Inter-Personal Interaction and Relationship (IPIR) Named Entity Recognition Methodology

# Introduction

# Assumptions

The author assumes that the reader is familiar with statistical and rule based Natural Language Processing (NLP) in general, in particular, stand-off techniques and tools including those that underly the UIMA and GATE software platforms, where the original text is not altered, but annotations, or markups are created that refer back to positional or character-based offsets within the original document. The author assumes the reader is familiar with NLP platforms that are built using a pipeline composed of *annotators*, where a representation of the input document and associated markups is passed from annotator to annotator. Each annotator adds, edits or deletes the markups that refer back to the original text. Each annotator handles a specific or atomic task that down-pipeline annotators rely upon. (This is the way of GATE and UIMA). Describing algorithms that utilize this technology stack are best described by the sequence of annotators in general, only delving into specific annotator algorithm details where needed.

For these task, the author assumes the reader is familiar with clinical medical terminologies used for Natural Language Processing, and in particular the Unified Medical Language System (UMLS), and the International Classification of Functioning Disability and Health and (ICF).

# Prior Related Software

# Prior Terminologies/Lexica Employed

# Terminology Sources

# Pipelines, Algorithms, and Rules

## Pre-processing Annotators

The IPIR NER built upon the Mental Functioning NER’s pipeline, which, in turn, is built upon a general NLP pipeline to segment text into constituent document-decomposition parts, which is built upon a UIMA infrastructure.

### Document Decomposition Pipeline

The Document Decomposition Pipeline identifies Sentences, Terms, Tokens, Section Names, Section Zones within the text.

#### Important: Term Lookup

Of note here: the secret sauce if you will: the terms identified are based first on dictionary/lexica lookup, where care has been taken to craft the content of the dictionaries, and where each of the dictionary entries contain term attributes including semantic categorization(s) for each term and the pedigree of where the term was sourced from. It is the semantic categorizations for each term identified within the text that is paramountly used down pipeline. While there are many ways to identify terms within the text, how and what is contained within the dictionaries is the salient part to each task.

### Mental Functioning NER Pipeline and Annotator

The mental functioning NER pipeline is a series of annotators to identify Mental Functioning Ontology (MFO) concepts within text. As such, it includes annotators that create Behavior Evidence, Emotion Evidence, and (financial, social, institutional) Support Evidence from the terms that have these semantic categories prior to an annotator that creates Mental Functioning Evidence and Mental Functioning Mention Markups. The Mental Functioning Ontology Annotator identifies and creates Mental Functioning Evidence Mark ups from terms that have semantic categories that generalize up to Mental Functioning within the MFO Ontology.

Specifically, a Mental Functioning Mention encapsulates the outer spans of any MFO evidence found within the scope of a sentence[1] or slot:value[2], or check-box[3]. Note that MFO mentions will not be made from within the bounds of a section name.

This is over-generative. Filters are applied to weed out spurious mentions. Underutilized but built in is a filter to filter out any mentions found within document types (identified from the document-decomposition pipeline) that indicate this document should not be processed. Currently, the pipeline does not identify any forbidden document types. In prior tasks, documents identified a-priori from their source HER were tagged as administrative or clinical, and those documents tagged as administrative were filtered out. The plan is when page and document classification become available, this feature will become important, particularly to ignore mentions that come from templated text coming from SSA forms found within the corpora we’ve been processing.

### Ad-hoc Filters

There are additional ad-hoc filters applied as well.

Documents that include SSA forms, where the templated text contains what look like mental functioning mentions, but are not, because the mentions are within templated text that are most likely to be instructions on how to fill out the form. Such mentions that included *see [x]* within the scope of the sentence are filtered out.

Similarly, mentions that can be attributed to the author of the document, not the client or patient are filtered out. These mentions have evidence that includes the use of pronouns like *we* and *our* along with use of pronouns like *you*, *your*, *his*, *her*, *she*, and *him*. Such mentions are also tagged with *Provider* attribution as a feature at the term and sentence level.

### Customization for inFACT Processing

The Mental Functioning Ontology pipeline is being embedded within the inFACT workflow, where it is important that the same segmentation is used across all the inFACT processing. As such, the functionality being described here assumes the input is GATE/Interlingua formatted files that include sentence mark-ups. An additional preprocessing annotator was added to the MFO pipeline that swaps in the inFACT sentences for the framework sentences. Also, upstream inFACT processing will have created *IPIR\_yes*, *IPIR\_no*, *Comcog\_yes*, and *Comcog\_no* annotations within the bounds of each sentence via trained LLM models.

For the purposes of inFACT, it is these inFACT labeled \_yes mark-up is then further processed down pipeline. The MFO evidence and mentions that overlays these markups are available for the downstream processing. Outside of inFACT, equivalent MFO markups that generalize up to IPIR and ComCog mentions have been generated and can be used independent of inFACT processing.

## The IPIR Sub-categorization Annotator

The IPIR Sub-Categorization annotator further sets 8 attributes to each IPIR\_yes.

For each IPIR\_yes sentence span, all the MFO generated evidence that generalizes to IPIR Activities from the Ontology, along with IPIR Participation evidence, behavior and Support evidence that cover that span are gathered and if there is any of these found, further processing is done.

There are 6 attributes set:

* ICF d710-d729 General Interpersonal Interactions
* ICF d730 Relating with Strangers
* ICF d740 Formal Relationships
* ICF d7400 Relating with Persons in Authority
* ICF d750 Informal Social Relationships
* ICF d760 Family Relationships
* ICF d770 Intimate Relationships
* ICF d779 Particular interpersonal relationships, other specified and unspecified

There had been a rule that if there was no evidence for an *IPIR\_yes* mention, AND there was one or less evidence of a person (pronouns, roles, person names, etc.) it was assumed that a provider was the author of the mention, and that the provider had the role of one in authority, so thus marked. However, it was observed that these mentions were contained to only some observation around dress and hygiene, so this rule was backed off because it produced too many false positives in our training and test sets.

The annotation guidelines specify that at least one of these attributes get set for an *IPIR\_yes* mention. If no evidence of any of these above attributes were found within an *IPIR\_ye*s, span, the *d779 other* attribute gets set to satisfy that criteria.

### Non Relevant Evidence

There is a check to throw out mentions that are not IPIR, ComCog, or Mental Functioning related due to some known context. For example, if the sentence has the word *invoice*, or *billing* in it, this is not an MFO mention. Within the IPIR terminologies, there are 162 terms marked as not IPIR related. More examples include *balance support*, *math support*, and with *respect to*. There are an additional 1389 terms marked as *NotMFO* which includes *case number*, *department of social services*, *social security act* along with terms like *swelling* and *please*.

These non-relevant evidence terms were garnered and added from frequency distributions of terms appearing in false positive mentions and never appearing in true positive or false negative mentions.

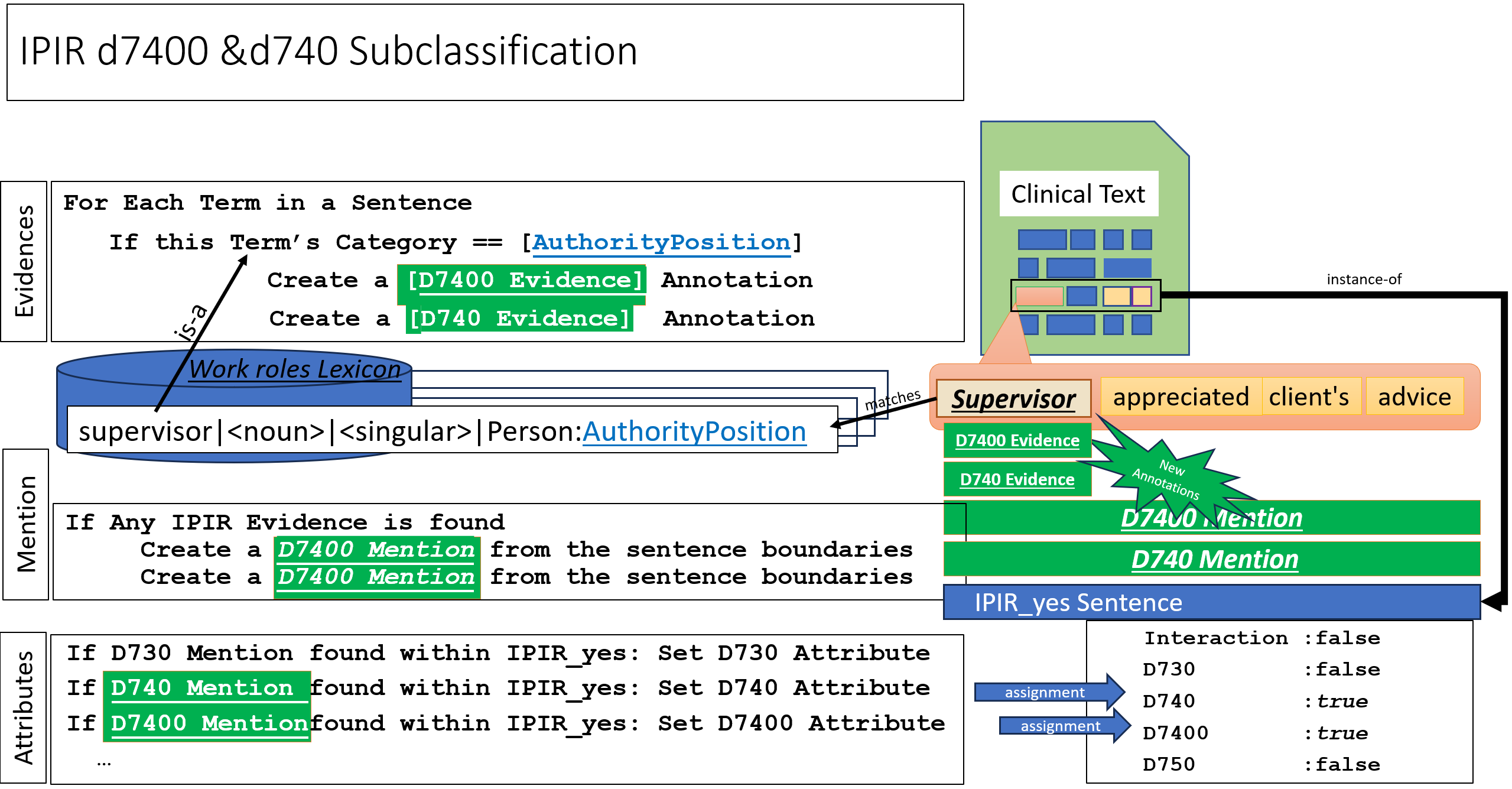
The IPIR annotator creates evidence markups for each term it deems IPIR evidence, and the sentence span with the evidence in it (that has not been filtered out by non-relevant counter evidence) is used to create one or more of the IPIR sub category mark up. This is in addition to having the IPIR\_yes attributes set. There are two reasons for the creation of these additional markups:

* Visualization: in GATE and UIMA, you cannot see attributes. Making them markups with spans allow one to turn that layer on or off to see each of the attributes.
* Evaluation: The existing evaluation functionality counts at the mark-up level, not at the attribute level. It was easier to create markups than to alter the complicated evaluation code.

ICF d710-d729 General Interpersonal Interaction markups are created when there is *Emotion Evidence* present. I had also tried creating interaction markups when the sentence span included person evidence and behavior evidence. While the resulting mentions looked appropriate, this caused too many false positives and was backed off.

ICF d730 Relating with Stranger markups are created when there is evidence of Stranger Relationship Evidence. These kinds of terms come from a lexicon of 100 terms like *shopper*, *public places*, and *visitor*. The bulk of these terms were garnered from examples seen in the data.

ICF d740 and ICF d7400 Formal Relationship markups are created when there are terms categorized as *Non-Authority* or *Sometimes-Authority*. These kinds of terms come from a person-role and title lexicon that was augmented with *non-Authority*, *Sometimes-Authority* and *Authority-Position* semantic categories, garnered from the Bureau of Labor Statistics public sources. Those terms marked with Authority-Positionalso had d7400 evidence and mentions made.



*Figure* x: IPIR d7400 & d740 Subclassification

ICF d750 Informal Social Relationship markups are created when there is *Informal Relationship Evidence* found in the span of the *IPIR yes* sentence. If there is no such evidence, but there is the presence of pronouns, an informal social relationship markup is made. This works only because we already know this is an *IPIR\_yes* sentence, so cases where the pronouns relate to possessions rather than to other people would have already been filtered out. The UMLS terms categorized with family history and person roles were the impetus for the informal relationships lexicon employed here. There are currently 125 terms in this resource.

ICF d760 Family Relationship markups are created when there is *Family History Evidence* found in the span of the *IPIR\_yes* sentence. The underlying terminology used here are the terms in the UMLS categorized with their Family-History semantic type.

ICF d770 Intimate Relationship markups are created when there is *Intimate Relationship Evidence* found in the span of the IPIR\_yes sentence. A small lexicon of 130 terms were created mostly from the guideline examples, observations, introspection and use of thesauri to garner adjectives and nouns that indicate an intimacy. Such terms include *significant other*, *old lady*, and *inseparable*.

ICF d779 Particular interpersonal relationships, other specified and unspecified markups are also created when there is evidence of *Other-Relationship* Evidence found in the span of IPIR\_yes sentence. There were only 30 or so terms identified with this tag, which included *others*, *nobody*, and *withdrawn*. The guidelines and training data were source for this terminology.

# Annotation Guidelines, Selected Corpus for Training and Testing, Annotation Efforts

The domain expert annotators created a list containing likely IPIR words within pages that have IPIR mentions in them. From a set of 1.5 million SSA claimant pages from 2019[x], 400 pages were selected based on a uniform frequency distribution among the presence of n number of words for annotation purposes. That is, some records were selected with many instances of IPIR words, and some were selected with only a few instances of IPIR words. The set was randomly segmented into a small set for inter-rater-reliability (IRR) (20 pages), development set (80 pages), training set(160 pages), and a validation set (80 pages).

Two domain expert annotators meticulously developed annotation guidelines for IPIR (see x), honed the definitions and assessed their internalization of the guidelines into annotation efforts resulting with IRR= 0.70. Once the IIR task was complete and the IIR acceptable, the rest of the samples annotated by one or the other annotator (not duplicative annotation) in an effort favoring getting annotations expediently to the modelers to model.

# Dictionary Lookup and Lexicon Sources

The dictionary lookup process involves loading all lexica into a hash table. Framework uses a lookup mechanism that starts with the span of a sentence, works from right to left (yes, right to left), that first uses all the tokens as a key to be looked up in the hash. When failing, the left-most token is dropped and that key is looked up in the hash. This continues until a key is found. Once found, the tokens consumed change right side of the window to be searched. This insures maximal span matches, favoring matches where the head of a term is found. The English language is structured where the important parts of terms or the head a phrase or term is the last or right part. This right to left window technique favors capturing at least the important parts of terms when there is ambiguity.

Framework comes with and runs with the following default lexica (unless otherwise tuned)

* UMLS SPECIALIST 2022 Lexicon
* Document Titles
* CCDA Section Names
* Page Header and Page Footer Evidence
* Common Clinical Slot Names
* Date and Time Terms
* Labs Terminology (from LOINC)
* Blood Panel Terminology (from LOINC)
* Units of Measure Terminology
* Person Evidence Terms
* ~~USPS Address Terms~~
* ~~Clinical Demographics Terms~~

For the ComCog task, additional lexica were curated, with two approaches used to gather the additional terminology.

Terms thought to be the most general Comcog classes from the MF Ontology were gathered. Each of these were used as seed terms to traverse through the 2022 UMLS to gather all descendent terms to be incorporated into the terminology used for lookup. Communication and Cognition activities is poorly covered in the UMLS, and this process was disappointing, only picking up ICF Activities and Participation subclasses, for the most part.

It was recognized that communication and cognition activities evidence are likely to be verbs and adverbs. An effort was made to classify the University of Colorado’s VerbNet classes into IPIR and Comcog categories manually. That effort was expanded to have each of the Comcog categories (about 162 classes) to further be subclassified into the 20 subclasses manually. This was done by 3 domain expert annotators and one software engineer (me). This was a daunting task involving classifying over 5000 lemmas resulting from the 162 classes. There was reason for classifying at the lemma level rather than class level. The classes were too inclusive, with most classes involving lemmas (in VerbNet terminology: members) from the classes that didn’t really fit the semantic relatedness criteria we were classifying at. We had time constraints so the effort was a quick and rough effort with the understanding that we’d go back and do better curation when time allowed. For the most part, the VerbNet Comcog tagged lemmas is the terminology employed.

# In a Rule Based System, what does Training entail?

The ComCog NER was created off the codebase that was developed for the IPIR NER task, with the new terminology swapped in.

Once the terminology was roughly in place, the NER was coded and the The Comcog training set was then used to tune the Comcog subcategorization task, involving looking at each category, addressing the more prevailing issue more fp’s than tn’s or visa-versa. For tn’s, term frequencies from the missed mentions were created, the top (manually recognized) salient terms from the list were considered to be added to the lexicon, where each played with to see if each added term added significant false positives or not. Those that did not were kept in the lexicon. For false positives, my evaluation tool includes the evidence that caused the false positive. Each term from the top frequency false positive evidence were removed. Those removed that did not incur too many false negatives as a consequence were kept out of the lexicon. Removed is an overstatement: for the term in consideration, the Comcog subcategorization category was commented out, so the term, while still known, would not be used to classify that subcategory.

# Efficacy

The efficacy for IPIR subcategories is below. The results for the training and validation are being provided to give an indication of the stability and how much degradation there was between training and testing sets. The intuition behind looking at the degradation also gives an impression of how similar the distribution of reportable classes between the training and testing sample. Given the small amount of overall data, it is nearly impossible to keep a balance between the sets for each subcategory.

*Table* x: IPIR Subcategory Efficacy At the Sentence Level, Training (n=671 pages)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Class** | **F-Score** | **Recall** | **Precision** | **Total** | **TP** | **FP** | **FN** |
| **Interaction** | 0.572 | 0.466 | 0.740 | 1145 | 533 | 187 | 612 |
| **D730** | 0.519 | 0.444 | 0.625 | 45 | 20 | 12 | 25 |
| **D740** | 0.757 | 0.834 | 0.694 | 885 | 738 | 326 | 147 |
| **D7400** | 0.530 | 0.449 | 0.647 | 49 | 22 | 12 | 27 |
| **D750** | 0.587 | 0.609 | 0.567 | 419 | 255 | 195 | 164 |
| **D760** | 0.737 | 0.645 | 0.858 | 392 | 253 | 42 | 139 |
| **D770** | 0.566 | 0.455 | 0.747 | 123 | 56 | 19 | 67 |
| **D779** | 0.458 | 0.355 | 0.646 | 299 | 106 | 58 | 193 |
| **Totals** | 0.641 | 0.591 | 0.700 | 3357 | 1983 | 851 | 1374 |
| **Avg** | 0.675 | 0.608 | 0.789 | 479.571 | 283.286 | 121.571 | 196.286 |

*Table* x: IPIR Subcategory Efficacy at the Sentence Level, Testing (n=180 pages)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Class** | **F-Score** | **Recall** | **Precision** | **Total** | **TP** | **FP** | **FN** | **Notes** |
| **Interaction** | 0.541 | 0.453 | 0.671 | 247 | 112 | 55 | 135 | [1] |
| **D730** | 0.348 | 0.267 | 0.500 | 15 | 4 | 4 | 11 | [2] |
| **D740** | 0.735 | 0.798 | 0.681 | 203 | 162 | 76 | 41 | [3] |
| **D7400** | 0.348 | 0.267 | 0.500 | 15 | 4 | 4 | 11 | [4] |
| **D750** | 0.536 | 0.584 | 0.496 | 101 | 59 | 60 | 42 | [5] |
| **D760** | 0.659 | 0.602 | 0.727 | 93 | 56 | 21 | 37 | [6] |
| **D770** | 0.613 | 0.561 | 0.676 | 41 | 23 | 11 | 18 | [7] |
| **D779** | 0.430 | 0.329 | 0.619 | 79 | 26 | 16 | 53 | [8] |
| **Totals** | 0.600 | 0.562 | 0.644 | 794 | 446 | 247 | 348 |  |
| **Avg** | 0.526 | 0.483 | 0.609 |  |  |  |  |  |

1. Performance stable, training score was f=.57, testing f=.54
2. Drop noted from training score was f=.52, testing f=.35
3. Performance stable, training score was f=.75, testing f=.74
4. Drop noted from training score was f=.53, testing f=.35
5. Performance stable, training score was f=.59, testing f=.54
6. Drop noted from training score was f=.74, testing f=.66
7. Improvement noted from training score was f=.57, testing f=.61
8. Performance stable, training score was f=.46, testing f=.43

# Failure Analysis

# Discussion