Identificação de aspectos de candidatos eleitorais em comentários de notícias

Leonardo Augusto Sápiras, Karin Becker

Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

{lasapiras,karin.becker}@inf.ufrgs.br

Resumo. *É interessante identificar como os candidatos eleitorais são percebidos e avaliados pela população em relação a questões relacionadas com o seu cotidiano, tais como saúde, educação e segurança. A população tem crescentemente expresso sua posição na internet, através de fóruns, comentários, ou redes sociais. Obter e classificar esse conteúdo de opinião não é uma atividade simples, mas pode ser realizada através de técnicas de Mineração de Opinião. Este artigo apresenta os resultados parciais de uma pesquisa sobre mineração de opiniões em nível de aspecto, usando como fonte de dados comentários expressos sobre notícias. Ao contrário de revisões de produtos, este tipo de fonte de dados não permite pressupor nem a existência de opiniões, nem seus alvos. O artigo descreve um estudo de caso envolvendo as eleições municipais de São Paulo de 2012, o qual focou na extração dos aspectos alvo de opiniões emitidas em comentários de leitores um jornal on-line. Nossos melhores resultados foram obtidos nos experimentos com técnicas de co-ocorrência (EMIM e phi-squared).*

Abstract. *It is usefull to determine how electoral candidates are evaluated by the population on issues related to their daily lives (e.g. health, education, security). People is increasingly sharing their opinions on the internet through forums, comments, or social networks. The non-trivial problem of identifying and classifying this opinionated content is the concern of Opinion Mining. This paper presents partial results of a research on aspect-based opinion mining, using comments expressed about news as data source. Unlike product reviews, this type of data source does not allows assumptions about opinionated content, nor opinion targets. The paper describes a case study involving the 2012 municipal elections of São Paulo, and it is focused on the extraction of the target aspects of the opinions expressed by online newspaper readers. Our best results were obtained in the experiments with techniques of co-occurrence (EMIM and phi-squared).*

1. Introdução

A mineração de opinião é uma área de estudo que analisa opiniões, sentimentos, emoções de pessoas sobre entidades e seus aspectos, combinando técnicas de mineração de dados e de processamento de linguagem natural [Liu 2012]. Uma das áreas de aplicação mais consolidadas na mineração de opinião é a revisão de produtos e serviços [Tsytsarau and Palpanas 2012]. Inicialmente a análise da subjetividade era realizada em nível de documento, procurando a opinião sobre a entidade alvo da revisão.

Quando um documento possui várias opiniões, estas são agregadas de acordo com alguma função que identifique a opinião prevalecente (e.g. diferença entre opiniões positivas e negativas). Como evolução, a mineração de opinião em revisão de produtos passou a ser realizada em níveis mais detalhados, como o de aspecto [Hu and Liu 2004, Guo et al. 2009, Lin and He 2009, Qiu et al. 2011, Liu et al. 2013]. Com este nível de detalhe, é possível identificar que, por exemplo, sob o aspecto qualidade, a opinião de um produto alvo da revisão é considerada positiva, mas sob o aspecto preço, a opinião é negativa.

Os trabalhos em mineração de opinião vêm se estendendo a fontes de dados menos estruturados, tais como Twitter [Tumasjan et al. 2010], redes sociais e blogs [Castellanos et al. 2011], ou notícias [Kim and Hovy 2006]. Nestas fontes, as tarefas de encontrar o conteúdo de opinião, e o seu alvo, são bem mais complexas. Ao contrário de revisões de produtos, onde o objeto da revisão é a entidade alvo da opinião, documentos nestas mídias podem conter opiniões sobre múltiplas entidades, sobre aspectos específicos destas, ou mesmo, podem não conter nenhuma opinião. Desconhecemos trabalhos que visem identificar opinião em nível de aspecto em fontes não estruturadas como redes sociais ou comentários gerados por usuários.

Este artigo apresenta o desenvolvimento de um estudo de caso que busca identificar em comentários sobre notícias políticas, as entidades alvo e seus aspectos específicos sobre os quais opiniões são expressas. A plataforma eleitoral de candidato inclui propostas relevantes para a população em diversas áreas, tais como saúde, educação, segurança, que são exploradas durante a campanha com o intuito de angariar votos. Em nosso trabalho, entendemos isso como os aspectos de um candidato. Assim, partimos da premissa que além da percepção global de um candidato, é possível identificar o sentimento do público em relação a aspectos específicos deste. Por exemplo, deseja-se poder identificar que a percepção do público sobre um candidato X em relação à saúde é mais positiva que a do candidato Y, mas no que se refere à educação, a percepção é mais negativa.

Este trabalho complementa uma pesquisa em andamento, que busca determinar indicadores de evolução de sentimento sobre candidatos a eleições [Tumitan and Becker 2014]. No estudo de caso relatados em [Tumitan and Becker 2013], assume-se como entidades os candidatos a eleições, e identificam-se em comentários escritos na língua portuguesa, opiniões positivas ou negativas sobre estes. O presente artigo estende este estudo de caso, buscando técnicas que possam identificar se as opiniões são sobre os candidatos em geral (e.g. candidato X), ou sobre algum aspecto destes (e.g. saúde, educação). De acordo com a análise de um corpus de comentários sobre notícias políticas, o estudo traça estratégias para resolver o desafio de identificação de opinião em nível de aspecto. Então, desenvolve experimentos que comparam técnicas consolidadas em revisão de produtos, mas que não foram exploradas em contextos não tão bem estruturados. Os resultados apontam que abordagens baseadas em co-ocorrência são promissoras quanto à identificação de aspectos.

O restante deste artigo está estruturado como segue: a Seção 2 descreve os trabalhos relacionados; a Seção 3 apresenta os principais aspectos do estudo de caso desenvolvido; a Seção 4 descreve os experimentos realizados e os resultados obtidos e, por fim, na Seção 5 são descritos conclusões e trabalhos futuros.

2. Trabalhos relacionados

A mineração de opiniões é detalhada em surveys como [Tsytsarau and Palpanas 2012, Liu 2012]. Ela pode ser realizada em diferentes níveis, tais como documento, sentença, ou aspecto. Este trabalho interessa-se pelo nível de aspecto. Assim, uma opinião é definida formalmente como um quíntupla (e_i, a_{ij}, s_{ijkl}, h_k, t_l), onde [Liu 2012]:

e_i : nome de uma entidade;
a_{ij} : aspecto da entidade e_i. Um aspecto também pode ser denominado tópico;
s_{ijkl} : opinião sobre aspecto a_{ij} da entidade e_i;
h_k : entidade que expressa a opinão, também chamado de fonte de opinião;
t_l : tempo no qual a opinião foi expressa por h_k.

A opinião expressa s_{ijkl} sobre uma entidade ou aspecto, é medida em termos de uma polaridade, podendo ser classificada em classes, tais como: positiva, negativa ou neutra [Tsytsarau and Palpanas 2012]. As opiniões podem ser expressas de diferentes formas, sendo que a maioria dos trabalhos conseguem tratar adequadamente opiniões regulares, diretas e explícitas (e.g. "O candidato X é bom") [Liu 2012]. Outras formas de expressão são bem mais complexas (e.g. comparativas, implícitas), necessitando avançadas análises sintáticas e semânticas Um aspecto pode ser caracterizado através de um conjunto de termos ou expressões utilizados para expressá-lo. Na mineração de opiniões, geralmente entidades e seus aspectos correspondem a substantivos ou locuções substantivas [Hu and Liu 2004, Castellanos et al. 2011].

Um dos trabalhos pioneiros na mineração de opinião em nível de aspecto foi [Hu and Liu 2004], o qual propõe o uso de regras de associação para encontrar co-ocorrências frequentes de substantivos usados em revisões de um mesmo tipo de produto. Outros modelos mais sofisticados de co-ocorrência, tais como LSA (Latent Semantic Analysis) [Guo et al. 2009] e LDA (Latent Dietrich Analysis) [Lin and He 2009], também foram propostos com o mesmo fim. Dependências sintáticas são exploradas em [Qiu et al. 2011, Liu et al. 2013] a fim de melhorar o reconhecimento do alvo da opinião, e a polarização de sentenças, mas pressupõem a existência de bons analisadores sintáticos, os quais não estão disponíveis para a língua portuguesa. Estas abordagens funcionam bem em revisões de produtos, onde o aspecto alvo da opinião é explícito, geralmente único e que pertence a um domínio específico (e.g. cinema, computador), diferente de fontes de dados menos estruturadas, onde o conjunto de documentos pode apresentar opiniões sobre aspectos relacionadas a domínios que não tenham relação entre si.

LCI [Castellanos et al. 2011] e Observatório da Web[1] são ambientes de monitoração de tweets, sendo que apenas o primeiro aborda a análise de opiniões. Nestes, as entidades alvo são identificadas pelos termos usados para filtrar os tweets, e os aspectos são termos frequentes agrupados por diferentes medidas de similaridade ou relevância. Não foram encontrados trabalhos que se propõem a identificar aspectos em comentários de notícias.

3. Estudo de caso

3.1. Contexto e Objetivo

O estudo de caso apresentado neste artigo insere-se em uma pesquisa mais ampla, na qual se busca desenvolver técnicas para realizar mineração de opinião em nível de aspecto em

[1] http://www.observatorio.inweb.org.br/eleicoes2012/destaques/

fontes de dados menos estruturadas, como comentários, tweets, blogs ou jornais. Neste estudo de caso, complementamos a pesquisa já relatada em [Tumitan and Becker 2013], a qual identifica em comentários de leitores de um jornal on-line (Folha de São Paulo), a opinião expressa em relação a candidatos a eleições. Foram consideradas as eleições municipais de 2012 da cidade de São Paulo, e os três candidatos a prefeito mais comentados. O trabalho anteriormente realizado utiliza como fonte de opiniões os comentários escritos em português que contenham menções explícitas a estes candidatos. A análise foi realizada em nível de sentença, porque um comentário pode conter opiniões sobre diversos candidatos.

O estudo de caso relatado no presente artigo trata especificamente da identificação de aspectos da entidade, os quais são o real alvo da opinião. Foram considerados neste estudo de caso dois aspectos frequentemente associados a um plano eleitoral, os quais são relevantes à população: saúde e educação. O objetivo do estudo de caso é apresentar um conjunto de experimentos, desenvolvidos na busca da identificação das técnicas mais adequadas, e sobre quais tipos de documentos deveriam ser aplicadas.

Com os resultados deste estudo de caso, planejamos em uma etapa posterior da pesquisa poder avaliar não apenas a percepção global do candidato, mas a percepção em relação a aspectos específicos (e.g. a percepção de X no tocante a saúde é mais positiva que a do candidato Y no mesmo aspecto).

Com o presente estudo de caso, buscamos respostas às seguintes questões: (i) É possível detalhar a opinião sobre um candidato em termos de aspectos relevantes? (ii) Deve-se realizar a identificação dos aspectos levando em conta apenas as notícias, os comentários, ou a combinação de ambos?

3.2. Análise do corpus

O corpus utilizado consiste de notícias sobre as eleições municipais de São Paulo de 2012 com os respectivos comentários, extraídos da Folha de São Paulo[2]. O processo de extração de notícias e comentários é descrito com maiores detalhes em [Tumitan and Becker 2013]. O corpus é constituído de 407 notícias, associadas a 14.848 comentários. Tais comentários são divididos em 79.752 sentenças.

A pesquisa considera como fonte de opiniões os comentários dos leitores como reação a uma notícia. Cada comentário pode incluir opiniões sobre um ou mais candidatos, e utiliza-se o nível de sentença para buscar frases que contenham opiniões e menções a candidatos, os quais são então considerados como alvo da opinião. Ao considerar o nível de aspecto, surge o problema de como identificar um aspecto, e o que deve ser considerado como documento para sua identificação: a notícia, o comentário ou ambos.

Assim, busca-se saber se os comentários em uma notícia sobre o aspecto saúde (e.g. "candidato X promete duplicar os postos de saúde") são representativos da opinião que os leitores têm sobre o candidato no tocante a este aspecto, pois as seguintes situações são possíveis:

- O comentário enfatiza o aspecto da notícia (e.g. "confio que X acabará com as filas nos postos");

[2]www.folha.uol.com.br

- O comentário revela uma opinião sobre uma entidade não descrita na notícia (e.g. “Y também prometeu acabar com as filas de ônibus, mas não cumpriu”);
- O comentário revela uma opinião sobre um aspecto não mencionado na notícia (e.g. “X é corrupto”, “X e Y parecem estar mais preocupados com problemas pessoais de familiares”).

Note-se que muitos comentários podem também expressar alguma opinião sobre um candidato de forma implícita ou indireta, ou seja, sem mencionar o nome do candidato, nem o aspecto (e.g. “ponho fé”, considerando a notícia exemplificada acima). Há também comentários que expressam uma opinião explícita sobre um aspecto de um candidato que nem tem relação com a notícia (e.g. “segurança nunca foi seu forte quando X estava no poder”, em uma notícia sobre postos de saúde). As diferentes situações discutidas estão ilustradas na Figura 1. Sem uma análise profunda da semântica da sentença, não é possível tratar opiniões indiretas e implícitas, razão pela qual estes comentários e respectivas opiniões não serão abordados neste trabalho.

3.3. Anotação manual e Profiling

Em busca de respostas para as questões definidas em nosso estudo de caso, foi realizado um processo de anotação manual do corpus de sentenças e notícias. Todas as 407 notícias eleitorais do corpus foram anotadas por um único anotador, que avaliou se as notícias evocavam um dos tópicos analisados (i.e. saúde ou educação). Já os comentários foram anotados em nível de sentença por 3 anotadores. Foi anotada uma amostra de 2072 sentenças, onde cada anotador deveria avaliar se a sentença evocava um dos tópicos, quais candidatos mencionava explicitamente, se expressava uma opinião, e a respectiva polaridade. Os anotadores foram orientados a basear sua avaliação apenas no conteúdo explicitamente escrito, sem usar julgamento próprio ou conhecimento do domínio político para inferir entendimento. Assumiu-se que todo comentário que contivesse pelo menos uma sentença anotada como mencionando um dado tópico, por transitividade, também mencionava o tópico. Esta relação entre sentenças e comentários resultou em uma amostra de 487 comentários anotados. O nível de concordância entre os anotadores das sentenças é apresentado na Tabela 1.

Tabela 1. Concordância entre os anotadores das sentenças

Anotadores	Aspecto	Concordância
Anotadores 1 e 2	Saúde	90,7%
	Educação	84,6%
Anotadores 1 e 3	Saúde	86,6%
	Educação	80%
Anotadores 2 e 3	Saúde	89,6%
	Educação	83%
Entre todos anotadores	Saúde	77,2%
	Educação	65,6%

Apenas foram considerados como relevantes os comentários e sentenças onde houve concordância entre os três anotadores, definindo os casos de comentários/sentenças relevantes indicados na Figura 1 com a cor verde. Com base na anotação realizada, foram quantificadas as seguintes situações, representadas através da Figura 1:

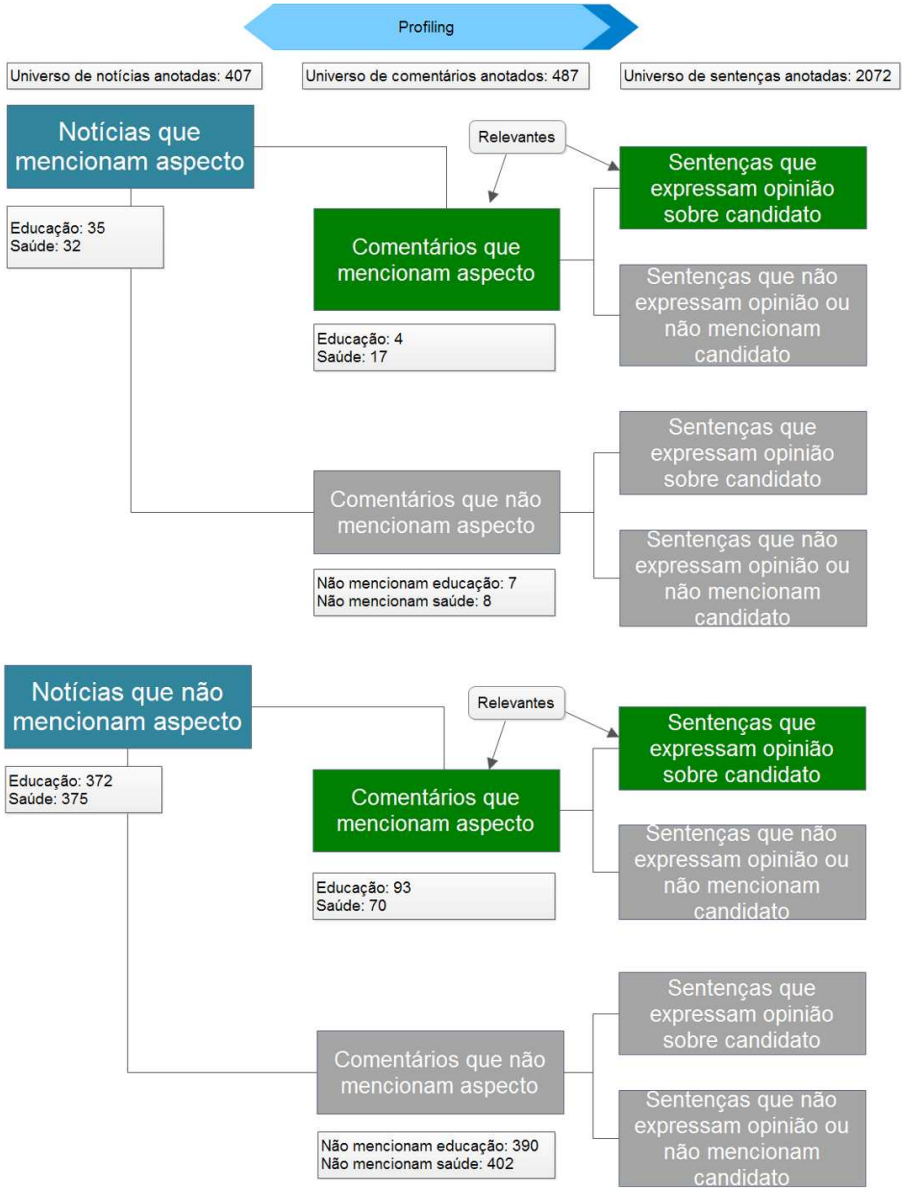

Figura 1. *Profiling* de notícias e comentários.

- Notícias que fazem menção a um aspecto;
- Notícias que não fazem menção a um aspecto;
- Comentários que fazem menção ao aspecto, e que estão relacionados a sentenças com opiniões com menções a pelo menos um candidato;
- Comentários que não fazem menção ao aspecto, e/ou que não estão relacionados a sentenças com opiniões com menções a pelo menos um candidato;

Na anotação, 35 notícias foram classificadas como relevantes para o aspecto Educação. Os anotadores entenderam dos 11 comentários anotados nestas notícias, apenas 4 comentários explicitamente expressavam opiniões sobre este aspecto. Das 32 notícias anotadas como Saúde, apenas 17 comentários foram anotados como sendo relevantes, em um conjunto de 25 comentários anotados. Por outro lado, se obteve uma quantidade maior de comentários relevantes para tais aspectos em notícias que não os

mencionavam, sendo 93 comentários para Educação (em um conjunto de 483 anotados), e 70 para saúde (em um conjunto de 472).

Ao analisar a proporção entre os comentários e as notícias com ou sem o tópico em questão, observamos uma ausência de padrões. No tocante à educação, há uma proporção maior de comentários sobre este aspecto em notícias que abordam outro tema (25%), do que em notícias específicas sobre educação (11%). Já para o aspecto saúde, esta proporção é inversa (21% e 53%, respectivamente). Além disto, observamos um número absoluto muito pequeno de comentários sobre os aspectos específicos em notícias sobre o mesmo aspecto, o qual é uma consequência do número igualmente baixo destas notícias. Assim, considerar somente estes comentários poderia comprometer os resultados. Por outro lado, há um número não negligenciável de comentários sobre os aspectos saúde e educação, relacionados a notícias totalmente diversas. Assim, optamos por extrair aspectos de comentário, sem levar em consideração o assunto abordado na notícia ao qual eles estão associados.

3.4. Abordagem proposta

Foi formulada uma abordagem para identificar os comentários que possuem menção aos aspectos analisados, com base em um conjunto de termos representativos de cada aspecto, utilizando *Palavras Sementes* e *co-ocorrência* de palavras sementes com palavras candidatas para encontrar *Termos Representativos* dos aspectos. As palavras sementes foram definidas pelos autores a partir do conhecimento do domínio, e estão listadas na Tabela 2. Elas são totalmente independentes do conjunto de Palavras Candidatas.

Tabela 2. Palavras sementes escolhidas.

	Aspecto	
	Saúde	Educação
Palavra	Saúde	Educação
	Hospital	Aula
	Médico	Aulas
	Médicos	Professores
	Vacina	Escola
	Vacinação	Escolas
	Hospitalar	-
	Doença	-
	Doenças	-
	SUS	-

A criação do conjunto de Termos Representativos foi feito com base no processo descrito na Figura 2, cujas etapas são descrita a seguir:

Geração de documentos de domínio: Para encontrar termos que representassem cada um dos aspectos, buscou-se notícias classificadas pelo jornal como sendo sobre Saúde e Educação. Foi realizada a extração de um corpus diferente de notícias, contendo 1000 notícias classificadas pela Folha de São Paulo com o rótulo Educação, e 1000 notícias com o rótulo Saúde.

Geração de palavras candidatas: De cada um destes *corpora* foram extraídos todas as palavras existentes, junto com suas respectivas frequências e classes gramaticais

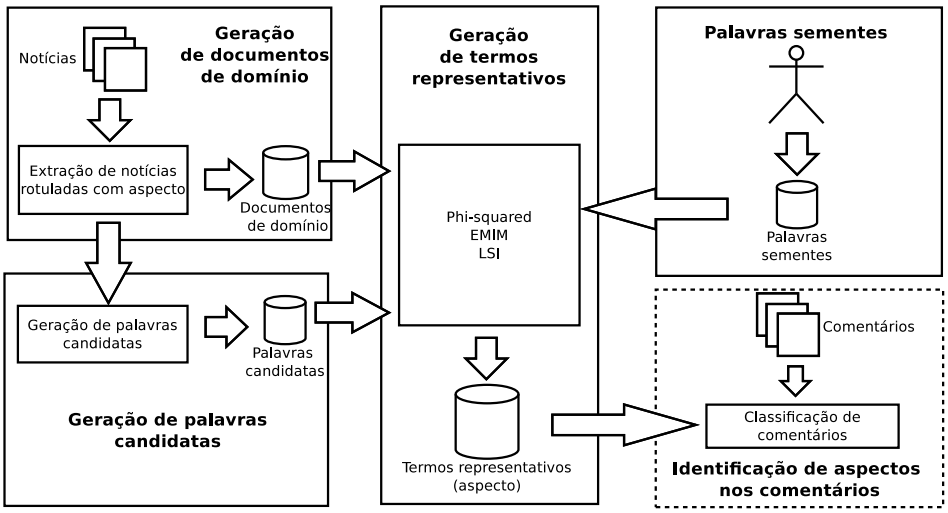

Figura 2. Processo de identificação de aspectos.

(*part-of-speech tags*). Para isso, foi utilizada a ferramenta NLTK [3]. *Stop words* foram eliminadas. Para a criação do conjunto de palavras candidatas de cada aspecto, utilizamos apenas substantivos e termos que apareciam no respectivo corpus (e.g. notícias sobre saúde) e que não apareciam no outro (e.g. notícias sobre educação). Uma proposta de melhoria futura para esta etapa é utilizar também as palavras que co-ocorram até um limite de vezes em ambos conjuntos.

Geração dos termos representativos: Utilizando uma combinação entre técnicas baseadas em frequencia, dependência ou independência de termos, foram selecionadasas palavras mais representativas para um determinado aspecto. As tecnicas de co-ocorrência usadas foram Indexação Semântica Latente [Deerwester et al. 1990], Phi-squared [Church and Gale 1991] and Expected Mutual Information Measure [Church and Hanks 1990].

Com posse dos termos representativos dos aspectos Saúde e Educação, nossa abordagem verifica quais são os comentários das notícias eleitorais que possuem tais termos em seu conteúdo.

4. Experimentos

Esta seção descreve as técnicas experimentadas para identificar os aspectos nos comentários sobre notícias políticas, bem como a avaliação dos resultados obtidos. Para isso foram implementadas abordagens utilizando co-ocorrência, que tinham como objetivo descobrir quais das palavras candidatas poderiam ser consideradas palavras representativas para um dado aspecto. Também é apresentada a avaliação de resultados para o conjunto de palavras sementes.

Durante a realização dos experimentos, todas as palavras passaram por um processo de limpeza, onde foram transformadas para minúsculo e os acentos foram removidos (e.g. “Vacinação” foi transformada em “vacinacao”). A razão é que muitos comentários foram escritos com erros de ortografia.

[3]Natural Language Toolkit - http://www.nltk.org/

4.1. Palavras Sementes

A primeira abordagem para determinar qual deveria ser o conjunto de termos representativos, considerou como tal conjunto apenas as palavras sementes. Logo, se um comentário possuir em seu conteúdo uma ou mais palavras sementes, ele passa a ser classificado como um comentário que menciona aspecto. Nas abordagens descritas a seguir, se verificou se os comentários possuíam em seu conteúdo um ou mais termos representativos de um aspecto.

4.2. Expected Mutual Information Measure

Essa abordagem é baseada na utilização da técnica Expected Mutual Information Measure (EMIM) [Church and Hanks 1990], que compara a probabilidade de observar duas palavras, *x* e *y*, junto com a probabilidade de observá-las independentemente. EMIM é definida através da Equação 1 onde, *a* é número de vezes que as palavras *x* e *y* co-ocorrem em um documento; *b* é o número de vezes que *x* ocorre em um documento e *y* não ocorre; *c* é o número de vezes que *y* ocorre é *x* não e, *d* é o número de vezes que nem *x* nem *y* ocorrem um documento. Em nossa abordagem, *x* representa uma palavra candidata e *y* representa uma palavra semente.

$$EMIM_{(x,y)} = \log_2 \frac{a(a+b+c+d)}{(a+b)(a+c)} \tag{1}$$

Caso o resultado de EMIM seja maior que zero, *x* passa a ser considerado um termo representativo. Experimentos preliminares demonstraram resultados insatisfatórios utilizando apenas esse critério. Buscando melhorar os resultados obtidos, mesclamos o resultado do EMIM com o resultado da quantidade de vezes que *x* e *y* co-ocorrem em todos os documentos de domínio. O experimento que apresentou a melhor avaliação de resultados para esta abordagem considerou a união entre o conjunto de palavras sementes com o conjunto de palavras candidatas onde *x* e *y* co-ocorrem no mínimo 10 vezes em todos os documentos de um dado aspecto.

4.3. Phi-squared

A técnica phi-squared (ϕ^2) é uma medida estatística que favorece uma alta ocorrência de eventos [Church and Gale 1991]. Phi-squared é definida através da Equação 2, cujas variáveis têm o mesmo significado das utilizadas na Equação 1.

$$\phi^2 = \frac{(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}, \text{onde } 0 \leq \phi^2 \leq 1 \tag{2}$$

Foram realizados diversos experimentos variando o resultado phi-squared com o *threshold*. Da mesma forma como realizado na abordagem EMIM, foi considerado também a quantidade de vezes que *x* e *y* co-ocorrem em todos os documentos. No final o experimento que apresentou a melhor avaliação de resultados considerou um phi-squared mínimo de 0,02 e uma co-ocorrência mínima entre *x* e *y* de 20 vezes. Por fim, o conjunto de termos representativos foi a união entre o conjunto resultante desta abordagem com o conjunto de palavras sementes do aspecto.

4.4. Indexação Semântica Latente

Esta abordagem utilizou uma técnica mista de Indexação Semântica Latente (LSI) combinada com a frequência em que as palavras candidatas e as palavras sementes co-ocorriam no corpus de notícias de domínio. O LSI implementa o modelo vetorial utilizando técnicas estatísticas para recuperar informações de um corpus com base no conteúdo semântico das consultas [Deerwester et al. 1990].

Assim, calculamos o LSI das palavras candidatas em relação aos documentos de domínio utilizando a biblioteca Python gensim[4], e realizamos diversos experimentos variando a similaridade resultante do LSI com a frequência em que as palavras sementes co-ocorriam no corpus de notícias sobre os aspectos.

4.5. Resultados

Para realizar a avaliação dos resultados, foram utilizadas as métricas de Precisão, Revocação, F-score e Acurácia [Aggarwal and Zhai 2012]. A Figura 3 apresenta um gráfico com a avaliação dos resultados das abordagens descritas acima em comentários sobre Saúde, e a Figura 4, os resultados para os comentários sobre Educação. Os experimento LSI_1 e LSI_2 consideraram como ponto de corte, LSI iguais a ou maiores a 0,75 e 0,8, respectivamente. Foram considerados como comentários relevantes em relação um aspecto apenas aqueles onde houve concordância entre todos os anotadores.

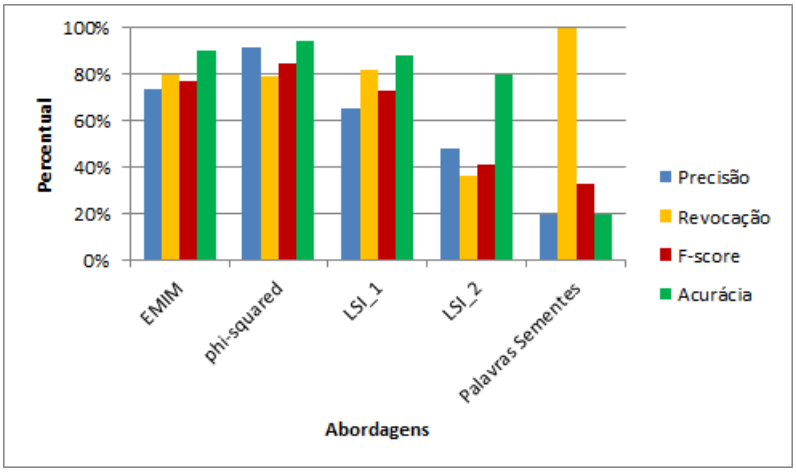

Figura 3. Avaliação de resultados para o aspecto saúde.

Com base no F-score, a abordagem que apresentou a melhor avaliação de resultado para o aspecto Saúde foi a phi-squared. Já para o aspecto Educação o melhor resultado foi obtido com a técnica EMIM. Realizamos também experimentos mesclando as técnicas EMIM e phi-squared, mas não houve melhoria nos resultados finais.

Como nós possuímos um corpus de notícias políticas rotuladas como relevantes para os aspectos analisados, realizamos também a validação de resultados usando tais notícias ao invés dos comentários. Apesar de termos descartado a utilização das notícias em nossa abordagem, o objetivo foi verificar o comportamento das técnicas em documentos que possuem uma qualidade melhor de escrita. Os resultados obtidos foram semelhantes aos realizados com os comentários, favorecendo as técnicas phi-squared e EMIM.

[4]gensim - http://radimrehurek.com/gensim/

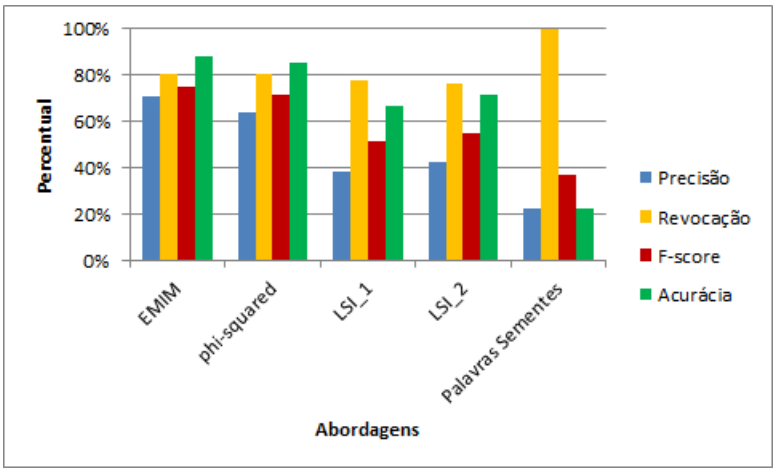

Figura 4. Avaliação de resultados para o aspecto educação.

5. Conclusões

Este artigo apresentou os resultados parciais de um estudo de caso para identificar a percepção da população sobre aspectos de candidatos eleitorais. O objetivo deste trabalho foi verificar se é possível identificar aspectos em fontes de dados fracamente estruturadas como notícias ou nos comentários. Com base no estudo de caso realizado para as eleições da cidade de São Paulo, foi possível verificar que é válido considerar apenas os comentários. Este trabalho também propôs uma abordagem para classificar tais comentários em relação aos aspectos Saúde e Educação. Experimentos realizados com diferentes técnicas demonstraram resultados satisfatórios para as técnicas de co-ocorrência EMIM e phi-squared.

No entanto, ainda é necessário realizar algumas melhorias na identificação dos aspectos, tais como a utilização de expressões idiomáticas, expressões multi-palavras e padrões léxico-sintáticos. Outra possibilidade de trabalho futuro, é o tratamento de opiniões irregulares e opiniões implícitas.

Para concluir o processo de polarização dos aspectos de candidatos eleitorais, iremos polarizar as sentenças utilizando as abordagem comparadas em [Tumitan and Becker 2014], a saber, baseada em dicionário e aprendiagem supervisionado. Dessa forma, será possível polarizar as sentenças não apenas em relação a um candidato mas também em relação aos aspectos mencionados comentário. Por fim, será necessário desenvolver a sumarização dos resultados obtidos.

Referências

Aggarwal, C. C. and Zhai, C. X. (2012). *Mining Text Data*. Springer US, Boston, MA.

Castellanos, M., Dayal, U., Hsu, M., Ghosh, R., Dekhil, M., Lu, Y., Zhang, L., and Schreiman, M. (2011). Lci: a social channel analysis platform for live customer intelligence. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, SIGMOD '11, pages 1049–1058, New York, NY, USA. ACM.

Church, K. W. and Gale, W. A. (1991). Concordances for llel text. In *Proceedings of the Seventh Annual Conference of the UW Centre for the New OED and Text Research*, pages 40–62.

Church, K. W. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Comput. Linguist.*, 16(1):22–29.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Guo, H., Zhu, H., Guo, Z., Zhang, X., and Su, Z. (2009). Product feature categorization with multilevel latent semantic association. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 1087–1096, New York, NY, USA. ACM.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '04, pages 168–177, New York, NY, USA. ACM.

Kim, S.-M. and Hovy, E. (2006). Extracting opinions, opinion holders, and topics expressed in online news media text. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, SST '06, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lin, C. and He, Y. (2009). Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 375–384, New York, NY, USA. ACM.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Liu, Q., Gao, Z., Liu, B., and Zhang, Y. (2013). A logic programming approach to aspect extraction in opinion mining. In *Proceedings of the 2013 IEEE/WIC/ACM International Conferences on Web Intelligence*, pages 276–283.

Qiu, G., Liu, B., Bu, J., and Chen, C. (2011). Opinion word expansion and target extraction through double propagation. *Computational Linguistics*, 37(1):9–27.

Tsytsarau, M. and Palpanas, T. (2012). Survey on mining subjective data on the web. *Data Min. Knowl. Discov.*, 24(3):478–514.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. In *ICWSM'10*.

Tumitan, D. and Becker, K. (2013). Tracking Sentiment Evolution on User-Generated Content: A Case Study on the Brazilian Political Scene. In *Anais do XXVIII Simpósion Brasileiro de Banco de Dados*, pages 135–144.

Tumitan, D. and Becker, K. (2014). Sentiment-based features for predicting election polls: a case study on the brazilian scenario. In *Proceedings of the 2014 IEEE/WIC/ACM International Conferences on Web Intelligence*, page 8p. IEEE Computer Society.