RECONHECIMENTO DE ENTIDADES NOMEADAS EM RELATÓRIOS DE INTELIGÊNCIA FINANCEIRA

Entity Named Recognition in Financial Intelligence Report

Jairo Santana[1], Diefferson K. Môro[1], Rogério de Aquino Silva[1], Vinicius F. C. Ramos[1], Gustavo Medeiros de Araújo[1]

(1) Universidade Federal de Santa Catarina, R. Eng. Agronômico Andrei Cristian Ferreira, s/n - Trindade, Florianópolis - SC, 88040-900

{jairo.santana, differson.moro, rogerriomp}@gmail.com, {v.ramos, gustavo.araujo}@ufsc.br

Resumo: O reconhecimento de entidades nomeadas é uma das subáreas do processamento de linguagem natural, mineração de textos e aprendizado de máquinas. Todas essas áreas fazem parte da grande área da inteligência artificial, muito utilizada em diversos problemas práticos do nosso dia a dia. Uma das competências da Polícia Federal é a investigação de crimes financeiros, em especial, a lavagem de dinheiro. Dentre os problemas encontrados na investigação policial, destacamos a análise dos Relatórios de Inteligência Financeira (RIF), escritos em português do Brasil, que são gerados pelo Conselho de Controle de Atividades Financeiras. O objetivo desta análise é identificar os atores envolvidos em esquemas de lavagem de dinheiro, mas, dependendo da complexidade do esquema, a identificação, por exemplo, desses atores e suas relações (sociedades, parentescos, “laranjas”, empresas “fantasmas”, etc) em um relatório, pode demandar um tempo significativo do policial envolvido na investigação. Este trabalho, portanto, visa apresentar resultados iniciais da automatização do reconhecimento de entidades nomeadas, importantes para a investigação policial, em RIFs. Identificamos, na literatura, uma grande lacuna para esse tipo de solução em textos em português. Os nossos resultados, ainda preliminares, demonstram que as ferramentas e os dados utilizados para o treinamento ainda precisam ser melhor trabalhados para que estes sejam mais significativos. Pudemos perceber que com poucos dados de treinamento conseguimos aumentar a precisão do reconhecimento de entidades de 14 para 27% e, em um teste com o framework RASA NLU, aumentamos a precisão para 60,98% de entidades reconhecidas corretamente, muito aquém dos 90% encontrados na literatura para outros idiomas.

Palavras-chave: Reconhecimento de Entidades Nomeadas; Mineração de Texto; Relatório de Inteligência Financeira; Processamento de Linguagem Natural.

Abstract: The named entity recognition is a subarea of natural language processing, text mining, and machine learning. These areas are part of the artificial intelligence area, very used in different kind of daily practical problems. One of the competencies of the Brazilian Federal Police is to investigate financial crimes, especially money laundering. Among the problems encountered in the police investigation, we highlight the analysis of the Financial Intelligence Reports, written in Brazilian Portuguese, which are generated by the Financial Activities Control Council. The aim of this analysis is to identify the actors involved in money laundering schemes, but depending on the complexity of the scheme, the identification, for example, of these actors and their relationships (societies, kinship, "oranges", "ghost" companies, etc.) in a report, may require significant time from the police man involved in the investigation. The main objective of this paper is to present initial results of the automation of named entity recognition, important for the police investigation, in Financial Intelligence Reports. We identified, in the literature, a large gap for this type of solution in Portuguese texts. Our preliminary results demonstrate that the tools and data used for training still need to be better explored to make them more meaningful. We could see that, with a few training dataset, we were able to increase the accuracy of the recognition of entities from 14 to 27% and, using the Rasa NLU framework, we got a 60.98% precision, very below the 90% found in the literature for other languages.

Keywords: Named Entity Recognition; Text Mining; Financial Intelligence Report; Natural Processing Language.

1. Introdução

Entre as diversas atribuições da Polícia Federal, encontra-se a investigação de crimes de lavagem de dinheiro (BRASIL, 1998), que pode ser definido como "a atividade de investir, ocultar, substituir ou transformar e restituir o dinheiro de origem ilícita aos circuitos econômico-financeiros legais, incorporando-o a qualquer tipo de negócio como se fosse obtido de forma lícita" (CALLEGARI, 2001 apud LUSTOSA, 2009). Lustosa (2009) refere-se a esse crime como uma maneira genérica de ocultação da origem de dinheiro ou bens advindos de atividades delitivas, seja por um processo ou um conjunto de operações, e a sua respectiva integração ao sistema financeiro por meio de operações capazes de converter o dinheiro sujo em dinheiro limpo.

Esse complexo esquema de lavagem de dinheiro pode ser dividido em três fases: ocultação, dissimulação e integração. A ocultação consiste basicamente em esconder e afastar o ativo da origem ilícita, para que se possa evitar que seja rastreado. Uma das maneiras de fazê-la é realizar diversas transações com valores fracionados, que desobriga a sua comunicação às autoridades financeiras (MENDRONI, 2015). A próxima etapa é a dissimulação, que tem por fim disfarçar a origem criminosa dos valores, camuflando evidências através de uma série de complexas transações financeiras internacionais em países que não cooperam com o combate à lavagem de dinheiro, os chamados *paraísos fiscais* (BRAGA, 2010). Por fim, a fase de integração é onde se tem os benefícios dos ativos como se fossem lícitos, seja através da compra de bens ou no investimento em empresas comerciais criadas e operando de forma legal.

O Conselho de Controle de Atividade Financeiras (COAF) é o responsável pela análise das evidências de crime de lavagem de dinheiro ao ser comunicado de operações suspeitas. Confirmadas as evidências, haverá uma troca de informações com as autoridades competentes (MARQUES, 2014), da forma que está descrito no artigo 15 da Lei nº 9.613/1998.

A Polícia Federal, o Ministério Público Federal, as polícias civis e os ministérios públicos estaduais são consideradas as autoridades competentes para receberem os informes do COAF, conforme o contexto da operação. Esses órgãos poderão bloquear a operação financeira suspeita, iniciar uma investigação criminal e, até mesmo, propor a ação penal (ARAS, 2007). O meio pelo qual o COAF disponibilizará seus informes às autoridades competentes é através do Relatório de Inteligência Financeira (RIF).

Os RIFs enviados para Polícia Federal são recebidos pela área de repressão a crimes financeiro e inicialmente avaliados, detectando qual será a delegacia responsável pela análise do relatório. O RIF será analisado detalhadamente de forma a identificar se realmente possuem indícios suficientes que retratam uma ação criminosa e, por fim, a instauração de um procedimento investigativo formal (POLÍCIA FEDERAL, 2013). O RIF é um documento escrito em linguagem natural, sem um padrão de estrutura e apresenta um relato conforme a compreensão do analista do COAF no formato digital PDF – Portable Document Format.

Atualmente, as áreas de inteligência da PF analisam o RIF de modo manual, identificando entidades, valores, operações realizadas e possíveis vínculos entre as entidades (POLÍCIA FEDERAL, 2013). Conforme a complexidade do relato no RIF, apenas uma leitura é feita para se realizar a análise ou, nos casos mais complexos, suas informações deverão ser tabuladas, armazenadas em planilhas eletrônicas ou em banco de dados, possibilitando sua leitura

por softwares analíticos que permitem gerar diagramas de relacionamentos entre as entidades, para, finalmente, facilitar a detecção de Organizações Criminosas. É primordial que a detecção e tabulação de entidades e outras informações aconteça com a máxima acurácia possível para não haver a possibilidade de se comprometer uma correta análise na fase posterior.

Nesse contexto, esse estudo tem como objetivo automatizar o processo de reconhecimento de entidade nomeada (REN), em especial, a identificação dos atores envolvidos nos RIFs utilizando ferramentas e técnicas de Processamento de Linguagem Natural (PLN), mineração de textos e aprendizado de máquina.

2. Trabalhos Relacionados

Pires (2017) apresenta em sua dissertação de mestrado alguns resultados sobre abordagens e configurações de ferramentas e corpus anotados em língua portuguesa para a pesquisa em um sistema de busca de notícias, da Universidade do Porto. O sistema de notícias chama-se SIGARRA. Segundo o autor, a ferramenta CoreNLP utilizada com o corpus HAREM obteve o melhor resultado para o Reconhecimento de Entidades Nomeadas (REN), chegando a 56% de precisão. Após essa identificação, o autor criou um conjunto de testes com 905 notícias anotadas e 12644 entidades anotadas. Nesse caso, o sistema implementado atingiu 86,86% em relação à *f-measure*.

Fonseca et al (2017) apresentam uma ferramenta para o REN baseada no *NameFinder*, uma das classes da ferramenta OpenNLP. Os autores utilizam o Corpus Amazônia para treinar o modelo e o Corpus HAREM para validar o modelo. Os resultados apontam para uma precisão de 58,65%, com recall de 56,60% e f-measure de 57,61% para as entidades do tipo Pessoa. Os autores ainda mostram resultados de diversas outras ferramentas de NER para o português e concluem que esses resultados são compatíveis com essas ferramentas, mas ainda aquém dos resultados encontrados em outros idiomas.

O diferencial do nosso trabalho é o uso de um contexto determinado, reconhecimento de entidades nomeadas em relatórios de inteligência financeira, o que poderá nos auxiliar na identificação dessas entidades. Além disso, propomos a comparação de outros modelos para o treinamento que não os apontados na literatura.

3. Procedimentos Metodológicos

O reconhecimento de entidades nomeadas (REN) é um dos principais elementos do PLN. Ele é essencial para várias etapas do PLN, dentre elas a classificação de uma frase ou a checagem de vínculos entre as entidades.

Neste contexto, este trabalho fez uma análise de algumas ferramentas e técnicas utilizadas para a REN em textos de inteligência financeira (RIF) escritos em português e compartilhados pelo COAF.

A primeira etapa do nosso processo é a identificação das ferramentas e técnicas utilizadas, através de uma pesquisa exaustiva na literatura, para o REN em língua portuguesa.

A segunda etapa visa a escolha e posterior utilização das técnicas e ferramentas identificadas na primeira etapa para o REN em relatórios de inteligência financeira (RIF). É importante ressaltar que as entidades mais importantes a serem identificadas no RIF são: pessoas (físicas e jurídicas), valores e datas.

As etapas do processo de REN seguem os algoritmos e técnicas utilizadas para tal finalidade, sendo as mais utilizadas as técnicas de aprendizado de máquina

supervisionados, dentre elas citamos o SVM, Redes Bayesianas e redes neurais. Uma das etapas é o treinamento realizados pelos algoritmos, para tanto, é necessário utilizar bases de dados com marcações de entidades para ajudar na identificação automática dessas entidades. Portanto, é fundamental encontrarmos, na literatura, as principais bases de dados anotadas em português.

Será necessária uma etapa intermediária para a conversão dos textos em PDF em texto puro para facilitar a manipulação dos dados.

As métricas utilizadas e que devem aparecer nos resultados são: i) a quantidade de palavras marcadas corretamente como entidade, ii) a quantidade de palavras marcadas indevidamente como entidade e iii) a quantidade de entidades omitidas.

4. Resultados

Abordamos a primeira etapa do nosso processo de revisão da literatura com uma revisão sistemática com o objetivo principal de identificação das principais ferramentas, técnicas e algoritmos utilizados para o reconhecimento de entidades nomeadas em textos em português. Esta etapa está em processo de publicação.

Identificamos na literatura 4 principais ferramentas de análise de textos que fazem o REN para o português, são elas: Stanford CoreNLP[1], OpenNLP[2], spaCy[3] e NLTK[4]. Além disso, identificamos, também, na literatura as principais bases de dados anotadas em português, que são: HAREM[5], BOSQUE[6], FLORESTA[7] e AMAZÔNIA[8].

Identificadas as principais bases de dados anotadas em português e as principais ferramentas utilizadas para o REN em português, procuramos por dados na literatura que nos permitisse escolher qual ferramenta e qual base de dados utilizar para o nosso objetivo final: o reconhecimento de entidades nomeadas em relatórios de inteligência financeira. Optamos, portanto, por utilizar a ferramenta spaCy e a base de dados HAREM por dois motivos principais: os resultados na literatura mostram o spaCy como uma boa ferramenta de REN e o HAREM é uma das bases de dados mais completas e anotadas para o português.

Alguns problemas foram superados e os relatamos para fins de pesquisa: a última versão do HAREM não é compatível com a última versão do spaCy (2.0.12), isto nos levou a desenvolver um script para a conversão dos dados para o spaCy (disponível após o blind review). O spaCy tem as entidades/categorias anotadas em inglês e o HAREM está anotado em português, portanto, fizemos a conversão das categorias no HAREM.

Para a conversão dos RIF do formato PDF para texto, utilizamos a biblioteca Pdfminer.

[1] Encontrada em https://nlp.stanford.edu/software/ acesso em 29 de setembro de 2018.
[2] Encontrada em http://opennlp.apache.org/ acesso em 29 de setembro de 2018.
[3] Encontrada em https://spacy.io/ acesso em 29 de setembro de 2018.
[4] Encontrada em https://www.nltk.org/ acesso em 29 de setembro de 2018.

[5] Encontrada em http://www.linguateca.pt/HAREM/PacoteRecursosSegundoHAREM.zip acesso em 25 de setembro de 2018.
[6] Encontrado em https://www.linguateca.pt/Floresta/ficheiros/Bosque_CP_7.5_cgde_22032016.conll.gz acesso em 29 de setembro de 2018.
[7] Encontrado em https://www.linguateca.pt/Floresta/ficheiros/FlorestaVirgem_CP.conll.gz acesso em 29 de setembro de 2018.
[8] Encontrado em https://www.linguateca.pt/Floresta/ficheiros/amazonia.conll.gz acesso em 29 de setembro de 2018.

Os nossos testes foram baseados nos seguintes modelos de treinamentos:

- Modelo 1 - modelo original disponibilizado no SpaCy, versão utilizada 2.0.0;
- Modelo 2 - modelo gerado com base no original, acrescido do treino de 500 iterações dos 'dados de treino' e 20 iterações nos dados dos arquivos de Localidades;
- Modelo 3 - modelo gerado com base no original, acrescido do treino de 20 iterações dos 'dados de treino';
- Modelo 4 - modelo em branco, acrescido do treino de 20 iterações dos 'dados de treino';
- Modelo 5 - novo modelo do HAREM, acrescido de 2 iterações com os dados de treinamento;
- Modelo 6 - novo modelo do HAREM, acrescido de 110 iterações com os dados de treino.
- Modelo 7 - modelo do spaCy com o treinamento junto ao framework Rasa MLU[9]

Segundo Pires (2017), os melhores resultados com o spaCy e o HAREM, para a classificação de entidades nomeadas como notícias em uma base de dados chamada SIGARRA, foram obtidos com 110 iterações de treinamento. Repetimos os mesmos treinamentos, entretanto, é importante ressaltar que as versões, da ferramenta e da base anotada, usadas no trabalho são diferentes.

Todos os RIF usados no treinamento tinham apenas 182 entidades marcadas, sendo 80 do tipo Pessoa (física ou jurídica), 38 do tipo Moeda, 14 do tipo Localidade, 38 do tipo Documento e apenas 12 do tipo Data.

Mostramos os resultados na Tabela 1, após os treinamentos, aplicamos o REN para um dos RIF.

Tabela 1: Resultados da Identificação.

Modelo	Precisão	Recall	f-score
1	1.041	1.587	1.257
2	0	0	0
3	27.586	38.095	32.000
4	14.754	28.571	19.459
5	8.641	11.111	9.722
6	27.272	33.333	30

Fonte: Elaborado pelo autor.

O modelo 7 foi testado em separado, pois não tínhamos total domínio do framework Rasa NLU, pois ele é utilizado em Chatbot para REN. O teste foi realizado a partir do modelo do spaCy e o Rasa NLU como treinamento em cima dos dados do RIF. Neste treinamento, conseguimos identificar 162 entidades, com uma precisão de 60,98%.

A precisão de todos os testes é muito baixa, entretanto, o Modelo 7 chega muito próximo aos valores encontrados por Fonseca et al (2017).

Sabemos que estes primeiros testes são importantes para reconhecermos o nosso corpus e, também, para entendermos os pontos de melhoria dos modelos e dos treinos. Esses resultados estão muito aquém dos resultados encontrados na literatura, seja para a língua portuguesa, seja para a língua inglesa. Cabe ressaltar que o Modelo 3 utilizou os mesmos dados usados para treinamento e validação dos resultados, por isso ele chegou a um resultado próximo ao Modelo 6.

[9] Encontrado em: https://rasa.com/docs/nlu/. Acessado em 01 de outubro de 2018.

5. Conclusão ou Considerações Finais

Compreendemos que a proposta de automatização de detecção de entidades nomeadas e seus vínculos em RIF é viável utilizando-se as ferramentas e bases de dados existentes. Entretanto, ainda não temos bons resultados com eles, visto que a identificação correta dessas entidades ainda não superou os 90% de acurácia, o que é um grande problema para a análise dos RIF por parte das instituições envolvidas, como é o caso da Polícia Federal. Os modelos treinados, apesar de um melhor desempenho, também não são satisfatórios. Sabemos que é necessário identificar mais entidades em outros RIF para aumentar a qualidade do treinamento. Isto deve levar ao REN com maior precisão, em especial as entidades do tipo Pessoa.

Uma outra proposta para aumentar a assertividade é a criação de expressões regulares e heurísticas que envolvam a identificação de entidades do tipo moeda, localização, data, CPF e CNPJ. Desta maneira, poderemos focar no reconhecimento de entidades do tipo pessoa física e jurídica, para, em um segundo momento estudarmos as correferências entre essas entidades.

Com trabalhos futuros, sugerimos a adaptação das outras bases de dados anotadas (FLORESTA, AMAZÔNIA e BOSQUE) para serem utilizadas como treinamento das principais ferramentas de NER: Stanford CoreNLP, OpenNLP, spaCy e NLTK. Após os testes com todas as bases e ferramentas, buscaremos melhorias nos algoritmos para conseguirmos uma acurácia acima dos 95%, conforme já é apresentado na literatura para o idioma inglês.

6. Referências

ARAS, V. Sistema nacional de combate à lavagem de dinheiro e de recuperação de ativos. Revista Jus Navigandi, no 1411, Teresina, 2007.

BRASIL. Lei da Lavagem de Dinheiro, 1998. Disponível em: <http://www.planalto.gov.br/ccivil_03/Leis/L9613.htm. Acesso em jun 2018.

CALLEGARI, A. L. Imputação Objetiva. Lavagem de dinheiro e outros temas do Direito Penal. Porto Alegre: Livraria do Advogado, 2001.

COAF. Relatório de Inteligência Financeira - RIF. Disponível em: <http://coaf.fazenda.gov.br/menu/a-inteligencia-financeira/relatorio-de-inteligencia-financeira-rif>. Acesso em: 1 jul. 2018.

FONSECA, Evandro B., CHIELE, Gabriel C., Vieira, Renata e VANIN, Aline A. Reconhecimento de Entidades Nomeadas para o Português Usando o OpenNLP. Anais do Encontro Nacional de Inteligência Artificial e Computacional (ENIAC 2015), s. pp. Disponível em: <http://www.lbd.dcc.ufmg.br/colecoes/eniac/2015/011.pdf>. Acesso em: 01 out. 2018.

LUSTOSA, D. S. de M. Aspectos gerais do crime de lavagem de dinheiro (Lei 9.613/98). Âmbito Jurídico, Rio Grande, XII, n. 70, 2009.

MARQUES, N. J. F. O papel do COAF no combate ao crime de lavagem de dinheiro. Conteudo Juridico, Brasilia-DF, 2014.

POLÍCIA FEDERAL. Manual Prático de Combate à Lavagem de Dinheiro e aos Crimes Financeiros (reservado) Brasília - DF, 2013.

PIRES, André R. O. Named entity extraction from Portuguese web text. Faculdade de Engenharia da Universidade do Porto, 2017.