Detecção de Posicionamento em Dados Textuais

O estado da arte e considerações

José Lucas Silva Freitas
Universidade Federal de Campina Grande
Campina Grande, Paraíba
jslucassf@gmail.com

ABSTRACT

The task of stance detection in textual data can be understood as given some text, automatically identifiyng wheter the author is in favor of, against or even neutral about a certain target. The target may be an idea, a person, a company or even a product. For instance, automatically detecting a politician's positioning towards a certain government policy on his speeches. The text to be analysed can even come from a normal person posting in social media platforms like Twitter or Facebook.

In this work, many papers that address this problem were studied. Our goal being to better understand the nature of the problem, listing the techniques that were already used in dealing with it and identifiyng possible open areas that may be explored in future works.

The solution to this problem may be really important for various systems based on interaction with users, based on text written by humans. One of the main applications that were mentioned in the works that we studied was the detection of fake news, mostly due to the troublesome social and political moment, in which we see ourselves nowadays.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

KEYWORDS

Texto, Posicionamento, Linguagem, notícias

ACM Reference Format:
José Lucas Silva Freitas. 2018. Detecção de Posicionamento em Dados Textuais: O estado da arte e considerações. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, **??** pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUÇÃO

Politicamente e socialmente, vivemos tempos conturbados. A cada dia novos temas polêmicos são trazidos para discussão da sociedade e cada vez mais, as pessoas se sentem no dever de se posicionarem. Entender tais posicionamentos pode ser muito útil, para entender o espírito do posicionamento da sociedade em relação aos temas mais importantes. Além disto, entender o posicionamento das pessoas pode ter várias outras aplicações, como desenvolvimento de aplicações específicas, e mais recentemente ajudar na detecção de notícias falsas, as *fake news*

Em reconhecimento da importância deste problema, vários trabalhos utilizando diversas técnicas para detecção automática de posicionamento

foram publicados. A associação de linguística computacional (ACL) organiza anualmente o *SemEval*, [1] um workshop com vários desafios na área da análise de semântica computacional, dentre estes desafios, algumas tarefas relacionadas a detecção de posicionamento já foram propostas. Estudar a contribuição de todos estes trabalhos e os métodos por eles utilizados é muito importante pois possibilita o entendimento do estado da arte e a identificação de possíveis territórios de pesquisa ainda não explorados.

Iniciamos com uma visão geral da área da mineração de textos. Explorando o estado da arte no processamento de textos, mostramos de uma forma geral, como são realizadas as tarefas de processamento de texto. Então falamos um pouco mais sobre a aplicação em alguns contextos, especialmente das fake news. Prosseguimos para esclarecer as diferenças entre duas tarefas da análise semântica, a detecção de posicionamento e a detecção de sentimento. Nas próximas sessões, exploramos abordagens que usam os textos, regras, e outras informações relevantes, como as informações temporais e a estrutura das conversas (quando analisamos textos produzidos em redes sociais).

2 MINERAÇÃO DE TEXTO

A Mineração de texto é uma área multidisciplinar que envolve recuperação da informação, análise de dados, visualização, classificação, aprendizado de máquina e etc. Quando tentamos detectar o posicionamento dos autores em um texto, estamos implicitamente buscando extrair informações semânticas importantes e não triviais e isto é exatamente como podemos definir mineração de texto, o processo de busca e extração de informações importantes em texto. Desta forma, se torna muito importante entender inicialmente o estado atual do nosso entendimento em técnicas e conceitos de mineração de texto, pois os mesmos serão aplicados no contexto do nosso problema.

O processamento de informações armazenadas de forma não estruturada é essencial para negócios em geral. Alguns estudos [Referência] já demonstraram que cerca de 80% de todas as informações potencialmente úteis para negócios está armazenada em formato não estruturado. Como texto sempre foi a forma mais natural de transmissão e armazenamento de conhecimento humano, a maior parte destes dados não estruturados está em formato textual, portanto o desenvolvimento de técnicas de processamento de texto possui um enorme potencial de geração de valor, tanto para empresas quanto para usuários finais das novas tecnologias que poderão ser desenvolvidas.

2.1 Um Framework Conceitual

[Referência] apresenta um framework conceitual para tarefas de mineração de texto que consiste em dois componentes:

(1) *Refinamento de texto*
(2) *Destilação de conhecimento*

A etapa de **refinamento de texto**, consiste em uma etapa inicial, onde ocorre a transformação dos documentos de texto em um tipo de representação intermediária de mais fácil processamento, alguns exemplos são organizações relacionais dos documentos, grafos de conceitos ou até contagem de palavras. O formato de representação intermediária pode ser baseado em documentos, onde cada entidade nos dados representa um documento ou em conceitos, onde cada entidade representa um conceito ou objeto de interesse no estudo.

Já na etapa de **destilação de conhecimento**, a representação intermediária dos documentos, produzida na fase anterior, é finalmente processada, possibilitando a dedução dos padrões e informações desejadas. Nesta etapa, várias técnicas de processamento são utilizadas, como visualização de documentos, análise de texto.

A ideia principal na abordagem baseada em visualização, é o agrupamento de documentos com base em suas similaridades e apresentar os grupos definidos fazendo uso de algum tipo de representação gráfica. A interpretação da semântica dos

[1] https://aclweb.org/aclwiki/SemEval_Portal

grupos ainda é uma tarefa a ser realizada por humanos.

Análise de texto é na verdade apenas um termo genérico que engloba muitas abordagens diferentes. Processamento de Linguagem Natural (NLP), Sumarização, extração de informação e etc. Aqui destacamos o crescente uso de técnicas de aprendizado de máquina, que vem se popularizando na pesquisa atual, devido ao avanço no poder de processamento dos computadores atuais.

2.2 Problemas em Aberto

Apesar dos avanços nos últimos anos, ainda enfrentamos alguns desafios na área. O concebimento de um formato intermediário eficiente para o processamento, o desenvolvimento de técnicas e algoritmos capazes de produzirem soluções mais independentes da linguagem e a exploração de formas de fazer uso de informações relativas ao domínio específico dos problemas, são exemplos de aspectos que necessitam de novas contribuições para no futuro.

3 FAKE NEWS E APLICAÇÕES

Sem dúvidas, a maior aplicação das técnicas de detecção de posicionamento atualmente está na verificação das chamadas *Fake News*. As notícias falsas tem causado muitas preocupações, principalmente em cenários eleitorais onde elas são suspeitas até mesmo de influenciar o resultado de campanhas presidenciais. [2]

Desta forma, existe o interesse em identificar a veracidade de notícias veiculadas na internet, mas como uma quantidade muito grande de informações circulam a cada dia, a automatização deste processo é desejada.

Vários estudos [Referência] sugerem que detectar o posicionamento dos autores das notícias e de outros portais pode ser um primeiro passo essencial para verificar a veracidade das mesmas. Profissionais da academia e várias empresas ao redor do mundo realizam o *Fake News Challenge* [3], um desafio que objetiva abordar este problema, e também sugere a detecção de posicionamento como primeiro passo para a verificação de veracidade de notícias.

Entretanto as aplicações podem ser ainda maiores. Alguns trabalhos [Referência] já fizeram uso de estruturas similares para realizar não só a verificação das notícias oficialmente publicadas, mas também para boatos que circulam as redes sociais.

4 DETECÇÃO DE POSICIONAMENTO X DETECÇÃO DE SENTIMENTO

A detecção de sentimento em texto, é um campo com vários trabalhos publicados, porém é importante ressaltar que como citado por [Referência], detecção de posicionamento é uma tarefa relacionada mas diferente da análise de sentimento. Em tarefas de análise de sentimento, buscamos determinar se um texto é positivo, negativo ou neutro. Geralmente extraímos também o alvo daquela opinião. Entretanto para detecção de posicionamento, buscamos identificar a concordância com um assunto pré-estabelecido. O assunto alvo da opinião pode não ser mencionado no texto, ou até não ser o sujeito do texto.

Estas são tarefas distintas porém são relacionadas. Em [Referência] por exemplo, regras de sentimento são utilizadas para auxiliar na detecção de posicionamento. Uma pergunta permanece em aberto entretanto, qual o nível de correlação entre as duas tarefas?

5 ABORDAGENS BASEADAS APENAS NO TEXTO

Durante a primeira etapa no framework que foi citado, tomamos decisões sobre o formato dos

[2] https://www.washingtonpost.com/news/the-fix/wp/2018/04/03/a-new-study-suggests-fake-news-might-have-won-donald-trump-the-2016-election/

[3] http://www.fakenewschallenge.org/

dados que serão processados, ou seja, decidimos sobre a representação intermediária. Uma decisão que pode ser feita, é se usaremos apenas os dados textuais dos documentos ou se faremos uso de outros tipos de informação que sejam releventes. Em [Referência], os autores realizam a classificação de *tweets* realizando o processamento apenas no conteúdo textual dos mesmos. A abordagem faz uso de diferentes classes de n-gramas como representação intermediária e gera modelos preditivos utilizando um algoritmo de classificação (SVM).

6 REGRAS DE CLASSIFICAÇÃO

O trabalho citado anteriormente [Referência] também faz uso de regras de sentimento em uma primeira etapa do processo de classificação. Regras de sentimento. Estas, fazem uso da presença de alguns tipos de n-gramas no texto que podem ser indicativos do sentimento do autor para com os alvos da opinião. Os n-gramas alvo, são aqueles usados para representar direta ou indiretamente os alvos da questão. Por exemplo, nomes próprios. Os n-gramas chave, são os que servem para deixar explícito a opinião do autor. Por exemplo *hashtags* ou termos como "readyforhillary".

Cada n-grama é então separado em dois tipos, favoráveis e contrários em relação ao alvo principal. Desta forma, criam-se pares ngrama-tipo, que compõem as regras de sentimento utilizadas para classificar os tweets. As regras são no formato:

- Presença de n-grama chave favorável e ausência de n-grama chave contrário implica em posicionamento Favorável
- Presença de n-grama chave contrário e ausência de n-grama chave favorável implica em posicionamento Contrário

O conjunto das regras geradas, são então utilizadas para classificar o posicionamento dos tweets.

Este é um processo semi automático pois necessita de supervisão humana para a anotação dos n-gramas, e no trabalho citado é utilizado como um primeiro passo para rotulação um conjunto de dados que será útil para treinar um modelo preditivo.

Ainda sobre análise de sentimento, esta abordagem se apoia na suposição de que o posicionamento pode ser expresso na forma de sentimentos positivos ou negativos em relação aos alvos do texto.

7 UTILIZANDO OUTRAS INFORMAÇÕES CONTEXTUAIS

Outros trabalhos abordam o problema de detecção de posicionamento fazendo uso de outras informações além dos textos sendo analisados. [Referência] se aproveita da estrutura das conversas em redes sociais enquanto [Referência] considera informações temporais das postagens no Twitter.

7.1 Estrutura das Conversas em Redes Sociais

As redes sociais possibilitam uma estruturação nas conversas, que segundo [Referência], pode influenciar o poder de classificação de nossos métodos. A intuição, é que após um rumor ser compartilhado, as postagens resposta ao texto original podem auxiliar a verificação da informação pois seus posicionamentos podem representar contra argumentações ou apoios à ideia inicial. Outro trabalho que também faz uso de informações relacionadas a estrutura das conversas, é [Referência]. Nele os autores ainda chamam atenção sobre o fato de que as abordagens baseadas em regras, que foram citadas anteriormente, possuam a falha de não poderem ser generalizáveis para diferentes contextos.

7.2 Informações temporais

Em [Referência] os autores estudam mais uma vez dados de redes sociais e levantam a hipótese de que a utilização de dados temporais afeta positivamente a precisão dos métodos de detecção. Uma ideia importante é que a ocorrência de um

tweet influencia a taxa em que futuros tweets serão publicados.

8 CONCLUSÕES

Este artigo buscou analisar a literatura com o objetivo de organizar o conhecimento relacionado ao estado da arte na pesquisa de detecção de posicionamento em dados textuais, suas aplicações e desafios. Vimos que esta é uma tarefa com aplicações especiais na verificação de veracidade em notícias e rumores.