PUCRS

ESCOLA POLITÉCNICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO
MESTRADO EM CIÊNCIA DA COMPUTAÇÃO

BERNARDO SCAPINI CONSOLI

ENRICHING PORTUGUESE WORD EMBEDDINGS WITH VISUAL INFORMATION

Porto Alegre
2021

PÓS-GRADUAÇÃO - *STRICTO SENSU*

Pontifícia Universidade Católica
do Rio Grande do Sul

PONTIFICAL CATHOLIC UNIVERSITY OF RIO GRANDE DO SUL
SCHOOL OF TECHNOLOGY
COMPUTER SCIENCE GRADUATE PROGRAM

ENRICHING PORTUGUESE WORD EMBEDDINGS WITH VISUAL INFORMATION

BERNARDO SCAPINI CONSOLI

Master Thesis submitted to the Pontifical Catholic University of Rio Grande do Sul in partial fulfillment of the requirements for the degree of Master in Computer Science.

Advisor: Prof. Dra. Renata Vieira

Porto Alegre
2021

Ficha Catalográfica

C755e Consoli, Bernardo Scapini

Enriching portuguese word embeddings with visual information / Bernardo Scapini Consoli. – 2021.
102 p.
Dissertação (Mestrado) – Programa de Pós-Graduação em Ciência da Computação, PUCRS.

Orientadora: Profa. Dra. Renata Vieira.

1. Word Embeddings. 2. Multimodal. 3. Portuguese. 4. Geosciences. 5. Named Entity Recognition. I. Vieira, Renata. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da PUCRS
com os dados fornecidos pelo(a) autor(a).
Bibliotecária responsável: Clarissa Jesinska Selbach CRB-10/2051

BERNARDO SCAPINI CONSOLI

ENRICHING PORTUGUESE WORD EMBEDDINGS WITH VISUAL INFORMATION

This Master Thesis has been submitted in partial fulfillment of the requirements for the degree of Master in Computer Science, of the Computer Science Graduate Program, School of Technology of the Pontifical Catholic University of Rio Grande do Sul

Sanctioned on March 19, 2021.

COMMITTEE MEMBERS:

Prof. Dr. Rafael Heitor Bordini (PPGCC/PUCRS)

Prof^{a}. Dr^{a}. Aline Marin Paes Carvalho (IC/UFF)

Prof. Dra. Renata Vieira (PPGCC/PUCRS - Advisor)

I dedicate this work to my family, friends and mentors who helped me on my way to achieving my dreams.

“The limits of my language means the limits of my world.”
(Ludwig Wittgenstein)

ENRIQUECENDO WORD EMBEDDINGS DA LÍNGUA PORTUGUESA COM INFORMAÇÕES VISUAIS

RESUMO

Essa dissertação foca no enriquecimento de word embeddings pré-treinados na língua Portuguesa com o uso de informações visuais. Essas informações foram extraídas de imagens retratando certos termos do vocabulário e embeddings visuais "imaginadas" para termos sem dados de imagem. Essas embeddings enriquecidas foram testadas contra seus modelos textuais originais em tarefas comuns de PLN, sendo elas: relação entre palavras, predição de analogias, reconhecimento de entidades nomeadas e similaridade de sentenças. Essas tarefas foram utilizadas para descobrir se o enriquecimento tem impacto sobre a performance dos embeddings nas tarefas em questão. Os resultados demonstram um aumento de desempenho para algumas tarefas, o que indica que o enriquecimento com dados visuais é útil para tarefas de PLN baseadas em word embeddings.

Palavras-Chave: word embeddings, multimodal, português, geociências, reconhecimento de entidades nomeadas, similaridade de sentenças, relacionamento de palavras.

ENRICHING PORTUGUESE WORD EMBEDDINGS WITH VISUAL INFORMATION

ABSTRACT

This dissertation focuses on the enrichment of existing Portuguese word embeddings with visual information in the form of visual embeddings. This information was extracted from images portraying given vocabulary terms and imagined visual embeddings learned for terms with not image data. These enriched embeddings were tested against their text-only counterparts in common NLP tasks, namely: word relatedness, analogy prediction, named entity recognition, and sentence similarity. These tasks were used to ascertain whether the enrichment has an impact on the embedding's performance the above mentioned tasks. The results show an increase in performance for several tasks, which indicates that visual information fusion for word embeddings can be useful for word embedding based NLP tasks.

Keywords: word embeddings, multimodal, portuguese, geosciences, named entity recognition, sentence similarity, word relatedness.

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

NLP – Natural Language Processing
PLN – Processamento de Linguagem Natural
NER – Named Entity Recognition
NILC – Núcleo Interinstitucional de Linguística Computacional
USP – Universidade de São Paulo
UFRGS – Universidade Federal do Rio Grande do Sul
PUCRS – Pontifícia Universidade Católica do Rio Grande do Sul
MSE – Mean Standard Error
NN – Neural Network
WE – Word Embedding
BR – Brazilian Portuguese
PT – European Portuguese
BBPFT300 – BBP fastText model, 300-dimension vector space
NILCFT300 – NILC fastText model, 300-dimension vector space
NILCFT100 – NILC fastText model, 100-dimension vector space
NILCW2V100 – NILC Word2Vec model, 100-dimension vector space
PETROVECFT – PetroVec fastText model
PETROVECHYBRIDFT – PetroVec fastText model, Hybrid version
PETROVECW2V – PetroVec Word2Vec model
PETROVECHYBRIDW2V – PetroVec Word2Vec model, Hybrid version

CONTENTS

1. INTRODUCTION

Language modelling technologies have been dominated by semantic embedding models ever since Mikolov et al. (2013b) and Mikolov et al.'s (2013a) [30, 31] popularization of Word Embeddings, a concept which revolutionized the field of Natural Language Processing (NLP). The architecture presented by the authors, Word2Vec, has been used as basis for many works across the spectrum of NLP tasks, as attested by nearly 45,000 citations when accounting both of the aforementioned papers (as recorded by Google Scholar), mainly because of the fact that training this architecture only requires raw text, and no human-made annotation (the main obstacle in training machine learning models).

Many architectures based on the original intuition behind Word2Vec have become popular since 2013. The most prevalent, besides the original Word2Vec, are fastText [18] and GloVe [36]. An evolution upon the concept, taking into account the current context of a word, not just an amalgamation of all contexts with which it was trained, was introduced by Peters et al. (2018) [37], with their ELMO architecture, and popularized by Devlin et al.'s (2019) [14] BERT architecture. These Contextual Embeddings, as they are sometimes referred to, have taken off and are currently the bleeding edge technology in the field of semantic embeddings with the immense GPT-3 model, from OpenAI[1], achieving state-of-the-art results in multiple NLP tasks [5].

All of the mentioned embedding architectures have at least one model trained on Portuguese language corpora. The *Núcleo Interinstitucional de Linguística Computacional* (NILC), from the *Universidade de São Paulo* (USP), for example, has several Word2Vec, fastText and GloVe models for the Portuguese language available within their Word Embedding repository[2]. The Allen Institute for AI maintains an ELMO model repository which includes a Portuguese language model[3]. BERTimbau[44], a Portuguese language BERT model, was recently developed and added to the Hugging Face[4] library. These models, and others, have been used to advance the state-of-the-art in several Portuguese language NLP tasks [40, 26, 16].

Beyond these efforts to further enhance the usage of text in the training of word embedding models, be it Portuguese language text or otherwise, an effort to enrich these embeddings with other modes of information also arose. The most studied modes of information used to enhance Word Embeddings are the visual mode (composed of images and video), and the audio mode (composed of sounds, spoken language, music, etc.). These efforts spurred the creation of multimodal embedding fusion architectures, used to join embeddings of disparate modes into a single embedding representing all fused knowledge. An

[1] https://openai.com/
[2] http://www.nilc.icmc.usp.br/embeddings
[3] https://allennlp.org/elmo
[4] https://huggingface.co/

example of this is the concatenation based architecture of Bruni et al. (2014) [6], which arrived at promising results after proposing that multiple embeddings of different modes could be concatenated, resulting in a higher-dimensional space, for them to be enhanced for better use in NLP tasks.

It is also important to note that this dissertation is inserted within the context of the *"Geologia Digital: Busca digital de dados geocientíficos heterogêneos"* (Digital Geology: Digital search of heterogeneous geoscientific data) project. This project, a result of Petrobras' partnership with the *Universidade Federal do Rio Grande do Sul* (UFRGS) and the *Pontifícia Universidade Católica do Rio Grande do Sul* (PUCRS), has as a central objective the research and development of Information Retrieval technology for internal usage within Petrobras' large and heterogeneous databases. This is where Word Embeddings come in, as several studies posit that they are more fit for use in industry than more computationally intensive Contextual Embeddings [38, 4], and they can be used to expand search terms through semantic similarity and relatedness. This approach to the problem also inspired the study into the possibility of the enrichment of Word Embeddings with visual data previously mentioned, which might enable the development of tools that take advantage of the images being processed by the Visual Data sub-teams within the *Geologia Digital* project for use with textual data.

The goal of this work is to study the possibility of usage of visual data to enrich textual data within word embeddings for use within NLP tasks in the Portuguese language. The main hypothesis presented herein is that fusing textual information with visual information will enhance results for traditionally text-only tasks. To test it, experiments in four NLP tasks were performed. Of these four tasks, two had test corpora for both a generic news domain and a specific geosciences domain, while the other two only had test corpora for a generic news domain. This is because of the nascent nature of data digitalization and organization within Petrobras, which is just beginning their efforts into creating proper test corpora for their domains of interest.

This dissertation contributes to the literature and Petrobras' aims by developing an intrinsic NLP test corpus for the geociensces domain and enhancing an already existing extrinsic NLP test corpus for the same domain. It further tests several generic news domain and specific geosciences domain word embedding models on appropriate domain test corpora, both with and without visual information enrichment, thus confirming that visual enrichment of word embeddings is a viable strategy even for text-only tasks.

The remaining chapters of this dissertation are arranged in the following manner: Chapter 2 presents a systematic review into multimodality within semantic embeddings; Chapter 3 presents the tools, resources and methods used to develop the multimodal embeddings studied in this work; Chapter 4 presents the testing methodology for the embedding models; Chapter 5 presents the results achieved for the tests; and Chapter 6 presents the conclusions reached with this study, discusses the results, and deliberates on the possibilities for future work.

2. SYSTEMATIC REVIEW OF MULTIMODAL EMBEDDINGS

A systematic review was performed in order to appropriately ground this research in the state-of-the-art for multimodal semantic embeddings, and take full advantage of already developed tools and methodologies. This review focused on the creation of multimodal embeddings with a bias toward the textual and visual modalities.

This chapter is structured around the literature review, and presents its aspects in the following manner: the objective, which guided the construction of the review; the search plan, which guided the search for existing studies, and its results; and the review questions, the major focuses of the review, and their answers.

2.1 The Objective and Research Questions

The objective of this literature revision is to systematically review and analyse the current state of the use of multimodality in the creation of semantically significant embeddings such as Word2Vec [31], Flair Embeddings [2], ELMO Embeddings [37] and BERT [14]. Given that the planned application of the multimodality for this work will be in textual-visual fusion, other modalities are understood to be less important to the review process.

This objective is meant to explicit a focus on the textual modality, as the preliminary objective of this work is the creation of semantic embeddings. It is to be noted that the review has a bias toward the visual modality, but does not completely discard other possible modalities, such as user data and audio, that may prove to increase semantic significance.

With the above objective as a guiding directive, five research questions were asked:

1. To what tasks are these embeddings mainly applied?
2. How were the embeddings constructed?
3. How were they evaluated?
4. What resources were used in the creation of the embeddings?
5. To what extent has multimodality been implemented in the creation of semantic embeddings? Has any implementation been successful?

These questions then guided the creation of the Search Plan, described in Section 2.2. Additionally, the answers to the research questions are presented in Section 2.3.

2.2 The Search Plan

The search plan includes the delineation of the search terms, of the databases that will be searched, and of the eligibility criteria. Additionally, the search terms must be translated into search strings compatible with each database's advanced search function.

Firstly, the search terms were created. These sprang from two main search terms: *multimodal* and *embeddings*. From these, synonyms and certain suitably related terms were added to the search terms, as presented below. Note that both the original terms and the new terms were used during the searches.

- Terms related to "Multimodal":
  - Modality;
- Terms related to "Embeddings":
  - Distributional Semantics;
  - Language Model;
  - Word Space;
  - Semantic Vector Space;

Secondly, the databases to be searched were defined, and their respective search strings were developed. The chosen databases were ACM Digital Library [1], IEEE Digital Library [2], and SCOPUS [3]. These were chosen as they are the biggest Computer Science related repositories available through PUCRS's bought licenses, and have the most robust search functions. Table 2.1 presents the search strings used for each database.

Lastly, the eligibility criteria were developed. These are composed of two sub-classes of criteria: the *Inclusion Criteria*, which must be met in order for a work to be included in the review; and the *Exclusion Criteria*, none of which can be met if a work is to be included in the review. The criteria are presented below.

[1] https://dl.acm.org/
[2] https://ieeexplore.ieee.org
[3] https://www.scopus.com/

Database	Search String
ACM Digital Library	recordAbstract: (+(multimodal modality) +("distributional semantics" embedding "language model" "word space" "semantic vector space"))
SCOPUS	ABS((multimodal OR modality) AND (distributional AND semantics OR embedding OR language AND model OR word AND space OR semantic AND vector AND space))
IEEE Digital Library	(("Abstract":"multimodal" OR "Abstract":"modality") AND ("Abstract":"’distributional semantics’" OR "Abstract":"embedding" OR "Abstract":"’language model’" OR "Abstract":"’word space’" OR "Abstract":"’semantic vector space’"))

Table 2.1 – Search Strings used for database searches.

- Inclusion Criteria:
  - Publication was an academic, peer-reviewed study;
  - Publication was a study pertaining to the field of Natural Language Processing;
  - Publication was a study making use of multimodal semantically significant embeddings;
- Exclusion Criteria:
  - Publication in a language other than English or Portuguese;
  - Publication’s full text neither made freely available by the author nor accessible via the licenses at PUCRS’s disposal;
  - Publication published before 2015;

These criteria were chosen to ensure that the works reviewed are recent, on topic and present reliable information, and ensure that the works are accessible to the reviewer. This final step of the search plan ended with the manual screening of the abstract of each of the works collected during the database search according to these criteria, and resulted in the review corpus.

The search plan resulted in the recovery of 250 works through the automated database search, and reduced to 111 works after the manual screening. These were then fully read and analyzed by the author of this dissertation.

2.3 Answering the Research Questions

This section presents the answers to the research questions as found by reading the review corpus collected during the systematic review. It is believed that the corpus is complete enough to provide an adequate picture of the current state-of-the-art in this research field, as well as provide a foundation from which this work will achieve its final goal.

To what tasks are these embeddings mainly applied?

Multimodal semantic embeddings see extensive use in video related tasks, such as video captioning [21], video understanding and video event recognition [19, 22], video hyperlinking [25], and video recommendation [19]. This is not unexpected, as video is an inherently multimodal medium, usually combining the visual and audio modalities. It is also common to consider the textual modality in video, through either video descriptions or speech transcription. It is thus common to see works proposing ways to better embed these data modalities and attempt to make their interaction and fusion more effective.

The use of multimodal embeddings for recommendation extended beyond video. Works used it for fashion recommendation [24, 23], product recommendation [32], and music recommendation [34].

These embeddings have also been used to provide information to machine learning models performing prediction tasks. The use of multimodal embeddings for such tasks, which are usually treated unimodally, has resulted in better results for certain domains. Some of these are social media popularity prediction [9, 46], and data classification prediction [29].

Semantic embeddings have also seen recent use in network embeddings. This task consists of the learning of low-dimensional vector representations for network nodes while preserving their structural information, and is mainly implemented so that off-the-shelf machine learning models become easily able to use this network information in downstream tasks [27]. A few approaches have begun to use the content information of content-rich network nodes to inform the embedding process, alongside a node networks' structural information. Structural and content information form the two main information modalities of these works [27, 47].

Finally, multimodal embeddings are also used in multimedia information retrieval. Several works focus on cross-modal retrieval [20, 17, 45]. Information retrieval is considered one of the more important multimodal tasks, given the glut of multimedia data available on the internet from which specific data must be retrieved.

How were the embeddings constructed?

The literature reveals two main ways in which multimodal embeddings are constructed: individually and simultaneously. That is, either learning is performed individually (an embedding is learned for each modality, and then these are fused) [11], or simultaneously (all modalities are learned at the same time in the same space). Henceforth, the former method will be referred to as *Post-Learning Fusion*, while the latter method will be referred to as *Simultaneous Learning*.

Post-learning fusion is divided into two further methods: early fusion and late fusion. Early fusion is performed at the representation level, and three methods of early fusion were found in the literature: feature concatenation, auto-encoder fusion, and cross-modal mapping. Feature concatenation is performed through the concatenation of all single modality fusion embedding vector pairs (that is, a textual feature vector representing a concept will be concatenated with a visual feature vector representing that same context) into a single, longer, multimodal feature vector [22, 19]. Auto-encoder fusion is performed through the use of auto-encoders fed with pre-trained single modality embeddings, thus generating a single feature vector which can then be extracted from the auto-encoder's last hidden layer [43]. Cross-modal mapping is performed through the learning of a certain amount of pre-mapped multimodal inputs and predicting those that do not have examples in both modalities [11]. Late fusion is performed at the level of prediction scores, and it is performed through an averaging of single modality predictions [22].

Lazaridous et al. (2015) [28] introduced the first instance found during the review of simultaneous learning semantic embedding model, based on Mikolov et al.'s (2013) [31] skip-gram architecture. They extended Mikolov et al.'s (2013) models to present relevant visual feature vectors alongside textual data during training for a subset of target words. This model has been shown to further propagate visual information to representations of words which were not trained with visual features.

How were they evaluated?

Most of the literature consisted of using multimodality to improve the performance of downstream tasks, such as those presented in answering the first research question, *To what tasks are these embeddings mainly applied?*. As such, the evaluation of the embeddings was extrinsic. That is, the evaluation metric was whether or not its addition to the systems performing the downstream task affected their performance.

Lazaridou et al. (2015) [28] were the only ones to perform intrinsic tests, using general semantic benchmarks such as concept relatedness (also known as semantic relatedness) [6] or semantic similarity. These are usually used to evaluate word embeddings, but multimodal embeddings were shown by Lazaridou et al. (2015) to outperform word embeddings on these tasks.

What resources were used in the creation of the embeddings?

Traditional textual feature vector building architectures, such as Word2Vec [31] and FastText [18], were often used when extracting textual features from text corpora [45, 19, 9]. Though these features were usually created with direct use of the aforementioned word embedding architectures, some authors chose to develop their own word embedding architectures, based on the traditional ones.

Visual features were extracted differently depending on whether the visual knowledge was presented in the form of images or video. Papers working with images used a variety of neural networks which learn features from annotated images, such as Convolutional Neural Networks [17, 33] and auto-encoders [45]. Papers working with videos often cut the video into parts, and used specialized neural networks to better capture action flow [8, 22], which may not be taken into consideration when dealing with images. Videos themselves are also inherently multimodal, and the audio modality was usually added as input information to the neural network being used for learning in each work.

Also, when working with social networks, user data was often used as additional information when creating the embedding. User history was often used to cluster items that possess similar user bases [19, 46].

To what extent has multimodality been implemented in the creation of semantic embeddings? Has any been successful?

The use of multimodal semantic embeddings has become more widespread in the last few years. This can be clearly seen in Figure 2.1, which shows a graph of all works found using the same search strings as the systematic review in the ACM Digital Library, the IEEE Digital Library and SCOPUS by year, from 2000 until 2019.

Several of the reviewed papers achieved state-of-the-art results for their respective tasks using multimodal embeddings [21, 22, 20]. Several of these even claimed to be the first to employ multimodality in their respective tasks [24].

This increase in interest is often attributed to the recent, rapid advancement of neural network technologies [21], and the rapid growth of multimedia data available through the internet [20, 17].

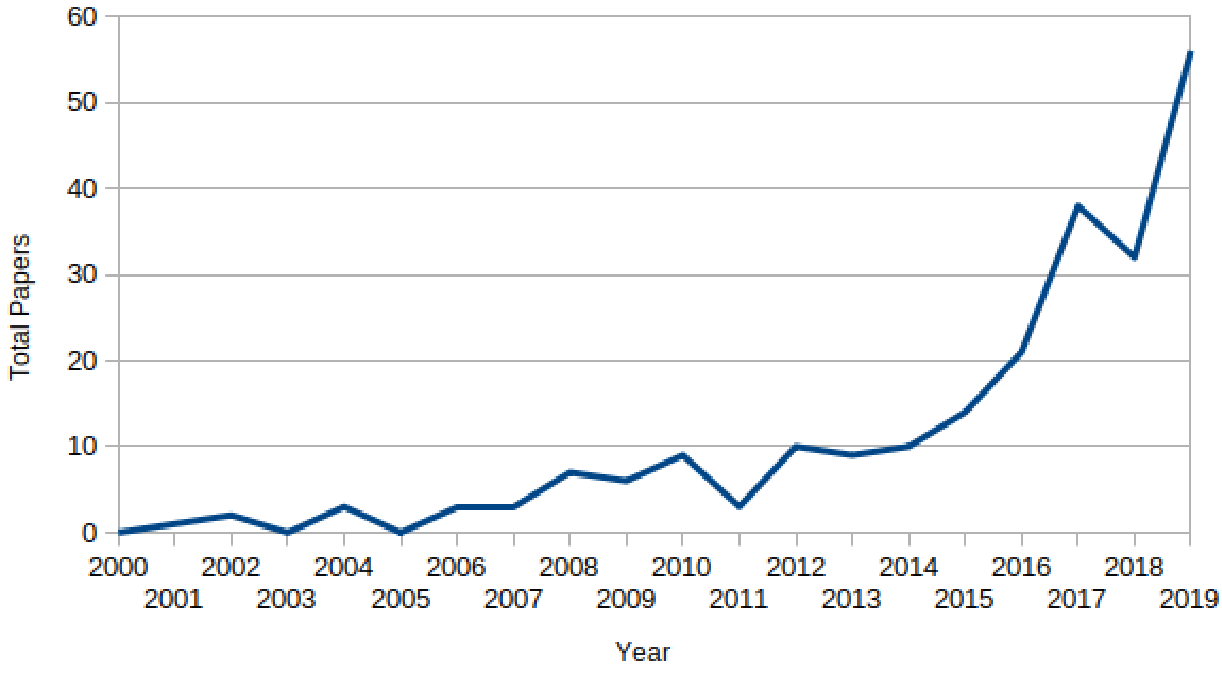

Figure 2.1 – Papers Related to Multimodal Semantic Embeddings by Year

2.4 Influence of the Research Questions on the Direction of the Dissertation

The three research questions that most influenced the direction of the dissertation were the following: "*How were the embeddings constructed?*"; "*How were the they evaluated?*"; and "*What resources were used in the creation of the embeddings?*". The other two questions served another purpose, having helped to establish the rising importance and the many uses of multimodal embedding technology, which helped to establish the reasoning behind the pursuit of this topic in this dissertation.

"*How were the embeddings constructed?*" was the question that helped inform the possible fusion architectures that would be used for the dissertation. Specifically, two post-learning architectures were chosen: Concatenation and Auto-encoding, examples of which were used in the work of Guo et al. (2019) [19] and Silberer et al. (2014) [43], respectively. For the dissertation, both of these architectures were reinforced with Collel et al.'s (2017) Imagined Embedding cross-modal mapping neural network [11] in order to broaden the limited visual embedding vocabulary at hand.

How were the they evaluated?" was the question that eventually led to the idea of using the multimodal embeddings for common NLP tasks, in order to see if it was possible to improve the effectiveness of Word Embeddings without having to rely on more computation intensive architectures, such as Contextual Embeddings. The answer to this question drew mostly from the work of Lazaridou et al. (2015) [28], who evaluated multimodal embeddings on the word relatedness and sentence similarity tasks.

"*What resources were used in the creation of the embeddings?*" was the question that enabled the beginning of the resource gathering process that led to the creation of new resources or finding of already developed resources for the Portuguese language that would enable the successful training and testing of the proposed architectures. It also informed what could and could not be done within the time-frame of the dissertation and the *Geologia Digital* project.

The last two questions establish that multimodality has been a rising interest over the past few decades, as the internet matures, and additionally expound on the kinds of tasks these architectures might be used for, beyond the common NLP tasks explored within this dissertation. This sets a clear path forward for future work on this area.

These and other papers representative of the literature found during the systematic review are presented in a Tables 2.2 and 2.3. These briefly answer the research questions for each listed paper.

Reference	Application	Methodology	Evaluation	Resources
Guo et al. (2016) [21]	Video Captioning	Word Embeddings are fused with visual features extracted from CNN using LSTM	Extrinsic, using BLUE, METEOR, and CIDEr on Youtube2Text video description corpus	Built their own, except for evaluation corpora
Guo et al. (2019) [19]	Video Recommendation	DeepWalk user data, Word Embeddings for user click history and text, CNNs for Video and Audio features fused in a Batch Norm layer	Extrinsic, using the ICME 2019 Short Video Understanding and Reccomendation Challenge dataset	Word2Vec, FastText, ResNet, DeepWalk
Wang et al. (2019) [45]	Cross-modal IR	Word Embeddings and CNN visual features fused using Batch-based Triplet Loss	Extrinsic, Cross-modal IR with Flickr8k, Flickr30k, and Microsoft-COCO	Word2Vec, ResNet
Collell et al. (2017) [11]	Multimodal embedding generation	Word Embeddings and visual features extracted from CNNs learned, mapped where both exist for same concept, predicted from available data otherwise	Intrinsic, semantic relatedness with the MEN and Wordsim353, semantic similarity through Sem-Sim, Simlex999, Wordsim353 and SimVerb-3500, visual similarity through VisSim	ImageNet, GloVe, MatConvNet
Lazaridou et al. (2015) [28]	Multimodal embedding generation	Visual features are introduced during training of skip-gram word embedding model.	Intrinsic, semantic relatedness with the MEN, semantic similarity through Sem-Sim, Simlex999, and visual similarity through VisSim	Built their own, except for evaluation corpora

Table 2.2 – Details on some of the more relevant papers found during the review.

Reference	Application	Methodology	Evaluation	Resources
Zhang et al. (2018) [46]	Popularity Prediction	Attention mechanisms are used to learn attended embeddings for both visual and textual modalities, then another attention mechanism is used to judge the importance of each for individual users	Extrinsic, using a social image dataset they built from images present in the Flickr datatset	VGGNet
Habibian et al. (2017) [22]	Event Recognition	Features for audio, visual, motion and textual information are separately trained and are then fused using their Video2vec fusion technique	Extrinsic, using zero-shot or few-shot event recognition in the TRECVID Multimedia Event Detection corpus, and the Columbia Consumer Video collection	VideoStory46k, Google Inception, ImageNet
Han et al. (2017) [24]	Fashion Recommendation	Textual and visual features are associated in a joint representation space	Extrinsic, using several fill-in-the-blank and prediction datasets	Built their own
Oramas et al. (2017) [34]	Music Recommendation	Text and audio features were extracted, then combined via late fusion	Extrinsic, using recommendation tests based on the Echo Next Taste Profile Subset and the Million Song Dataset	Word2Vec, librosa

Table 2.3 – Details on some of the more relevant papers found during the review, continued.

3. DEVELOPING THE MULTIMODAL EMBEDDINGS

Two kinds of resources are needed to create post-learning fusion multimodal embedding models: unimodal embeddings in the desired modalities and a fusion architecture. In this work, the unimodal embeddings will encompass the textual and visual modes, both of which will then be fused using different architectures to form several multimodal embedding models. This work will develop models using generic corpora (corpora extracted from news sources, fiction literature, and various websites across the internet), and geosciences corpora (geosciences related theses, journals and bulletins). These extra, domain specific models based on geosciences corpora were created during the course of the *Geologia Digital* project, and will be used to test the efficacy of the presented multimodal embedding architectures on domain-specific corpora, as opposed to generic domain corpora.

The greatest challenge to be overcome, regardless of corpora domain, is the disparity between available textual and visual information. The abundance of text knowledge often overshadows visual knowledge. Some of the works presented in Chapter 2 postulate solutions to this problem, and these architectures will be used, for the first time, to create multimodal embeddings for the Portuguese language.

This chapter will explore both the process of acquiring and developing unimodal embeddings (textual and visual) in the generic and geosciences domains, and the architectures used to fuse the unimodal embeddings into multimodal embedding models.

3.1 Unimodal Embeddings

The planned experiments will require textual and visual embeddings in both the generic and the geosciences domains. For textual embeddings, four corpora were acquired: two generic and two focused on the geosciences domain. For visual embeddings, two corpora were acquired: one generic and one in the geosciences domain, though unfortunately the geociences corpus proved to not be robust enough for use in the creation of multimodal embeddings.

3.1.1 Textual Embeddings

All textual embeddings used in this work are word embeddings based on either the Word2Vec [31] or fastText [18] architectures. The reason for this choice was the need for multimodal solutions for these specific architectures in the *Geologia Digital* project, as they would be best suited for deployment within existing architecture in Petrobras' systems.

Their choice of architecture was rooted in the fact that contextualized models, such as BERT or ELMO, significantly increase computational requirements for both training and inference when compared to non-contextual models, such as Word2Vec and fastText [38, 4]. This makes contextual embeddings less appealing in industrial scenarios, since, as per Polgnano et al (2020) [38], it is yet unclear whether the accuracy increase delivered by contextual embedding is worth the performance issues associated with them.

The generic embeddings used for this work are NILC's word embeddings[1] [26] and BBP corpus word embeddings[2] [40]. Three versions of NILC's embeddings were used: the 100 feature word2vec version and the 100 and 300 feature fastText versions. These three were deemed to be adequate for studying the effect of different parameters when adding multimodality to textual models. Only the 300 feature fastText version of BBP was used, as it was the only one readily available for download. This final BBP model was chosen as a means to study how different text embedding training corpora within the same domain affected multimodal fusion.

Two new geosciences domain embeddings were developed during the course of this study as part of a collaboration with experts from Petrobras' CENPES research nucleus through the *Geologia Digital* project: PetroVec and PetroVec-Hybrid[3]. These models were thoroughly tested using both intrinsic and extrinsic tasks, and the results were compiled into an article published in the Computers in Industry journal[4] [16]. These are the current state-of-the-art models for the Portuguese language in the Geosciences domain.

Table 3.1 has details for each of the generic and geosciences embeddings presented in this section.

3.1.2 Visual Embeddings

The visual embeddings were somewhat harder to acquire. No pre-trained Portuguese term paired visual embeddings were found, nor were there any image-term pair datasets like ImageNet [13] available. This meant that either translation or development of a new dataset would be required for the acquisition of visual embeddings, both generic and geociences domain specific.

The generic visual embedding, henceforth referred to as ImageNet embedding, is derived from Collell et al.'s (2017) [11] work, as they made their original visual embeddings created using ImageNet freely available[5]. The individual embeddings were paired with English language terms from the English language WordNet, however, and so needed to be

[1] http://www.nilc.icmc.usp.br/embeddings
[2] https://github.com/jneto04/ner-pt
[3] https://github.com/Petroles/Petrovec
[4] https://www.sciencedirect.com/science/article/abs/pii/S0166361520305819
[5] https://liir.cs.kuleuven.be/software_pages/imagined_representation_aaai.php

Table 3.1 – Corpora and token totals for each of the text corpora used for training text embedding models.

Corpus	Sources	Vocabulary	Token Number
NILC	LX-Corpus, Wikipedia, GoogleNews, SuIMDB-PT, G1, PLN-Br, Public domain literature Lacio-web, e-books, Mundo Estranho, CHC, FAPESP, Digitalized Textbooks, Folhinha, NILC subcorpus, Para Seu Filho Ler, SARESP	929,605	1,395,926,282
BBP	BlogSet-BR, brWaC, Portuguese Wikipedia	553,637	4,900,352,063
Petrovec	Petrobras' Bulletin of Geosciences and Petroleum Production, ANP bulletins, ANP technical reports, Theses and dissertations on the Oil and Gas domain, Proceedings of the Rio Oil and Gas Conference	161,842	85,725,834
Petrovec-Hybrid	All Petrovec corpora, All publicly available NILC texts (Roughly a quarter of the collection)	440,692	451,021,003

translated before use with Portuguese language textual embeddings. In order to translate the English terms, OpenWordNet-PT [35], an open Brazilian WordNet available online[6], was used. Since the codes used to refer to each term in both WordNets were the same, and Collell et al. (2017) also shared the WordNet code for each term, about 5000 of the term-visual embedding pairs were successfully translated into Brazilian Portuguese unigrams. This resulted in what we believe to be the first visual embedding dataset paired with Brazilian Portuguese terms, made available in this project's GitHub page [7].

Domain specific embeddings for the geoscience domain had to be developed from the ground up. Firstly, all unigram terms from the Petroleum Abstracts thesaurus were extracted and used in a mass image scraping effort through Google Images. These terms include names for rocks, tools and physical structures, both natural and man-made. The next step of this effort involved finding suitable image search links. In Google Images, every search link is unique, and so can be reused to find the same images as the first time it was

[6]http://wn.mybluemix.net/

[7]https://github.com/bsconsoli/Enriching-Portuguese-Word-Embeddings-with-Visual-Information

found through the usual methods of online search. A manual search was thus performed to find the most representative image collections within search links, with two to four links being selected per term. The first hundred images of each link were then scraped automatically, using respectful scraping etiquette. Google Images sorts by relevance, and those too far removed from the beginning of the list tend to off-topic subjects. Repeat images, obviously off-topic images and non-photographic images (eg. drawings or 3D computer generated images) were then removed during a manual sweep.

This corpus had little oversight from domain experts, despite association with the *Geologia Digital* project. The breadth of expertise necessary to evaluate every term was simply too manpower intensive, and only some classes of image, such as those pertaining to some rocks and small tools, were able to be checked by experts, for a total of not even 100 of the over 1000 terms in the corpus.

Possibly because of this lackluster curation, or perhaps because the around 100 images found for each term were simply too few in number, this corpus was ultimately unable to be used in training a good image classification neural network. The attempted training resulted in poorly differentiated features unable to accurately portray the terms in such a way that they might be useful in multimodal fusion. This, alongside preliminary word relatedness results showing that fusions using the ImageNet embedding presented above improved both generic and geosciences textual embeddings to a similar degree, led to a decision that the use of this corpus for multimodal fusion was best left for future work, once the corpus had been properly curated and extended by a discussed follow-up to the *Geologia Digital* project.

3.1.3 Dealing with the Information Gap

The great imbalance between visual embeddings and text embeddings becomes clear when comparing the roughly 5000 terms of the ImageNet embedding to the textual embedding vocabularies shown in Table 3.1. In order to ameliorate this problem, the "imagined embeddings" architecture described in Collell et al. (2017) [11] was used. As exemplified in Figure 3.1, textual embedding-visual embedding pairs are created for the terms present in the visual embedding vocabulary, w, and used to train a feed-forward neural network. It does this by inputting the textual embedding $\vec{l_x}$ into the NN, and expecting the visual embedding $\vec{v_x}$ as an output, where the w_x is the term being learned. Once this textual-visual translation, f, is learned by the network, it can be extrapolated into terms without visual counterparts, creating "Imagined" visual embeddings for the entire vocabulary represented by the textual embedding that was translated.

In Collel et al.'s work, they developed three imagined models were trained for each available word embedding, each trained to a different epoch (25, 50, 100). All other parameters were kept the same between all training instances, as Collell et al. (2017) revealed that

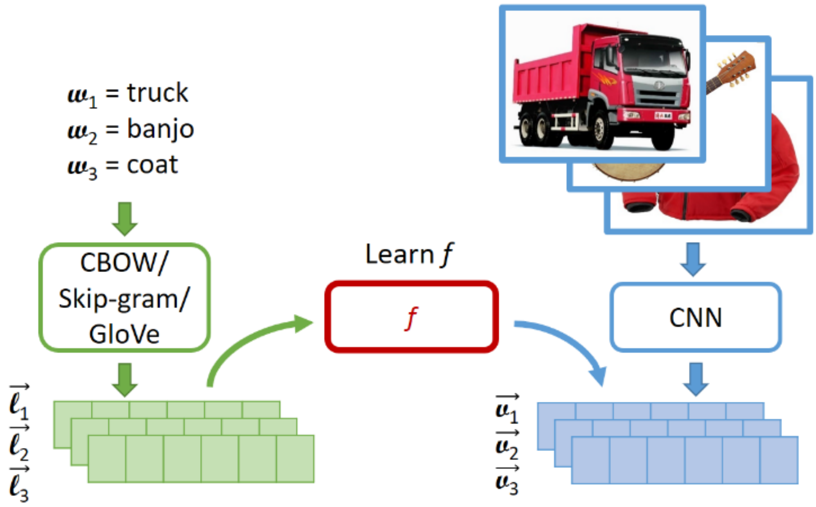

Figure 3.1 – Example of the architecture used by Collell et al. (2017). The imagined representations are the outputs of a text-to-vision mapping, *f*. Image created by Collell et al. (2017) [11]

they did not significantly affect the final prediction. All parameters are presented in Table 3.2.

Table 3.2 – Imagined Embedding neural network parameters

Parameter	Value
Dropout	0.25
Learning Rate	0.1
Optimizer	SGD
Loss	MSE
Hidden Layers	1, 200 nodes
Activation Function	TanH

Notably, the work discusses that while these imagined embeddings are valuable aggregates to common embeddings, substituting the textual embeddings completely with these "imagined embeddings" yields worse results. Additionally, in a follow-up paper, Collell et al. (2018) [10] highlighted several problems with this architecture, such as the fact that they do not fully mimic the behaviour of proper visual embeddings to the desired degree. It remains, however, that when combined with the original textual embeddings, these "imagined embeddings" do positively affect results in intrinsic tasks such as Word Relatedness.

3.2 Multimodal Fusion Techniques

Of the many fusion techniques presented in Chapter 2, the ones chosen for this project were two examples of the early fusion architecture. Early fusion techniques seek to create new embeddings to represent all fused modes in a single vector before beginning the process of using them in any downstream task. This kind of fusion was chosen because it is the most flexible, with models developed using it able to be simply plugged into many already existing solutions for downstream tasks without requiring modifications to the architecture. It was deemed that this would facilitate a wider array of testing while not having been shown to be definitively superior or inferior to other fusion strategies in the literature.

In order to perform this kind of fusion, it is helpful to ensure that all fused embeddings are in the same scale, so that none can overly influence the result simple because it is presented in a larger scale than another. To do this, a mathematical process called Standardization was performed on the embeddings, making it so all features were scaled according to a standard deviation of 1 and had a mean of 0. Another version of these embeddings was created where, after standardization, they were also normalized, so that all values fit between -1 and 1. This second version was created mostly to test the machine learning and whether it would learn better with unbounded or bounded feature values.

The two early fusion techniques used in this work are concatenation and auto-encoding, explained in detail below.

3.2.1 Concatenation Fusion

Concatenation fusion is a rather simple process: you concatenate one mode's embeddings to the end of another mode's embeddings. Though simple, it effectively packages all necessary data into a single vector space by expanding the dimensionality of said space.

This fusion technique's greatest weakness, the fact that should one embedding in a certain mode not have a pair in another (as often happens with text-image multimodality, eg. you have textual embeddings but not visual) you cannot create the multimodal embeddings, is completely solved by the imagined embeddings explained in Section 3.1.

As such, the development of this embedding required the prediction of a imagined visual embedding for each word in the vocabulary, which was then concatenated with its originating word embedding. This resulted in multimodal embeddings with larger feature pools with which to draw from. Figure 3.2 presents the architecture of the concatenated fusion used for every word embedding in this work.

3.2.2 Auto-encoding Fusion

Auto-encoding fusion is performed by a Neural Network trained to predict an output by using the output itself as an input. Once this is done, one of the hidden layers of this network with less features than the original input is extracted to serve as an embedded

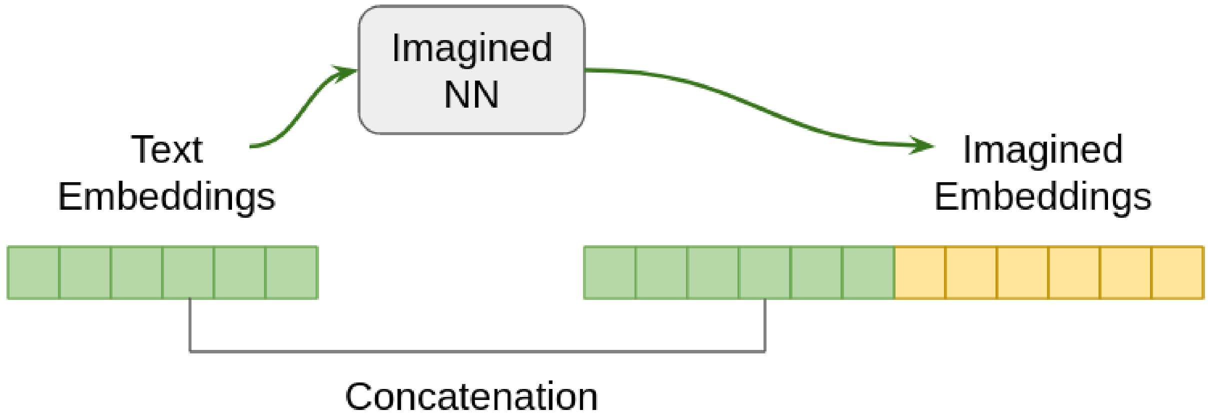

Figure 3.2 – Simplified concatenation fusion architecture.

version of the input. This serves to both shorten the final embedding, and to fuse several embeddings together. This fusion will, in theory, keep the most important features and fuse less important features together to hopefully make them more impactful.

This architecture has been used to lessen the impact of the gap between textual and visual information in the literature [43]. In this instance, whenever there was no visual pair for the textual embedding, a zeroed vector was appended to the textual embedding for the purposes of auto-encoding. The architecture presented below is a bit different, as it offers a new possibility: using imagined embeddings to fill the knowledge gap and offer complete feature vectors for auto-encoding.

As such, imagined visual embeddings were predicted from each embedding in the each model's vocabulary, and paired with its originating embedding. These embeddings were then passed through an auto-encoding neural network, and the resulting Auto-encoded vectors were used as the final multimodal embeddings. Figure 3.3 presents the architecture of the Auto-encoded fusion used for every multimodal word embedding in this work, while Table 3.3 presents the parameters for the auto-encoding neural network.

Table 3.3 – Auto-encoding Embedding neural network parameters.

Parameter	Value
Learning Rate	0.001
Optimizer	Adam
Loss	MSE
Hidden Layers	4, explained in text
Activation Function	ReLU in between layers TanH as output

The hidden layers are divided into two encoding layers and two decoding layers. The first encoding layer has the initial input node size of the concatenated textual-visual feature vector and an output node size of the feature vector of the textual model plus half the

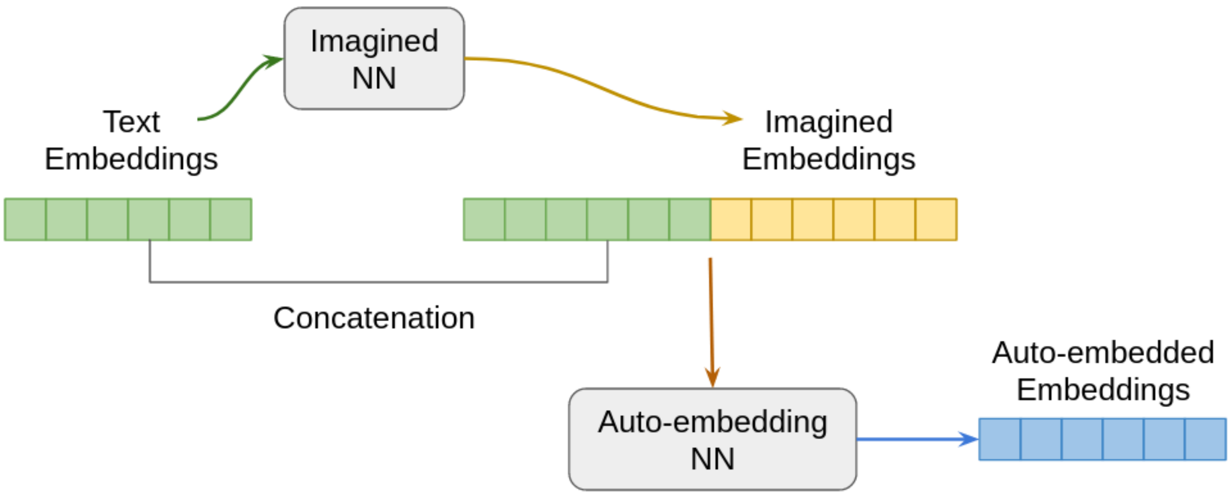

Figure 3.3 – Simplified auto-encoding fusion architecture.

feature vector of the visual model. The second layer has the input node size of the previous output, and an output the size of the feature vector of the textual embedding. The output of this second layer is extracted and used as the Auto-encoded textual-visual embedding. The decoder is used only during training, and its two hidden layers are the same as the encoder's, but in reverse order.

3.3 Multimodal Embeddings

Several different textual-visual multimodal embeddings were created using the unimodal embeddings and multimodal fusion techniques explained above. The model combinations are presented in Table 3.4.

Note that the act of training to different epochs was simply due to a lack of time and computational resources that would be required to train the best model for each individual task. As such, the best performing model out of each group can be taken to best represent the capabilities of the multimodal embedding fusion in question.

Table 3.4 – Each text embedding model was used to train 14 multimodal models. To reiterate, the textual models are: the BBP model, three NILC models, and the two PetroVec models. This makes for a total of 84 multimodal models trained in total.

Model	Fusion Architecture	Scaling Algorithm	Epochs Trained
All	Concatenation	Normalized	25
			50
			100
		Standardized	25
			50
			100
			150
	Auto-encoding	Normalized	25
			50
			100
		Standardized	25
			50
			100
			150

4. QUALITY EVALUATION STRATEGY FOR THE MULTIMODAL EMBEDDINGS

Each multimodal embedding underwent a number of intrinsic and extrinsic tests in order to ascertain their reliability when used in NLP tasks in the generic domain and, where possible, the geosciences domain. This Chapter will present the tests and their set-up, while the results will be discussed in the following chapter.

4.1 Intrinsic Tests

Intrinsic tests for semantic embeddings measure how closely the embeddings are able to predict human use of language. This does not mean that embeddings with the best scores in intrinsic tests will also achieve the best scores in downstream extrinsic tests, however.

4.1.1 Word Relatedness

Word Relatedness is the intrinsic task of giving a score to how closely related two terms are. These tasks are usually scored via Spearman correlation, which assigns a Real number score between -1 and 1. The closer the score to -1 if the predictions are the exact opposite of the annotation, the closer to 0 if the predictions are completely unrelated to the annotation and the closer to 1 if the predictions line up perfectly with the annotation. The more representative of human understanding of the terms an embedding is, the closer the Spearman score comes to 1.

These tests should be tailored to the domain of the models being tested, as certain words can have different meanings depending on context. Since the focus of this project is not whether certain models do better in certain domains, the models were only tested on their respective domains in order to ascertain whether the impact of adding visual embeddings would be similar in these distinct circumstances.

This task, alongside other kinds of relatedness tests, is particularly important in the context of the *Geologia Digital* project, as these embeddings will be used for search term expansion within Information Retrieval systems, and good Word Relatedness scores are essential for models that are intended to be used in such a manner.

A custom code was written for this task, and is shared across domains. It uses the Gensim python library to extract the Cosine distance between each word pair as a relatedness measurement, and compares them to their respective annotated relatedness scores using the Spearman Correlation method. The code can be accessed in the GitHub page for this project[1].

Generic Domain

The test corpus used for generic domain word relatedness testing, MEN [6], was translated from the English language to the Portuguese language with the help of DeepL Translate[2]. The machine translations were checked individually to ensure some degree of uniformity, but the corpora should be considered Silver standard nonetheless.

MEN is a set of English word pairs, 3000 in total, each assigned a relatedness judgement (which ranged from 0, not at all related, to 50, incredibly related. These judgements were collected via crowdsourcing using Amazon's Mechanical Turk platform. The words were randomly selected from a subset created by separating all those that appeared at least 700 times in a combined ukWaC/Wackypedia corpus, and at least 50 times in the open-sourced subset of the ESP game dataset. Before the final selection, word pair semantic relatedness scores were predicted by a pre-trained embedding model to ensure that a balanced range of relatedness levels was represented in the dataset. Table 4.1 presents a few examples of word pairs present in the translated MEN corpus.

Table 4.1 – Four examples of word pairs from the translated MEN corpus.

Word 1	Word 2	Relatedness
rio (river)	água (water)	49.0
répteis (reptiles)	serpente (serpent)	45.0
banda (band)	metal (metal)	27.0
recém-nascido (newborn)	construção (construction)	6.0

Geosciences Domain

The test corpus for the geosciences domain, henceforth called GeoSim, was developed as part of the *Geologia Digital* project, and was used to test the PetroVec word embeddings [16]. It was developed in collaboration with several industry experts, Geology students and a PhD in Geology. Its main focus is Oil and Gas, a sub-domain of the geosciences domain, and can be considered a Gold standard corpus.

GeoSim is composed of 1500 word pairs annotated in a Likert scale from 1 to 7, which were later normalized to a number between 0 and 1 for ease of use. All words were chosen from those present in the Portuguese version of the Petroleum Abstracts Exploration and Production Thesaurus[3], provided by the Petrobras team from the *Geologia Digital*

[1] https://github.com/bsconsoli/Enriching-Portuguese-Word-Embeddings-with-Visual-Information
[2] https://www.deepl.com/translator
[3] https://www.pa.utulsa.edu/products/tulsadatabase/thesaurus

project. These word pairs were picked randomly from pools of pre-made word pairs which were themselves separated using the relationships between words present in the Petroleum Abstracts Thesaurus. This was done to ensure a good distribution between very related, somewhat related and dissimilar word pairs, similarly to how the MEN corpus was developed. Table 4.2 presents a few examples of word pairs present in the GeoSim corpus.

Table 4.2 – Four examples of word pairs from the translated GeoSim corpus.

Word 1	Word 2	Relatedness
zoologia (zoology)	insetos (insects)	0.810
controle (control)	regulamentacao (regulation)	0.714
procedimento (procedure)	programa (program)	0.524
contabilidade (accounting)	inferior (inferior)	0.190

4.1.2 Analogy Prediction

Hartmann et al. (2017) [26] published an analogy prediction test set, divided into Brazilian Portuguese and European Portuguese halves, alongside their initial publication of their NILC word embeddings. The test gives a related word pair and a single word from which it must predict a pair analogous to the first.

The code used to run these tests was made available alongside the test set itself. It can be found in the associated paper's GitHub page[4]. It measures accuracy by counting how many correct predictions were achieved by the model against the total number of predictions.

Used to intrinsically test the NILC embedding models, the test set is composed of several categories of analogies, both semantic and syntactic. The first two examples in Table 4.3 are of semantic analogies, while the latter two are of syntactic analogies.

Table 4.3 – Four examples, two semantic and two syntactic, of word pairs from the Analogy Prediction corpus, translated to English.

Analogy	Example	Prediction
capital city/nation	Berlin/Germany	Rome/?
national currency/nation	Euro/Germany	Real/?
singular/plural	apple/apples	car/?
present continuous/past simple	dancing/danced	falling/?

In general, Word Embedding models have more difficulty achieving high scores for semantic analogies, and generally do much better with syntactic analogies.

[4]https://github.com/nathanshartmann/portuguese_word_embeddings

4.2 Extrinsic Tests

Extrinsic tests measure the reliability of embeddings in helping achieve greater results in downstream tasks. A number of such tasks were chosen for this purpose, and though the list is not exhaustive, it should serve to ascertain how multimodality can be expected to affect the performance of these models.

4.2.1 Semantic Similarity in Short Sentences

Semantic similarity requires that a model give a numerical value to how semantically similar two sentences are, with the lower similarity extreme being that the sentences are completely different, and the higher similarity extreme being that the sentences are paraphrases. The ASSIN [15] sentence similarity corpus was used for this task in this work.

The code used for the tests is the same as was used by Hartmann et al. (2017) [26], available in the publication's GitHub page[5]. The architecture uses a linear regression algorithm trained on two features: the cosine simirity of the TF-IDF of each sentence and the cosine similarity between the sum of each sentence's word embeddings.

ASSIN is a Portuguese language corpus annotated for both textual inference and semantic similarity. It is composed of sentence pairs annotated with whether or not one implies the other (textual inference) and how similar they are (annotated from 1, completely different, to 5, paraphrases). Figure 4.1 presents two example pairs extracted from the ASSIN corpus.

```
<pair entailment="Entailment" id="2504" similarity="4.25">
    <t>A estreia do Brasil na Copa América está marcada para o dia 14 de junho, contra o Peru.</t>
    <h>O time estreia na Copa América contra o Peru.</h>
</pair>
<pair entailment="None" id="2505" similarity="2.25">
    <t>Funcionamento normal, lojas e praça de alimentação de 10h às 22h.</t>
    <h>Já na segunda-feira do feriado, o shopping abrirá das 9h às 19h.</h>
</pair>
```

Figure 4.1 – Example of sentence pairs from the ASSIN corpus.

The similarity scores were used for testing the multimodal models. As previously discussed, semantic similarity is particularly important for the *Geologia Digital* project.

[5]https://github.com/nathanshartmann/portuguese_word_embeddings

4.2.2 Named Entity Recognition

Named Entity Recognition (NER) requires that, given a set of classes for named entities, a model recognize and classify said entities within raw text, usually by use of tags. Word embeddings can be used to parse the text input into the model, using the feature vectors in its tagging predictions. Two annotated corpora were used to evaluate the multimodal embeddings: HAREM [39] for generic domain embeddings; and GeoCorpus 3.0 [16] for geosciences domain embeddings.

The code used for the NER tests was developed by Santos et al. (2019) [40], available in the paper's GitHub page[6]. It uses an LSTM-CRF neural network architecture to train a sequence tagger using the Flair Toolkit to perform a NER task based on the supplied training and test corpora.

The HAREM corpora are a set of corpora produced during the HAREM workshops, and include First HAREM, MiniHAREM and Second HAREM. This work used First HAREM, as the training dataset, and MiniHAREM, as the testing dataset. All HAREM corpora are annotated in the same way, and have two annotation scenarios: the selective scenario, annotated with only the three classic NER classes (Person, Location and Organization); and the complete scenario, annotated with a total of ten different classes of named entity, including those which comprise the selective scenario. Figure 4.2 demonstrates an example of the HAREM corpus.

<EM ID="556" CATEG="PESSOA" TIPO="INDIVIDUAL">Leonardo</EM> nasceu a <EM ID="557" CATEG="TEMPO" TIPO="DATA">15 de Abril de 1452</EM> , na pequena cidade de <EM ID="558" CATEG="LOCAL" TIPO="HUMANO">Vinci</EM> (...)

Figure 4.2 – A snippet of a sentence from the First HAREM, to exemplify its annotation.

GeoCorpus 3.0 is a NER corpus in the Oil and Gas domain, a sub-domain of the geosciences. More specifically, its texts are about Brazilian sedimentary basins, and it is annotated with thirty classes of named entity, though only ten were judged to have enough instances for use with machine learning architectures. As GeoCorpus does not have an established baseline within the literature, as is the case with HAREM, it was tested using 10-fold cross-validation. Figure 4.3 demonstrates an example of GeoCorpus 3.0.

Os dois andares do <EM CATEG="epoca">Lopingiano</EM> devem o seu nome a localidades chinesas nas quais os fósseis e <EM CATEG="unidadeEstratigrafica">estratos</EM> (...)

Figure 4.3 – A snippet of a sentence from GeoCorpus 3.0, to exemplify its annotation.

[6] https://github.com/jneto04/ner-pt

4.3 On the Construction of test sets

Two of the presented test sets, GeoSim and GeoCorpus 3.0, were a result of the *Geologia Digital* project. The author of this dissertation, Bernardo Consoli, led both the effort for the construction of GeoSim and the effort for the revision of GeoCorpus 3.0.

As previously mentioned, GeoSim was developed specifically to test the PetroVec set of Word Embeddings, both against each other and against generic Word Embedding models trained on News text corpora. GeoCorpus 3.0 was revised to make the corpus overall more consistent in its annotation.

A more detailed overview of the work that was performed on these two corpora are present in Appendixes A and B.

Additionally, machine-assisted translations were performed for the MEN corpus which was, as mentioned above, originally constructed for the English language. It is worthy of note that the English-Portuguese translation ran into a few unavoidable issues. The first is the fact that some English words translate into the same term in Portuguese, but have slightly different connotations in English. An example can be found with the words **football** and **soccer**, both of which translate to the Portuguese word ***futebol***, and lose meaning distinctly apparent in American English. Another issue is in words with multiple meanings, and which have different translations depending on context. This is the case of the word **crane**, which can either be a bird (translated to ***grou*** in Portuguese) or a piece of construction machinery (translated to ***guindaste*** in Portuguese), which makes the translator have to choose one of the possible translations without appropriate context, thus losing the meaning of the original English word. In the case of different possible translations, the words chosen by DeepL Translate were not changed by human translators. This means that these tests will not be perfect and will be affected by the language and culture in which they were annotated.

All of these mentioned corpora are either linked to or available for download on this dissertation's GitHub page[7].

[7] https://github.com/bsconsoli/Enriching-Portuguese-Word-Embeddings-with-Visual-Information

5. RESULTS

The results for the tests presented in Chapter 4 are discussed in this chapter. Each task will be discussed separately, and the analysis for each will be presented in the same three basic table structures.

The first table structure, the model comparison table, will present the name of the test corpus, the scaling algorithms used for each test set (Normalized or Standardized), and the specifics for each task. All tasks share a *Model* column, which gives an abbreviated name for each Word Embedding model and an *Architecture* column, which gives the type of the architecture used in the particular test, with each model being tested with three different architectures. Any other information given pertains to the scoring of the specific test, such as Spearman Correlation score, accuracy score or F-value.

The second table structure is the Architectures and Algorithms table. These are paired with their respective model comparison tables and present the number of times a particular architecture or scaling algorithm performed best for a given model. These tables are each composed of three subtables: Overall, which contains a sum of the Architecture scores found in the other two subtables as well as the score for each scaling algorithm; and the Normalized and Standardized subtables present the individual Architecture scores for each Scaling Algorithm.

Finally, the Overall table presents the sum for best performing architectures and scaling algorithms for the task in question, providing an overview for closing analysis. It is shaped like the Overall subtable of the Architecture and Algorithm tables. Both Architecture and Algorithm and Overall tables are presented because the large amount of tests for each task obfuscates important information by dent of sheer volume of data. These two tables condense relevant information into a more readable format which is easier to both analyse and reference.

The rest of the chapter is divided into the following sections: first, there is an analysis of the results for the Word Relatedness tests, which includes both a generic news test corpus and a specific geosciences test corpus; then, we have the Analogy Prediction task, which includes only a generic news test corpus, though it is divided into European Portuguese and Brazilian Portuguese; after that are the analyses of the Semantic Similarity of Sentences task, which again is composed of only the generic news corpus divided into European and Brazilian Portuguese tracks; the last task to be presented is the Named Entity Recognition task, which is composed of a generic news corpus, divided into two tracks with different categories, and a geosciences domain corpus.

5.1 Word Relatedness

Two word relatedness tests were performed using the multimodal models: MEN, for the generic news domain; and GeoSim, for the specific geosciences domain. The BBP and NILC models were tested using the MEN test set, while all PetroVec and PetroVec-Hybrid models were tested using the GeoSim test set.

5.1.1 MEN

The MEN test set was used to test BBP and NILC models, given that it is a generic domain dataset. The test set is presented in Section 4.1.1, but to reiterate, it is a collection of 3000 word pairs annotated with a relatedness score from 50 (most related) to 0 (least related). The objective of the semantic models is to score each word pair in order to rank them from most related to least related. The closer to the original ranking the model gets, the higher its Spearman Correlation, the chosen method for scoring these kinds of tests. As explained in Section 3.3, only the best results for each model will be considered during this analysis. The complete results for this test set are available in Appendix D.

Table 5.1 – The best Spearman Correlation results for each multimodal model and the results for their text-only counterparts for the MEN test set.

MEN					
Normalized			**Standardized**		
Model	**Architecture**	**Correlation**	**Model**	**Architecture**	**Correlation**
BBPFT300	Text-Only	0.607	**BBPFT300**	Text-Only	0.610
	Concatenated	0.622		**Concatenated**	**0.648**
	Auto-encoded	0.624		**Auto-encoded**	**0.649**
NILCFT100	Text-Only	0.588	**NILCFT100**	Text-Only	0.615
	Concatenated	0.626		**Concatenated**	**0.648**
	Auto-encoded	0.623		**Auto-encoded**	**0.649**
NILCW2V100	Text-Only	0.489	NILCW2V100	Text-Only	0.493
	Concatenated	0.530		Concatenated	0.518
	Auto-encoded	0.528		Auto-encoded	0.528
NILCFT300	Text-Only	0.567	NILCFT300	Text-Only	0.570
	Concatenated	0.595		Concatenated	0.586
	Auto-encoded	0.601		Auto-encoded	0.597

As can be seen in Table 5.1, the best results were achieved by both Concatenated and Auto-encoded versions of the Standardized BBP and NILC 100-dimensional fastText architectures, all of which have a Spearman Correlation of about 65 percentage points. This is a 3.5 percentage point increase from the best text-only model, the NILC 100-dimensional fastText model.

Furthermore, Table 5.2 presents the architectures and algorithms in terms of how many times each had the best performance when used in the tested models. This table shows that the best overall architecture is the Auto-encoded architecture, which only performs worse than the Concatenated architecture twice. It should be said, however, that the better performance is measured in fractions of percentage points, and only once (in the NILCW2V100 architecture) does the Auto-encoded architecture perform better by 1 or more percentage points. The two times that the Concatenated architecture performed better, it was similarly by fractions of a percentage point. This means that both architectures can be expected to perform similarly, with a slight advantage to the Auto-encoded architecture, within the realm of term relatedness.

Table 5.2 also shows that Standardization is the better scaling algorithm when it comes to this test. On average, Standardized models perform 1 percentage point better than Normalized models, with the largest performance improvement being 2.6 percentage points in favor of the Standardized model. Notably, while this is not presented in these tables, Normalization was also shown to negatively impacts the performance of Text-Only models, when compared to their non-scaled counterparts, while Standardization did not noticeably impact performance in Text-Only models.

Table 5.2 – The "*No. of best results*" column represents the number of times each architecture and scaling algorithm had the best results in a model. The Overall subtable presents a conglomeration of all results, while the Normalized and Standardized subtables present separated results for their respective scaling algorithms.

MEN - Overall			
Architecture	**No. of Best Results**	**Scaling Algorithm**	**No. of Best Results**
Text-Only	0	Normalized	3
Concatenated	2	Standardized	9
Auto-encoded	6	-	-
MEN - Normalized		**MEN - Standardized**	
Architecture	**No. of Best Results**	**Scaling Algorithm**	**No. of Best Results**
Text-Only	0	Text-Only	0
Concatenated	2	Concatenated	0
Auto-encoded	2	Auto-encoded	4

5.1.2 GeoSim

The GeoSim test set was used to test PetroVec and PetroVec-Hybrid models, given that it was created specifically for the geosciences domain. The test set is presented in Section 4.1.1, but to reiterate, it is composed of 1500 word pairs annotated from 7 (most related) to 1 (least related). The objective of the test is the same as the previous two: for the model to rank the word pairs in the rankings as the human annotation. The closer to the original ranking the model gets, the higher its Spearman Correlation, the chosen method for scoring these kinds of tests. As explained in Section 3.3, only the best results for each model will be considered during this analysis. The complete results for this test set are available in Appendix E.

Table 5.3 – The best results for each multimodal model and the results for their text-only counterparts for the GeoSim test set.

GeoSim					
Normalized			**Standardized**		
Model	**Architecture**	**Correlation**	**Model**	**Architecture**	**Correlation**
PetroVecFT	Text-Only	0.607	PetroVecFT	Text-Only	0.609
	Concatenated	0.611		Concatenated	0.615
	Auto-encoded	0.621		Auto-encoded	0.642
PetroVecHybridFT	Text-Only	0.607	**PetroVecHybridFT**	Text-Only	0.619
	Concatenated	0.629		Concatenated	0.633
	Auto-encoded	0.657		**Auto-encoded**	**0.667**
PetroVecW2V	Text-Only	0.608	PetroVecW2V	Text-Only	0.611
	Concatenated	0.613		Concatenated	0.613
	Auto-encoded	0.621		Auto-encoded	0.629
PetroVecHybridW2V	Text-Only	0.643	**PetroVecHybridW2V**	Text-Only	0.648
	Concatenated	0.660		Concatenated	0.655
	Auto-encoded	**0.667**		**Auto-encoded**	**0.664**

As can be seen in Table 5.3, the best results were achieved by the Auto-encoded architectures of the PetroVecHybridW2V and PetroVecHybridFT models, achieving results 1.9 percentage points higher than the best Text-Only architecture. Both scaling algorithms achieved the same highest result, though the Auto-encoded algorithm achieved it with two models while the Normalized algorithm only achieved it with one.

Table 5.4 shows that the Auto-encoded architecture performed better with every model. They achieved, on average, results 1.7 percentage points higher than Concatenated architectures. The Standardization scaling algorithm likewise performed better than the Normalization algorithm with every model, achieving results 0.5 percentage points higher than its counterpart.

The largest difference between a multimodal model's score when compared to their text-only counterpart's was nearly 5 percentage points, in the Auto-encoded Standardized fastText version of the PetroVec-Hybrid model. Finally, multimodal Hybrid models showed more improvement when compared to their textual counterparts than non-hybrid models, with Auto-encoded models improving fastText models more so than Concatenated models, and vice-versa for Word2Vec models.

5.1.3 Word Relatedness Task Overview

Table 5.5 makes it clear that the best performing architecture for this task was the Auto-encoded architecture, and the best performing scaling algorithm was the Standardiza-

Table 5.4 – The "*No. of best results*" column represents the number of times each architecture and scaling algorithm had the best results in a model. The Overall subtable presents a conglomeration of all results, while the Normalized and Standardized subtables present separated results for their respective scaling algorithms.

GeoSim - Overall			
Architecture	**No. of Best Results**	**Scaling Algorithm**	**No. of Best Results**
Text-Only	0	Normalized	2
Concatenated	0	Standardized	10
Auto-encoded	8	-	-
GeoSim - Normalized		**GeoSim - Standardized**	
Architecture	**No. of Best Results**	**Architecture**	**No. of Best Results**
Text-Only	0	Text-Only	0
Concatenated	0	Concatenated	0
Auto-encoded	4	Auto-encoded	4

tion algorithm. This is consistent across both geosciences domain and generic domain tests. Most importantly, regardless of domain, the fusion of Imagined Visual Embeddings based on the translated ImageNet corpus with the Word Embedding models described above results in an average increase in Correlation of 2.4 percentage points, proving that multimodality can improve tasks which use word semantic relatedness as a basis.

Table 5.5 – The "*No. of best results*" column represents the number of times each architecture and scaling algorithm had the best results in a model.

Overall			
Architecture	**No. of Best Results**	**Scaling Algorithm**	**No. of Best Results**
Text-Only	0	Normalized	5
Concatenated	2	Standardized	19
Auto-encoded	14	-	-

5.2 Analogy Prediction

As presented in Section 4.1.2, the Analogy Prediction dataset used for this test focused on two kinds of analogies: Semantic and Syntactic. These are each divided into a Brazilian Portuguese set and an European Portuguese set. To reiterate, the objective of this task is to accurately predict the second word of a pair, when given an example pair and the first word of the prediction pair (eg. Example: Berlin/Germany, Prediction: Paris/?). The accuracy of the model is then measured in a percentage, from 0 (completely inaccurate) to 100 (completely accurate). As explained in Section 3.3, only the best results for each model will be considered during this analysis. The complete results for this test set are available in Appendix F.

5.2.1 Brazilian Portuguese Test Set

As can be seen in Table 5.6, the best multimodal results were present an accuracy that is either very similar or somewhat worse than that achieved by their Text-Only counterparts. Once again, Normalized Text-Only models had a tendency to severely under perform, whereas Standardized models achieved similar results to the original, non-scaled versions of the Text-Only models. There also is no disparity between results of the separate syntactic and semantic tests, that is to day, when the multimodal fusion either enhances both, diminishes both or doesn't affect either. None of the models enhanced one result while diminishing the other.

Table 5.6 – The best accuracy results for each multimodal model and the results for their text-only counterparts for the Brazilian Portuguese Analogy Prediction test set.

ANALOGY PREDICTION TEST - BRAZILIAN PORTUGUESE									
NORMALIZED					**STANDARDIZED**				
Model	**Architecture**	**Syntactic**	**Semantic**	**Total**	**Model**	**Architecture**	**Syntactic**	**Semantic**	**Total**
BBPFT300	Textual	0.445	0.064	0.256	BBPFT300	Textual	0.447	0.064	0.257
	Concatenated	0.444	0.065	0.256		Concatenated	0.441	0.063	0.254
	Auto-encoded	0.395	0.053	0.225		Auto-encoded	0.382	0.047	0.216
NILCFT100	Textual	0.487	0.282	0.384	**NILCFT100**	**Textual**	**0.510**	**0.302**	**0.406**
	Concatenated	0.481	0.280	0.380		Concatenated	0.505	0.292	0.398
	Auto-encoded	0.495	0.301	0.398		**Auto-encoded**	**0.511**	**0.311**	**0.411**
NILCW2V100	Textual	0.247	0.077	0.162	NILCW2V100	Textual	0.255	0.080	0.167
	Concatenated	0.247	0.075	0.161		Concatenated	0.254	0.081	0.167
	Auto-encoded	0.239	0.072	0.155		Auto-encoded	0.235	0.080	0.157
NILCFT300	Textual	0.330	0.154	0.242	NILCFT300	Textual	0.332	0.158	0.245
	Concatenated	0.335	0.155	0.245		Concatenated	0.331	0.157	0.244
	Auto-encoded	0.285	0.142	0.214		Auto-encoded	0.299	0.143	0.221

Table 5.7, meanwhile reflects the results from the first table. It once again shows that both the Auto-encoded and Concatenated architectures are not helpful for this task, with neither achieving results that could be considered decisively better than their text-only counterpart. It also further reinforces the fact that the Standardization algorithm is the most appropriate for use with multimodal fusion.

5.2.2 European Portuguese Test Set

Table 5.8 presents a very similar picture to that of Table 5.6. The best multimodal results once again are not much higher than the best Text-Only results, while most others show a comparably large drop in accuracy. Once again Normalization tends to decrease Text-Only results when compared to Standardization, and whether Concatenation or Auto-encoding is less disruptive depends on the base text model.

Table 5.9 shows an interesting piece of information: together, the multimodal models have slightly more best results per model than the Text-Only architecture. These increases were by fractions of percentage points, however, and while statistical significance

Table 5.7 – The "*No. of best results*" column represents the number of times each architecture and scaling algorithm had the best results in a model. The Overall subtable presents a conglomeration of all results, while the Normalized and Standardized subtables present separated results for their respective scaling algorithms.

BR - Overall			
Architecture	**No. of Best Results**	**Scaling Algorithm**	**No. of Best Results**
Text-Only	5	Normalized	3
Concatenated	1	Standardized	9
Auto-encoded	2	-	-
BR - Normalized		**BR - Standardized**	
Architecture	**No. of Best Results**	**Scaling Algorithm**	**No. of Best Results**
Text-Only	2	Text-Only	3
Concatenated	1	Concatenated	0
Auto-encoded	1	Auto-encoded	1

Table 5.8 – The best accuracy results for each multimodal model and the results for their text-only counterparts for the European Portuguese Analogy Prediction test set.

EUROPEAN PORTUGUESE									
NORMALIZED					**STANDARDIZED**				
Model	**Modality**	**Syntactic**	**Semantic**	**Total**	**Model**	**Modality**	**Syntactic**	**Semantic**	**Total**
BBPFT300	Textual	0.448	0.057	0.254	BBPFT300	Textual	0.451	0.058	0.255
	Concatenated	0.448	0.058	0.254		Concatenated	0.444	0.057	0.251
	Auto-encoded	0.399	0.049	0.225		Auto-encoded	0.387	0.042	0.216
NILCFT100	Textual	0.485	0.274	0.379	**NILCFT100**	**Textual**	**0.509**	**0.293**	**0.401**
	Concatenated	0.481	0.273	0.377		Concatenated	0.505	0.284	0.394
	Auto-encoded	0.493	0.291	0.392		**Auto-encoded**	**0.509**	**0.298**	**0.403**
NILCW2V100	Textual	0.243	0.072	0.158	NILCW2V100	Textual	0.252	0.074	0.163
	Concatenated	0.243	0.070	0.156		Concatenated	0.250	0.075	0.163
	Auto-encoded	0.235	0.067	0.151		Auto-encoded	0.231	0.074	0.152
NILCFT300	Textual	0.322	0.140	0.231	NILCFT300	Textual	0.324	0.143	0.233
	Concatenated	0.327	0.139	0.233		Concatenated	0.322	0.144	0.233
	Auto-encoded	0.282	0.128	0.205		Auto-encoded	0.292	0.123	0.208

Table 5.9 – The "*No. of best results*" column represents the number of times each architecture and scaling algorithm had the best results in a model. The Overall subtable presents a conglomeration of all results, while the Normalized and Standardized subtables present separated results for their respective scaling algorithms.

PT - Overall			
Architecture	**No. of Best Results**	**Scaling Algorithm**	**No. of Best Results**
Text-Only	3	Normalized	3
Concatenated	2	Standardized	9
Auto-encoded	2	-	-
PT - Normalized		**PT - Standardized**	
Architecture	**No. of Best Results**	**Scaling Algorithm**	**No. of Best Results**
Text-Only	1	Text-Only	2
Concatenated	1	Concatenated	1
Auto-encoded	1	Auto-encoded	1

tests couldn't be adequately performed for this task because of the limited size of the set, it is safe to say that such a meager increase is not considered relevant within the scope of this dissertation. Furthermore, the table helps cement the fact that the Standardization algorithm should be expected to outperform the Normalization algorithm.

5.2.3 Analogy Prediction Task Overview

In general, the results between the two language-specific test sets were corroborative. The multimodal fusion failed to improve upon Text-Only results more than a few fractions of a percentage point, and the Normalization scaling algorithm caused a generalized drop in accuracy, even for Text-Only models. An interesting observation to be made is that, in both test sets, the best overall model, for both multimodal and Text-Only models, was NILCFT100, the 100-dimensional fastText model trained on the NILC text corpus. It outperformed the second-best model by around 15 percentage points for both languages.

That said, the poor results presented in an overview in Table 5.10 were somewhat expected as the image data used for the visual embeddings focused mostly on objects, while the Analogy Prediction tests focused on abstracts such as parentage, countries and currency for the Semantic half, and word forms for the Syntactic half. It is promising, however, that the previously mentioned best overall model, NILCFT100, achieved the best multimodality results when compared to their Text-Only counterpart. Perhaps with further testing, it might be ascertained that the better the original text-embedding, the more effective the imagined visual embedding fusion is.

Table 5.10 – The "*No. of best results*" column represents the number of times each architecture and scaling algorithm had the best results in a model.

Analogy Prediction - Overall			
Architecture	**No. of Best Results**	**Scaling Algorithm**	**No. of Best Results**
Text-Only	8	Normalized	6
Concatenated	3	Standardized	18
Auto-encoded	4	-	-

5.3 Semantic Similarity of Sentences

The ASSIN Semantic Similarity dataset is, as mentioned in Section 4.2.1, divided into two tracks, European Portuguese and Brazilian Portuguese. The objective of the task is to predict a number between 1 (unrelated sentences) and 5 (paraphrasing sentences) to represent the similarity between two short sentences. The task was evaluated using Pearson's Correlation and Mean Standard Error (MSE), as it was during the original ASSIN task. As explained in Section 3.3, only the best results for each model will be considered during this analysis. The complete results for this test set are available in Appendix G.

5.3.1 Brazilian Portuguese Test Set

As presented in Table 5.11, this task's results only have two decimal points of precision. This is because the tests were performed with the same script used for the original ASSIN task, and it output results as they are seen in the below tables. Of interest in this first table is the fact that the best performing model for this test set (Auto-encoded BBPFT300) used the Normalization scaling algorithm.

Table 5.11 – The best results for the Brazilian Portuguese track of the ASSIN task.

Brazilian Portuguese							
NORMALIZED				**STANDARDIZED**			
Model	**Architecture**	**Pearson**	**MSE**	**Model**	**Architecture**	**Pearson**	**MSE**
BBPFT300	Text-Only	0.56	0.52	BBPFT300	Text-Only	0.56	0.52
	Concatenated	0.56	0.52		Concatenated	0.57	0.51
	Auto-encoded	**0.59**	**0.50**		Auto-encoded	0.58	0.50
NILCFT100	Text-Only	0.53	0.55	NILCFT100	Text-Only	0.53	0.54
	Concatenated	0.54	0.54		Concatenated	0.54	0.54
	Auto-encoded	0.51	0.56		Auto-encoded	0.54	0.54
NILCW2V100	Text-Only	0.45	0.60	NILCW2V100	Text-Only	0.45	0.61
	Concatenated	0.47	0.60		Concatenated	0.46	0.60
	Auto-encoded	0.46	0.60		Auto-encoded	0.47	0.60
NILCFT300	Text-Only	0.49	0.58	NILCFT300	Text-Only	0.49	0.58
	Concatenated	0.50	0.57		Concatenated	0.50	0.57
	Auto-encoded	0.50	0.57		Auto-encoded	0.52	0.55

Table 5.12 presents results similar to those in the Word Relatedness tasks, with the Auto-encoding architecture performing in general better than the others, and the Standardization algorithm achieving better results on average.

Table 5.12 – The "*No. of best results*" column represents the number of times each architecture and scaling algorithm had the best results in a model. The Overall subtable presents a conglomeration of all results, while the Normalized and Standardized subtables present separated results for their respective scaling algorithms.

BR - Overall			
Architecture	**No. of Best Results**	**Scaling Algorithm**	**No. of Best Results**
Text-Only	0	Normalized	2
Concatenated	2	Standardized	4
Auto-encoded	4	-	-
BR - Normalized		**BR - Standardized**	
Architecture	**No. of Best Results**	**Architecture**	**No. of Best Results**
Text-Only	0	Text-Only	0
Concatenated	2	Concatenated	0
Auto-encoded	1	Auto-encoded	3

5.3.2 European Portuguese Test Set

As presented in Table 5.13, the same model as before (Auto-encoded BBPFT300) yielded the best results, though this time with the Standardization scaling algorithm rather than the Normalization algorithm.

Table 5.13 – The best results for the European Portuguese track of the ASSIN task.

European Portuguese							
NORMALIZED				**STANDARDIZED**			
Model	**Architecture**	**Pearson**	**MSE**	**Model**	**Architecture**	**Pearson**	**MSE**
BBPFT300	Text-Only	0.59	0.79	BBPFT300	Text-Only	0.59	0.79
	Concatenated	0.59	0.79		Concatenated	0.60	0.78
	Auto-encoded	0.58	0.79		**Auto-encoded**	**0.60**	**0.76**
NILCFT100	Text-Only	0.52	0.88	NILCFT100	Text-Only	0.53	0.86
	Concatenated	0.52	0.88		Concatenated	0.54	0.85
	Auto-encoded	0.52	0.88		Auto-encoded	0.55	0.85
NILCW2V100	Text-Only	0.47	0.93	NILCW2V100	Text-Only	0.47	0.93
	Concatenated	0.47	0.93		Concatenated	0.48	0.92
	Auto-encoded	0.48	0.92		Auto-encoded	0.49	0.91
NILCFT300	Text-Only	0.50	0.90	NILCFT300	Text-Only	0.50	0.90
	Concatenated	0.51	0.90		Concatenated	0.51	0.90
	Auto-encoded	0.50	0.90		Auto-encoded	0.52	0.88

Table 5.14 shows that positive results are skewed toward the Auto-encoded architecture and the Standardization scaling algorithm. This is corroborative with what we have already learned.

Table 5.14 – The "*No. of best results*" column represents the number of times each architecture and scaling algorithm had the best results in a model. The Overall subtable presents a conglomeration of all results, while the Normalized and Standardized subtables present separated results for their respective scaling algorithms.

PT - Overall			
Architecture	**No. of Best Results**	**Scaling Algorithm**	**No. of Best Results**
Text-Only	0	Normalized	0
Concatenated	1	Standardized	8
Auto-encoded	4	-	-
PT - Normalized		**PT - Standardized**	
Architecture	**No. of Best Results**	**Architecture**	**No. of Best Results**
Text-Only	0	Text-Only	0
Concatenated	1	Concatenated	0
Auto-encoded	1	Auto-encoded	3

5.3.3 Semantic Similarity Task Overview

The results found with these Semantic Similarity tests echo those found with the Word Relatedness tests of Section 5.1. This is expected, as these tasks are similar, though in a different scale. The tests found essentially the same results: the Auto-encoding architecture is superior; the Standardization algorithm is better suited to multimodality; and multimodal models outperform Text-Only models. This can all be ascertained through the compiled information in Table 5.15

One of the more interesting findings is that Concatenated and Auto-encoded models trained on the same textual-visual corpus had similar results in these tests, though results show that the Concatenation architecture worked better with Normalization algorithm than the Standardization algorithm, and the opposite is true for the Auto-encoded architecture.

Table 5.15 – The "*No. of best results*" column represents the number of times each architecture and scaling algorithm had the best results in a model.

Overall			
Architecture	**No. of Best Results**	**Scaling Algorithm**	**No. of Best Results**
Text-Only	0	Normalized	2
Concatenated	3	Standardized	12
Auto-encoded	8	-	-

5.4 Named Entity Recognition

The Named Entity Recognition task requires that, as mentioned in Section 4.2.2, a model recognize and tag a given set of classes within raw textual input. The two test sets used for this task in this work are HAREM, a corpus built from news domain texts that will serve as the generic test set, and GeoCorpus, a corpus built from geosciences domain texts that will serve as the geosciences test set. As explained in Section 3.3, only the best results for each model will be considered during this analysis. The complete results for this test set are available in Appendix H.

5.4.1 HAREM

The HAREM test set is composed of two tracks, which will be analysed separately at first and then as part of an overview. It was used to test the models trained on the NILC and BBP text corpora.

Selective Track

The Selective track is the smaller of the two, including only the five most populated named entity categories within the test set. Table 5.16 shows that while multimodality did not manage to improve the best F-score achieved by the best Text-Only model, Standardized NILCFT300, it did raise another model, Standardized NILCW2V100, to tie with this score. Otherwise, whatever increases in F-score as a result of the multimodal fusion that can be observed are minimal at best for this task.

Table 5.16 – The best results for the Selective track of the HAREM task.

HAREM SELECTIVE									
NORMALIZED					**STANDARDIZED**				
Model	**Architecture**	**Precision**	**Recall**	**F1**	**Model**	**Architecture**	**Precision**	**Recall**	**F1**
BBPFT300	Text-Only	0.733	0.679	0.705	BBPFT300	Text-Only	0.734	0.680	0.706
	Concatenated	0.741	0.677	0.708		Concatenated	0.738	0.668	0.701
	Auto-encoded	0.745	0.650	0.694		Auto-encoded	0.728	0.653	0.688
NILCFT100	Text-Only	0.737	0.671	0.702	NILCFT100	Text-Only	0.716	0.691	0.703
	Concatenated	0.733	0.655	0.692		Concatenated	0.735	0.679	0.706
	Auto-encoded	0.739	0.656	0.695		Auto-encoded	0.731	0.691	0.710
NILCW2V100	Text-Only	0.736	0.659	0.696	**NILCW2V100**	Text-Only	0.727	0.690	0.708
	Concatenated	0.740	0.653	0.694		**Concatenated**	**0.746**	**0.686**	**0.715**
	Auto-encoded	0.755	0.650	0.699		**Auto-encoded**	**0.733**	**0.697**	**0.714**
NILCFT300	Text-Only	0.739	0.678	0.707	**NILCFT300**	**Text-Only**	**0.740**	**0.690**	**0.714**
	Concatenated	0.740	0.673	0.705		**Concatenated**	**0.741**	**0.691**	**0.715**
	Auto-encoded	0.769	0.635	0.696		Auto-encoded	0.737	0.665	0.699

Table 5.17 shows that the Standardization algorithm is superior to the Normalization algorithm, which once again has a tendency to worsen Text-Only results. As for the individual architectures, the Text-Only architecture was usually matched or slightly outperformed by the multimodal architectures, though the Concatenation architecture seems to be slightly superior in this regard. Few of the improvements upon the base Text-Only models was recorded to have been of over 0.5 percentage point increase, however, so it cannot be definitely concluded that the proposed multimodal architectures helped in this task.

Table 5.17 – The "*No. of best results*" column represents the number of times each architecture and scaling algorithm had the best results in a model. The Overall subtable presents a conglomeration of all results, while the Normalized and Standardized subtables present separated results for their respective scaling algorithms.

Selective - Overall			
Architecture	**No. of Best Results**	**Scaling Algorithm**	**No. of Best Results**
Text-Only	3	Normalized	2
Concatenated	3	Standardized	10
Auto-encoded	2	-	-
Selective - Normalized		**Selective - Standardized**	
Architecture	**No. of Best Results**	**Architecture**	**No. of Best Results**
Text-Only	2	Text-Only	1
Concatenated	1	Concatenated	2
Auto-encoded	1	Auto-encoded	1

Total Track

The Total track is the larger of the two, including all ten named entity categories present in the First HAREM test set. Table 5.18 shows that quite a few models tied for best score, and F-score of about 64 percentage points. This score was a tie among Text-Only and multimodal architectures. The Total track's results are similar to the Selective track's in shape: the multimodal models did not dramatically improve F-scores when compared to their Text-Only counterparts; many models were only worsened by multimodal fusion; and most Standardized models performed better than Normalized models.

Table 5.18 – The best results for the Total track of the HAREM task.

HAREM TOTAL									
NORMALIZED					STANDARDIZED				
Model	**Architecture**	**Precision**	**Recall**	**F1**	**Model**	**Architecture**	**Precision**	**Recall**	**F1**
BBPFT300	Text-Only	0.679	0.585	0.628	**BBPFT300**	**Text-Only**	**0.685**	**0.602**	**0.641**
	Concatenated	**0.694**	**0.597**	**0.642**		Concatenated	0.675	0.586	0.627
	Auto-encoded	0.689	0.573	0.626		Auto-encoded	0.678	0.571	0.619
NILCFT100	Text-Only	0.688	0.566	0.621	NILCFT100	**Text-Only**	**0.682**	**0.602**	**0.640**
	Concatenated	0.691	0.580	0.631		Concatenated	0.686	0.594	0.637
	Auto-encoded	0.710	0.573	0.634		**Auto-encoded**	**0.693**	**0.594**	**0.639**
NILCW2V100	Text-Only	0.687	0.579	0.628	NILCW2V100	Text-Only	0.675	0.595	0.633
	Concatenated	0.674	0.569	0.617		Concatenated	0.686	0.592	0.635
	Auto-encoded	0.677	0.585	0.628		Auto-encoded	0.676	0.597	0.634
NILCFT300	**Text-Only**	**0.684**	**0.600**	**0.639**	**NILCFT300**	Text-Only	0.667	0.599	0.631
	Concatenated	0.686	0.592	0.636		**Concatenated**	**0.680**	**0.606**	**0.641**
	Auto-encoded	0.712	0.577	0.637		Auto-encoded	0.690	0.591	0.637

Table 5.19 shows these traits more clearly. Roughly half the time multimodality strictly worsens the final result for this track, and Standardized models perform better on average. Finally, the Concatenation architecture tends to perform better than Auto-encoding architecture overall for this track.

Table 5.19 – The "*No. of best results*" column represents the number of times each architecture and scaling algorithm had the best results in a model. The Overall subtable presents a conglomeration of all results, while the Normalized and Standardized subtables present separated results for their respective scaling algorithms.

Total - Overall			
Architecture	**No. of Best Results**	**Scaling Algorithm**	**No. of Best Results**
Text-Only	4	Normalized	4
Concatenated	3	Standardized	8
Auto-encoded	1		
Total - Normalized		**Total - Standardized**	
Architecture	**No. of Best Results**	**Architecture**	**No. of Best Results**
Text-Only	2	Text-Only	2
Concatenated	1	Concatenated	2
Auto-encoded	1	Auto-encoded	0

HAREM Overview

In general, there was little to no improvement seen across both HAREM tracks when using multimodal fusion to augment embeddings. As seen in Table 5.20, when combined, the multimodal architectures achieved slightly better results than Text-Only models for roughly half of the models, while Text-Only outperformed multimodal by larger margins for the other half of all models. The table also reinforces the superiority of the Standardization algorithm over the Normalization algorithm.

Table 5.20 – The "*No. of best results*" column represents the number of times each architecture and scaling algorithm had the best results in a model.

HAREM Overall			
Architecture	**No. of Best Results**	**Scaling Algorithm**	**No. of Best Results**
Text-Only	7	Normalized	6
Concatenated	6	Standardized	18
Auto-encoded	3	-	-

5.4.2 GeoCorpus

GeoCorpus is a geosciences domain NER test set with 10 named entity categories. It was used to test models trained on the PetroVec text corpora. Table 5.21 tells a different story than the HAREM tables. Not only do the multimodal architectures generally outperform the Text-Only models, they outperform them by upwards of 3.2 percentage points. The highest results are generally achieved using the Concatenation architecture, the opposite of what happened in previous tasks where multimodal architectures outperformed the Text-Only models. This task was also the only instance where a Text-Only model that did not perform the best out of all Text-Only models yielded a multimodal model that performed best overall.

Table 5.21 – The best results for the GeoCorpus test set.

GEOCORPUS									
NORMALIZED					**STANDARDIZED**				
Model	**Architecture**	**Precision**	**Recall**	**F1**	**Model**	**Architecture**	**Precision**	**Recall**	**F1**
PetroVecFT	Text-Only	0.792	0.763	0.777	**PetroVecFT**	Text-Only	0.827	0.811	0.818
	Concatenated	0.822	0.758	0.789		**Concatenated**	**0.860**	**0.841**	**0.850**
	Auto-encoded	0.808	0.780	0.794		Auto-encoded	0.820	0.830	0.825
PetroVecHybridFT	Text-Only	0.783	0.735	0.758	PetroVecHybridFT	Text-Only	0.827	0.826	0.827
	Concatenated	0.836	0.795	0.815		Concatenated	0.858	0.827	0.842
	Auto-encoded	0.792	0.791	0.792		Auto-encoded	0.822	0.836	0.829
PetroVecW2V	Text-Only	0.803	0.766	0.784	PetroVecW2V	Text-Only	0.817	0.810	0.813
	Concatenated	0.831	0.796	0.813		Concatenated	0.841	0.799	0.819
	Auto-encoded	0.800	0.760	0.780		Auto-encoded	0.825	0.817	0.821
PetroVecHybridW2V	Text-Only	0.795	0.753	0.773	PetroVecHybridW2V	Text-Only	0.814	0.808	0.811
	Concatenated	0.837	0.796	0.815		Concatenated	0.851	0.827	0.838
	Auto-encoded	0.780	0.772	0.776		Auto-encoded	0.818	0.829	0.823

Table 5.22 once again reinforces the notion that the Standardization algorithm is better than the Normalization algorithm. It also confirms that the Concatenation architecture is on average better than the Auto-encoding architecture on average, and that multimodal models generally achieved better results for this task.

Table 5.22 – The "*No. of best results*" column represents the number of times each architecture and scaling algorithm had the best results in a model. The Overall subtable presents a conglomeration of all results, while the Normalized and Standardized subtables present separated results for their respective scaling algorithms.

GeoCorpus - Overall			
Architecture	**No. of Best Results**	**Scaling Algorithm**	**No. of Best Results**
Text-Only	0	Normalized	0
Concatenated	6	Standardized	12
Auto-encoded	2		
GeoCorpus - Normalized		**GeoCorpus - Standardized**	
Architecture	**No. of Best Results**	**Architecture**	**No. of Best Results**
Text-Only	0	Text-Only	0
Concatenated	3	Concatenated	3
Auto-encoded	1	Auto-encoded	1

5.4.3 Named Entity Recognition Task Overview

The HAREM and GeoCorpus tests produced different results. Whereas models for both HAREM tracks showed very little improvement as a result of multimodality, the GeoCorpus models' performance was substantially enhanced. Some similarities between the two is that the Concatenation architecture achieved better results on average, and that the best scaling algorithm was Standardization.

A possible explanation for this is that the Imagined Visual Embeddings do not have much of an impact in larger vocabulary such as those of the NILC and BBP text corpora trained models, whereas using the smaller PetroVec corpora trained models results in better embeddings overall. This is somewhat mirrored in the Word Relatedness test, the only other test where both domains were represented. Though both generic and geosciences test corpora saw an increase in correlation as a result of the multimodal fusion, the increases in the geosciences test were more pronounced, with the largest text-only to multimodal increase within the same model being almost twice the same for the generic test set.

Though more work must go into discovering the reason of the disparity between the HAREM and GeoCorpus results, the results presented in Table 5.23 represent the overall findings of this study: multimodality can improve upon Text-Only results for the NER task; the Concatenation architecture is better suited for NER; and the Standardization scaling algorithm continues to result in superior scores for this task.

Table 5.23 – The "*No. of best results*" column represents the number of times each architecture and scaling algorithm had the best results in a model.

Overall			
Architecture	**No. of Best Results**	**Scaling Algorithm**	**No. of Best Results**
Text-Only	7	Normalized	6
Concatenated	12	Standardized	30
Auto-encoded	5		

6. DISCUSSION AND CONCLUSIONS

This dissertation presented the results of a study into the usefulness of visual data when used in conjunction with textual data for NLP tasks in a general news domain and a specific geosciences domain. It involved the development of corpora and word embedding architectures which were then put through a test battery for multimodal Word Embedding models which included the following tasks: Word Relatedness, Sentence Similarity, Analogy Prediction and Named Entity Recognition. These results revealed some aspects of textual-visual multimodal fusion for Word Embeddings within NLP tasks for the Portuguese language, a field in which it is most common to study purely textual Word Embedding models.

The dissertation takes inspiration from the works of Bruni et al. (2014) [6], and their concatenation based multimodal fusion architecture; Silberer et al. (2014) [43], and their auto-encoding multimodal fusion architecture; and Collell et al. (2017) [11], and their Imagined Embeddings cross-modal mapping neural network, for visual vocabulary expansion. It takes a different tack from previous work by exploring the possibility of use of this technology beyond the English language, using resources for the Portuguese language, and also by exploring its use in specific knowledge domains, such as the geosciences domain presented within this work.

The testing performed in this dissertation further adds to the literature by empirically showing that multimodal fusion can improve Portuguese Language Word Embeddings in the tasks presented above, with the exception of Analogy Prediction for which the results were in favor of Text-Only models in the experimental settings presented herein. The Word Relatedness and Sentence Similarity tasks corroborate each other's results in expected ways, given that the tasks are essentially connected. The Named Entity recognition task, on the other hand, presented some contradictory results between its test sets, and its overall results do not entirely match the findings of the other tasks.

The Word Relatedness and Sentence Similarity tasks in particular were the largest focus for the *Geologia Digital* project. The project was interested in word embedding-based term expansion for information retrieval architectures, and this study shows that the addition of Imagined Visual Embeddings in the form of a multimodal fusion with Word Embeddings results in semantic distances that more closely correlate to human intuition, which can be helpful to a information search engine. Furthermore, tests showed that even though the images used as a base for the Imagined Embeddings did not belong to the geosciences domain, they still provided a substantial enhancement to the PetroVec embeddings.

The Analogy Prediction test was another test of interest to the project, though there was no geosciences domain corpus with which to test the PetroVec embeddings. As mentioned before, the poor results obtained for this task were expected, though it is hard to say whether or not usage with PetroVec's more focused vocabulary would not elicit better results in this area, as may have been the case for the Named Entity Recognition tests.

The NER tests themselves provided some interestingly mixed results. While the multimodality did not seem to offer much value to models working on the HAREM test corpora, the fused models did quite well on the GeoCorpus test set. In both tasks with both generic and geosciences test corpora available, the PetroVec models seemed to enhance results more, comparatively, on their respective tests. This points to a specific quality in the corpus which allows it to better integrate with the Imagined Visual Embeddings, such as the possibility that a more focused corpus results in better imagined embeddings. Regardless, the matter warrants further research in future work.

Another notable characteristic to be discussed in the results is the fact that the Auto-encoding architecture showed itself to be superior to the Concatenation architecture in both the Word Relatedness and the Sentence Similarity tasks, but inferior in the NER task. Further study is required to explain this, though it surely pertains to the details of how each neural network learns their respective tasks, and which of the input embedding's characteristics they value most.

Some limitations encountered during the research and development for this dissertation was a lack of training and testing resources for the Portuguese language in both the general news and geosciences domains. This meant that several resources had to be translated from English, collected and annotated from the ground up, and revised for use in the project. This also resulted in less testing within the geosciences domain than might otherwise be desirable, as two tasks only had appropriate Portuguese language tests for the general news domain. Another limitation of this work was the difficulty in providing statistical significance testing. While the difference between some results is great enough that it could safely be assumed that they are significant, confirmation for these and the closer results would help ground the study. This aspect was not added to this documentation since the complexity of the task demanded time that was not available, and we plan to tackle this issue in a future continuation of this research.

Additionally, it must be added that the use of traditional static word embedding models rather than contextualized models was deliberate. The reason was the need for multimodal solutions for these specific architectures in the *Geologia Digital* project. The choice of architecture was rooted in the fact that contextualized models, such as BERT or ELMO, significantly increase computational requirements for both training and inference when compared to non-contextual models, such as Word2Vec and fastText [38, 4]. This makes contextual embeddings less appealing in industrial scenarios, since, as per Polignano et al. (2020) [38], it is yet unclear whether the accuracy increase delivered by contextual embedding is worth the performance issues associated with them.

As for future work, an interesting avenue would be to branch off to Contextual Embeddings such as BERT. The *Geologia Digital* project has, in the past month, begun to experiment with such Contextual Embeddings for the geosciences domain, and a complete comparative evaluation against Word Embeddings will be possible as soon as these new embeddings are completed. Other techniques will have to be used as far as multimodal fusion is concerned, such as the contextual fusion proposed in EViLBERT [7], which takes complex images and full sentence descriptions as input, rather than simple image/word pairs. These paths of research are interesting possibilities for future work in the geosciences domain and for the Portuguese Language itself. Such research would also be able to tackle the question of which kind of embedding, Contextual or Static, excels in this context.

Another future work would involve an analysis of tasks which attempt to use text to aid in visual tasks, such as object detection/recognition. This would require that the images in question came with an accompanying text which might be analyzed. There could also be tests in inherently multimodal tasks, such as text-image pairing or text-image retrieval.

It must be noted that the work performed for this dissertation and summarised above would not have been possible without first building a knowledge base about multimodal semantic models and their use in different domains so that the author could realistically complete the proposed study. As part of this knowledge base building effort, the author was involved in the research for several academic papers on the field of semantic embeddings. In chronological order, they were: **Multidomain Contextual Embeddings for Named Entity Recognition**, published in the Proceedings of the IberLEF 2019 workshop by Santos et al. (2019) [42], which studied the use of contextual embeddings in the NER task for several non-standard domains such as the Legal and Medical domains; **Word Embedding Evaluation in Downstream Tasks and Semantic Analogies**, published in the Proceedings of the LREC 2020 conference by Santos et al. (2020) [41], which studied the use of different training corpora in word embedding models and made a new training corpus freely available; **Embeddings for Named Entity Recognition in Geoscience Portuguese Literature**, published in the Proceedings of the LREC 2020 conference by Consoli et al. (2020) [12], which was focused on the development of appropriate test corpora for a number of NLP tasks in the geosciences domain; **Portuguese Word Embeddings for the Oil and Gas Industry: development and evaluation**, published in the Computers in Industry journal by Gomes et al. (2021) [16] which explored the use of domain-specific embeddings against general domain embeddings when it comes to the oil and gas sub-domain of the geosciences domain.

REFERENCES

[1] Agirre, E.; Alfonseca, E.; Hall, K. B.; Kravalova, J.; Pasca, M.; Soroa, A. “A study on similarity and relatedness using distributional and wordnet-based approaches”. In: Proceedings of the 7th Conference of the North American Chapter of the Association of Computational Linguistics on Human Language Technologies, 2009, pp. 19–27.

[2] Akbik, A.; Blythe, D.; Vollgraf, R. “Contextual string embeddings for sequence labeling”. In: Proceedings of the 27th International Conference on Computational Linguistics, 2018, pp. 1638–1649.

[3] Amaral, D. O. F. “Reconhecimento de entidades nomeadas na área da geologia: Bacias sedimentares brasileiras”, Doctoral thesis, Programa de Pós-Graduação em Ciência da Computação, PUCRS, 2017, 109p.

[4] Arora, S.; May, A.; Zhang, J.; Ré, C. “Contextual embeddings: When are they worth it?” In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 2650–2663.

[5] Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; Amodei, D. “Language models are few-shot learners”. In: Proceedings of the 33rd Conference on Advances in Neural Information Processing Systems, 2020, pp. 25.

[6] Bruni, E.; Tran, N.; Baroni, M. “Multimodal distributional semantics”, *Journal of Artificial Intelligence Research*, vol. 49, Jan 2014, pp. 1–47.

[7] Calabrese, A.; Bevilacqua, M.; Navigli, R. “Evilbert: Learning task-agnostic multimodal sense embeddings”. In: Proceedings of the 29th International Joint Conference on Artificial Intelligence, 2020, pp. 481–487.

[8] Cao, D.; Yu, Z.; Zhang, H.; Fang, J.; Nie, L.; Tian, Q. “Video-based cross-modal recipe retrieval”. In: Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 1685–1693.

[9] Chen, J.; Liang, D.; Zhu, Z.; Zhou, X.; Ye, Z.; Mo, X. “Social media popularity prediction based on visual-textual features with xgboost”. In: Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 2692–2696.

[10] Collell, G.; Moens, M. “Do neural network cross-modal mappings really bridge modalities?” In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018, pp. 462–468.

[11] Collell, G.; Zhang, T.; Moens, M. “Imagined visual representations as multimodal embeddings”. In: Proceedings of the 31st AAAI Conference on Artificial Intelligence, 2017, pp. 4378–4384.

[12] Consoli, B. S.; Santos, J.; Gomes, D.; Cordeiro, F.; Vieira, R.; Moreira, V. “Embeddings for named entity recognition in geoscience portuguese literature”. In: Proceedings of the 12th Language Resources and Evaluation Conference, 2020, pp. 4625–4630.

[13] Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; Li, F. “Imagenet: A large-scale hierarchical image database”. In: Proceedings of the 23rd IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.

[14] Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. “BERT: pre-training of deep bidirectional transformers for language understanding”. In: Proceedings of the 17th Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies, 2019, pp. 4171–4186.

[15] Fonseca, E. R.; Santos, L. B.; Criscuolo, M.; Aluísio, S. M. “Visão geral da avaliação de similaridade semântica e inferência textual”, *Linguamática*, vol. 8, Dec 2016, pp. 3–13.

[16] Gomes, D. S. M. and Cordeiro, F. C. and Consoli, B. S. and Santos, N. L. and Moreira, V. P. and Vieira, R. and Moraes, S. and Evsukoff, A. G. “Portuguese word embeddings for the oil and gas industry: Development and evaluation”, *Computers in Industry*, vol. 124, Jan 2021, pp. 1–44.

[17] Gong, X.; Huang, L.; Wang, F. “Deep semantic correlation learning based hashing for multimedia cross-modal retrieval”. In: Proceedings of the 18th IEEE International Conference on Data Mining, 2018, pp. 117–126.

[18] Grave, E.; Mikolov, T.; Joulin, A.; Bojanowski, P. “Bag of tricks for efficient text classification”. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, 2017, pp. 427–431.

[19] Guo, D.; Hong, J.; Luo, B.; Yan, Q.; Niu, Z. “Multi-modal representation learning for short video understanding and recommendation”. In: Proceedings of the 20th IEEE International Conference on Multimedia & Expo, 2019, pp. 687–690.

[20] Guo, W.; Huang, H.; Kong, X.; He, R. “Learning disentangled representation for cross-modal retrieval with deep mutual information estimation”. In: Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 1712–1720.

[21] Guo, Z.; Gao, L.; Song, J.; Xu, X.; Shao, J.; Shen, H. T. “Attention-based LSTM with semantic consistency for videos captioning”. In: Proceedings of the 26th ACM Conference on Multimedia Conference, 2016, pp. 357–361.

[22] Habibian, A.; Mensink, T.; M., S. C. G. “Video2vec embeddings recognize events when examples are scarce”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, Jul 2017, pp. 2089–2103.

[23] Han, X.; Wu, Z.; Huang, P. X.; Zhang, X.; Zhu, M.; Li, Y.; Zhao, Y.; Davis, L. S. “Automatic spatially-aware fashion concept discovery”. In: Proceedings of the 16th IEEE International Conference on Computer Vision, 2017, pp. 1472–1480.

[24] Han, X.; Wu, Z.; Jiang, Y.; Davis, L. S. “Learning fashion compatibility with bidirectional lstms”. In: Proceedings of the 25th ACM on Multimedia Conference, 2017, pp. 1078–1086.

[25] Hao, Y.; Ngo, C.; Huet, B. “Neighbourhood structure preserving cross-modal embedding for video hyperlinking”, *IEEE Transactions on Multimedia*, vol. 22, Jan 2020, pp. 188–200.

[26] Hartmann, N.; Fonseca, E. R.; Shulby, C.; Treviso, M. V.; Rodrigues, J. S.; Aluísio, S. M. “Portuguese word embeddings: Evaluating on word analogies and natural language tasks”. In: Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology, 2017, pp. 122–131.

[27] He, Z.; Liu, J.; Li, N.; Huang, Y. “Learning network-to-network model for content-rich network embedding”. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019, pp. 1037–1045.

[28] Lazaridou, A.; Pham, N. T.; Baroni, M. “Combining language and vision with a multimodal skip-gram model”. In: Proceedings of the 13th Conference of the North American Chapter of the Association of Computational Linguistics on Human Language Technologies, 2015, pp. 153–163.

[29] Li, D.; Huang, M.; Li, X.; Ruan, Y.; Yao, L. “Mfecnn: Mixture feature embedding convolutional neural network for data mapping”, *IEEE Transactions on Nanobioscience*, vol. 17, Jul 2018, pp. 165–171.

[30] Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. “Efficient estimation of word representations in vector space”. In: Proceedings of the 1st International Conference on Learning Representations, 2013, pp. 12.

[31] Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; Dean, J. “Distributed representations of words and phrases and their compositionality”. In: Proceedings of the 27th Annual Conference on Neural Information Processing Systems, 2013, pp. 3111–3119.

[32] Nedelec, T.; Smirnova, E.; Vasile, F. “Specializing joint representations for the task of product recommendation”. In: Proceedings of the 2nd Workshop on Deep Learning for Recommender Systems, 2017, pp. 10–18.

[33] Niu, Z.; Zhou, M.; Wang, L.; Gao, X.; Hua, G. “Hierarchical multimodal LSTM for dense visual-semantic embedding”. In: Proceedings of the 16th IEEE International Conference on Computer Vision, 2017, pp. 1899–1907.

[34] Oramas, S.; Nieto, O.; Sordo, M.; Serra, X. “A deep multimodal approach for cold-start music recommendation”. In: Proceedings of the 2nd Workshop on Deep Learning for Recommender Systems, 2017, pp. 32–37.

[35] Paiva, V.; Rademaker, A.; Melo, G. “Openwordnet-pt: An open brazilian wordnet for reasoning”. In: Proceedings of the 24th International Conference on Computational Linguistics, 2012, pp. 353–360.

[36] Pennington, J.; Socher, R.; Manning, C. D. “Glove: Global vectors for word representation”. In: Proceedings of the 19th Conference on Empirical Methods in Natural Language Processing, 2014, pp. 1532–1543.

[37] Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. “Deep contextualized word representations”. In: Proceedings of the 16th Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies, 2018, pp. 2227–2237.

[38] Polignano, M.; Gemmis, M.; Semeraro, G. “Contextualized BERT sentence embeddings for author profiling: The cost of performances”. In: Proceedings of the 20th International Conference on Computational Science and its Applications, 2020, pp. 135–149.

[39] Santos, D.; Cardoso, N. “A golden resource for named entity recognition in portuguese”. In: Proceeding of the 7th International Conference on the Computational Processing of Portuguese, 2007, pp. 69–79.

[40] Santos, J.; Consoli, B. S.; Santos, C. N.; Terra, J.; Collovini, S.; Vieira, R. “Assessing the impact of contextual embeddings for portuguese named entity recognition”. In: Proceedings of the 8th Brazilian Conference on Intelligent Systems, 2019, pp. 437–442.

[41] Santos, J.; Consoli, B. S.; Vieira, R. “Word embedding evaluation in downstream tasks and semantic analogies”. In: Proceedings of the 12th Language Resources and Evaluation Conference, 2020, pp. 4828–4834.

[42] Santos, J.; Terra, J.; Consoli, B. S.; Vieira, R. “Multidomain contextual embeddings for named entity recognition”. In: Proceedings of the Iberian Languages Evaluation

Forum co-located with 35th Conference of the Spanish Society for Natural Language Processing, 2019, pp. 434–441.

[43] Silberer, C.; Lapata, M. “Learning grounded meaning representations with autoencoders”. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014, pp. 721–732.

[44] Souza, F.; Nogueira, R.; Lotufo, R. “Bertimbau: Pretrained BERT models for brazilian portuguese”. In: Proceedings of the 9th Brazilian Conference on Intelligent Systems, 2020, pp. 403–417.

[45] Wang, S.; Guo, D.; Xu, X.; Zhuo, L.; Wang, M. “Cross-modality retrieval by joint correlation learning”, *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 15, Jul 2019, pp. 1–16.

[46] Zhang, W.; Wang, W.; Wang, J.; Zha, H. “User-guided hierarchical attention network for multi-modal social image popularity prediction”. In: Proceedings of the 27th World Wide Web Conference, 2018, pp. 1277–1286.

[47] Zheng, C.; Pan, L.; Wu, P. “Multimodal deep network embedding with integrated structure and attribute information”, *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, May 2020, pp. 1437–1449.

APPENDIX A – GEOSIM

Intrinsic testing is an important method of evaluation for Word Embedding based semantic models. These kinds of tests best serve their purpose if they were created with the model's use domain in mind, which means that already-existing tests for the generic news domain are not appropriate for evaluating models which focus on the geosciences, such as the PetroVec models of the *Geologia Digital* project. This prompted the development of a new Word Relatedness intrinsic test specifically build with the likely use scenario of the PetroVec models in mind.

A.1 Related Work

The development methodology for this test corpus was inspired by the work of Aguirre et al. (2009) [1] and Bruni et al. (2014) [6]. Aguirre et al. (2009) used the annotations of a handful of experts in the form of a 1 (least similar) to 7 (most similar) Likert scale to compose their corpus, which had a about 500 total annotated pairs. These pairs were divided into two tracks, as half of the pairs were annotated for similarity and the other half for relatedness.

Bruni et al. (2014), on the other hand, chose to eschew the similarity test altogether, citing that the relatedness test was more relevant to the testing of Word Embeddings, given the inherent nature of semantic vector spaces. To annotate their corpus, they used the Amazon's crowdsourcing tool, Mechanical Turk[1], to gather the results of 50 pair to pair comparisons for each pair, with the pair whose words are most related to each other gaining one point per comparison (eg. if the pair "Cat - Tiger" were compared to "car - lake", the first would receive one point and the latter no points, and the both would each be compared to 49 independent pairs to round out the 50 total comparisons). The most similar pairs had scores closest to 50, while the least similar pairs had scores closes to 0.

A.2 Development Methodology

The word pairs for GeoSim were generated from the words within Petroleum Abstract's Exploration and Production Controlled Vocabulary Thesaurus[2]. The relations between words recorded in the thesaurus were used to balance pairs into three categories: strong relationships (categorized as hypernym or hyponym); weak relationships (catego-

[1] https://www.mturk.com/
[2] https://www.pa.utulsa.edu/products/tulsadatabase/thesaurus

rized as other); and not related (no relationship between the words). These categories were created to ensure that the chosen corpus would have a plethora of strong and weak relationships to use as part of model testing, rather than the almost certainly purely weak relationships that would be present in a a truly random pair corpus.

Each pair was then annotated with the aid of a 1 (least similar) to 7 (most similar) Likert scale, the same as was used for Aguirre et al. (2009), by three experts in the Geosciences field affiliated with the *Geologia Digital* project. This resulted in a list of 1500 word pairs, each annotated with three scores. The scores were then averaged, achieving the final score of the corpus for each word pair.

APPENDIX B – GEOCORPUS 3.0

GeoCorpus is an evaluation corpus for the Portuguese language that collects several scientific works in the field of Geology. This corpus, originally developed by Daniela Amaral [3], contains works whose theme is geological entities (GE) related to the Brazilian Sedimentary Basin subdomain. The collected texts are essentially theses, dissertations, articles and bulletins from Petrobras' geosciences publications. They were recovered and selected from the geological terms of the Chronostratigraphic table, which contains names of sedimentary rocks, names of Brazilian sedimentary basins, the terms related to Tectonics, Sedimentation and Magmatism and stratigraphic units.

In this report, the changes that were made to this corpus in order to improve it will be presented and discussed.

B.1 Modifications

The corpus contained several problems that had a negative impact on machine learning experiments carried out with it. Below are the modifications made.

B.1.1 Removing empty categories

Description: There were some categories in GeoCorpus with an empty identifier.

Example: <EM CATEG=" ">quartzo</EM>

Solution: The empty categories were removed from the corpus in order to standardize them. It is important to note that this change does not impact the annotation.

B.1.2 Removal of nested categories

Description: Some annotated terms contained nested categories.

Example: <EM CATEG="EstruturaSedimentar"><EM CATEG="baciaSedimentar">Bacia do Paraná </EM></EM>

Solution: In these cases, only the more specialized category was removed, leaving only the more generic of the two.

B.1.3 Correction of category annotation

Description: Some entities have been categorized into more than one class.

Example: <EM CATEG="ERA">Neoarqueano</EM> ...
<EM CATEG="PERIODO">Neoarqueano</EM>

Solution: In these cases the annotation was redone. We sent the identified cases to an expert, who assigned the correct class to the entity.

B.1.4 Removing duplicate lines

Description: There were some repeated lines in GeoCorpus.

Example: Lines 766 and 776 were the same, containing the same sentence: "Grãos de silicato de zircônio incrustados em rochas metamórficas do grupo Warrawoona na Austrália ocidental foram datados em até 4 , 4 bilhões de anos , indicando que por essa época uma crosta estava se consolidando."

Solution: Identical lines have been removed from GeoCorpus, since repetitions tend to hinder machine learning. Altogether, 73 lines were removed, containing a total of 51 entities.

B.1.5 Correction of improperly broken lines

Description: There was an improper line break pattern in GeoCorpus. In some sentences with a comma or opening parentheses, there was a new line segmenting the sentence into two parts.

Solution: As not all lines with parentheses or commas had the incorrect line break, the phrases that presented this break were corrected manually.

B.1.6 Other category removal

Description: There was a category called ’others’ in GeoCorpus, with 737 entities.

Solution: All entities in this category were passed on to an expert, and recategorized into more specific categories, for entities that did not fit into the existing categories, new categories were created indicated by the expert. This was deemed necessary because the class ’others’ was very wide only hindered automatic classification.

B.1.7 Standardization of categories

Description: There was no pattern in the name of the corpus categories, some were all capitalized, others in lowercase, and those with compound words alternated.

Solution: All categories are named with the Camel Case standard.

B.1.8 Entities without annotation

Description: 2913 entities in GeoCorpus that should be annotated and were not were identified. These entities were annotated one or more times in the corpus, but in certain instances were not annotated.

Example:
Sentence 1: ... de <EM CATEG="sedimentaresSiliciclasticas">**quartzo**</EM>.
Sentence 2: ... em adição a outros minerais detríticos como o **quartzo**. In this example, the ’quartzo’ entity appears categorized in the first sentence, but in the second, it is not.

Solution: All entities that should be categorized were categorized using a script. It was possible to use a script because the words that were not categorized had no problem of context or double meaning, which would make them appear at one time categorized and at another time not.

B.2 Analysis of GeoCorpus

After modifying the corpus in an attempt to obtain a better result, an analysis was performed on the entities of the modified corpus, together with an analysis on the entities of the unmodified corpus, for comparison purposes. With that, we want to compare and contrast the two versions of GeoCorpus.

In the original Geocorpus we had a sum of 6126 registered entities, divided into 20 classes, with the necessary modifications made, the impact we had on the number of categories and classes is observed. In the modified Geocorpus, we have a total of 8954 registered entities, an increase of 2828 entities, divided into 30 classes, an increase of 10 classes. These numbers are presented in Table B.1.

The number of unique entities of the new GeoCorpus was also analyzed. With this table, we can see that within the 8954 total entities noted, there are a total of 1229 distinct entities, a number well below the total number of entities, which demonstrates that there are many repetitions of the same entities, something that we have to take into account. These numbers are presented in Table B.2.

Table B.1 – Comparison GeoCorpus: Original version x GeoCorpus: Revised version

Class	**#Instances(Original)**	**#Instances(Revised)**
Time		
age	796	799
eon	288	256
era	326	414
epoch	650	687
period	637	714
Rocks		
metamorphics	197	378
magmatics	222	582
siliciclasticSedimentary	741	1102
carbonateSedimentary	240	355
chemicalSedimentary	5	12
organicSedimentary	22	22
Constituents and Properties of Rocks		
sedimentaryRockConstituent	0	112
mineral	0	212
fossils	0	132
sedimentaryStructure	0	86
geologicalStructure	0	78
Site		
basinsGeologicalContext	262	663
sedimentationEnvironment	0	146
bentonic	13	27
planktonic	44	112
oilField	0	6
Elements of Stratigraphy		
sedimentaryBasin	243	552
stratigraphicUnit	578	764
geotectonicUnit	0	28
stratigraphy	0	247
formation	18	0
Others		
petroleumSystem	0	93
basinStructure	40	0
geomorphology	0	54
granulometry	67	129
chemicalElement	0	26
methodologicalProcedure	0	166
other	737	0
Sum	**6126**	**8954**

Table B.2 – Unique Entities by class, organized into superclasses - Revised Version

Class	**#Unique Instances**
Time	
age	84
epoch	74
period	62
era	47
eon	20
Rocks	
metamorphics	59
magmatics	58
siliciclasticSedimentary	160
carbonateSedimentary	77
chemicalSedimentary	4
organicSedimentary	1
Constituents and Properties of Rocks	
sedimentaryRockConstituent	24
mineral	6
fossils	29
sedimentaryStructure	28
sedimentaryBasin	83
geologicalStructure	19
Site	
sedimentationEnvironment	32
basinsGeologicalContext	121
bentonic	4
planktonic	9
oilField	2
Elements of Stratigraphy	
stratigraphicUnit	153
geotectonicUnit	8
stratigraphy	29
Others	
geomorphology	6
chemicalElement	3
granulometry	13
methodologicalProcedure	4
petroleumSystem	10
Sum	**1229**

APPENDIX C – READING THE FULL RESULTS

The results presented in the following appendices are organized by fusion architecture, scaling algorithm and model. Multimodal models are compared with text-only models used to build their imagined visual embeddings, as was the case in the main body of this dissertation.

Additionally, each result will be highlight using the following colors: dark yellow representing text-only result; green representing multimodal result that is at least 1 percentage point higher than the paired text-only result; light yellow representing a multimodal result that is less than 1 percentage point lower or higher than the paired text-only result; and red representing a multimodal result that is at least 1 percentage point lower than the text-only paired result.

APPENDIX D – MEN FULL RESULTS

Table D.1 – Complete results for the MEN dataset.

Concatenated Models			
Normalized		Standardized	
Model	Spearman	Model	Spearman
BBPFT300-NORM	0.607	BBPFT300-STDZ	0.610
BBPFT-NORM_25	0.620	BBPFT-STDZ_25	0.640
BBPFT-NORM_50	0.621	BBPFT-STDZ_50	0.648
BBPFT-NORM_100	0.622	BBPFT-STDZ_100	0.643
		BBPFT-STDZ_150	0.642
NILCFT100-NORM	0.588	NILCFT100-STDZ	0.615
NILCFT100-NORM_25	0.611	NILCFT100-STDZ_25	0.646
NILCFT100-NORM_50	0.620	NILCFT100-STDZ_50	0.648
NILCFT100-NORM_100	0.626	NILCFT100-STDZ_100	0.643
		NILCFT100-STDZ_150	0.642
NILCW2V100-NORM	0.489	NILCW2V100-STDZ	0.493
NILCW2V100-NORM_25	0.502	NILCW2V100-STDZ_25	0.513
NILCW2V100-NORM_50	0.516	NILCW2V100-STDZ_50	0.518
NILCW2V100-NORM_100	0.530	NILCW2V100-STDZ_100	0.514
		NILCW2V100-STDZ_150	0.517
NILCFT300-NORM	0.567	NILCFT300-STDZ	0.570
NILCFT300-NORM_25	0.580	NILCFT300-STDZ_25	0.583
NILCFT300-NORM_50	0.588	NILCFT300-STDZ_50	0.588
NILCFT300-NORM_100	0.595	NILCFT300-STDZ_100	0.583
		NILCFT300-STDZ_150	0.586
Auto-encoded Models			
Normalized		**Standardized**	
Model	**Spearman**	**Model**	**Spearman**
BBPFT300-NORM	0.607	BBPFT300-STDZ	0.610
BBPFT-NORM_25_AE	0.604	BBPFT-STDZ_25_AE	0.636
BBPFT-NORM_50_AE	0.580	BBPFT-STDZ_50_AE	0.649
BBPFT-NORM_100_AE	0.624	BBPFT-STDZ_100_AE	0.641
		BBPFT-STDZ_150_AE	0.642
NILCFT100-NORM	0.588	NILCFT100-STDZ	0.615
NILCFT100-NORM_25_AE	0.621	NILCFT100-STDZ_25_AE	0.644
NILCFT100-NORM_50_AE	0.620	NILCFT100-STDZ_50_AE	0.649
NILCFT100-NORM_100_AE	0.623	NILCFT100-STDZ_100_AE	0.640
		NILCFT100-STDZ_150_AE	0.638
NILCW2V100-NORM	0.489	NILCW2V100-STDZ	0.493
NILCW2V100-NORM_25_AE	0.506	NILCW2V100-STDZ_25_AE	0.527
NILCW2V100-NORM_50_AE	0.528	NILCW2V100-STDZ_50_AE	0.528
NILCW2V100-NORM_100_AE	0.527	NILCW2V100-STDZ_100_AE	0.518
		NILCW2V100-STDZ_150_AE	0.521
NILCFT300-NORM	0.567	NILCFT300-STDZ	0.570
NILCFT300-NORM_25_AE	0.584	NILCFT300-STDZ_25_AE	0.597
NILCFT300-NORM_50_AE	0.571	NILCFT300-STDZ_50_AE	0.576
NILCFT300-NORM_100_AE	0.601	NILCFT300-STDZ_100_AE	0.585
		NILCFT300-STDZ_150_AE	0.574

APPENDIX E – GEOSIM FULL RESULTS

Table E.1 – Full results table for the GeoSim testset.

Concatenated Models			
Normalized		**Standardized**	
Model	**Spearman**	**Model**	**Spearman**
PetroVecFT	0.607	PetroVecFT	0.609
PetroVecFT-25	0.606	PetroVecFT-25	0.613
PetroVecFT-50	0.607	PetroVecFT-50	0.613
PetroVecFT-100	0.611	PetroVecFT-100	0.615
		PetroVecFT-150	0.605
PetroVecHybridFT	0.607	PetroVecHybridFT	0.619
PetroVecHybridFT-25	0.605	PetroVecHybridFT-25	0.633
PetroVecHybridFT-50	0.621	PetroVecHybridFT-50	0.632
PetroVecHybridFT-100	0.629	PetroVecHybridFT-100	0.627
		PetroVecHybridFT-150	0.626
PetroVecW2V	0.608	PetroVecW2V	0.611
PetroVecW2V-25	0.611	PetroVecW2V-25	0.613
PetroVecW2V-50	0.610	PetroVecW2V-50	0.613
PetroVecW2V-100	0.613	PetroVecW2V-100	0.612
		PetroVecW2V-150	0.610
PetroVecHybridW2V	0.643	PetroVecHybridW2V	0.648
PetroVecHybridW2V-25	0.652	PetroVecHybridW2V-25	0.651
PetroVecHybridW2V-50	0.658	PetroVecHybridW2V-50	0.652
PetroVecHybridW2V-100	0.660	PetroVecHybridW2V-100	0.652
		PetroVecHybridW2V-150	0.655
Auto-encoded Models			
Normalized		**Standardized**	
Model	**Spearman**	**Model**	**Spearman**
PetroVecFT	0.607	PetroVecFT	0.609
PetroVecFT-25	0.621	PetroVecFT-25	0.642
PetroVecFT-50	0.619	PetroVecFT-50	0.636
PetroVecFT-100	0.582	PetroVecFT-100	0.637
		PetroVecFT-150	0.629
PetroVecHybridFT	0.607	PetroVecHybridFT	0.619
PetroVecHybridFT-25	0.643	**PetroVecHybridFT-25**	**0.667**
PetroVecHybridFT-50	0.657	PetroVecHybridFT-50	0.655
PetroVecHybridFT-100	0.652	PetroVecHybridFT-100	0.646
		PetroVecHybridFT-150	0.652
PetroVecW2V	0.608	PetroVecW2V	0.611
PetroVecW2V-25	0.617	PetroVecW2V-25	0.629
PetroVecW2V-50	0.621	PetroVecW2V-50	0.626
PetroVecW2V-100	0.619	PetroVecW2V-100	0.629
		PetroVecW2V-150	0.628
PetroVecHybridW2V	0.643	PetroVecHybridW2V	0.648
PetroVecHybridW2V-25	**0.664**	**PetroVecHybridW2V-25**	**0.664**
PetroVecHybridW2V-50		PetroVecHybridW2V-50	0.656
PetroVecHybridW2V-100	**0.667**	PetroVecHybridW2V-100	0.652
		PetroVecHybridW2V-150	**0.662**

APPENDIX F – ANALOGY PREDICTION FULL RESULTS

Table F.1 – Results for Concatenated models in the Analogies dataset.

BRAZILIAN PORTUGUESE							
NORMALIZED				**STANDARDIZED**			
Model	**Syntactic**	**Semantic**	**Total**	**Model**	**Syntactic**	**Semantic**	**Total**
BBPFT	0.445	0.064	0.256	BBPFT	0.447	0.064	0.257
BBPFT_25	0.443	0.064	0.255	BBPFT_25	0.441	0.063	0.254
BBPFT_50	0.444	0.065	0.256	BBPFT_50	0.436	0.065	0.251
BBPFT_100	0.443	0.065	0.255	BBPFT_100	0.410	0.057	0.235
				BBPFT_150	0.391	0.066	0.230
NILCFT100	0.487	0.282	0.384	**NILCFT100**	**0.510**	**0.302**	**0.406**
NILCFT100_25	0.481	0.280	0.380	**NILCFT100_25**	**0.505**	**0.292**	**0.398**
NILCFT100_50	0.479	0.275	0.377	**NILCFT100_50**	**0.505**	**0.278**	**0.391**
NILCFT100_100	0.477	0.272	0.374	NILCFT100_100	0.447	0.256	0.351
				NILCFT100_150	0.415	0.221	0.318
NILCW2V100	0.247	0.077	0.162	NILCW2V100	0.255	0.080	0.167
NILCW2V100_25	0.247	0.075	0.161	NILCW2V100_25	0.254	0.081	0.167
NILCW2V100_50	0.239	0.072	0.156	NILCW2V100_50	0.245	0.077	0.161
NILCW2V100_100	0.234	0.072	0.153	NILCW2V100_100	0.227	0.073	0.150
				NILCW2V100_150	0.206	0.074	0.140
NILCFT300	0.330	0.154	0.242	NILCFT300	0.332	0.158	0.245
NILCFT300_25	0.335	0.155	0.245	NILCFT300_25	0.331	0.157	0.244
NILCFT300_50	0.330	0.153	0.241	NILCFT300_50	0.330	0.159	0.244
NILCFT300_100	0.330	0.156	0.243	NILCFT300_100	0.324	0.157	0.240
				NILCFT300_150	0.323	0.161	0.242
EUROPEAN PORTUGUESE							
NORMALIZED				**STANDARDIZED**			
Model	**Syntactic**	**Semantic**	**Total**	**Model**	**Syntactic**	**Semantic**	**Total**
BBPFT	0.448	0.057	0.254	BBPFT	0.451	0.058	0.255
BBPFT_25	0.447	0.058	0.253	BBPFT_25	0.444	0.057	0.251
BBPFT_50	0.448	0.058	0.254	BBPFT_50	0.439	0.059	0.250
BBPFT_100	0.447	0.058	0.253	BBPFT_100	0.412	0.051	0.232
				BBPFT_150	0.393	0.057	0.226
NILCFT100	0.485	0.274	0.379	**NILCFT100**	**0.509**	**0.293**	**0.401**
NILCFT100_25	0.481	0.273	0.377	**NILCFT100_25**	**0.505**	**0.284**	**0.394**
NILCFT100_50	0.479	0.267	0.372	NILCFT100_50	0.504	0.270	0.387
NILCFT100_100	0.476	0.263	0.369	NILCFT100_100	0.446	0.242	0.343
				NILCFT100_150	0.415	0.213	0.314
NILCW2V100	0.243	0.072	0.158	NILCW2V100	0.252	0.074	0.163
NILCW2V100_25	0.243	0.070	0.156	NILCW2V100_25	0.250	0.075	0.163
NILCW2V100_50	0.234	0.069	0.152	NILCW2V100_50	0.241	0.073	0.157
NILCW2V100_100	0.231	0.067	0.149	NILCW2V100_100	0.224	0.069	0.146
				NILCW2V100_150	0.203	0.066	0.134
NILCFT300	0.322	0.140	0.231	NILCFT300	0.324	0.143	0.233
NILCFT300_25	0.327	0.139	0.233	NILCFT300_25	0.323	0.143	0.233
NILCFT300_50	0.323	0.137	0.230	NILCFT300_50	0.322	0.144	0.233
NILCFT300_100	0.322	0.141	0.231	NILCFT300_100	0.317	0.141	0.229
				NILCFT300_150	0.316	0.145	0.230

Table F.2 – Results for Auto-encoded models in the Analogies dataset.

PORTUGUÊS BRASILEIRO							
NORMALIZED				STANDARDIZED			
Model	**Syntactic**	**Semantic**	**Total**	**Model**	**Syntactic**	**Semantic**	**Total**
BBPFT	0.445	0.064	0.256	BBPFT	0.447	0.064	0.257
BBPFT_25	0.391	0.054	0.224	BBPFT_25	0.382	0.047	0.216
BBPFT_50	0.395	0.053	0.225	BBPFT_50	0.377	0.048	0.214
BBPFT_100	0.392	0.058	0.226	BBPFT_100	0.359	0.049	0.205
				BBPFT_150	0.351	0.052	0.203
NILCFT100	0.487	0.282	0.384	**NILCFT100**	**0.510**	**0.302**	**0.406**
NILCFT100_25	0.495	0.301	0.398	**NILCFT100_25**	**0.511**	**0.311**	**0.411**
NILCFT100_50	0.491	0.290	0.390	**NILCFT100_50**	**0.504**	**0.301**	**0.402**
NILCFT100_100	0.486	0.288	0.387	NILCFT100_100	0.434	0.279	0.356
				NILCFT100_150	0.405	0.241	0.323
NILCW2V100	0.247	0.077	0.162	NILCW2V100	0.255	0.080	0.167
NILCW2V100_25	0.239	0.072	0.155	NILCW2V100_25	0.235	0.080	0.157
NILCW2V100_50	0.229	0.071	0.150	NILCW2V100_50	0.216	0.070	0.143
NILCW2V100_100	0.235	0.072	0.153	NILCW2V100_100	0.188	0.069	0.129
				NILCW2V100_150	0.171	0.065	0.118
NILCFT300	0.330	0.154	0.242	NILCFT300	0.332	0.158	0.245
NILCFT300_25	0.285	0.142	0.214	NILCFT300_25	0.299	0.143	0.221
NILCFT300_50	0.283	0.138	0.210	NILCFT300_50	0.293	0.144	0.219
NILCFT300_100	0.283	0.141	0.212	NILCFT300_100	0.294	0.132	0.213
				NILCFT300_150	0.282	0.139	0.210
PORTUGUÊS EUROPEU							
NORMALIZED				**STANDARDIZED**			
Model	**Syntactic**	**Semantic**	**Total**	**Model**	**Syntactic**	**Semantic**	**Total**
BBPFT	0.448	0.057	0.254	BBPFT	0.451	0.058	0.255
BBPFT_25	0.394	0.049	0.223	BBPFT_25	0.387	0.042	0.216
BBPFT_50	0.399	0.049	0.225	BBPFT_50	0.382	0.045	0.214
BBPFT_100	0.395	0.050	0.224	BBPFT_100	0.363	0.044	0.204
				BBPFT_150	0.355	0.048	0.202
NILCFT100	0.485	0.274	0.379	**NILCFT100**	**0.509**	**0.293**	**0.401**
NILCFT100_25	0.493	0.291	0.392	**NILCFT100_25**	**0.509**	**0.298**	**0.403**
NILCFT100_50	0.490	0.279	0.384	NILCFT100_50	0.502	0.291	0.396
NILCFT100_100	0.485	0.277	0.381	NILCFT100_100	0.432	0.259	0.346
				NILCFT100_150	0.404	0.226	0.315
NILCW2V100	0.243	0.072	0.158	NILCW2V100	0.252	0.074	0.163
NILCW2V100_25	0.235	0.067	0.151	NILCW2V100_25	0.231	0.074	0.152
NILCW2V100_50	0.225	0.067	0.146	NILCW2V100_50	0.211	0.064	0.137
NILCW2V100_100	0.230	0.067	0.148	NILCW2V100_100	0.185	0.064	0.125
				NILCW2V100_150	0.168	0.059	0.113
NILCFT300	0.322	0.140	0.231	NILCFT300	0.324	0.143	0.233
NILCFT300_25	0.282	0.128	0.205	NILCFT300_25	0.292	0.123	0.208
NILCFT300_50	0.276	0.126	0.201	NILCFT300_50	0.287	0.125	0.206
NILCFT300_100	0.276	0.126	0.201	NILCFT300_100	0.289	0.119	0.204
				NILCFT300_150	0.277	0.121	0.199

APPENDIX G – ASSIN FULL RESULTS

Table G.1 – The complete results for the Brazilian Portuguese ASSIN track.

Concatenated Models					
NORMALIZED			STANDARDIZED		
Model	**Pearson**	**MSE**	**Model**	**Pearson**	**MSE**
BBPFT	0.56	0.52	BBPFT	0.56	0.52
BBPFT_25	0.56	0.52	BBPFT_25	0.57	0.52
BBPFT_50	0.56	0.52	BBPFT_50	0.56	0.52
BBPFT_100	0.56	0.52	**BBPFT_100**	**0.57**	**0.51**
			BBPFT_150	0.57	0.52
NILCFT100	0.53	0.55	NILCFT100	0.53	0.54
NILCFT100_25	0.54	0.54	NILCFT100_25	0.54	0.54
NILCFT100_50	0.54	0.54	NILCFT100_50	0.54	0.54
NILCFT100_100	0.53	0.54	NILCFT100_100	0.54	0.54
			NILCFT100_150	0.53	0.55
NILCW2V100	0.45	0.60	NILCW2V100	0.45	0.61
NILCW2V100_25	0.47	0.60	NILCW2V100_25	0.45	0.60
NILCW2V100_50	0.46	0.60	NILCW2V100_50	0.45	0.60
NILCW2V100_100	0.46	0.60	NILCW2V100_100	0.46	0.60
			NILCW2V100_150	0.46	0.60
NILCFT300	0.49	0.58	NILCFT300	0.49	0.58
NILCFT300_25	0.50	0.57	NILCFT300_25	0.50	0.57
NILCFT300_50	0.50	0.57	NILCFT300_50	0.49	0.57
NILCFT300_100	0.50	0.57	NILCFT300_100	0.50	0.57
			NILCFT300_150	0.49	0.57
Auto-encoded Models					
NORMALIZED			STANDARDIZED		
Model	**Pearson**	**MSE**	**Model**	**Pearson**	**MSE**
BBPFT	0.56	0.52	BBPFT	0.56	0.52
BBPFT_25	0.58	0.50	BBPFT_25	0.58	0.50
BBPFT_50	0.57	0.51	BBPFT_50	0.57	0.51
BBPFT_100	**0.59**	**0.50**	BBPFT_100	0.58	0.50
			BBPFT_150	0.58	0.50
NILCFT100	0.53	0.55	NILCFT100	0.53	0.54
NILCFT100_25	0.51	0.56	NILCFT100_25	0.54	0.54
NILCFT100_50	0.51	0.56	NILCFT100_50	0.52	0.55
NILCFT100_100	0.51	0.56	NILCFT100_100	0.53	0.55
			NILCFT100_150	0.53	0.55
NILCW2V100	0.45	0.60	NILCW2V100	0.45	0.61
NILCW2V100_25	0.46	0.60	NILCW2V100_25	0.46	0.60
NILCW2V100_50	0.46	0.60	NILCW2V100_50	0.46	0.60
NILCW2V100_100	0.46	0.60	NILCW2V100_100	0.46	0.60
			NILCW2V100_150	0.47	0.60
NILCFT300	0.49	0.58	NILCFT300	0.49	0.58
NILCFT300_25	0.50	0.57	NILCFT300_25	0.52	0.56
NILCFT300_50	0.50	0.57	NILCFT300_50	0.52	0.56
NILCFT300_100	0.50	0.57	NILCFT300_100	0.52	0.56
			NILCFT300_150	0.52	0.55

Table G.2 – The complete results for the European Portuguese ASSIN track.

Concatenated Models					
NORMALIZED			**STANDARDIZED**		
Model	**Pearson**	**MSE**	**Model**	**Pearson**	**MSE**
BBPFT	0.59	0.79	BBPFT	0.59	0.79
BBPFT_25	0.59	0.79	BBPFT_25	0.59	0.78
BBPFT_50	0.59	0.79	BBPFT_50	0.59	0.79
BBPFT_100	0.59	0.79	**BBPFT_100**	**0.60**	**0.78**
			BBPFT_150	0.59	0.79
NILCFT100	0.52	0.88	NILCFT100	0.53	0.86
NILCFT100_25	0.52	0.88	NILCFT100_25	0.53	0.86
NILCFT100_50	0.52	0.88	NILCFT100_50	0.53	0.86
NILCFT100_100	0.52	0.88	NILCFT100_100	0.54	0.85
			NILCFT100_150	0.53	0.86
NILCW2V100	0.47	0.93	NILCW2V100	0.47	0.93
NILCW2V100_25	0.47	0.93	NILCW2V100_25	0.48	0.93
NILCW2V100_50	0.46	0.94	NILCW2V100_50	0.48	0.93
NILCW2V100_100	0.47	0.94	NILCW2V100_100	0.48	0.92
			NILCW2V100_150	0.48	0.92
NILCFT300	0.50	0.90	NILCFT300	0.50	0.90
NILCFT300_25	0.51	0.90	NILCFT300_25	0.51	0.90
NILCFT300_50	0.51	0.90	NILCFT300_50	0.51	0.90
NILCFT300_100	0.51	0.90	NILCFT300_100	0.51	0.90
			NILCFT300_150	0.51	0.90
Auto-encoded Models					
NORMALIZED			**STANDARDIZED**		
Model	**Pearson**	**MSE**	**Model**	**Pearson**	**MSE**
BBPFT	0.59	0.79	BBPFT	0.59	0.79
BBPFT_25	0.56	0.81	**BBPFT_25**	**0.60**	**0.76**
BBPFT_50	0.57	0.80	BBPFT_50	0.60	0.77
BBPFT_100	0.58	0.79	BBPFT_100	0.60	0.77
			BBPFT_150	0.60	0.77
NILCFT100	0.52	0.88	NILCFT100	0.53	0.86
NILCFT100_25	0.52	0.88	NILCFT100_25	0.55	0.85
NILCFT100_50	0.52	0.88	NILCFT100_50	0.53	0.86
NILCFT100_100	0.51	0.88	NILCFT100_100	0.54	0.85
			NILCFT100_150	0.54	0.86
NILCW2V100	0.47	0.93	NILCW2V100	0.47	0.93
NILCW2V100_25	0.47	0.93	NILCW2V100_25	0.49	0.91
NILCW2V100_50	0.48	0.92	NILCW2V100_50	0.49	0.91
NILCW2V100_100	0.47	0.93	NILCW2V100_100	0.49	0.91
			NILCW2V100_150	0.49	0.91
NILCFT300	0.50	0.90	NILCFT300	0.50	0.90
NILCFT300_25	0.50	0.90	NILCFT300_25	0.49	0.90
NILCFT300_50	0.49	0.90	NILCFT300_50	0.52	0.88
NILCFT300_100	0.50	0.91	NILCFT300_100	0.50	0.89
			NILCFT300_150	0.51	0.88

APPENDIX H – NER FULL RESULTS

Table H.1 – The results of the BBP300 and NILCFT100 concatenated models of the Selective HAREM Track.

Concatenated Models - HAREM SELECTIVE									
NORMALIZED					STANDARDIZED				
Model	Category	Precision	Recall	F1	Model	Category	Precision	Recall	F1
BBPFT300	LOC	0.717	0.704	0.711	BBPFT300	LOC	0.714	0.703	0.709
	ORG	0.617	0.573	0.594		ORG	0.626	0.570	0.597
	PER	0.754	0.634	0.688		PER	0.762	0.635	0.693
	TMP	0.866	0.856	0.861		TMP	0.868	0.853	0.860
	VAL	0.783	0.718	0.749		VAL	0.755	0.736	0.745
	Total	0.733	0.679	0.705		Total	0.734	0.680	0.706
BBPFT_25	LOC	0.701	0.692	0.697	BBPFT_25	LOC	0.709	0.669	0.688
	ORG	0.608	0.559	0.583		ORG	0.632	0.549	0.587
	PER	0.773	0.630	0.694		PER	0.766	0.619	0.685
	TMP	0.889	0.856	0.872		TMP	0.862	0.845	0.853
	VAL	0.780	0.730	0.754		VAL	0.778	0.742	0.760
	Total	0.734	0.673	0.702		Total	0.738	0.661	0.697
BBPFT_50	LOC	0.711	0.707	0.709	BBPFT_50	LOC	0.708	0.684	0.696
	ORG	0.630	0.563	0.595		ORG	0.635	0.538	0.583
	PER	0.777	0.631	0.696		PER	0.765	0.634	0.693
	TMP	0.874	0.859	0.866		TMP	0.860	0.848	0.854
	VAL	0.783	0.718	0.749		VAL	0.785	0.739	0.762
	Total	0.741	0.677	0.708		Total	0.738	0.668	0.701
BBPFT_100	LOC	0.708	0.698	0.703	BBPFT_100	LOC	0.706	0.660	0.682
	ORG	0.636	0.582	0.608		ORG	0.623	0.540	0.578
	PER	0.774	0.641	0.701		PER	0.752	0.596	0.665
	TMP	0.853	0.856	0.855		TMP	0.888	0.853	0.870
	VAL	0.777	0.715	0.744		VAL	0.779	0.736	0.757
	Total	0.737	0.681	0.708		Total	0.735	0.651	0.690
					BBPFT_150	LOC	0.714	0.679	0.696
						ORG	0.629	0.517	0.567
						PER	0.749	0.600	0.667
						TMP	0.892	0.859	0.875
						VAL	0.770	0.727	0.748
						Total	0.738	0.653	0.693
NILCFT100	LOC	0.720	0.692	0.706	NILCFT100	LOC	0.702	0.734	0.718
	ORG	0.632	0.566	0.597		ORG	0.612	0.570	0.590
	PER	0.786	0.650	0.711		PER	0.767	0.669	0.715
	TMP	0.841	0.836	0.839		TMP	0.825	0.836	0.830
	VAL	0.725	0.672	0.698		VAL	0.692	0.690	0.691
	Total	0.737	0.671	0.702		Total	0.716	0.691	0.703
NILCFT100_25	LOC	0.722	0.679	0.700	NILCFT100_25	LOC	0.717	0.710	0.714
	ORG	0.618	0.545	0.579		ORG	0.632	0.552	0.590
	PER	0.780	0.630	0.697		PER	0.761	0.637	0.694
	TMP	0.806	0.797	0.801		TMP	0.841	0.839	0.840
	VAL	0.715	0.653	0.683		VAL	0.771	0.745	0.758
	Total	0.727	0.651	0.687		Total	0.735	0.679	0.706
NILCFT100_50	LOC	0.715	0.663	0.688	NILCFT100_50	LOC	0.724	0.684	0.703
	ORG	0.624	0.538	0.578		ORG	0.592	0.534	0.562
	PER	0.776	0.621	0.690		PER	0.756	0.637	0.692
	TMP	0.816	0.816	0.816		TMP	0.858	0.856	0.857
	VAL	0.747	0.669	0.706		VAL	0.761	0.752	0.756
	Total	0.731	0.646	0.686		Total	0.729	0.670	0.698
NILCFT100_100	LOC	0.718	0.681	0.699	NILCFT100_100	LOC	0.704	0.691	0.697
	ORG	0.637	0.545	0.588		ORG	0.580	0.492	0.532
	PER	0.761	0.615	0.680		PER	0.759	0.632	0.690
	TMP	0.831	0.819	0.825		TMP	0.865	0.853	0.859
	VAL	0.758	0.699	0.727		VAL	0.734	0.736	0.735
	Total	0.733	0.655	0.692		Total	0.720	0.661	0.689
					NILCFT100_150	LOC	0.719	0.691	0.705
						ORG	0.607	0.520	0.560
						PER	0.777	0.647	0.706
						TMP	0.838	0.833	0.836
						VAL	0.748	0.736	0.742
						Total	0.733	0.668	0.699

Table H.2 – The results of the NILCW2V100 and NILCFT300 concatenated models of the Selective HAREM Track.

Concatenated Models - HAREM SELECTIVE									
NORMALIZED					STANDARDIZED				
Model	**Category**	**Precision**	**Recall**	**F1**	**Model**	**Category**	**Precision**	**Recall**	**F1**
NILCW2V100	LOC	0.709	0.701	0.705	**NILCW2V100**	LOC	0.715	0.713	0.714
	ORG	0.661	0.543	0.596		ORG	0.610	0.571	0.590
	PER	0.746	0.605	0.669		PER	0.769	0.668	0.715
	TMP	0.831	0.831	0.831		TMP	0.826	0.842	0.834
	VAL	0.798	0.703	0.747		VAL	0.748	0.730	0.739
	Total	0.736	0.659	0.696		**Total**	**0.727**	**0.690**	**0.708**
NILCW2V100_25	LOC	0.722	0.683	0.702	**NILCW2V100_25**	LOC	0.740	0.702	0.720
	ORG	0.670	0.559	0.610		ORG	0.628	0.545	0.584
	PER	0.755	0.607	0.673		PER	0.760	0.663	0.708
	TMP	0.826	0.816	0.821		TMP	0.863	0.850	0.856
	VAL	0.770	0.678	0.721		VAL	0.784	0.767	0.775
	Total	0.740	0.653	0.694		**Total**	**0.746**	**0.686**	**0.715**
NILCW2V100_50	LOC	0.719	0.688	0.703	**NILCW2V100_50**	LOC	0.725	0.703	0.714
	ORG	0.628	0.515	0.566		ORG	0.636	0.568	0.600
	PER	0.735	0.590	0.654		PER	0.787	0.665	0.721
	TMP	0.816	0.825	0.820		TMP	0.836	0.836	0.836
	VAL	0.778	0.709	0.742		VAL	0.756	0.721	0.738
	Total	0.727	0.645	0.684		**Total**	**0.743**	**0.684**	**0.712**
NILCW2V100_100	LOC	0.712	0.675	0.693	NILCW2V100_100	LOC	0.721	0.665	0.692
	ORG	0.617	0.517	0.562		ORG	0.585	0.497	0.538
	PER	0.749	0.611	0.673		PER	0.750	0.634	0.687
	TMP	0.855	0.819	0.837		TMP	0.844	0.842	0.843
	VAL	0.754	0.669	0.709		VAL	0.767	0.736	0.751
	Total	0.728	0.643	0.683		Total	0.726	0.653	0.687
					NILCW2V100_150	LOC	0.708	0.687	0.697
						ORG	0.587	0.499	0.540
						PER	0.768	0.641	0.699
						TMP	0.839	0.825	0.832
						VAL	0.750	0.709	0.729
						Total	0.724	0.657	0.688
NILCFT300	LOC	0.726	0.710	0.718	**NILCFT300**	LOC	0.728	0.710	0.719
	ORG	0.639	0.556	0.594		ORG	0.628	0.556	0.589
	PER	0.767	0.645	0.700		PER	0.754	0.663	0.705
	TMP	0.841	0.819	0.830		TMP	0.872	0.845	0.858
	VAL	0.760	0.736	0.748		VAL	0.776	0.767	0.772
	Total	0.739	0.678	0.707		**Total**	**0.740**	**0.690**	**0.714**
NILCFT300_25	LOC	0.719	0.700	0.709	**NILCFT300_25**	LOC	0.718	0.692	0.705
	ORG	0.626	0.552	0.587		ORG	0.619	0.554	0.585
	PER	0.758	0.615	0.679		PER	0.762	0.670	0.713
	TMP	0.864	0.842	0.853		TMP	0.878	0.856	0.867
	VAL	0.790	0.727	0.757		VAL	0.779	0.755	0.766
	Total	0.738	0.668	0.701		**Total**	**0.739**	**0.686**	**0.711**
NILCFT300_50	LOC	0.725	0.703	0.714	**NILCFT300_50**	LOC	0.729	0.703	0.716
	ORG	0.626	0.570	0.597		ORG	0.630	0.570	0.598
	PER	0.770	0.618	0.685		PER	0.755	0.664	0.707
	TMP	0.867	0.850	0.859		TMP	0.875	0.850	0.863
	VAL	0.767	0.718	0.742		VAL	0.777	0.761	0.769
	Total	0.740	0.673	0.705		**Total**	**0.741**	**0.691**	**0.715**
NILCFT300_100	LOC	0.722	0.683	0.702	**NILCFT300_100**	LOC	0.731	0.700	0.715
	ORG	0.618	0.549	0.581		ORG	0.619	0.554	0.585
	PER	0.763	0.628	0.689		PER	0.750	0.652	0.698
	TMP	0.872	0.831	0.851		TMP	0.855	0.853	0.854
	VAL	0.775	0.730	0.752		VAL	0.799	0.779	0.789
	Total	0.738	0.664	0.699		**Total**	**0.739**	**0.685**	**0.711**
					NILCFT300_150	LOC	0.717	0.683	0.700
						ORG	0.616	0.549	0.580
						PER	0.754	0.646	0.696
						TMP	0.852	0.845	0.848
						VAL	0.787	0.782	0.785
						Total	0.734	0.677	0.704

Table H.3 – The results of the BBP300 and NILCFT100 auto-encoded models of the Selective HAREM Track.

Auto-encoded - HAREM SELECTIVE									
NORMALIZED					**STANDARDIZED**				
Model	**Category**	**Precision**	**Recall**	**F1**	**Model**	**Category**	**Precision**	**Recall**	**F1**
BBPFT300	LOC	0.717	0.704	0.711	BBPFT300	LOC	0.714	0.703	0.709
	ORG	0.617	0.573	0.594		ORG	0.626	0.570	0.597
	PER	0.754	0.634	0.688		PER	0.762	0.635	0.693
	TMP	0.866	0.856	0.861		TMP	0.868	0.853	0.860
	VAL	0.783	0.718	0.749		VAL	0.755	0.736	0.745
	Total	0.733	0.679	0.705		Total	0.734	0.680	0.706
BBPFT_25	LOC	0.713	0.695	0.704	BBPFT_25	LOC	0.716	0.690	0.703
	ORG	0.624	0.561	0.591		ORG	0.605	0.559	0.581
	PER	0.757	0.558	0.642		PER	0.753	0.554	0.638
	TMP	0.903	0.870	0.886		TMP	0.852	0.859	0.855
	VAL	0.761	0.693	0.726		VAL	0.772	0.748	0.760
	Total	0.737	0.651	0.691		Total	0.728	0.653	0.688
BBPFT_50	LOC	0.701	0.694	0.697	BBPFT_50	LOC	0.698	0.688	0.693
	ORG	0.655	0.536	0.590		ORG	0.608	0.536	0.570
	PER	0.772	0.561	0.650		PER	0.764	0.582	0.661
	TMP	0.867	0.864	0.866		TMP	0.870	0.853	0.862
	VAL	0.809	0.727	0.766		VAL	0.766	0.733	0.749
	Total	0.745	0.650	0.694		Total	0.727	0.654	0.688
BBPFT_100	LOC	0.705	0.700	0.702	BBPFT_100	LOC	0.708	0.684	0.696
	ORG	0.619	0.536	0.575		ORG	0.631	0.526	0.574
	PER	0.738	0.532	0.618		PER	0.729	0.543	0.622
	TMP	0.863	0.839	0.851		TMP	0.854	0.859	0.856
	VAL	0.783	0.730	0.756		VAL	0.775	0.752	0.763
	Total	0.727	0.641	0.681		Total	0.728	0.642	0.682
					BBPFT_150	LOC	0.715	0.671	0.692
						ORG	0.592	0.534	0.562
						PER	0.745	0.574	0.648
						TMP	0.849	0.856	0.852
						VAL	0.755	0.748	0.752
						Total	0.722	0.648	0.683
NILCFT100	LOC	0.720	0.692	0.706	NILCFT100	LOC	0.702	0.734	0.718
	ORG	0.632	0.566	0.597		ORG	0.612	0.570	0.590
	PER	0.786	0.650	0.711		PER	0.767	0.669	0.715
	TMP	0.841	0.836	0.839		TMP	0.825	0.836	0.830
	VAL	0.725	0.672	0.698		VAL	0.692	0.690	0.691
	Total	0.737	0.671	0.702		Total	0.716	0.691	0.703
NILCFT100_25	LOC	0.725	0.670	0.696	NILCFT100_25	LOC	0.701	0.716	0.709
	ORG	0.650	0.515	0.575		ORG	0.610	0.552	0.580
	PER	0.803	0.624	0.702		PER	0.791	0.667	0.723
	TMP	0.831	0.819	0.825		TMP	0.852	0.864	0.858
	VAL	0.759	0.684	0.719		VAL	0.738	0.742	0.740
	Total	0.750	0.647	0.694		Total	0.731	0.691	0.710
NILCFT100_50	LOC	0.713	0.683	0.698	NILCFT100_50	LOC	0.699	0.718	0.709
	ORG	0.619	0.510	0.559		ORG	0.620	0.564	0.591
	PER	0.789	0.640	0.706		PER	0.797	0.663	0.724
	TMP	0.839	0.842	0.841		TMP	0.814	0.831	0.822
	VAL	0.766	0.681	0.721		VAL	0.746	0.730	0.738
	Total	0.739	0.656	0.695		Total	0.729	0.688	0.708
NILCFT100_100	LOC	0.721	0.665	0.692	NILCFT100_100	LOC	0.692	0.714	0.703
	ORG	0.642	0.482	0.550		ORG	0.611	0.534	0.570
	PER	0.791	0.618	0.694		PER	0.776	0.640	0.701
	TMP	0.822	0.811	0.817		TMP	0.832	0.825	0.828
	VAL	0.729	0.660	0.692		VAL	0.726	0.690	0.708
	Total	0.741	0.633	0.683		Total	0.720	0.669	0.694
					NILCFT100_150	LOC	0.690	0.714	0.702
						ORG	0.617	0.536	0.574
						PER	0.783	0.642	0.706
						TMP	0.817	0.848	0.832
						VAL	0.742	0.724	0.733
						Total	0.723	0.676	0.699

Table H.4 – The results of the NILCW2V100 and NILCFT300 auto-encoded models of the Selective HAREM Track.

Auto-encoded - HAREM SELECTIVE									
NORMALIZED					**STANDARDIZED**				
Model	**Category**	**Precision**	**Recall**	**F1**	**Model**	**Category**	**Precision**	**Recall**	**F1**
NILCW2V100	LOC	0.709	0.701	0.705	NILCW2V100	LOC	0.715	0.713	0.714
	ORG	0.661	0.543	0.596		ORG	0.610	0.571	0.590
	PER	0.746	0.605	0.669		PER	0.769	0.668	0.715
	TMP	0.831	0.831	0.831		TMP	0.826	0.842	0.834
	VAL	0.798	0.703	0.747		VAL	0.748	0.730	0.739
	Total	0.736	0.659	0.696		Total	0.727	0.690	0.708
NILCW2V100_25	LOC	0.746	0.683	0.713	NILCW2V100_25	LOC	0.722	0.698	0.710
	ORG	0.654	0.503	0.568		ORG	0.606	0.543	0.573
	PER	0.760	0.611	0.678		PER	0.773	0.662	0.713
	TMP	0.883	0.831	0.856		TMP	0.833	0.833	0.833
	VAL	0.774	0.724	0.748		VAL	0.744	0.730	0.737
	Total	0.755	0.650	0.699		Total	0.731	0.678	0.703
NILCW2V100_50	LOC	0.718	0.696	0.707	**NILCW2V100_50**	LOC	0.702	0.726	0.714
	ORG	0.637	0.499	0.560		ORG	0.596	0.547	0.570
	PER	0.765	0.624	0.687		PER	0.794	0.688	0.737
	TMP	0.848	0.833	0.841		TMP	0.864	0.842	0.853
	VAL	0.784	0.690	0.734		VAL	0.763	0.748	0.755
	Total	0.741	0.653	0.695		**Total**	**0.733**	**0.697**	**0.714**
NILCW2V100_100	LOC	0.723	0.679	0.701	NILCW2V100_100	LOC	0.711	0.703	0.707
	ORG	0.628	0.494	0.553		ORG	0.614	0.564	0.588
	PER	0.783	0.620	0.692		PER	0.772	0.663	0.713
	TMP	0.874	0.825	0.849		TMP	0.851	0.842	0.847
	VAL	0.762	0.727	0.744		VAL	0.756	0.733	0.745
	Total	0.746	0.650	0.695		Total	0.731	0.685	0.707
					NILCW2V100_150	LOC	0.721	0.713	0.717
						ORG	0.613	0.536	0.572
						PER	0.785	0.658	0.716
						TMP	0.845	0.816	0.831
						VAL	0.745	0.733	0.739
						Total	0.736	0.678	0.706
NILCFT300	LOC	0.726	0.710	0.718	**NILCFT300**	LOC	0.728	0.710	0.719
	ORG	0.639	0.556	0.594		ORG	0.628	0.556	0.589
	PER	0.767	0.645	0.700		PER	0.754	0.663	0.705
	TMP	0.841	0.819	0.830		TMP	0.872	0.845	0.858
	VAL	0.760	0.736	0.748		VAL	0.776	0.767	0.772
	Total	0.739	0.678	0.707		**Total**	**0.740**	**0.690**	**0.714**
NILCFT300_25	LOC	0.728	0.662	0.693	NILCFT300_25	LOC	0.719	0.687	0.702
	ORG	0.645	0.536	0.586		ORG	0.598	0.517	0.554
	PER	0.782	0.581	0.667		PER	0.734	0.592	0.655
	TMP	0.869	0.822	0.845		TMP	0.866	0.842	0.854
	VAL	0.766	0.712	0.738		VAL	0.752	0.736	0.744
	Total	0.749	0.640	0.690		Total	0.723	0.651	0.686
NILCFT300_50	LOC	0.753	0.649	0.697	NILCFT300_50	LOC	0.716	0.687	0.701
	ORG	0.663	0.492	0.565		ORG	0.612	0.534	0.571
	PER	0.808	0.607	0.693		PER	0.785	0.616	0.691
	TMP	0.861	0.825	0.843		TMP	0.863	0.839	0.851
	VAL	0.773	0.712	0.741		VAL	0.747	0.770	0.758
	Total	0.769	0.635	0.696		Total	0.737	0.665	0.699
NILCFT300_100	LOC	0.714	0.681	0.697	NILCFT300_100	LOC	0.712	0.690	0.701
	ORG	0.668	0.485	0.562		ORG	0.607	0.517	0.558
	PER	0.791	0.542	0.643		PER	0.770	0.626	0.691
	TMP	0.850	0.848	0.849		TMP	0.849	0.845	0.847
	VAL	0.773	0.712	0.741		VAL	0.755	0.773	0.764
	Total	0.751	0.627	0.683		Total	0.731	0.666	0.697
					NILCFT300_150	LOC	0.719	0.703	0.711
						ORG	0.607	0.547	0.575
						PER	0.773	0.625	0.691
						TMP	0.837	0.828	0.832
						VAL	0.751	0.748	0.750
						Total	0.730	0.671	0.699

Table H.5 – The results of the BBP300 concatenated model for the Total HAREM Track.

Concatenated Models - HAREM TOTAL									
NORMALIZED					STANDARDIZED				
Model	**Category**	**Precision**	**Recall**	**F1**	**Model**	**Category**	**Precision**	**Recall**	**F1**
BBPFT300	ABS	0.285	0.188	0.226	BBPFT300	ABS	0.286	0.193	0.230
	ACO	0.225	0.180	0.200		ACO	0.152	0.140	0.146
	COI	0.396	0.130	0.195		COI	0.493	0.210	0.294
	LOC	0.703	0.696	0.699		LOC	0.708	0.703	0.706
	OBR	0.302	0.138	0.190		OBR	0.309	0.112	0.164
	ORG	0.612	0.579	0.595		ORG	0.616	0.568	0.591
	OTR	0.000	0.000	0.000		OTR	0.000	0.000	0.000
	PER	0.757	0.636	0.691		PER	0.766	0.677	0.718
	TMP	0.851	0.856	0.854		TMP	0.841	0.850	0.846
	VAL	0.747	0.696	0.721		VAL	0.754	0.770	0.762
	Total	0.679	0.585	0.628		Total	0.685	0.602	0.641
BBPFT_25	ABS	0.296	0.198	0.237	BBPFT_25	ABS	0.242	0.157	0.191
	ACO	0.200	0.120	0.150		ACO	0.158	0.120	0.136
	COI	0.382	0.130	0.194		COI	0.362	0.154	0.217
	LOC	0.699	0.698	0.699		LOC	0.696	0.689	0.693
	OBR	0.295	0.122	0.173		OBR	0.227	0.080	0.118
	ORG	0.628	0.579	0.602		ORG	0.623	0.564	0.592
	OTR	0.000	0.000	0.000		OTR	0.000	0.000	0.000
	PER	0.754	0.648	0.697		PER	0.746	0.651	0.695
	TMP	0.870	0.848	0.858		TMP	0.843	0.862	0.852
	VAL	0.757	0.709	0.732		VAL	0.764	0.755	0.759
	Total	0.685	0.587	0.632		Total	0.675	0.586	0.627
BBPFT_50	ABS	0.310	0.223	0.260	BBPFT_50	ABS	0.267	0.178	0.213
	ACO	0.237	0.180	0.205		ACO	0.105	0.080	0.091
	COI	0.396	0.130	0.195		COI	0.373	0.136	0.199
	LOC	0.714	0.709	0.712		LOC	0.698	0.679	0.688
	OBR	0.333	0.138	0.196		OBR	0.238	0.080	0.120
	ORG	0.627	0.593	0.609		ORG	0.594	0.540	0.566
	OTR	0.000	0.000	0.000		OTR	0.000	0.000	0.000
	PER	0.782	0.654	0.713		PER	0.740	0.645	0.689
	TMP	0.850	0.848	0.849		TMP	0.849	0.856	0.852
	VAL	0.766	0.715	0.740		VAL	0.752	0.736	0.744
	Total	0.694	0.597	0.642		Total	0.670	0.575	0.619
BBPFT_100	ABS	0.292	0.178	0.221	BBPFT_100	ABS	0.242	0.162	0.195
	ACO	0.233	0.140	0.175		ACO	0.118	0.080	0.095
	COI	0.465	0.124	0.195		COI	0.359	0.142	0.204
	LOC	0.705	0.710	0.708		LOC	0.705	0.687	0.696
	OBR	0.380	0.186	0.250		OBR	0.259	0.080	0.122
	ORG	0.633	0.571	0.601		ORG	0.618	0.550	0.582
	OTR	0.000	0.000	0.000		OTR	0.000	0.000	0.000
	PER	0.744	0.639	0.687		PER	0.758	0.648	0.699
	TMP	0.840	0.845	0.842		TMP	0.840	0.845	0.842
	VAL	0.757	0.715	0.735		VAL	0.759	0.755	0.757
	Total	0.688	0.589	0.635		Total	0.680	0.580	0.626
					BBPFT_150	ABS	0.275	0.183	0.220
						ACO	0.217	0.200	0.208
						COI	0.438	0.130	0.200
						LOC	0.701	0.689	0.695
						OBR	0.216	0.058	0.092
						ORG	0.605	0.550	0.576
						OTR	0.000	0.000	0.000
						PER	0.746	0.642	0.690
						TMP	0.835	0.842	0.838
						VAL	0.747	0.752	0.749
						Total	0.677	0.579	0.624

Table H.6 – The results of the NILCFT100 concatenated model for the Total HAREM Track.

Concatenated Models - HAREM TOTAL									
NORMALIZED					STANDARDIZED				
Model	Category	Precision	Recall	F1	Model	Category	Precision	Recall	F1
NILCFT100	ABS	0.272	0.127	0.173	NILCFT100	ABS	0.214	0.152	0.178
	ACO	0.391	0.180	0.247		ACO	0.324	0.220	0.262
	COI	0.487	0.117	0.189		COI	0.448	0.161	0.236
	LOC	0.708	0.690	0.699		LOC	0.710	0.725	0.717
	OBR	0.241	0.069	0.107		OBR	0.203	0.069	0.103
	ORG	0.627	0.573	0.599		ORG	0.617	0.587	0.602
	OTR	0.000	0.000	0.000		OTR	0.000	0.000	0.000
	PER	0.759	0.641	0.695		PER	0.770	0.685	0.725
	TMP	0.806	0.811	0.809		TMP	0.846	0.853	0.850
	VAL	0.694	0.641	0.667		VAL	0.718	0.718	0.718
	Total	0.688	0.566	0.621		Total	0.682	0.602	0.640
NILCFT100_25	ABS	0.287	0.168	0.212	NILCFT100_25	ABS	0.197	0.147	0.169
	ACO	0.208	0.100	0.135		ACO	0.300	0.180	0.225
	COI	0.578	0.161	0.251		COI	0.460	0.142	0.217
	LOC	0.703	0.675	0.689		LOC	0.730	0.710	0.720
	OBR	0.241	0.069	0.107		OBR	0.273	0.080	0.124
	ORG	0.596	0.557	0.576		ORG	0.607	0.570	0.588
	OTR	0.000	0.000	0.000		OTR	0.000	0.000	0.000
	PER	0.763	0.665	0.711		PER	0.760	0.673	0.714
	TMP	0.857	0.831	0.844		TMP	0.813	0.825	0.819
	VAL	0.737	0.696	0.716		VAL	0.766	0.773	0.770
	Total	0.688	0.576	0.627		Total	0.686	0.594	0.637
NILCFT100_50	ABS	0.301	0.188	0.231	NILCFT100_50	ABS	0.204	0.147	0.171
	ACO	0.185	0.100	0.130		ACO	0.194	0.120	0.148
	COI	0.485	0.099	0.164		COI	0.489	0.142	0.220
	LOC	0.716	0.702	0.709		LOC	0.712	0.718	0.715
	OBR	0.373	0.117	0.178		OBR	0.244	0.058	0.094
	ORG	0.609	0.557	0.582		ORG	0.598	0.540	0.567
	OTR	0.000	0.000	0.000		OTR	0.000	0.000	0.000
	PER	0.766	0.662	0.710		PER	0.764	0.674	0.716
	TMP	0.835	0.814	0.824		TMP	0.864	0.859	0.861
	VAL	0.725	0.687	0.706		VAL	0.733	0.742	0.738
	Total	0.691	0.580	0.631		Total	0.685	0.590	0.634
NILCFT100_100	ABS	0.221	0.147	0.177	NILCFT100_100	ABS	0.263	0.183	0.216
	ACO	0.273	0.120	0.167		ACO	0.276	0.160	0.203
	COI	0.500	0.111	0.182		COI	0.489	0.142	0.220
	LOC	0.730	0.675	0.701		LOC	0.703	0.696	0.699
	OBR	0.351	0.106	0.163		OBR	0.255	0.064	0.102
	ORG	0.560	0.580	0.570		ORG	0.605	0.513	0.555
	OTR	0.250	0.071	0.111		OTR	0.000	0.000	0.000
	PER	0.754	0.629	0.686		PER	0.745	0.654	0.697
	TMP	0.861	0.822	0.841		TMP	0.867	0.864	0.866
	VAL	0.733	0.706	0.719		VAL	0.725	0.752	0.738
	Total	0.680	0.570	0.620		Total	0.684	0.581	0.628
					NILCFT100_150	ABS	0.248	0.178	0.207
						ACO	0.412	0.140	0.209
						COI	0.346	0.111	0.168
						LOC	0.709	0.691	0.700
						OBR	0.213	0.053	0.085
						ORG	0.573	0.503	0.536
						OTR	0.000	0.000	0.000
						PER	0.743	0.651	0.694
						TMP	0.839	0.842	0.841
						VAL	0.734	0.745	0.740
						Total	0.675	0.571	0.619

Table H.7 – The results of the NILCW2V100 concatenated model for the Total HAREM Track.

Concatenated Models - HAREM TOTAL									
NORMALIZED					STANDARDIZED				
Model	Category	Precision	Recall	F1	Model	Category	Precision	Recall	F1
NILCW2V100	ABS	0.287	0.168	0.212	NILCW2V100	ABS	0.216	0.152	0.179
	ACO	0.294	0.200	0.238		ACO	0.172	0.100	0.127
	COI	0.465	0.124	0.195		COI	0.349	0.179	0.237
	LOC	0.715	0.713	0.714		LOC	0.730	0.721	0.725
	OBR	0.397	0.122	0.187		OBR	0.255	0.075	0.115
	ORG	0.609	0.557	0.582		ORG	0.592	0.573	0.582
	OTR	0.333	0.071	0.118		OTR	0.000	0.000	0.000
	PER	0.746	0.637	0.687		PER	0.749	0.672	0.708
	TMP	0.829	0.819	0.824		TMP	0.846	0.839	0.843
	VAL	0.732	0.678	0.704		VAL	0.726	0.730	0.728
	Total	0.687	0.579	0.628		Total	0.675	0.595	0.633
NILCW2V100_25	ABS	0.216	0.137	0.168	NILCW2V100_25	ABS	0.260	0.173	0.207
	ACO	0.222	0.160	0.186		ACO	0.189	0.140	0.161
	COI	0.378	0.105	0.164		COI	0.448	0.161	0.236
	LOC	0.707	0.681	0.693		LOC	0.723	0.696	0.709
	OBR	0.351	0.138	0.199		OBR	0.317	0.101	0.153
	ORG	0.587	0.566	0.576		ORG	0.605	0.584	0.594
	OTR	0.000	0.000	0.000		OTR	0.000	0.000	0.000
	PER	0.738	0.626	0.678		PER	0.751	0.662	0.704
	TMP	0.840	0.816	0.828		TMP	0.867	0.850	0.859
	VAL	0.762	0.696	0.728		VAL	0.735	0.724	0.730
	Total	0.672	0.569	0.616		Total	0.686	0.592	0.635
NILCW2V100_50	ABS	0.235	0.142	0.177	NILCW2V100_50	ABS	0.264	0.173	0.209
	ACO	0.237	0.180	0.205		ACO	0.350	0.140	0.200
	COI	0.333	0.099	0.152		COI	0.431	0.154	0.227
	LOC	0.705	0.685	0.695		LOC	0.719	0.692	0.705
	OBR	0.358	0.128	0.188		OBR	0.365	0.101	0.158
	ORG	0.593	0.570	0.581		ORG	0.584	0.550	0.567
	OTR	0.250	0.071	0.111		OTR	0.250	0.071	0.111
	PER	0.745	0.634	0.685		PER	0.747	0.680	0.712
	TMP	0.833	0.805	0.819		TMP	0.836	0.848	0.842
	VAL	0.747	0.672	0.708		VAL	0.752	0.752	0.752
	Total	0.674	0.569	0.617		Total	0.685	0.592	0.635
NILCW2V100_100	ABS	0.205	0.137	0.164	NILCW2V100_100	ABS	0.212	0.157	0.181
	ACO	0.125	0.080	0.098		ACO	0.188	0.120	0.146
	COI	0.340	0.099	0.153		COI	0.426	0.161	0.233
	LOC	0.747	0.675	0.709		LOC	0.721	0.709	0.715
	OBR	0.319	0.080	0.128		OBR	0.212	0.058	0.092
	ORG	0.594	0.513	0.551		ORG	0.611	0.549	0.578
	OTR	0.000	0.000	0.000		OTR	0.000	0.000	0.000
	PER	0.732	0.626	0.675		PER	0.757	0.650	0.699
	TMP	0.851	0.825	0.838		TMP	0.832	0.850	0.841
	VAL	0.760	0.641	0.696		VAL	0.737	0.730	0.733
	Total	0.684	0.550	0.610		Total	0.680	0.583	0.628
					NILCW2V100_150	ABS	0.248	0.183	0.211
						ACO	0.220	0.180	0.198
						COI	0.339	0.130	0.188
						LOC	0.708	0.689	0.698
						OBR	0.296	0.085	0.132
						ORG	0.602	0.570	0.585
						OTR	0.000	0.000	0.000
						PER	0.741	0.645	0.689
						TMP	0.868	0.856	0.862
						VAL	0.745	0.755	0.750
						Total	0.675	0.586	0.628

Table H.8 – The results of the NILCFT300 concatenated model for the Total HAREM Track.

Concatenated Models - HAREM TOTAL

NORMALIZED

Model	Category	Precision	Recall	F1
NILCFT300	ABS	0.241	0.193	0.214
	ACO	0.282	0.220	0.247
	COI	0.418	0.142	0.212
	LOC	0.717	0.734	0.725
	OBR	0.333	0.117	0.173
	ORG	0.616	0.579	0.597
	OTR	0.200	0.071	0.105
	PER	0.761	0.651	0.702
	TMP	0.858	0.833	0.845
	VAL	0.745	0.745	0.745
	Total	0.684	0.600	0.639
NILCFT300_25	ABS	0.245	0.173	0.202
	ACO	0.313	0.200	0.244
	COI	0.358	0.148	0.210
	LOC	0.711	0.723	0.717
	OBR	0.232	0.069	0.107
	ORG	0.621	0.566	0.592
	OTR	0.333	0.071	0.118
	PER	0.753	0.647	0.696
	TMP	0.883	0.853	0.868
	VAL	0.752	0.733	0.742
	Total	0.686	0.592	0.636
NILCFT300_50	ABS	0.219	0.162	0.187
	ACO	0.317	0.260	0.286
	COI	0.323	0.130	0.185
	LOC	0.726	0.691	0.708
	OBR	0.242	0.080	0.120
	ORG	0.605	0.573	0.589
	OTR	0.250	0.071	0.111
	PER	0.754	0.651	0.699
	TMP	0.867	0.845	0.856
	VAL	0.741	0.727	0.734
	Total	0.679	0.585	0.629
NILCFT300_100	ABS	0.248	0.168	0.200
	ACO	0.290	0.180	0.222
	COI	0.422	0.117	0.184
	LOC	0.712	0.703	0.708
	OBR	0.260	0.101	0.146
	ORG	0.608	0.564	0.586
	OTR	0.167	0.071	0.100
	PER	0.757	0.630	0.688
	TMP	0.868	0.856	0.862
	VAL	0.757	0.736	0.747
	Total	0.686	0.583	0.630

STANDARDIZED

Model	Category	Precision	Recall	F1
NILCFT300	ABS	0.235	0.198	0.215
	ACO	0.180	0.180	0.180
	COI	0.333	0.161	0.217
	LOC	0.723	0.721	0.722
	OBR	0.238	0.080	0.120
	ORG	0.597	0.577	0.587
	OTR	0.000	0.000	0.000
	PER	0.748	0.663	0.703
	TMP	0.832	0.842	0.837
	VAL	0.746	0.748	0.747
	Total	0.667	0.599	0.631
NILCFT300_25	ABS	0.263	0.203	0.229
	ACO	0.220	0.180	0.198
	COI	0.354	0.173	0.232
	LOC	0.727	0.718	0.723
	OBR	0.258	0.090	0.134
	ORG	0.607	0.579	0.593
	OTR	0.333	0.143	0.200
	PER	0.747	0.675	0.709
	TMP	0.844	0.853	0.848
	VAL	0.771	0.764	0.767
	Total	0.680	0.606	0.641
NILCFT300_50	ABS	0.284	0.223	0.250
	ACO	0.303	0.200	0.241
	COI	0.288	0.142	0.190
	LOC	0.709	0.697	0.703
	OBR	0.253	0.101	0.145
	ORG	0.597	0.568	0.582
	OTR	0.286	0.143	0.191
	PER	0.747	0.662	0.702
	TMP	0.860	0.864	0.862
	VAL	0.733	0.718	0.726
	Total	0.670	0.594	0.630
NILCFT300_100	ABS	0.226	0.183	0.202
	ACO	0.167	0.140	0.152
	COI	0.338	0.154	0.212
	LOC	0.722	0.722	0.722
	OBR	0.226	0.075	0.112
	ORG	0.590	0.543	0.566
	OTR	0.000	0.000	0.000
	PER	0.748	0.667	0.705
	TMP	0.861	0.842	0.851
	VAL	0.748	0.739	0.744
	Total	0.671	0.592	0.629
NILCFT300_150	ABS	0.240	0.178	0.204
	ACO	0.244	0.220	0.232
	COI	0.300	0.130	0.181
	LOC	0.716	0.715	0.715
	OBR	0.139	0.053	0.077
	ORG	0.611	0.557	0.583
	OTR	0.111	0.071	0.087
	PER	0.742	0.661	0.699
	TMP	0.856	0.856	0.856
	VAL	0.721	0.715	0.718
	Total	0.667	0.589	0.626

Table H.9 – The results of the BBP300 auto-encoded model for the Total HAREM Track.

Auto-encoded Models - HAREM TOTAL									
NORMALIZED					**STANDARDIZED**				
Model	**Category**	**Precision**	**Recall**	**F1**	**Model**	**Category**	**Precision**	**Recall**	**F1**
BBPFT300	ABS	0.285	0.188	0.226	BBPFT300	ABS	0.286	0.193	0.230
	ACO	0.225	0.180	0.200		ACO	0.152	0.140	0.146
	COI	0.396	0.130	0.195		COI	0.493	0.210	0.294
	LOC	0.703	0.696	0.699		LOC	0.708	0.703	0.706
	OBR	0.302	0.138	0.190		OBR	0.309	0.112	0.164
	ORG	0.612	0.579	0.595		ORG	0.616	0.568	0.591
	OTR	0.000	0.000	0.000		OTR	0.000	0.000	0.000
	PER	0.757	0.636	0.691		PER	0.766	0.677	0.718
	TMP	0.851	0.856	0.854		TMP	0.841	0.850	0.846
	VAL	0.747	0.696	0.721		VAL	0.754	0.770	0.762
	Total	0.679	0.585	0.628		Total	0.685	0.602	0.641
BBPFT_25	ABS	0.295	0.208	0.244	BBPFT_25	ABS	0.232	0.183	0.205
	ACO	0.345	0.200	0.253		ACO	0.174	0.160	0.167
	COI	0.400	0.086	0.142		COI	0.392	0.124	0.188
	LOC	0.692	0.687	0.689		LOC	0.690	0.673	0.682
	OBR	0.406	0.207	0.275		OBR	0.279	0.090	0.137
	ORG	0.599	0.556	0.576		ORG	0.600	0.550	0.574
	OTR	0.000	0.000	0.000		OTR	0.000	0.000	0.000
	PER	0.727	0.572	0.641		PER	0.752	0.581	0.656
	TMP	0.899	0.853	0.875		TMP	0.864	0.859	0.861
	VAL	0.762	0.696	0.728		VAL	0.749	0.733	0.741
	Total	0.678	0.567	0.617		Total	0.668	0.562	0.611
BBPFT_50	ABS	0.300	0.198	0.239	BBPFT_50	ABS	0.253	0.188	0.216
	ACO	0.276	0.160	0.203		ACO	0.243	0.180	0.207
	COI	0.412	0.086	0.143		COI	0.345	0.124	0.182
	LOC	0.710	0.707	0.708		LOC	0.708	0.679	0.693
	OBR	0.376	0.186	0.249		OBR	0.313	0.112	0.165
	ORG	0.613	0.547	0.578		ORG	0.589	0.556	0.572
	OTR	0.000	0.000	0.000		OTR	0.000	0.000	0.000
	PER	0.748	0.583	0.656		PER	0.736	0.580	0.648
	TMP	0.853	0.850	0.852		TMP	0.849	0.856	0.852
	VAL	0.787	0.724	0.754		VAL	0.746	0.721	0.733
	Total	0.689	0.573	0.626		Total	0.669	0.565	0.612
BBPFT_100	ABS	0.329	0.228	0.270	BBPFT_100	ABS	0.285	0.198	0.234
	ACO	0.333	0.160	0.216		ACO	0.233	0.200	0.215
	COI	0.317	0.080	0.128		COI	0.408	0.124	0.190
	LOC	0.701	0.695	0.698		LOC	0.687	0.682	0.684
	OBR	0.414	0.191	0.262		OBR	0.319	0.122	0.177
	ORG	0.584	0.568	0.576		ORG	0.608	0.561	0.584
	OTR	0.000	0.000	0.000		OTR	0.000	0.000	0.000
	PER	0.748	0.569	0.646		PER	0.763	0.583	0.661
	TMP	0.849	0.845	0.847		TMP	0.849	0.859	0.854
	VAL	0.739	0.693	0.715		VAL	0.771	0.742	0.756
	Total	0.675	0.568	0.617		Total	0.678	0.571	0.619
					BBPFT_150	ABS	0.275	0.208	0.237
						ACO	0.191	0.180	0.186
						COI	0.385	0.124	0.187
						LOC	0.719	0.683	0.700
						OBR	0.200	0.075	0.109
						ORG	0.598	0.536	0.566
						OTR	0.000	0.000	0.000
						PER	0.733	0.578	0.647
						TMP	0.883	0.870	0.876
						VAL	0.748	0.745	0.747
						Total	0.674	0.565	0.615

Table H.10 – The results of the NILCFT100 auto-encoded model for the Total HAREM Track.

Auto-encoded Models - HAREM TOTAL									
NORMALIZED					STANDARDIZED				
Model	Category	Precision	Recall	F1	Model	Category	Precision	Recall	F1
NILCFT100	ABS	0.272	0.127	0.173	NILCFT100	ABS	0.214	0.152	0.178
	ACO	0.391	0.180	0.247		ACO	0.324	0.220	0.262
	COI	0.487	0.117	0.189		COI	0.448	0.161	0.236
	LOC	0.708	0.690	0.699		LOC	0.710	0.725	0.717
	OBR	0.241	0.069	0.107		OBR	0.203	0.069	0.103
	ORG	0.627	0.573	0.599		ORG	0.617	0.587	0.602
	OTR	0.000	0.000	0.000		OTR	0.000	0.000	0.000
	PER	0.759	0.641	0.695		PER	0.770	0.685	0.725
	TMP	0.806	0.811	0.809		TMP	0.846	0.853	0.850
	VAL	0.694	0.641	0.667		VAL	0.718	0.718	0.718
	Total	0.688	0.566	0.621		Total	0.682	0.602	0.640
NILCFT100_25	ABS	0.299	0.147	0.197	NILCFT100_25	ABS	0.227	0.173	0.196
	ACO	0.429	0.180	0.254		ACO	0.270	0.200	0.230
	COI	0.706	0.074	0.134		COI	0.510	0.154	0.237
	LOC	0.729	0.689	0.708		LOC	0.709	0.722	0.715
	OBR	0.432	0.186	0.260		OBR	0.245	0.064	0.101
	ORG	0.624	0.520	0.567		ORG	0.607	0.566	0.586
	OTR	0.000	0.000	0.000		OTR	0.143	0.071	0.095
	PER	0.758	0.650	0.700		PER	0.767	0.670	0.716
	TMP	0.858	0.833	0.845		TMP	0.837	0.842	0.839
	VAL	0.750	0.699	0.724		VAL	0.734	0.745	0.740
	Total	0.710	0.573	0.634		Total	0.682	0.597	0.637
NILCFT100_50	ABS	0.299	0.147	0.197	NILCFT100_50	ABS	0.260	0.173	0.207
	ACO	0.409	0.180	0.250		ACO	0.235	0.160	0.191
	COI	0.619	0.080	0.142		COI	0.658	0.154	0.250
	LOC	0.711	0.681	0.695		LOC	0.686	0.715	0.700
	OBR	0.404	0.112	0.175		OBR	0.288	0.090	0.138
	ORG	0.589	0.526	0.556		ORG	0.601	0.557	0.578
	OTR	0.333	0.071	0.118		OTR	0.250	0.071	0.111
	PER	0.785	0.661	0.717		PER	0.775	0.663	0.715
	TMP	0.808	0.819	0.814		TMP	0.843	0.850	0.847
	VAL	0.755	0.699	0.726		VAL	0.755	0.755	0.755
	Total	0.701	0.569	0.628		Total	0.686	0.595	0.637
NILCFT100_100	ABS	0.300	0.107	0.157	NILCFT100_100	ABS	0.270	0.188	0.222
	ACO	0.242	0.160	0.193		ACO	0.167	0.180	0.173
	COI	0.632	0.074	0.133		COI	0.571	0.124	0.203
	LOC	0.703	0.681	0.692		LOC	0.705	0.721	0.713
	OBR	0.341	0.080	0.129		OBR	0.333	0.080	0.129
	ORG	0.610	0.520	0.561		ORG	0.598	0.550	0.573
	OTR	0.500	0.071	0.125		OTR	0.167	0.071	0.100
	PER	0.789	0.640	0.706		PER	0.766	0.663	0.711
	TMP	0.863	0.822	0.842		TMP	0.857	0.845	0.851
	VAL	0.743	0.709	0.725		VAL	0.726	0.724	0.725
	Total	0.709	0.560	0.626		Total	0.683	0.591	0.634
					NILCFT100_150	ABS	0.228	0.157	0.186
						ACO	0.346	0.180	0.237
						COI	0.537	0.136	0.217
						LOC	0.712	0.704	0.708
						OBR	0.327	0.085	0.135
						ORG	0.617	0.573	0.594
						OTR	0.250	0.071	0.111
						PER	0.770	0.678	0.721
						TMP	0.863	0.853	0.858
						VAL	0.713	0.724	0.718
						Total	0.693	0.594	0.639

Table H.11 – The results of the NILCW2V100 auto-encoded model for the Total HAREM Track.

Auto-encoded Models - HAREM TOTAL									
NORMALIZED					STANDARDIZED				
Model	Category	Precision	Recall	F1	Model	Category	Precision	Recall	F1
NILCW2V100	ABS	0.287	0.168	0.212	NILCW2V100	ABS	0.216	0.152	0.179
	ACO	0.294	0.200	0.238		ACO	0.172	0.100	0.127
	COI	0.465	0.124	0.195		COI	0.349	0.179	0.237
	LOC	0.715	0.713	0.714		LOC	0.730	0.721	0.725
	OBR	0.397	0.122	0.187		OBR	0.255	0.075	0.115
	ORG	0.609	0.557	0.582		ORG	0.592	0.573	0.582
	OTR	0.333	0.071	0.118		OTR	0.000	0.000	0.000
	PER	0.746	0.637	0.687		PER	0.749	0.672	0.708
	TMP	0.829	0.819	0.824		TMP	0.846	0.839	0.843
	VAL	0.732	0.678	0.704		VAL	0.726	0.730	0.728
	Total	0.687	0.579	0.628		Total	0.675	0.595	0.633
NILCW2V100_25	ABS	0.228	0.168	0.193	NILCW2V100_25	ABS	0.149	0.112	0.128
	ACO	0.345	0.200	0.253		ACO	0.256	0.200	0.225
	COI	0.375	0.093	0.149		COI	0.368	0.154	0.217
	LOC	0.733	0.704	0.718		LOC	0.717	0.718	0.718
	OBR	0.409	0.144	0.213		OBR	0.241	0.069	0.107
	ORG	0.606	0.513	0.556		ORG	0.586	0.570	0.578
	OTR	0.000	0.000	0.000		OTR	1.000	0.071	0.133
	PER	0.748	0.628	0.683		PER	0.770	0.680	0.722
	TMP	0.865	0.836	0.851		TMP	0.849	0.842	0.845
	VAL	0.733	0.699	0.716		VAL	0.748	0.758	0.753
	Total	0.690	0.570	0.624		Total	0.676	0.597	0.634
NILCW2V100_50	ABS	0.226	0.193	0.208	NILCW2V100_50	ABS	0.199	0.147	0.169
	ACO	0.370	0.200	0.260		ACO	0.250	0.140	0.180
	COI	0.333	0.093	0.145		COI	0.510	0.154	0.237
	LOC	0.702	0.720	0.711		LOC	0.702	0.713	0.707
	OBR	0.380	0.186	0.250		OBR	0.279	0.090	0.137
	ORG	0.638	0.571	0.603		ORG	0.613	0.564	0.588
	OTR	0.333	0.071	0.118		OTR	1.000	0.071	0.133
	PER	0.750	0.621	0.680		PER	0.750	0.654	0.699
	TMP	0.850	0.831	0.840		TMP	0.851	0.839	0.845
	VAL	0.728	0.696	0.712		VAL	0.741	0.730	0.736
	Total	0.677	0.585	0.628		Total	0.679	0.588	0.631
NILCW2V100_100	ABS	0.156	0.076	0.102	NILCW2V100_100	ABS	0.210	0.152	0.177
	ACO	0.235	0.160	0.191		ACO	0.191	0.180	0.186
	COI	0.421	0.049	0.088		COI	0.433	0.161	0.234
	LOC	0.728	0.652	0.688		LOC	0.709	0.702	0.705
	OBR	0.434	0.122	0.191		OBR	0.261	0.064	0.103
	ORG	0.615	0.485	0.542		ORG	0.594	0.561	0.577
	OTR	0.500	0.071	0.125		OTR	0.250	0.071	0.111
	PER	0.756	0.620	0.682		PER	0.740	0.690	0.714
	TMP	0.870	0.831	0.850		TMP	0.816	0.825	0.820
	VAL	0.751	0.703	0.726		VAL	0.722	0.733	0.728
	Total	0.702	0.543	0.612		Total	0.667	0.592	0.627
					NILCW2V100_150	ABS	0.187	0.132	0.155
						ACO	0.297	0.220	0.253
						COI	0.415	0.136	0.205
						LOC	0.711	0.728	0.720
						OBR	0.419	0.096	0.156
						ORG	0.595	0.534	0.563
						OTR	0.000	0.000	0.000
						PER	0.765	0.662	0.710
						TMP	0.844	0.842	0.843
						VAL	0.724	0.724	0.724
						Total	0.682	0.588	0.631

Table H.12 – The results of the NILCFT300 auto-encoded model for the Total HAREM Track.

Auto-encoded Models - HAREM TOTAL									
NORMALIZED					**STANDARDIZED**				
Model	**Category**	**Precision**	**Recall**	**F1**	**Model**	**Category**	**Precision**	**Recall**	**F1**
NILCFT300	ABS	0.241	0.193	0.214	NILCFT300	ABS	0.235	0.198	0.215
	ACO	0.282	0.220	0.247		ACO	0.180	0.180	0.180
	COI	0.418	0.142	0.212		COI	0.333	0.161	0.217
	LOC	0.717	0.734	0.725		LOC	0.723	0.721	0.722
	OBR	0.333	0.117	0.173		OBR	0.238	0.080	0.120
	ORG	0.616	0.579	0.597		ORG	0.597	0.577	0.587
	OTR	0.200	0.071	0.105		OTR	0.000	0.000	0.000
	PER	0.761	0.651	0.702		PER	0.748	0.663	0.703
	TMP	0.858	0.833	0.845		TMP	0.832	0.842	0.837
	VAL	0.745	0.745	0.745		VAL	0.746	0.748	0.747
	Total	0.684	0.600	0.639		Total	0.667	0.599	0.631
NILCFT300_25	ABS	0.218	0.208	0.213	NILCFT300_25	ABS	0.265	0.198	0.227
	ACO	0.306	0.220	0.256		ACO	0.321	0.180	0.231
	COI	0.486	0.105	0.173		COI	0.333	0.142	0.199
	LOC	0.730	0.698	0.714		LOC	0.720	0.702	0.711
	OBR	0.388	0.165	0.231		OBR	0.349	0.122	0.181
	ORG	0.625	0.543	0.581		ORG	0.584	0.529	0.555
	OTR	0.333	0.071	0.118		OTR	0.250	0.071	0.111
	PER	0.757	0.609	0.675		PER	0.727	0.619	0.668
	TMP	0.860	0.831	0.845		TMP	0.881	0.853	0.867
	VAL	0.758	0.712	0.734		VAL	0.734	0.727	0.730
	Total	0.686	0.575	0.625		Total	0.674	0.577	0.622
NILCFT300_50	ABS	0.263	0.178	0.212	NILCFT300_50	ABS	0.206	0.173	0.188
	ACO	0.212	0.140	0.169		ACO	0.182	0.160	0.170
	COI	0.552	0.099	0.168		COI	0.316	0.111	0.164
	LOC	0.744	0.671	0.706		LOC	0.728	0.710	0.719
	OBR	0.420	0.154	0.226		OBR	0.280	0.075	0.118
	ORG	0.638	0.515	0.570		ORG	0.597	0.552	0.574
	OTR	0.500	0.071	0.125		OTR	0.500	0.071	0.125
	PER	0.772	0.667	0.715		PER	0.766	0.629	0.690
	TMP	0.869	0.825	0.846		TMP	0.852	0.845	0.848
	VAL	0.758	0.758	0.758		VAL	0.739	0.755	0.747
	Total	0.712	0.577	0.637		Total	0.678	0.581	0.626
NILCFT300_100	ABS	0.285	0.168	0.211	NILCFT300_100	ABS	0.264	0.218	0.239
	ACO	0.302	0.260	0.280		ACO	0.250	0.180	0.209
	COI	0.517	0.093	0.157		COI	0.371	0.142	0.205
	LOC	0.726	0.673	0.699		LOC	0.727	0.704	0.715
	OBR	0.255	0.069	0.109		OBR	0.323	0.112	0.166
	ORG	0.622	0.526	0.570		ORG	0.633	0.545	0.586
	OTR	0.333	0.071	0.118		OTR	0.200	0.071	0.105
	PER	0.770	0.594	0.671		PER	0.761	0.648	0.700
	TMP	0.902	0.828	0.863		TMP	0.875	0.853	0.864
	VAL	0.756	0.712	0.733		VAL	0.737	0.758	0.747
	Total	0.706	0.555	0.621		Total	0.690	0.591	0.637
					NILCFT300_150	ABS	0.240	0.188	0.211
						ACO	0.238	0.200	0.217
						COI	0.347	0.105	0.161
						LOC	0.733	0.709	0.721
						OBR	0.288	0.101	0.150
						ORG	0.611	0.550	0.579
						OTR	0.000	0.000	0.000
						PER	0.757	0.641	0.694
						TMP	0.851	0.839	0.845
						VAL	0.749	0.752	0.750
						Total	0.684	0.585	0.631

Table H.13 – The results of the PetroVecFT concatenated model for the GeoCorpus task.

Concatenated Models									
NORMALIZED					STANDARDIZED				
Model	Category	Precision	Recall	F1	Model	Category	Precision	Recall	F1
PetroVecFT	baciaSedimentar	0.738	0.730	0.733	PetroVecFT	baciaSedimentar	0.816	0.835	0.824
	contextoGeologicoDeBacia	0.742	0.754	0.747		contextoGeologicoDeBacia	0.796	0.825	0.810
	epoca	0.832	0.724	0.774		epoca	0.870	0.772	0.818
	idade	0.814	0.700	0.753		idade	0.764	0.722	0.741
	magmaticas	0.768	0.805	0.786		magmaticas	0.839	0.817	0.827
	metamorficas	0.821	0.739	0.776		metamorficas	0.823	0.824	0.823
	periodo	0.813	0.667	0.732		periodo	0.870	0.740	0.799
	sedimentaresCarbonaticas	0.794	0.835	0.814		sedimentaresCarbonaticas	0.820	0.896	0.856
	sedimentaresSiliciclasticas	0.829	0.847	0.838		sedimentaresSiliciclasticas	0.845	0.847	0.846
	unidadeEstratigrafica	0.767	0.795	0.780		unidadeEstratigrafica	0.817	0.870	0.843
	total	0.792	0.763	0.777		total	0.827	0.811	0.818
PetroVecFT_25	baciaSedimentar	0.770	0.794	0.780	PetroVecFT_25	baciaSedimentar	0.803	0.812	0.807
	contextoGeologicoDeBacia	0.810	0.599	0.664		contextoGeologicoDeBacia	0.804	0.785	0.790
	epoca	0.790	0.682	0.732		epoca	0.869	0.726	0.784
	idade	0.812	0.639	0.710		idade	0.834	0.736	0.779
	magmaticas	0.814	0.685	0.736		magmaticas	0.819	0.817	0.813
	metamorficas	0.892	0.693	0.771		metamorficas	0.843	0.788	0.808
	periodo	0.827	0.663	0.736		periodo	0.843	0.745	0.789
	sedimentaresCarbonaticas	0.773	0.705	0.727		sedimentaresCarbonaticas	0.803	0.818	0.809
	sedimentaresSiliciclasticas	0.827	0.811	0.818		sedimentaresSiliciclasticas	0.826	0.856	0.841
	unidadeEstratigrafica	0.781	0.780	0.780		unidadeEstratigrafica	0.820	0.818	0.817
	total	0.805	0.713	0.754		total	0.825	0.790	0.807
PetroVecFT_50	baciaSedimentar	0.796	0.793	0.794	PetroVecFT_50	baciaSedimentar	0.851	0.817	0.833
	contextoGeologicoDeBacia	0.805	0.703	0.750		contextoGeologicoDeBacia	0.816	0.810	0.813
	epoca	0.837	0.724	0.775		epoca	0.913	0.807	0.857
	idade	0.831	0.663	0.736		idade	0.838	0.766	0.799
	magmaticas	0.815	0.756	0.782		magmaticas	0.855	0.823	0.838
	metamorficas	0.839	0.778	0.807		metamorficas	0.823	0.877	0.849
	periodo	0.874	0.701	0.777		periodo	0.899	0.766	0.826
	sedimentaresCarbonaticas	0.795	0.819	0.807		sedimentaresCarbonaticas	0.788	0.852	0.819
	sedimentaresSiliciclasticas	0.842	0.833	0.837		sedimentaresSiliciclasticas	0.861	0.889	0.874
	unidadeEstratigrafica	0.791	0.805	0.798		unidadeEstratigrafica	0.783	0.838	0.807
	total	0.822	0.758	0.789		total	0.843	0.824	0.833
PetroVecFT_100	baciaSedimentar	0.764	0.783	0.774	PetroVecFT_100	baciaSedimentar	0.848	0.854	0.851
	contextoGeologicoDeBacia	0.813	0.744	0.775		contextoGeologicoDeBacia	0.861	0.857	0.859
	epoca	0.824	0.692	0.752		epoca	0.887	0.792	0.836
	idade	0.799	0.710	0.751		idade	0.835	0.729	0.778
	magmaticas	0.784	0.760	0.771		magmaticas	0.824	0.824	0.823
	metamorficas	0.829	0.681	0.737		metamorficas	0.901	0.872	0.886
	periodo	0.816	0.656	0.723		periodo	0.867	0.783	0.822
	sedimentaresCarbonaticas	0.815	0.825	0.818		sedimentaresCarbonaticas	0.834	0.868	0.851
	sedimentaresSiliciclasticas	0.845	0.837	0.841		sedimentaresSiliciclasticas	0.887	0.888	0.887
	unidadeEstratigrafica	0.777	0.765	0.771		unidadeEstratigrafica	0.846	0.881	0.863
	total	0.806	0.748	0.776		total	0.861	0.835	0.848
					PetroVecFT_150	baciaSedimentar	0.802	0.838	0.818
						contextoGeologicoDeBacia	0.823	0.831	0.826
						epoca	0.926	0.867	0.895
						idade	0.795	0.763	0.779
						magmaticas	0.879	0.841	0.859
						metamorficas	0.846	0.832	0.838
						periodo	0.896	0.814	0.852
						sedimentaresCarbonaticas	0.833	0.883	0.857
						sedimentaresSiliciclasticas	0.877	0.871	0.874
						unidadeEstratigrafica	0.887	0.873	0.879
						total	0.860	0.841	0.850

Table H.14 – The results of the PetroVecHybridFT concatenated model for the GeoCorpus task.

Concatenated Models									
NORMALIZED					STANDARDIZED				
Model	**Category**	**Precision**	**Recall**	**F1**	**Model**	**Category**	**Precision**	**Recall**	**F1**
PetroVecHybridFT	baciaSedimentar	0.728	0.749	0.738	PetroVecHybridFT	baciaSedimentar	0.806	0.804	0.804
	contextoGeologicoDeBacia	0.787	0.651	0.712		contextoGeologicoDeBacia	0.760	0.817	0.786
	epoca	0.789	0.708	0.745		epoca	0.893	0.870	0.881
	idade	0.807	0.718	0.758		idade	0.766	0.750	0.757
	magmaticas	0.749	0.751	0.750		magmaticas	0.859	0.839	0.848
	metamorficas	0.801	0.650	0.716		metamorficas	0.783	0.807	0.794
	periodo	0.797	0.698	0.744		periodo	0.905	0.812	0.855
	sedimentaresCarbonaticas	0.778	0.773	0.775		sedimentaresCarbonaticas	0.740	0.818	0.777
	sedimentaresSiliciclasticas	0.802	0.829	0.815		sedimentaresSiliciclasticas	0.862	0.850	0.856
	unidadeEstratigrafica	0.780	0.750	0.764		unidadeEstratigrafica	0.821	0.862	0.841
	total	0.783	0.735	0.758		total	0.827	0.826	0.827
PetroVecHybridFT-25	baciaSedimentar	0.736	0.741	0.738	PetroVecHybridFT_25	baciaSedimentar	0.825	0.806	0.815
	contextoGeologicoDeBacia	0.778	0.643	0.685		contextoGeologicoDeBacia	0.780	0.851	0.811
	epoca	0.804	0.610	0.679		epoca	0.864	0.760	0.808
	idade	0.775	0.677	0.721		idade	0.824	0.731	0.774
	magmaticas	0.768	0.672	0.712		magmaticas	0.848	0.834	0.840
	metamorficas	0.730	0.621	0.670		metamorficas	0.855	0.864	0.859
	periodo	0.729	0.621	0.670		periodo	0.833	0.730	0.778
	sedimentaresCarbonaticas	0.680	0.711	0.694		sedimentaresCarbonaticas	0.824	0.855	0.839
	sedimentaresSiliciclasticas	0.795	0.815	0.805		sedimentaresSiliciclasticas	0.867	0.882	0.874
	unidadeEstratigrafica	0.780	0.772	0.775		unidadeEstratigrafica	0.845	0.867	0.855
	total	0.778	0.699	0.734		total	0.838	0.819	0.828
PetroVecHybridFT_50	baciaSedimentar	0.727	0.729	0.728	PetroVecHybridFT_50	baciaSedimentar	0.832	0.848	0.840
	contextoGeologicoDeBacia	0.795	0.706	0.748		contextoGeologicoDeBacia	0.827	0.835	0.831
	epoca	0.820	0.708	0.759		epoca	0.882	0.795	0.836
	idade	0.824	0.707	0.760		idade	0.859	0.730	0.788
	magmaticas	0.816	0.747	0.780		magmaticas	0.835	0.825	0.828
	metamorficas	0.798	0.652	0.716		metamorficas	0.888	0.876	0.882
	periodo	0.827	0.703	0.756		periodo	0.885	0.749	0.809
	sedimentaresCarbonaticas	0.798	0.782	0.789		sedimentaresCarbonaticas	0.810	0.876	0.841
	sedimentaresSiliciclasticas	0.808	0.832	0.820		sedimentaresSiliciclasticas	0.897	0.894	0.895
	unidadeEstratigrafica	0.784	0.792	0.788		unidadeEstratigrafica	0.832	0.851	0.841
	total	0.800	0.744	0.771		total	0.858	0.827	0.842
PetroVecHybridFT_100	baciaSedimentar	0.772	0.760	0.765	PetroVecHybridFT_100	baciaSedimentar	0.671	0.695	0.683
	contextoGeologicoDeBacia	0.812	0.801	0.804		contextoGeologicoDeBacia	0.624	0.271	0.307
	epoca	0.869	0.743	0.801		epoca	0.733	0.326	0.398
	idade	0.838	0.717	0.770		idade	0.474	0.248	0.280
	magmaticas	0.839	0.795	0.817		magmaticas	0.369	0.241	0.271
	metamorficas	0.875	0.801	0.835		metamorficas	0.429	0.321	0.355
	periodo	0.855	0.718	0.779		periodo	0.638	0.432	0.503
	sedimentaresCarbonaticas	0.814	0.847	0.830		sedimentaresCarbonaticas	0.212	0.224	0.218
	sedimentaresSiliciclasticas	0.873	0.873	0.873		sedimentaresSiliciclasticas	0.746	0.522	0.586
	unidadeEstratigrafica	0.822	0.861	0.840		unidadeEstratigrafica	0.783	0.487	0.565
	total	0.836	0.795	0.815		total	0.692	0.392	0.469
					PetroVecHybridFT_150	baciaSedimentar	0.798	0.796	0.797
						contextoGeologicoDeBacia	0.801	0.824	0.812
						epoca	0.894	0.794	0.841
						idade	0.842	0.736	0.784
						magmaticas	0.841	0.838	0.839
						metamorficas	0.851	0.849	0.849
						periodo	0.884	0.798	0.838
						sedimentaresCarbonaticas	0.859	0.858	0.858
						sedimentaresSiliciclasticas	0.877	0.883	0.880
						unidadeEstratigrafica	0.859	0.875	0.867
						total	0.852	0.827	0.839

Table H.15 – The results of the PetroVecW2V concatenated model for the GeoCorpus task.

Concatenated Models									
NORMALIZED					STANDARDIZED				
Model	**Category**	**Precision**	**Recall**	**F1**	**Model**	**Category**	**Precision**	**Recall**	**F1**
PetroVecW2V	baciaSedimentar	0.750	0.761	0.755	PetroVecW2V	baciaSedimentar	0.774	0.771	0.773
	contextoGeologicoDeBacia	0.785	0.768	0.776		contextoGeologicoDeBacia	0.748	0.827	0.785
	epoca	0.792	0.663	0.721		epoca	0.867	0.749	0.804
	idade	0.744	0.694	0.716		idade	0.832	0.723	0.773
	magmaticas	0.810	0.773	0.791		magmaticas	0.808	0.841	0.824
	metamorficas	0.850	0.779	0.813		metamorficas	0.830	0.837	0.832
	periodo	0.839	0.666	0.741		periodo	0.843	0.751	0.794
	sedimentaresCarbonaticas	0.847	0.846	0.846		sedimentaresCarbonaticas	0.796	0.846	0.820
	sedimentaresSiliciclasticas	0.841	0.842	0.842		sedimentaresSiliciclasticas	0.852	0.869	0.860
	unidadeEstratigrafica	0.792	0.856	0.823		unidadeEstratigrafica	0.813	0.866	0.839
	total	0.803	0.766	0.784		total	0.817	0.810	0.813
PetroVecW2V_25	baciaSedimentar	0.812	0.862	0.836	PetroVecW2V_25	baciaSedimentar	0.647	0.689	0.667
	contextoGeologicoDeBacia	0.781	0.765	0.770		contextoGeologicoDeBacia	0.643	0.389	0.418
	epoca	0.887	0.825	0.854		epoca	0.579	0.261	0.300
	idade	0.803	0.731	0.763		idade	0.338	0.304	0.320
	magmaticas	0.919	0.767	0.835		magmaticas	0.386	0.335	0.358
	metamorficas	0.747	0.824	0.783		metamorficas	0.381	0.334	0.354
	periodo	0.858	0.772	0.809		periodo	0.562	0.463	0.504
	sedimentaresCarbonaticas	0.799	0.829	0.814		sedimentaresCarbonaticas	0.393	0.357	0.374
	sedimentaresSiliciclasticas	0.831	0.831	0.830		sedimentaresSiliciclasticas	0.725	0.553	0.609
	unidadeEstratigrafica	0.849	0.780	0.813		unidadeEstratigrafica	0.636	0.671	0.649
	total	0.831	0.796	0.813		total	0.635	0.451	0.508
PetroVecW2V_50	baciaSedimentar	0.824	0.808	0.815	PetroVecW2V_50	baciaSedimentar	0.790	0.768	0.778
	contextoGeologicoDeBacia	0.806	0.734	0.768		contextoGeologicoDeBacia	0.789	0.774	0.781
	epoca	0.805	0.645	0.715		epoca	0.799	0.703	0.747
	idade	0.753	0.625	0.682		idade	0.799	0.648	0.714
	magmaticas	0.789	0.733	0.759		magmaticas	0.845	0.755	0.797
	metamorficas	0.771	0.774	0.771		metamorficas	0.788	0.727	0.754
	periodo	0.879	0.664	0.756		periodo	0.849	0.682	0.756
	sedimentaresCarbonaticas	0.811	0.778	0.793		sedimentaresCarbonaticas	0.832	0.845	0.837
	sedimentaresSiliciclasticas	0.845	0.808	0.826		sedimentaresSiliciclasticas	0.871	0.845	0.858
	unidadeEstratigrafica	0.810	0.827	0.818		unidadeEstratigrafica	0.843	0.817	0.829
	total	0.812	0.739	0.774		total	0.824	0.759	0.790
PetroVecW2V_100	baciaSedimentar	0.771	0.770	0.770	PetroVecW2V_100	baciaSedimentar	0.803	0.843	0.822
	contextoGeologicoDeBacia	0.793	0.822	0.806		contextoGeologicoDeBacia	0.785	0.739	0.761
	epoca	0.745	0.678	0.707		epoca	0.846	0.703	0.766
	idade	0.756	0.702	0.727		idade	0.801	0.640	0.709
	magmaticas	0.818	0.747	0.779		magmaticas	0.773	0.775	0.770
	metamorficas	0.858	0.782	0.817		metamorficas	0.786	0.757	0.767
	periodo	0.862	0.646	0.738		periodo	0.888	0.715	0.792
	sedimentaresCarbonaticas	0.844	0.817	0.830		sedimentaresCarbonaticas	0.801	0.826	0.813
	sedimentaresSiliciclasticas	0.847	0.850	0.848		sedimentaresSiliciclasticas	0.844	0.848	0.846
	unidadeEstratigrafica	0.818	0.830	0.824		unidadeEstratigrafica	0.818	0.810	0.813
	total	0.806	0.767	0.786		total	0.816	0.767	0.790
					PetroVecW2V_150	baciaSedimentar	0.782	0.802	0.791
						contextoGeologicoDeBacia	0.793	0.787	0.789
						epoca	0.903	0.751	0.819
						idade	0.817	0.720	0.764
						magmaticas	0.857	0.796	0.825
						metamorficas	0.837	0.820	0.827
						periodo	0.863	0.724	0.787
						sedimentaresCarbonaticas	0.829	0.880	0.854
						sedimentaresSiliciclasticas	0.880	0.870	0.875
						unidadeEstratigrafica	0.841	0.838	0.838
						total	0.841	0.799	0.819

Table H.16 – The results of the PetroVecHybridW2V concatenated model for the GeoCorpus task.

Concatenated Models									
NORMALIZED					STANDARDIZED				
Model	**Category**	**Precision**	**Recall**	**F1**	**Model**	**Category**	**Precision**	**Recall**	**F1**
PetroVecHybridW2V100	baciaSedimentar	0.760	0.789	0.774	PetroVecHybridW2V100	baciaSedimentar	0.756	0.755	0.755
	contextoGeologicoDeBacia	0.800	0.749	0.774		contextoGeologicoDeBacia	0.768	0.827	0.796
	epoca	0.794	0.660	0.720		epoca	0.852	0.771	0.809
	idade	0.741	0.703	0.721		idade	0.806	0.744	0.773
	magmaticas	0.776	0.755	0.764		magmaticas	0.788	0.846	0.815
	metamorficas	0.818	0.758	0.786		metamorficas	0.781	0.784	0.782
	periodo	0.863	0.641	0.735		periodo	0.853	0.735	0.790
	sedimentaresCarbonaticas	0.806	0.776	0.791		sedimentaresCarbonaticas	0.847	0.840	0.843
	sedimentaresSiliciclasticas	0.829	0.834	0.832		sedimentaresSiliciclasticas	0.839	0.871	0.855
	unidadeEstratigrafica	0.790	0.834	0.812		unidadeEstratigrafica	0.840	0.867	0.854
	total	0.795	0.753	0.773		total	0.814	0.808	0.811
PetroVecHybridW2V_25	baciaSedimentar	0.672	0.714	0.692	PetroVecHybridW2V_25	baciaSedimentar	0.850	0.858	0.854
	contextoGeologicoDeBacia	0.677	0.676	0.668		contextoGeologicoDeBacia	0.827	0.838	0.832
	epoca	0.763	0.515	0.606		epoca	0.888	0.775	0.827
	idade	0.713	0.583	0.596		idade	0.821	0.739	0.778
	magmaticas	0.789	0.644	0.699		magmaticas	0.823	0.834	0.828
	metamorficas	0.828	0.680	0.735		metamorficas	0.859	0.876	0.867
	periodo	0.851	0.564	0.677		periodo	0.881	0.716	0.790
	sedimentaresCarbonaticas	0.783	0.665	0.706		sedimentaresCarbonaticas	0.848	0.875	0.861
	sedimentaresSiliciclasticas	0.821	0.793	0.805		sedimentaresSiliciclasticas	0.897	0.896	0.897
	unidadeEstratigrafica	0.664	0.725	0.691		unidadeEstratigrafica	0.812	0.845	0.828
	total	0.740	0.662	0.694		total	0.852	0.822	0.837
PetroVecHybridW2V_50	baciaSedimentar	0.807	0.843	0.824	PetroVecHybridW2V_50	baciaSedimentar	0.815	0.811	0.813
	contextoGeologicoDeBacia	0.775	0.778	0.776		contextoGeologicoDeBacia	0.799	0.830	0.814
	epoca	0.836	0.697	0.760		epoca	0.892	0.774	0.829
	idade	0.816	0.673	0.737		idade	0.827	0.723	0.771
	magmaticas	0.836	0.788	0.811		magmaticas	0.792	0.822	0.806
	metamorficas	0.762	0.801	0.778		metamorficas	0.818	0.803	0.809
	periodo	0.891	0.702	0.785		periodo	0.879	0.774	0.823
	sedimentaresCarbonaticas	0.804	0.829	0.815		sedimentaresCarbonaticas	0.845	0.851	0.848
	sedimentaresSiliciclasticas	0.827	0.825	0.826		sedimentaresSiliciclasticas	0.865	0.876	0.870
	unidadeEstratigrafica	0.805	0.846	0.824		unidadeEstratigrafica	0.831	0.866	0.848
	total	0.817	0.776	0.796		total	0.836	0.815	0.826
PetroVecHybridW2V_100	baciaSedimentar	0.802	0.826	0.814	PetroVecHybridW2V_100	baciaSedimentar	0.817	0.804	0.810
	contextoGeologicoDeBacia	0.811	0.811	0.810		contextoGeologicoDeBacia	0.811	0.845	0.827
	epoca	0.849	0.712	0.773		epoca	0.920	0.786	0.847
	idade	0.815	0.725	0.762		idade	0.834	0.724	0.775
	magmaticas	0.842	0.802	0.816		magmaticas	0.811	0.829	0.819
	metamorficas	0.861	0.847	0.850		metamorficas	0.822	0.797	0.808
	periodo	0.915	0.707	0.796		periodo	0.874	0.780	0.824
	sedimentaresCarbonaticas	0.829	0.851	0.835		sedimentaresCarbonaticas	0.844	0.856	0.850
	sedimentaresSiliciclasticas	0.866	0.863	0.864		sedimentaresSiliciclasticas	0.877	0.886	0.881
	unidadeEstratigrafica	0.793	0.842	0.814		unidadeEstratigrafica	0.844	0.856	0.850
	total	0.837	0.796	0.815		total	0.847	0.819	0.833
					PetroVecHybridW2V_150	baciaSedimentar	0.828	0.818	0.823
						contextoGeologicoDeBacia	0.804	0.835	0.818
						epoca	0.900	0.782	0.837
						idade	0.829	0.748	0.786
						magmaticas	0.835	0.828	0.831
						metamorficas	0.809	0.816	0.812
						periodo	0.879	0.769	0.820
						sedimentaresCarbonaticas	0.845	0.844	0.844
						sedimentaresSiliciclasticas	0.883	0.896	0.889
						unidadeEstratigrafica	0.866	0.884	0.875
						total	0.851	0.827	0.838

Table H.17 – The results of the PetroVecFT auto-encoded model for the GeoCorpus task.

Auto-encoded Models									
NORMALIZED					STANDARDIZED				
Model	**Category**	**Precision**	**Recall**	**F1**	**Model**	**Category**	**Precision**	**Recall**	**F1**
	baciaSedimentar	0.738	0.730	0.733		baciaSedimentar	0.816	0.835	0.824
	contextoGeologicoDeBacia	0.742	0.754	0.747		contextoGeologicoDeBacia	0.796	0.825	0.810
	epoca	0.832	0.724	0.774		epoca	0.870	0.772	0.818
	idade	0.814	0.700	0.753		idade	0.764	0.722	0.741
	magmaticas	0.768	0.805	0.786		magmaticas	0.839	0.817	0.827
PetroVecFT	metamorficas	0.821	0.739	0.776	PetroVecFT	metamorficas	0.823	0.824	0.823
	periodo	0.813	0.667	0.732		periodo	0.870	0.740	0.799
	sedimentaresCarbonaticas	0.794	0.835	0.814		sedimentaresCarbonaticas	0.820	0.896	0.856
	sedimentaresSiliciclasticas	0.829	0.847	0.838		sedimentaresSiliciclasticas	0.845	0.847	0.846
	unidadeEstratigrafica	0.767	0.795	0.780		unidadeEstratigrafica	0.817	0.870	0.843
	total	0.792	0.763	0.777		total	0.827	0.811	0.818
	baciaSedimentar	0.742	0.765	0.753		baciaSedimentar	0.810	0.810	0.810
	contextoGeologicoDeBacia	0.762	0.716	0.737		contextoGeologicoDeBacia	0.762	0.869	0.812
	epoca	0.848	0.738	0.788		epoca	0.881	0.739	0.803
	idade	0.826	0.718	0.767		idade	0.817	0.719	0.764
	magmaticas	0.769	0.840	0.802		magmaticas	0.795	0.831	0.813
PetroVecFT_25	metamorficas	0.843	0.783	0.810	PetroVecFT_25	metamorficas	0.812	0.853	0.832
	periodo	0.848	0.721	0.779		periodo	0.829	0.754	0.789
	sedimentaresCarbonaticas	0.759	0.837	0.796		sedimentaresCarbonaticas	0.786	0.875	0.828
	sedimentaresSiliciclasticas	0.819	0.839	0.828		sedimentaresSiliciclasticas	0.879	0.881	0.880
	unidadeEstratigrafica	0.728	0.794	0.760		unidadeEstratigrafica	0.815	0.865	0.839
	total	0.791	0.775	0.783		total	0.822	0.819	0.820
	baciaSedimentar	0.796	0.843	0.819		baciaSedimentar	0.832	0.839	0.835
	contextoGeologicoDeBacia	0.810	0.736	0.771		contextoGeologicoDeBacia	0.771	0.840	0.804
	epoca	0.840	0.720	0.775		epoca	0.874	0.776	0.822
	idade	0.767	0.751	0.757		idade	0.780	0.753	0.766
	magmaticas	0.783	0.820	0.801		magmaticas	0.810	0.848	0.828
PetroVecFT_50	metamorficas	0.878	0.835	0.855	PetroVecFT_50	metamorficas	0.815	0.889	0.850
	periodo	0.824	0.679	0.744		periodo	0.844	0.784	0.812
	sedimentaresCarbonaticas	0.765	0.824	0.793		sedimentaresCarbonaticas	0.787	0.858	0.821
	sedimentaresSiliciclasticas	0.827	0.846	0.836		sedimentaresSiliciclasticas	0.859	0.858	0.858
	unidadeEstratigrafica	0.806	0.792	0.798		unidadeEstratigrafica	0.807	0.882	0.843
	total	0.808	0.780	0.794		total	0.820	0.830	0.825
	baciaSedimentar	0.771	0.817	0.793		baciaSedimentar	0.807	0.841	0.823
	contextoGeologicoDeBacia	0.783	0.725	0.753		contextoGeologicoDeBacia	0.792	0.846	0.817
	epoca	0.828	0.717	0.768		epoca	0.876	0.751	0.808
	idade	0.742	0.738	0.739		idade	0.783	0.749	0.765
	magmaticas	0.772	0.787	0.779		magmaticas	0.833	0.833	0.833
PetroVecFT_100	metamorficas	0.854	0.797	0.824	PetroVecFT_100	metamorficas	0.801	0.852	0.825
	periodo	0.841	0.689	0.758		periodo	0.817	0.747	0.780
	sedimentaresCarbonaticas	0.808	0.821	0.814		sedimentaresCarbonaticas	0.791	0.878	0.832
	sedimentaresSiliciclasticas	0.832	0.840	0.836		sedimentaresSiliciclasticas	0.856	0.867	0.861
	unidadeEstratigrafica	0.792	0.814	0.803		unidadeEstratigrafica	0.827	0.866	0.845
	total	0.801	0.773	0.787		total	0.822	0.820	0.821
						baciaSedimentar	0.817	0.845	0.830
						contextoGeologicoDeBacia	0.793	0.850	0.821
						epoca	0.884	0.776	0.826
						idade	0.724	0.743	0.731
						magmaticas	0.809	0.827	0.818
					PetroVecFT_150	metamorficas	0.831	0.862	0.845
						periodo	0.834	0.736	0.782
						sedimentaresCarbonaticas	0.832	0.904	0.867
						sedimentaresSiliciclasticas	0.869	0.865	0.867
						unidadeEstratigrafica	0.816	0.894	0.853
						total	0.822	0.826	0.824

Table H.18 – The results of the PetroVecHybridFT auto-encoded model for the GeoCorpus task.

Auto-encoded Models									
NORMALIZED					STANDARDIZED				
Model	**Category**	**Precision**	**Recall**	**F1**	**Model**	**Category**	**Precision**	**Recall**	**F1**
	baciaSedimentar	0.728	0.749	0.738		baciaSedimentar	0.806	0.804	0.804
	contextoGeologicoDeBacia	0.787	0.651	0.712		contextoGeologicoDeBacia	0.760	0.817	0.786
	epoca	0.789	0.708	0.745		epoca	0.893	0.870	0.881
	idade	0.807	0.718	0.758		idade	0.766	0.750	0.757
	magmaticas	0.749	0.751	0.750		magmaticas	0.859	0.839	0.848
PetroVecHybridFT	metamorficas	0.801	0.650	0.716	PetroVecHybridFT	metamorficas	0.783	0.807	0.794
	periodo	0.797	0.698	0.744		periodo	0.905	0.812	0.855
	sedimentaresCarbonaticas	0.778	0.773	0.775		sedimentaresCarbonaticas	0.740	0.818	0.777
	sedimentaresSiliciclasticas	0.802	0.829	0.815		sedimentaresSiliciclasticas	0.862	0.850	0.856
	unidadeEstratigrafica	0.780	0.750	0.764		unidadeEstratigrafica	0.821	0.862	0.841
	total	0.783	0.735	0.758		total	0.827	0.826	0.827
	baciaSedimentar	0.775	0.770	0.772		baciaSedimentar	0.809	0.819	0.814
	contextoGeologicoDeBacia	0.738	0.730	0.734		contextoGeologicoDeBacia	0.754	0.853	0.800
	epoca	0.825	0.783	0.802		epoca	0.883	0.829	0.855
	idade	0.761	0.704	0.730		idade	0.809	0.746	0.775
	magmaticas	0.729	0.746	0.737		magmaticas	0.814	0.849	0.831
PetroVecHybridFT-25	metamorficas	0.778	0.717	0.746	PetroVecHybridFT_25	metamorficas	0.808	0.831	0.819
	periodo	0.843	0.710	0.770		periodo	0.875	0.811	0.841
	sedimentaresCarbonaticas	0.680	0.779	0.726		sedimentaresCarbonaticas	0.780	0.868	0.822
	sedimentaresSiliciclasticas	0.816	0.815	0.816		sedimentaresSiliciclasticas	0.858	0.883	0.870
	unidadeEstratigrafica	0.755	0.804	0.778		unidadeEstratigrafica	0.806	0.862	0.833
	total	0.776	0.760	0.768		total	0.822	0.836	0.829
	baciaSedimentar	0.771	0.771	0.771		baciaSedimentar	0.814	0.801	0.808
	contextoGeologicoDeBacia	0.795	0.763	0.778		contextoGeologicoDeBacia	0.767	0.851	0.806
	epoca	0.817	0.756	0.785		epoca	0.889	0.854	0.871
	idade	0.785	0.793	0.788		idade	0.774	0.760	0.766
	magmaticas	0.766	0.856	0.808		magmaticas	0.830	0.837	0.833
PetroVecHybridFT_50	metamorficas	0.847	0.794	0.819	PetroVecHybridFT_50	metamorficas	0.808	0.868	0.837
	periodo	0.857	0.695	0.767		periodo	0.891	0.802	0.843
	sedimentaresCarbonaticas	0.722	0.783	0.751		sedimentaresCarbonaticas	0.720	0.814	0.764
	sedimentaresSiliciclasticas	0.817	0.856	0.836		sedimentaresSiliciclasticas	0.860	0.864	0.862
	unidadeEstratigrafica	0.753	0.816	0.783		unidadeEstratigrafica	0.805	0.871	0.837
	total	0.792	0.791	0.792		total	0.821	0.833	0.827
	baciaSedimentar	0.729	0.754	0.741		baciaSedimentar	0.784	0.780	0.782
	contextoGeologicoDeBacia	0.759	0.745	0.751		contextoGeologicoDeBacia	0.754	0.863	0.805
	epoca	0.834	0.720	0.772		epoca	0.884	0.817	0.849
	idade	0.811	0.690	0.745		idade	0.809	0.768	0.787
	magmaticas	0.731	0.804	0.765		magmaticas	0.832	0.870	0.850
PetroVecHybridFT_100	metamorficas	0.809	0.691	0.744	PetroVecHybridFT_100	metamorficas	0.873	0.827	0.848
	periodo	0.801	0.729	0.763		periodo	0.861	0.789	0.823
	sedimentaresCarbonaticas	0.782	0.817	0.799		sedimentaresCarbonaticas	0.787	0.846	0.815
	sedimentaresSiliciclasticas	0.823	0.863	0.842		sedimentaresSiliciclasticas	0.842	0.876	0.859
	unidadeEstratigrafica	0.758	0.777	0.767		unidadeEstratigrafica	0.821	0.882	0.850
	total	0.784	0.764	0.774		total	0.823	0.833	0.828
						baciaSedimentar	0.775	0.770	0.772
						contextoGeologicoDeBacia	0.727	0.849	0.783
						epoca	0.899	0.823	0.859
						idade	0.790	0.747	0.767
						magmaticas	0.815	0.842	0.828
					PetroVecHybridFT_150	metamorficas	0.836	0.825	0.830
						periodo	0.866	0.809	0.836
						sedimentaresCarbonaticas	0.808	0.850	0.829
						sedimentaresSiliciclasticas	0.851	0.876	0.863
						unidadeEstratigrafica	0.816	0.871	0.843
						total	0.818	0.829	0.823

Table H.19 – The results of the PetroVecW2V auto-encoded model for the GeoCorpus task.

Auto-encoded Models									
NORMALIZED					STANDARDIZED				
Model	Category	Precision	Recall	F1	Model	Category	Precision	Recall	F1
PetroVecW2V	baciaSedimentar	0.750	0.761	0.755	PetroVecW2V	baciaSedimentar	0.774	0.771	0.773
	contextoGeologicoDeBacia	0.785	0.768	0.776		contextoGeologicoDeBacia	0.748	0.827	0.785
	epoca	0.792	0.663	0.721		epoca	0.867	0.749	0.804
	idade	0.744	0.694	0.716		idade	0.832	0.723	0.773
	magmaticas	0.810	0.773	0.791		magmaticas	0.808	0.841	0.824
	metamorficas	0.850	0.779	0.813		metamorficas	0.830	0.837	0.832
	periodo	0.839	0.666	0.741		periodo	0.843	0.751	0.794
	sedimentaresCarbonaticas	0.847	0.846	0.846		sedimentaresCarbonaticas	0.796	0.846	0.820
	sedimentaresSiliciclasticas	0.841	0.842	0.842		sedimentaresSiliciclasticas	0.852	0.869	0.860
	unidadeEstratigrafica	0.792	0.856	0.823		unidadeEstratigrafica	0.813	0.866	0.839
	total	0.803	0.766	0.784		total	0.817	0.810	0.813
PetroVecW2V_25	baciaSedimentar	0.781	0.764	0.772	PetroVecW2V_25	baciaSedimentar	0.818	0.794	0.805
	contextoGeologicoDeBacia	0.794	0.773	0.782		contextoGeologicoDeBacia	0.757	0.847	0.799
	epoca	0.810	0.644	0.717		epoca	0.839	0.753	0.793
	idade	0.748	0.661	0.701		idade	0.815	0.755	0.782
	magmaticas	0.777	0.781	0.779		magmaticas	0.844	0.844	0.844
	metamorficas	0.771	0.782	0.775		metamorficas	0.810	0.850	0.829
	periodo	0.850	0.686	0.759		periodo	0.851	0.728	0.784
	sedimentaresCarbonaticas	0.811	0.850	0.830		sedimentaresCarbonaticas	0.814	0.861	0.837
	sedimentaresSiliciclasticas	0.844	0.823	0.833		sedimentaresSiliciclasticas	0.876	0.891	0.883
	unidadeEstratigrafica	0.784	0.851	0.816		unidadeEstratigrafica	0.814	0.845	0.829
	total	0.800	0.760	0.780		total	0.825	0.817	0.821
PetroVecW2V_50	baciaSedimentar	0.793	0.803	0.797	PetroVecW2V_50	baciaSedimentar	0.791	0.800	0.795
	contextoGeologicoDeBacia	0.782	0.756	0.768		contextoGeologicoDeBacia	0.751	0.831	0.788
	epoca	0.791	0.625	0.698		epoca	0.873	0.747	0.805
	idade	0.729	0.668	0.696		idade	0.812	0.744	0.775
	magmaticas	0.770	0.757	0.764		magmaticas	0.791	0.840	0.815
	metamorficas	0.765	0.769	0.765		metamorficas	0.847	0.830	0.838
	periodo	0.826	0.658	0.733		periodo	0.816	0.733	0.771
	sedimentaresCarbonaticas	0.809	0.845	0.827		sedimentaresCarbonaticas	0.816	0.876	0.845
	sedimentaresSiliciclasticas	0.857	0.826	0.841		sedimentaresSiliciclasticas	0.869	0.878	0.873
	unidadeEstratigrafica	0.790	0.860	0.823		unidadeEstratigrafica	0.811	0.873	0.841
	total	0.796	0.756	0.775		total	0.818	0.816	0.817
PetroVecW2V_100	baciaSedimentar	0.808	0.804	0.806	PetroVecW2V_100	baciaSedimentar	0.817	0.832	0.824
	contextoGeologicoDeBacia	0.782	0.748	0.764		contextoGeologicoDeBacia	0.768	0.831	0.798
	epoca	0.808	0.562	0.660		epoca	0.869	0.745	0.802
	idade	0.717	0.646	0.679		idade	0.813	0.738	0.772
	magmaticas	0.767	0.776	0.771		magmaticas	0.825	0.845	0.835
	metamorficas	0.760	0.769	0.764		metamorficas	0.816	0.901	0.856
	periodo	0.836	0.610	0.705		periodo	0.872	0.747	0.804
	sedimentaresCarbonaticas	0.789	0.806	0.797		sedimentaresCarbonaticas	0.806	0.850	0.827
	sedimentaresSiliciclasticas	0.842	0.809	0.825		sedimentaresSiliciclasticas	0.871	0.880	0.875
	unidadeEstratigrafica	0.786	0.847	0.815		unidadeEstratigrafica	0.797	0.863	0.829
	total	0.792	0.737	0.763		total	0.827	0.820	0.824
					PetroVecW2V_150	baciaSedimentar	0.828	0.820	0.824
						contextoGeologicoDeBacia	0.753	0.815	0.783
						epoca	0.876	0.724	0.792
						idade	0.840	0.715	0.771
						magmaticas	0.798	0.818	0.808
						metamorficas	0.770	0.820	0.794
						periodo	0.857	0.744	0.796
						sedimentaresCarbonaticas	0.801	0.884	0.840
						sedimentaresSiliciclasticas	0.857	0.865	0.861
						unidadeEstratigrafica	0.796	0.841	0.818
						total	0.820	0.803	0.811

Table H.20 – The results of the PetroVecHybridW2V auto-encoded model for the GeoCorpus task.

Auto-encoded Models									
NORMALIZED					STANDARDIZED				
Model	**Category**	**Precision**	**Recall**	**F1**	**Model**	**Category**	**Precision**	**Recall**	**F1**
PetroVecHybridW2V	baciaSedimentar	0.760	0.789	0.774	PetroVecHybridW2V	baciaSedimentar	0.756	0.755	0.755
	contextoGeologicoDeBacia	0.800	0.749	0.774		contextoGeologicoDeBacia	0.768	0.827	0.796
	epoca	0.794	0.660	0.720		epoca	0.852	0.771	0.809
	idade	0.741	0.703	0.721		idade	0.806	0.744	0.773
	magmaticas	0.776	0.755	0.764		magmaticas	0.788	0.846	0.815
	metamorficas	0.818	0.758	0.786		metamorficas	0.781	0.784	0.782
	periodo	0.863	0.641	0.735		periodo	0.853	0.735	0.790
	sedimentaresCarbonaticas	0.806	0.776	0.791		sedimentaresCarbonaticas	0.847	0.840	0.843
	sedimentaresSiliciclasticas	0.829	0.834	0.832		sedimentaresSiliciclasticas	0.839	0.871	0.855
	unidadeEstratigrafica	0.790	0.834	0.812		unidadeEstratigrafica	0.840	0.867	0.854
	total	0.795	0.753	0.773		total	0.814	0.808	0.811
PetroVecHybridW2V_25	baciaSedimentar	0.796	0.777	0.786	PetroVecHybridW2V_25	baciaSedimentar	0.786	0.821	0.803
	contextoGeologicoDeBacia	0.762	0.759	0.760		contextoGeologicoDeBacia	0.803	0.849	0.825
	epoca	0.776	0.668	0.718		epoca	0.831	0.802	0.815
	idade	0.751	0.638	0.689		idade	0.747	0.746	0.745
	magmaticas	0.768	0.755	0.761		magmaticas	0.806	0.841	0.823
	metamorficas	0.711	0.716	0.712		metamorficas	0.826	0.833	0.830
	periodo	0.849	0.691	0.762		periodo	0.867	0.765	0.812
	sedimentaresCarbonaticas	0.776	0.785	0.781		sedimentaresCarbonaticas	0.866	0.890	0.878
	sedimentaresSiliciclasticas	0.790	0.788	0.789		sedimentaresSiliciclasticas	0.849	0.866	0.857
	unidadeEstratigrafica	0.754	0.842	0.796		unidadeEstratigrafica	0.807	0.890	0.847
	total	0.775	0.744	0.759		total	0.818	0.829	0.823
					PetroVecHybridW2V_50	baciaSedimentar	0.781	0.775	0.778
						contextoGeologicoDeBacia	0.774	0.856	0.812
						epoca	0.873	0.789	0.829
						idade	0.806	0.773	0.788
						magmaticas	0.799	0.841	0.819
						metamorficas	0.787	0.847	0.816
						periodo	0.867	0.744	0.800
						sedimentaresCarbonaticas	0.801	0.814	0.808
						sedimentaresSiliciclasticas	0.867	0.877	0.872
						unidadeEstratigrafica	0.807	0.875	0.839
						total	0.820	0.821	0.820
PetroVecHybridW2V_100	baciaSedimentar	0.786	0.801	0.793	PetroVecHybridW2V_100	baciaSedimentar	0.784	0.774	0.779
	contextoGeologicoDeBacia	0.756	0.776	0.766		contextoGeologicoDeBacia	0.753	0.842	0.795
	epoca	0.786	0.710	0.745		epoca	0.847	0.794	0.819
	idade	0.746	0.681	0.712		idade	0.786	0.769	0.776
	magmaticas	0.758	0.793	0.774		magmaticas	0.797	0.872	0.833
	metamorficas	0.773	0.780	0.776		metamorficas	0.813	0.859	0.835
	periodo	0.817	0.711	0.760		periodo	0.857	0.768	0.809
	sedimentaresCarbonaticas	0.810	0.829	0.819		sedimentaresCarbonaticas	0.825	0.857	0.841
	sedimentaresSiliciclasticas	0.820	0.810	0.815		sedimentaresSiliciclasticas	0.839	0.859	0.849
	unidadeEstratigrafica	0.753	0.848	0.798		unidadeEstratigrafica	0.829	0.879	0.853
	total	0.780	0.772	0.776		total	0.813	0.826	0.819
					PetroVecHybridW2V_150	baciaSedimentar	0.781	0.782	0.782
						contextoGeologicoDeBacia	0.733	0.807	0.768
						epoca	0.830	0.804	0.816
						idade	0.805	0.731	0.766
						magmaticas	0.781	0.851	0.814
						metamorficas	0.797	0.852	0.822
						periodo	0.862	0.766	0.811
						sedimentaresCarbonaticas	0.817	0.852	0.834
						sedimentaresSiliciclasticas	0.853	0.849	0.851
						unidadeEstratigrafica	0.786	0.876	0.828
						total	0.806	0.815	0.811

Pontifícia Universidade Católica do Rio Grande do Sul
Pró-Reitoria de Graduação
Av. Ipiranga, 6681 - Prédio 1 - 3º. andar
Porto Alegre - RS - Brasil
Fone: (51) 3320-3500 - Fax: (51) 3339-1564
E-mail: prograd@pucrs.br
Site: www.pucrs.br