Identificação e classificação de comentários tóxicos utilizando processamento de linguagens naturais e técnicas de aprendizagem profunda

Miguel Angelo Cece de Castro Neto[1], **Paulo Alves dos Santos Junior**[1]
Alceu de Souza Britto Junior[2]

[1] Departamento de Informática – Universidade Estadual de Ponta Grossa (UEPG)
Avenida General Carlos Cavalcanti, 4748
CEP 84030-900 – Ponta Grossa, PR – Brasil

[2]Programa de Pós-Graduação em Informática (PPGIa) – Pontificia Universidade Católica do Paraná (PUCPR)
Rua Imaculada Conceição, 1155
CEP 80215-901 – Curitiba, PR – Brasil

miguelceccineto@gmail.com, contato@pauloalvesjr.com, alceu@ppgia.pucpr.br

Abstract. *The article addresses the problem of multiclass sentiment analysis at the sentence level. Recurrent neural networks and their demonstrations have demonstrated successful modeling of sentiment classifiers and last generation in language modeling with the same demonstration demonstrating efficiency and performance in various tasks. The datasheet of the learned paper transfer, larger database, has proven effective in small databases. This work proposes the use ages between architecture in a with modeling of language to a classification of the sentiment in comment.*

Resumo. *Este artigo aborda o problema da análise de sentimentos multiclasse no nível de sentença. Redes neurais recorrentes e suas variações demonstraram sucesso na modelagem de classificadores de sentimentos e recentemente a modelagem de linguagens através dessa arquitetura demonstrou-se eficaz atingindo o estado da arte em várias tarefas. O uso da técnica de transferência de aprendizado, proveniente de banco de dados maiores, se demonstrou eficiente em banco de dados pequenos. Nesse trabalho propomos o uso destas variações arquitetura em conjunto com a modelagem de linguagem para a classificação de sentimento em comentários.*

1. Introdução

Discutir assuntos na internet pode ser difícil. Ameaças, ofensas pessoais, e assédio, podem criar um ambiente tóxico dentro de fóruns e redes sociais, e até afastar usuários. As plataformas lutam para combater efetivamente discussões tóxicas e limitar ou desligar usuários com essas práticas, porém essa é uma tarefa de difícil automação exigindo grande quantidade de dados e métodos capazes de analisar de maneira eficaz o contexto e significado de sentenças.

Recentemente, em resposta ao crescimento do uso de redes sociais, comentários são publicados massivamente. Como a escrita compõe grande parte de todos os dados

gerados pela humanidade, que até então, apenas serviam para consultas, hoje servem como base de dados para o desenvolvimento de sistemas de processamento de linguagens naturais. Aplicando técnicas de aprendizagem profunda e processamento de linguagens naturais podemos automatizar processos análise e classificação de textos com uma boa taxa de acerto, comparado à classificação por métodos bayesianos, por exemplo.

No problema a ser discutido nesse artigo, com base de dados referente a comentários tóxicos, temos basicamente a análise de sentimentos conforme as classes rotuladas no mesmo, sendo utilizada para representar em quais classes de toxicidade determinada sentneça pertence. Formalmente, o objetivo da análise de sentimentos é extrair a seguinte sêxtupla:

$$(a\lambda 1, a\lambda 2, a\lambda 3, a\lambda 4, a\lambda 5, a\lambda 6)$$

Onde $a\lambda i$ se refere a probabilidade de cada classe, que respectivamente representam: Tóxico; Muito Tóxico; Obsceno; Ameaça; Insulto; Ódio de Identidade.

A análise de sentimentos é determinada comumente de maneira binária (positiva e negativa), porém abordagens multiclasse também são possíveis. Em uma abordagem multiclasse determinado comentário pode possuir diversos rótulos simultaneamente. Para definir a ativação de determinada classe, é definido um limiar. Tal limiar só é válido devido a função de ativação utilizada como saída na rede.

A utilização de redes neurais artificiais voltou a popularizar-se no reconhecimento de padrões em imagens devido a evolução de hardware, tais métodos de aprendizagem de máquina combinados ao processamento de linguagens naturais, permitem a avaliação automática de padrões e extração de conhecimento de bases de texto. A identificação de comentários ofensivos em textos é um derivado do caso geral de análise de sentimentos, na era da internet, identificar e classificar sentenças permite tomada de decisões e maior controle e filtragem de conteúdo.

Esse trabalho tem como principal interesse, o estudo e experimento de técnicas de aprendizagem profunda aplicada à área de processamento de linguagens naturais. Abordagens de aprendizagem de máquina tradicionais necessitam de intervenção humana para definição de características, existe a possibilidade de delegar tal tarefa à algoritmos simples de extração de características línguisticas ou sintáticas disponíveis na literatura, como utilizando contagem, ou processamento esparso. A tarefa realizada utilizando métodos clássicos, por fim utilizam algum método superficial de aprendizado, como por exemplo, máquinas de vetores de suporte [Cortes and Vapnik 1995], naive bayes e naive bayes SVM [Wang and Manning 2012].

Embora os métodos descritos anteriormente já tenham sido validados, o foco desse trabalho é utilizar um método automátizado de extração de características que seja eficaz. Para a análise de sentimentos em comentários e classificação de toxicidade, utilizaremos utilizaremos redes neurais recorrentes em conjunto com estruturas celulares específicas, como células recorrentes bloqueadas [Pascanu et al. 2013] e células de longa memória de curto termo [Hochreiter and Schmidhuber 1997]. Abordaremos técnicas de transferência de aprendizado e construção de modelos de linguagem. A estrutura recorrente é uma variação da rede neural clássica para aceitar entradas com tamanhos e saídas arbitrários,

possuindo também registro de contexto em relação a posição do elemento em seu conjunto de entrada. Redes recorrentes comuns não são efetivas devido a problemas de esquecimento de aprendizado ou pouca capacidade de aprendizado devido a limitação imposta pela quantidade de parâmetros. Para resolver essa situação existem diferentes células que podem ser incorporadas nas arquiteturas, como por exemplo células recorrentes bloqueadas e células de longa memória de curto termo.

2. Trabalho Relacionado

A análise de sentimentos está sob grande aprimoramento devido a aplicação de redes recorrentes [Karparthy 2015], redes convolucionais [Y. LeCun and Haffner 1998] e transferência de aprendizado [Howard and Ruder 2018].

O principal problema do método superficial citado acima, é que ele não é capaz de identificar o contexto de palavras, não sendo uma opção robusta para classificação.

Redes recorrentes e suas variações são aplicadas utilizando comumente matrizes que definem significado próximos ao real para cada palavra, como por exemplo Word2Vec [Mikolov et al. 2013] ou GloVe (Socher et al. 2014). Tais métodos provem contexto utilizando palavras próximas (modelo skip-gram) para modelas matrizes e consequentemente contexto para redes recorrentes.

Transferência de aprendizado é uma técnica onde um modelo já treinado é reajustado para detectar padrões diferentes dos iniciais. Geralmente usado quando a base de dados disponível é menor do que a usada no modelo já treinado. Trabalhos recentes propõem a utilização de modelos de linguagem universais, que sofrerão afinação (fine tuning) para atender uma tarefa específica [Howard and Ruder 2018] com a possibilidade de utilização de dados não supervisionados para gerar um aumento de performance (Radford et al. 2018). A utilização de dados não rotulados permite reduzir drasticamente a necessidade de exemplos rotulados. O estado da arte da maioria das tarefas na área de linguagens naturais está relacionado ao uso de transferência de aprendizado e aprendizado não supervisionado, que é um dos maiores desafios atualmente.

2.1. Funções de ativação

Funções de ativação não lineares, dão capacidades não lineares para redes neurais [Y. LeCun and Haffner 1998].

2.1.1. ReLU - Rectified Linear Activation Function

Dada pela fórmula:

$$ReLU(Z) = max(0, Z) \tag{1}$$

A imagem abaixo descreve expressão da ReLU em um intervalo de 5 a –5.

É a função de otimização padrão recomendada para uso na maioria das redes neurais. Aplicando essa função no output de uma transformação linear produz uma transformação não linear. A função permanece muito próxima de ser linear, contudo,

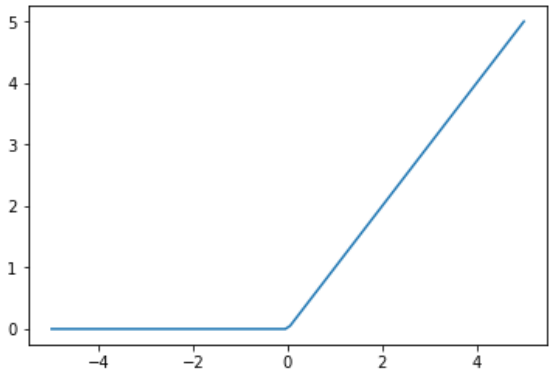

Figura 1. Unidade Linear Retificada

no sentido de que é uma função linear por partes com duas peças lineares. Como as unidades lineares retificadas são quase lineares, elas preservam as propriedades que tornam os modelos lineares fáceis de otimizar com métodos baseados em gradiente. Eles também preservam muitas das propriedades que fazem modelos lineares generalizarem sistemas a partir de componentes mínimos [Goodfellow et al. 2016].

2.1.2. Sigmoide

A função sigmoide é dada pela formula:

$$\sigma(Z) = 1/(1 + e^{-Z}) \tag{2}$$

A imagem abaixo descreve a curva sigmoidal em um intervalo de 5 a –5.

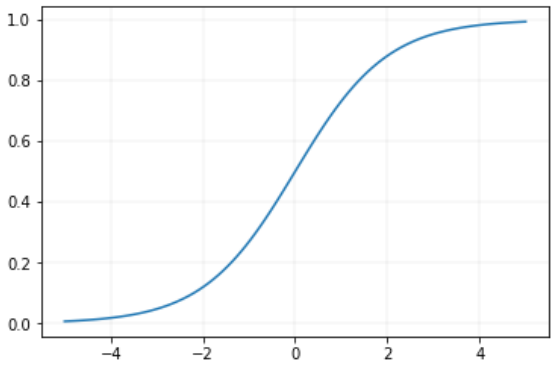

Figura 2. Curva Sigmoidal

Aplicando essa função no output de uma transformação linear, teremos como resultado um valor entre 0 e 1.

Além de a função sigmoide ser utilizada na estrutura de células de longa memória de curto termo e células recorrentes bloqueadas, também pode ser usada na saída do modelo. Poderíamos por exemplo, na classificação de sentimentos em texto, considerar valores próximos de 0 um sentimento negativo, e próximos de 1, positivo. Os valores

maiores iguais a 0.5 serão classificados como positivos, enquanto 0.5 serão classificados como negativos.

2.1.3. Tangente Hiperbólica

Dada pela formula:

$$\sigma(Z) = 1/(1 + e^{-Z}) \tag{3}$$

A imagem abaixo descreve a curva da tangente hiperbolica em um intervalo de 5 a -5.

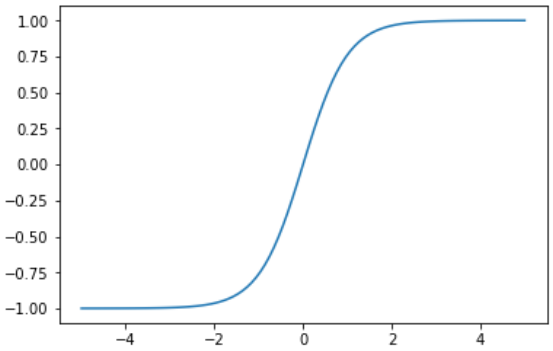

Figura 3. Curva Tangente Hiperbólica

Produz um resultado entre 1 e -1, e é usada na estrutura de células simples de redes recorrentes, células de longa memória de curto termo e células recorrentes bloqueada. Em comparação a função de ativação sigmoide, possui a vantagem de ser simétrica em relação ao eixo x.

2.2. Otimizadores

2.2.1. Escolhendo o Algoritmo de Otimização

Infelizmente ainda não existe um consenso sobre qual é o melhor algoritmo de otimização. Schaul et al. (2014) apresentou uma comparação valiosa com um grande número de algoritmos de otimização através de uma ampla variedade de tarefas de otimização. Enquanto os resultados sugeriam que a família de algoritmos com taxa de aprendizado adaptativa (como RMSProp e AdaDelta) apresentavam desempenho robusto, não houve um único algoritmo que emergiu sobre os outros.

Atualmente, os algoritmos de otimização mais populares são Gradiente Descendente Estocástico, Gradiente Descendente Estocástico com Momento, RMSProp, RMSProp com Momento, AdaDelta e Adam. A escolha de qual algoritmo de otimização usar, parece depender mais na familiaridade do desenvolvedor com o com o algoritmo (para afinar os hiperparametros mais facilmente) [Goodfellow et al. 2016].

2.2.2. Otimizador Adam

Com a utilização de minibatches, o progresso no gradiente tende a oscilar, logo o aprendizado pela rede se torna lento. Várias soluções foram propostas para solucionar esse problema, e a mais popular é uma junção das anteriores. O otimizador Adam [Kingma and Ba 2014] utiliza a estimação do primeiro e do segundo momento dos gradientes.

2.3. Função de Custo

Um aspecto importante no desenvolvimento de redes neurais profundas é na escolha de função de custo. Felizmente, as funções de custo para redes neurais são mais ou menos as mesmas da de outros modelos paramétricos, assim como modelos lineares [Goodfellow et al. 2016].

2.3.1. Entropia Cruzada

O cálculo do custo por entropia cruzada é dado pela seguinte formula:

$$CE = -\sum_{x} p(x) \log q(x) \tag{4}$$

2.4. Regularização

Um problema central em aprendizado de máquina é como fazer um algoritmo que irá ter um bom desempenho não somente nos dados de treino, mas também em novos dados inseridos nele. Muitas estratégias usadas em aprendizado de máquina são explicitamente desenvolvidas para reduzir o erro no conjunto de teste, possivelmente à custa de um erro maior no conjunto de treino. Essas estratégias são conhecidas coletivamente como regularização [Goodfellow et al. 2016].

O Dropout oferece pouco custo computacional, porém um método poderoso para regularização de uma ampla família de modelos (Srivastava et al. 2014).

Especificamente, dropout treina o conjunto consistindo através de sub-redes geradas pela remoção de unidades que fazem parte das camadas intermediarias ou de entrada como ilustra a figura abaixo: [Goodfellow et al. 2016].

2.5. Inicialização de parâmetros

Talvez a única propriedade conhecida com completa certeza é que parâmetros precisam ”quebrar a simetria”entre as diferentes unidades. Se duas unidades internas com a mesma função de ativação, estão conectadas às mesmas entradas, então essas unidades devem ter diferentes parâmetros iniciais. Caso elas tenham os mesmos parâmetros iniciais então um algoritmo de aprendizado determinístico aplicado à um modelo e função de custo também determinísticos, irá conseguintemente atualizar essas duas unidades da mesma forma. Mesmo se o modelo ou o algoritmo de treino for capaz de usar aleatoriedade para computar diferentes atualizações para diferentes unidades (por exemplo, dropout), na maioria das vezes é melhor inicializar cada unidade para computar uma função diferente de todas as outras unidades. Isso pode ajudar a garantir que nenhum padrão de entrada

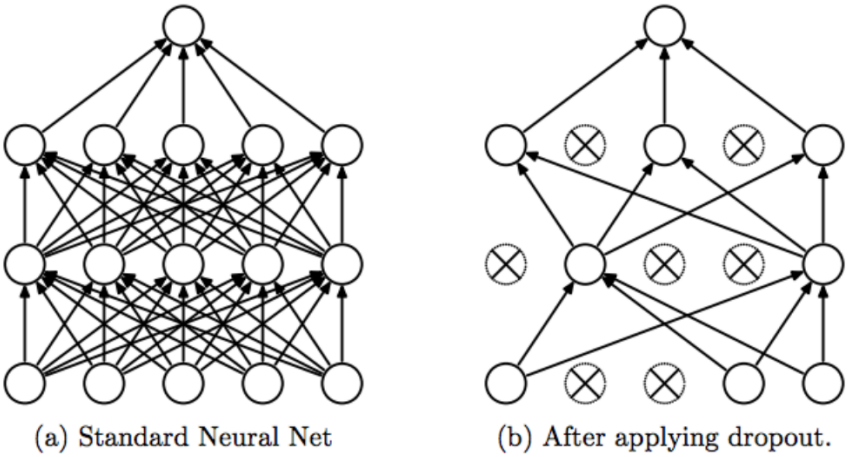

Figura 4. Dropout

seja perdido no espaço nulo da forward propagation e nenhum padrão de gradiente seja perdido no espaço nulo da back-propagation. O objetivo de ter cada unidade computando uma funçao diferente motiva a inicialização aleatória de parâmetros (Goodfellow et al. 2017).

Comumente, estabelecemos constantes os valores dos biases escolhidos de maneira heurística, e inicializamos apenas os pesos aleatoriamente (Goodfellow et al. 2017).

2.6. Word Embeddings

Word embedding é o nome coletivo de um conjunto de técnicas de modelagem de linguagem e de aprendizado de recursos no processamento de linguagem natural, em que palavras ou frases do vocabulário são mapeadas para vetores de números reais. Conceitualmente, envolve uma incorporação matemática de um espaço com uma dimensão por palavra para um espaço vetorial contínuo com uma dimensão muito menor.

2.7. Redes Neurais Recorrentes

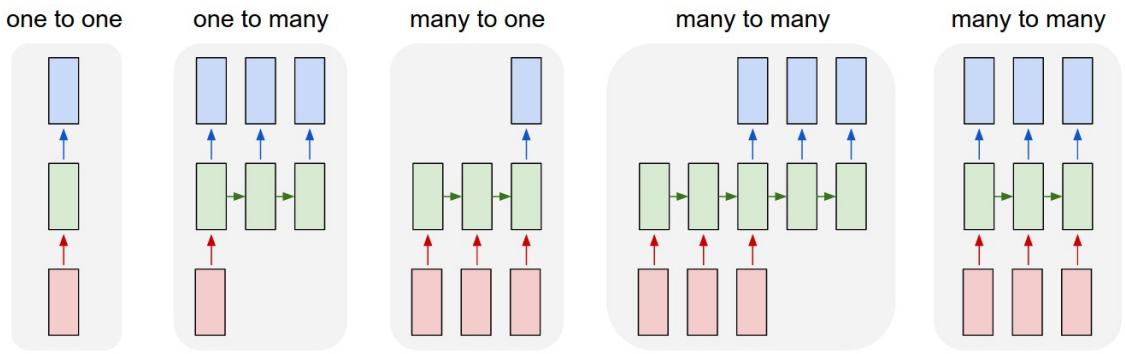

Figura 5. Possibilidades de entrada de saída em redes reccorentes. [Karparthy 2015]

2.7.1. (GRU)

Células recorrentes bloqueadas são definidas pelas seguintes formulas:

$$\text{Ç} < t >= tanh(Wc[\Gamma_r * c < t-1 >, x < t >] + b_c) \tag{5}$$

$$\Gamma_u = \sigma(W_u[c < t-1 >, x < t >] + b_u) \tag{6}$$

$$\Gamma_r = \sigma(W_r[c < t-1 >, x < t >] + b_r) \tag{7}$$

$$c < t >= \Gamma_u * \text{Ç} < t > +(1 - \Gamma_u) * c < t-1) \tag{8}$$

$$a < t >= c < t > \tag{9}$$

2.7.2. Longa memória de curto termo (LSTM)

Células de longa memória de curto termo são definidas pelas seguintes formulas:

$$\text{Ç} < t >= tanh(Wc[a < t-1 >, x < t >] + b_c) \tag{10}$$

$$\Gamma_u = \sigma(W_u[c < t-1 >, x < t >] + b_u) \tag{11}$$

$$\Gamma_f = \sigma(W_f[c < t-1 >, x < t >] + b_f) \tag{12}$$

$$\Gamma_o = \sigma(W_o[c < t-1 >, x < t >] + b_o) \tag{13}$$

$$c < t >= \Gamma_u * \text{Ç} < t > +(\Gamma_f) * c < t-1) \tag{14}$$

$$a < t >= \Gamma_o * c < t > \tag{15}$$

3. Metodologia

3.1. Organização dos dados

O banco de dados utilizou da técnica de separação de treino e teste, onde 90% dos dados foi atribuída para treino e 10% para validação. Não existe uma métrica fixa para a definição da quantidade a ser repartida, porém essa deve ser feita de acordo com a quantidade de dados rotulados. Como a validação segue apenas para verificar se não está acontecendo sobreajuste aos dados de treinamento, foi escolhido 10% como dos hiperparametros. Essa técnica é conhecida como holdout. Em abordagens clássicas de aprendizado de máquina, normalmente é utilizada a validação cruzada, porém, em casos de aprendizagem profunda isso se torna inviável devido a quantidade de processamento. Como modelos de

aprendizagem profunda possuem grande volume de dados, a abordagem mais popular em relação a desempenho e resultado é escolhida nesse trabalho.

A segmentação a seguir representa como os dados foram separados:

Figura 6. Separacão de dados

A segmentação de testes não está inclusa nos dados públicos. Tal teste é feita através de dados não rotulados na plataforma Kaggle. É possível realizar um submissão com os rotulos e obter a acurácia. Essa segmentação não é de acesso público para evitar que modelos sobreajustem seus parâmetros, assim o resultado retornado garante a validade de generalização do modelo.

3.2. Definição do modelo inicial

Um modelo inicial foi definido para obter o primeiro resultado e utilizar uma técnica bastante popular no segmento de aprendizagem profunda. A técnica de iteração, consiste na ideia de definir uma arquitetura inicial, e definir hiperparâmetros rapidamente. Como ponto de partida, foi utilizado uma arquitetura bastante popular e extremamente efetiva, sendo sem transferência de aprendizado, o estado da arte. A rede foi definida da seguinte forma:

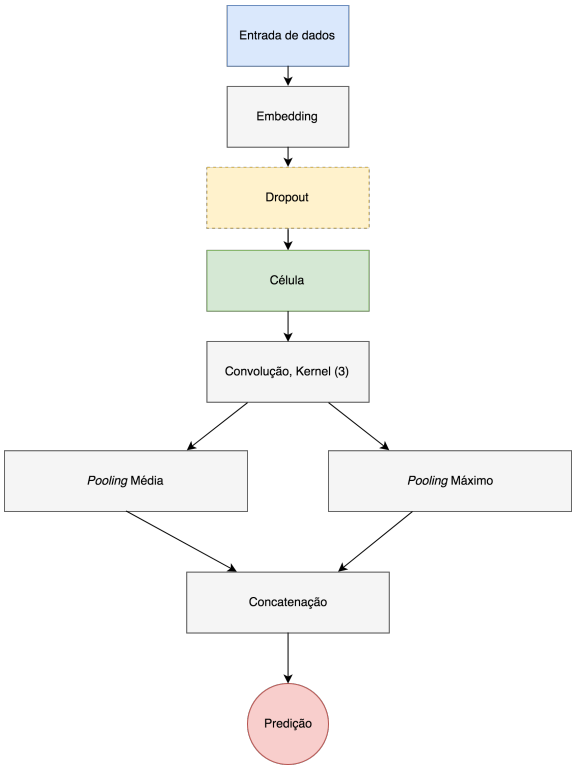

Figura 7. Estrutura genérica de rede

Embedding se refere a essa codificação. Utilizamos o conceito de vetores globais para representação de palavras (GloVe) pré-treinados em um crawler global.

Dropout previne o encaixe exagerado da função aos dados, logo faz o modelo generalizar melhor.

Célula define qual componente interno da rede neural recorrente será utilizado. Nessa arquitetura utilizamos variações entre 128 e 256 nós.

Convolução utiliza uma janela deslizante unidimensional com um filtro de tamanho três. Busca encontrar características e padrões.

Pooling diminuem tamanho da saída da convolução.

Concatenção junta os tensores de saída dos poolings.

Predição faz uma rede neural totalmente conectada e utiliza a função de ativação sigmoid, sendo que retornos maior que 0.5 serão considerados verdadeiros, e menores como falso.

Os hiperparâmetros do otimizar foram definidos através dos valores indicados no artigo do mesmo. A taxa de aprendizado utiliza do processo de iteração, e o valor utilizado nas análises preliminares é arbitrário, sendo inicialmente 1x10-3. Os pesos das células foram inicializador através da inicialização Glorot (Bengio Glorot, 2010).

4. Análises Preliminares

Em primeira instância os modelos obtiveram resultados satisfatórios, alcançando acurácia próxima à melhor submissão deste problema.

QUADRO

Podemos observar que nesse conjunto de dados a bidirecionalidade não trás benefícios nestas instâncias. Outro fator de relevante é o desempenho do GRU ter sobresaido em relação a LSTM. O GRU é um modelo ”simplificado”da LSTM tradicional.

Análises utilizando modelos de linguagem e transferência de aprendizado ainda estão sendo desenvolvidas.

5. Considerações Finais

Os modelos testados apresentaram um resultado satisfatório, porém não melhor que o melhor resultado atual. Técnicas como transferência de aprendizado e modelo de linguagem não foram aplicadas. Há a possibilidade de reelaborar a célula que constrói a rede neural recorrente para atender melhor a este caso. Talvez a aplicação de uma rede de capsulas (Hilton et al. 2017) proporcionaria um resultado superior devido a não variação em translações, uma das características desta arquitetura.

6. Referências Bibliográficas

Referências

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*.

Howard, J. and Ruder, S. (2018). Fine-tuned language models for text classification. *CoRR*, abs/1801.06146.

Karparthy, A. (2015). The unreasonable effectiveness of recurrent neural networks.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Pascanu, R., Gülçehre, Ç., Cho, K., and Bengio, Y. (2013). How to construct deep recurrent neural networks. *CoRR*, abs/1312.6026.

Wang, S. and Manning, C. D. (2012). Baselines and bigrams: Simple, good sentiment and topic classification.

Y. LeCun, L. Bottou, Y. B. and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*.