	80.56	80.56	77.78	77.78	77.78	80.56	80.56
	F1-Score	57.78	67.5	74.36	75.32	75.32	74.67	73.68	75.68	77.33	77.33
Usage	Precision	74.77	74.19	75.63	77.39	76.47	76.67	76.47	76.07	77.39	77.78
	Recall	72.81	80.7	78.95	78.07	79.82	80.7	79.82	78.07	78.07	79.82
	F1-Score	73.78	77.31	77.25	77.73	78.11	78.63	78.11	77.06	77.73	78.79
Weakness	Precision	76.19	85	83.33	83.33	83.33	78.95	73.68	82.35	77.78	83.33
	Recall	66.67	70.83	62.5	62.5	62.5	62.5	58.33	58.33	58.33	62.5
	F1-Score	71.11	77.27	71.43	71.43	71.43	69.77	65.12	68.29	66.67	71.43
Macro_AVG		**64.91**	**70.08**	**71.92**	**71.32**	**71.93**	**71.25**	**72.13**	**72.49**	**72.1**	**72.89**

Table 24: Class-10 Disagreement measure performance

Class-11	Thr-range	Thr-1	Thr-3	Thr-5	Thr-7	Thr-9	Thr-11	Thr-13	Thr-15	Thr-17	Thr-19
Basis	Precision	61.9	65.22	78.95	76.19	68.42	68.18	66.67	70	72.22	70
	Recall	52	60	60	64	52	60	56	56	52	56
	F1-Score	56.52	62.5	68.18	69.57	59.09	63.83	60.87	62.22	60.47	62.22
CoCoGM	Precision	76.6	74.51	78	78.43	79.59	79.59	76.47	76.47	79.59	79.59
	Recall	72	76	78	80	78	78	78	78	78	78
	F1-Score	74.23	75.25	78	79.21	78.79	78.79	77.23	77.23	78.79	78.79
CoCoRes	Precision	77.27	75	75	69.23	69.23	69.23	69.23	72	69.23	69.23
	Recall	73.91	78.26	78.26	78.26	78.26	78.26	78.26	78.26	78.26	78.26
	F1-Score	75.56	76.6	76.6	73.47	73.47	73.47	73.47	75	73.47	73.47
CoCoXY	Precision	66.67	64.71	73.33	71.43	62.5	64.71	70.59	72.22	70.59	75
	Recall	60.87	47.83	47.83	43.48	43.48	47.83	52.17	56.52	52.17	52.17
	F1-Score	63.64	55	57.89	54.05	51.28	55	60	63.41	60	61.54
Future	Precision	80	100	100	100	90	90	90	90	90	90
	Recall	61.54	76.92	76.92	76.92	69.23	69.23	69.23	69.23	69.23	69.23
	F1-Score	69.57	86.96	86.96	86.96	78.26	78.26	78.26	78.26	78.26	78.26
Motivation	Precision	57.69	68.89	64.58	73.81	73.81	75	73.17	69.77	68.18	68.18
	Recall	68.18	70.45	70.45	70.45	70.45	68.18	68.18	68.18	68.18	68.18
	F1-Score	62.5	69.66	67.39	72.09	72.09	71.43	70.59	68.97	68.18	68.18
Neutral	Precision	77.88	82.19	80.36	78.11	78.88	79.4	79.4	79.65	78.72	78.81
	Recall	80	81.82	81.82	82.73	83.18	84.09	84.09	83.64	84.09	84.55
	F1-Score	78.92	82	81.08	80.35	80.97	81.68	81.68	81.6	81.32	81.58
Similar	Precision	74.29	72.73	71.88	71.88	73.53	73.53	73.53	71.88	71.88	71.88
	Recall	83.87	77.42	74.19	74.19	80.65	80.65	80.65	74.19	74.19	74.19
	F1-Score	78.79	75	73.02	73.02	76.92	76.92	76.92	73.02	73.02	73.02
Support	Precision	53.85	53.33	66.67	61.54	57.14	66.67	66.67	66.67	61.54	61.54
	Recall	46.67	53.33	53.33	53.33	53.33	53.33	53.33	53.33	53.33	53.33
	F1-Score	50	53.33	59.26	57.14	55.17	59.26	59.26	59.26	57.14	57.14
Usage	Precision	77.78	74.6	75.19	76.42	76.03	77.97	77.5	77.5	77.5	78.15
	Recall	79.82	82.36	85.09	82.46	80.7	80.7	81.58	81.58	81.58	81.58
	F1-Score	78.79	78.33	79.84	79.32	78.3	79.31	79.49	79.49	79.49	79.83
Weakness	Precision	88.89	84.21	78.95	82.35	78.95	76.19	82.35	75	83.33	88.24
	Recall	66.67	66.67	62.5	58.33	62.5	66.67	58.33	62.5	62.5	62.5
	F1-Score	76.19	74.42	69.77	68.29	69.77	71.11	68.29	68.18	71.43	73.17
Macro_Avg		**69.52**	**71.73**	**72.54**	**72.13**	**70.37**	**71.73**	**71.46**	**71.51**	**71.05**	**71.56**

Table 25: Class-11 Disagreement measure performance

All classes Disagreement Macro Avg Performance		Thrsh-1	Thrsh-3	Thrsh-5	Thrsh-7	Thrsh-9	Thrsh-11	Thrsh-13	Thrsh-15	Thrsh-17	Thrsh-19
Class-6	Macro_Avg	72.74	72.61	73.43	74.15	74.63	73.64	74.87	76.74	77.09	79.53
Class-7	Macro_Avg	64.65	70.33	70.39	73.36	77.06	76.31	77	77.54	77.08	76.2
Class-8	Macro_Avg	67.01	72.43	72.21	71.9	72.31	72.8	73.96	73.39	72.9	72.96
Class-9	Macro_Avg	66.19	68.26	69.46	70.36	71.33	71.61	71.7	70.37	70.73	71.54
Class-10	Macro_Avg	64.91	70.08	71.92	71.32	71.93	71.25	72.13	72.49	72.1	72.89
Class-11	Macro_Avg	69.52	71.73	72.54	72.13	70.37	71.73	71.46	71.51	71.05	71.56

Table 26: All classes' final disagreement performance

Threshold	Thr-19	Thr-15	Thr-13	Thr-13	Thr-19	Thr-5
Class	Class-6	Class-7	Class-8	Class-9	Class-10	Class-11
Percentage	79.53	77.54	73.96	71.70	72.89	72.54

Table 27: Highest disagreement performance of all classes

Classification performance of highest disagreement measures

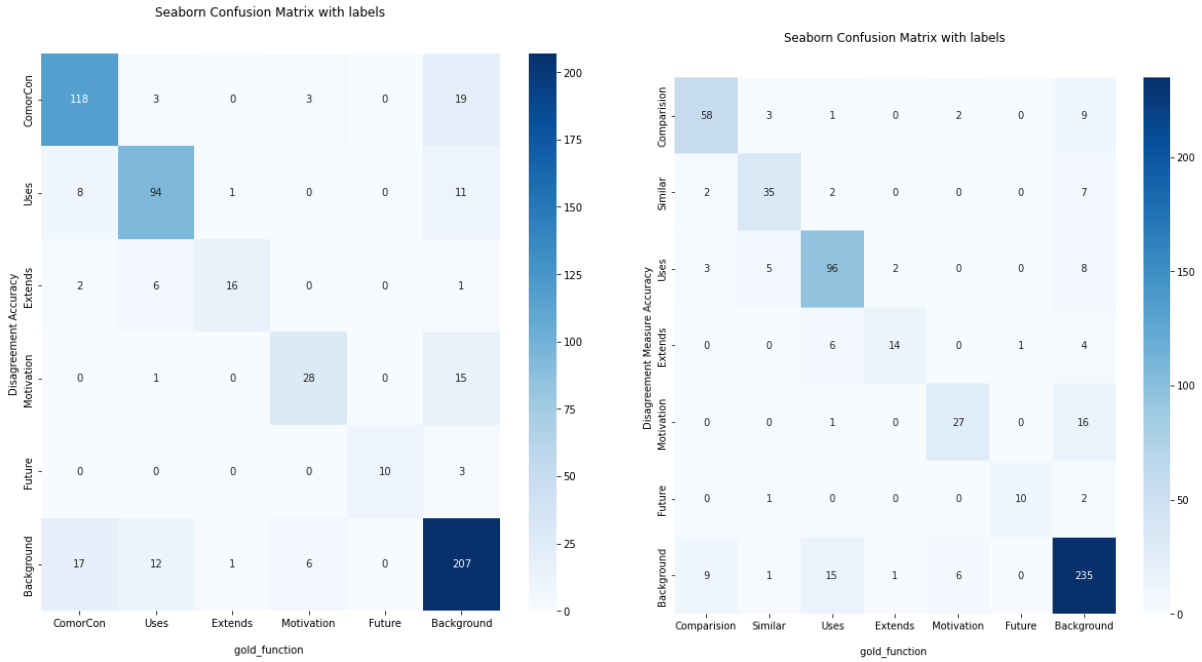


Figure 38: Class-6 (Thr-19) and class -7 (Thr-15) disagreement performances

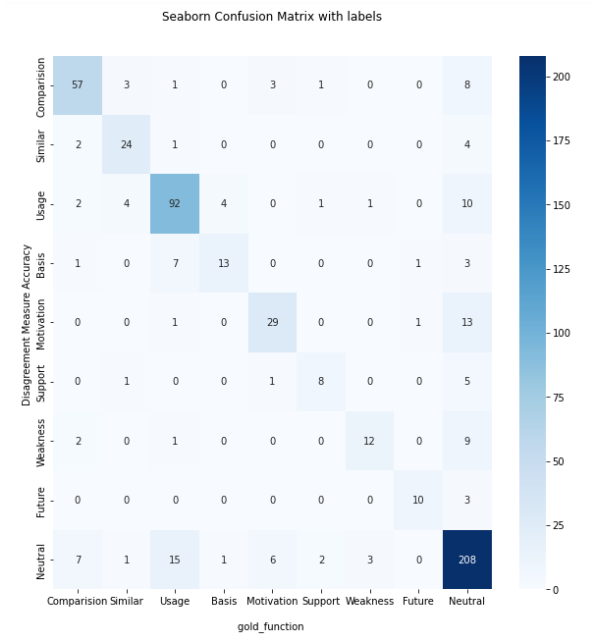


Figure 39: Class-9 Threshold 13 disagreement performance

7. Model's Performance and it results:

In this section models performance has been discussed. All required sklearn packages has imported to create the confusion matrix and compute Precision, Recall, and F1 score in order to evaluate model performance

Classification Report:

A classifications report is a performance review indicator. After training the model classification and confusion matrix can be used to test the performance. It can be used to demonstrate our performing test model's accuracy, recall, F1 Score, and support.

Precision:
The percentage of accurately predicted positive readings to total positive significant observations is known as precision. This will helpful to find out which class label has high precision. The low false percentage is related to high precision.

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

After performing the ensemble process to best epochs we see higher precision 100% (Future) in 7, 10 and 11 class scheme. And 91.67 % in 8 class scheme. Lowest precision 47.06% is observed at 9 and 10 class scheme.

Recall:
As a consequence, Recall calculates how many True Positives our model catches by rating it as Good (True Positive). With this logic, we see that when False Negative has a high cost, Recall will be the model measure we use to select the best model.

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

The highest best epochs recall performance 88.39 % observed at 7 class scheme.

F1-Score:
It combines precision and recall into a single measure by finding mean between the two. When it's time to strike a balance between precision and recall, when there is an uneven class distribution, the F1 Score is a better measure to use.

$$\text{F1} = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

The highest F1 score 83.33% with 76.19 % macro average observed at best epoch's 6-class scheme (Table 29). After performing the disagreement measure, best macro 79.53% achieved in 6 class scheme,77.08 in class-7 scheme and 72.54 in class 11 scheme (Table 30).

Below tables shows the final performance of each class. After performing the ensemble majority voting to all epochs in five model seeds of 11 variants we achieved best 74.79 higher performance in 6 class scheme (Table28).

All epochs Ensemble		2_1_02	2_1_06	2_1_07	2_1_08	2_1_09	2_1_10	3_2_06	3_2_06_a	3_2_08	3_2_09	3_2_10
Class-6	Macro_Avg	67.1	69.17	70.38	71	73.45	71.51	70.73	71.81	73.79	**74.79**	74.74
Class-7	Macro_Avg	69.63	69.81	72.31	72.94	74.09	73.48	70.11	72.81	73.15	74.36	**74.64**
Class-8	Macro_Avg	69.04	72.24	70.35	69.79	70.44	71.56	69.86	68.98	71.02	70.95	72.98
Class-9	Macro_Avg	66.15	67.75	69.57	69.44	69.72	68.74	68.44	69.85	68.86	69.55	67.16
Class-10	Macro_Avg	68.4	67.14	69.89	68.78	68.11	68.88	68.44	67.96	68.89	67.95	69.01
Class-11	Macro_Avg	66.95	69.59	67.43	66.75	65.85	67.33	67.77	67.64	69.42	66.18	67.76

Table 28: All epoch's ensemble macro result:

We achieved macro 76.19% in 6 class scheme for best epochs (Table 29).

Best Epoch Ensemble		2_1_02	2_1_06	2_1_07	2_1_08	2_1_09	2_1_10	3_2_06	3_2_06_a	3_2_08	3_2_09	3_2_10
Class-6	Macro_Avg	70.49	73.97	74.22	74.11	74.61	74.4	73.95	73.96	75.33	73.82	**76.19**
Class-7	Macro_Avg	68.72	69.06	73.21	73.46	74.31	74.61	71.67	72.98	72.78	75.11	75.98
Class-8	Macro_Avg	72.06	72.38	70.08	73.14	70.81	**74.19**	70.39	73.53	71.02	72.54	73.79
Class-9	Macro_Avg	70.94	69.34	69.34	69.47	69.38	72.32	69.48	70.11	70.73	70.49	71.9
Class-10	Macro_Avg	69.17	69.68	71.91	68.81	70.24	72.62	70.95	68.68	70.38	69.2	70.3
Class-11	Macro_Avg	65.46	66.53	67.31	67.68	67.14	68.93	69.22	68.23	70.23	66.82	70.68

Table 29: Best epoch ensemble macro score

We achieved disagreement measure score 79.53 macro in 6 class scheme (Table 30).

Disagreement Measure performance

All classes Disagreement Macro Avg Performance		Thrsh-1	Thrsh-3	Thrsh-5	Thrsh-7	Thrsh-9	Thrsh-11	Thrsh-13	Thrsh-15	Thrsh-17	Thrsh-19
Class-6	Macro_Avg	72.74	72.61	73.43	74.15	74.63	73.64	74.87	76.74	77.09	**79.53**
Class-7	Macro_Avg	64.65	70.33	70.39	73.36	77.06	76.31	77	77.04	**77.08**	76.2
Class-8	Macro_Avg	67.01	72.43	72.21	71.9	72.31	72.8	73.96	73.39	72.9	72.96
Class-9	Macro_Avg	66.19	68.26	69.46	70.36	71.33	71.61	71.7	70.37	70.73	71.54
Class-10	Macro_Avg	64.91	70.08	71.92	71.32	71.93	71.25	72.13	72.49	72.1	72.89
Class-11	Macro_Avg	69.52	71.73	72.54	72.13	70.37	71.73	71.46	71.51	71.05	71.56

Table 30: Disagreement measure ensemble macro score

8. Discussion on difficulties

This article presents the obstacles that the model faced during development and how they have been overcome. Previously, many authors tested a series of Deep learning models for citation context analysis by experimenting with different ways of encoding features. But each model produced different outcomes using the same dataset. And no single model achieved the good performance. So, to avoid such problems first we have checked the differences between various deep learning models outcomes and created ensemble techniques to improve the performance. After that we tested disagreement measure between each pair of classifiers to combine the all classifiers. And based on this we find the classifier which is most different from others. Finally voting classifier is used to combine the several base models' predictions in order to improve the performance. After applying these procedures we succeeded best 76.19 % macro average for best epochs in 6-class(Table 29) scheme and disagreement score 79.53 % in 6 class scheme (Table 30).

9. Project Management

9.1 Project Schedule

This section outlines the initial approach I established for doing this research, as well as the changes I made to it as the research continued. First week was setup to examine the previous literatures, and completed in-time. Second week, continued studies on the dataset and compared with previous dataset for better understanding. Consistent attempts to apply the ensemble model for all epochs in the 3rd and 4th week. In 5th week, Observation and record of all results in excel file during 5th week. Later 6th week, spent time to analyse classifier diversities to choose weak classifiers to combine. Summarizing 7th and 8th week developed disagreement measure to improve the performance. 9th and 10th week recorded disagreement results and compared with previous models. 11th week dedicated to study, edit thesis and prepare PPT. Overall, 12th and 13th week dedicated to report writing and Presentation.

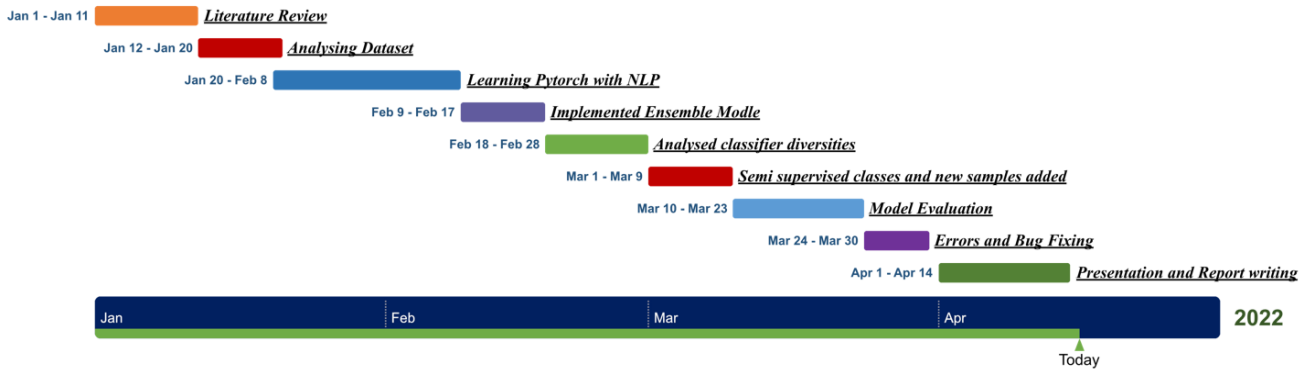


Figure 40: Original Project Time line

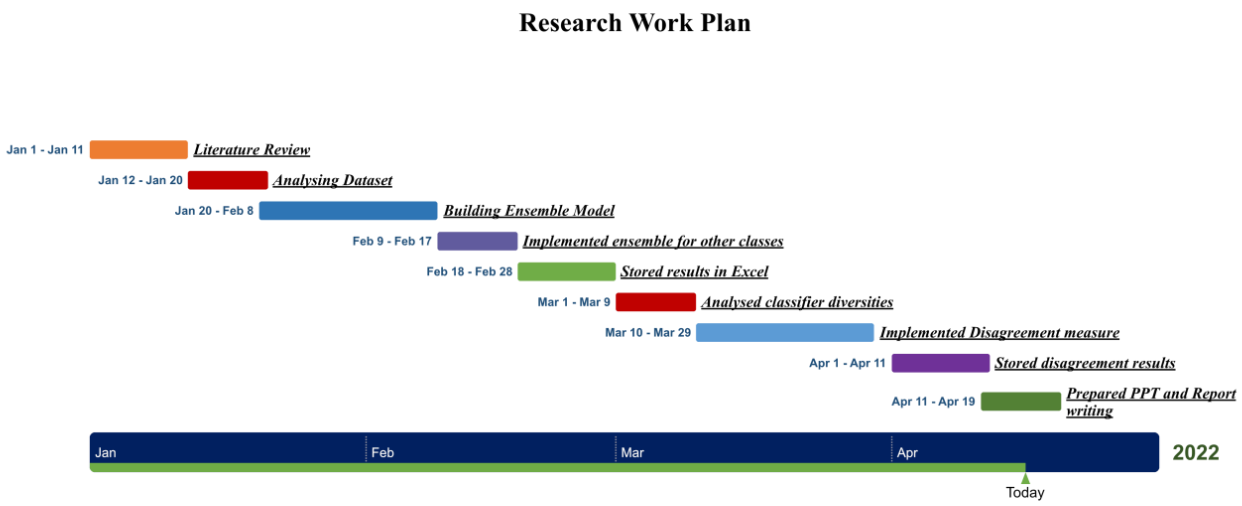


Figure 41: Modified Project Timeline

9.2 Risk Management

This section describes the risks that were identified and faced when developing and training the model.

9.2.1 Risk

It is not an easy task to improve the performance based on prior results. Because the outcomes of each model different. We must examine the classification performance of various models before deploying approaches.

9.2.2 Analysis:

Above ensemble and disagreement result tables (Table 28, 29 and 30) shows the per-class results of five models that performed well using at least one annotation strategy. Larger or even less intellectually challenging citation functions, such as "Neutral"/"Background" and "Usage"/"Uses," were simpler to detect. Similar findings were made by Teufel et al. (2006b, 2010) and Cohan et al. (2019). For example, the best models for future citations achieved 83.33 % F1 with 6-class scheme (Table 13, Best epoch performance). And also future higher precision 100% (Future) in 7, 10 and 11 class scheme, 91.67% in 8 class scheme. The highest performances could reach 86.96 % F1 with 100% precision (Table 14, best epoch performance). Neutral reached 82.11 % F1 with 80.70 % precision in class 10 scheme (Table 16).

9.3 Quality Management

The techniques which were used to improve the citation context performance is ensemble majority voting and diversity measures. The evaluations are done on validation data to test the performance.

9.4 Social, Legal, Ethical and Professional Considerations

Professor Dr. Xiaorui provided the required dataset to evaluate the model performance, and he annotated the citation sentence using Teufel's citation scheme.

10 Conclusion

This part displays the system's accomplishments as well as future work that could be done on the developed model.

10.1 Achievements:

Ensemble approaches for citation function classification performance have been researched and refined throughout this work. When compared to prior writers who employed the citation approach, we can find that the best epoch model ensemble performance has improved.

11 Future work

In future Semi supervised models will be used to test the citation function performance. There are several semi supervised classes like extension, similar, support and usage will be used to train the models. Final performance will be evaluated by adding different percentages of new samples. To make the findings comparable, I'll use the datasets and techniques outlined in the literature review to test my constructed models.

12. References

Aljohani, N.R., Fayoumi, A., & Hassan, S. (2021a). A novel focal-loss and class-weight-aware convolutional neural network for the classification of in-text citations. Journal of Information Science (Mar, 2021), 1–14. https://doi.org/10.1177/0165551521991022

Aljohani, N.R., Fayoumi, A., & Hassan, S. (2021b). An in-text citation classification predictive model for a scholarly search system. Scientometrics, 126, 5509–5529. https://doi.org/10.1007/s11192-021-03986-z

Beltagy, I., & Cohan, A. (2019). In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing *(EMNLP-IJCNLP)*, pages 3615–3620.

Cohan, A., Ammar, W., & Field Cady. (2019). In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota. Association for Computational Linguistics, pages 3586–3596

David, O., & Richard M. (1999). Popular ensemble methods: an empirical study. Journal of Artificial Intelligence Research, 11:169–198

Ding, Y., Zhang, G., & Zhai, C. (2014). Content-based citation analysis: The next generation of citation analysis. Journal of the Association for Information Science and Technology, 65(9), 1820-1833. https://doi.org/10.1002/asi.23256

Gabriele, Z., & Cunningham, P. (2001). Using Diversity in Preparing Ensembles of Classifiers Based on Different Feature Subsets to Minimize Generalization Error. Machine Learning: ECML 2001, 2167(1995):576–587.

Ghosal, T.,Tiwary, P., & Stahl, C.(2022). Towards establishing a research lineage via identification of Science Studies (Advance publication). https://doi.org/10.1162/qss_a_00170

Jurgens, D., & Jurafsky, D. (2018). Measuring the Evolution of a Scientific Field through Citation Frames. Transactions of the Association for Computational Linguistic, Pages-391–406. https://doi.org/10.1162/tacl_a_00028

Lauscher, A., Glavaš, G., & Eckert, K. (2017). Investigating convolutional networks and domain-specific embeddings for semantic classification of citations. In Proceedings of the 6th international workshop on mining scientific publications pages 24-28.

Sagi, O., & Rokach,L.,(2018). "Ensemble learning: A survey." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 8.4.

Schafer, U., & Kasterka, U., (2010). Scientific ¨ authoring support: A tool to navigate in typed citation graphs. In Proceedings of the NAACL-HLT 2010 Workshop on Computational Linguistics and Writing, pages 7–14.

Teufel, S., Siddharthan, A., & Tidhar, D. (2006a). An annotation scheme for citation function. In Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue (SIGdial'06), 80–87. https://aclanthology.org/W06-1312

Teufel, S., Siddharthan, A., & Tidhar, D. (2006b). Automatic classification of citation function. In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP'06), 103–110. https://aclanthology.org/W06-1613

Teufel, S. (2010). The Structure of Scientific Articles: Applications to Citation Indexing and Summarization. Centre for the Study of Language & Information.

Teufel, S. (2017). Do "Future Work" sections have a purpose? Citation links and entailment for global scientometric questions. In Proceedings of the 2nd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries co-located with the 40th ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'17). http://ceur-ws.org/Vol-1888/paper1.pdf

Xiaojin Zhu and Andrew B. Goldberg. 2009. Introduction to Semi-Supervised Learning. Morgan and Claypool.

Appendix A: Project Presentation

Introduction

Citation Context Analysis:

Citation context analysis is a key task in analysing scientific reports, particularly for understanding the flow of knowledge dissemination in the innovation process. Citations are discovering why previous works are cited how prior work affect future researchers' uptake ?

Why we need citation context analysis ?

- To see if the present study was inspired by the cited work.
- To demonstrate the intellectual roots of the current study.
- Use of something that has previously been presented by another study as part of the present authors survey.
- To add to the present research by providing more information.

Teufel designed 13 annotation schemes, and Jurgens devised six based on Teufel's annotation schemes, which I employed in my research.

Teufels citation	Jurgens functions	Description
Weak	Weakness	The cited approach flaws
CoCoGM	Comparison Or Contrast	Goals or approaches that contrast or compare (neutral).
CoCoRO	Comparison Or Contrast	In the Results, there is a contrast/comparison (neutral).
CoCo-	Comparison Or Contrast	Unfavorable Contrast/Comparison (modern work is superior than that referenced).
CoCoXY	Background	Compare and contrast the two cited strategies.
PBas	Basis	As a preliminary step, the author refers to the mentioned work.
PUse	Usage	The author employs techniques, methods, information, and ideas.
PModi	Extends	The effort of the author and the work mentioned are compatible/ supportive of one another.
PMot	Motivation	Used to encourage people to work on the present paper
Psim	Similar	The work of the authors and the work mentioned are comparable.
Psup	Support	The author's work and the quoted work are compatible and complement each other.
Neut	Neutral	Delivers information related to this domain
Future	Future	Source of future work.

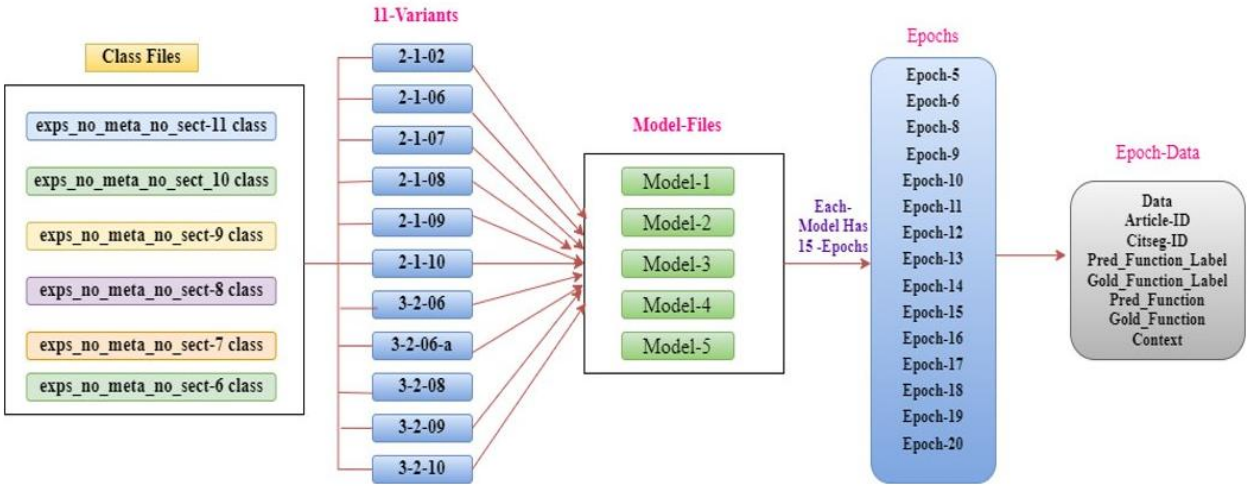
Background Study

- Over the last fifty years, a vast number of research publication shave focused on citation activity. Citations are main component of a scientific field's current condition, indicating how researchers organize their work and affecting future researchers' uptake.
- Teufel and his colleagues published the biggest dataset of 2,829 assessed citations in 2006. These citations were taken from the ACL anthology reference corpus of data science articles. The authors of the article generated the 12 different sorts of annotations that were employed in their citation categorization scheme. Most of the subsequent work in this arena has been based on this dataset and the categorization system used.
- Lauscher employed CNNs to categorise citation functions instead of typical machine learning models that need extensive feature engineering. Although the proposed citation scheme's sentiment classes appear to be acceptable for citation sentiment categorization, the citation functions are less useful in strengthening citation analysis methodologies than previous schemes.
- Jurgens enhanced the accuracy of overall CFC from 54.9 percent macro F1 on a dataset using an application-friendly six-class annotation system, that is "Background", "Future" (Work), "Comparison Or Contrast", "Motivation", "Uses", and "Extends".
- Munkhdalai used a simple citation system and a deep learning architecture to develop a solution for end-to-end citation function categorization. Support vector machine(SVM) ,Long short-term memory(LSTM), LSTM + Global Attention, Bi-LSTM, and Bi-LSTM + Global Attention were the models utilised. Regarding citation context, unidirectional LSTMs with global attention had the highest F1 score of 75.86 percent. And for the citation phrase, bidirectional LSTMs with CAN (Compositional attention network) received the greatest F1 score of 68.61 percent.
- Gabriele Zenobi proposes an approach for constructing classifier ensembles that prioritises variability between ensemble members. The extensive experiments from that study show that ensembles based on diversity are more accurate than ensembles relying on error rate. This also shown that by using ensemble approaches with a bigger, clean dataset and 11 different model outputs, we can boost classification accuracy.

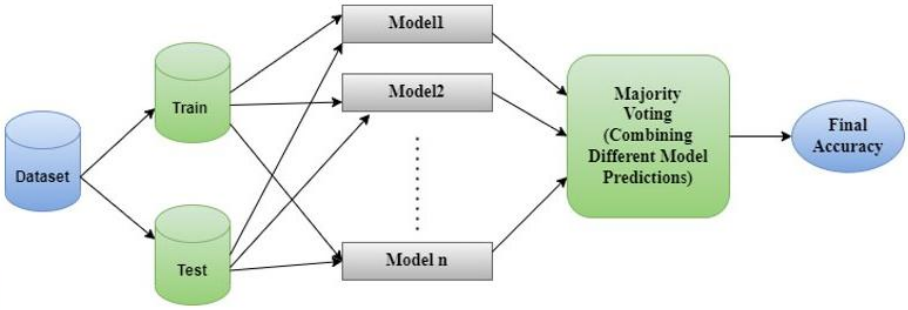
Objective

- Previously, we have trained a series of Deep learning models for citation context analysis by experimenting with different ways of encoding features; We observed different models performed differently using the same dataset, and no single model achieved high performance.
- As a result, we believed that we could investigate the differences between various Deep learning models and create an ensemble techniques to out of proper analysis.

- **Dataset-Overview**



Methodology



- In this report we choose ensemble techniques to improve the Prediction performance. Because ensemble techniques are believed to be the best advanced approach for several machine learning problems.
- We used python programming language to create the code. It has several libraries including Pandas, NumPy, TensorFlow, Seaborn, and Scikit-learn.
- We imported Scikit-learn packages to generate the confusion matrix and display the categorization result
- As part of this study, I performed three procedures.

  - Ensemble applied for best epochs of each model seed
  - Ensemble applied for All epochs of each model seed
  - Disagreement measure performed for respective model seeds to check the final accuracy.

Methodology: Ensemble

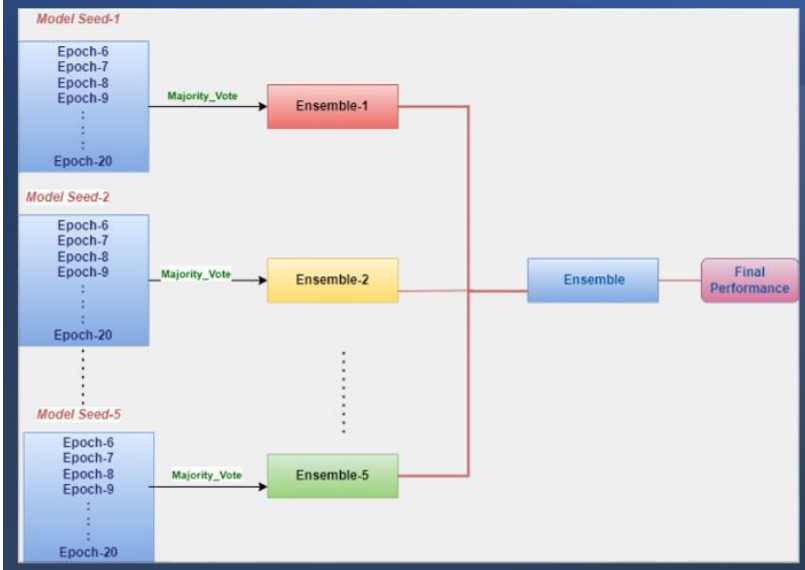
Why Ensemble Learning?

There are two primary reasons to utilise an ensemble over a single model:

1. Performance: When compared to single model an ensemble can generate higher predictions and produce better results.

2. Robustness: The spread of forecasts and prediction performance are reduced by using an ensemble.

Ensemble Process for All Epochs



Rule:

- To build an ensemble we need to analyze the similarities and differences between the performances of the variant models.

- **Procedure:**
- Each Model Seed has 15 epochs
- Ensemble Majority voting is performed to combine the multiple predictions of each model seed epochs.
- Using each model outcomes again ensemble is performed to predict the final performance.

Class 6 and 11 Ensemble Performance

Class-6

	precision	recall	f1-score	support
ComorCon	0.7708	0.7762	0.7735	143
Uses	0.7917	0.8333	0.8120	114
Extends	0.9231	0.4800	0.6316	25
Motivation	0.6512	0.6364	0.6437	44
Future	0.9091	0.7692	0.8333	13
Background	0.7809	0.8066	0.7935	243
accuracy			0.7766	582
macro avg	0.8045	0.7170	0.7479	582
weighted avg	0.7797	0.7766	0.7748	582

Class -11

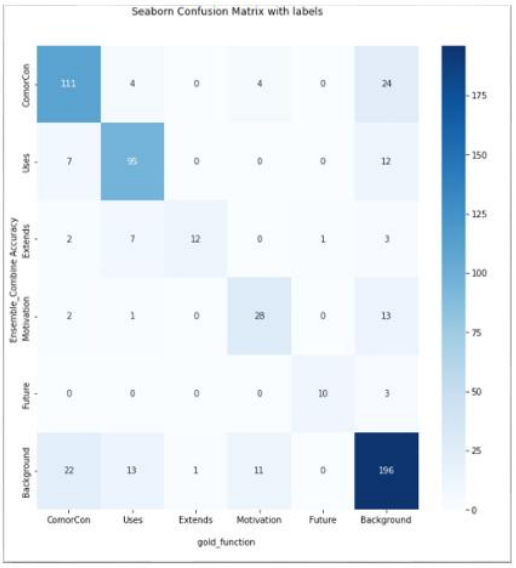
	precision	recall	f1-score	support
Basis	0.6667	0.4800	0.5581	25
CoCoGM	0.7600	0.7600	0.7600	50
CoCoRes	0.6923	0.7826	0.7347	23
CoCoXY	0.6500	0.5652	0.6047	23
Future	0.9000	0.6923	0.7826	13
Motivation	0.5962	0.7045	0.6458	44
Neutral	0.7783	0.8136	0.7956	220
Similar	0.7097	0.7097	0.7097	31
Support	0.5714	0.5333	0.5517	15
Usage	0.8214	0.8070	0.8142	114
Weakness	0.7895	0.6250	0.6977	24
accuracy			0.7509	582
macro avg	0.7214	0.6794	0.6959	582
weighted avg	0.7523	0.7509	0.7495	582

The above results are ensemble performance of class-6 using '3-2-09' variant and class –7 using '3-2-10' variant of all **epochs.**

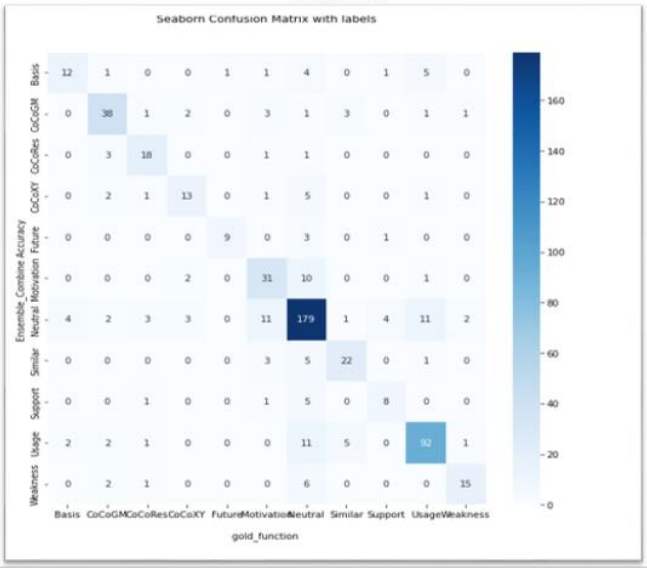
Class-6 (3-2-09) - F1-score of 74.79 was achieved
Class-11(2-1-06) - F1-score of 69.59was achieved

Class –6 and 11 performance confusion matrix

Class-6



Class-11



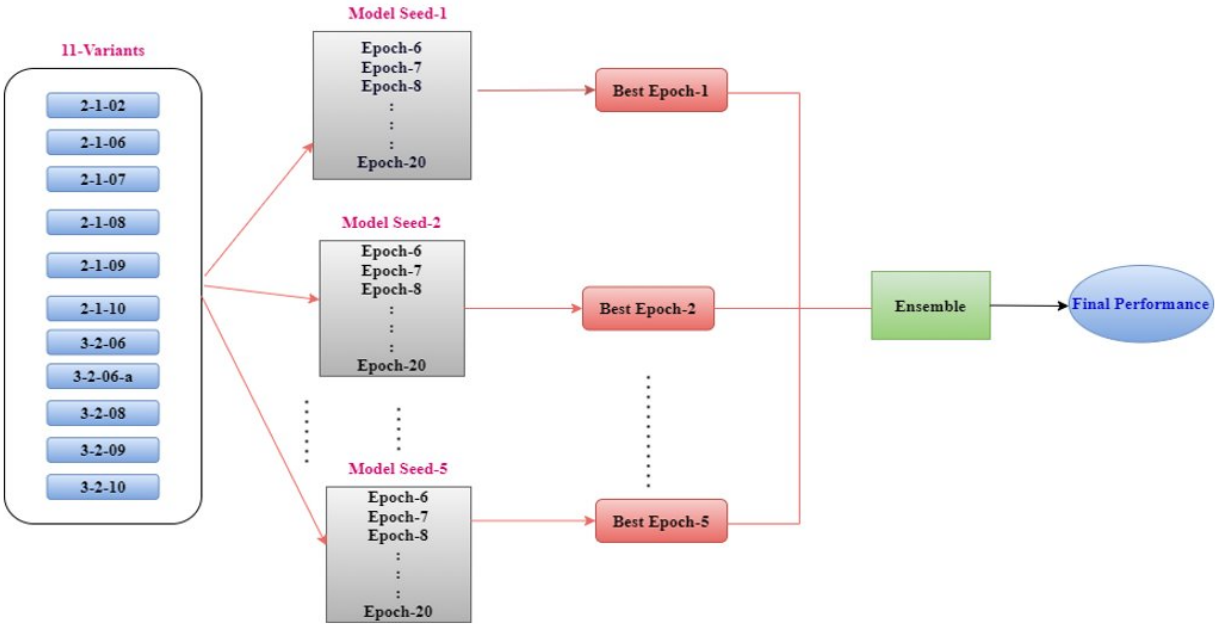
Class-11 All Variants Performance

		Ensemble 2_1_02	Ensemble 2_1_06	Ensemble 2_1_07	Ensemble 2_1_08	Ensemble 2_1_09	Ensemble 2_1_10	Ensemble 3_2_06	Ensemble 3_2_06_a	Ensemble 3_2_08	Ensemble 3_2_09	Ensemble 3_2_10
Basis	**Precision**	64.71	66.67	66.67	63.16	57.14	70.59	68.42	75	60	61.11	68.18
	Recall	44	48	40	48	48	48	52	48	48	44	60
	F1-Score	52.38	55.81	50	54.55	52.17	57.14	59.09	58.54	53.33	51.16	63.83
CoCoGM	**Precision**	75	76	75	70.59	73.47	73.58	74.47	72.55	75.47	77.08	72.73
	Recall	66	76	78	72	72	78	70	74	80	74	80
	F1-Score	70.21	76	76.47	71.29	72.73	75.73	72.16	73.27	77.67	75.51	76.19
CoCoRes	**Precision**	72	69.23	60.71	68	65.38	60.71	62.96	65.38	66.67	60.71	65.38
	Recall	78.26	78.26	73.91	73.91	73.91	73.91	73.91	73.91	78.26	73.91	73.91
	F1-Score	75	73.47	66.67	70.83	69.39	66.67	68	69.39	72	66.67	69.39
CoCoXY	**Precision**	58.33	65	69.23	55.56	66.67	62.5	75	64.71	68.75	52.94	61.54
	Recall	60.87	56.52	39.13	43.48	34.78	43.48	52.17	47.83	47.83	39.13	34.78
	F1-Score	59.57	60.47	50	48.78	45.71	51.28	61.54	55	56.41	45	44.44
Future	**Precision**	81.82	90	90	90	90	90	81.82	90	81.82	81.82	90.91
	Recall	69.23	69.23	69.23	69.23	69.23	69.23	69.23	69.23	69.23	69.23	76.92
	F1-Score	75	78.26	78.26	78.26	78.26	78.26	75	78.26	75	75	83.33
Motivation	**Precision**	59.57	59.62	58.49	65.22	62.5	65.96	63.04	65.85	69.05	68.89	60.42
	Recall	63.64	70.45	70.45	68.18	68.18	70.45	65.91	61.36	65.91	70.45	65.91
	F1-Score	61.54	64.58	63.92	66.67	65.22	68.13	64.44	63.53	67.44	69.66	63.04
Neutral	**Precision**	76.19	77.83	77.49	75.86	76.07	77.53	77.49	74.15	77.02	77.73	78.73
	Recall	80	81.36	81.36	80	80.91	80	81.36	79.55	82.27	80.91	79.09
	F1-Score	78.05	79.56	79.38	77.88	78.41	78.75	79.38	76.75	79.56	79.29	78.91
Similar	**Precision**	75.86	70.97	71.88	70.59	70	68.57	70.97	73.33	74.29	66.67	70.97
	Recall	70.97	70.97	74.19	77.42	67.74	77.42	70.97	70.97	83.87	70.97	70.97
	F1-Score	73.33	70.97	73.02	73.85	68.85	72.73	70.97	72.13	78.79	68.75	70.97
Support	**Precision**	57.14	57.14	70	58.33	58.33	54.55	53.33	61.54	77.78	53.33	60
	Recall	53.33	53.33	46.67	46.67	46.67	40	53.33	53.33	46.67	53.33	40
	F1-Score	55.17	55.17	56	51.85	51.85	46.15	53.33	57.14	58.33	53.33	48
Usage	**Precision**	74.8	82.14	75.41	78.07	74.38	75.21	77.97	75	78.76	75	74.8
	Recall	80.7	80.7	80.7	78.07	78.95	79.82	80.7	81.58	78.07	78.95	80.7
	F1-Score	77.64	81.42	77.97	78.07	76.6	77.45	79.31	78.15	78.41	76.92	77.64
Weakness	**Precision**	70.59	78.95	87.5	66.67	73.68	82.35	66.67	72.22	71.43	77.78	72.73
	Recall	50	62.5	58.33	58.33	58.33	58.33	58.33	54.17	62.5	58.33	66.67
	F1-Score	58.54	69.77	70	62.22	65.12	68.29	62.22	61.9	66.67	66.67	69.57
Macro_Avg		66.95	69.59	67.43	66.75	65.85	67.33	67.77	67.64	69.42	66.18	67.76

Class-6 and 11 variants F1-Score with Accuracy

		2_1_02	2_1_06	2_1_07	2_1_08	2_1_09	2_1_10	3_2_06	3_2_06_a	3_2_08	3_2_09	3_2_10
Class-6												
ComorCon	F1-Score	77.86	77.82	77.4	78.75	79.3	78.47	77.24	78.62	80.28	77.35	78.23
Uses	F1-Score	75.21	75.42	78.79	77.25	78.6	78.41	78.97	78.97	80.69	81.2	79.48
Extends	F1-Score	35.9	36.84	50	48.65	60.47	51.28	56.41	52.38	53.33	63.16	57.14
Motivation	F1-Score	62.79	66.67	64.29	62.79	64.37	65.12	63.53	64.44	67.47	64.37	68.24
Future	F1-Score	72	80	72	78.26	78.26	75	69.23	76.92	80	83.33	84.62
Background	F1-Score	78.86	78.28	79.84	80.32	79.68	80.8	79.02	79.5	80.98	79.35	80.74
	Macro_Avg	67.1	69.17	70.38	71	73.45	71.51	70.73	71.81	73.79	74.79	74.74
Class-11												
Basis	F1-Score	52.38	55.81	50	54.55	52.17	57.14	59.09	58.54	53.33	51.16	63.83
CoCoGM	F1-Score	70.21	76	76.47	71.29	72.73	75.73	72.16	73.27	77.67	75.51	76.19
CoCoRes	F1-Score	75	73.47	66.67	70.83	69.39	66.67	68	69.39	72	66.67	69.39
CoCoXY	F1-Score	59.57	60.47	50	48.78	45.71	51.28	61.54	55	56.41	45	44.44
Future	F1-Score	75	78.26	78.26	78.26	78.26	78.26	75	78.26	75	75	83.33
Motivation	F1-Score	61.54	64.58	63.92	66.67	65.22	68.13	64.44	63.53	67.44	69.66	63.04
Neutral	F1-Score	78.05	79.56	79.38	77.88	78.41	78.75	79.38	76.75	79.56	79.29	78.91
Similar	F1-Score	73.33	70.97	73.02	73.85	68.85	72.73	70.97	72.13	78.79	68.75	70.97
Support	F1-Score	55.17	55.17	56	51.85	51.85	46.15	53.33	57.14	58.33	53.33	48
Usage	F1-Score	77.64	81.42	77.97	78.07	76.6	77.45	79.31	78.15	78.41	76.92	77.64
Weakness	F1-Score	58.54	69.77	70	62.22	65.12	68.29	62.22	61.9	66.67	66.67	69.57
	Macro_Avg	66.95	69.59	67.43	66.75	65.85	67.33	67.77	67.64	69.42	66.18	67.76

2. Ensemble Process for best epochs



Class-6 and 11 Best Epochs Ensemble Performance

	precision	recall	f1-score	support
ComorCon	0.7887	0.7832	0.7860	143
Uses	0.7946	0.7807	0.7876	114
Extends	0.8235	0.5600	0.6667	25
Motivation	0.7250	0.6591	0.6905	44
Future	0.9091	0.7692	0.8333	13
Background	0.7808	0.8354	0.8072	243
accuracy			0.7852	582
macro avg	0.8036	0.7313	0.7619	582
weighted avg	0.7859	0.7852	0.7839	582

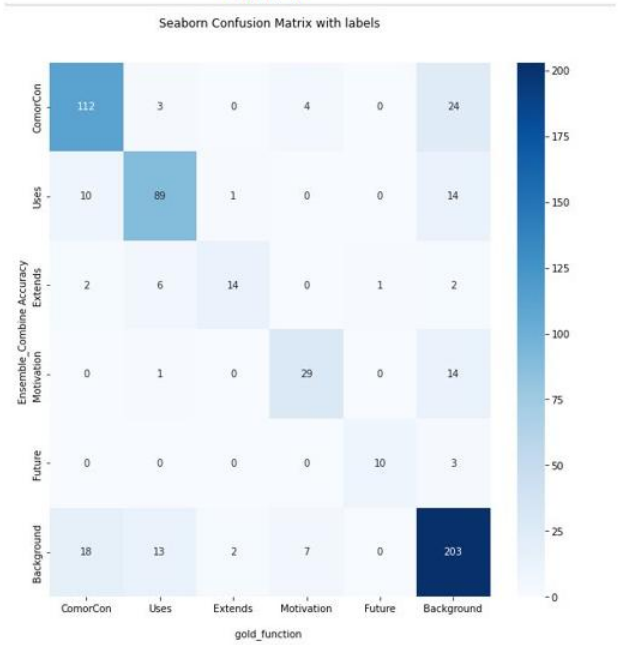
Class-6

	precision	recall	f1-score	support
Basis	0.6818	0.6000	0.6383	25
CoCoGM	0.7843	0.8000	0.7921	50
CoCoRes	0.7273	0.6957	0.7111	23
CoCoXY	0.6429	0.3913	0.4865	23
Future	1.0000	0.7692	0.8696	13
Motivation	0.6977	0.6818	0.6897	44
Neutral	0.7895	0.8182	0.8036	220
Similar	0.7097	0.7097	0.7097	31
Support	0.5333	0.5333	0.5333	15
Usage	0.7559	0.8421	0.7967	114
Weakness	0.8421	0.6667	0.7442	24
accuracy			0.7595	582
macro avg	0.7422	0.6825	0.7068	582
weighted avg	0.7587	0.7595	0.7564	582

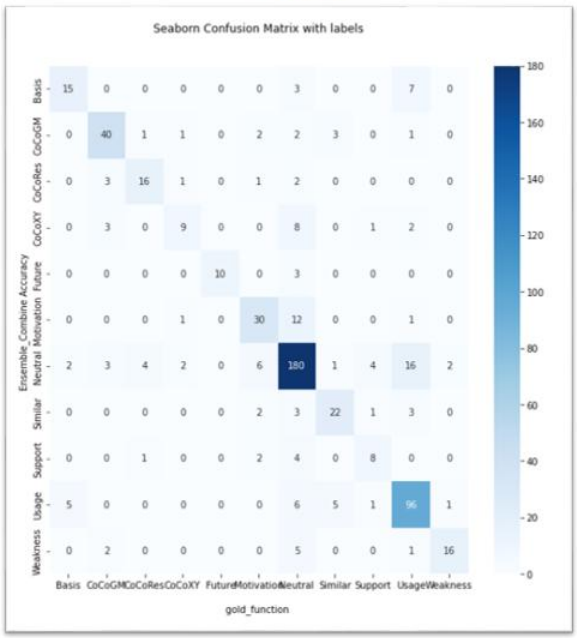
Class-11

Class-6 and 7 Confusion Matrix

Class-6



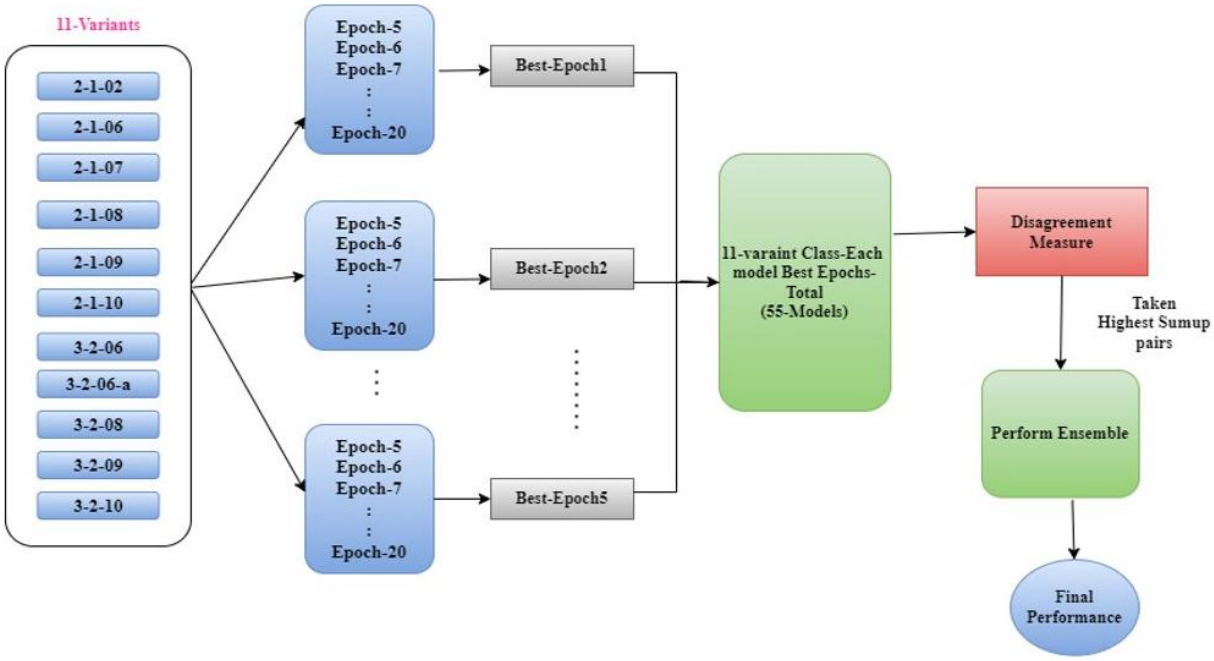
Class-11



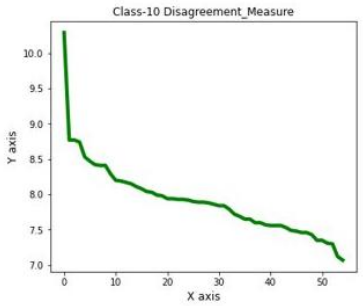
Class 6 and 11 F1-Score and Accuracy Results

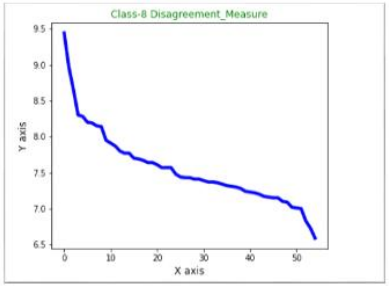
	Class-6	2_1_02	2_1_06	2_1_07	2_1_08	2_1_09	2_1_10	3_2_06	3_2_06_a	3_2_08	3_2_09	3_2_10
ComorCon	F1-Score	76.12	78.35	76.55	77.46	79.72	77.46	77.13	79.33	80.84	77.74	78.6
Uses	F1-Score	77.73	76.6	78.22	81.55	79.83	79.13	80.69	80.52	83.05	80	78.76
Extends	F1-Score	53.33	58.54	63.41	60	61.9	60	61.54	56.41	57.14	53.66	66.67
Motivation	F1-Score	62.65	67.5	66.67	62.07	64.29	68.29	65	66.67	69.88	67.47	69.05
Future	F1-Score	74.07	83.33	80	83.33	81.82	80	80	80	80	83.33	83.33
Background	F1-Score	79.02	79.51	80.48	80.24	80.08	81.51	79.35	80.82	81.06	80.71	80.72
	Macro_Avg	70.49	73.97	74.22	74.11	74.61	74.4	73.95	73.96	75.33	73.82	76.19
	Class-11											
Basis	F1-Score	51.16	50	45	51.16	53.33	52.38	66.67	61.9	57.78	51.16	63.83
CoCoGM	F1-Score	67.37	72.73	75.25	70.59	71.15	69.81	71.58	73.27	79.21	75	79.21
CoCoRes	F1-Score	72	73.91	66.67	73.47	70.83	70.59	66.67	60.39	73.47	68	71.11
CoCoXY	F1-Score	52.38	52.17	57.14	50	53.66	60	55	53.66	57.14	55.81	48.65
Future	F1-Score	75	78.26	75	78.26	83.33	81.82	75	81.82	75	75	86.96
Motivation	F1-Score	61.7	64.52	64.44	65.93	63.74	65.17	68.18	65.12	66.67	68.89	68.97
Neutral	F1-Score	76.96	77.88	79.65	78.4	78.56	80.54	80.09	76.61	79.82	80	80.36
Similar	F1-Score	71.19	71.19	74.19	75.76	70	72.73	75	70.97	82.54	66.67	70.97
Support	F1-Score	53.33	50	53.85	53.85	53.85	59.26	59.26	58.06	56	48.28	53.33
Usage	F1-Score	78.48	79.65	77.82	78.95	76.47	79.31	78.81	77.82	78.26	76.39	79.67
Weakness	F1-Score	60.47	61.54	71.43	68.09	63.64	66.67	65.12	61.9	66.67	69.77	74.42
	Macro_Avg	65.46	66.53	67.31	67.68	67.14	68.93	69.22	68.23	70.23	66.82	70.68

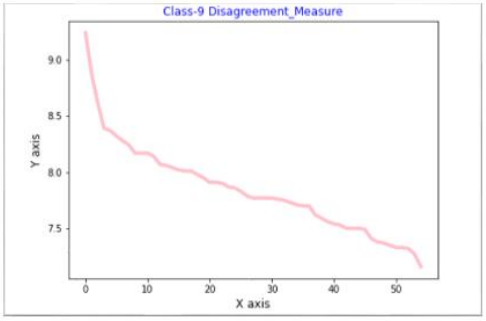
3. Disagreement Measure



Curve Of Sum of Disagreement all Base classifiers







Class 8 Disagreement Performance

Threshold Range		Thrsh-1	Thrsh-3	Thrsh-5	Thrsh-7	Thrsh-9	Thrsh-11	Thrsh-13	Thrsh-15	Thrsh-17	Thrsh-19
ComorCon	Precision	72.86	77.78	78.87	78.87	79.17	81.69	80.56	82.86	80.56	80.56
	Recall	68	74.67	74.67	74.67	76	77.33	77.33	77.33	77.33	77.33
	F1-Score	70.34	76.19	76.71	76.71	77.55	79.45	78.91	80	78.91	78.91
Similar	Precision	50.85	58	60.42	70	71.79	71.79	74.36	73.68	72.97	72.97
	Recall	83.33	80.56	80.56	77.78	77.78	77.78	80.56	77.78	75	75
	F1-Score	63.16	67.44	69.05	73.68	74.67	74.67	77.33	75.68	73.97	73.97
Uses	Precision	75.96	76.99	76.99	79.28	78.63	77.31	77.31	78.81	78.45	77.97
	Recall	69.3	76.32	76.32	77.19	80.7	80.7	80.7	81.58	79.82	80.7
	F1-Score	72.48	76.65	76.65	78.22	79.45	78.97	78.97	80.17	79.13	79.31
Basis	Precision	48	64	66.67	56	63.16	66.67	73.68	63.64	61.9	68.42
	Recall	48	64	64	56	48	56	56	56	52	52
	F1-Score	48	64	45.31	56	54.55	60.87	63.64	59.57	56.52	59.09
Motivation	Precision	60.87	65.22	56.36	60	66.67	68.29	69.05	68.29	69.05	66.67
	Recall	63.64	68.18	70.45	68.18	63.64	63.64	65.91	63.64	65.91	63.64
	F1-Score	62.22	66.67	62.63	63.83	65.12	65.88	67.44	65.88	67.44	65.12
Future	Precision	78.57	85.71	80	78.57	76.92	75	81.82	81.82	81.82	81.82
	Recall	84.62	92.31	92.31	84.62	76.92	69.23	69.23	69.23	69.23	69.23
	F1-Score	81.48	88.89	85.71	81.48	76.92	72	75	75	75	75
Weak	Precision	72.22	60.87	66.67	70	82.35	82.35	82.35	82.35	87.5	87.5
	Recall	54.17	58.33	58.33	58.33	58.33	58.33	58.33	58.33	58.33	58.33
	F1-Score	61.9	59.57	62.22	63.64	68.29	68.29	68.29	68.29	70	70
Background	Precision	77.24	82.01	82.13	81.67	79.85	80.53	80.23	80.38	79.78	79.78
	Recall	75.7	78.09	76.89	81.67	83.67	84.06	84.06	84.86	84.86	84.86
	F1-Score	76.46	80	79.42	81.67	81.71	82.26	82.1	82.56	82.24	82.24
Macro_Avg		67.01	72.43	72.21	71.9	72.31	72.8	73.96	73.39	72.9	72.96

Class-10 Disagreement Measure Results

Tested highest sumup all base classifers with different threshold range to predict the performance.

	Threshold_Range	Thrsh-1	Thrsh-3	Thrsh-5	Thrsh-7	Thrsh-9	Thrsh-11	Thrsh-13	Thrsh-15	Thrsh-17	Thrsh-19
Basis	Precision	46.43	54.55	74	68.18	70	70	75	75	70	73.68
	Recall	52	48	56	60	56	56	60	60	56	56
	F1-Score	49.06	51.06	62.22	63.83	62.22	62.22	66.67	66.67	62.22	63.64
CoCoGM	Precision	57.63	69.81	80.43	77.78	78.72	79.17	80.85	79.59	81.25	81.25
	Recall	68	74	74	70	74	76	76	78	78	78
	F1-Score	62.39	71.84	77.08	73.68	76.29	77.55	78.35	78.79	79.59	79.59
CoCoRes	Precision	58.62	64.29	63.33	63.33	67.86	63.33	65.52	65.52	65.52	65.52
	Recall	68	72	76	76	76	76	76	76	76	76
	F1-Score	62.96	67.92	69.09	69.09	71.7	69.09	70.37	70.37	70.37	70.37
CoCoXY	Precision	50	50	60	68.42	73.33	75	76.47	68.42	72.22	72.22
	Recall	60.87	56.52	52.17	56.52	47.83	52.17	56.52	56.52	56.52	56.52
	F1-Score	54.9	53.06	55.81	61.9	57.89	61.54	65	61.9	63.41	63.41
Future	Precision	83.33	91.67	90.91	81.82	83.33	81.82	90	90	81.82	81.82
	Recall	76.92	84.62	76.92	69.23	76.92	69.23	69.23	69.23	69.23	69.23
	F1-Score	80	88	83.33	75	80	75	78.26	78.26	75	75
Motivation	Precision	58.49	66.67	68.18	66.67	66.67	63.64	64.44	67.44	69.05	69.05
	Recall	70.45	68.18	68.18	63.64	63.64	63.64	65.91	65.91	65.91	65.91
	F1-Score	63.92	67.42	18.18	65.12	65.12	63.64	65.17	66.67	67.44	67.44
Neutral	Precision	81.28	83.17	79.74	78.24	79.17	79.15	79.24	79.17	78.93	79.67
	Recall	66.67	75.88	81.14	82.02	83.33	81.58	82.02	83.33	83.77	84.21
	F1-Score	73.25	79.36	80.43	80.09	81.2	80.35	80.6	81.2	81.28	81.88
Similar	Precision	48.15	61.36	69.05	70.73	70.73	71.79	70	73.68	74.36	74.36
	Recall	72.22	75	80.56	80.56	80.56	77.78	77.78	77.78	80.56	80.56
	F1-Score	57.78	67.5	74.36	75.32	75.32	74.67	73.68	75.68	77.33	77.33
Usage	Precision	74.77	74.19	75.63	77.39	76.47	76.67	76.47	76.07	77.39	77.78
	Recall	72.81	80.7	78.95	78.07	79.82	80.7	79.82	78.07	78.07	79.82
	F1-Score	73.78	77.31	77.25	77.73	78.11	78.63	78.11	77.06	77.73	78.79
Weakness	Precision	76.19	85	83.33	83.33	83.33	78.95	73.68	82.35	77.78	83.33
	Recall	66.67	70.83	62.5	62.5	62.5	62.5	58.33	58.33	58.33	62.5
	F1-Score	71.11	77.27	71.43	71.43	71.43	69.77	65.12	68.29	66.67	71.43
Macro_AVG		64.91	70.08	71.92	71.32	71.93	71.25	72.13	72.49	72.1	72.89

Class-11 Best epoch Disagreement Measure

		Thrsh-1	Thrsh-3	Thrsh-5	Thrsh-7	Thrsh-9	Thrsh-11	Thrsh-13	Thrsh-15	Thrsh-17	Thrsh-19
Basis	Precision	61.9	65.22	78.95	76.19	68.42	68.18	66.67	70	72.22	70
	Recall	52	60	60	64	52	60	56	56	52	56
	F1-Score	56.52	62.5	68.18	69.57	59.09	63.83	60.87	62.22	60.47	62.22
CoCoGM	Precision	76.6	74.51	78	78.43	79.59	79.59	76.47	76.47	79.59	79.59
	Recall	72	76	78	80	78	78	78	78	78	78
	F1-Score	74.23	75.25	78	79.21	78.79	78.79	77.23	77.23	78.79	78.79
CoCoRes	Precision	77.27	75	75	69.23	69.23	69.23	69.23	72	69.23	69.23
	Recall	73.91	78.26	78.26	78.26	78.26	78.26	78.26	78.26	78.26	78.26
	F1-Score	75.56	76.6	76.6	73.47	73.47	73.47	73.47	75	73.47	73.47
CoCoXY	Precision	66.67	64.71	73.33	71.43	62.5	64.71	70.59	72.22	70.59	75
	Recall	60.87	47.83	47.83	43.48	43.48	47.83	52.17	56.52	52.17	52.17
	F1-Score	63.64	55	57.89	54.05	51.28	55	60	63.41	60	61.54
Future	Precision	80	100	100	100	90	90	90	90	90	90
	Recall	61.54	76.92	76.92	76.92	69.23	69.23	69.23	69.23	69.23	69.23
	F1-Score	69.57	86.96	86.96	86.96	78.26	78.26	78.26	78.26	78.26	78.26
Motivation	Precision	57.69	68.89	64.58	73.81	73.81	75	73.17	69.77	68.18	68.18
	Recall	68.18	70.45	70.45	70.45	70.45	68.18	68.18	68.18	68.18	68.18
	F1-Score	62.5	69.66	67.39	72.09	72.09	71.43	70.59	68.97	68.18	68.18
Neutral	Precision	77.88	82.19	80.36	78.11	78.88	79.4	79.4	79.65	78.72	78.81
	Recall	80	81.82	81.82	82.73	83.18	84.09	84.09	83.64	84.09	84.55
	F1-Score	78.92	82	81.08	80.35	80.97	81.68	81.68	81.6	81.32	81.58
Similar	Precision	74.29	72.73	71.88	71.88	73.53	73.53	73.53	71.88	71.88	71.88
	Recall	83.87	77.42	74.19	74.19	80.65	80.65	80.65	74.19	74.19	74.19
	F1-Score	78.79	75	73.02	73.02	76.92	76.92	76.92	73.02	73.02	73.02
Support	Precision	53.85	53.33	66.67	61.54	57.14	66.67	66.67	66.67	61.54	61.54
	Recall	46.67	53.33	53.33	53.33	53.33	53.33	53.33	53.33	53.33	53.33
	F1-Score	50	53.33	59.26	57.14	55.17	59.26	59.26	59.26	57.14	57.14
Usage	Precision	77.78	74.6	75.19	76.42	76.03	77.97	77.5	77.5	77.5	78.15
	Recall	79.82	82.36	85.09	82.46	80.7	80.7	81.58	81.58	81.58	81.58
	F1-Score	78.79	78.33	79.84	79.32	78.3	79.31	79.49	79.49	79.49	79.83
Weakness	Precision	88.89	84.21	78.95	82.35	78.95	76.19	82.35	75	83.33	88.24
	Recall	66.67	66.67	62.5	58.33	62.5	66.67	58.33	62.5	62.5	62.5
	F1-Score	76.19	74.42	69.77	68.29	69.77	71.11	68.29	68.18	71.43	73.17
Macro_Avg		69.52	71.73	72.54	72.13	70.37	71.73	71.46	71.51	71.05	71.56

Summary

Ensemble Performed for all epochs

- class-6 '3-2-09' was achieved highest accuracy of 74.79%

Ensemble performed for best epochs and got below accuracy

- Class-6- (3-2-10)- 76.19%
- Class-7- (3-2-10)- 75.98
- Class-8- (2-1-10)- 74.19

Disagreement Measure

- Class-6 got 74.87% at threshold range 13
- Class-8 –73.96% at threshold range -13
- Class –11 got 72.54 at Threshold range 5

Conclusion

- Ensemble majority voting has been developed to combine the multiple model predictions and the overall performance was improved.
- The disagreement measure between each pair of classifiers also calculated to combine all the classifiers and based on this we found a classifier which is most different from others.
- We achieved highest F1 score and Accuracy for different classes.
- The findings obtained using our technique are not comparable because the data collection and citation methods used by previous writers were different,.
- In the future work, we will apply Semi supervised models to test and increase the citation function performance

Appendix B: Certificate of Ethical Approval

Certificate of Ethical Approval

Applicant:	Jyothi Yendamuri
Project Title:	Citation Analysis Using Semi-Supervised and Ensemble Techniques

This is to certify that the above named applicant has completed the Coventry University Ethical Approval process and their project has been confirmed and approved as Low Risk

Date of approval:	14 Feb 2022
Project Reference Number:	P133961

Appendix C: Source Code Links

Google Drive Links:

Ensemble All Epochs Code Google drive link

https://drive.google.com/drive/folders/1hqjKle4wjjacEbfbDVHbaJD5YW1JMnF9?usp=sharing

Ensemble Best Epochs Source code:

https://drive.google.com/drive/folders/1IKqFZnfa9VdLcHovpqZdrsMikWNgyL6q?usp=sharing

(OR)

GitHub Links:

Ensemble all epoch's and best epochs source code links

To view GitHub files please use 'nbviewer' application.

https://github.com/JyothiYendamuri/Citation-Function-Classification-Using-Ensemble.git

https://github.com/JyothiYendamuri/Citation-classification-repository

Nbviewer site link

https://nbviewer.org/

Nbviewer link example:

https://nbviewer.org/github/JyothiYendamuri/Citation-Function-Classification-Using-Ensemble/blob/main/Class10_All%20Epochs_Final_Ensemble1.ipynb