Inteligência Analítica Aplicada ao
Processo de Participação Social

Análise de Sentimentos

InWeb/UFMG CEWEB/NIC.BR MJ

Agosto de 2015

Sumário

1 Contexto **2**
1.1 Descrição . . . 2
1.2 Motivação . . . 2
1.3 Entrada . . . 2
1.4 Saída . . . 2
1.5 Estratégia . . . 2

2 Solução Proposta **3**
2.1 Análise de Sentimentos . . . 3
2.2 Extração de Regras *Offline* . . . 3
2.3 Predição de Sentimentos . . . 4
2.3.1 Extração de Regras *Online* com Projeção de Dados . . . 5
2.3.2 Estendendo Modelos de Classificação Dinamicamente . . . 5

3 Estudo de Caso **6**
3.1 Preparação dos Dados . . . 6
3.2 Resultados Preliminares . . . 6

4 Próximos Passos **7**

1 Contexto

1.1 Descrição

O Anteprojeto de Lei de Proteção de Dados Pessoais segue um processo complexo no qual há participação social na construção e elaboração da lei. Nesse caso, uma proposta de lei foi enviada a debate público e cidadãos enviaram contribuições ao anteprojeto de lei por meio de comentários. Embora tal processo de elaboração de leis seja mais transparente ao público, ele exige a capacidade de absorver e avaliar centenas ou até mesmo milhares de comentários. Nós abordamos o problema de reduzir o trabalho manual necessário para analisar comentários através da classificação automática dos mesmos. Tal classificação é feita de acordo com aspectos e opiniões (i.e., positivas ou negativas) expressos nos comentários. Nesse sentido, definimos um aspecto como sendo qualquer questão específica da proposta de lei, sob a qual os cidadãos podem expressar opiniões.

1.2 Motivação

Classificar comentários com base em aspectos e opiniões oferece uma alternativa eficiente para o entendimento das contribuições e a absorção de sugestões. O objetivo fundamental da tarefa de classificação de comentários é permitir a identificação de reclamações, reivindicações e argumentações pertinentes, o que pode adicionar informações valiosas capazes de mudar a percepção sobre certos aspectos da proposta de lei.

1.3 Entrada

Tendo em vista que os comentários são extensos e numerosos, optamos por produzir classificadores utilizando técnicas de Aprendizado de Máquina e Processamento de Linguagem Natural. Dessa forma, a entrada consiste de um conjunto de comentários. Uma pequena parte do conjunto de comentários devem servir como fonte de exemplos, ou seja, devem vir acompanhados de rótulos que identifiquem os aspectos de interesse no comentário, bem como a opinião associada ao aspecto (i.e., positiva ou negativa).

1.4 Saída

A saída consiste dos mesmos comentários recebidos como entrada, porém acompanhados de rótulos relacionados a aspectos e opiniões. Esses rótulos são associados automaticamente aos comentários durante o processo de classificação. Dessa forma, o analista poderá navegar pelo comentários seguindo filtragens por aspectos e opiniões, bem como focalizar seus esforços em aspectos associados a opiniões majoritariamente negativas.

1.5 Estratégia

Nossa estratégia para classificação automática baseada em aspectos e opiniões é baseada no uso de regras de associação [1]. Tal decisão é suportada pela necessidade de interpretar as classificações. Mais especificamente, o analista estará interessado não apenas em realizar a classificação, mas também em entender o porque da opinião ser negativa ou positiva. Modelos de classificação baseados em regras de associação são facilmente interpretáveis, além de serem altamente precisos.

2 Solução Proposta

Esta seção provê definições para os principais conceitos empregados neste projeto.

2.1 Análise de Sentimentos

A tarefa de análise de sentimentos, no contexto deste projeto, é definida como a seguir. Temos como entrada uma pequena semente de treinamento (referenciado como $\mathcal{D}$), o qual consiste de um conjunto de registros na forma de $< d, s_i >$, onde d é um comentário (representado como uma lista de termos, q_1, q_2, ..., q_n) e s_i é o sentimento implícito em d. Comentários em $\mathcal{D}$ são unicamente identificados e a variável sentimento s assume seus valores de um conjunto pré-definido e discreto de possibilidades (e.g., s_1, s_2, ..., s_k).

A semente de treinamento é utilizada para construir uma função relacionando padrões textuais nos comentários aos seus respectivos sentimentos. Uma sequência de comentários futuros, ordenados cronologicamente, (referenciadas como $\mathcal{T}$) consiste de um registro $< t, ? >$ para o qual somente os termos no comentário t são conhecidos, enquanto o sentimento expresso em t é desconhecido. Os modelos de classificação obtidos a partir de $\mathcal{D}$ são utilizados para mensurar os sentimentos para cada comentário em $\mathcal{T}$.

Definição 1. Uma regra de sentimento é uma regra de associação especializada $\mathcal{X} \rightarrow s_i$, onde o antecedente $\mathcal{X}$ é um conjunto de termos, e o consequente s_i é o sentimento previsto. O domínio para $\mathcal{X}$ é o vocabulário de $\mathcal{D}$. A cardinalidade da regra $\mathcal{X} \rightarrow s_i$ é dada pelo número de termos no antecedente, que é $|\mathcal{X}|$. O suporte de $\mathcal{X}$, que é denotado como $\sigma(\mathcal{X})$, é o número de comentários em $\mathcal{D}$ tendo $\mathcal{X}$ como um subconjunto. A confiança da regra $\mathcal{X} \rightarrow s_i$, denotada como $\theta(\mathcal{X} \rightarrow s_i)$, é a probabilidade condicional do sentimento s_i dados os termos em $\mathcal{X}$, que é calculada de acordo com a Equação 1.

$$\theta(\mathcal{X} \rightarrow s_i) = \frac{\sigma(\mathcal{X} \cup s_i)}{\sigma(\mathcal{X})} \tag{1}$$

2.2 Extração de Regras *Offline*

A abordagem mais simples para a aprendizagem de sentimento utilizando regras de sentimento é a *offline*, onde um conjunto de regras é extraído a partir dos dados de treinamento $\mathcal{D}$, e então, essas regras compõem o modelo de classificação.

Definição 2. O modelo de classificação é denotado como $\mathcal{R}$ e este é composto por um conjunto de regras $\mathcal{X} \rightarrow s_i$ extraída de $\mathcal{D}$. O modelo é representado como um conjunto de entidades na forma $<$*chave*, *valor*$>$, onde *chave*$=\{\mathcal{X}, s_i\}$ e *valor*$=\{\sigma(\mathcal{X}), \sigma(\mathcal{X} \cup s_i), \theta(X \rightarrow s_i)\}$. Cada entidade no conjunto corresponde a uma regra e a *chave* é utilizada para viabilizar o acesso rápido às propriedades das regras.

O processo de extração é dividido em 2 passos: contagem para definição do suporte e cálculo da confiança. Uma vez que o suporte $\sigma(\mathcal{X})$ é conhecido, é simples computar a confiança $\theta(\mathcal{X} \rightarrow s_i)$ para a regra correspondente [3].

Geralmente, o cálculo do suporte para o conjunto de termos em $\mathcal{D}$ inicia com a exploração de todos os comentários em $\mathcal{D}$ e cálculo do suporte de cada termo isoladamente. Na próxima

iteração, conjuntos de tamanho 2 são enumerados utilizando os conjuntos de termos de tamanho 1, e seus valores de suporte são calculados acessando os dados de treinamento. A pesquisa pelos conjuntos de termos prossegue, e o processo de enumeração é repetido até os valores de suporte, para todos os conjuntos de termos em $\mathcal{D}$, serem finalmente calculados.

Obviamente, o número de regras aumenta exponencialmente com o tamanho do vocabulário (i.e., o número de termos distintos em $\mathcal{D}$) e restrições de custo computacional devem ser impostas durante a extração de regras. Tipicamente, a espaço de pesquisa para regras é restringido a partir de poda de regras que não aparecem frequentemente em $\mathcal{D}$ (i.e., abordagem de suporte mínimo). Enquanto tais restrições fazem a extração de regra factível, elas também levam a perdas nos modelos de classificação, uma vez que algumas regras são podadas e não são incluídas em $\mathcal{R}$.

2.3 Predição de Sentimentos

Uma vez que o modelo de classificação $\mathcal{R}$ é extraído a partir de $\mathcal{D}$, regras são coletivamente utilizadas para mensurar os sentimentos das novos comentários que chegam através de $\mathcal{T}$. Basicamente, o modelo é interpretado como uma votação, na qual cada regra $\{\mathcal{X} \rightarrow s_i\} \in \mathcal{R}$ é um voto dado por $\mathcal{X}$ para o sentimento s_i. Dada um comentário $t \in \mathcal{T}$, uma regra $\mathcal{X} \rightarrow s_i$ somente é considerada como um voto válido se esta regra é aplicável para t.

Definição 3. Uma regra $\{\mathcal{X} \rightarrow s_i\} \in \mathcal{R}$ é dita ser aplicável para o comentário $t \in \mathcal{T}$ se $\mathcal{X} \subseteq t$. Ou seja, se todos termos em $\mathcal{X}$ estão presentes em t.

Nem toda regra em $\mathcal{R}$ é aplicável a um comentário específico $t \in \mathcal{T}$. Eventualmente, o modelo pode conter muitas regras que não são aplicáveis a nenhum comentário em $\mathcal{T}$. Estas regras são ditas inúteis, e o conjunto de todas as regras inúteis em $\mathcal{R}$ são denotas como $\mathcal{R}_\emptyset$.

Denotamos como $\mathcal{R}_t$ o conjunto de todas as regras em $\mathcal{R}$ que são aplicáveis para o comentário $t \in \mathcal{T}$. Assim, somente e todas as regras em $\mathcal{R}_t$ são consideradas como votos válidos quando estiverem mensurando os sentimentos no comentário t. Portanto, para m comentários futuros em $\mathcal{T} = \{t_1, t_2, \ldots, t_m\}$ o modelo de classificação $\mathcal{R}$ pode ser decomposto como $\{\mathcal{R}_{t_1} \cup \mathcal{R}_{t_2} \cup \ldots \cup \mathcal{R}_{t_m} \cup \mathcal{R}_\emptyset\}$. Regras em $\mathcal{R}_\emptyset$ representam um desperdício de recurso computacional, e podem poluir o modelo de classificação com informações irrelevantes, idealmente $|\mathcal{R}_\emptyset| = 0$.

Além disso, denotamos como $\mathcal{R}_t^{s_i}$ o subconjunto de $\mathcal{R}_t$ contendo apenas regras predizendo o sentimento s_i. Votos em $\mathcal{R}_t^{s_i}$ têm pesos diferentes, dependente da confiança das regras correspondentes. É calculada a média dos votos ponderados, por $\theta(\mathcal{X} \rightarrow s_i)$, para o sentimento s_i, dando uma pontuação para o sentimento s_i a respeito ao comentário t, como mostrado na Equação 2:

$$s(t, s_i) = \frac{\sum \theta(\mathcal{X} \rightarrow s_i)}{|\mathcal{R}_t^{s_i}|} \tag{2}$$

Finalmente, os pontos são normalizados, conforme expresso pela função $\hat{p}(s_i|t)$, como mostrado na Equação 3. A função de pontuação estima a probabilidade do sentimento s_i como a atitude implícita no comentário t.

$$\hat{p}(s_i|t) = \frac{s(t, s_i)}{\sum_{j=0}^{k} s(t, s_j)} \tag{3}$$

2.3.1 Extração de Regras *Online* com Projeção de Dados

Até agora discutimos a extração de regras *offline*, entretanto, a extração de regras *online* oferece várias vantagens. Uma dessas vantagens é que os classificadores se tornam capazes de extrair eficientemente regras a partir de $\mathcal{D}$ sem a aplicação de poda baseada em suporte. A ideia por trás de extração de regras *online* é evitar completamente a extração de regras inúteis, projetando os dados de treinamento sob demanda [2]. Mais especificamente, a extração de regras é adiada até que um comentário $t \in \mathcal{T}$ é dado. Em seguida, os termos em t são utilizados como um filtro que configura os dados de treino em $\mathcal{D}$ de uma maneira que apenas regras que são aplicáveis a t podem ser extraídas. Este processo de filtragem produz um conjunto de treinamento projetado, denotado por $\mathcal{D}_t$, que contém apenas termos que estão presentes no comentário t. A projeção de treino sob demanda assegura que $|\mathcal{R}_\emptyset| = 0$, evidenciando que apenas as regras inúteis não estão incluídas no modelo de classificação $\mathcal{R}$.

2.3.2 Estendendo Modelos de Classificação Dinamicamente

Com a extração de regras *online*, nós estendemos o modelo de classificação $\mathcal{R}$ dinamicamente à medida que os comentários em $\mathcal{T}$ são processados. Inicialmente $\mathcal{R}$ está vazio, um submodelo $\mathcal{R}_{t_i}$ é anexado a $\mathcal{R}$ a cada momento que o classificador processa um novo comentário t_i. Assim, após o processamento de uma sequência de m comentários $\{t_1, t_2, \ldots, t_m\}$, o modelo $\mathcal{R}$ é $\{\mathcal{R}_{t_1} \cup \mathcal{R}_{t_2} \cup \ldots \cup \mathcal{R}_{t_m}\}$.

Produzir um submodelo $\mathcal{R}_t$ envolve extração de regras a partir de $\mathcal{D}_t$. Esta operação tem um custo computacional significativo, uma vez que é necessário executar vários acessos a $\mathcal{D}$. Diferentes comentários em $\mathcal{T} = \{t_1, t_2, \ldots, t_m\}$ podem exigir diferentes submodelos $\{\mathcal{R}_{t_1}, \mathcal{R}_{t_2}, \ldots, \mathcal{R}_{t_m}\}$, mas diferentes submodelos podem compartilhar algumas regras (i.e., $\{\mathcal{R}_{t_i} \cap \mathcal{R}_{t_j}\} \neq \emptyset$). Neste caso, a memorização é muito eficaz para evitar a replicação de trabalho, reduzindo o número de operações de acesso a dados.

Dessa forma, antes de extrair a regra $\mathcal{X} \rightarrow s_i$, o classificador verifica se esta regra já está em $\mathcal{R}$. Se uma entrada é encontrada com a chave correspondente a $\{\mathcal{X}, s_i\}$, então a regra em $\mathcal{R}$ é utilizada ao invés de extraí-la a partir de $\mathcal{D}_t$. Se ela não for encontrada, a regra é extraída a partir de $\mathcal{D}_t$ e depois ela é inserida em $\mathcal{R}$. As principais etapas deste processo são resumidas no Algoritmo 1.

Algorithm 1 Extração de Regras *Online*

Require: comentário $t \in \mathcal{T}$ e $\mathcal{D}$
Ensure: $\mathcal{R}_t$ e $\mathcal{R}$

1: $\mathcal{D}_t \Leftarrow \mathcal{D}$ projetado de acordo com os termos em t
2: $\mathcal{R}_t \Leftarrow$ regras $\{\mathcal{X} \rightarrow s_i\} \notin \mathcal{R}$, extraídas a partir de $\mathcal{D}_t$
3: $\mathcal{R} \Leftarrow \mathcal{R}_t \cup \mathcal{R}$

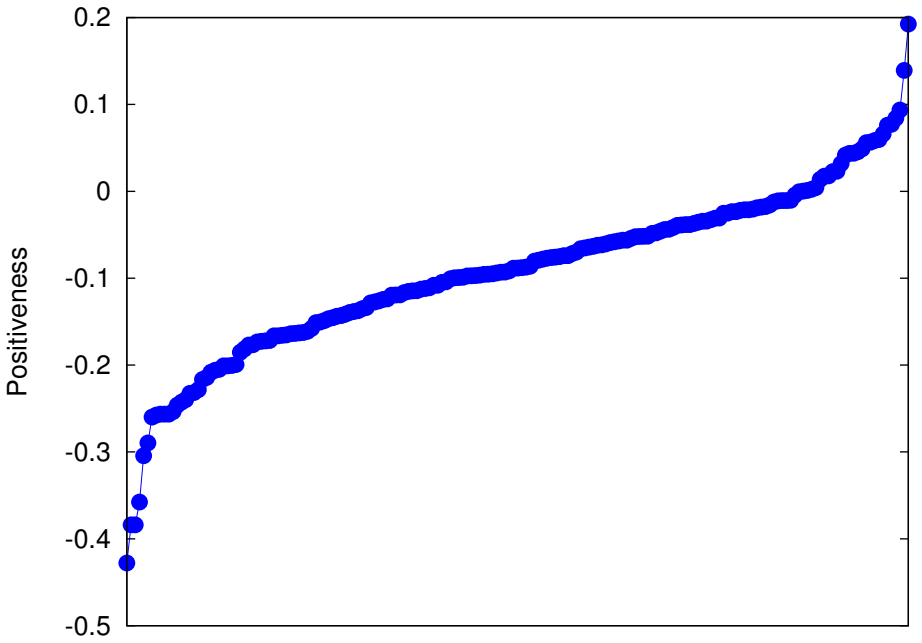

Figura 1: Distribuição de Sentimento.

3 Estudo de Caso

3.1 Preparação dos Dados

A base utilizada para a avaliação apresentada a seguir, tem 232 items comentáveis, dos quais 188 foram de fato comentados por 1020 comentários enviados por 227 usuários. Foi realizada a rotulação de um conjunto de comentários, de forma a associar rótulos positivos ou negativos aos comentários. A rotulação seguiu uma amostragem dos comentários disponíveis, resultando em um conjunto de treino composto por 187 comentários positivos, e 184 comentários negativos.

3.2 Resultados Preliminares

Nesta seção apresentamos resultados preliminares após a execução do algoritmo discutido na Seção 2.

A Figura 1 mostra que a maioria dos itens comentáveis está associada com opiniões negativas. A Tabela 1 mostra os extremos, com o item comentável mais negativo e mais positivo. Item 221:

- Art. 50. As infrações realizadas por pessoas jurídicas de direito privado às normas previstas nesta Lei ficam sujeitas às seguintes sanções administrativas aplicáveis por órgão competente:
  - IV – bloqueio dos dados pessoais

Item 46:

Tabela 1: Itens Comentáveis.

ID	Positividade	Comentário
221	-0.427881	Termo vago e que implica a impossibilidade da livre realização de atividade econômica. Sugerimos sua retirada sob pena de ser objeto de questionamento judicial.
155	0.0840085	ele traz um conteúdo importante qual seja a cooperação internacional que ao meu ver e algo imprescindível dado o fato da realidade que vivemos hoje sobre diversos crimes internacionais

- IV – princípio do livre acesso, pelo qual deve ser garantida consulta facilitada e gratuita pelos titulares sobre as modalidades de tratamento e sobre a integralidade dos seus dados pessoais;
  - paragrafo 1^o Os órgãos públicos darão publicidade às suas atividades de tratamento de dados por meio de informações claras, precisas e atualizadas em veículos de fácil acesso, preferencialmente em seus sítios eletrônicos, respeitando o princípio da transparência disposto no inciso VI.

4 Próximos Passos

Há três próximos passos imediatos. O primeiro é trabalhar com a base completa, que recentemente foi expandida com dados dos PDFs. O segundo ponto é discutir uma taxonomia que seja capaz de estabelecer os aspectos de interesse. O terceiro passo é realizar a visualização do itens comentáveis sob o ponto de vista do sentimento expresso pelos comentaristas.

Referências

[1] Roberto Lourenco Jr., Adriano Veloso, Adriano Pereira, Wagner Meira Jr., Renato Ferreira, and Srinivasan Parthasarathy. Economically-efficient sentiment stream analysis. In *The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 637–646, 2014.

[2] Adriano Veloso, Wagner Meira Jr.. Demand-Driven Associative CLassification. In *Springer Briefs in Computer Science*, 2011.

[3] Mohammed Zaki, Srinivasan Parthasarathy, Mitsunoru Ogihara. New Algorithms for Fast Discovery of Association Rules. In *The 3rd International Conference on Knowledge Discovery and Data mining*, 283–286, 1997.