Detecção de Categorias de Aspectos Utilizando Redes Neurais Profundas em Avaliações Online

Bruno Á. Souza[1], **Alice A. F. Menezes**[1], **Carlos M. S. Figueiredo**[1,2], **Fabíola G. Nakamura**[1], **Eduardo F. Nakamura**[1]

[1]Universidade Federal do Amazonas (UFAM) – Manaus, AM – Brasil

[2]Universidade do Estado do Amazonas (UEA) – Manaus, AM – Brasil

{bruno.abia,alice.menezes,fabiola,nakamura}@icomp.ufam.edu.br

cfigueiredo@uea.edu.br

Abstract. *Virtual environments such as online stores (e.g. Amazon, Google Play and Booking) adopt a collaborative strategy of evaluation and reputation, where users classify products and services. User's opinion represents the satisfaction level of a rated item. The set of ratings of an item is a reference to its reputation/quality. Therefore, the automatic identification of a usersatisfaction related to an item, considering its textual evaluation, is a tool with singular economic potential. With deep learning researches evolution in sentiment analysis based in aspects, opportunities to apply several neural networks in this context arisen. However, the data representation models applied in these works focus only on Embeddings pre-trained networks as a way to perform feature extraction. In this way, this work aims to present a comparison between data representation techniques and deep networks approaches, to analyze which of them have better results in classifying categories of aspects. Thus, we can seethat TF-IDF with a Convolution Neural Network (CNN) had an F1 measure of 0.93%, being at least 0.02% higher than the others approaches applied in this work.*

Resumo. *Ambientes virtuais, como lojas online (e.g. Amazon, Google Play e Booking), adotam uma estratégia colaborativa de avaliação e reputação, onde os usuários classificam os produtos e serviços. A opinião do usuário representa o grau de satisfação em relação ao item avaliado. O conjunto de avaliações de um item é referencial de sua reputação/qualidade. Portanto, a identificação automática da satisfação do usuário em relação a um item, considerando sua avaliação textual, é uma ferramenta com potencial econômico singular. Com a evolução das pesquisas relacionadas a aprendizagem profunda na área de análise de sentimentos baseada em aspectos, têm surgido oportunidades de aplicar diversas redes neurais neste contexto. Porém, os modelos de representação de dados aplicados nesses pesquisas focam unicamente no uso de redes pré-treinadas de Embeddings como forma de realizar a extração de características dos dados. Desta forma, este trabalho tem como objetivo apresentar uma comparação entre técnicas de representação de dados e abordagens redes profundas, a fim de verificar qual apresenta melhor resultado na tarefa de classificar categorias de aspectos. Com isso, conseguimos observar que o uso de TF-IDF com uma Rede Neural Convolucional (CNN) apresentou uma me-*

dida F1 de 0,93%, sendo pelo menos 0,02% superior as demais aplicadas neste trabalho.

1. Introdução

Ambientes virtuais, como lojas online (e.g. Amazon, Google Play, Booking) e redes sociais, permitem que usuários avaliem produtos, serviços e compartilhem suas experiências. Esse compartilhamento de opiniões define de forma colaborativa a reputação tanto dos estabelecimentos quanto dos produtos/serviços disponibilizados. Essa realimentação dos clientes representa uma ferramenta importante para que as empresas possam identificar oportunidades de melhoria e crescimento. Portanto, identificar automaticamente a polaridade ou sentimento de uma avaliação realizada por um usuário (e.g. uma a cinco estrelas ou simplesmente positivo/negativo) é um problema com grande potencial econômico e estratégico para empresas [de Paula et al. 2017, Liu 2012, Ye et al. 2009].

A maioria dos trabalhos da literatura tem adotado técnicas de processamento de linguagem natural e algoritmos de aprendizagem de máquina (SVM, Naive Bayes e Max Entropy) para realizar as inferências de sentimentos e categorias expressas nos textos [Souza et al. 2016, Almeida et al. 2016, Araújo et al. 2013, Stiilpen Junior and Merschmann 2016, Santos and Moura 2016, Ye et al. 2009]. Porém, estas técnicas possuem limitações quanto ao retorno dos dados classificados, pois realizam a classificação apenas em positivo ou negativo.

Recentemente, técnicas de classificação de sentimento em nível de aspecto tem sido exploradas [Pavlopoulos 2014, Pontiki et al. 2015, Gulaty 2016], pois além de obterem o mesmo retorno dos algoritmos tradicionais também informam quais são os contextos que estão sendo tratados na sentença. Podemos citar como exemplo, a análise de uma entidade loja. Neste caso, podemos descobrir aspectos como atendimento e qualidade dos produtos e preços. Assim, dada uma sentença e um aspecto existente em um texto, o objetivo principal deste tipo de abordagem é inferir a polaridade/sentimento (positivo, negativo ou neutro) relacionado a este aspecto analisado. Em um exemplo prático, analisemos a revisão "*A comida é excelente, mas o atendimento é horrível*". O sentimento referente ao aspecto "*comida*" é positivo, enquanto o aspecto "*atendimento*" é negativo.

Segundo [Pontiki et al. 2016] esse problema de análise de sentimentos baseada em aspectos pode ser dividias em 4 subtarefas:

- Subtarefa 1: Dada uma sentença t, a abordagem ser capaz de reconhecer os aspectos existentes no texto;
- Subtarefa 2: Dados os aspectos extraídos do texto, ser capaz de classificar em positivo, negativo ou neutro, respectivamente, cada informação coletada;
- Subtarefa 3: Dado um determinado aspecto no texto em que existe uma categoria associada, como "A comida deste restaurante é ruim", o aspecto seria **comida** e a categoria seria **qualidade da comida**;
- Subtarefa 4: Dadas as categorias extraídas do texto, ser capaz de classificar em positivo, negativo ou neutro respectivamente cada informação coletada;

Como solução para estas subtarefas, Redes Neurais Convolucionais (CNN), Redes Neurais Recorrentes (RNN) e *Long Short-Term Memory* (LSTM) têm sido utilizadas

para resolver este problema no âmbito de Aprendizagem Profunda. Estas redes vem obtendo resultados representativos para este tipo de classificação de sentimento baseada em aspectos.

Neste trabalho, buscamos solucionar a subtarefa 3 (detecção de categorias). Assim, apresentamos as seguintes contribuições: (i) a análise de modelos de representação de dados para textos aplicados a tarefa de classificação; e (ii) a aplicação de 3 redes profundas (LSTM, CNN e RNN) para realizar a subtarefa de reconhecimento de categoria da análise de sentimentos baseada em aspectos.

O restante deste trabalho está dividido da seguinte forma: na Seção 2, são apresentados os trabalhos relacionados de análise de sentimentos baseada em aspectos; na Seção 3, é descrita a abordagem proposta para este trabalho; na Seção 4, são apresentados os experimentos e os resultados obtidos; por fim, a Seção 5 apresenta a conclusão e trabalhos futuros.

2. Trabalhos Relacionados

Atualmente, a análise de sentimentos baseada em aspectos tem sido assunto de muitos trabalhos na literatura dentro da área de mineração de textos [Kim 2014, Nguyen and Shirai 2015, Poria et al. 2016]. O objetivo principal desta área é a execução de duas tarefas principais: (i) reconhecimento de qual assunto (tópico) está sendo tratado em uma sentença s, onde $s \geq 1$, ou seja, onde pode existir um ou mais tópicos sendo abordados; (ii) a inferência de polaridade sob o texto que está sendo analisado (positivo, negativo ou neutro).

Abordagens anteriores realizam essa extração utilizando uma estratégia de classificação de múltiplas classes [Pontiki et al. 2016]. Este tipo de abordagem possui algumas limitações, pois em sua maioria apresentam dependências de domínio em relação a base de dados nas quais estavam sendo aplicadas. Já na análise de sentimentos eram utilizados diferentes classificadores com uma ampla variedade de recursos, como o uso de unigramas, *bag-of-words* com TF-IDF, *pos-tag* e sentenças léxicas.

Atualmente, trabalhos como o de [Xu et al. 2017] demonstram o uso de Redes Neurais Convolucionais (CNN) para a análise de sentimentos baseado em aspectos. Neste contexto, os autores demonstram a possibilidade de executar esta estratégia independente de domínio. Em seus experimentos, os autores comparam sua abordagem com o SVM e uma LSTM, alcançando uma acurácia de 76.90% nas avaliações relacionadas a computadores e 68.34% nas avaliações a respeito de restaurantes, ficando abaixo apenas do SVM, que obteve 80% na primeira base de dados e 72.1% na segunda. No trabalho de [Wang and Liu 2015], os autores também usaram CNN para este tipo de inferência. Em seus resultados, os autores conseguiram uma medida de F1 de 51% na detecção de aspectos, enquanto na análise de sentimentos obtiveram uma acurácia de 78%.

No trabalho de [Wang et al. 2016] é apresentado o uso de Redes Neurais Recursivas combinadas para a execução das duas tarefas. Em seus experimentos, os autores comparam seus resultados com outras técnicas como LSTM e CNN. Como resultado, os autores obtiveram melhor desempenho na mineração de opinião na base de dados de avaliações de restaurantes, atingindo uma medida F1 de 84.11%. Na base de avaliações de computadores, obtiveram melhor resultado que os demais métodos das duas tarefas, atingindo uma medida F1 acima de 78% nas classificações.

Diferente das pesquisas já realizadas, este trabalho utiliza outras representações de dados como entrada para as redes profundas, a fim de verificar se há melhora nos resultados da classificação das categorias dos aspectos no domínio de avaliações de restaurantes.

3. Abordagem Proposta

A abordagem proposta na elaboração deste trabalho consiste das seguintes etapas (ilustradas na Figura 1): (i) pré-processamento dos dados, a fim de retirar possíveis ruídos e termos não representativos para as avaliações; (ii) extração das características dos textos para entrada das redes profundas; (iii) treinamento das redes profundas para classificação das categorias existentes nos textos.

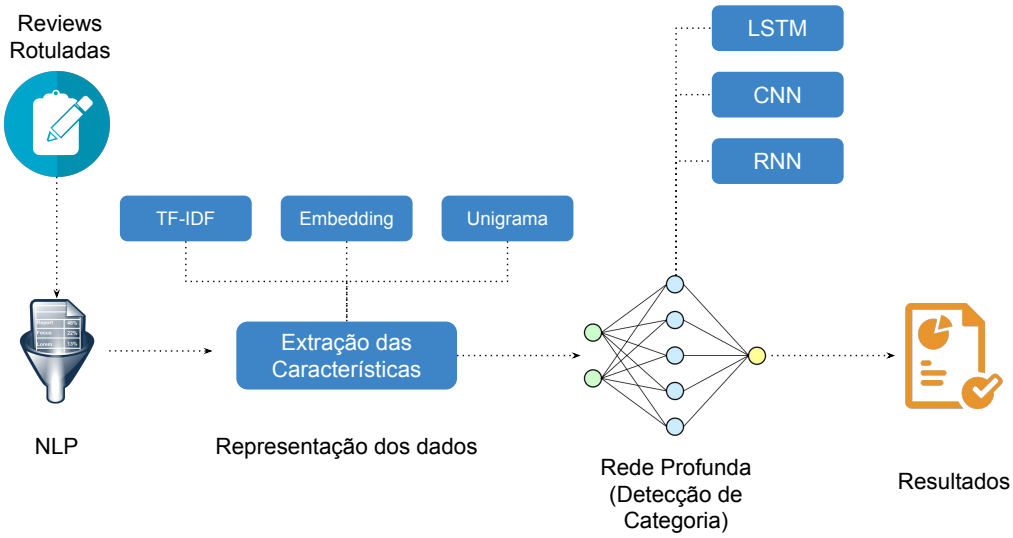


Figure 1. Arquitetura da abordagem utilizada.

3.1. Pré-processamento

Nesta etapa, os textos pertencentes a base de dados foram submetidos a uma filtragem, a fim de remover conteúdos com pouco valor semântico. Este processo consiste em separar os *tweets* em *tokens* (segmentos de sentenças), para logo em seguida, remover as menções, URLs e *emoticons* de cada um deles. Posteriormente, os *tokens* são normalizados, isto é, são submetidos a transformações (e.g., tratamento de pontuação, limpeza de caracteres especiais). Após a normalização, aqueles *tokens* que forem *stopwords* (palavras que podem ser considerados irrelevantes para o contexto estudado) são descartados e, por fim, os afixos dos *tokens* restantes são eliminados (*stemming*)

3.2. Extração de Características

Nesta etapa, nós selecionamos as principais técnicas de extração de características (TF-IDF, Unigramas e *Embedding*), a fim de comparar qual técnica fornecida como entrada para as redes neurais melhora o desempenho da classificação das categorias dos aspectos. Neste contexto, consideramos o total de dimensões extraídas por cada abordagem, de

forma que a rede pré-treinada utilizada possui 400 dimensões. A abordagem utilizando *bag-of-words* com TF-IDF construiu uma matriz com 2.480 dimensões e a abordagem com Unigramas obteve a mesma quantidade de dimensões, porém com valores em nível de atributo diferentes. Vale ressaltar que isso ocorre devido as abordagens terem formas diferentes de realizar o processo de extração de características.

Em relação as técnicas utilizadas, o TF-IDF (*Term Frequency and Inverse Document Frequency*) representa a distribuição ponderada dos termos, onde o TF representa a contagem de um termo t dentro de um documento d, e o IDF representa a distribuição de probabilidade observando o mesmo termo t, mas dessa vez observando a relação desse termo em toda base de dados. Os Unigramas, por sua vez, representam a contagem absoluta de um termo t dentro de toda a base de dados. Por fim, as redes de *Embeeding* assumem características distintas da estratégia de *bag-of-words*, pois enquanto o TF-IDF e os Unigramas assumem que os termos dentro de uma coleção são independentes, os *Embeedings* assumem que existe associações probabilísticas entre palavras e, com isso, estas redes conseguem construir processos de extração de características preservando a integridade do texto e com dimensões menores que a abordagem de *bag-od-words*.

3.3. Classificação

Após a construção da matriz, executamos a comparação de três redes profundas (CNN, RNN e LSTM), a fim de verificar qual apresenta melhor resultado na tarefa de classificação das categorias dos aspectos. Estas redes foram selecionadas porque observamos que nos trabalhos existentes no estado da arte, estas são as mais utilizadas para classificar tal análise semântica [Poria et al. 2016, Kim 2014, Nguyen and Shirai 2015]. Para obtenção e comparação de resultados, utilizamos os mesmos hiperparâmetros aplicados nas pesquisas mapeadas na seção dos trabalhos relacionados.

4. Experimentos e Resultados

A seguir, apresentamos uma descrição da base de dados utilizada neste trabalho, os experimentos realizados e os resultados obtidos após avaliação do método proposto.

4.1. Base de Dados

A base de dados utilizada neste artigo é disponibilizada no site SemEval (conferência focada em pesquisas de análise semântica). Neste contexto, esta base é composta por 2.290 avaliações de restaurantes[1] em inglês, onde os dados já se encontram particionados em amostras de treino e teste. Com isso, a base selecionada é composta por 1.708 amostras que serão utilizadas para treino e 582 amostras para teste. No total, foram reconhecidas 12 categorias referentes aos aspectos presentes, conforme ilustrado na Figura 2.

Conforme observado na Figura 2, as 3 categorias mais presentes na base de dados são FOOD#QUALITY, RESTAURANT#GENERAL e SERVICE#GENERAL, pois no contexto de restaurante é comum que os usuários comentem sobre qualidade de comida, nome de restaurantes e qualidade do serviço. Outra característica que pode ser destacada é que cada texto pode possuir mais de uma categoria presente em seu contexto. Com isso, adotamos a mesma estratégia de [Pontiki et al. 2016], onde o problema pode ser tratado com multi-rótulo, ou seja, a lista de rótulos será convertida em uma matriz de rótulos.

[1] `http://alt.qcri.org/semeval2014/task4/index.php?id=data-and-tools`

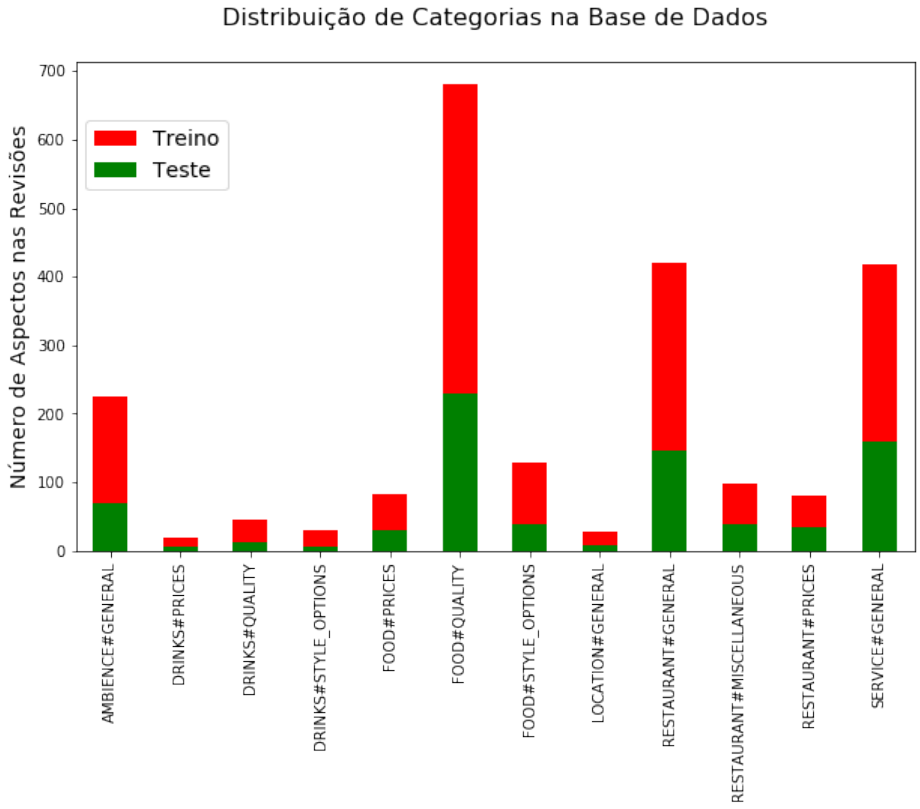


Figure 2. Distribuição dos dados dentro da coleção.

4.2. Classificação de Texto

Para a classificação de texto utilizamos a abordagem de organização dos rótulos de cada documento, adotando uma estratégia multi-classe. Assim, assumimos que cada documento pode possuir mais de uma categoria associada [Pontiki et al. 2016]. Para a tarefa treinamento utilizamos 3 redes profundas para classificação das categorias dos aspectos (LSTM, CNN, RNN) e, além disso, variamos os modelos de representação dados (TF-IDF, Unigramas e *Embeeding*), com o objetivo de verificar qual apresenta melhor desempenho estatístico segundo as métricas de avaliação. Durante o processo de classificação dos textos, utilizamos 100 épocas para medir o desempenho de aprendizado das redes profundas.

Os parâmetros utilizados nas redes podem ser observados na Tabela 1. Vale ressaltar que todos os parâmetros foram selecionados de forma empírica, onde alteramos esses parâmetros, a fim de verificar qual apresenta melhor resultado.

Table 1. Tabela de parâmetros usados nas redes profundas.

	Função de perda	Funções de ativação	Otimizador
LSTM	Categorical Crossentropy	Softmax	Rmsprop
CNN	Categorical Crossentropy	Relu/Softmax	Adam
RNN	Categorical Crossentropy	Softmax	Adam

4.3. Métricas de Avaliação

Para avaliar a eficácia dos classificadores na tarefa de identificação de categorias de aspectos, utilizamos as seguintes métricas: precisão, revocação, F1 e acurácia. Enquanto a precisão consiste na fração de documentos atribuídos a uma determinada classe que realmente pertencem a esta classe (segundo o conjunto de teste), a revocação representa a fração de todos os documentos que pertencem a uma determinada classe (segundo o conjunto de teste) e foram atribuídas corretamente a esta classe pelo classificador. Já a métrica F1 pode ser definida como uma medida que busca relacionar as métricas de precisão e revocação a fim de obter uma medida de qualidade que equilibre a importância relativa destas duas métricas. Esta medida pode ser atingida atraves da média harmônica entre a precisão e a revocação.

4.4. Resultados

Em nossos experimentos, medimos o desempenho estatístico das redes profundas alterando os modelos de representação de dados ao longo de 100 épocas, a fim de verificar se há melhora nos resultados da classificação dos textos. Para a avaliação do uso de *Embeeding*, utilizamos uma rede pré-treinada com 400 dimensões do Yelp, desenvolvida para realizar o processo de extração de características em avaliações. Para o uso do TF-IDF e dos Unigramas, utilizamos o processamento direto sobre os textos existentes em nossa base dados.

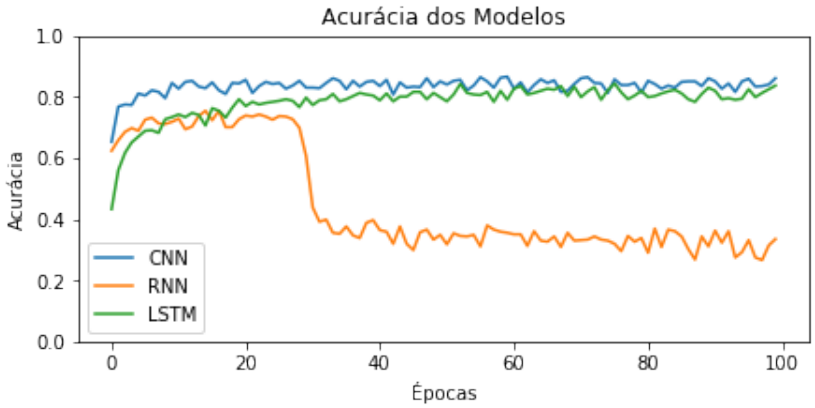
Como resultado, conseguimos observar que a rede profunda CNN apresentou melhor resultado em todas as variações de representação de dados (conforme ilustrado nas Figuras 3, 4 e 5), com destaque para o resultado com TF-IDF, onde obtivemos o melhor desempenho na classificação de categorias de aspectos (Figura 3(b)). Vale ressaltar que não eliminamos termos com baixa frequência e, com isso, a matriz de representação manteve 2.290 dimensões. Já o uso de Unigramas apresentou resultado similar ao do TF-IDF, porém é possível observar que a RNN não obteve resultados satisfatórios, pois nos três casos esta rede apresentou acurácia inferior a 50%.

Os experimentos que utilizam *Embeeding* em sua maioria apresentam resultados superiores ao uso de *bag-of-words* (TF-IDF ou Unigramas), pois essa abordagem tem uma característica de observar correlações entre os termos e, além disso, montar representações de dados com dimensões menores que as demais, preservando todas as características semânticas. Em nossos experimentos, foi possível observar que o uso de *Embeeding* apresentou uma medida F1 de 91%, enquanto o uso com TF-IDF apresentou 93%.

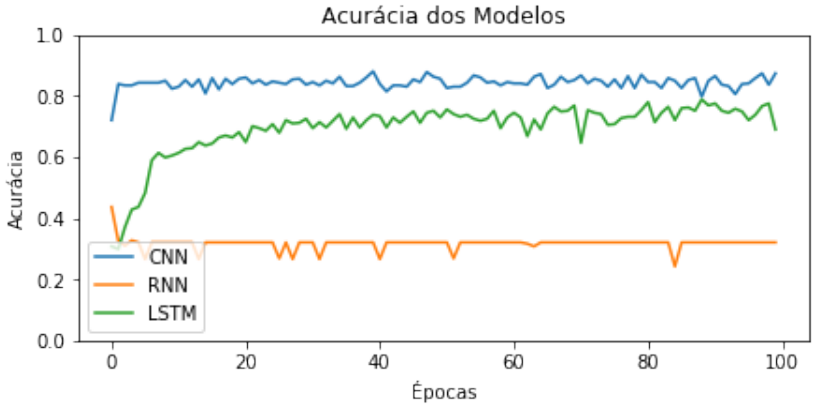
Vale ressaltar, que as métricas de precisão, revocação e F1 são apresentadas para a média das classes, pois estamos tratando esse problema com uma estratégia multi-classes. Assim, os valores representam todas as 12 classes detectadas. Outra característica importante que pode ser destacada é o aprendizado da rede profunda LSTM, pois com o passar das épocas é possível observar como sua acurácia aumenta. Com isso, é possível estimar que, algum ponto ao longo das épocas aconteça uma convergência de resultados com a rede que apresentou melhor resultado (CNN).

5. Conclusões e Diretrizes Futuras

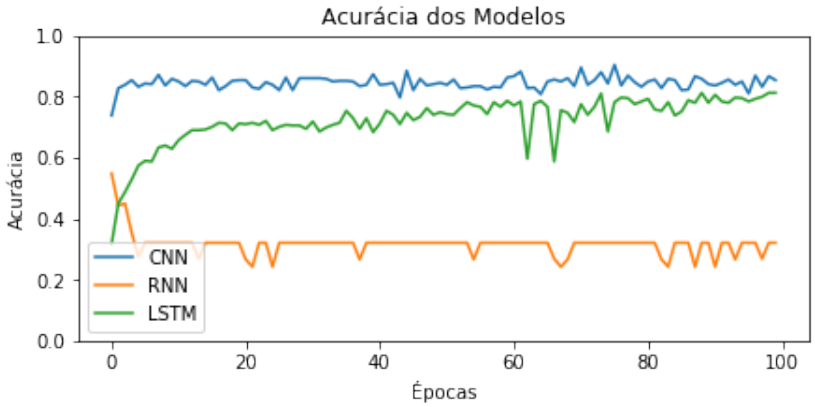
Neste trabalho, realizamos a classificação das categorias dos aspectos utilizando uma abordagem de aprendizagem profunda, onde comparamos 3 técnicas de representação



(a) Acurácia com uso de *Embeeding*.



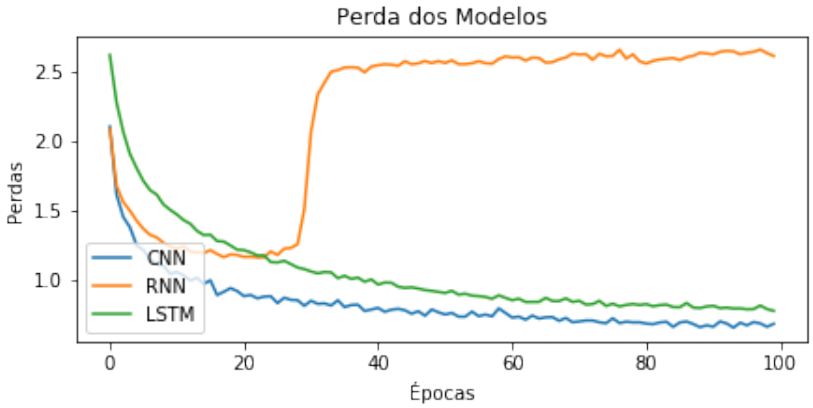
(b) Acurácia com uso de TF-IDF.



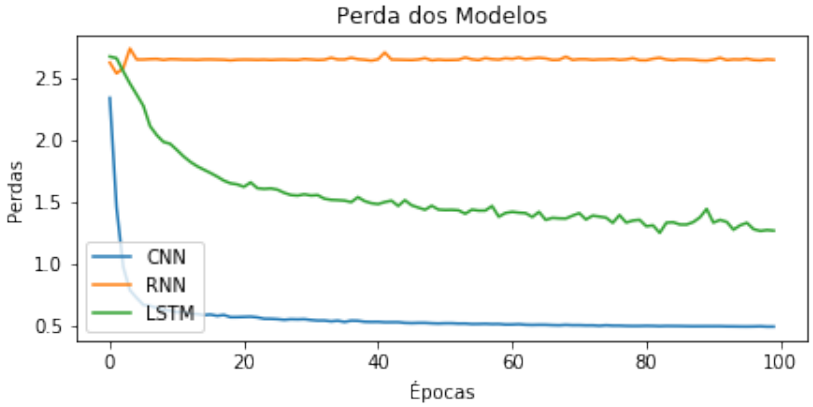
(c) Acurácia com uso de Unigramas.

Figure 3. Acurácia das redes variando os modelos de representação de dados.

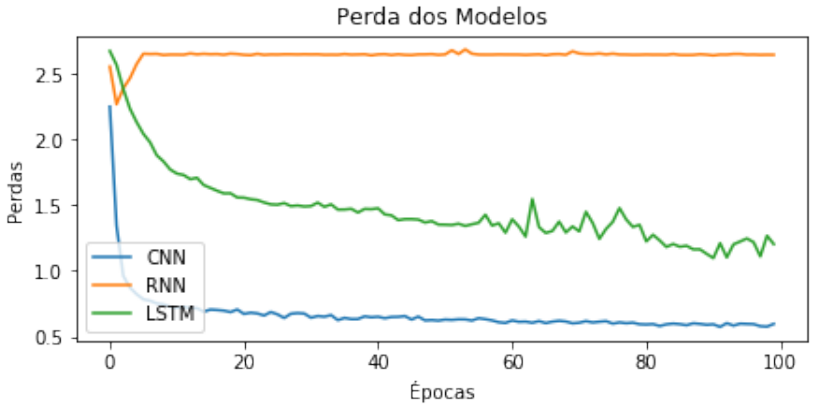
de dados (Unigramas, TF-IDF e *Embeedings*) combinadas as redes profundas, que tem sido utilizadas para realizar classificação na área de análise de sentimentos baseadas em aspectos no domínio de avaliação de restaurantes. Em nossos resultados, podemos observar que a abordagem utilizando TF-IDF combinada com a rede profunda CNN apresentou uma medida F1 de 0,93%, sendo superior ao uso de *Embeeding* com a mesma rede profunda. Porém, se observarmos em um contexto de todas as redes analisadas, o uso de



(a) Perdas com uso de *Embeeding*.



(b) Perdas com uso de TF-IDF.

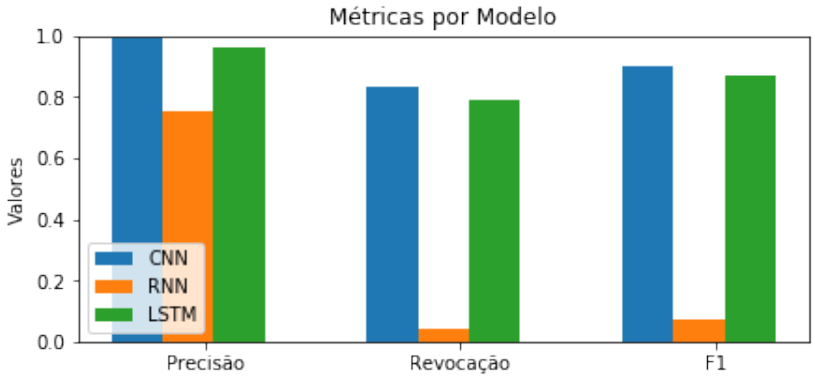


(c) Perdas com uso de Unigramas.

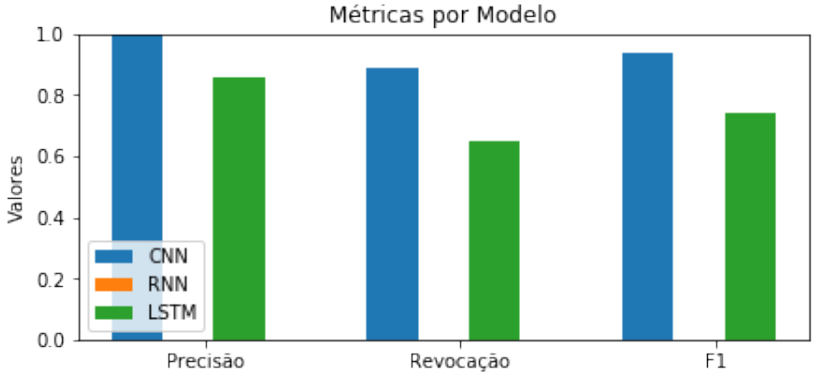
Figure 4. Perdas das redes variando os modelos de representação de dados.

Embeeding gerou uma capacidade de aprendizado mais eficiente nas 3 redes profundas que o uso de TF-IDF.

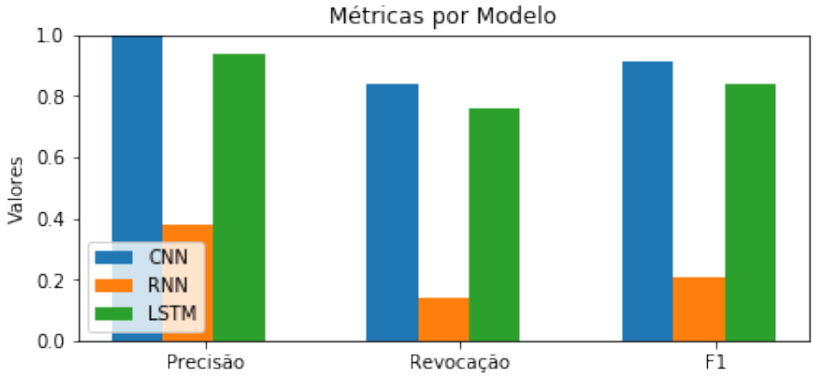
Vale ressaltar que tal classificação ainda apresenta muitas oportunidades de estudo, como o uso de outros modelos de representação de dados e a utilização de métodos estatísticos de extração de tópicos para reconhecimento de categorias. Com isso, como diretrizes futuras, podemos alterar as redes pré-treinadas de *Embedding*, a fim de veri-



(a) Métricas com uso de *Embeeding*.



(b) Métricas com uso de TF-IDF.



(c) Métricas com uso de Unigramas.

Figure 5. Métricas (Precisão, Revocação e F1) das redes variando os modelos de representação de dados.

ficar se a uma melhora na eficiência do modelo utilizando redes com maiores dimensões. Também podemos inserir a tarefa de análise de sentimentos sobre os aspectos detectados, com o objetivo de fornecer uma classificação com mais características sobre os textos analisados.

6. Referências

Almeida, T. G., Souza, B. A., Menezes, A. A., Figueiredo, C., and Nakamura, E. F. (2016). Sentiment analysis of portuguese comments from foursquare. In *Proceedings*

of the 22nd Brazilian Symposium on Multimedia and the Web, pages 355–358. ACM.

Araújo, M., Gonçalves, P., Benevenuto, F., and Cha, M. (2013). Métodos para análise de sentimentos no twitter. In *Proceedings of the 19th Brazilian symposium on Multimedia and the Web (WebMedia'13)*.

de Paula, H. L., Souza, B. A., Nakamura, F. G., and Nakamura, E. F. (2017). Quantificando a importância de emojis e emoticons para identificaç ao de polaridade em avaliaç oes online.

Gulaty, M. (2016). *Aspect-Based Sentiment Analysis*. PhD thesis, Dublin, National College of Ireland.

Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Nguyen, T. H. and Shirai, K. (2015). Phrasernn: Phrase recursive neural network for aspect-based sentiment analysis. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2509–2514.

Pavlopoulos, I. (2014). Aspect based sentiment analysis. *Athens University of Economics and Business*.

Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., Mohammad, A.-S., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., et al. (2016). Semeval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 19–30.

Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., and Androutsopoulos, I. (2015). Semeval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495.

Poria, S., Cambria, E., and Gelbukh, A. (2016). Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems*, 108:42–49.

Santos, R. L. d. S. and Moura, R. S. (2016). Extraç ao de métricas e analise de sentimentos em comentarios web no domınio de hotéis.

Souza, B. A., Almeida, T. G., Menezes, A. A., Nakamura, F. G., Figueiredo, C., and Nakamura, E. F. (2016). For or against?: Polarity analysis in tweets about impeachment process of brazil president. In *Proceedings of the 22nd Brazilian Symposium on Multimedia and the Web*, pages 335–338. ACM.

Stiilpen Junior, M. and Merschmann, L. H. C. (2016). A methodology to handle social media posts in brazilian portuguese for text mining applications. In *Proceedings of the 22nd Brazilian Symposium on Multimedia and the Web*, pages 239–246. ACM.

Wang, B. and Liu, M. (2015). Deep learning for aspect-based sentiment analysis.

Wang, W., Pan, S. J., Dahlmeier, D., and Xiao, X. (2016). Recursive neural conditional random fields for aspect-based sentiment analysis. *arXiv preprint arXiv:1603.06679*.

Xu, L., Lin, J., Wang, L., Yin, C., and Wang, J. (2017). Deep convolutional neural network based approach for aspect-based sentiment analysis. *Adv Sci Technol Lett*, 143:199–204.

Ye, Q., Zhang, Z., and Law, R. (2009). Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert systems with applications*, 36(3):6527–6535.