自然语言中的否定和不确定性检测

项目报告

姓名：*刘希文*

学校：中国人民大学

提交日期：2024.12.18

摘要

本报告探讨了在自然语言处理中检测否定和不确定性的方法，结合了基于规则和深度学习的技术。通过对不同方法的比较和分析，提出了在医疗文本中有效识别否定和不确定性线索的解决方案。

目录

1. 引言

本项目的主要目标是提高在医疗报告文本中否定和不确定性线索的检测能力，这在改善临床决策和辅助疾病预测与诊断方面具有重大意义。项目结合了基于规则和深度学习技术来实现这一目标。

基于规则的方法涉及创建预定义的语言规则和模式，专门设计用于识别和提取否定和不确定性线索的实例。另一方面，深度学习方法利用神经网络模型。这些模型使用标记过的数据进行训练，其中输入的文本样本与指示否定和不确定性线索的存在及其范围的注释配对。

2. 数据集描述

在我们的项目中，我们使用了一个包含否定和不确定性样本文本和注释的JSON文件（文本内容基本上是医疗记录）。这个数据集被分成两个主要子集：训练集和验证集。训练集占整个数据集的70%，其余的30%被分配给验证集。这种划分确保我们有足够的数据用于训练模型，同时也留出了一个单独的数据集用于评估其性能。

JSON文件包含各种文本样本，这些样本可以是句子或段落，以及相应的注释，指示每个文本中否定和不确定性的存在。这些注释使我们能够准确地标记和分类文本中的否定和不确定性实例。
注：数据集的主要语言为西班牙语，由于本项目探讨的是语言的内在逻辑，与语言本身无关，故是何语言并不影响。另：选择此数据集的原因是数据组织格式高度简洁有序，几乎无需预处理步骤；

---

3. 基于规则的方法

检测过程被分为两个不同的步骤：提示词检测和范围检测。每个步骤将在专门的部分中分别解释，评估是在第一次诊断上进行的。

3.1 提示词检测

提示词检测涉及检查JSON文件以识别NEG和UNC标签，这些标签表明相关词的存在。这些词被提取并存入一个字典中，连同它们对应的标签。

```
{'keyword': 'no', 'category': NEG}
{'keyword': 'sospechar', 'category': UNC}
```

在创建包含相关词的词元和它们对应的标签的字典后，我们遍历输入文本的所有标记，检查它们是否出现在字典中。如果一个标记在字典中找到，它被存储，允许我们将其分类为线索。

```
for token in doc:
    if token.lemma_ in dict:
        negation_indices.add(token.i)
```

3.2 范围检测

为了确定最有效的范围检测方法，我们采用了三种不同的方法。每种方法都单独评估其性能，并确定最佳方法。

3.2.1 固定大小窗口

这种初始方法是最基本的方法。它涉及创建一个固定大小的窗口，其中大小可以调整。在我们的实现中，我们使用了一个大小为4的窗口，这意味着范围是由检测到线索后的随后四个词形成的。

实际上，这种基本方法可能在所有情况下表现不佳。其中一个局限性是它可能在范围内包含太多的词，这可能导致噪声或错误的解释。

```
Predicted scope: ['hijos', 'tiene', 'un', 'hermano']
Real scope: ['hijos', 'O', 'O', 'O']
```

相反的情况也可能发生，即在范围内考虑的词不够。这可能导致错过重要的上下文信息或未能捕获句子的完整含义。

```
Predicted scope: ['alteraciones', 'en', 'el', 'contenido', 'o', 'o']
Real scope: ['alteraciones', 'en', 'el', 'contenido', 'del', 'pensamiento']
```

理想的方法是如果所有范围都有相同的大小，但在这种情况下并非如此。下图显示了获得的定量结果。

```
True Positives: 27
True Negatives: 0
False Positives: 5
False Negatives: 18
+-----------+----------+
| Measure   | Score    |
| Precision | 0.84     |
| Recall    | 0.60     |
| F1-score  | 0.70     |
| Accuracy  | 0.54     |
+-----------+----------+
```

图1：定量结果

3.2.2 查找模式：名词+形容词

我们设计的第二种方法在前一种方法的基础上进行了改进。不是考虑线索后的固定数量的词，而是扩展范围直到我们遇到一个名词后跟一个形容词。换句话说，范围包括检测到的线索，并继续直到遇到这种特定的名词-形容词模式。

然而，这种方法仍然面临与前一种方法类似的挑战。在某些情况下，它可能包括过多的词，因为实际的范围可能不遵循名词后跟形容词的模式。因此，该方法继续添加词直到它识别出这样的模式，可能导致范围超出必要。

```
Predicted scope: ['hijos,, '.', 'tiene', 'un', 'hermano', 'con', 'el', 'que', 'tiene', 'contacto', 'er
```

```
Real scope : [ ' hijos ' , 'o ' , 'o ' , 'o ' , 'o ' , 'o ' , 'o ' , 'o ' , 'o ' , 'o ' , 'o ' ,
'o ' , 'o ' , 'o ' , 'o ' , 'o ' , 'o ']
```

实际上，第二种方法也可能未能包括所有必要的词在范围内。这可能是由于多种因素造成的，比如遇到一个名词后跟两个形容词而不是一个，遇到连词引入额外信息，或者遇到其他偏离特定名词形容词模式的语言结构。这些可能性突出了准确确定范围的挑战和需要进一步改进方法的必要性。

```
Predicted scope: ['alergias', 'mediamentosas', 'o']
Real scope: ['alergias', 'mediamentosas', 'conocidas']
```

尽管提到了限制，第二种方法在某些情况下仍然可以取得成功的结果。

```
Predicted scope: ['alteraciones', 'en', 'la', 'sensopercepcion', 'ni', 'otras', 'alteraciones',
Real scope: ['alteraciones', 'en', 'la', 'sensopercepcion', 'ni', 'otras', 'alteraciones', 'dent
```

下图显示了获得的定量结果。

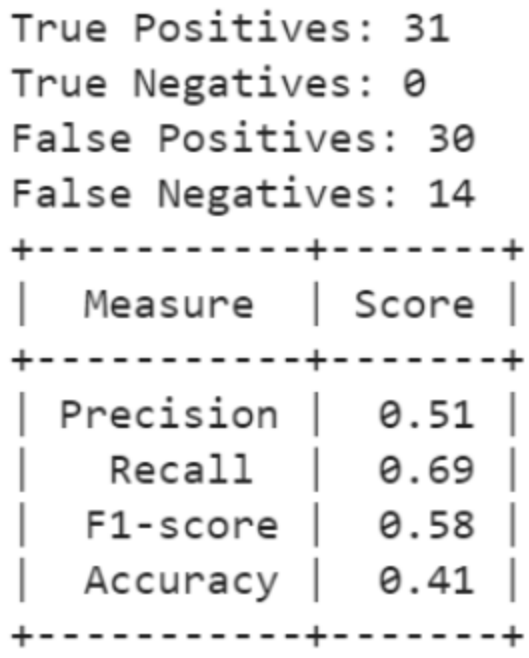

图2：定量结果

结果比前一种方法更糟。假阳性的出现是一个重大挑战，导致整体准确性下降。这突出了进一步改进和优化以减少假阳性和提高线索检测和范围识别精度的必要性。

3.2.3 引入依存树

我们使用Spacy Stanza库引入依存树，以增强受否定和不确定性线索影响的词的检测。通过分析句子中词之间的关系，我们旨在获得更全面的上下文影响理解。

定量评估证实了引入依存树的积极影响。

```
True Negatives: 1
True Positives: 54
False Positives: 5
False Negatives: 28
+-----------+-------+
| Measure | Score |
| Precision | 0.92 |
| Recall | 0.66 |
| F1-score | 0.77 |
| Accuracy | 0.62 |
+-----------+-------+
```

图3：定量报告

定性分析揭示了包含依存树的有希望的结果。细化的范围更好地捕获了预期的上下文，允许更准确地识别受否定和不确定性线索影响的词。这里我们有一个系统性能的例子：

```
no valorables . - tc abdominal : glandula pancreatica de pequeño tamaño , atrofica , con lipomatosis
difusa , sin identificar se lesiones focales ni
dilatacion significativa de el conducto pancreatico .
ureterohidronefrosis bilateral secundaria a globo vesical , observando se una vejiga de paredes trabeculadas
probablemente en relacion a patologia prostatica . evolucio clinica a su llegada a urgencias estable ,
afebril , destacando a la exploracion fisica sequedad mucosa . electrocardiograma en el
que destacan t negativas en di y avl y
d2 sin disponer de ecgs previos y equilibrio
acido-base con acidosis metabolica e hiperglucemia &gt ;
750mg/dl con cetonas altas . bajo la sospecha de cetoacidosis
diabetica se inicia sueroterapia con reposicion de
potasio y perfusion de insulina . analitica que
evidencia minima insuficiencia renal asi como leve
elevacion de troponina i en meseta . se solicita
valoracion por cardiologia que realiza ecoscopia sin evidenciar disfuncion sistolica
aparente . en planta permanece estable . revisando analiticas previas ambulatorias , en marzo se objetivaba
alteracion de glucemia en ayunas ( 190mg/dl ) , sin recibir tratamiento .
se solicita analitica con hba1c de 13.9% y
funcion tiroidal que es normal ; marcadores
tumorales negativos . ampliamos estudio con tc abdominal que descarta patologia tumoral
```

图4：样本文本

3.3 结论

总之，我们的方法从基础发展到更复杂的方法，以提高线索检测和范围识别。在这个特定的数据集中，分析词之间的关系并利用从依存树派生的模式获得了最佳结果。这种方法使我们能够更准确地捕获上下文，并实现更高的精度。

然而，需要注意的是，方法的有效性可能会根据特定数据集而变化。在某些情况下，最初的基本方法可能会获得令人满意的结果，同时需要较少的计算资源。考虑数据的性质并尝试不同的技术以确定最合适的方法至关重要。

---

4. 深度学习方法

4.1 基于字符的方法

我们将基于字符的方法视为我们LSTM模型的基线，允许我们从最低到更高级别探索标记复杂性。虽然基于字符的模型可能需要更长的训练时间，但它具有更好的泛化潜力，并能够处理混合语言以及词汇表外的词，包括拼写错误或混合语言词。

4.1.1 数据准备

在数据准备阶段，我们从JSON文件中提取文本和注释以创建数据集。每个文本都在字符级别进行标记，每个标记都被分配了相应的标签（NEG, UNC, NSCO, USCO, 或 O）。然后，数据集被分成训练集（70%）和验证集（30%）。最后，我们构建了两个数据加载器，以便于在训练和验证期间高效处理和处理数据集。

4.1.2 模型架构

对于模型，我们使用了简单的LSTM（长短期记忆）架构。模型的结构如下：

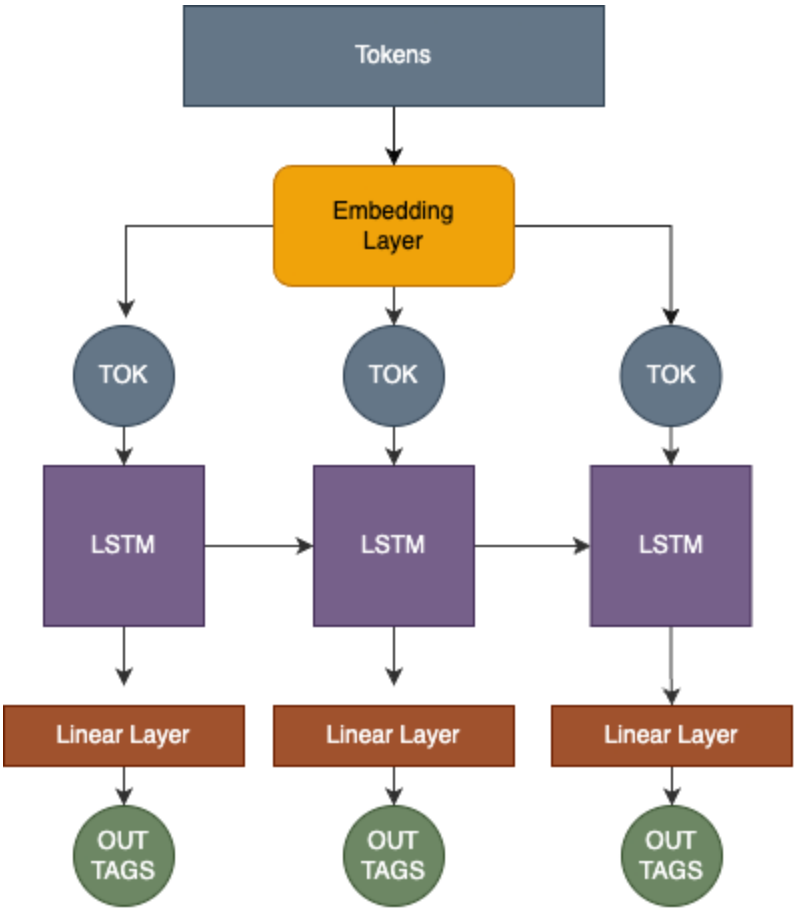

图5：模型架构

后续项目的实施中使用了相同的LSTM模型架构。

4.1.3 模型训练

在基于字符的模型训练过程中，我们使用选定的数据集迭代优化模型参数。我们调整了各种超参数，以探索它们对模型性能的影响。具体来说，我们在训练阶段尝试了字符窗口、隐藏大小和字符嵌入大小。这个阶段的目标是确定产生最佳性能的配置。

4.1.4 问题和解决方案：不平衡数据

最初，由于训练数据集中的不平衡数据，我们的模型结果不佳。“OTHER”标签的普遍性导致大多数标记被分类为如此，导致在准确检测否定和不确定性线索和范围方面遇到困难。

经过实验，我们发现修改损失函数可以改善结果。最初，我们将交叉熵损失应用于所有输出标记，包括“OTHER”标记。然而，这种不平衡问题导致模型优先将大多数标记分类为“OTHER”，而不是正确识别重要的标记。为了解决这个问题，我们通过为非“OTHER”标签分配更高的权重来调整损失函数。

```
loss_full = criterion(y_pred.transpose(1, 2), y)
loss_specif = criterion_specific(y_pred.transpose(1, 2), y)
loss = loss_full*0.55 + loss_specif*0.45
loss.backward()
```

4.1.5 模型评估

下图表示了验证数据预测的混淆矩阵，展示了在修改损失函数后性能的显著提高。虽然最初的结果不尽人意，但对损失函数的调整显著提高了模型的性能。

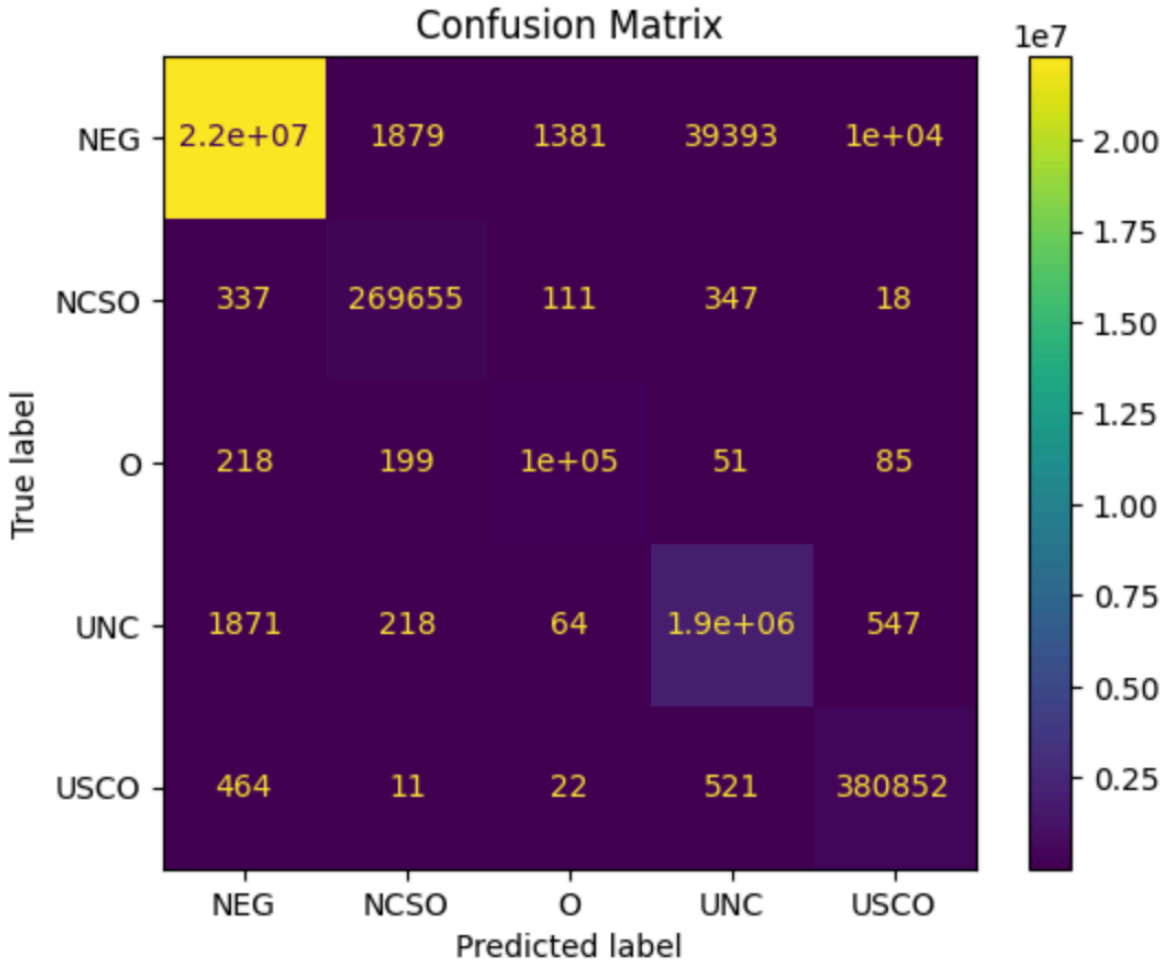

图6：混淆矩阵

在定性结果中，我们观察到“no”这个词在触发其他标记的检测中起着关键作用。

```
NONE USCO UNC NSCO NEG
el paciente no presenta leucoplasia severa en el dorso de la lengua.
```

```
NONE USCO UNC NSCO NEG
el paciente presenta leucoplasia severa en el dorso de la lengua.
```

```
NONE USCO UNC NSCO NEG
aquesta setmana no presentava cap síntoma de febril.
```

4.1.6 结论

基于这些结果，可以得出基于字符的方法没有取得高水平的成功，因为它在许多情况下失败了。在后续部分中，我们将专注于提高深度学习模型的性能。

4.2 基于词的方法

为了尝试提高算法的性能，我们决定探索不同的方法。通过将词视为独立实体并理解它们在句子上下文中的关系，我们预计能够提高准确性并丰富对文本内容的理解。这个项目中遵循的过程如下所述：

4.2.1 数据准备

在数据准备阶段，我们利用JSON文件中的文本和注释创建了一个大型数据集。每个文本都被标记为单词，并分配了相应的标签（NEG, UNC, NSCO, USCO, 或 O）。再次，数据集被分成两个子集：训练集（70%）和验证集（30%）。为了在训练和验证期间高效处理和处理数据集，我们实现了两个数据加载器。

4.2.2 模型训练

对于模型训练，我们使用了与下文（基于字符）描述的相同的模型架构。训练过程涉及将训练集的数据通过数据加载器输入模型，使用指定的损失函数和优化器优化模型参数，并通过反向传播迭代更新模型权重。目标是训练模型以根据输入文本准确预测否定和不确定性标签。

4.2.3 模型评估

为了评估训练模型的性能，我们使用验证集进行了评估。我们计算了准确率、精确度、召回率和F1分数等性能指标，以衡量模型在正确分类否定和不确定性实例方面的有效性。

此外，我们还生成了一个混淆矩阵，以提供模型预测的详细概览，包括真正例、假正例、真负例和假负例。

以下是我们对验证集评估的一些定性发现，展示了模型如何能够准确地检测到大多数线索和范围。

Accuracy: 97.49%
Classification Report:
precision & recall & f1-score & support
NEG & 0.90 & 0.98 & 0.94 & 3500
NSCO & 0.75 & 0.98 & 0.85 & 10571
0 & 1.00 & 0.98 & 0.99 & 157076
UNC & 0.93 & 0.83 & 0.88 & 497
USCO & 0.86 & 0.88 & 0.87 & 1399
accuracy & & & &
macro avg & 0.89 & 0.93 & 0.90 & 173043
weighted avg & 0.98 & 0.97 & 0.98 & 173043

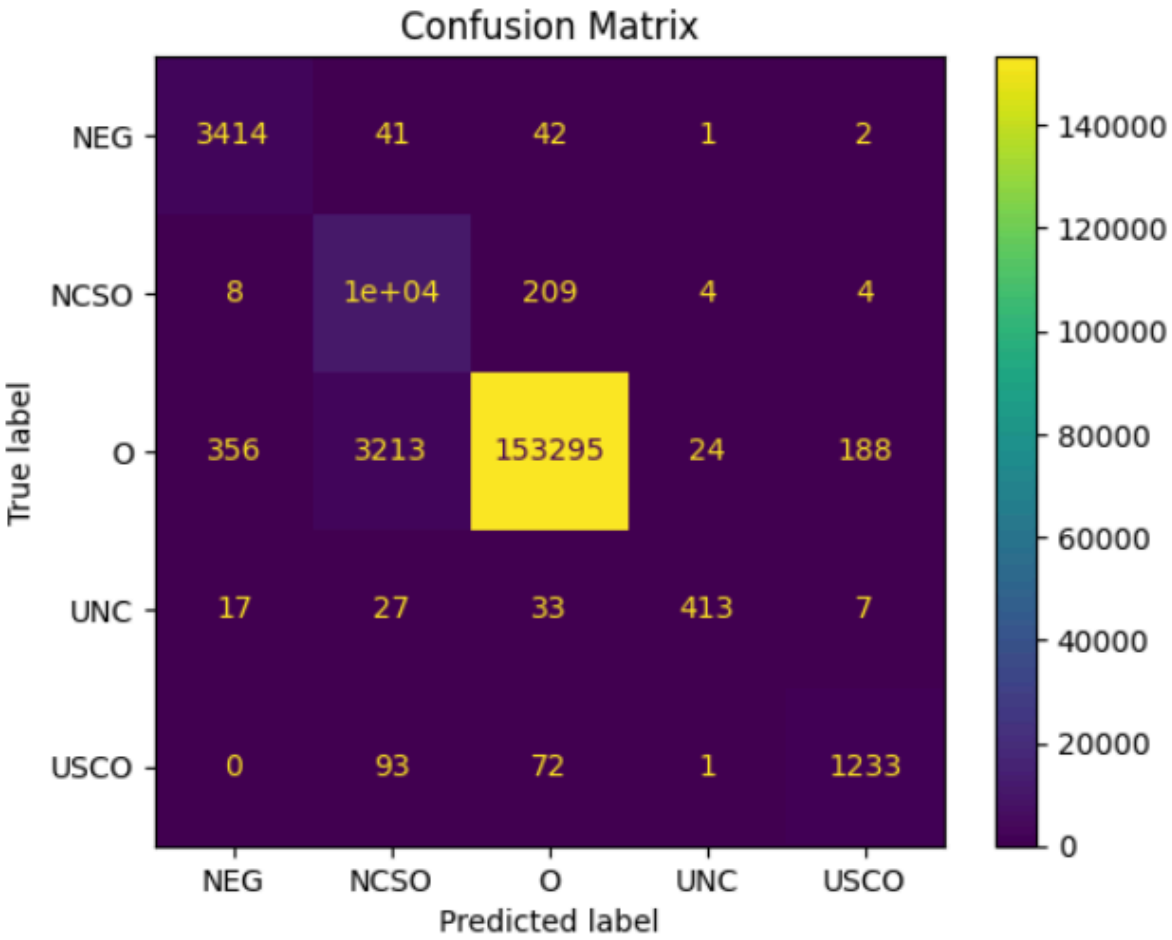

图7：混淆矩阵

4.2.4 结论

根据观察到的结果，我们的基于词的方法在准确识别线索和范围方面取得了成功，使我们能够以高精度实现我们的目标。此外，我们有效地解决了项目中遇到的挑战，特别是不平衡数据问题。通过实施适当的损失函数，我们成功地解决了这一挑战。

4.3 微调BERT

最初，我们对微调BERT寄予厚望，考虑到它在POS和NER标记任务中的卓越表现。然而，我们的结果没有达到预期。我们怀疑“OTHER”标签和其他相关标签之间的不平衡可能是主要的根本原因。尽管尝试了修改损失函数，结果仍然令人失望。

与基于字符的方法类似，我们构建了一个数据集，每个标记都被分配了以下标签之一：'OTHER', 'USCO', 'UNC', 'NSCO', 或 'NEG'。

模型本身是一个简单的'bert-base-multilingual-cased'包装器。它在输出中结合了一个基本的线性层，将BERT的最后隐藏状态转换为所需的标签输出大小。还应用了Dropout正则化以提高性能。

```
estudio . antecedents - sin alergias medicamentosas conocidas . - fumador de 2 paquetes/dia durante mas de 50 años ( fe 100 paq/año ) .
- alcohol : 1 copa de vino diaria y 7 cervezas a el dia ( enolismo 80 g/dia ) . - hipertension arterial esencial en tratamiento farmacologico con dos farmacos con correcto control
tensional . - poliposis colonica por lo que sigue controles en ccee de digestivo de huvh . fcs ( 6/10 ) polipectomia de 5 lesiones polipoideas . ap de colon ascendente : adenoma
tubular y tubulo-vellosos , alguno con focos superficiales de displasia de alto grado . ap colon a 15 cm de margen anal : adenoma tubulo-velloso con displasia de bajo grado
**ultima colonoscopia en enero de 2013 : sin evidencias de hallazgos patologicos salvo a nivel de sigma , mucosa discretamente
eritematosa sugestiva de sigmoiditis leve . diverticulosis de sigma no complicada . lesion submucosa a 90
cms de el margen anal sugestiva de lipoma . hemorroides externas . - aneurisma de aorta ascendente predominantemente tubular diagnosticado en
2013 de manera incidental mediante tc toracoabdominal realizado ambulatoriamente por sindrome constitucional . siguio controles en la unidad de patologia aortica de cardiologia de
huvu ( dra. ********** ) siendo dado de alta en enero de 2014 para seguimiento ambulatorio con ecografia de control cada 2 años . *ultima ett en mayo de 2013 : aa ( 48 mm ) y
raiz aortica ( 39 mm ) dilatadas . insuficiencia aortica ligera-moderada ii . ventriculo izquierdo ligeramente hipertrofico con funcion sistolica conservada . *ultima angiornm en
octubre de 2013 : dilatacion de la porcion tubular de la aorta ascendente ( 47mm ) con morfologia de la raiz aortica conservada y aorta descendente no dilatada
. - litiasis renal bilateral . no disponemos de mas informacion clinica . - esquizofrenia diagnosticada hace
unos 15 años . en seguimiento ambulatorio por psiquiatra de zona . - parkinsonismo vascular diagnosticado en junio de 2016 a raiz de cuadro de bradicinesia y trastorno de la marcha
. en tratamiento farmacologico y en seguimiento por la utm de neurologia de huvh ( dr. ********** ) solicitando se valoracion por ncr en septiembre de 2017 dada la aparicion de
la triada de hakim con hallazgo de hidrocefalia en la rnm de craneo de abril de 2017 . se decidio ingreso para registro de la pic. *tc craneal en agosto de 2016 : marcada atrofia
cerebral de predominio subcortical , signos de leucoaraiosis , un infarto lacunar cronico en territorio de vascularizacion de arterias perforantes dependientes de la circulacion
anterior asi como un pequeño infarto cronico en territorio de vascularizacion de arteria cerebelosa superior derecha . *rnm craneal en abril de 2017 : moderat grau d'atrofia
corticosubcortical global . acusada hidrocefalia supratentorial de caracteristiques croniques , amb estenosi de el terç mitja de
l'aqüeducte de silvi malgrat aquest persisteix permeable . no s'evidencien signes d'hidrocefalia cronica de l'adult
. moderada desmielinitzacio de substancia blanca profunda de probable origen hipoxic cronic . petit infart lacunar cronic a el
```

图7：样本文本

```
class BERT_Tagger(nn.Module):
    def __init__(self, bert, output_dim, dropout):
        super().__init__()
        self.bert = bert
        embedding_dim = bert.config.to_dict()['hidden_size']
        self.fc = nn.Linear(embedding_dim, output_dim)
        self.dropout = nn.Dropout(dropout)

    def forward(self, tokens):
        bert_out = self.bert(tokens)["last_hidden_state"]
        predictions = self.fc(bert_out)
        return predictions
```

尽管最初的承诺，微调BERT并没有产生预期的结果。显然，需要进一步的调查和替代策略来克服不平衡问题带来的挑战，并提高模型的性能。

4.4 最终评估

在我们评估各种方法的过程中，我们得出结论，更简单的模型通常展现出更好的整体性能。尽管基于规则的方法简单，但它展现了可称赞的计算效率和速度。然而，它在捕获深度学习方法擅长的复杂模式方面存在困难。在深度学习方法中，我们发现基于词的方法，结合基本的LSTM，取得了显著的好结果。

虽然我们承认，如果有更多时间和努力，BERT模型可能会学习到复杂的模式并优于基于词的方法，但这将以增加训练期间的计算强度和资源需求为代价。

这个项目为我们提供了宝贵的洞见，突出了使用基于规则的简单专家系统的有效性。它是我们整个学期对人工智能主要关注点的一个显著转变，让我们了解到在特定情境下同样有效的替代方法论。