Proposta recebida em Outubro 2022 e aceite para publicação em Janeiro 2023.

Extracção de Relações de Apoio e Oposição em Títulos de Notícias de Política em Português

Extraction of Support and Opposition Relationships in Portuguese Political News Headlines

David S. Batista

Resumo

Títulos de notícias de política relatam com frequência relações de apoio ou oposição entre personalidades, por exemplo: *"Marques Mendes critica estratégia de Rui Rio"* ou *"Costa reafirma confiança em Centeno."* Neste trabalho analisámos milhares de títulos arquivados, identificando os que expressam relações de apoio ou oposição e associando as personalidades políticas com o seu identificador na Wikidata, resultando assim num grafo semântico. O grafo permite responder a interrogações envolvendo personalidades políticas e partidos. Descrevemos o processo de geração do grafo e tornamo-lo disponível, assim como uma colecção de dados anotada manualmente, que permitiu treinar classificadores de aprendizagem supervisionada para identificar as relações expressas nos títulos e ligar as personalidades com a Wikidata.

Palavras chave

extracção de relações semânticas, dados anotados, web semântica, ciência política

Abstract

Political news headlines often report supportive or opposing relationships between personalities, for example: *"Marques Mendes criticizes Rui Rio's strategy"* or *"Costa reaffirms confidence in Centeno."* In this work we analyzed thousands of archived titles, identifying those that express supportive or opposing relationships, and associating the political personalities with their identifier on Wikidata, thus resulting in a semantic graph. The graph allows answering questions involving political personalities and parties. We describe the graph generation process and make it available together with a labelled dataset, which allowed supervised learning classifiers to be trained to identify the relationships expressed in the titles and link the personalities with Wikidata.

Keywords

semantic relationship extraction, annotated dataset, semantic web, political science

1. Introdução

Os títulos de notícias relacionados com política ou políticos relatam com frequência interacções envolvendo duas ou mais personalidades políticas. Muitas dessas interacções correspondem a relações de apoio ou oposição de uma personalidade para uma outra personalidade, por exemplo:

- *"Marques Mendes critica estratégia de Rui Rio"*
- *"Catarina Martins pede a demissão do governador Carlos Costa"*
- *"Sócrates foi às bases apelar ao voto em Soares"*

A análise de um grande número deste tipo de relações ao longo do tempo permite vários estudos, por exemplo: encontrar quais as grandes comunidades de apoio ou oposição em função dos governos no poder, ou encontrar as grandes alianças e oposições e as suas dinâmicas. Pode-se também explorar individualmente uma personalidade ao longo do tempo, por exemplo, comparando as relações de apoio ou oposição antes de tomar posse em determinado cargo público com as relações depois de ter assumido o cargo, ou ver que relações de apoio subitamente emergiram. Uma base de dados reunindo notícias expressando relações de apoio ou oposição entre personalidades políticas pode ser usada para rapidamente reunir uma colecção de notícias contendo ou envolvendo personalidades e partidos políticos específicos, por exemplo, para auxiliar numa tarefa de jornalismo de investigação.

Tendo um método automático para extrair relações e podendo aplicá-lo a uma colecção de dados abrangendo longos períodos de tempo permitiria concretizar os exemplos descritos anteriormente.

Neste trabalho apresentamos um método para extrair relações de apoio ou oposição entre personalidades políticas e descrevemos os resultados da aplicação do mesmo a uma colecção

DOI: 10.21814/lm.15.1.386

de notícias abrangendo um período de cerca de 25 anos. Durante o processo de extracção das relações ligámos as personalidades políticas envolvidas com o seu identificador na Wikidata (Malyshev et al., 2018) enriquecendo assim a relação com informação associada à personalidade (e.g.: afiliação política, cargos públicos exercidos, legislaturas, relações familiares, etc.).

Todas as relações extraídas são representadas sob a forma de triplos semânticos seguindo a norma *Resource Description Framework* (RDF) (Schreiber & Raimond, 2014). As personalidades políticas envolvidas, representadas pelo seu identificador na Wikidata, são ligadas através de uma relação de oposição ou apoio representada pela notícia que dá suporte à relação. Esta estrutura dá assim origem a um gráfico semântico, sendo então possível formular interrogações SPARQL (Prud'hommeaux et al., 2013) envolvendo a informação da Wikidata associada a cada personalidade e as relações extraídas dos títulos de notícias, por exemplo:

- Listar todas as notícias onde a personalidade X se opõe à personalidade Y
- Listar os membros de um determinado partido que apoiaram alguma personalidade específica
- Listar os membros de um determinado partido apoiados/opostos por membros de um outro partido
- Listar personalidades que estão ligadas através de uma relação familiar e de uma relação de oposição/apoio
- Listar personalidades que fazem parte do mesmo Governo e que estão envolvidas numa relação de oposição/apoio

As principais contribuições deste trabalho são:

- um grafo semântico ligando personalidades políticas representadas na Wikidata através de uma relação de oposição ou apoio suportada por uma notícia
- um conjunto de dados anotados utilizado para treinar os classificadores de extracção de relações sentimento direccionado de títulos de notícias, e também para ligar as personalidades mencionadas à Wikidata
- um interface web que permite explorar o gráfico semântico

Este artigo está organizado da seguinte forma: na Secção 2 referimos trabalho relacionado, na Secção 3 descrevemos a base de conhecimento usada no suporte de ligação das personalidades políticas à Wikidata. A Secção 4 refere e descreve as fontes de notícias utilizadas. Na Secção 5 detalhamos o conjunto de dados anotados e na Secção 6 os classificadores de aprendizagem supervisionada desenvolvidos. Na Secção 7 descrevemos o processo de extracção de triplos RDF e a construção do gráfico semântico. Finalmente, na Secção 8 reunimos as conclusões deste trabalho e apresentamos algumas ideias para trabalho futuro.

2. Trabalho Relacionado

A análise de sentimento, no contexto de Processamento de Linguagem Natural, tem sido maioritariamente alvo de estudo em conteúdo gerado em redes sociais (Zimbra et al., 2018) ou na avaliação de produtos ou serviços (Pontiki et al., 2016). Nestes domínios o autor do texto e o alvo da opinião são explícitos. No contexto de análise de notícias de política, onde existe com frequência um sentimento expresso entre actores políticos sob a forma de relações de apoio ou oposição (Balahur et al., 2009, 2010), as abordagens de análise de sentimento a produtos ou serviços não se aplicam, dado que a direcção da relação de sentimento tem que ser considerada.

Nesta secção descrevemos recursos semelhantes aos que produzimos neste trabalho, que tornamos públicos, e abordagens para a tarefa de extracção de sentimento direccionado em texto de notícias de política.

2.1. Recursos e dados anotados

Sarmento et al. (2009) propõem um método para a criação automática de um corpus para detecção de um sentimento positivo ou negativo para com uma personalidade política, e aplicam o método a comentários a notícias de jornais *on-line*. Neste recurso a origem do sentimento, pressupõem-se, é do comentador.

Moreira et al. (2013) disponibilizam uma ontologia descrevendo actores políticos, os seus cargos e partidos políticos afiliados, usando fontes de informação oficiais e informação recolhida da *web* para adicionar nomes alternativos às personalidades presentes na ontologia.

de Arruda et al. (2015) criaram um corpus de notícias políticas em português do Brasil, anotando cada parágrafo com o sentimento segundo duas dimensões: o actor político referido pelo parágrafo, e o sentimento dessa referência: positivo, negativo ou neutro. Neste recurso fica em aberto qual é a origem do sentimento. (Baraniak & Sydow, 2021) disponibilizam corpora se-

melhante, anotando o sentimento para com uma personalidade política em textos de jornais *online*, para o Inglês e Polaco.

2.2. Extracção de sentimento direccionado em texto de notícias

Vários autores exploraram métodos para extrair sentimento envolvendo actores políticos. De notar que muitos dos trabalhos transformam a tarefa de detecção de sentimento numa tarefa de detecção de uma relação entre entidades mencionadas (Bassignana & Plank, 2022).

Alguns exploram estas relações num contexto de política internacional, i.e.: os actores são nações referidas em texto de notícias de política, sendo que algumas dessas relações implicitamente têm um sentimento positivo ou negativo. O'Connor et al. (2013) propõem um modelo não supervisionado baseado em *topic models* e padrões linguísticos para identificar relações, de forma aberta, descrevendo conflitos entre nações referenciadas em artigos de notícias em Inglês. Han et al. (2019) propõem também um modelo não supervisionado para gerar descritores de relações para pares de nações mencionadas em notícias em Inglês. O modelo proposto estende o trabalho de Iyyer et al. (2016) integrando informação linguística (i.e.: predicados verbais e substantivos comuns e próprios) por forma a identificar o contexto das relações.

Liang et al. (2019) define a tarefa de extracção de relações de culpabilidade para textos em Inglês: dado um artigo d e um conjunto de entidades E, presentes no artigo, detectar se existe uma relação de culpabilidade (s, t), onde $s, t \in E$, quando s culpa t com base no artigo d, sendo que há $|E| \cdot (|E| - 1)$, possíveis relações de culpabilidade. Para detectar estas relações os autores propõem três modelos. O modelo *Entity Prior* extrai informação sobre entidades, tentando capturar um *prior* sobre quem é susceptível de culpar quem sem informação adicional. O modelo *Context* faz uso da informação de contexto da frase onde duas entidades ocorrem para determinar a presença de uma relação de culpabilidade. O modelo *Combined* combina a informação dois modelos anteriores num único modelo. Os autores aplicam esta abordagem num corpus com 998 artigos de notícias e com cerca de 3 entidades por artigo reportando uma macro-média F_1 de 0,70 com o modelo *Combined*.

Park et al. (2021) propõe uma estrutura de relações para detectar o sentimento e a direcção: dada uma frase s referindo duas entidades p e q, detectar qual a relação de sentimento entre p e q de entre cinco possíveis: neutra, p tem uma opinião positiva ou negativa em relação a q, ou q tem uma opinião positiva ou negativa em relação a p. No seu trabalho os autores usam múltiplos modelos transformando a tarefa de extracção de sentimentos em sub-tarefas que respondem a perguntas de sim/não para cada um dos 5 sentimentos possíveis, combinando depois os vários resultados num resultado final. Esta abordagem é aplicada para Inglês num corpus criado pelos autores contendo frases de artigos de notícias contento pelo menos duas entidades. Os pares de entidades estão anotadas com um dos 5 sentimentos possíveis. Os autores reportam uma macro-média de F_1 de 0,68.

3. Base de Conhecimento

Dado que as personalidades envolvidas nas relações a extrair são personalidades políticas relevantes, começámos por construir uma base de conhecimento a partir da Wikidata (Malyshev et al., 2018). Fazendo interrogações SPARQL ao *endpoint* público[1] recolhemos o identificador de todas as:

- pessoas que são ou foram afiliadas a algum partido político português
- pessoas portuguesas nascidas depois de 1935 cuja profissão seja: *juiz, economista, advogado, funcionário público, político, empresário ou banqueiro*
- pessoas que têm ou tiveram pelo menos um cargo de uma lista de cargos públicos portugueses previamente seleccionados (e.g.: *ministro, líder do partido, embaixador*, etc.)

Para além dos resultados destas interrogações, seleccionámos manualmente alguns identificadores de personalidades não abrangidas pelas interrogações SPARQL definidas acima, muitas delas de um contexto político internacional, mas que interagem com personalidades portuguesas. Acrescentámos também todos os identificadores de partidos políticos a que as personalidades recolhidas estão afiliadas. Este processo resultou num total de 1 757 personalidades e de 37 partidos políticos. De notar que alguns dos partidos incluídos são partidos já extintos e/ou de um contexto internacional. Para cada um dos identificadores das personalidades e partidos descarregámos a página correspondente na Wikidata utilizando um outro *endpoint*[2] público.

[1] https://query.wikidata.org

[2] https://www.wikidata.org/wiki/Special:EntityData?

Título	**Relação**
Sá Fernandes acusa **António Costa** de defender interesses corporativos	Ent_1-opõe-se-Ent_2
Joana Mortágua: declarações de **Cavaco** são "uma série de disparates"	Ent_1-opõe-se-Ent_2
Passos Coelho é acusado de imaturidade política por **Santos Silva**	Ent_2-opõe-se-Ent_1
Durão Barroso defende **Paulo Portas** como "excelente ministro"	Ent_1-apoia-Ent_2
Armando Vara escolhido por **Guterres** para coordenar autárquicas	Ent_2-apoia-Ent_1
Manuel Alegre recebe apoio de **Jorge Sampaio**	Ent_2-apoia-Ent_1
Rui Tavares e **Ana Drago** eleitos nas primárias do LIVRE	outra
Teresa Zambujo reconhece vitória de **Isaltino Morais**	outra
CDS acusa **Marcelo Rebelo de Sousa** de pôr em causa relação com **Cavaco**	outra

Tabela 1: Exemplos de títulos e das relações manualmente anotadas correspondentes.

Para cada personalidade política seleccionámos: o seu identificador na Wikidata, o seu nome mais comum e os nomes alternativos, i.e.: combinações de nomes próprios e apelidos. Com base nestes três campos, criámos um índice no ElasticSearch (Gormley & Tong, 2015) usando a sua configuração de omissão, não fazendo uso de qualquer funcionalidade extra tais como analisadores de n-gramas.

4. Fontes de dados

A principal fonte de notícias foi o arquivo da *web* portuguesa (Gomes et al., 2013). Usando a API pública de pesquisa recolhemos páginas arquivadas restringindo os resultados a ocorrências de nomes reunidos na Secção 3 e a 45 domínios `.pt` associados a diversas fontes de informação: jornais *on-line*, *websites* de estações de televisão e rádio, e portais agregadores de conteúdos.

Uma segunda fonte de notícias foi a colecção CHAVE[3](Santos & Rocha, 2004, 2001), contendo os artigos do jornal PÚBLICO publicados entre 1994 e 1995. Finalmente, foram também adicionados alguns artigos não arquivados pelo arquivo.pt, retirados directamente das secções *Mundo*, *Política* e *Sociedade* do site publico.pt.

Deste processo resultou uma colecção de cerca de 13,7 milhões de títulos de artigos publicados entre 1994 e 2022. De seguida foi aplicado um pré-processamento de modo a remover notícias com: títulos duplicados, títulos com menos de 4 palavras, e títulos ou URLs contendo palavras que fazem partem de uma lista pré-definida (e.g.: *desporto*, *celebridades*, *artes*, *cinema*, etc.) que sugerem um outro contexto que não política. Este pré-processamento resultou em 1,3 milhões de títulos distintos, cerca de 10% dos dados inicialmente recolhidos.

[3] https://www.linguateca.pt/CHAVE/

5. Colecção de Relações Anotadas

De forma a poder treinar classificadores de aprendizagem supervisionada para identificar as relações presentes nos títulos das notícias, e fazer a ligação das personalidades com a Wikidata, anotámos manualmente títulos com: as menções a personalidades, os identificadores na Wikidata e a relação entre as personalidades mencionadas.

Começámos por pré-processar todos os títulos recolhidos usando o pacote de software spaCy 3.0 (Honnibal et al., 2020), com o modelo `pt_core_news_lg-3.0.0` fizemos o reconhecimento de entidades mencionadas do tipo `PESSOA`. Para cada entidade reconhecida tentamos encontrar o seu identificador correspondente na Wikidata fazendo uma interrogação ao índice descrito na Secção 3 e assumindo que na lista de resultados o primeiro é o identificador correcto associado à entidade. Seleccionámos depois os títulos para anotação, incluindo apenas títulos referindo pelos menos duas personalidades.

No processo de anotação todos os títulos foram carregados para ferramenta de anotação Argilla,[4] e usando o interface gráfico fomos seleccionando títulos para anotar.

Para cada título corrigimos sempre que necessário as entidades reconhecidas e os seus identificadores na Wikidata, caso existam. Anotámos a relação existente: **oposição** ou **apoio**, e a direcção da mesma. Quando nenhuma das duas se verifica a relação é anotada como **outra**. A Tabela 1 demonstra alguns exemplos das relações anotadas. O processo de anotação foi feito por um anotador. Nas situações mais ambíguas, por exemplo, onde é necessária a informação no texto da notícia para decidir, as relações foram anotados como **outra**.

Este processo resultou num conjunto de dados contendo 3.324 títulos anotados. Para cada título anotámos apenas duas personalidades e a

[4] https://github.com/argilla-io/argilla

relação entre elas, mesmo que os títulos contenham referências a mais do que duas personalidades. A Tabela 2 caracteriza os dados em termos de número de relações e direcção. A maioria dos títulos contém uma relação de **oposição** ou **outra**, e a grande maioria das relações têm uma direcção da primeira para a segunda entidade, $\text{Ent}_1 \rightarrow \text{Ent}_2$.

Relação	**$\text{Ent}_1 \rightarrow \text{Ent}_2$**	**$\text{Ent}_1 \leftarrow \text{Ent}_2$**	**Total**
opõe-se	1 155	102	1 257
apoia	717	44	761
outra	-	-	1 306
Total	1 872	146	3 324

Tabela 2: Relações por classe e direcção.

O rácio de relações de oposição para com relações de apoio é de 1,6, este valor é semelhante com os dados para Inglês disponibilizados por Park et al. (2021), onde esta mesma relação entre as duas classes é de 1,8. Em termos de representatividade das classes, agregadas por sentimento, os dois conjuntos de dados também são semelhantes, sendo **outra** a classe mais presente, seguindo-se **oposição** e por último **apoio**.

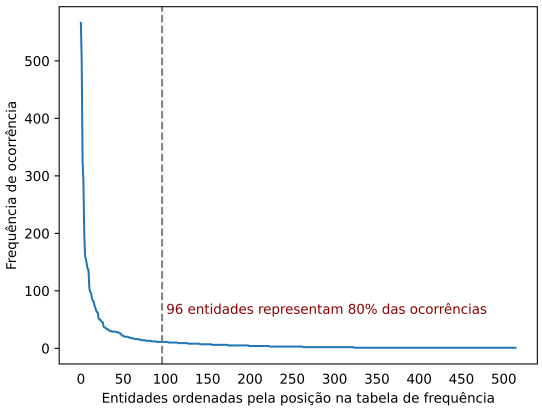

Figura 1: Distribuição de frequências de ocorrências das personalidades nos títulos anotados.

Das 6.648 menções a nomes de personalidades políticas anotadas, 515 são distintas e têm um identificador na Wikidata. Um total de 129 entidades distintas, identificadas por agregação da *string* que as refere no título, não estão associadas a um identificador por não estarem presentes na Wikidata.

Analisando a frequência de ocorrência de cada entidade observa-se que há pequeno número de entidades responsáveis por uma grande parte de todas as ocorrências de entidades nos dados anotados. Como mostra a Figura 1 existe um número pequeno de entidades frequentes, e uma longa lista de entidades pouco frequentes, em concreto, 96 personalidades distintas, i,.e.: 19% das personalidades, são responsáveis por 80% das menções a personalidades nos dados.

Em termos de número de palavras contidas nos títulos, excluindo palavras que fazem parte das entidades, há uma mediana de 8 palavras com um máximo de 22 e um mínimo de 1. Este conjunto de dados anotados encontram-se online[5] sob o formato JSON como ilustrado na Figura 2.

```
{"title": "Ana Gomes defende Durão Barroso",
 "label": "ent1_supports_ent2",
 "date": "2002-05-11 08:26:00",
 "url": "http://www.publico.pt/141932",
 "ent1": "Ana Gomes",
 "ent2": "Durão Barroso",
 "ent1_id": "Q2844986'",
 "ent2_id": "Q15849"}
```

Figura 2: Exemplo de um título anotado sob o formato JSON.

6. Processo de Extracção de Relações

O processo de extração de triplos RDF a partir dos títulos das notícias envolve 4 sub-processos:

- reconhecimento de entidades-mencionadas do tipo `PESSOA`
- ligação das entidades com um identificador na Wikidata
- classificação do tipo de relação
- classificação da direcção da relação

6.1. Reconhecimento de Entidades

O reconhecimento de entidades mencionadas é baseado num método híbrido, combinando regras com um modelo supervisionado.

Usando a componente `EntityRuler`[6] do spaCy 3.0, definimos uma série de regras combinando padrões baseados nos nomes de todas as personalidades da base de conhecimento descrita na Secção 3. Para detectar as entidades do tipo `PESSOA` este classificador aplica primeiro as regras e de seguida o modelo supervisionado para Português do spaCy. Em situações de desacordo entre as duas abordagens as entidades marcadas com regras têm prioridade.

[5] https://github.com/politiquices/data-releases
[6] https://spacy.io/usage/rule-based-matching

A Tabela 3 mostra a performance para as três abordagens sob o conjunto de dados anotado.

Abordagem	**P**	**A**	**F_1**
Regras	0,99	0,42	0,59
Modelo	0,97	0,91	0,94
Regras+Modelo	0,97	0,92	0,94

Tabela 3: (P)recisão, A(brangência) e F_1 para a componente de REM combinando regras e o modelos supervisionado.

6.2. Ligação com a Wikidata

O algoritmo para associar personalidades com identificadores na Wikidata tem duas fases. Numa primeira fase o algoritmo tenta apenas usar o título da notícia, se este processo falhar, tenta então usar possíveis referências às personalidades no texto da notícia.

O algoritmo começa por fazer uma interrogação à base de conhecimento (BC), usando a referência à personalidade presente no título, gerando assim uma lista de candidatos para uma determinada personalidade. Se a lista contém apenas um candidato e a similaridade (Jaro, 1989) para com a personalidade referida no título é de pelo menos 0,8 esse candidato é seleccionado. Se houver mais do que um candidato, o algoritmo filtra apenas aqueles com uma similaridade de 1,0 e se houver apenas um esse é o candidato seleccionado. Em qualquer outro caso nenhum candidato é retornado.

O Algoritmo 1 descreve o procedimento que usa apenas o título da notícia.

```
def title_only(ent, candidates):
    if len(candidates) == 1:
      if jaro(ent,candidates[0]) >= 0.8:
        return candidates[0]
    else:
      filtered = exact(ent, candidates):
        if len(filtered) == 1:
          return candidates[0]
    return None
```

Algoritmo 1: Ligação com a Wikidata usando apenas o título.

Se nenhum candidato for gerado na primeira fase ou nenhum for seleccionado da lista de candidatos o algoritmo tenta expandir as entidades mencionadas no título com base no texto da notícia, explorando um padrão: uma personalidade mencionada no título por uma versão curta do seu nome (e.g.: apenas o apelido) é normalmente referida no texto da notícia por um nome mais completo.

O algoritmo identifica todas as pessoas mencionadas no texto da notícia, usando a componente descrita na Secção 6.1, e seleciona apenas as que têm pelo menos um nome em comum com o nome da personalidade referida no título, gerando assim uma entidade expandida, e assumindo que corresponde à mesma entidade referida no título.

Se do processo resulta apenas uma entidade expandida e se há uma similaridade de 1.0 com um dos candidatos anteriormente seleccionados da BC, esse candidato é escolhido. Caso contrário a entidade expandida é usada para fazer uma interrogação à BC e recolher uma nova lista de candidatos. Se nessa lista apenas há um candidato e a sua similaridade é de pelo menos 0.8 para com a entidade expandida, esse candidato é escolhido. Se há mais do que um candidato e apenas um tem uma similaridade 1.0 com a entidade expandida, esse é escolhido.

```
def article_text(expanded, candidates):
  if len(expanded) == 1:
    filtered = exact(expanded[0], candidates)
        if len(filtered) == 1:
          return filtered[0]

    x_candidates = get_candidates(expanded)
    if len(x_candidates) == 1:
      if jaro(expanded, x_candidates[0])>=0.8:
        return x_candidates[0]

    filtered = exact(expanded, x_candidates)
    if len(filtered) == 1:
      return matches[0]

  if len(expanded) > 1:
    filtered = []
    for e in expanded:
      exact_candidates = exact(e, candidates)
          for c in exact_candidates:
            filtered.append(c)
    if len(filtered) == 1:
      return filtered[0]

  return None
```

Algoritmo 2: Ligação com a Wikidata usando o texto da notícia para expandir as entidades reconhecidas no título.

Se do processo de expansão resultam várias entidades expandidas, filtramos candidatos da BC com similaridade 1.0 para com a entidade expandida, se existir apenas um, esse candidato

é o escolhido. Em qualquer outro caso aqui não descrito nenhum candidato é selecionado.

O Algoritmo 2 descreve este procedimento usando o texto da notícia.

Os resultados desta abordagem sobre o conjunto de dados anotados são descritos na Tabela 4. A classificação *incorrecta* corresponde a personalidades que não foram associadas ao identificador correcto na Wikidata, *não desambiguada* para as que o algoritmo não conseguiu seleccionar um identificador único de entre todos os candidatos ou a BC não retornou nenhum resultado.

Na Tabela 4 são reportadas duas avaliações, a primeira coluna descreve os resultados para o algoritmo base, sem mapeamentos. A segunda coluna considera a ambiguidade que uma referência pode ter em termos de personalidades que representa. Por exemplo, nos dados anotados, todas as menções a *Cavaco* correspondem à personalidade *Cavaco Silva*, com base nisto o algoritmo mapeia todas as referências a *Cavaco* para *Cavaco Silva*. Da mesma forma, todas as menções a *Marques Mendes* correspondem à personalidade *Luís Marques Mendes*. Fazendo uso destes mapeamentos reduzimos o número de entidades para as quais o algoritmo não consegue encontrar um identificador.

Classificação	**base**	**mapeamentos**
correcta	5 059	5 136
incorrecta	43	43
não desambiguada	246	169
Exactidão	0,93	0,96

Tabela 4: Resultados da avaliação do algoritmo de ligação com a Base de Conhecimento.

6.3. Classificador de Tipo de Relação

Optámos por decompor a tarefa de classificação da relação em duas tarefas: classificação do tipo de relação e direcção da relação, por oposição a desenvolver um único classificador que teria que distinguir de entre 5 classes possíveis, e com classes muito desequilibradas em termos de representatividade. Esta secção descreve o classificador desenvolvido para detectar o tipo de relação presente num título, tendo três classes possíveis: **opõe-se**, **apoia** e **outra**. Todas as experiências foram feitas com uma avaliação cruzada de quatro partições.[7]

[7] https://github.com/politiquices/data-releases

Avaliámos diferentes abordagens para a classificação supervisionada das relações presentes nos títulos, nomeadamente: um classificador SVM (Cortes & Vapnik, 1995) com um *kernel* linear, uma rede neuronal recorrente do tipo LSTM (Hochreiter & Schmidhuber, 1997), e uma rede neuronal do tipo *transformer*, o DistilBERT (Sanh et al., 2019).

Para o classificador SVM utilizámos como *features* uma abordagem baseada em vectores TF-IDF (Salton & Buckley, 1988), fazendo um pré-processamento do título, usando um padrão, de modo a identificar o contexto relevante, i.e.: o contexto no título que contém informação que descreve a relação: `<Ent1 X Ent2 contexto>` onde `X` = {*"diz a", "responde a", "sugere a", "diz que", "afirma que", "espera que", "defende que", "considera que", "sugere que", "quer saber se", "considera", "manda"*}. Sempre que o padrão não se verifica usamos todas as palavras do título para construir o vector, excepto o nome das personalidades.

A rede neuronal recorrente LSTM foi usada numa arquitectura bidireccional, ou seja, são usadas duas redes LSTM, ambas com uma dimensão de 128, uma lendo o título da primeira para a última palavra e outra da última para a primeira palavra, sendo que os dois estados finais de cada LSTM são concatenados e passados a um *layer* linear. Usamos *embeddings* pré-treinados para Português baseados no método FastText (*skip-gram*) de dimensão 50 (Hartmann et al., 2017). A rede foi treinada por 5 *epochs* com um *batch size* de 8.

O modelo DistilBERT foi treinado tendo como base um modelo pré-treinado para o Português (Abdaoui et al., 2020), sendo depois afinado no conjunto de dados anotado, i.e.: os pesos de todos os *layers* pré-treinados foram actualizados tendo em conta a tarefa de classificação da relação. A rede foi treinada durante 5 *epochs* com um *batch size* de 8.

A Tabela 5 descreve os resultados para os vários classificadores. Não há diferenças muito acentuadas em termos de performance entre os 3 classificadores, embora a abordagem usando o DistilBERT tenha alcançado os melhores resultados. Ao analisar os resultados notámos que há relações difíceis de classificar correctamente, nomeadamente as que contêm expressões idiomáticas, por exemplo:

- *José Lello diz que Nogueira Leite quer "abifar uns tachos"*
- *Louçã diz que "António Borges é o grilo falante" de Passo Coelho*

Relação	P	A	F_1
opõe-se	0,71	0,69	0,70
outra	0,69	0,69	0,69
apoia	0,65	0,69	0,67
Macro-Média	0,69	0,69	0,69

(a) SVM com um kernel linear.

Relação	P	A	F_1
opõe-se	0,75	0,64	0,69
outra	0,65	0,75	0,70
apoia	0,65	0,62	0,63
Macro-Média	0,69	0,68	0,68

(b) LSTM bidireccional.

Relação	P	A	F_1
opõe-se	0,74	0,76	0,75
outra	0,72	0,71	0,72
apoia	0,72	0,71	0,71
Macro-Média	0,73	0,72	0,72

(c) DistilBERT pré-treinado em Português.

Tabela 5: (P)recisão, (A)brangência e F_1 para uma avaliação com 4-partições e validação cruzada com diferentes classificadores.

Outras relações são ambíguas e difíceis de classificar sem mais nenhum outro contexto do que aquele presente no título. No conjunto de dados que tornamos público, todos os títulos contêm um URL para o texto da notícia.

Os resultados obtidos com as abordagens descritas, para os dados em Português, estão em linha com os resultados reportados anteriormente em dados em Inglês (Liang et al., 2019; Park et al., 2021).

6.4. Classificador de Direcção da Relação

O classificador da direcção tem duas classes possíveis. Como mostra a Tabela 2, o conjunto de dados tem um viés para com a classe $Ent_1 \rightarrow Ent_2$ representando 91,5% dos dados. Assim, optámos por desenvolver uma abordagem baseada em regras para detectar apenas a classe $Ent_1 \leftarrow Ent_2$, e sempre que nenhuma das regras se verifica o classificador atribui a classe $Ent_1 \rightarrow Ent_2$.

Definimos regras baseadas em padrões construídos com informação morfológica e sintáctica (Nivre et al., 2020) extraída do título com o spaCy, usando o mesmo modelo que o descrito na Secção 5. Extraímos a informação morfo-sintáctica de todas as palavras, incluindo para os verbos informação sobre a conjugação: a pessoa e o número. Os padrões definidos foram os seguintes:

- **VOZ_PASSIVA**: procuramos por padrões `<VERB><ADP>`, um verbo seguido de uma proposição. Verificamos se a voz passiva está presente e envolve as personalidades mencionadas no título: se a entidade Ent_1 tem uma dependência para com o verbo do tipo **acl**, se o verbo tem uma dependência para com a Ent_1 do tipo **nsubj:pass** ou se o verbo tem uma dependência para com a Ent_2 do tipo **obl:agent**.
- **VERBO_ENT2**: detecta o padrão morfológico `<PUNCT><VERB>`Ent_2`<EOS>`, um sinal de pontuação seguido de um verbo, e terminando com a Ent_2, restringido o verbo a ser conjugado na 3ª pessoa do singular do presente do indicativo, e onde `<EOS>` representa o final do título, significando que Ent_2 é a última palavra no texto do título.
- **NOUN_ENT2**: verifica se o padrão `<ADJ>?<NOUN><ADJ>?<ADP>`Ent_2`<EOS>` está presente no título, i.e.: um substantivo podendo ser precedido ou sucedido de um ou mais adjectivos terminando com a Ent_2, sendo que o substantivo é restrito a uma lista de substantivos pré-definida.

A Tabela 6 mostra alguns exemplos de títulos de notícias e das regras que foram aplicadas para detectar a direcção $Ent_1 \leftarrow Ent_2$. As regras são aplicadas de forma sequencial, pela mesma ordem aqui descritas, se nenhum dos padrões é detectado no título o classificador atribui a classe $Ent_1 \rightarrow Ent_2$.

A Tabela 7 contém os resultados deste classificador para o conjunto de dados anotados.

Os resultados mostram que o método proposto classifica correctamente grande parte da direcção das relações $Ent_1 \leftarrow Ent_2$, a única classe para as quais foram desenvolvidas regras, sem prejuízo para com a classe $Ent_1 \rightarrow Ent_2$.

7. Grafo Semântico

Os componentes descritos na secção anterior formam o processo de extracção de triplos RDF a partir dos títulos de notícias recolhidos.

O processo de extracção começa por fazer o reconhecimento de personalidades no título da notícia e a sua ligação com o identificador de cada personalidade na Wikidata. O processo de extracção continua se ambas as personalidades reconhecidas foram ligadas com um identificador

Título	Regra Aplicada
Marques Júnior elogiado por **Cavaco Silva** pela "integridade de carácter"	VOZ_PASSIVA
Passos Coelho é acusado de imaturidade política por **Santos Silva**	VOZ_PASSIVA
António Costa vive no "país das maravilhas"acusa **Assunção Cristas**	VERBO_ENT2
Passos Coelho "insultou 500 mil portugueses", acusa **José Sócrates**	VERBO_ENT2
Maria Luís Albuquerque sob críticas de **Luís Amado**	NOUN_ENT2
André Ventura diz-se surpreendido com perda de apoio de **Cristas**	NOUN_ENT2

Tabela 6: Exemplos de títulos e respectivas regras usadas para detectar a direcção da relação.

Direction	P	A	F_1	#Títulos
$Ent_1 \rightarrow Ent_2$	0,99	1,00	0,99	1 488
$Ent_1 \leftarrow Ent_2$	0,95	0,84	0,89	129
weighted avg.	0,98	0,98	0,98	1 517

Tabela 7: (P)recisão, A(brangência) e F_1 usando 3 regras baseadas em padrões.

na Wikidata, caso contrário o título é descartado. O tipo de relação presente no título é detectado com o modelo DistilBERT. Se a relação entre as personalidades no título da notícia não for classificada como **outra** o classificador da direcção da relação é também aplicado ao título, caso contrário o título é descartado.

Para todos os títulos considerados o resultado final é um triplo RDF ligando as personalidades através de uma relação de oposição ou apoio suportada por uma notícia. Os triplos RDF gerados são indexados num motor SPARQL (Jena, 2015) juntamente com um sub-grafo da Wikidata descrito na Secção 3.

O grafo gerado tem um total de 680 personalidades políticas, 107 partidos políticos e 10.361 notícias cobrindo um período de 25 anos, Está disponível *on-line* no formato *Terse RDF Triple Language*[8] e poder ser também explorado através de um interface textitweb.[9]

8. Conclusões e Trabalho Futuro

Este trabalho descreve em detalhe o processo de construção de um grafo semântico a partir de títulos de notícias de política.

Através de interrogações SPARQL e fazendo referência às várias propriedades, retiradas da Wikidata de cada personalidade, consegue-se explorar relações de apoio e oposição através de agregações por partidos políticos, cargos públicos, governos constitucionais, assembleias constituintes, entre outras, podendo assim formular-se interrogações mais complexas, por exemplo: *"Ministros do XXII Governo Constitucional que foram opostos por personalidades do PCP ou BE."*, obtendo-se como resposta a lista de ministros e os artigos que dão suporte às relações de oposição vindas do BE.

Umas das limitações deste trabalho prende-se com o título da notícia não conter informação suficiente para perceber que tipo de relação ou sentimento existe de uma personalidade para outra, ou a presença de expressões idiomáticas, que tornam difícil a classificação automática. Como trabalho futuro gostaríamos de explorar o texto da notícia de forma a complementar o título para melhorar a detecção da relação. Com base também no texto da notícia as relações poderiam ser enriquecidas, categorizando-as em tópicos, dando mais uma dimensão à relação, um contexto para o sentimento de suporte ou oposição.

Alguns títulos contêm uma relação mútua, por exemplo: *"Sócrates e Alegre trocam acusações sobre co-incineração"* ou *"Pinto da Costa rebate críticas de Pacheco Pereira"*, poderiam ser classificados com a direcção $Ent_2 \leftrightarrow Ent_1$, indicando neste caso que ambas as personalidades se acusam mutuamente.

Este trabalho também deixa em aberto oportunidades de realizar diversos estudos com base na estrutura do grafo, por exemplo: encontrar comunidades de apoio e oposição em função do tempo e verificar quais as mudanças dentro dessas comunidades. Pode-se também estudar triângulos políticos: se duas personalidades políticas, X e Y, sempre acusam ou defendem uma terceira personalidade Z, qual será a relação típica expectável entre X e Y?

Agradecimentos

Gostaríamos de agradecer ao Nuno Feliciano por todos os comentários dados durante a elaboração deste trabalho e à equipa do Arquivo.PT por disponibilizar acesso aos dados arquivados através de uma API e pela consideração deste trabalho para os prémios Arquivo.PT 2021. Ao Edgar Fe-

[8] https://github.com/politiquices/data-releases
[9] https://www.politiquices.pt

lizardo e ao Tiago Cogumbreiro pelas revisões extensivas ao artigo, e também aos revisores Sérgio Nunes e José Paulo Leal por todos os comentários e correcções apontadas.

Referências

Abdaoui, Amine, Camille Pradel & Grégoire Sigel. 2020. Load what you need: Smaller versions of mutililingual BERT. Em *Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*, 119–123. 10.18653/v1/2020.sustainlp-1.16.

de Arruda, Gabriel Domingos, Norton Trevisan Roman & Ana Maria Monteiro. 2015. An annotated corpus for sentiment analysis in political news. Em *10th Brazilian Symposium in Information and Human Language Technology (STIL)*, 101–110.

Balahur, Alexandra, Ralf Steinberger, Erik van der Goot, Bruno Pouliquen & Mijail Kabadjov. 2009. Opinion mining on newspaper quotations. Em *International Joint Conference on Web Intelligence and Intelligent Agent Technology*, vol. 3, 523–526. 10.1109/WI-IAT.2009.340.

Balahur, Alexandra, Ralf Steinberger, Mijail Kabadjov, Vanni Zavarella, Erik van der Goot, Matina Halkia, Bruno Pouliquen & Jenya Belyaeva. 2010. Sentiment analysis in the news. Em *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, 655–662.

Baraniak, Katarzyna & Marcin Sydow. 2021. A dataset for sentiment analysis of entities in news headlines (SEN). *Procedia Computer Science* 192. 3627–3636. 10.1016/j.procs.2021.09.136.

Bassignana, Elisa & Barbara Plank. 2022. What do you mean by relation extraction? A survey on datasets and study on scientific relation classification. Em *60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, 67–83. 10.18653/v1/2022.acl-srw.7.

Cortes, Corinna & Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* 20(3). 273–297. 10.1007/BF00994018.

Gomes, Daniel, David Cruz, João Miranda, Miguel Costa & Simão Fontes. 2013. Search the past with the Portuguese Web Archive. Em *22nd International World Wide Web Conference*, 10.1145/2487788.2487934.

Gormley, Clinton & Zachary Tong. 2015. *Elasticsearch: The definitive guide*. O'Reilly Media.

Han, Xiaochuang, Eunsol Choi & Chenhao Tan. 2019. No permanent Friends or enemies: Tracking relationships between nations from news. Em *Conference of the North American Chapter of the ACL (NAACL)*, 1660–1676. 10.18653/v1/N19-1167.

Hartmann, Nathan, Erick Fonseca, Christopher Shulby, Marcos Treviso, Jéssica Silva & Sandra Aluísio. 2017. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. Em *11th Brazilian Symposium in Information and Human Language Technology (STIL)*, 122–131.

Hochreiter, Sepp & Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9(8). 1735–1780. 10.1162/neco.1997.9.8.1735.

Honnibal, Matthew, Ines Montani, Sofie Van Landeghem & Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in Python. 10.5281/zenodo.1212303.

Iyyer, Mohit, Anupam Guha, Snigdha Chaturvedi, Jordan Boyd-Graber & Hal Daumé III. 2016. Feuding families and former Friends: Unsupervised learning for dynamic fictional relationships. Em *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, 1534–1544. 10.18653/v1/N16-1180.

Jaro, Matthew A. 1989. Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical Association* 84(406). 414–420. 10.1080/01621459.1989.10478785.

Jena, Apache. 2015. A free and open source java framework for building semantic web and linked data applications. Available online: https://jena.apache.org/ (accessed on 20 November 2022).

Liang, Shuailong, Olivia Nicol & Yue Zhang. 2019. Who blames whom in a crisis? Detecting blame ties from news articles using neural networksa. Em *AAAI Conference on Artificial Intelligence*, vol. 33 01, 655–662. 10.1609/aaai.v33i01.3301655.

Malyshev, Stanislav, Markus Krötzsch, Larry González, Julius Gonsior & Adrian Bielefeldt. 2018. Getting the most out of Wikidata: Semantic technology usage in Wikipedia's knowledge graph. Em *17th International*

Semantic Web Conference (ISWC), 376–394. 10.1007/978-3-030-00668-6_23.

Moreira, Silvio, David S Batista, Paula Carvalho, Francisco M Couto & Mario J Silva. 2013. Tracking politics with POWER. *Program: electronic library and information systems* 47(2). 120–135. 10.1108/00330331311313708.

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers & Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. Em *12th Language Resources and Evaluation Conference (LREC)*, 4034–4043.

O'Connor, Brendan, Brandon M. Stewart & Noah A. Smith. 2013. Learning to extract international relations from political context. Em *51st Annual Meeting of the ACL*, 1094–1104.

Park, Kunwoo, Zhufeng Pan & Jungseock Joo. 2021. Who blames or endorses whom? entity-to-entity directed sentiment extraction in news text. Em *Findings of the Association for Computational Linguistics (ACL-IJCNLP)*, 4091–4102. 10.18653/v1/2021.findings-acl.358.

Pontiki, Maria, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra & Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. Em *10th International Workshop on Semantic Evaluation (SemEval)*, 19–30. 10.18653/v1/S16-1002.

Prud'hommeaux, Eric, Steve Harris & Andy Seaborne. 2013. SPARQL 1.1 query language. W3C Technical Report, http://www.w3.org/TR/sparql11-query.

Salton, Gerard & Chris Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management* 24(5). 513–523. 10.1016/0306-4573(88)90021-0.

Sanh, Victor, Lysandre Debut, Julien Chaumond & Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. Em *Fifth Workshop on Energy Efficient Training and Inference of Transformer Based Models (EMC^2)*, on–line.

Santos, Diana & Paulo Rocha. 2001. Evaluating CETEMPúblico, a free resource for Portuguese. Em *39th Annual Meeting of the Association for Computational Linguistics*, 450–457. 10.3115/1073012.1073070.

Santos, Diana & Paulo Rocha. 2004. CHAVE: Topics and questions on the Portuguese participation in CLEF. Em *Working Notes for CLEF*, on–line.

Sarmento, Luís, Paula Carvalho, Mário J. Silva & Eugénio de Oliveira. 2009. Automatic creation of a reference corpus for political opinion mining in user-generated content. Em *1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion*, 29–36. 10.1145/1651461.1651468.

Schreiber, Guus & Yves Raimond. 2014. RDF 1.1 Primer. W3C Technical Report, https://www.w3.org/TR/rdf11-primer/.

Zimbra, David, Ahmed Abbasi, Daniel Zeng & Hsinchun Chen. 2018. The state-of-the-art in Twitter sentiment analysis: A review and benchmark evaluation. *ACM Transactions on Management Information Systems* 9(2). 10.1145/3185045.